KATEDRA GEOINFORMATIKY
Univerzita Palackého v Olomouci

Vydavatelství
Univerzity
Palackého

POKROČILÉ ZPRACOVÁNÍ GEODAT

Karel Macků

Přírodovědecká fakulta

Katedra geoinformatiky

POKROČILÉ ZPRACOVÁNÍ GEODAT

Mgr. Karel MACKŮ, Ph.D.

Univerzita Palackého v Olomouci

2023

Tato publikace vznikla s podporou Erasmus+ Program, Jean Monnet Module.
Project No. 620791-EPP-1-2020-1-CZ-EPPJMO-MODULE UrbanDM - Data mining and analyzing of urban structures as contribution to European Union studies.

Podpora Evropské komise při tvorbě této publikace nepředstavuje souhlas s obsahem, který odráží pouze názory autorů, a Komise nemůže být zodpovědná za jakékoliv využití informací obsažených v této publikaci.

Za návrh šablony dokumentu děkuji Mgr. Jakubu Koníčkovi.
Děkuji recenzentům RNDr. Pavlíně Netrdové, Ph.D. a doc. Ing. Zdeně Dobešové, Ph.D. za podnětné připomínky k textu.

With the support of the
Erasmus+ Programme
of the European Union

Odborní recenzenti:
RNDr. Pavlína Netrdová, Ph.D.
doc. Ing. Zdena DOBEŠOVÁ, Ph.D.

1. vydání

DOI: 10.5507/prf.23.24463209
ISBN 978-80-244-6320-9 (online: iPDF)

OBSAH

ÚVOD 4
1 EXPLORATORNÍ ANALÝZA VÍCEROZMĚRNÝCH DAT 5
1.1 GRAFICKÁ EXPLORATORNÍ ANALÝZA 5
1.2 VYŠETŘOVÁNÍ ODLEHLÝCH HODNOT 8
1.2.1 Vícerozměrné vyšetřování odlehlých hodnot 11
1.2.2 Robustní metody určování odlehlých hodnot 13
1.2.3 Odlehlé hodnoty a prostor 14
2 PROSTOROVÁ STATISIKA 16
2.1 ANALÝZY BODOVÝCH DAT 17
2.2 ANALÝZY AREÁLOVÝCH DAT 23
2.2.1 Vyhlazování areálových dat 23
2.2.2 Prostorová autokorelace 28
3 METODA MONTE CARLO 32
4 PROSTOROVĚ VÁŽENÉ METODY 35
4.1 OPTIMALIZACE JÁDRA 37
4.2 VYBRANÉ PROSTOROVĚ VÁŽENÉ METODY 38
5 PROSTOROVÉ REGRESNÍ MODELY 43
5.1 GLOBÁLNÍ PROSTOROVÉ MODELY 44
5.2 LOKÁLNÍ PROSTOROVÉ MODELY 48
6 ZÁVĚR 51
7 POUŽITÉ ZDROJE 52

Pokud není uvedeno jinak, obrázky jsou vytvořeny autorem publikace K. Macků.

ÚVOD

Během výuky předmětu KGI/POGEO „Pokročilé zpracování geodat“ v rámci magisterského programu Geoinformatika a kartografie na Katedře geoinformatiky Přírodovědecké fakulty UP mají studenti pro studijní potřeby k dispozici přednáškové prezentace a podrobné manuály ke cvičením, vedených převážně v software R. Doposud však chyběl ucelený text, který by blíže vysvětloval a komentoval teoretický rámec přednášených metod. Hlavním smyslem tohoto učebního textu je vyplnit tuto mezeru a poskytnout studentům geoinformatiky a příbuzných oborů materiál, který by sloužil jako doplněk k vedeným přednáškám. Text nepokrývá celý sylabus předmětu KGI/POGEO, zaměřuje se především na témata spojená s disciplínou prostorové statistiky. Text je samozřejmě vhodný i pro další zájemce z geovědní praxe, kteří chtějí rozšířit své teoretické znalosti v oblasti metod prostorových analýz.

Cílem učebního textu je přiblížit čtenářům srozumitelnou cestou analýzy, které jsou v geoinformatické praxi méně obvyklé a často vnímané jako příliš složité a obtížně aplikovatelné. Jak je však demonstrováno dále, tyto metody mají své využití i mimo akademickou praxi a dokážou při analýze dat odhalit zajímavé prostorové chování zkoumaných jevů. V ukázkách zpracování dat byla mimo jiné použita data popisující statistické jevy ve státech Evropské unie získaná z databáze Eurostat, OECD Regional Database, Evropská agentury pro životní prostředí a dalších. Účelem je, aby studenti získali povědomí o datech a tématech týkajících se Evropské Unie a států Evropy a možnostech jejich analýzy a interpretace.

1 EXPLORATORNÍ ANALÝZA VÍCEROZMĚRNÝCH DAT

Exploratorní analýza dat (dále jen EDA) je významným krokem procesu zpracování dat, ať už prostorových nebo neprostorových. Jejím účelem je odhalit a shrnout hlavní charakteristiky zkoumaných dat, například pomocí popisu rozdělení pravděpodobnosti, základních popisných statistik, nebo identifikace odlehlých hodnot, které mohou v dalších fázích analýzy způsobovat problémy. Významný statistik John Tukey ve své práci *Exploratory data analysis* (Tukey, 1977) říká, že „*Exploratory data analysis is a detecitve work – numerical detective work or graphical detective work.*". Jeho tvrzení velmi stručně vystihuje význam této fáze zpracování dat – pomocí EDA se snažíme odhalit charakteristické chování dat, na základě kterého dále formulujeme hypotézy o příčinách pozorovaných jevů, ověřujeme předpoklady statistických konfirmačních metod a také přinášíme odůvodnění pro výběr vhodných statistických nástrojů a technik, které budou pro ověření hypotéz použity.

Do exploratorní analýzy můžeme zahrnout celou řadu numerických i grafických technik. V prvé řadě bychom mohli mluvit o aplikaci základních popisných statistik – výpočtu charakteristik polohy a variability, díky kterým lze získat základní přehled o zpracovávaných datech. Dále je vhodné zaměřit se na rozdělení pravděpodobnosti a ověřit normalitu dat, která je důležitým předpokladem velkého množství statistických analýz. Tyto úkony jsou předmětem základních statistických kurzů, nyní se je pokusíme rozšířit o další metody, zaměřené především na vícerozměrná data.

Rozšířením exploratorní analýzy o prostorovou složku dat označujeme jako exploratorní prostorovou analýzu (ESDA – Exploratory Spatial Data Analysis), která se zabývá problémem zjišťování prostorových vlastností souborů dat, kde pro každou hodnotu atributu existuje lokační údaj (Haining et al., 1998). ESDA slouží k identifikaci neobvyklých pozorování (včetně detekce chyb a prostorových odlehlých hodnot), popisu prostorových vzorů (náhodných/ shlukových/rovnoměrných) nebo pro formulování hypotéz o zkoumaných datech. ESDA lze také použít při prostorovém modelování k posouzení kvality řešeného modelu. Řada metod představených v této publikaci může být chápána také jako součást exploratorní prostorové analýzy (např. analýza autokorelace).

1.1 GRAFICKÁ EXPLORATORNÍ ANALÝZA

Významnou částí EDA je samotná vizualizace dat. Ne nadarmo se říká, že jeden obrázek vydá za tisíc slov, a vhodná vizualizace může mnohdy rychle odhalit důležité informace, které jsou v datech ukryty. Zde lze opět odkázat na Johna Tukeyho, který vizualizaci komentuje slovy „*using visualisation to find meaning in your data*" (Tukey, 1977). Věřil, že grafická prezentace informace hraje nesmírnou roli. Vhodná vizualizace může být nápomocná k pochopení struktury dat, zlepšit rozhodovací procesy založené na datech a získat objektivnější přístup k řešenému problému (Yau, 2013). Vizualizace nenahrazuje statistické testy, nicméně přináší cenné poznání, na základě kterého mohou být vzneseny nové hypotézy pro další testovaní a analýzu dat.

V případě vícerozměrných dat je nutno volit metody, které dokážou přenést informaci z n-rozměrného prostoru do takového prostředí, které je lidským vnímáním snadno přijatelné. Úskalím vizualizace vícerozměrných dat mohou být rozdílné jednotky jednotlivých veličin. Ty při společné vizualizaci způsobují dominanci takových veličin, které číselně nabývají nejvyšších absolutních hodnot (např. veličiny s teoreticky neomezeným rozsahem hodnot, jako je např. příjem domácností vs. veličiny vyjádřené relativně v procentech, kterou je třeba míra nezaměstnanosti). Úpravou pomocí standardizace se všechny veličiny převádějí na stejné bezrozměrné jednotky, kde nezáleží na původních jednotkách ani pozorovaných velikostech. Standardizované veličiny se tak stávají vzájemně srovnatelnými a použitelnými pro společnou vizualizaci a analýzu. Standardizaci lze provádět různými metodami, jako příklad uveďme nejčastěji používanou standardizaci směrodatnou odchylkou [1] nebo standardizaci rozpětím [2]:

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad [1] \qquad\qquad y_{ij} = \frac{x_{ij} - min\,(x_j)}{max(x_j) - min\,(x_j)} \qquad [2]$$

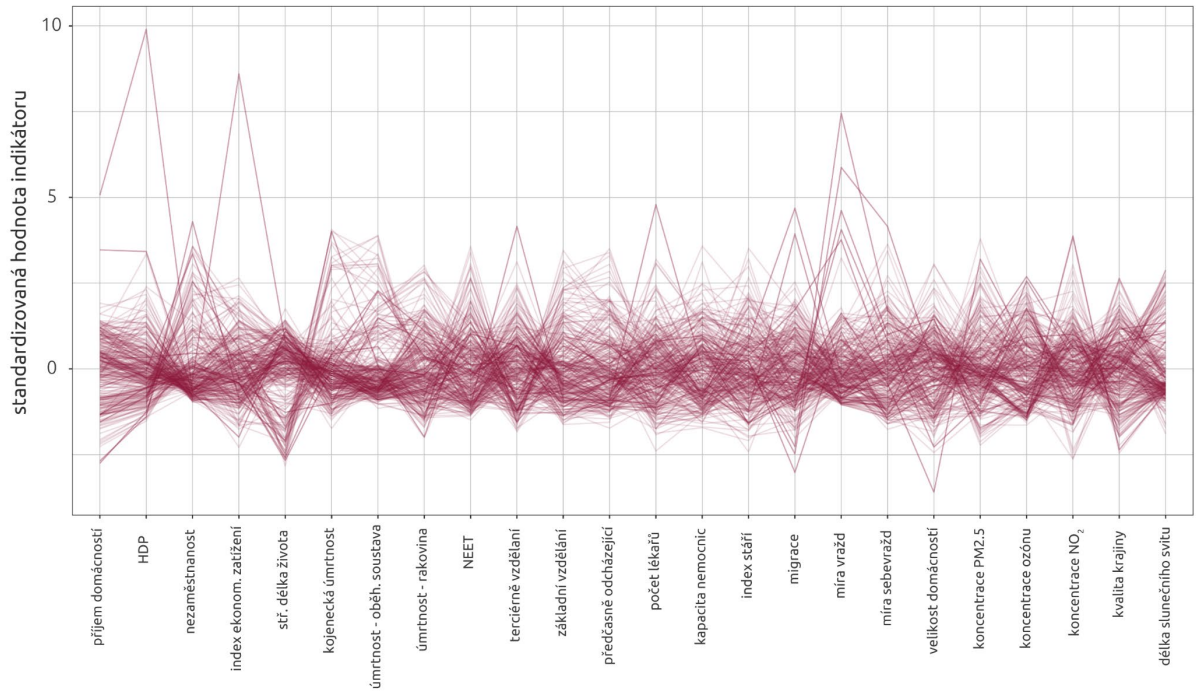
, kde y_{ij} je nová standardizovaná hodnota, x_{ij} je původní hodnota i-tého řádku j-té veličiny, $\bar{x}_j$ je výběrový průměr j-té veličiny, s_j směrodatná odchylka j-té veličiny, a min/max jsou minimální a maximální hodnota j-té veličiny.

Příkladem vícerozměrných grafických metod může být metoda paralelních os nebo metoda heatmapy, které zde budou demonstrovány na datové sadě 24 indikátorů kvality života v Evropě v roce 2015. Prostorové rozlišení je definováno administrativní klasifikací NUTS 2, datová sada zahrnuje 271 těchto administrativních jednotek především z území států EU, ale také z vybraných nečlenských zemí. Zdrojem dat je databáze Eurostat, OECD Regional Database, Evropská agentura pro životní prostředí, služba Copernicus a německá agentura Deutscher Wetterdienst. Datová sada vznikla v rámci disertační práce K. Macků, její detailní popis je k dispozici v textu práce (Macků, 2020).

Metoda *paralelních os* představena na Obr. 1 je pro svou jednoduchost velmi intuitivním nástrojem popisu vícerozměrných dat. Na ose y jsou vyneseny standardizované hodnoty indikátorů rozmístěných na ose x, které umožňují relativní porovnání profilů jednotlivých záznamů (v geoinformatické terminologii se jedná o jeden řádek atributové tabulky) a odhalení některých trendů v datech. Z Obr. 1 je patrna např. výrazná nepřímá úměra mezi HDP a mírou nezaměstnanosti (druhá a třetí osa), podobně mezi střední délkou života a mírou kojenecké úmrtnosti (pátá a šestá osa). V řadě indikátorů lze pozorovat separaci jednotlivých záznamů do několika skupin podobného charakteru – především u střední délky života, kde má řada záznamů výrazně nižší hodnoty než ostatní. Podobně kombinace indikátorů NEET[1] a podílu terciérně vzdělaných obyvatel naznačují, že v datech se budou vyskytovat minimálně dvě dobře odlišitelné skupiny regionů. Lze tedy usuzovat, že data mají potenciál pro klasifikaci, např. pomocí metody shlukové analýzy. Paralelní osy mohou také sloužit jako doplněk pro analýzu odlehlých hodnot, jelikož můžeme sledovat, ve kterých indikátorech daný záznam nejvíc vybočuje.

[1] Zkratka pro *youth Not in Employment, Education or Training* – vyjadřuje podíl mladých obyvatel ve věku 15 až 24 let, kteří jsou mimo vzdělávací systém, neabsolvují žádnou odbornou přípravu, ani nejsou zaměstnaní.

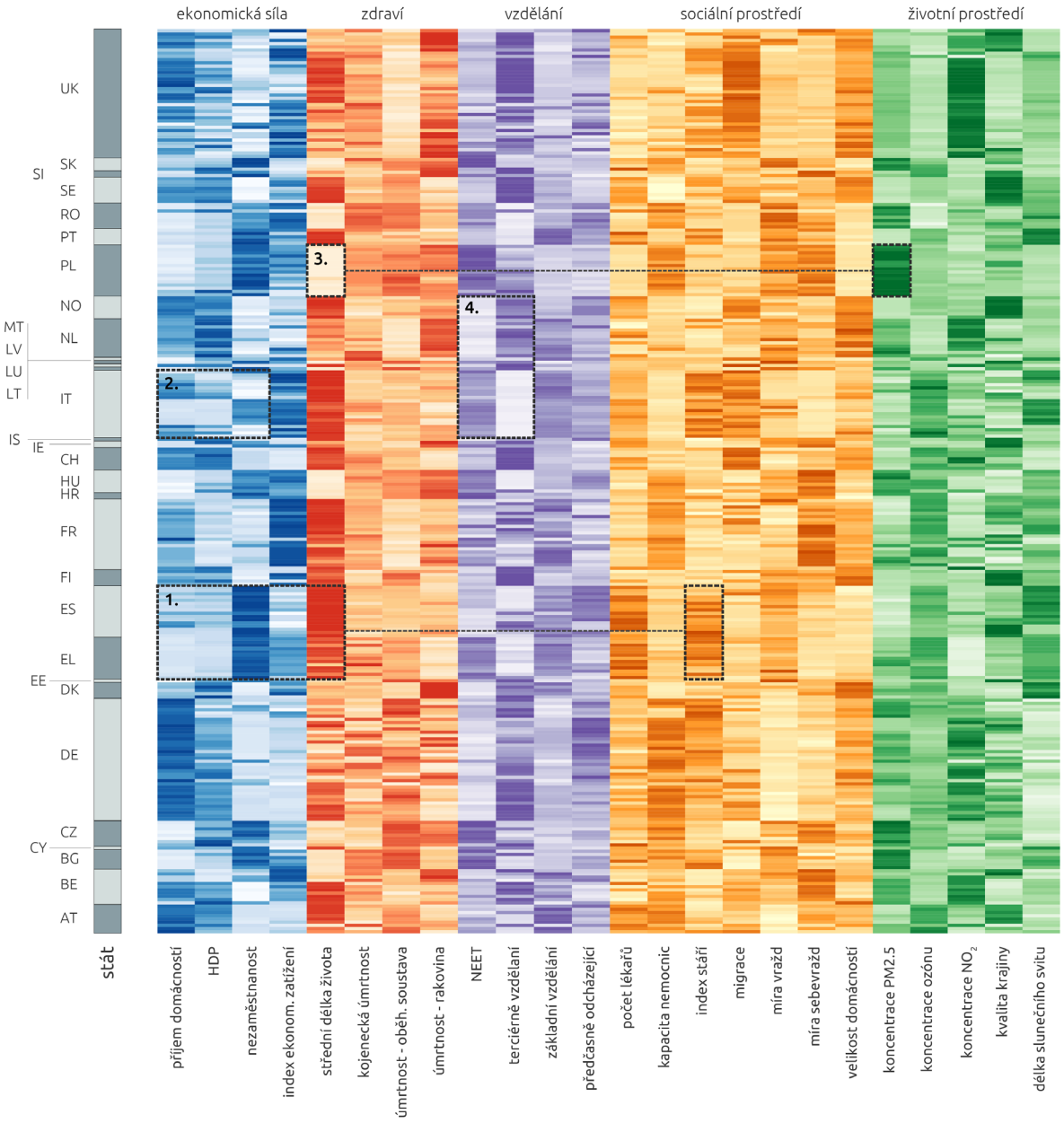
Může být vznesena námitka, že paralelní osy jsou zavádějícím typem vizualizace, jelikož svým grafickým provedením prezentují kontinuitu (spojitá linie) v případě na sebe nijak nenavazujících diskrétních témat. Nicméně, právě samotný průběh a tvar spojující linie umožňuje pozorovat podobnostmi/rozdílnosti mezi jednotlivými záznamy zkoumané datové sady. Volba pořadí sledovaných veličin nehraje roli, pořadí může být libovolně měněno, třeba za účelem seskupení tematicky blízkých veličin. Další rozšíření metody paralelních os spočívá např. v barevném rozlišení příslušnosti jednotlivých záznamů k různým kategoriím. Informace o této příslušnosti musí být předem ve vstupních datech uložena.



Obr. 1 Průběh standardizovaných hodnot indikátorů zobrazený metodou paralelních os

Metoda *heatmapy* používá pro vizualizaci standardizované hodnoty veličiny intenzitu barvy v matici o rozměrech odpovídajících počtu sledovaných veličin a počtu sledovaných záznamů. Jelikož komunikačním prostředkem je spojitá intenzita barvy, je prakticky nemožné přesně odečítat konkrétní hodnoty v buňkách. Proto je metoda vhodnější pro zobrazení obecnějšího relativního přehledu o numerických datech, a odhalení některých typických chování. První situace (1) na Obr. 2 zobrazuje záznamy, které mají stejný trend ve skupině indikátorů (nízké hodnoty HDP a příjmu domácností, spíše nižší hodnoty indexu ekonomického zatížení a vysoké hodnoty nezaměstnanosti, střední délky života a indexu stáří). Tyto záznamy NUTS 2 přísluší ke Španělsku a Řecku, kromě atributové podobnosti mají i společný geograficky aspekt, a to lokalizaci v jižní části Evropy. Podobně jako vizualizace paralelními osami podporují předpoklad, že data jsou vhodná k typizaci, dle prostorové lokalizace však také k regionalizaci. V situaci (2) se vyskytují rozdílné hodnoty indikátorů příjmu domácností, HDP a nezaměstnanosti v rámci jednoho státu, což poukazuje na silnou vnitrostátní variabilitu. V situacích (3) a (4) jsou vidět vztahy mezi dvěma indikátory v rámci jedné domény, nebo i napříč doménami, které mohou být následně popsány pomoci korelace. Konkrétně situace (3) evokuje závislost mezi koncentrací částic PM2,5 a střední délkou života, která však bude prostorově variabilní, neboť ne ve všech regionech je tento vztah tak silný

jako v případě Polska (zvýrazněném v situačním rámečku 3). Podobně můžeme pozorovat vzájemný negativní vztah mezi indikátory NEET a mírou terciérně vzdělaného obyvatelstva v situačním rámečku 4. Detailní průzkum těchto vztahů a vyhodnocení, zdali jsou skutečné nebo se může jednat o falešné korelace je už úkolem dalšího hlubšího zkoumání.



Obr. 2 Vizualizace standardizovaných hodnot indikátorů heatmapou

1.2 VYŠETŘOVÁNÍ ODLEHLÝCH HODNOT

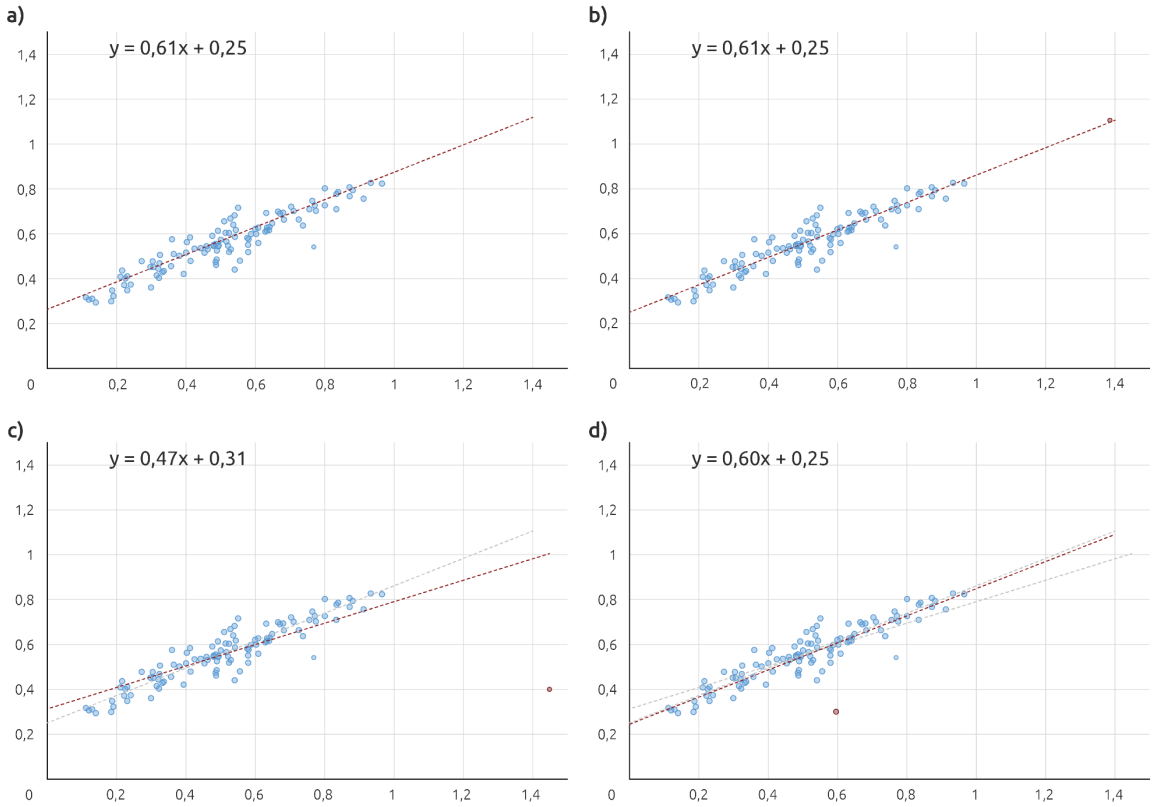
„An observation (or subset of observation) which appears to be inconsistent with the remainder of that set of data“ (Barnett & Lewis, 1978)

Reálná data velmi často obsahuji odlehlé hodnoty (outliery), tedy takové záznamy, které se výrazně liší od zbytku datového souboru. Příčiny vzniku outlierů mohou být různé: mohou vznikat chybami při sběru dat (především u veličin, jejichž měření je více technického

charakteru) nebo popisují skutečný stav. Jedná se tedy o záznamy, které jsou svými hodnotami mimořádné a zasluhují větší pozornost. Jelikož odlehlá hodnoty mohou výrazně ovlivnit spolehlivost a přesnost výsledků aplikovaných analýz, je vhodné tyto záznamy identifikovat a zvážit možnosti zacházení s nimi (Dixon, 1950). Je pak důležitou otázkou, jak s outliery dále naložit. V nejjednodušším případě záznamy s odlehlými hodnotami odstraňujeme za účelem získání co nejreprezentativnějšího vzorku. Tento přístup je vhodný např. v situaci, kdy jsme si jistí, že odlehlá hodnota je hrubou chybou. Dalším řešením je úprava odlehlé hodnoty tak, aby se více blížila reprezentativním záznamům. Příkladem takového řešení je metoda winsorizace, kdy jsou odlehlé hodnoty nahrazeny konkrétním percentilem. Touto změnou však modifikujeme datovou sadu, a nepracujeme se skutečnými záznamy. Posledním řešením je v navazujících analýzách aplikovat metody výpočtu, které dokážou částečně negativní vliv odlehlých hodnot potlačit (viz robustní metody v kapitole 1.2.2). V prvním kroku je však nezbytné uvědomit si potencionální přítomnost těchto odlehlých hodnot a následně je umět spolehlivě identifikovat. Vyšetřování odlehlých hodnot může být shrnuto do tří fází:

- identifikace odlehlé hodnoty,
- vyhodnocení podstaty odlehlé hodnoty,
- řešení odlehlé hodnoty.

Vliv odlehlých hodnot na konkrétní analýzu může být zanedbatelný, avšak někdy zcela zásadní – v závislosti na typu analýzy, množství odlehlých hodnot, a především jejich poloze ve vícerozměrném prostoru. Představme si následující dvourozměrná data, nad kterými je sestaven lineární regresní model (Obr. 3a):



Obr. 3 Vliv odlehlých hodnot na výsledek regresního modelu: červená šrafovaná linie představuje regresní model sestavený z dat, která obsahují různé outliery. Šedá linie znázorňuje regresní model ze všech předchozích případů (pro porovnání rozdílů).

V případě, že datový bod lze považovat za odlehlý v obou dimenzích a leží ve směru hlavního trendu dat, výsledný regresní model nijak nezkreslí (Obr. 3b). V případě, že se jedná o odlehlou hodnotu pouze ve vektoru prediktoru (tzv. *leverage point*), může být vliv na regresní model významný (Obr. 3c). V poslední situaci (Obr. 3d) bychom při pohledu z jednotlivých dimenzí datový bod jako odlehlý neoznačili, chápeme-li však data vícerozměrně, pravděpodobně už se o odlehlou hodnotu jednat bude. Jeho vliv na konkrétní regresní model je však v současné poloze minimální, jelikož nemá výrazný extrém ve směru prediktoru (není to *leverage point*).

Nejprve se zběžně podívejme na jednoduché možnosti identifikace odlehlých hodnot jednorozměrných dat. Vykazují-li data normální rozdělení, lze aplikovat *pravidlo tří sigma*, které očekává, že všechny hodnoty leží maximálně ve vzdálenosti 3 směrodatných odchylek od průměru (pokrývající cca 99,7 % dat). Body mimo tento interval lze považovat za odlehlé.

Jednorozměrné outliery lze identifikovat také graficky, např. metodou boxplotu. Pomocí 1,5 násobku mezikvartilového rozpětí přičteného/odečteného od horního/dolního kvartilu získáváme hraniční hodnoty, vůči kterým vymezujeme hodnoty odlehlé (Dawson, 2011). Výhodou tohoto přístupu je robustnost vůči narušení normality rozdělení, lze použít také různé variace, např. s 9. a 91. percentilem.

Dále mohou být aplikovány statistické testy pro odlehlé hodnoty: **Grubbsův test odlehlých hodnot** je statistický test, který předpokládá u dat normální rozdělení. Je zaměřený pouze na

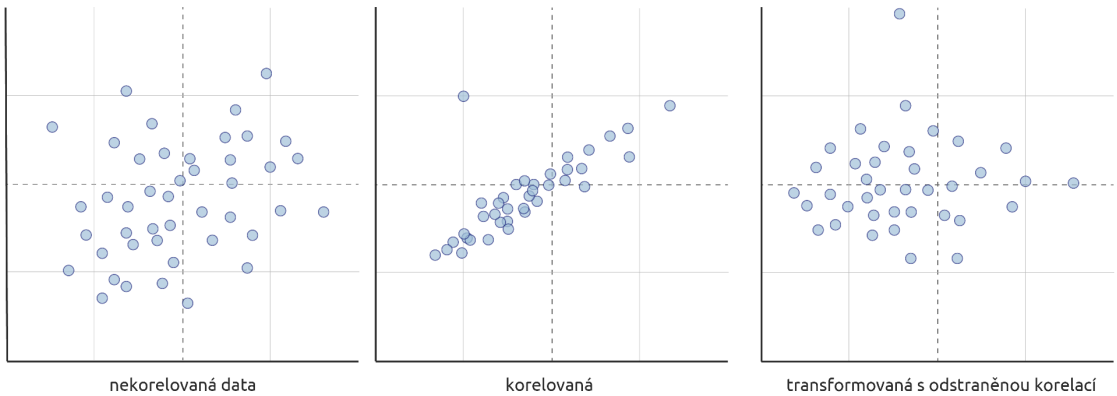
ohodnocení nejnižší a nejvyšší hodnoty sledované veličiny, počítá testové kritérium pomocí průměru a směrodatné odchylky, testové kritérium je pak porovnáno s prahovou hodnotou na stanovené hranici alfa (tabelované hodnoty vycházející z t − rozdělení). Neparametrickou alternativou ke Grubbsovu testu je **Dixonův test.** Testová statistika je založena na výpočtu variačního rozpětí, významnost zjištěné hodnoty je určena porovnání s tabulkovou kritickou hodnotou pro příslušné n výběrového souboru a zvolenou α. Dixonův test je vhodný i pro malé soubory (i do 10 hodnot).

1.2.1 Vícerozměrné vyšetřování odlehlých hodnot

U vícerozměrného vyšetřování odlehlých hodnot obecně čelíme několika problémům. Především u vícerozměrného souboru nestačí vyšetřovat každou veličinu samostatně, např. sestavit sadu boxplotů pro všechny veličiny. Potřebujeme sofistikovanější metodu, která dokáže vyhodnotit i vzájemné vztahy napříč vícerozměrným prostorem. Za takových podmínek může být datový záznam outlierem navzdory naší intuici (viz příklad 3d v úvodu kapitoly).

Mahalanobisova vzdálenost

Metrika vzdálenosti popsána P. Mahalanobisem (Mahalanobis, 1936) je nejklasičtějším nástrojem pro identifikaci odlehlých hodnot ve vícerozměrném prostoru. Tato metoda vychází z obecného konceptu sledování vzdálenosti mezi dvěma záznamy ve vícerozměrném prostoru. Běžně používáme např. Euklidovskou vzdálenost, potažmo z nějakého důvodu můžeme použít libovolnou jinou metriku. Zmíněná Euklidovská vzdálenost funguje pro popsání vzdálenosti mezi dvěma body dobře, pokud nejsou vstupní veličiny silně korelovány – vnímání vzdálenosti mezi dvěma body je ve všech směrech stejné. Jakmile však veličiny vykazují silnou korelaci, vzniká v datech přirozený směrový trend. Díky němu pomyslný krok o jednotku v jednom směru (třeba ve směru hlavní korelace, probíhající přibližně osou prvního a třetího kvadrantu) není zcela srovnatelný s krokem o jednotku v jiném směru (třeba ve směru osy druhého a čtvrtého kvadrantu). Tento vliv je potřeba nějak eliminovat, což dokáže právě Mahalanobisova vzdálenost (viz Obr. 4).



Obr. 4 Transformace korelovaných dat Mahalanobisovou vzdáleností

Základní princip metody spočívá ve výpočtu rozdílu od střední hodnoty, který se dělí kovarianční maticí. Tento krok můžeme považovat za ekvivalent k vícerozměrné standardizaci dat: pokud jsou data významně korelována, kovariance bude vysoká a jejím dělením se zredukuje vzdálenost. Úprava tedy transformuje sloupce na nekorelované proměnné, standardizuje hodnoty tak, aby se jejich rozptyl rovnal 1. Závěrem se spočítá Euklidovská vzdálenost. Pokud by byla data zcela nekorelovaná, Mahalanobisova vzdálenost by byla rovna Euklidovské.

Vzdálenost lze počítat způsobem „každý s každým", tedy klasickou matici vzdáleností mezi jednotlivými záznamy, nebo počítat vzdálenost každého záznamu od centroidu, tedy vůči optimálnímu „průměrnému" záznamu. Na tomto principu pak určujeme odlehlé hodnoty, jelikož hledáme záznamy, které se nejvíce liší od centroidu:

$$d_{M(i)} = \sqrt{(x_i - \bar{x}_n)^T C^{-1} (x_i - \bar{x}_n)}$$

$$d_{M(i)} = \sqrt{\begin{pmatrix} x_i - \bar{x}_n \\ y_i - \bar{y}_n \end{pmatrix}^T C^{-1} \begin{pmatrix} x_i - \bar{x}_n \\ y_i - \bar{y}_n \end{pmatrix}} \quad \textit{(příklad pro dvourozměrná data)}$$

, kde $d_{M(i)}$ je Mahalanobisova vzdálenosti pro i-tý záznam, C je kovarianční matice, x_i je vektor záznamů a $\bar{x}_n$ je průměrový vektor centroidu.

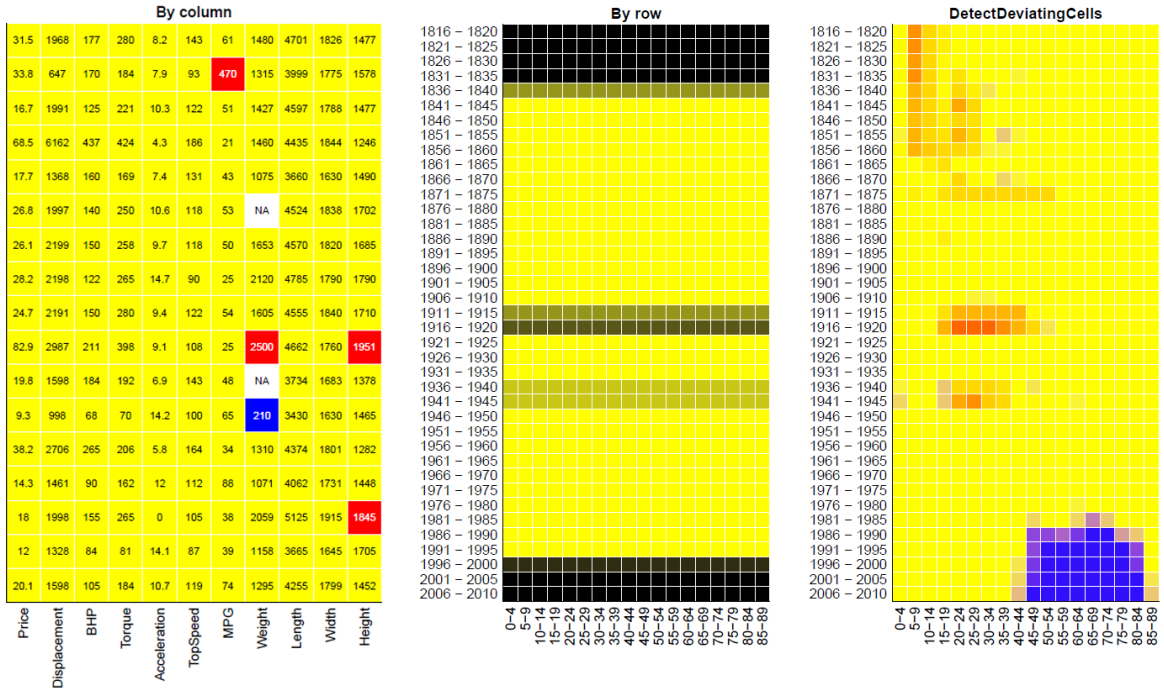
Pro vyhodnocení odlehlosti nestačí pouze spočítat vzdálenost od centroidu, je nutné také určit kritickou hodnotu vzdálenosti, od které se již záznamy budou považovat za odlehlé. K určení prahové vzdálenosti lze přistoupit různě: pouhým expertním posouzením (nejméně vhodné řešení), jednoduchým ukazatelem – podobně jako při tvorbě boxplotu lze použít kvartily a 1,5 násobek mezikvartilovlého rozpětí, nebo porovnání s teoretickým χ^2 rozdělení o n stupních volnosti a hladině významnosti α, jak uvádějí Hubert & Debruyne (2010).

Metoda Deviating Data Cells (DDC)

Mahalanobisova vzdálenost umožňuje identifikovat vybočující záznamy tzv. „*by row*" přístupem – pouze označuje záznamy (řádky v tabulce), který jsou z nějakého důvodu vybočující. Nijak však příčinu odlehlosti neobjasňuje. Pro pochopení příčiny odlehlosti je nutné prohlížet jednotlivé atributy odlehlého záznamu a pokoušet se jejich odlehlost vysvětlit průzkumem vstupních dat. Pohled „*by row*" zároveň nemusí při větším počtu sledovaných atributů, ze kterých bude pouze malý počet vybočujících, označit celý záznam jako odlehlý. Metoda DDC neoznačuje jako odlehlé hodnoty celé záznamy (by row), ale pouze výrazně vybočující atributy v kontextu vzájemných vztahů mezi všemi atributy (tzv. *cellwise* přístup). Rozdíly mezi různými přístupy shrnuje Obr. 5.

Princip metody lze zjednodušeně shrnout do několika kroků: nejprve je provedena robustní standardizace vstupních atributů (jelikož princip robustní standardizace je komplikovaný a přesahující úroveň tohoto textu, není zde dále vysvětlován. Zájemci jej však mohou najít v pracích Maronna et al. (2006); Rousseeuw & Bossche (2018)). Následně je na atributy aplikovaná jednorozměrná detekce odlehlých hodnot (*columnwise* přístup – přístup hodnotící

jednotlivé indikátory), odlehlé záznamy jsou následně vyloučeny z dalších výpočtů. Mezi všemi indikátory jsou hledány významné korelace (vyšší než stanovená prahová hodnota, ve výchozím nastavení s hodnotou 0,5). Z těchto párů jsou postupně tvořeny jednoduché robustní regresní modely, které predikují hodnoty jednotlivých atributů pro všechny buňky. Z predikovaných hodnot je počítán vážený průměr, a ten porovnán s původní pozorovanou hodnotu. Residua rozdílu mezi predikovanou a pozorovanou hodnotou jsou vyhodnocena vůči očekávané prahové hodnotě odvozené z χ^2 rozdělení. Je-li pozorovaná hodnota signifikantně vyšší/nižší než očekávaná, buňka je označena jako outlier. Díky tomuto přístupu poskytuje metoda hlubší porozumění struktuře dat a dokáže odhalit konkrétní problémové atributy.



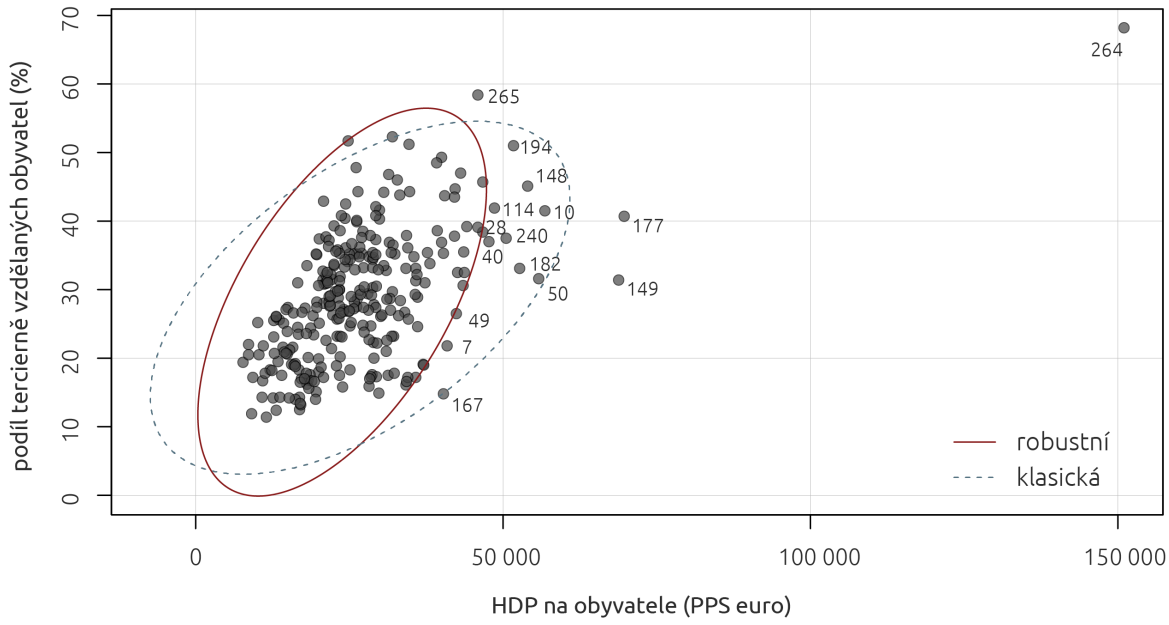
Obr. 5 Rozdílné přístupy k identifikaci odlehlých hodnot (zdroj: Rousseeuw & Bossche, 2018). V přístupu DDC (vpravo) odstíny červené indikují vyšší hodnotu, než je očekávaná; odstíny modré nižší hodnotu než je očekávaná. Přesnou legendu bohužel autoři neuvádějí.

1.2.2 Robustní metody určování odlehlých hodnot

Pokud některá z představených metod odhalí v datech odlehlé hodnoty, lze očekávat, že navazující analýzy budou těmito metodami do určité míry ovlivněné. Jak moc, to už závisí na řadě aspektů – množství odlehlých hodnot, velikost odlehlosti, směrové uspořádání atd. V každém případě je namístě uvědomění si negativního vlivu odlehlých hodnot a pokusit se jej co nejvíce zredukovat. Můžeme se rozhodnout odlehlé hodnoty zcela vypustit, a pak je problém vyřešen. Pokud však vypuštění jistého počtu záznamů výrazně narušuje celistvost datové sady (tento aspekt může být důležitý především v případě prostorových dat, kdy vypuštění některých administrativních jednotek způsobí narušení prostorové kontinuity), je vhodné odlehlé hodnoty zachovat a používanou analýzu upravit tak, aby vliv odlehlosti byl co nejmenší. Takovéto varianty metody označujeme jako **robustní**.

Robustní lze obecně popsat jako „necitlivý“ k malým odchylkám od ideálních předpokladů, na kterých je metoda odhadu optimalizována. Takovou optimalizací může být např. často vyžadované normální rozdělení v případě řady statistických metod. Jednu robustní metodu známe už ze základů statistiky – medián je vlastně robustní odhad střední hodnoty ve chvíli, kdy jsou data nějakým způsobem odchýlená od normálního rozdělení.

Představenou Mahalanobisovu vzdálenost můžeme konstruovat také robustním způsobem. To znamená, že identifikace odlehlých hodnot není samotnými odlehlými hodnotami ovlivněna, výsledek je tedy více citlivý a přísněji filtruje vybočující záznamy, jak je demonstrováno na Obr. 6. Robustní přístup je vhodnější také v situaci, kdy data nesplňuji podmínku normálního rozdělení (Varmuza & Filzmoser, 2009). Detaily postupu výpočtu nebudou v tomto textu rozebírány, princip výpočtu je popsán např. v Filzmoser et al. (2005).



Obr. 6 Porovnání vymezení odlehlých hodnot ve dvourozměrném prostorou pomocí klasického a robustního přístupu

1.2.3 Odlehlé hodnoty a prostor

Dvě uvedené metody vyšetřování odlehlých záznamů bohužel nijak nepracují s prostorovou informací, identifikují odlehlé hodnoty pouze na globální úrovni – v kontextu všech pozorování vstupních veličin a bez zahrnutí prostorové informace. Vztáhneme-li však informaci o odlehlosti k prostoru, můžeme dospět k závěru, že hodnota, která byla označena jako odlehlá, v určitém lokálním (geografickém) kontextu vybočující není. Nebo naopak v lokálním pohledu můžeme identifikovat odlehlé hodnoty, které však v globálním pohledu nejsou nijak výjimečné. Jako příklad uvádí Filzmoser & Gregorich (2020) datovou sadu průměrné roční teploty a srážkového úhrnu z pozorovacích stanic rozmístěných po celé Evropě. Pozorují 10 vybraných lokálních outlierů, které vždy vybočují pouze tehdy, uvažujeme-li kontext okolí 10 nejbližších pozorování. V globálním vnímání všech vstupních dat však většina těchto záznamů jako odlehlá vnímána není.

Podobně jako jsme schopni pomocí prostorové autokorelace (lokální analýza LISA) identifikovat odlehlé hodnoty jednorozměrného atributu, přenesením tohoto problému do vícerozměrné dimenze se zachováním prostorové složky se metody přestavené Filzmoserem et al. (2014) dívají na vícerozměrná data pouze v lokálním měřítku, a hodnotí, zdali je konkrétní záznam odlehlý pouze v kontextu svého okolí. Mohou přitom nastat následující případy:

- lokální, avšak ne globální outlier;
- globální, avšak ne lokální outlier;
- lokální i globální outlier;
- ani lokální ani globální outlier.

Přístup je založený na kombinaci výpočtu Mahalanobisovy vzdálenosti pro hodnocení odlehlosti atributů a Euklidovské (geografické) vzdálenosti pro vyhodnocení prostorové blízkosti (sousedství založené na vzdálenosti). Mahalanobisova vzdálenost je počítána mezi všemi páry záznamů, pro lokální ověření odlehlosti jsou však porovnávány pouze takové záznamy, které spadají do vymezeného sousedství.

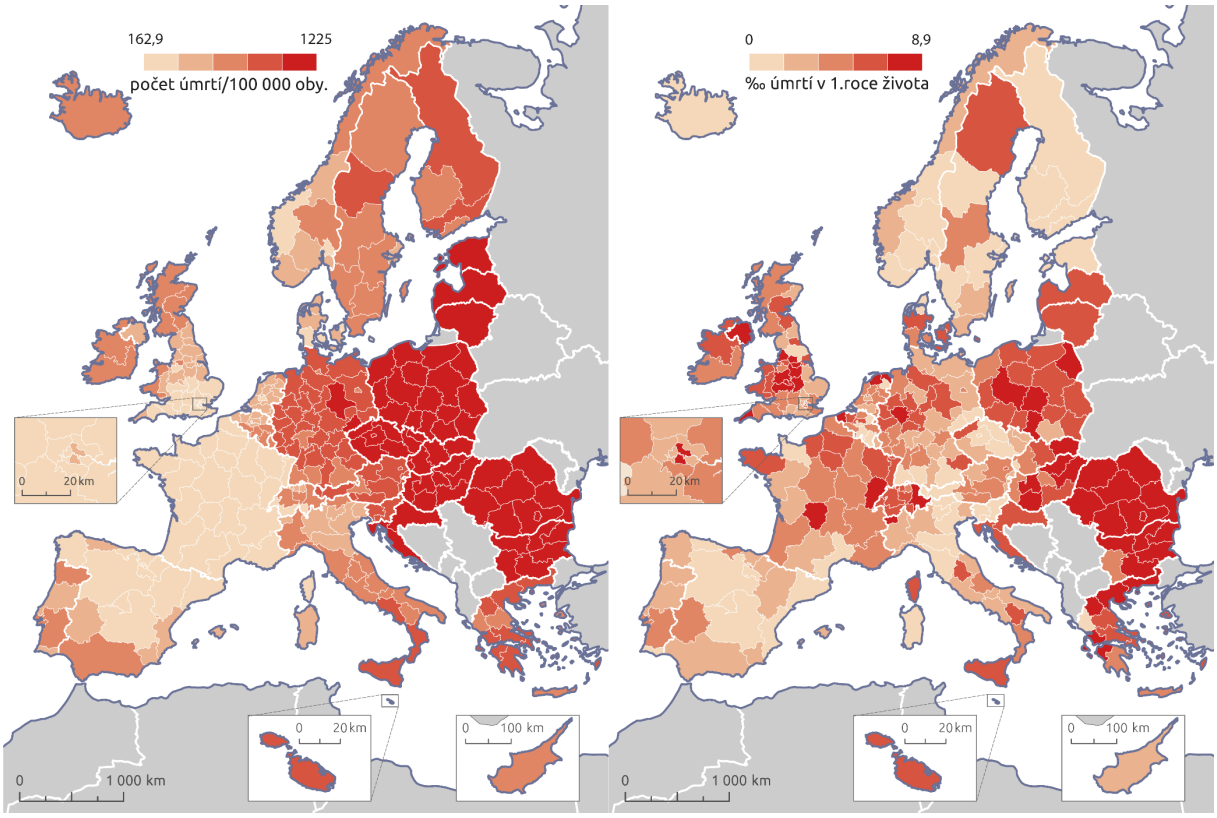
2 PROSTOROVÁ STATISIKA

Pojem prostorová statistika referuje na aplikaci statistických metod a konceptů na data, která obsahují libovolně zachycenou prostorovou informaci (Unwin, 2009). Metody prostorové statistiky nabízejí způsoby, jak využít prostorovou informaci k detekci a kvantifikaci vzorů a míry propojení mezi sledovými jevy (Waller & Gotway, 2004). Podle Gelfanda et al. (2010) se prostorová statistika zabývá jevy, jejichž prostorová poloha je sama o sobě předmětem zájmu, nebo přímo přispívá ke stochastickému modelu daného jevu. Prostorová informace hraje v celé analýze klíčovou roli – prostor není pouze doplňující informace, která může anebo nemusí být využita, ale figuruje zde jako rozšiřující proměnná. V úlohách prostorové statistiky je vždy sledovaný jev hodnocen v kontextu jeho prostorové informace. Díky využití statistických principů umožňují nástroje prostorové statistiky nejen aplikovat různé metody pro kvantifikaci vybraných jevů, ale podložit je také mírou statistické spolehlivosti získaných výsledků. Většina úloh prostorové statistiky odhaluje a kvantifikuje *prostorové vzory* v datech. Taková úloha může být třeba pouze omezena na zjištění, zdali je rozložení výskytů/hodnot sledovaného jevu v prostoru náhodné, nebo vykazuje nějaký trend či shlukování. Řešení se touto cestou pokouší zachytit zvláštní aspekt prostorových dat, kterým je funkční vztah mezi hodnotou proměnné a její polohou.

Většina charakteristických prostorových vzorů v datech vyplývá z konceptu dvou hlavních tzv. efektů prostorových dat:

- *Efekt I. řádu* se zaměřuje na prostorový trend v datech. Říká, že pravděpodobnost výskytu jevu se liší v každém místě zkoumaného prostoru, a náhodná veličina je proto prostorově závislá. Prostorová závislost a variabilita může být pozorována jako závislost na nějaké příčinné proměnné, která nemusí být známa, ale v prostoru se mění. Typicky se sledovaná veličina v prostoru charakteristicky uspořádává, např. hodnoty se postupně zvyšují ve směru od východu na západ. Toto chování v prostoru označujeme jako trend (Obr. 7 – vlevo).
- *Efekt II. řádu* je výsledkem interakcí, kdy přítomnost jednoho pozorování (nebo konkrétní hodnoty sledovaného jevu) zvyšuje pravděpodobnost výskytu dalších pozorování (nebo podobných hodnot sledovaného jevu) v bezprostřední blízkosti. Popisuje tedy závislost výskytu hodnoty náhodné veličiny na výskytu hodnot v jejím okolí, týká se působení vzájemných vlivů mezi pozorováními. Výskyt jevu na jednom místě zvyšuje pravděpodobnost výskytu jevu v okolních místech. Je-li v datech přítomný efekt II. řádu, podobné hodnoty (vysoké nebo nízké) vytvářejí prostorové shluky (Obr. 7 – vpravo), které je možné kvantifikovat např. nástroji pro měření prostorové autokorelace.

Rozlišení mezi těmito hlavními efekty může být někdy komplikované, neboť se do jisté míry vzájemně prolínají: samotný prostorový trend je zachytitelný metodami prostorové autokorelace, podobně trend také splňuje předloženou definici efektu druhého řádu. Doplňujícím rozlišujícím ukazatelem potom může být měřítko pohledu, kdy v globálním pohledu pozorujeme trendovou prostorovou proměnlivost, zatímco v lokálním pohledu se můžeme zaměřit více na kolísání způsobené prostorovou autokorelací.



Obr. 7 Příklady efektů I. a II. řádu: vlevo – jev s typickým geografickým trendem (úmrtnost důsledkem nemocí oběhové soustavy); vpravo - jev zatíženým autokorelací (kojenecká úmrtnost)

Prostorová složka dat sice přináší obohacující informaci, zároveň ale práci s daty komplikuje. Příkladem takovýchto komplikací mohou být jevy známé jako *Modifiable Areal Unit Problem* (Longley et al., 2005), *Edge Effect* (hraniční efekt) nebo závislost analýzy na zvoleném měřítku. Také samotná podstata prostorových dat narušuje základní statistické předpoklady: při testování statistických hypotéz běžně očekáváme, že analyzovaný soubor je souborem výběrovým, tedy pouze jednou možnou realizací celé populace. V případě prostorových dat je však zpracovávány soubor konečný a dá se říct, že skutečně představuje celou populaci (např. nezaměstnanost ve všech obcích České republiky). Proto u vyhodnocování prostorově-statistických testů většinou nespoléháme na očekávané teoretické rozdělení pravděpodobnosti, ale musíme jej nějakým způsobem nasimulovat a následně vyhodnotit míru významnosti (viz kapitola 3 o metodě Monte Carlo). Všechny tyto specifické aspekty prostorových dat budou postupně popisovány v následujících částech tohoto studijního textu.

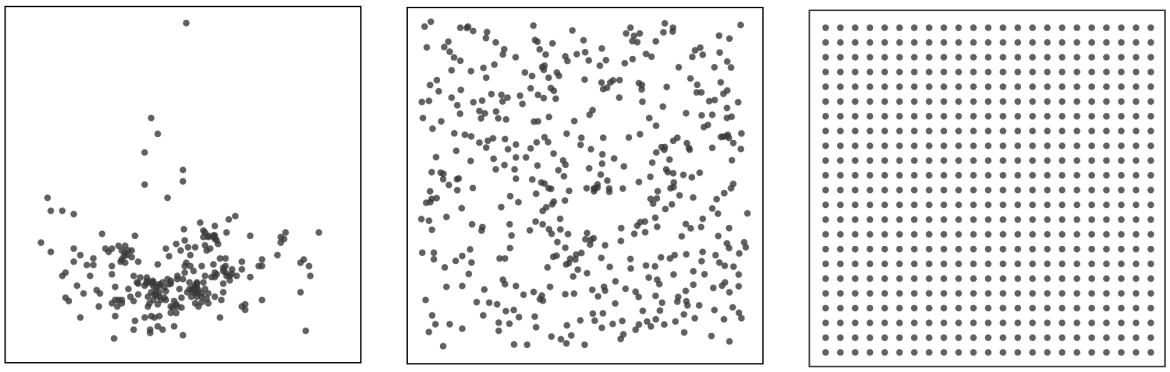
2.1 ANALÝZY BODOVÝCH DAT

Analýzy bodových dat lze začít modifikací jednoduchých popisných statistik pro potřeby specifik prostorových dat. Příkladem takových ukazatelů jsou průměrný střed, mediánový střed, směrodatná vzdálenost nebo elipsa směrodatné odchylky (viz Horák (2015)). Tyto ukazatele jsou prostorovou obdobou základních popisných statistik – také ukazují jednoduchou charakteristiku polohy nebo variability. Důležitou vlastností je, že primárně pracují pouze se samotnou polohou výskytu bodového jevu a jsou tudíž počítány nad souřadnicemi vstupních bodů. Metody mohou být modifikovány v podobě tzv. vážených

metod, kde do výpočtu vstupují jako váhy hodnoty sledované veličiny. Tyto techniky analýzy bodového vzoru byly populární hlavně v dobách, kdy výpočetní technika neumožňovala žádné složitější výpočty. Přestože mohou poskytnout základní informaci o bodové struktuře, mohou být pro některé výzkumné otázky příliš triviální a cennější informace o pozorovaném vzoru zůstávají skryté. K podrobnějšímu prozkoumání bodových vzorů lze použít výkonnější analýzy.

Před popisem samotných metod pro kvantifikaci prostorových vzorů bodových dat je nutné vysvětlit základní teoretické východiska pro studium bodových jevů, též označovaných jako *bodové procesy*. Uvažujme, že našim cílem je vyhodnotit prostorovou distribuci (tedy vzor) zkoumaných dat. Na teoretické úrovni očekáváme v zásadě tři možné varianty:

- **shluková**: jednotlivé záznamy mají tendenci se k sobě seskupovat, vytvářejí shluky (Obr. 8a),
- **náhodná**: záznamy jsou v prostoru rozmístěna bez jasného vzoru (Obr. 8b),
- **rovnoměrná**: jednotlivé záznamy jsou v prostoru rozmístěny pravidelně, až nepřirozeně (Obr. 8c).



Obr. 8 Tři základní varianty prostorové distribuce bodových dat: shluková (a), náhodná (b), pravidelná (c)

Pro určení pozorované distribuce bodů a pro testování jejího typu potřebujeme mít teoretické modely distribuce bodů, které odpovídají jistým typovým situacím se zřejmou interpretací (Horák, 2015). Při hodnocení charakteru bodového procesu se v praxi testuje význam rozdílu mezi pozorovanou strukturou a některým z teoretických modelů náhodné distribuce. V tomto místě by čtenáře měla napadanou otázka, jak tedy vypadá náhodná distribuce bodových jevů.

COMPLETE SPATIAL RANDOMNESS – homogenní Poissonův proces

Jaká je tedy podoba náhodného bodového procesu? Nejčastěji se setkáváme s modelem kompletní prostorové náhodnosti - *Complete Spatial Randomness (CSR*) (Illian et al., 2008). V jeho jádru nacházíme Poissonovo rozdělení pravděpodobnosti, které popisuje pravděpodobnost výskytu jevu v určitém časovém či objemovém intervalu.

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Mezi vlastnosti Poissonova rozdělení patří časová i prostorová nezávislost, a stejná pravděpodobnost výskytu jevu pro každý časový okamžik. Poissonovo rozdělení je ve své základní variantě jednorozměrné, nicméně jeho vlastnosti lze přenést do dvourozměrného prostoru a modelovat pomocí něj prostorovou náhodnost. Z hodnoceného území vybíráme náhodné oblasti A o stejné velikosti $|A|$, jejich distribuce $X(A_1), \dots, X(A_i)$ jsou nezávislé na poloze a jejich rozdělení pravděpodobností má Poissonovo rozdělení:

$$f_{X(A)}(x) = \frac{(\lambda A)^x}{x!} e^{-\lambda A}$$

kde $\lambda = \frac{n}{|A|}$ můžeme chápat jako intenzitu výskytu (průměrný počet výskytů na plochu). Takový model má dvě důležité vlastnosti (Horák, 2015):

1) je homogenní, nevykazuje tedy žádný efekt I. řádu (trend). Parametr λ je konstantní, pravděpodobnost výskytu jevu je v každém místě stejná,
2) sleduje Poissonovo rozdělení, počet výskytů v sousedících oblastech na sobě nijak nezávisí (nevykazuje prostorovou autokorelaci a není v něm přítomen efekt II. řádu).

Nejjednodušší alternativou k CSR je heterogenní Poissonův proces. V tomto případě je konstantní intenzita výskytu jevu λ je nahrazena funkcí intenzity $\lambda(s)$, stále platí zachování prostorové nezávislosti (počet výskytů v sousedících oblastech na sobě nijak nezávisí). Závislostí λ proces zachycuje projev efektu I. řádu – trend. Rozšířením heterogenního procesu o prvek náhodnosti vzniká tzv. Coxův proces, kde intenzita $\lambda(s)$ kolísá náhodně, ne deterministicky.

Jakmile je definována podoba náhodného bodového procesu, lze různými metodami ověřovat, jestli mu jsou pozorovaná data dostatečně podobná (jsou náhodná) nebo ne – mají tedy shlukový nebo rovnoměrný prostorový vzor.

Average Nearest Neighbor – metoda nejbližší vzdálenosti

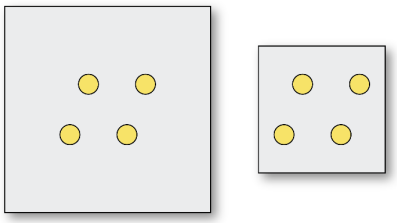
Metoda porovnává průměrné nejbližší vzdálenosti mezi pozorovanými body vůči referenční teoretické průměrné vzdálenosti, která by byla očekávána u náhodně rozmístěných prvků. Pokud je průměrná vzdálenost menší než teoretická očekávaná, prvky pravděpodobně tíhnou ke shlukování. Pokud je pozorovaná vzdálenost větší než očekávaná, jsou prvky rozprostřeny rovnoměrně.

Ukazatel se vypočítá jako podíl pozorované průměrné vzdálenosti a očekávané. Pro každý pozorovaný prvek se zjistí jeho vzdálenost k nejbližšímu sousedovi a tyto hodnoty se průměrují.

$$ANN = \frac{\overline{D}_O}{\overline{D}_E} = \frac{\frac{1}{n}\sum_{i=1}^{n} d_i}{\frac{0.5}{\sqrt{\frac{n}{A}}}}$$

kde $\bar{D}_O$ je pozorovaná průměrná nejbližší vzdálenost, $\bar{D}_E$ je očekávaná průměrná nejbližší vzdálenosti, d_i je vzdálenost k nejbližšímu sousedu pro prvek i, a A je (ve výchozím nastavení) plocha minimálního ohraničujícího pravoúhelníku všech bodů, jejichž počet je n.

Metoda je zásadně citlivá na velikost pozorované plochy (A) – v případě požadavku srovnatelnosti hodnocení více datových sad je nejlepší fixně nastavit zájmovou oblast. Stejně rozmístěné prvky totiž mohou být ohodnoceny jako rozptýlené nebo naopak shlukované v kontextu celkové plochy, jak demonstruje Obr. 9.



Obr. 9 Vliv velikosti ohraničujícího obdélníku (A) na relativní vnímání prostorového vzoru: vlevo tendence rozmístění ke shlukování; vpravo rovnoměrné rozmístění (převzato z https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/average-nearest-neighbor.htm)

Statistická významnost výsledku analýzy je odvozena převedením na z-skóre, tedy normované normální rozdělení. Vhodnějším řešením je odvození statistické významnosti pomocí permutací Monte Carlo simulace (viz níže).

Analýza kvadrantů

Základním principem kvadrantových analýz je překrytí bodové vrstvy uměle vytvořeným gridem o konstantní velikosti buňky (případně lze experimentovat s náhodně rozmístěnými buňkami nebo náhodným výběrem z pravidelné sítě). Nad počty pozorovaných výskytů se následně provádí vyhodnocení, např. pomocí *indexu disperze* (Horák, 2015). Ten vychází z vlastností Poissonova rozdělení, kde se pro náhodný jev očekává $E(X) = \sigma^2 = \lambda$. Z jednotlivých počtů pozorování x_i v m buňkách se spočítá aritmetický průměr $\bar{x}$ a rozptyl σ^2, výsledný index se sestaví podle vztahu:

$$VMR = \frac{var(X)}{E(X)} = \frac{\sigma^2}{\bar{x}}$$

Je-li výsledkem zlomku hodnota blízká 1 (střední hodnota i rozptyl jsou přibližně stejné), jde o náhodnou distribuci, kde pozorované parametry odpovídají parametrům Poissonova procesu. Je-li VMR větší než 1, hodnota indikuje shlukování, hodnoty menší než 1 naopak ukazují na pravidelný vzor. Pro otestování statistické významnosti lze použít t-test nebo χ^2 test dobré shody (viz (Horák, 2015)).

G funkce

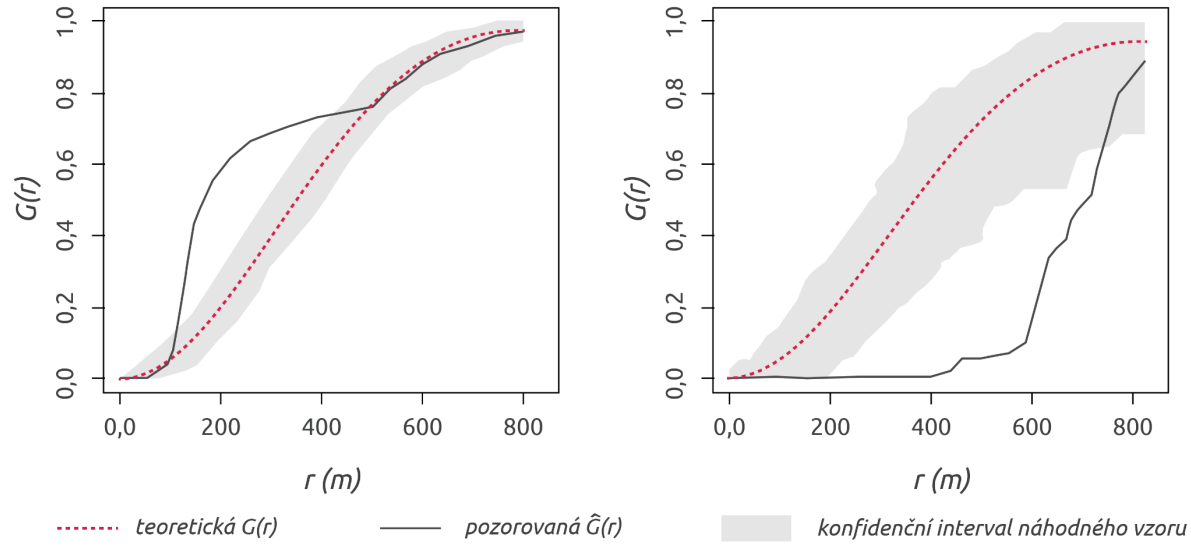
Funkce G zkoumá kumulativní frekvenci výskytu vzdáleností k nejbližšímu sousedu při určité prahové hodnotě vzdálenosti.

$$\hat{G}(r) = \frac{1}{n}\sum_{i=1}^{n} I_i\,,\quad I_i = \begin{cases} 1 & d_i \in \{d_i : d_i < r,\ \forall i\} \\ 0 \end{cases}$$

Odhad $\hat{G}$ funkce reprezentuje počet prvků ve vzdálenosti do zvoleného prahu r, normalizované počtem celkových prvků n (průměrná hodnota). Jednoduše změříme vzdálenost d_i od každého bodu k jeho nejbližšímu sousedu. Pokud je vzdálenost menší než stanovaný práh, bod se započítá do celkové sumy, ta se podělí celkovým počtem bodů. Práh se běžně stanovuje v postupných vzdálenostních krocích (třeba po 100 metrech), následně se pro každou vzdálenost vynesou hodnoty do grafu. Pozorované hodnoty lze porovnat s očekávanou G funkcí homogenního Poissonova procesu pro konkrétní hodnoty r:

$$G(r) = 1 - e^{\lambda \pi r^2}$$

Pokud je $\hat{G}(r) > G(r)$ (pozorovaná funkce větší než očekávaná), pozorujeme více blízkých bodů, než se očekávalo podle CSR a dochází tedy ke shlukování (viz Obr. 10 vlevo). Je-li naopak $\hat{G}(r) < G(r)$, lze vzor považovat za pravidelně rozptýlený (viz Obr. 10 vpravo).



Obr. 10 Grafické vyhodnocení G funkce (převzato a upraveno z Rossiter (2020))

Ripleyho K funkce

Ripleyho K funkce je další metoda zachycující efekt II. řádu bodových jevů, která vyhodnocuje tendenci k prostorové závislosti výskytu bodového jevu v libovolné vzdálenosti (Ripley, 1977). Pracuje s výskytem jevu v definovaném okolí (je zde tedy vidět jistá podobnost např. k Moranovu I kritériu – dále vysvětleno v kapitole 2.2.2). Počet pozorovaných hodnot je porovnáván s očekáváným počtem pozorování, které by bylo zaznamenáno u CSR procesu. Pokud je pozorovaná hodnota větší než očekávaná náhodná vycházející z CSR, pravděpodobně dochází ke shlukování; naopak pokud je hodnota nižší než očekávaná náhodná, body jsou pravděpodobně rovnoměrně rozprostřeny (Waller & Gotway, 2004).

Hodnota K funkce je teoreticky počítána spojitě pro všechny hodnoty vzdálenosti. Základní myšlenka výpočtu je jednoduchá – spočívá v podílu průměrného počtu pozorovaných bodů

ve vzdálenosti h (*spatial lag* – N_h) a průměrné intenzity výskytu λ (počet výskytů na jednotku plochy). Jelikož je λ vyjádřena jako podíl počtu prvků na ploše, svým způsobem odstraňuje závislost na hustotě odhadu v konkrétním analyzovaném intervalu vzdálenosti.

$$K_h = \frac{1}{\lambda} E(N_h) = \frac{1}{\lambda} \cdot \frac{1}{n} \cdot \sum N_h$$

Pro jednodušší zpracování se vzdálenost h diskretizuje do intervalů. Pro výpočet střední hodnoty je tedy nutné pro každou hodnotu vzdálenosti h zjistit průměrný počet sousedů nad všemi pozorovanými body. Pro interpretaci je pozorované hodnoty K funkce nutné porovnat s očekávanou hodnotou, jaká by byla zaznamenána nad CSR. která je popsána vztahem:

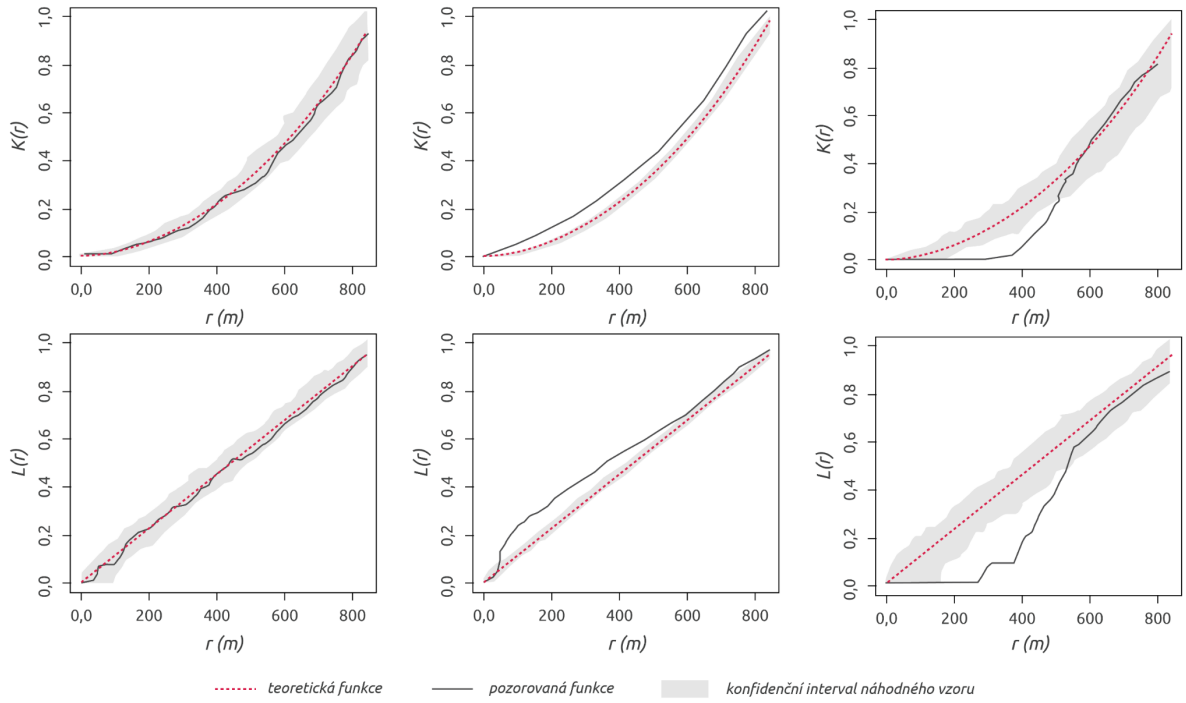
$$K'_h = \pi h^2$$

, kde h je opět hodnota vzdálenosti. Výsledek je interpretován podobně jako v případě G funkce, pozorované a očekávané hodnoty lze vynést do grafu a sledovat jejich vzájemný vztah. S rostoucí hodnotou h současně narůstá rozptyl v průměrném počtu pozorovaných bodů. Rozdíly mezi pozorovanou a očekávanou funkcí mohou být při kvadratickém tvaru očekávané funkce často špatně rozpoznatelné. Tyto nedostatky se proto dále odstraňují transformací pomocí tzv. L funkce, která původní očekávanou K funkci linearizuje:

$$L_h = \sqrt{\frac{K_h}{\pi}} - h$$

Výsledné hodnoty jsou pak vůči přímce lépe porovnatelné (Obr. 11). Hodnoty, kde pozorovaná funkce překročí očekávanou hodnotu odhadu L_h indikují shlukování, zatímco hodnoty menší než očekávaná značí disperzi.

G funkce a K funkce (potažmo její alternativní vyjádření jako L funkce) jsou na první pohled velmi podobné analýzy. Základní rozdíl můžeme pozorovat v prací s prvky ve vymezeném okolí. Zatímco G funkce zkoumá podíl nejbližších sousedů v určitých krocích vzdáleností, K funkce hodnotí komplexněji intenzitu výskytu bodů, jelikož pozoruje jejich průměrný počet v určitém vzdálenostním intervalu.



Obr. 11 Interpretace původní (nahoře) a linearizované (dole) K funkce: vlevo – náhodné rozmístění, uprostřed – tendence ke shlukování bodů, vpravo – rovnoměrné rozmístění. Převzato a upraveno z Rossiter (2020)

2.2 ANALÝZY AREÁLOVÝCH DAT

Zatímco u bodových dat jsme se soustředili především na polohu jednotlivých záznamů, u areálových dat budeme kromě polohy hodnotit také atributovou složku. Představené analýzy budou opět úzce souviset s efekty I. a II. řádu.

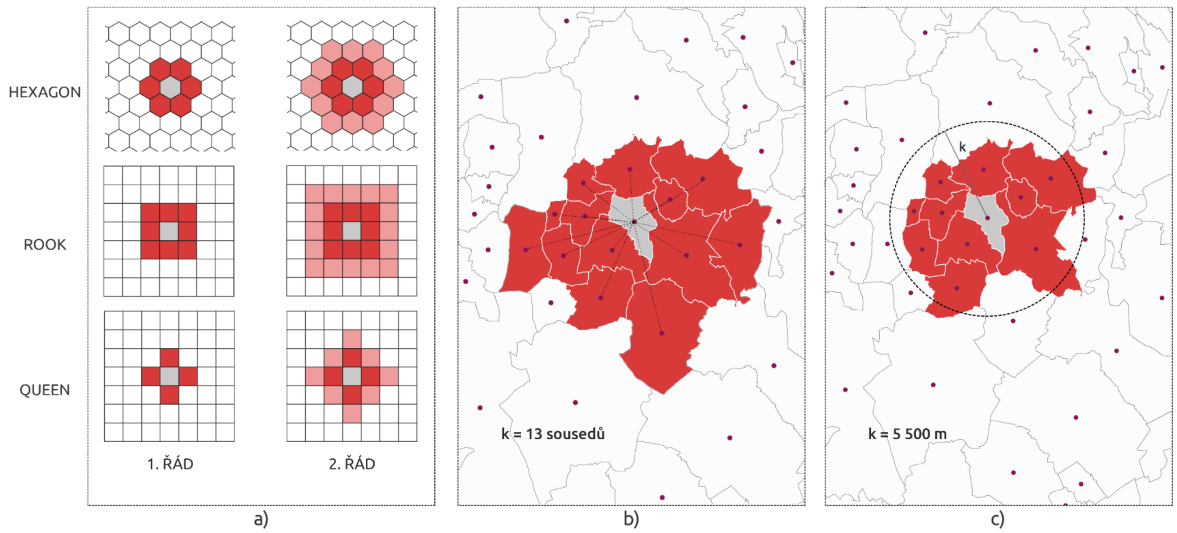
2.2.1 Vyhlazování areálových dat

Shlazování dat běžně používáme u jednorozměrných dat (nejčastěji u časových řad) – máme-li např. meteorologická data o srážkách, jsou zatížené různým kolísáním a výkyvy, jako jsou lokální bouřky. Chceme-li pozorovat dlouhodobý trend, data vyhladíme např. pomocí klouzavého průměru, kdy vypočítáme hodnotu srážek pro daný den jako průměr hodnot z pohyblivého okna, např. z okolních 7 dnů (tzv. týdenní klouzavý průměr). Stejným principem můžeme vyhlazovat i geografická data – průměrujeme hodnotu pozorované náhodné veličiny na základě vymezeného okolí, nyní však dvourozměrného (geografického). Cílem této operace je zvýraznění efektu I. řádu a potlačení náhodné složky, která může někdy zobrazení trendu utlumovat. Vyhlazování má samozřejmě smysl pouze v případě, kdy sledovaný jev vykazuje přirozenou kontinuitu. O vyhlazování dat se nejčastěji mluví u polygonových dat, samozřejmě není problémem je aplikovat i na bodová data – v analýzách jsou koneckonců polygony pro potřeby výpočtu reprezentovány svým centroidem.

Jelikož základem prostorového vyhlazování je vymezení prostorového „okolí“ (nebo také sousedství) pro každý sledovaný prvek, je nezbytné nejprve popsat základní způsoby

vymezování okolí, které se dále aplikují ve většině prostorových metod. Sousedství vymezujeme nejčastěji jedním ze tří způsobů (graficky jsou znázorněny na Obr. 12):

- *Topologicky*: jako sousedící se považují pouze ty záznamy, které se skutečně topologicky dotýkají, a to buď sdílením hrany (tzv. *rook* – podle možnosti pohybu věže ve hře šachy) nebo sdílením hrany nebo alespoň jednoho lomového bodu (tzv. *queen* – pohyb figurky dáma). Můžeme dále rozlišovat sousedství 1., 2. nebo vyššího řádu podle toho, jestli uvažujeme pouze přímé sousedy, nebo dále sousedy sousedů pro zajištění většího dosahu.
- *Počtem k nejbližších sousedů*: v této možnosti seřadíme všechny okolní prvky vzestupně podle vzdálenosti (mezi centroidy pozorovaného prvku a jeho sousedy) a jako sousedy považujeme *k* nejbližších prvků.
- *Vzdáleností*: sousedství je určeno pomocí vzdálenosti jednotlivých záznamů (jejich centroidů). Stanoví se prahová hodnota, a pokud je vzdálenost mezi pozorovaným prvkem a centroidem libovolného sousedního prvku menší než tato hodnota, považujeme jej za souseda.



Obr. 12 Vymezení sousedství topologicky (a), pomocí k nejbližších sousedů (b) a pomocí vzdálenosti (c)

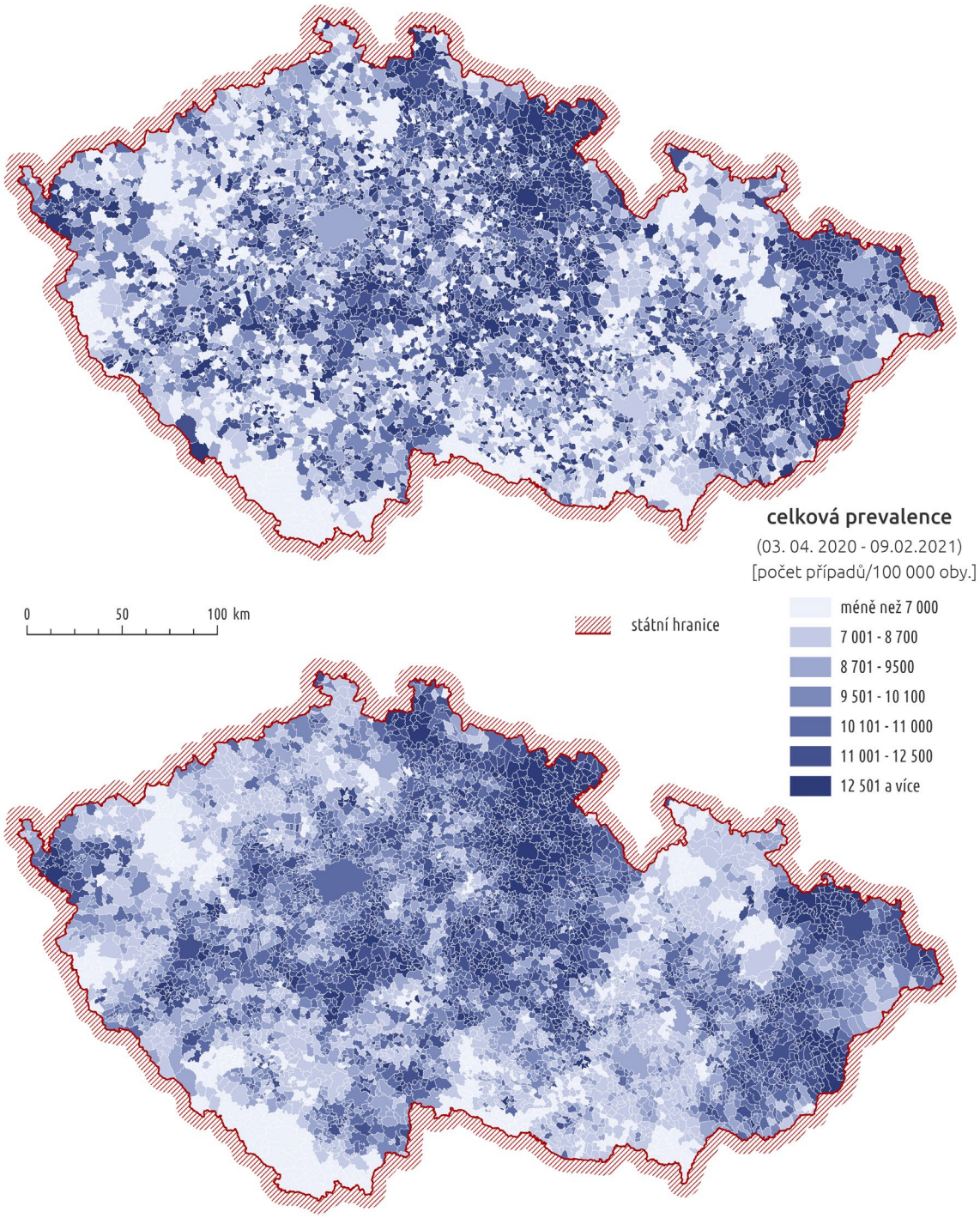
Tyto způsoby vymezení sousedství lze dále rozšiřovat a více specifikovat – např. modifikací vzdálenosti pomocí jádrové funkce. Tyto metody budou dále popsány v kapitole 4.

Aplikací principu sousedství na všechny pozorované prvky vzniká matice prostorových vah, která je základním kamenem všech metod pracujících se sousedstvím. Jedná se o matici typu „každý s každým“, ve které je zakódována informace o sousedství. V nejjednodušším případě: pokud jsou dva prvky chápány jako sousedé, mají v matici hodnotu 1. Prvky, které spolu nesousedí, mají přiřazenu hodnotu 0. V praktických úlohách se nejčastěji používají řádkově standardizované váhy (w_{ij}): původní hodnota váhy (1) je dělena řádkovým součtem (všechna sousedství jednoho prvku), součet standardizovaných řádkových vah je pak roven 1. Matici lze dále modifikovat: rozlišovat, jestli bude řádkově standardizována, jestli bude samotný sledovaný prvek zahrnut do sousedství nebo ne, atd. (rozbor těchto nastavení nabízí např. Anselin (2018) v dokumentaci k software GeoDa).

Aplikací matice prostorových vah a průměrováním hodnot z okolí pak vznikají modifikovaná data, která Anselin (2018) označuje termínem *spatially lagged variable* (česká terminologie v této disciplíně není stále pevně daná, mohli bychom si vystačit s označením prostorově vyhlazená proměnná). Anselin uvádí vzorec pouze jako sumu sousedních hodnot, pro získání klouzavého průměru je však nutno sumu dále dělit váhami:

$$Wy_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}$$

kde w_{ij} je matice prostorových vah a y_j je hodnota sledovaného jevu v lokalitě j. Sousedící prvky s vlivem mají nenulovou hodnotu, nesousedící prvky bez vlivu mají nulovou hodnotu. Již touto jednoduchou aplikací tak vzniká prostorová varianta klouzavých průměrů, které vyhlazují původní data a zvýrazňují jejich prostorový trend. Rozdíl mezi původními a shlazenými daty je jasně patrný na Obr. 13.



Obr. 13 Výskyt případů nemoci COVID-19 v obcích Česka: nahoře původní data, dole prostorově vyhlazená data pomocí sousedství typu královna I. řádu (zdroj dat: MZČR)

Empirické Bayesovo vyhlazování

Vyhlazování dat lze provádět i na atributové úrovni pomocí statistických modelů, například tzv. Bayesovským vyhlazováním. V situacích, kdy pracujeme s poměrovými daty (relativní údaje, podíly) můžeme pozorovat nestabilitu variability přímo závislou na velikosti vztažené populace. Jelikož je metoda nejčastěji používána nad zdravotnickými daty, bude demonstrována na případu výskytu libovolné nemoci. Základní chybou by bylo vyjádřit

skutečný počet nakažených v absolutních počtech, protože by nebyla zohledněna velikost dotčené populace, kterou nemoc může postihnout. Uvažujme např. onemocnění rakovinou prsu – absolutní čísla za administrativní jednotku (obec) nám moc informace o riziku onemocnění neposkytnou. Vhodnější je vyjádřit údaj relativně jako počet případů na celkový počet obyvatel v území. Dále je nutné se zamyslet, jestli je volba populace, ke které informaci vztahujeme správná: muži zjevně tímto onemocněním trpět nemohou, proto dává mnohem větší smysl relativizovat vůči počtu žen žijících v obci. Tímto přepočtem získáme hrubou míru rizika rakoviny prsu.

V situaci, kdy je potencionální populace malá – v obci žije např. jen 20 žen a jedna z nich má rakovinu prsu, získáváme poměrně vysoké relativní riziko, které s každým dalším případem výrazně stoupá. Pokud bychom poměr porovnali jinou obcí, kde žije třeba 100 000 žen, bude zjištěná míra dost stabilní, pár nových případů ji nijak významně nevychýlí. Vzniká tedy otázka, jestli je možné tuto nestabilitu způsobenou velikostí ohrožené populace nějak podchytit a zohlednit.

Empirické Bayesovo vyhlazování využívá k výpočtu pravděpodobnosti kombinaci skutečných pozorovaných dat s vhodným apriorním odhadem pro získání nové, upravené posteriorní pravděpodobnosti sledovaného jevu. Díky tomu jsou záznamy s nízkou stabilitou výrazněji upraveny (vyhlazeny) než záznamy s vysokou stabilitou.

Mějme pozorované hodnoty relativního rizika r_i, s neznámou apriorní distribucí a jejími parametry (střední hodnota a průměr). Parametry nahradíme tzv. Poisson-Gamma modelem, kde Poissonovo rozdělení popisuje pozorované počty událostí a Gamma rozdělení s parametry $\Gamma(\alpha, \beta)$ popisuje průměrné riziko π (např. průměrná incidence v celém zájmovém území). Kombinací apriorního Gamma rozdělení popisující průměrné riziko a Poissonova rozdělení skutečně pozorovaných počtů událostí vzniká nové posteriori rozdělení, kde je míra rizika upravena (vyhlazena) pomocí parametrů Gamma rozdělení (α, β).

$$E(\pi) = \frac{O + \alpha}{P + \beta}, \qquad Var(\pi) = \frac{O + \alpha}{(O + \beta)^2}$$

Jinými slovy, nový odhad rizika π_i upravuje původní hodnotu pomocí parametrů α, β z Gamma rozdělení, odhadnutých z pozorovaných dat. Vyhlazená hodnota pozorovaných rizik se vypočítá z původního relativního rizika r_i a odhadu θ, kterým může být průměrné globální riziko:

$$\pi_i = w_i r_i + (1 - w_i)\theta$$

$$w_i = \frac{\sigma^2}{\sigma^2 + \frac{\mu}{P_i}}, \quad r_i = \frac{O_i}{P_i}$$

, kde w_i jsou váhy vycházející z populace P_i, r_i je hrubá míra rizika v místě i; μ, σ^2 je průměr a rozptyl apriorní distribuce (tyto parametry odhadujeme pomocí Poisson-Gamma modelu z dat), P_i je celková populace v místě i a O_i počet pozorovaných případů v místě i. Technika Bayesovského vyhlazování v podstatě spočívá ve výpočtu váženého průměru mezi relativním rizikem v každé pozorované jednotce a globálním průměrným rizikem, přičemž váhy jsou úměrné ohrožené populaci.

Je nutno si uvědomit, že výsledné hodnoty rizika už nejsou skutečně pozorované, ale upraveným odhadem, který redukuje míru zkreslení vznikající malou populací. Empirické Bayesovo vyhlazování lze kombinovat také s prostorovou informací, kdy se místo globálního relativního rizika využívá lokální relativní riziko, vypočítané pouze z části území, která je definována maticí prostorových vah. Tímto přístupem se do shlazování zakomponuje prostorový trend, a získá se tak přesnější výsledek. Zahrnutí prostorou do shlazování lze řešit např. v software GeoDa (viz Anselin (2018)).

2.2.2 Prostorová autokorelace

Jelikož hodnocení prostorové autokorelace je jedna ze základních prostorově-statistických metod, bude jejímu vysvětlení věnována větší pozornost. Pro pochopení prostorové autokorelace můžeme vyjít z obyčejné korelace – ta popisuje míru lineární závislosti dvou náhodných veličin. Změnou z korelace na autokorelaci dostáváme variantu závislosti jedné veličiny na sobě samé. Toto nejasné vysvětlení můžeme chápat jako hodnocení, jak moc veličina v jednom místě koreluje s hodnotami stejné veličiny v jiném místě. Příkladem v jednorozměrném prostorou je autokorelace časové řady: vymezíme určitý časový úsek, který porovnáváme s jiným. Pokud tento princip rozšíříme do geografického prostoru, lze hodnotit, jestli veličina v nějaké části zájmového území koreluje s hodnotami v jiné části území. Tento prostorový vztah je založen na známém Toblerově prvním geografickém zákonu: „*Everything is related to everything else, but near things are more related than distant things*“ (Tobler, 1970). Analýza autokorelace tedy umožňuje zjistit, jestli se v území vyskytují statisticky významné oblasti podobných (vysokých nebo nízkých) hodnot. Těmto oblastem budeme říkat prostorové shluky. Problém identifikace shluků podobných hodnot můžeme přeformulovat do statistických hypotéz, které jsou analýzou autokorelace testovány:

H_0: výskyt hodnot jevu v prostoru vykazuje náhodný vzor,

H_A: výskyt hodnot jevu v prostoru vykazuje nenáhodný vzor – shlukování nebo pravidelnost.

Autokorelaci kardinálních dat lze kvantifikovat několika ukazateli, nejpoužívanějšími z nich jsou Moranovo I, Getis-Ord G nebo Gearyho C. Důležité je rozlišit měřítkovou úroveň, na které autokorelaci hodnotíme. Ptáme-li se, jestli se v zájmovém území vůbec vyskytují shluky podobných hodnot, používáme *globální* míry autokorelace. Pokud na tuto otázku dostaneme pozitivní odpověď, má smysl se ptát, kde konkrétně se tyto shluky nacházejí, a jestli se jedná o shluky vysokých nebo nízkých hodnot. Proto aplikujeme míry *lokální* autokorelace (Livings & Wu, 2020).

Globální metody

Nejprve se zaměřme na statistiku globální *Moranovo I* kritérium, zavedenou Moranem (1948) a dále rozpracovanou Cliffem & Ordem (1973). Indikátor se počítá podle vzorce:

$$I = \frac{\sum_i \sum_j w_{ij} c_{ij}}{s^2} \cdot \frac{n}{W} = \frac{\sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \cdot \frac{n}{\sum_{i \neq j} \sum_j w_{ij}}$$

, kde z_i a z_j popisují hodnotu náhodné veličiny ve sledovaných lokalitách i a j, $\bar{z}$ zastupuje celkovou průměrnou hodnotu veličiny v celém zájmovém území, w_{ij} je vyjádřením matice prostorových vah mezi místy *i* a *j*, a *n* je celkový počet prvků ve zkoumaném území.

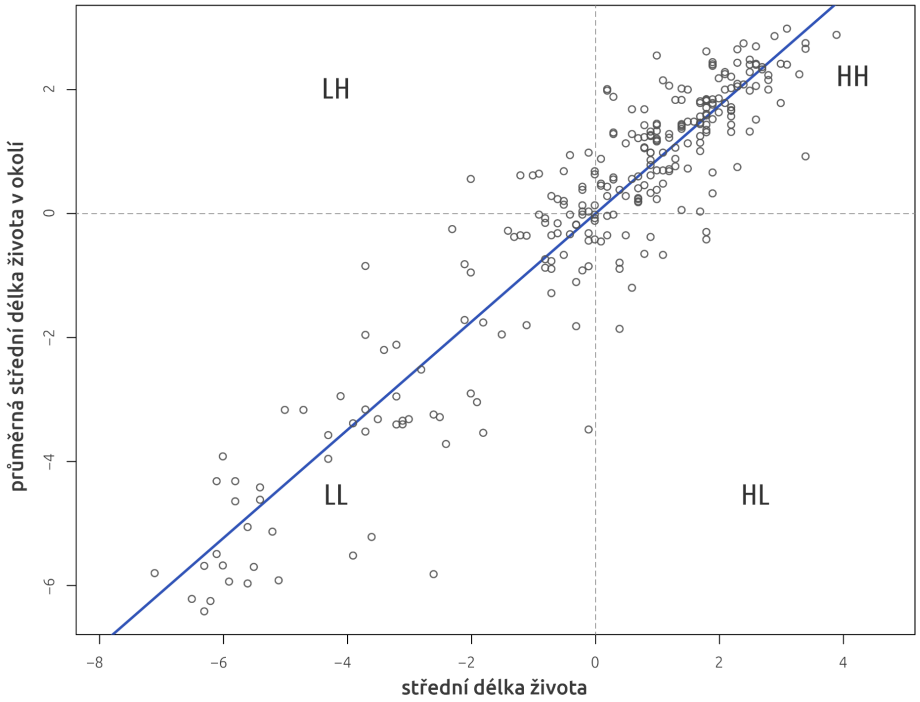
Součin $c_{ij} = (z_i - \bar{z})(z_j - \bar{z})$ (tzv. *cross produkt* – vektorový součin) zachycuje podobnost/nepodobnost lokalit z_i a z_j vůči celkové průměrné hodnotě $\bar{z}$. Jsou-li hodnoty z_i a z_j obě menší nebo obě větší než průměrná hodnota, dostáváme pozitivní součin; je-li právě jedna z hodnot menší než průměrná hodnota, dostáváme negativní součin. Zároveň čím větší odchylky jsou, tím větší je výsledný součin. Toto vyhodnocení je nutné provést pro kombinaci všech záznamů v zájmovém území (z_i) a s jejich okolím (z_j) (proto se ve vzorci vyskytuje dvojitá suma). Matice prostorových vah w_{ij} specifikuje vymezení sousedství – určuje tedy pro konkrétní lokalitu z_i její sousedy z_j a také kvantifikuje váhově jejich vliv, dle vybrané metody definice sousedství. Jmenovatel vzorce reprezentuje výběrový rozptyl (po úpravě vzorce je *n* převedeno do čitatele). Ve výsledku tedy dostáváme podíl dvou odlišností: celková odlišnost pozorované hodnoty v lokalitě i od globálního průměru v kombinaci s odlišností hodnot v okolí lokality *i* globálního průměru; a celková odlišnost pozorovaných hodnot od průměrné globální hodnoty.

Hodnoty globálního Moranova I se pohybují přibližně na intervalu (-1, 1), kdy kladné hodnoty indikují přítomnost autokorelace – tedy výskyt shluků podobných hodnot, zatímco záporná hodnota I poukazuje na zápornou autokorelaci – spíše na rovnoměrnou distribuci hodnot v prostoru. Hraniční hodnota pro určení statisticky významné autokorelace může být |0,3|, výsledek je však nutno podpořit statistickým testem (princip určování statistické významnosti pomocí Monte Carlo simulace bude dále popsán v kapitole 3).

Vztah pozorované hodnoty náhodné veličiny v lokalitě i a průměrné hodnoty v jejím okolí (dle Anselina (1996) označované jako *spatial lag*) lze graficky vykreslit v tzv. Moranově scatterplotu. Jak uvádí Anselin (1996), typicky se zobrazují hodnoty odchylek jednotlivých záznamů od průměru (v obou osách), stejný výsledek však získáme i při zobrazení původních nepřepočítaných hodnot (tímto způsobem graf vykresluje např. funkce *moran.plot* z balíku *spdep* v prostředí R). Prostor tohoto bodového grafu můžeme přirozeně rozdělit nulou do čtyř kvadrantů, jak je znázorněno na Obr. 14. Jednotlivé kvadranty reprezentují charakteristické skupiny pozorování:

- **HH**: nadprůměrná hodnota v lokalitě obklopená nadprůměrnými hodnotami v blízkém okolí (*hot spot*),
- **LL**: podprůměrná hodnota v lokalitě obklopená podprůměrnými hodnotami v blízkém okolí (*cold spot*),

- **HL**: nadprůměrná hodnota v lokalitě obklopená podprůměrnými hodnotami v blízkém okolí (*prostorový outlier*),
- **LH**: podprůměrná hodnota v lokalitě obklopená nadprůměrnými hodnotami v blízkém okolí (*prostorový outlier*).



Obr. 14 Moranův scatterplot – použita data střední délky života v evropských regionech NUTS 2 v roce 2015 (zdroj dat: Eurostat)

Vypočítáme-li lineární regresi mezi pozorovanými hodnotami a průměrnou hodnotou v jejich blízkého okolí, koeficient regresní přímky odpovídá hodnotě Moranova I kritéria (Spurná, 2008). Tento poznatek názorně podtrhuje princip prostorové autokorelace – cílem je vypozorovat závislost mezi hodnotami v konkrétní lokalitě a hodnotami v jejich okolí.

Lokální metody

Pokud se na globální úrovni podaří zamítnout nulovou hypotézu o náhodné distribuci hodnot sledovaného jevu, usuzujeme, že někde v zájmovém území dochází ke shlukování. Na otázku, kde konkrétně se tyto shluky nacházejí, a jestli jsou to shluky vysokých nebo nízkých hodnot, odpoví až lokální varianty metod pro hodnocení autokorelace. Princip lokálního Moranova I navrhl Anselin (1995), nazýváme jej také zkratkou LISA (Local Indicators of Spatial Association). Lokální Moranovo I pro lokalitu i je určeno vzorcem:

$$I_i = \frac{(z_i - \bar{z})}{S_i^2} \sum_{j \neq i} w_{ij}(z_j - \bar{z}) = r_i \sum_{j \neq i} w_{ij} r_j$$

Proti globální statistice ve vzorci pozorujeme pouze jednoduchou sumu, kde r_i zastupuje standardizovanou hodnotu náhodné veličiny v místě i a r_j zastupuje rozdíl hodnot z okolí místa i vůči celkovému globálnímu průměru náhodné veličiny (detailní rozbor výpočtu je

dostupný např. z oficiální dokumentace software GeoDa – Anselin (2020)). Vypočítané hodnoty je nutno prověřit testem pro statistickou významnost a spočítat pravděpodobnost výsledku pomocí odhadu střední hodnoty a rozptylu (Horák, 2015).

S ověřováním statistické významnosti úzce souvisí pojem *False Discovery Rate* (FDR), navržený Benjaminim & Hochbergem (1995). Během vyhodnocování významnosti výsledků výpočtu se snažíme eliminovat chybu I. druhu pomocí volby vhodné hladiny α. Jelikož je významnost výsledkem permutačního procesu, mohou některé výsledky být falešně pozitivní. Za účelem eliminovat tyto nepřesnosti lze na výpočet aplikovat korekce, jako je např. FDR nebo Bonferroni korekce.

Při aplikaci FDR korekce jsou nejprve záznamy seřazeny vzestupně dle hodnoty p-value, následně je přidána nová proměnná $FDR = i * \alpha/n$, kde i je pořadí záznamu, n počet pozorování a α je požadovaná hladina významnosti. Jako významné jsou pak označeny pouze záznamy, kde je nově vypočítané FDR větší než původní p-value.

Getis–Ord G

Globální Moranovo I není jedinou statistikou autokorelace pro kardinální data, kromě Gearyho C můžeme použít další indikátor pojmenovaný podle svých autorů A. Getise a J. K. Orda jako *Getis-Ord G* (Getis & Ord, 1992). Podobně jako u Moranova I lze metodu aplikovat na globální úrovni a testovat nulovou hypotézu o náhodné distribuci hodnot sledovaného jevu. Větší smysl má ale hledat konkrétní shluky nízkých nebo vysokých hodnot pomocí lokální varianty. Obecně se Getis-Ordova analýza snaží zodpovědět stejnou otázku jako LISA, liší se pouze přístupem řešení, který demonstrují následující vzorce:

$$G = \frac{\sum_{i \neq j} w_{ij} x_j}{\sum_{i \neq j} x_j} \qquad G^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}$$

Statistika se skládá z poměru prostorově váženého průměru hodnot ve vymezeném sousedství ku sumě všech hodnot. Ve výpočtech většinou narazíme na dvě varianty této metody, označené jako G a G*. Jejich rozdíl spočívá pouze v tom, že varianta G při vyhodnocování podobnosti mezi lokalitou i a jejím okolím nezahrnuje samotnou lokalitu i do okolí, naopak varianta G* považuje lokalitu i za součást okolí.

Interpretace výsledků je velmi jednoduchá: hodnota větší než průměr (nebo kladná hodnota v případě použití z-score) naznačuje shluk vysokých hodnot, hodnota menší než průměr (nebo negativní hodnota v případě použití z-score) označuje shluk nízkých hodnot. Na rozdíl od analýzy LISA, Getis-Ordovo G neumí určit prostorové outliery. Zamítnutím nulové hypotézy prokazujeme shlukující se vzorec nízkých nebo vysokých hodnot (identifikuje pouze typ shlukování). Další rozdíl proti analýze LISA vyplývající z představeného vzorce je takový, že Getis-Ordovo G hodnotí, jestli je okolí prvku odlišné od všech prvků. LISA se ptá, jestli je okolí prvku odlišné od všech hodnot a zároveň jestli je hodnota prvku odlišná od jeho okolí. Stejně jako ostatní prostorové statistiky, signifikance výsledků je ověřená pomocí randomizace permutací.

3 METODA MONTE CARLO

Metoda Monte Carlo je označení pro koncept numerického řešení relativně složitých matematických, fyzikálních a jiných problémů pomocí principu velkého množství opakovaných simulací náhodných pokusů (Fabian & Kluiber, 1998).

Základní myšlenka metody Monte Carlo spočívá v řešení problémů pomocí kombinace prvku náhodnosti experimentu a dostatečně velkého počtu opakování toho experimentu (simulace). Představme si situaci, kdy bychom neměli žádnou znalost teorie pravděpodobnosti, ale přece jen by nás zajímalo, s jakou pravděpodobností hodíme dvěma šestistěnnými kostkami v součtu hodnotu sedm. Ve skutečnosti víme, že existuje celkem 36 různých kombinací, z nichž šest vyústí v součet hodnot sedm. Matematicky bychom tedy dokázali spočítat, že pravděpodobnost je 6/36, tedy přibližně 16,67. Bez znalosti teorie pravděpodobnosti bychom tuto hodnotu neuměli spočítat, mohli bychom ji však zjistit prostým opakováním experimentu. Například bychom 100x hodili dvěma kostkami, a pokud by třeba 18x padl součet sedm, tvrdili bychom, že pravděpodobnost je 18 %. Postupným zvyšováním opakování počtu pokusů by se naše empiricky pozorovaná pravděpodobnost zpřesňovala a stávala by se více spolehlivou. Pokud zopakujeme pokus např. milionkrát, je už počet opakování náhodného pokusu dostatečně reprezentativní pro to, abychom pozorovanou pravděpodobnost považovali za stabilní a dostatečně přesnou. S nekonečným zvyšováním počtu opakování by se přibližovala k teoreticky určené hodnotě.

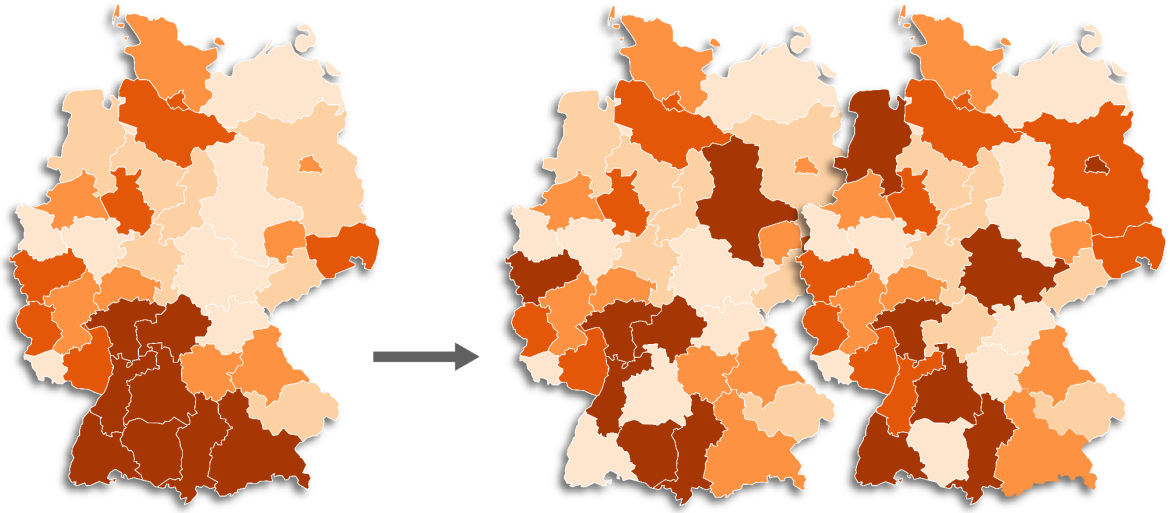
APLIKACE METODY NA PROSTOROVÁ DATA

Jakým způsobem tato simulační metoda souvisí s prostorovými daty? Význam metody Monte Carlo pro prostorová data je především při procesu určování statistické významnosti výsledků metod prostorové statistiky. Jak již bylo naznačeno v předchozích kapitolách, při práci s prostorovými daty většinou očekáváme, že sledovaný výběrový soubor je zároveň i celou populací. Pro data není používáno teoretické rozdělení, kterým by se dalo při vyhodnocování výsledků řídit, míra pravděpodobnosti je proto řešena simulací různých možných scénářů metodou Monte Carlo.

Princip si představíme na určení statistické významnosti vypočtené hodnoty Moranova I. Při hodnocení prostorové autokorelace předpokládá nulová hypotéza náhodné rozmístění hodnot sledované veličiny v prostoru, každá hodnota má stejnou pravděpodobnost vyskytovat se v libovolné části území a není nijak ovlivněná svým okolím. Kontrující alternativa naopak předpokládá prostorovou závislost (autokorelaci) výskytu hodnoty na svém okolí. Vypočtená testová statistika musí být porovnána s nějakým rozdělením, aby mohla být hypotéza vyhodnocena. Lze vycházet z normálního rozdělení testové statistiky, ale to samozřejmě nemusí být splněno a následné inference mohou být chybné.

Alternativním přístupem je simulace možných situací pro sestavení teoretického rozdělení a jeho následné porovnání se empiricky pozorovanou testovou statistikou. K takovému výsledku se dostaneme pomocí *randomizace permutací*: permutace jsou různé možnosti uspořádání, v případě prostorových dat se tedy jedná o všechny možnosti, jak pozorované

hodnoty náhodné veličiny uspořádat v zájmovém území. Jinými slovy – geometrické prvky (polygony) zůstávají fixní, pouze se na nich náhodně promíchávají hodnoty atributů (Obr. 15). Z každé takovéto permutace je možno vypočítat testovou statistiku I.



Obr. 15 Příklad originálních dat jevu střední délka života německých regionů NUTS 2 (vlevo) a jejich dvou náhodných permutací (vpravo)

Pokud ze všech možných permutací (kterých je $n!$ v případě souboru s n záznamy) náhodně vybereme dostatečně velké množství variant (náhodný výběr označujeme slovem *randomizace*), lze z nich sestavit náhodnou veličinu, jejíž rozdělení je využito jako *referenční rozdělení* pro pozorovanou sledovanou testovou statistiku. Tento přístup je dostatečně robustní i vůči potencionálním narušením předpokladů normality apod.

Z referenčního rozdělení pak dostáváme pseudo p-hodnotu:

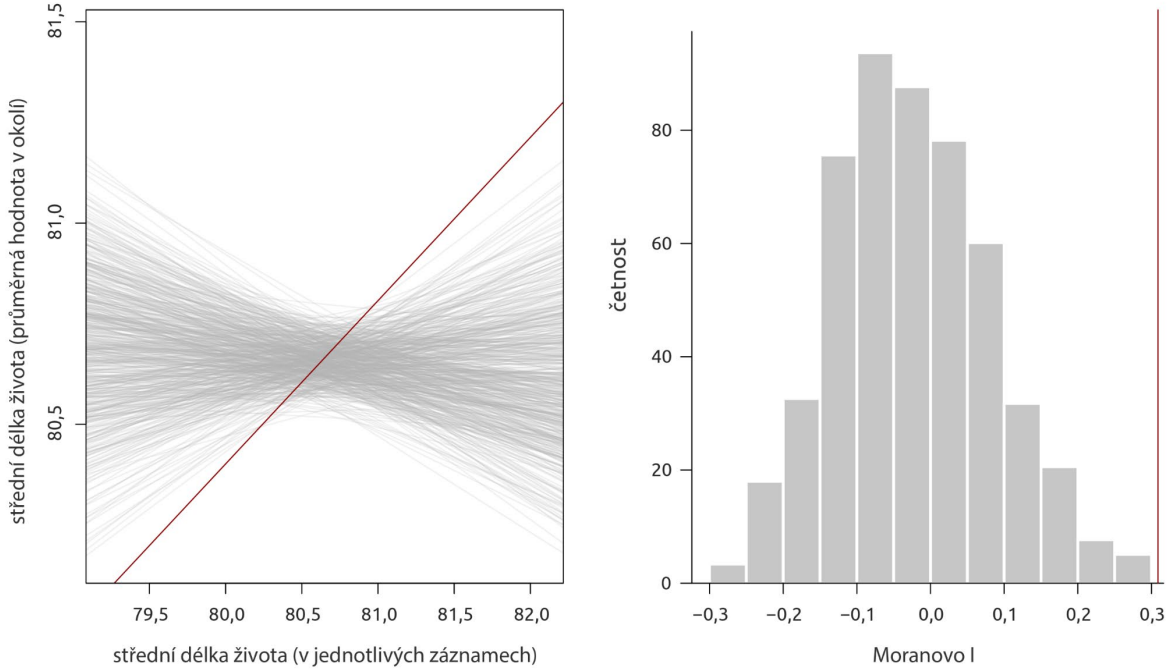
$$p = \frac{R + 1}{M + 1}$$

, kde M je celkový počet náhodných permutací, ze který se sestavuje referenční rozdělení a R je počet z těchto permutací, kdy byla pozorována hodnota větší než hodnota statistiky pro náš zájmový soubor. Pseudo p-hodnota je pouze shrnutím výsledků z referenčního rozdělení a neměla by být interpretována jako analytická p-hodnota v běžných statistických metodách. Kvalita ohodnocení závisí částečně na počtu permutací: např. vyjde-li p-hodnota 0,001 z 99 permutací, nelze jednoznačně určit, jestli je lepší nebo horší než p-hodnota 0,01 z 999 permutací.

Grafické znázornění porovnání hodnoty Moranova I pro skutečně pozorovaná data (červená linie) vůči 499 náhodným permutacím (šedé linie) je představeno na Obr. 16. Jednotlivé permutace dohromady formují teoretické rozdělení, které je zobrazeno také histogramem. Na histogramu je patrné, že skutečná pozorovaná hodnota Moranova I je dostatečně vysoká (extrémní), proto je vysoce nepravděpodobné, že by náhodná veličina byla v prostoru

rozmístěna náhodně. Zamítáme nulovou hypotézu o náhodném rozmístění a usuzujeme, že data vykazují autokorelaci.

Stejný princip je využíván např. při vyhodnocování významnosti Ripleyho K funkce.



Obr. 16 Simulace hodnot Moranova I pro 499 náhodných permutací proměnné střední délka života v německých regionech NUTS 2 (šedě) a Moranovo I pro původní pozorovaná data (červeně).

4 PROSTOROVĚ VÁŽENÉ METODY

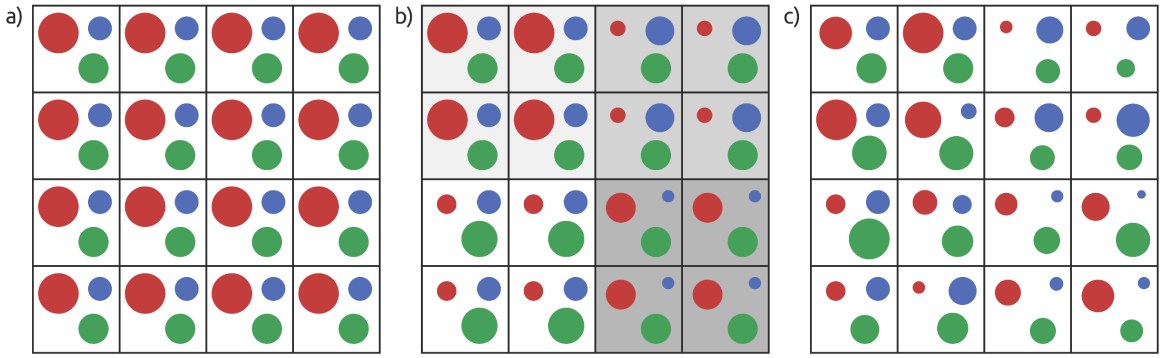
Z předchozí lekce o prostorové statistice jsme si odnesli základní poznatky o specifikách prostorových dat, která díky své prostorové složce sice přinášejí novou, obohacující informaci, zároveň však komplikují práci s daty, neboť narušují některé základní statistické předpoklady. Přítomnost efektů I. a II. řádu často může způsobovat, že výsledky zjištěné kvantitativní analýzou nebudou platné v celém sledovaném území. To by nebylo nijak zvláštní, např. z regresních modelů jsme si také vědomi faktu, že pro některé záznamy model nadhodnocuje, některé podhodnocuje, a že tuto informaci máme uchovanou v modelových residuích. Zkoumáním těchto residuí regresního modelu v kontextu geografického prostoru bychom pozorovali dvě možné situace: pokud by rozmístění kladných a záporných residuí bylo v prostoru náhodné, pak je vše v pořádku. S velkou pravděpodobností bychom však zjistili, že hodnoty residuí vykazují autokorelaci. Jejich shlukování indikuje, že odchylky od modelu jsou nějakým způsobem nenáhodné (prostorově závislé), tudíž se sestavený model chová v každé části sledovaného území trochu jinak. Proto je vhodné mít k dispozici postupy, jak se s problémem této prostorové závislosti vypořádat.

Výše popsané chování označujeme jako prostorovou *nestacionaritu* – výsledky libovolné analýzy nejsou prostorově homogenní (stacionární), nýbrž jsou závislé na konkrétní lokalitě, jsou rozdílné v různých částech území. Toto chování je důsledkem efektů I. a II. řádu, které jsou pro většinu prostorových dat typické. Matematicky může být nestacionarita vnímána jako další proměnná, kterou je vhodné zachytit a nějakým způsobem popsat.

Příčiny nestacionárního chování mohou být různé – většinou jsou důsledkem skutečně jiného chování jevu napříč územím, které je způsobeno různými lokálními kontextuálními faktory. Na vině může být ale také nevhodně navržený model, ve kterém např. chybí některé důležité vysvětlující proměnné, které by doplnily chybějící informace ústící v nestacionaritu. Poslední aspekt může být čistě numerický, náhodný. Ten by však nikdy neměl mít takovou sílu, aby vyústil ve výrazný prostorový vzor.

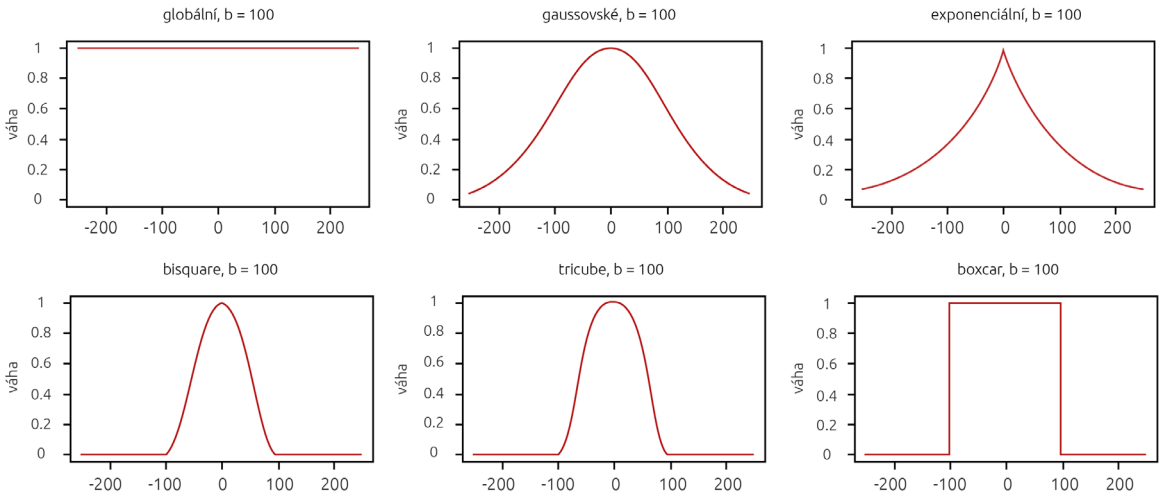
Jedním ze způsobů, jak se s nestacionaritou vypořádat, je aplikace *prostorově vážených metod* (dále bude používána zkratka GW metody vycházející z anglického označení *geographically weighted*). Jestliže předpokládáme, že sledovaný jev se chová rozdílně v různých subprostorech celého sledovaného území, řešením by bylo sledovat jev skutečně v těchto dílčích částech a hodnotit v každém subprostoru samostatně, a tímto způsobem variabilitu v prostoru zachytit. Pro každou jednotlivý záznam by byl použit samostatný model/odhad parametrů, který nejlépe vystihuje lokální podmínky a chování sledovaného jevu pouze v kontextu daného subprostorou. Zjednodušená aplikace této myšlenky může mít podobu překrytí území libovolným gridem, a následně hodnocení samostatně pouze skupiny prvků, které přísluší k jedné buňce gridu (jak je znázorněno na Obr. 17). Takovým přístupem jsou výsledky zkreslené volbou rozlišení gridu, je proto potřeba tuto myšlenku ještě trochu vylepšit. Aplikace GW metod je nejčastěji vztahována k polygonovým datům, které reprezentují prostorovou kontinuitu. V případě datové sady o n záznamech je pro každý z těchto n záznamů vytvořen lokální model pracující pouze s k záznamů z vymezeného sousedství. Lokální model je platný pouze pro momentálně hodnocený záznam, a díky

výpočtu založeném na vymezeném okolí skutečně zachycuje jen lokální chování sledovaného jevu. V nejjednodušším případě si můžeme představit příklad z kapitoly 2, kde byly představené prostorové klouzavé průměry (*spatially lagged variable*). Nejedná se o nic jiného než o lokální výpočet průměru založený na vymezeném okolí, tedy nejjednodušší formu GW metody.



Obr. 17 Rozdělení prostoru všech objektů pomocí gridu. V případě a) není aplikován žádný prostorový přístup, sestavený model (abstraktně reprezentován velikostí barevných kruhů) je všude stejný. V případě b) je prostor rozdělen na 4 plochy (různé odstíny šedi, které mohou představovat např. vyšší administrativní celky) a v každé z nich pozorujeme mírně odlišné chování. Poslední varianta c) je pouze prostorově detailnější. Je patrné, že sobě blízké skupiny jsou si také podobné vlastnostmi modelu.

Zásadním faktorem ovlivňujícím výpočet GW metod je způsob vymezení okolí. Z předchozí kapitoly známe základní vymezení sousedství pomocí topologického dotyku, *k* nejbližších sousedů anebo pomocí fixní vzdálenosti. Nyní budou tyto možnosti ještě rozšířeny o tzv. *jádrovou funkci* zahrnující vliv vzdálenosti – ne všechny hodnoty zahrnuté v sousedství zde mají stejnou váhu, ale s rostoucí vzdáleností jejich vliv klesá dle dané funkční závislosti. Tímto nástrojem lze přidat blízkým prvkům větší význam než prvkům vzdáleným. Nejčastěji používané jádrové funkce vykresluje Obr. 18.



Obr. 18 Různé typy jádrových funkcí při zachování stejného dosahu (vzdálenost b = 100 m)

Z hlediska terminologie se u GW metod často mluví o *pohyblivém jádře* (*moving spatial kernel*). Nejedná se však o nic jiného než o způsob vymezení sousedství. Analogicky k základním konceptům sousedství jsou rozlišovány dva typy jádra:

- *Fixní jádro* pracuje s pevně nastavenou vzdáleností, mění se pouze počet prvků, které při stanovené vzdálenosti do výpočtu vstupují. Tato metoda je vhodná pro jevy s přibližně pravidelným rozmístěním a podobnou plochou jednotlivých územních jednotek. Problémy vznikají v územích s nepravidelnými tvary a různými velikostmi ploch.
- *Adaptivní jádro* pracuje s konstantně vymezeným počtem prvků (k sousedů), pro zachování počtu prvků musí měnit velikost svého dosahu (poloměr jádra). Díky tomu je zajištěna konstantní kvalita výpočtu, při rozdílné velikosti sledovaných jednotek však může mít interpretace hodnoceného jevu obtížná.

Na oba typy jádra je pak aplikována zvolená jádrová funkce. Volba dosahu jádra (vzdálenost u fixního a počet sousedů u adaptivního) má zásadní vliv – čím větší má jádro dosah, tím větší shlazení výsledků zajišťuje. Dosah jádra má přitom větší vliv než volba tvaru jádrové funkce. V extrémním případě by bylo možné např. použít adaptivní jádro zahrnující všechny prvky ve sledovaném území, pouze vliv jádrové funkce by do výpočtu přinášel zohlednění prostorové variability. Naopak při volbě malých jader se může stát, že některé prostorově odlehlé lokality třeba nebudou mít žádného souseda, nebo budou mít tak málo sousedů, že provedený výpočet bude nespolehlivý a označený jako statisticky nevýznamný.

4.1 OPTIMALIZACE JÁDRA

Hlavní problém GW metod spočívá v otázce, jak vhodně vybrat typ, dosah jádra a jádrovou funkci. Tyto parametry mohou samozřejmě být určeny expertní znalostí výzkumníka. Pokud však taková znalost není k dispozici, může si výzkumník pomoci nějakým výpočetním kritériem, které na základně numerických indikátorů určí nejvhodnější parametry tak, aby např. minimalizovaly chyby a nepřesnosti modelu. Jako příklad si představme metodu *cross validace* a *AIC*.

Cross validace

Princip cross validace (CV) spočívá ve vypuštění jednoho prvku z datového souboru, natrénování hodnoceného modelu nad daty bez tohoto prvku, predikci chybějící hodnoty pomocí sestaveného modelu a následně porovnání predikované hodnoty s původní vypuštěnou hodnotou. V případě GW metod se provádí sestavení modelu na základě prvků z vymezeného okolí, vypuštěná hodnota je hodnota momentálně sledované lokality. Tento výpočet zopakuje pro všechny z n prvků v zájmovém území, a sleduje se celková hodnota predikované chyby podle vzorce:

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(h)]^2$$

, kde y_i je skutečná hodnota veličiny v místě i a $\hat{y}_{\neq i}(h)$ je predikovaná hodnota veličiny v místě i při použití velikosti jádra h. Následně se celý proces opakuje pro jiné nastavení hodnot sousedství a zjišťuje se, pro jaké sousedství má CV nejmenší chybu. Jako optimální je taková hodnota vzdálenosti (h), kde je suma odchylek minimální, což maximalizuje predikční schopnost modelu.

Adjustované Akaike informační kritérium (AIC)

Poněkud složitějším ukazatelem kvality modelu je adjustované Akaikeho informační kritérium (Akaike, 1974). AIC hledá kompromis mezi prediktivní sílou modelu a jeho složitostí. AIC bere v úvahu různý počet stupňů volnosti v různých modelech, díky čemuž je možné přesněji porovnávat jejich relativní výkony (Fotheringham et al., 2002). AIC se spočítá podle vztahu:

$$AIC_C(h) = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{\frac{n + tr(S)}{n - 2 - tr(S)}\right\}$$

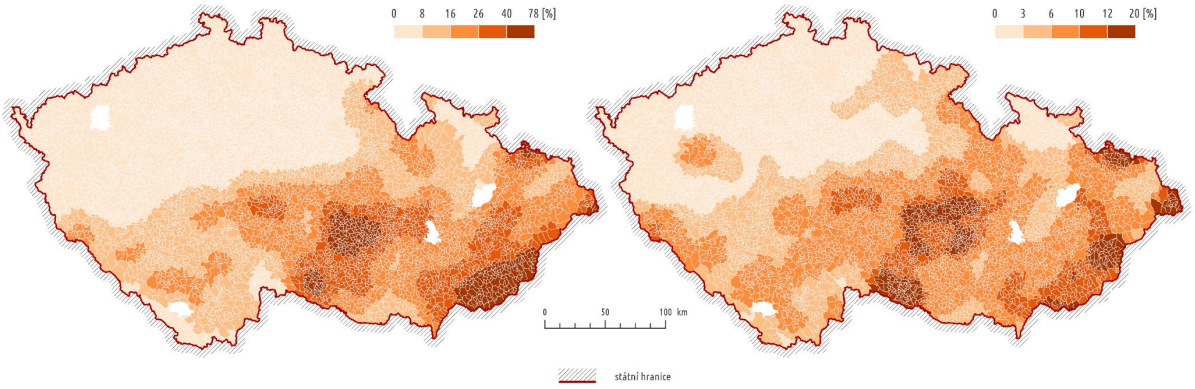
, kde n je velikost souboru, $\hat{\sigma}$ odhad směrodatné odchylky residuí, $tr(S)$ stopa matice popisující vztah predikované a pozorované hodnoty y. Bližší vysvětlení principu výpočtu AIC je představeno například v Fotheringham et al. (2002). Pro praktické využití je dostačující informace, že model s nejnižší hodnotou AIC je považován za nejvhodnější.

Metody AIC a CV (a případně i další numerické ukazatele) pracují na stejném principu – sledují kvalitu modelu pro konkrétní hodnotu velikosti jádra a následně pro zpracování analýzy vybírají takové nastavení, které ústí v nejkvalitnější model. Je důležité si uvědomit, že se jedná pouze o numericky optimální hodnotu, navržené nastavení sousedství tedy např. nemusí dávat smysl v kontextu zkoumaného jevu.

4.2 VYBRANÉ PROSTOROVĚ VÁŽENÉ METODY

Prostorově vážená deskriptivní statistika jedné proměnné

Základní deskriptivní statistika (charakteristiky polohy a variability) slouží pro zjednodušený popis datového souboru, kdy vstupní data nahrazujeme globálně platným ukazatelem. Pro prostorová data může být tato redukce až moc velká, jelikož se v ní ztrácí veškerá prostorová variabilita. Využitím prostorově vážené popisné statistiky s použitím průměru vlastně aplikujeme prostorové klouzavé průměry, můžeme takto zvýraznit prostorové trendy, nebo vyhladit hodnoty proměnné třeba pro potřeby vizualizace. Rozšířením od další statistické charakteristiky (jako je rozptyl nebo směrodatná odchylka) lze dalšími způsoby pozorovat nestacionaritu variability sledovaného jevu (Obr. 19).

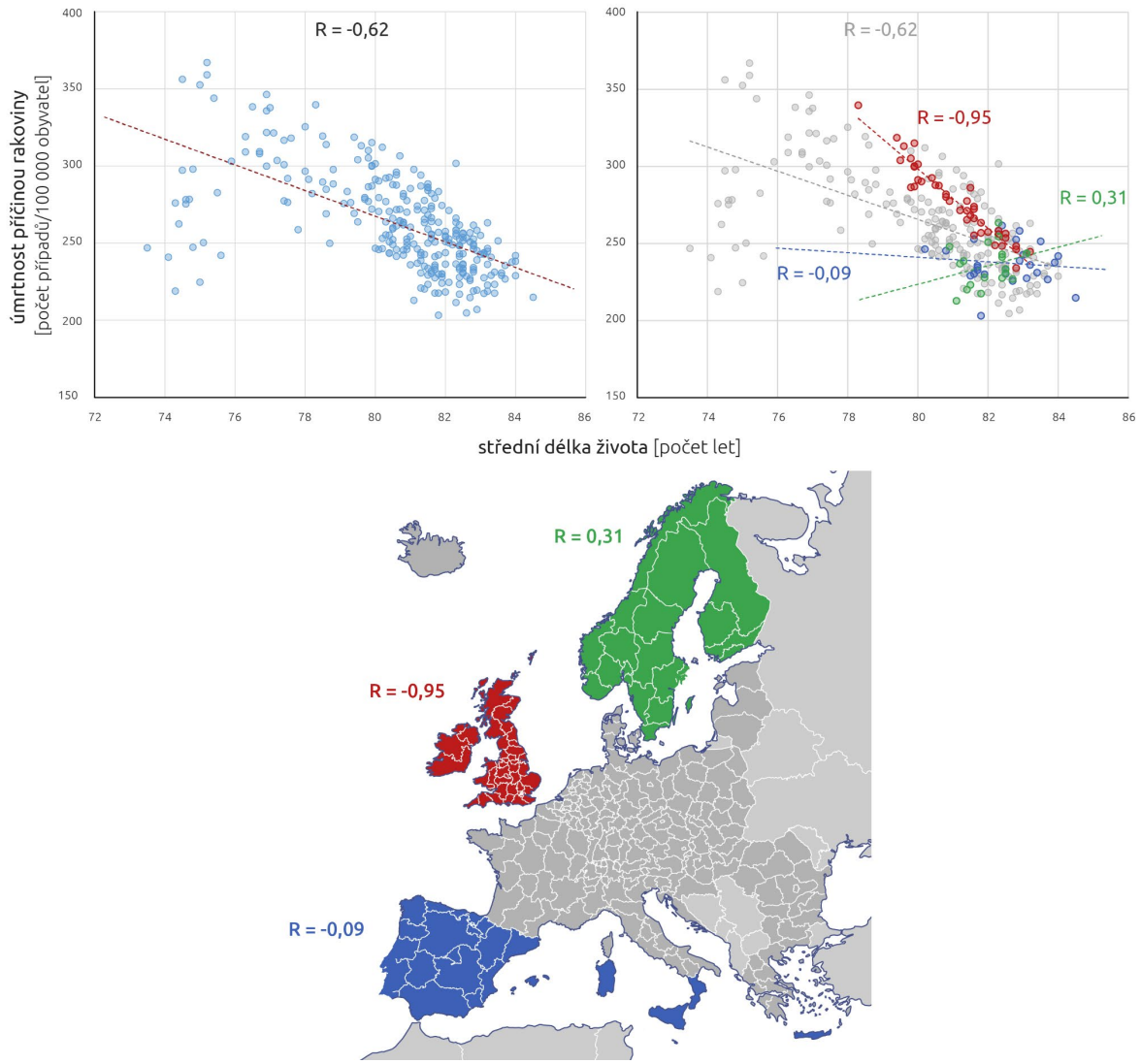


Obr. 19 Religiozita obcí ČR – prostorově vážený průměr (vlevo), prostorově vážená směrodatná odchylka (vpravo). Aplikováno fixní jádro o poloměru 11 km s bisquare jádrovou funkcí. Zdroj dat: ČSÚ, SLBD 2011

Prostorově vážená korelace dvou proměnných

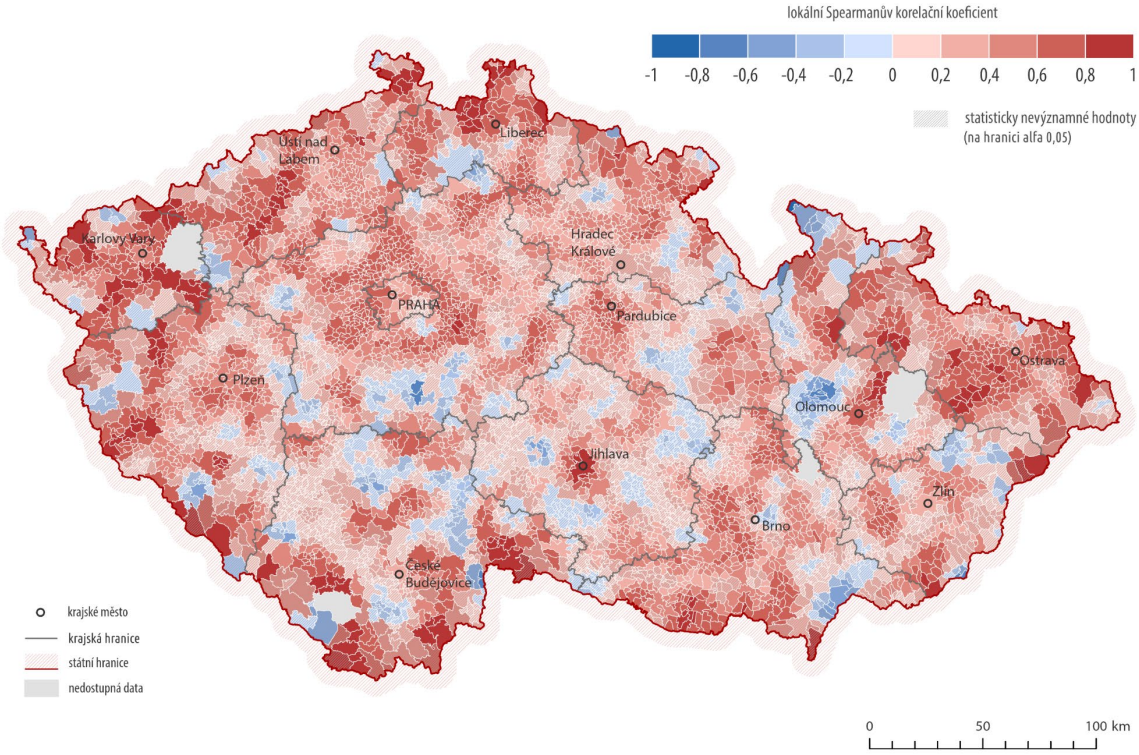
Také vztah dvou proměnných může být v prostoru proměnlivý. Pro základní motivaci může dobře posloužit Obr. 20, zobrazující vztah mezi střední délkou života a úmrtností příčinou rakoviny vyjádřený Spearmanovým korelačním koeficientem. V globálním trendu pozorujeme poměrně vysokou negativní korelaci (vlevo nahoře). V případě, kdy bychom identifikovali, že barevně odlišené podmnožiny záznamů jsou si také prostorově blízké (vpravo), pozorujeme, že tyto lokální hodnoty korelace se mohou výrazně lišit, mohou dokonce nabývat i zcela opačného trendu – což je případ zeleně označeného území regionů Skandinávie. Toto zjištění pak můžeme interpretovat slovy, že úmrtnost příčinou rakoviny není ve skandinávských regionech tak významnou příčinou úmrtí, proto jejím náhodným kolísáním vzniká pozitivní korelace s protichůdným jevem střední délky života.

Jiná situace může nastat tehdy, kdy hodnota globálního koeficientu vychází nízká nebo nevýznamná. Nicméně při lokálním průzkumu GW metodami bychom mohli narazit na lokality, kdy z důvodu nestacionarity opět nacházíme významné hodnoty korelace a tyto místa pak zasluhují větší pozornost při zkoumání závislostí jevů.



Obr. 20 Prostorová nestacionarita v případě korelace – Spearmanův korelační koeficient pro všechny záznamy NUTS 2 (vlevo nahoře), pro jednotlivé části území (vpravo nahoře a dole uprostřed)

Metoda prostorově vážené korelace je velmi jednoduchá a dokáže odhalit zajímavou variabilitu napříč územím. Je nutné si uvědomit, že hodnota korelačního koeficientu v konkrétní lokalitě *i* nevyjadřuje vztah dvou proměnných v tomto místě, ale v okolí toho místa. S tímto vědomím je nutné dále přistupovat k interpretaci. Je vhodné také neopomíjet hodnocení míry statistické významnosti, které se také pro jednotlivá pozorování liší. Nástroje pro prostorově vážené korelace jsou implementovány např. v R v balících *lctools* (Kalogirou, 2012), *GWmodel* (Gollini et al., 2015) nebo *GWpcor* (Percival & Tsutsumida, 2017). Ukázka výstupu prostorově vážené korelace je představena na Obr. 21.



Obr. 21 Příklad prostorově vážené korelace: vztah mezi procentuálním úspěchem SPOLU a podílem vysokoškolsky vzdělaných obyvatel ve volbách do Poslanecké sněmovny Parlamentu ČR v roce 2021. Bylo použito fixní jádro o dosahu 11 km s bisquare jádrovou funkcí.

S metodou korelace je úzce propojená regresní analýza. Není divu, že také regresní modelování má svoji prostorově váženou variantu – prostorově váženou regresi (GWR). Jelikož metoda prostorově vážené regrese je ze všech představených nejvíce používaná, bude podrobněji představena v další kapitole.

Prostorově vážená analýza hlavních komponent (GW PCA)

Prostorově vážená PCA není tak přímočará a snadná na interpretaci, jako např. prostorově vážené deskriptivní statistiky nebo korelace. Motivace pro metodu je však stejná – snaží se zachytit prostorovou variabilitu pomocí lokálních výpočtů PCA. V globální variantě PCA je sestaven jeden model, který redukuje vstupní data do nových, nekorelovaných dimenzí (komponent), které jsou seřazeny dle své významnosti a je vhodné je nějak interpretovat. Práce s lokální variantou je poněkud složitější, jelikož pro velikost souboru n by bylo nutné interpretovat n modelů PCA. GW PCA lze ale využít např. pro pozorování variability rozptylu, který lze pomocí PCA vysvětit. Lze se pak např. zaměřit na oblasti nízkého/vysokého vysvětlení, a ty zkoumat blíže. Dále lze pozorovat, jak se v prostoru mění dominance proměnných s největším vlivem na jednotlivé komponenty. Jelikož se komponenty vždy orientují ve směru největšího rozptylu, měla by právě dominance konkrétního atributu značit jeho velkou lokální variabilitu. Tato využití GW PCA představují ve svém článku Harris et al. (2015).

Poslední užitečný způsob využití GW PCA spočívá ve hledání lokálních outlierů. Identifikace odlehlých hodnot už byla popsána v kapitole 1.1, nyní bude představena jiná metoda, která nehodnotí data globálně, ale v lokálním měřítku. Postup je založen na metodě cross validace jednotlivých zpracovávaných záznamů: v lokálním modelu PCA jsou vždy vynechány hodnoty záznamu, který je momentálně hodnocen. Pomocí sestaveného modelu PCA jsou tyto hodnoty pak zpětně dopočítány. Porovnáním skutečné pozorované a modelované hodnoty vzniká míra rozdílnosti (tzv. *discrepancy*), kterou lze vyjádřit kvalitu lokálního modelu. Pokud je rozdíl velmi velký, znamená to, že se pozorovaný záznam výrazně odlišuje od svého okolí, a může být označen jako outlier. V procesu chybí metoda pro stanovení hraniční hodnoty odlehlosti, ale jelikož výsledkem je pouze jednorozměrný vektor discrepance, lze použít libovolnou jednorozměrnou metodou pro identifikaci odlehlých hodnot.

Softwarové možnosti výpočtu GW PCA jsou dost omezené, řešení nabízí např. již zmiňovaný balíček *GWmodel* pro R.

Kapitoly zaměřené na prostorovou statistiku a prostorově vážené metody se snaží zdůraznit výrazná specifika prostorových dat, která mohou komplikovat interpretaci a kvalitu výsledků standardních statistických metod. Pro zachycení lokální proměnlivosti je možné využít prostorově vážené metody, které konstruují n modelů na základě vymezeného okolí – jádra. Teoreticky může být každá metoda převedena na GW metodu, je jen nutné najít postup, jak matici prostorových vah aplikovat na samotný algoritmus výpočtu. Je nutné si uvědomit, že GW metody jsou nesmírně citlivé na parametrizaci: způsob vymezení sousedství, jeho velikost a použitou jádrovou funkci. Při různých nastaveních těchto parametrů vždy najdeme prostorovou variabilitu, někdy však může být zavádějící a někdy ji nelze jednoduše vysvětlit. Při používání GW metod by proto měla být věnována velká pozornost právě experimentování s parametrizací, metody je také vhodné kombinovat s jejich globální variantou pro porovnání získaných výsledků.

5 PROSTOROVÉ REGRESNÍ MODELY

Obecným úkolem regresního modelování je aproximovat data pomocí známého funkčního přepisu tak, aby bylo splněné nějaké vymezené kritérium přesnosti. Prostorové regresní modelování obohacuje klasické regresní modely o prostorový aspekt, jehož vliv na výsledek analýzy je většinou nezanedbatelný – jak už bylo demonstrováno na tématech předchozích kapitol, kdy byla řeč o efektech I. a II. řádu, a o prostorové nestacionaritě. Očekáváme, že důsledkem prostorových vztahů a kontinuity modelovaných jevů může docházet k prostorovým vlivům, kdy se hodnoty jevu ve dvou různých, ale geograficky sobě blízkých místech A a B vzájemně ovlivňují.

Připomeňme si pro začátek, jaké hlavní předpoklady od dat očekáváme při aplikaci regresních modelů:

Linearita: u regresního modelu očekáváme, že závisle proměnná *y* je dána lineární kombinací vstupních prediktorů. Není-li vztah mezi proměnnými skutečně lineární, výpovědní hodnota modelu je nedostatečná. Linearitu můžeme pozorovat na grafu vztahu mezi rezidui modelu a jednotlivými nezávisle proměnnými. Body by měly být rozloženy kolem nulové hodnoty, měly by vykazovat náhodné kolísání a neměly by se formovat v žádném specifickém tvaru, např. v podobě paraboly, která by indikovala kvadratickou závislost.

U lineárních modelů očekáváme **normalitu**, která má velký vliv na použití testů o parametrech modelu. Kromě ověření normality vstupních dat ji můžeme pozorovat také na rozdělení reziduí. Vykazují-li rezidua normální rozdělení, lze konstatovat, že model v predikci nenadhodnocuje, ani nepodhodnocuje a je celkově vyvážený.

Heteroskedasticita popisuje nežádoucí proměnlivý rozptyl reziduí v závislosti na změně hodnoty sledované veličiny. Pokud bychom např. sledovali vztah věku a příjmu na náhodném výběrovém souboru, s rostoucím věkem bude s největší pravděpodobností narůstat rozdílnost mezi jednotlivými osobami, které tvoří výběrový soubor. Obdobně v nekvalitním regresním modelu se mění rozptyl reziduí v závislosti na hodnotě modelované proměnné. Přítomnost heteroskedasticity může vyplývat přímo z charakteru sledovaného jevu, může být důsledkem přítomnosti odlehlých hodnot, anebo absencí důležitých vstupních vysvětlujících proměnných.

Poslední důležitým předpokladem, který je už přímo vztažený k prostorovým datům je **stacionarita** modelu. U nezávislého modelu očekáváme, že se bude chovat napříč celým zájmovým územím stejně. Pokud je však přítomna nestacionarita, poukazuje na proměnlivé chování modelovaného jevu napříč zájmovým územím. Naštěstí dokážeme nestacionaritu celkem úspěšně zachytit, a to pomocí pozorování reziduí v prostoru. Nezávislý model by měl různé hodnoty reziduí náhodně rozmístěné v prostoru, v opačném případě se budou podobné hodnoty shlukovat a rezidua budou vykazovat prostorovou autokorelaci. V takovém případě geografický prostor regresní model zkresluje, a je proto vhodné použít některou z technik prostorového regresního modelování, aby byl tento vliv potlačen a výsledný model skutečně prezentoval pouze ryzí chování závislosti zvolených proměnných.

Hlavní problém prostorového regresního modelování zůstává zakomponování prostorové závislosti do regresního vztahu, tedy konkrétní vyjádření v regresní rovnici. K problému může být přistupováno dvěma způsoby:

Globální řešení tvoří jeden globální model, ve kterém je pomocí prostorových vazeb prostorová závislost zahrnuta (modely typu SAR). V rámci globálních modelů se v ekonometrické teorii uvádí tři různé typy zachycení prostorových interakcí: (1) endogenní interakční účinky mezi závislou proměnnou, (2) exogenní interakční účinky mezi vysvětlujícími proměnnými a (3) interakční účinky mezi chybovými členy. Tyto modely jsou dále rozebírány v kapitole 5.

Lokální řešení pracuje pouze s lokálním výpočtem založeném na vymezeném okolí, fungující na již představeném principu geograficky vážených metod. Metoda geograficky vážené regrese je popsána v kapitole 5.2.

Pro bližší rozbory problémů prostorového regresního modelování lze nahlédnout do následujících prací (Anselin, 2003, 2021; Elhorst, 2010; Haining, 2003; INSEE Eurostat, 2018; LeSage & Pace, 2009).

5.1 GLOBÁLNÍ PROSTOROVÉ MODELY

Spatial Autoregressive model (SAR)

Jako SAR označujeme skupinu regresních modelů, jejichž hlavní myšlenkou je pro modelování sledované proměnné y v lokalitě i využít pouze hodnoty z blízkého okolí této lokality. Prediktorem jsou tedy pouze hodnoty pozorované proměnné v okolí, které jsou zachycené pomocí matice prostorových vah a parametrů. Pozor, nejedná se ale o GW přístup, konceptualizace se od GW metod odlišuje, protože výsledkem je pouze jeden globální model, platný pro celé území. Obecný zápis SAR modelu vypadá následovně:

$$y = \rho W y + \varepsilon$$

, kde autoregresní parametr ρ vyjadřuje prostorovou závislost a odhaduje se z matice prostorových vah W (řádkově standardizovány), ε je chybový vektor. Pro lepší představu je vhodné si uvést také vztah pro vyjádření konkrétního pozorování SAR modelu, které má podobu obdobnou ke *spatially lagged variable*:

$$y_i = \rho \sum_{j}^{n} W_{ij} y_j + \varepsilon_i$$

Pro autoregresní modelování se většinou používá jednoduché sousedství prvního řádu typu královna, není však problém nastavit libovolný koncept sousedství, který konkrétní software nabízí (např. v GeoDa lze aplikovat libovolnou vytvořenou matici prostorových vah). Tento obecný model (spíše jeho část) postrádá matici prediktorů a jejich regresní koeficienty, zabývá se pouze vyjádřením vztahu mezi predikované hodnotou proměnné y_i pomocí hodnot v jejím okolí y_j. Toto vyjádření se následně kombinuje s neprostorovým regresním modelem →

vzniká smíšený regresně-prostorový autoregresní model (*mixed regressive spatial autoregressive model - MRSA*), který má už své konkrétní implementace (modely SEM, SLM).

General nesting spatial model/Manskiho model (GNS)

Většina literatury postupně odvozuje jednotlivé prostorové regresní modely ze základního obecného vztahu, tzv. *General nesting spatial model* (nebo také *Manskiho model*). Ten má velmi komplexní podobu:

$$y = \alpha\iota_N + \rho Wy + \beta X + \theta WX + u = \alpha\iota_N + \rho Wy + \beta X + \theta WX + \lambda Wu + \varepsilon$$

, kde $\alpha\iota_N$ je konstantní parametr (trend); ρWy popisuje prostorové interakce proměnné y (jak bylo uvedeno pro obecný SAR model); βX je klasická matice prediktorů a jejich regresních koeficientů (stejně jako v neprostorové regresi); θWX zachycuje vnější prostorové interakce mezi prediktory a $u = \lambda Wu + \varepsilon$ jsou prostorově autokorelovaná rezidua. Podle toho, jak jsou vyjádřeny jednotlivé složky tohoto obecného modelu (konkrétně které z nich nabývají hodnoty 0), získáváme další varianty modelů.

Spatial Autoregressive Combined model (SAC)

Třída modelů SAC neobsahuje vyjádření prostorových interakcí mezi prediktory, ukryté pod sčítancem θWX. Pokud je tento prvek nulový, z obecného Manskiho modelu dostaneme následující variantu:

$$y = \alpha\iota_N + \rho Wy + \beta X + \boldsymbol{\theta WX} + \lambda Wu + \varepsilon =$$

$$= \alpha\iota_N + \rho Wy + \beta X + \lambda Wu + \varepsilon$$

Spatial Durbin Model (SDM)

Třída modelů SDM neobsahuje vyjádření prostorových autokorelovaných reziduí, ukryté pod sčítancem λWu. Očekává se pouze základní vektor náhodných reziduí ε. Všechny ostatní složky Manskiho modelu jsou zachovány (prostorové vlivy závisle a nezávisle proměnných). Zápis modelu má následující podobu:

$$y = \alpha\iota_N + \rho Wy + \beta X + \theta WX + \boldsymbol{\lambda Wu} + \varepsilon =$$

$$= \alpha\iota_N + \rho Wy + \beta X + \theta WX + \varepsilon$$

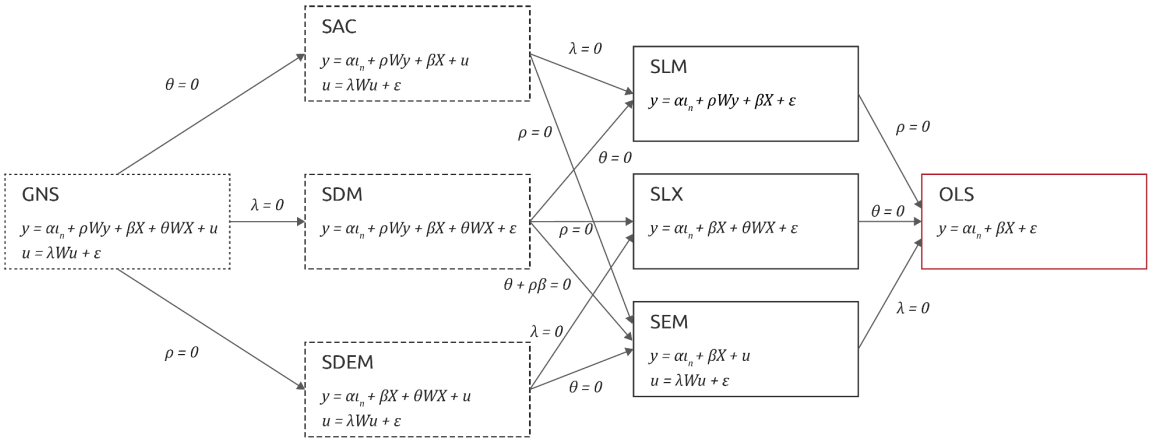
Spatial Durbin Error Model (SDEM)

Třída modelů SDEM neobsahuje prostorové interakce proměnné y odvozené ze SAR modelu a ukryté pod sčítancem ρWy. Prostorové interakce jsou zachovány na úrovni prediktorů (θWX) a prostorově autokorelovaných reziduí (λWu). Zápis modelu má následující podobu:

$$y = \alpha\iota_N + \boldsymbol{\rho Wy} + \beta X + \theta WX + \lambda Wu + \varepsilon =$$

$$= \alpha\iota_N + \beta X + \theta WX + \lambda Wu + \varepsilon$$

Z těchto stále ještě dost obecných modelů dále můžeme odvozovat jejich zjednodušené varianty tím, že některé parametry položíme rovny nule. Tím získáváme dva nejjednodušší modely, kterým budeme věnovat větší pozornost. Celou typologii modelů přehledně zachycuje schéma na Obr. 22.



Obr. 22 Třídění prostorových regresních modelů (převzato z: Halleck Vega & Elhorst, 2015)

Spatial Error Model (SEM)

Prostorový chybový model vnímá prostorový aspekt jako obtíž, která má negativní vliv na kvalitu sledovaných statistických inferencí, a proto se jej snažíme nějakým způsobem zbavit. Prostorové vztahy nejsou modelovány ani na úrovni závisle proměnné y, ani na úrovni matice prediktorů βX, ale jsou vyjádřeny v autokorelaci reziduí jako chybová složka.

$$y = \beta X + \varepsilon = \beta X + \lambda W\varepsilon + u$$

, kde X je matice prediktorů, β je vektor regresních koeficientů, ε je vektor prostorově autokorelovaných reziduí, který je dále rozložen na matici blízkosti W, vektor autokorelačních parametrů λ, vektor prostorově autokorelovaných reziduí u a vektor náhodných rezidui ε.

Prostorovou variabilitou u takového modelu popisujeme tu část modelu, kterou neumíme vysvětlit (residua). Model můžeme chápat jako neúplný – závislost poukazuje na to, že v něm nějaký prediktor chybí. Ten je však neznámý, neumím jej popsat, a proto tyto nevysvětlené vztahy vkládáme do reziduí. SEM model také neočekává výrazné interakce mezi zpracovanými proměnnými.

Spatial Lag Model (SLM)

V českém jazyce označujeme SLM jako prostorový intervalový model v podobě:

$$y = \beta X + \rho Wy + \varepsilon$$

, kde Wy je vektor prostorových vztahů závisle proměnné (základní SAR model), ρ je autoregresivní koeficient popisující vliv závisle proměnné v okolí na pozorovanou závislou proměnnou v místě i, X je matice prediktorů, β je vektor regresních koeficientů a ε je vektor reziduí s náhodným rozdělením.

Oproti SEM je v tomto případě prostorový aspekt vnímán jako podstata modelovaného jevu, a je proto snaha jej zachytit. Z tohoto důvodu je prostorová autokorelace vložena na úroveň vysvětlované proměnné, jedná se tedy o přímou aplikaci SAR modelu, pouze rozšířenou o matici vysvětlujících prediktorů. Pokud bychom se podívali hlouběji do výpočtu (Anselin, 2021), zjistíme, že ve skutečnosti není proměnná y v lokalitě i vysvětlována svým okolím, ale posloupností mocnin autoregresivního koeficientu ρ a matice prostorových vah.

Spatially Lagged X Model (SLX)

Regresní model s prostorově závislou proměnnou X (*spatially lagged variable*) má obecnou podobu:

$$y = \beta X + \theta WX + \varepsilon$$

, kde sčítanec θWX zachycuje prostorové interakce na úrovni prediktorů, X je matice prediktorů, β je vektor regresních koeficientů a ε je vektor rezidui s náhodným rozdělením. Tento typ modelu reflektuje vzájemné prostorové vlivy jednotlivých prediktorů, čím je schopný vyjádřit tzv. *spillover effect* (Halleck Vega & Elhorst, 2015). V porovnání s vlivem prostorového intervalu, spillover effect popisují autoři jako dopad změny vysvětlující proměnné x v jednotce i na závisle proměnnou y v jiné blízké jednotce j. Schopnost kvantifikovat míru tohoto vlivu je důležitým nástrojem při zkoumání regionálních vztahů určitých jevů.

Pokud by se u SEM a SLM prostorová složka opět vyloučila, výsledkem by byl jednoduchý neprostorový model lineární regrese. Pro řešení konkrétního problému můžeme uvažovat všechny představené modely. Hledáme pouze proces, který nám umožní vyhodnotit, která implementace je kvalitnější. Každý z modelů pojímá zachycení prostorových vazeb jiným způsobem, a volba nejvhodnějšího modelu závisí na zkušenosti výzkumníka a pochopení principu modelovaného problému. Jak uvádí Halleck Vega & Elhorst (2015), různé prostorové ekonometrické modely je obecně obtížné rozlišit bez předchozí znalosti skutečného procesu vzniku dat, který v praxi často není k dispozici.

Globální prostorové regresní modely nejsou v software příliš často implementovány. GeoDa nabízí pouze modelování SEM a SLM, pro ověření jejich vhodnosti nad zpracovávanými daty lze využít statistický test LaGrangeových multiplikátorů, a tak podpořit uživatelské rozhodování při volbě modelu. Při práci v R je k dispozici např. balík *spatialreg*, který v sobě má implementovány modely SEM, SLM a SLX.

5.2 LOKÁLNÍ PROSTOROVÉ MODELY

Geograficky vážená regrese (GWR)

Cíl geograficky vážené regrese je zcela stejný jako cíle jiných regresních modelů – vyjádřit způsob, jakým sada prediktorů ovlivňuje sledovanou závisle proměnnou. Zatímco předchozí skupina modelů založených na autoregresivním modelu fungovala jako jeden globální model (ve kterém byly zachyceny lokální proměnlivosti pomocí vymezení sousedství v prediktoru), GW regrese přejímá principy obecných GW metod, které byly představeny v dřívější kapitole 4. Neexistuje tedy jeden globální regresní model, ale pro n záznamů je sestaveno n lokálních regresních modelů, kde každý z nich je vypočten pouze na základě definovaného sousedství. Tímto způsobem dokáže model lépe zachytit prostorovou nestacionaritu a při vhodném nastavení mohou být některé lokální vztahy diametrálně odlišné od modelu globálního (viz tzv. Simpsonův paradox).

Pro konceptuální definici modelu uvádí Brunsdon et al. (1996) vztah:

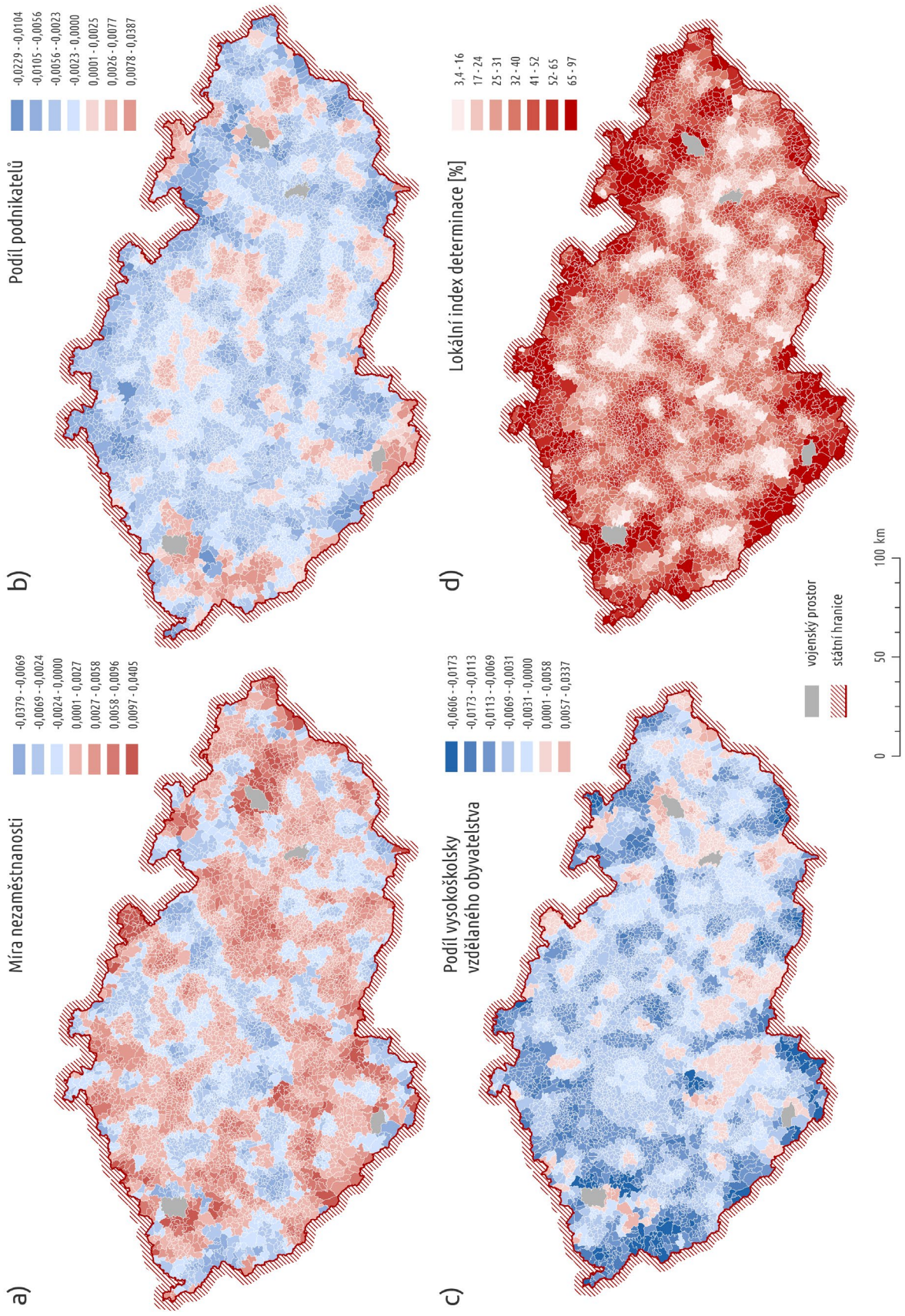
$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$$

, kde $\beta_k(u_i, v_i)$ je hodnota k-tého regresního koeficientu v lokalitě i o souřadnicích (u_i, v_i). Takto určené koeficienty jsou závislé na poloze a na svém okolí. Jelikož vliv sousedních bodů je přepočítán pomocí jádrové funkce, dochází k diskretizaci spojité funkce $\beta_k(u, v)$ pro lokalitu i. Lokální odhad probíhá váženou metodou nejmenších čtverců. V případě klasické OLS se jedná o minimalizační úlohu, kde je odchylka minimalizovaná pomocí hodnot koeficientů. V případě GWR je váha aplikovaná před hledáním minimální hodnoty, z čehož vyplývá proměnlivá odchylka a taky proměnlivá kvalita modelu.

GW regrese přejímá všechny výhody i nástrahy GW metod – musíme zde opět uvažovat nad parametrizací nastavení: volba fixního nebo adaptivního jádra, velikost jeho dosahu a vliv tvaru jádrové funkce (vliv tohoto parametru je nejslabší). V případě, že nejsme schopni expertně zvolit parametry jádra, GW regrese nabízí několik nástrojů pro jejich automatické určení – cross validace a adjustované Akaikeho informační kritérium. Tyto metriky kvality modelu jsou blíže popsány v kapitole 4.

Zatímco výstupem globálních regresních modelů je obecně jeden model se všemi svými parametry, v případě GWR dostáváme v každém zkoumaném prvku zájmového území samostatný model se všemi jeho výsledky (regresní koeficienty, index determinace R^2, atd.). Takto můžeme mapovat např. změny regresních koeficientů nebo změnu R^2 napříč zájmovým územím, a pozorovat variabilitu vztahů ve zkoumaném problému. Výsledkem je soubor odhadů lokálních parametrů pro každou jednotku zájmového území (regresní koeficienty, t-value) a lokální index determinace.

U GWR je nutné také ověřit spolehlivost vypočtených výsledů. Ta může někdy být nízká, třeba v důsledku volby malého sousedství, a tudíž malého počtu prvků ve výpočtu. Stejně jako u jiných prostorových metod, i zde se hojně využívá simulace metodou Monte Carlo.



Obr. 23 Výstupy prostorově vážené regrese: variabilita regresních koeficientů jednotlivých prediktorů (a–c) a lokálního indexu determinace (d) při modelování procentuálního úspěchu ANO v parlamentních volbách v roce 2021. Použito fixní jádro o poloměru 15 km, bisquare jádrová funkce.

Ukázkový příklad na Obr. 23 představuje výsledky GWR modelu mezi procentuálním úspěchem ANO v parlamentních volbách v roce 2021 a vybranými prediktory: míra nezaměstnanosti (NEZAM), počet podnikatelů (PODNIK) a zastoupení vysokoškolsky vzdělaných obyvatel v populaci (VYS). Neprostorový regresní model má následující podobu:

$$y = 0{,}0335 - 0{,}0048\,VYS - 0{,}0032\,PODNIK + 0{,}0029\,NEZM$$

, bylo dosaženo kvality $R^2 = 0{,}218$. Při použití GWR je jasně patrná prostorová proměnlivost všech prediktorů, v některé části území dokonce prediktory dosahují opačného znaménka než v obecném neprostorovém modelu. Je také zřejmé, že lokálním přístupem se často daří mnohem lépe popsat skutečné vztahy, což prokazují hodnoty R^2 vyšší než v globálním modelu (Obr. 23d).

6 ZÁVĚR

Předložený materiál se pokouší nabídnout teoretický přehled zaměřený na vybrané metody prostorové statistiky. Těmito tématy se zabývá jen pár autorů (a v českém prostředí obzvláště), proto si troufám tvrdit, že se jedná o ojedinělý studijní materiál.

Kombinace prostorových a statistických metod je pokročilý způsob, jak pracovat s geodaty. A především, je to způsob správný, jelikož neignoruje klíčový aspekt geodat, tedy jejich geografickou část. Oproti běžným, neprostorovým metodám je takto možné „vytěžit" z analyzovaných dat mnohem více informací, které společně s vhodnou vizualizací mohou přidat další díly do příběhu, který se datovou analýzou snažíme odhalit. Přestože rozšíření metod prostorové statistiky není příliš velké, jejich uplatnění dokládají příklady z geografie, ekonometrie, sociologie, demografie nebo politologie. Nyní je na vás, abyste k tomuto šíření dále přispívali.

Závěrem je nutno podotknout, že představený materiál rozhodně není kompletním výčtem všech existujících metod a neinformuje přesně o všech možnostech jejich nastavení. Na základě tohoto přehledu se však již čtenář může dále zaměřit na konkrétní analýzy a zdokonalovat se v nich samostudiem. Věřím, že současná literatura nebo internetové zdroje nabízí spoustu dílčích materiálů, obecně doporučuji sledovat např. práci zde několikrát citovaného Luca Anselina a jeho software GeoDa, nebo vybraná fóra k jazyku R (např. R-bloggers nebo RPubs), ve kterém lze řešit např. větší množství typů prostorových regresních modelů.

7 POUŽITÉ ZDROJE

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (1996). The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In *Spatial Analytical Perspectives on GIS* (pp. 111–125). Taylor and Francis.

Anselin, L. (2003). Spatial Econometrics. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics*. Blackwell Publishing Ltd.

Anselin, L. (2018). *Applications of Spatial Weights*. GeoDa: An Introduction to Spatial Data Science. https://geodacenter.github.io/workbook/4d_weights_applications/lab4d.html#creating-a-spatially-lagged-variable

Anselin, L. (2020). *Local Spatial Autocorrelation (1)*. GeoDa: An Introduction to Spatial Data Science. https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html

Anselin, L. (2021). *Spatial Models in Econometric Research*. https://doi.org/10.13140/RG.2.2.26447.20641

Barnett, V., & Lewis, T. (1978). *Outliers in Statistical Data*. J. Wiley & Sons.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Cliff, A. D., & Ord, J. K. (1973). *Spatial autocorrelation*. Pion Ltd.

Dawson, R. (2011). How Significant Is A Boxplot Outlier? *Journal of Statistics Education*, *19*(2).

Dixon, W. J. (1950). Analysis of Extreme Values. *The Annals of Mathematical Statistics*, *21*(4), 488–506. https://doi.org/10.1214/aoms/1177729747

Elhorst, J. P. (2010). Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*, *5*(1), 9–28. https://doi.org/10.1080/17421770903541772

Fabian, F., & Kluiber, Z. (1998). *Metoda Monte Carlo a možnosti jejího uplatnění* (1st ed.). Prospektrum spol s.r.o.

Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, *31*(5), 579–587. https://doi.org/10.1016/j.cageo.2004.11.013

Filzmoser, P., & Gregorich, M. (2020). Multivariate Outlier Detection in Applied Data Analysis: Global, Local, Compositional and Cellwise Outliers. *Mathematical Geosciences*, *52*(8), 1049–1066. https://doi.org/10.1007/s11004-020-09861-6

Filzmoser, P., Ruiz-Gazen, A., & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, *55*(1), 29–47. https://doi.org/10.1007/s00362-013-0524-z

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley Publishing, Inc.

Gelfand, A. E., Diggle, P. J., Fuentes, M., & Guttorp, P. (2010). *Handbook of Spatial Statistics*. CRC Press.

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, *24*(3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel : an R package for exploring spatial heterogeneity. *Journal of Statistical Software*, *63*(17), 1–50. https://doi.org/10.1080/10095020.2014.917453

Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

Haining, R., Wise, S., & Ma, J. (1998). Exploratory spatial data analysis in a geographic information system environment. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(3), 457–469.

Halleck Vega, S., & Elhorst, J. P. (2015). THE SLX MODEL. *Journal of Regional Science*, *55*(3), 339–363. https://doi.org/10.1111/jors.12188

Harris, P., Clarke, A., Juggins, S., Brunsdon, C., & Charlton, M. (2015). Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. *Geographical Analysis*, *47*(2), 146–172. https://doi.org/10.1111/gean.12048

Horák, J. (2015). *Prostorové analýzy dat* (6. vydání). VŠB-TU Ostrava, Hornicko-geologická fakulta, Institut geoinformatiky.

Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 36–43. https://doi.org/10.1002/wics.61

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Ltd.

INSEE Eurostat. (2018). *Handbook of Spatial Analysis*.

Kalogirou, S. (2012). Testing local versions of correlation coefficients. *Jahrbuch Für Regionalwissenschaft*, *32*(1), 45–61. https://doi.org/10.1007/s10037-011-0061-y

LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics* (3rd ed.). CRC Press.

Livings, M., & Wu, A.-M. (2020). Local Measures of Spatial Association. In J. P. Wilson (Ed.), *Geographic Information Science & Technology Body of Knowledge*. https://doi.org/10.22224/gistbok/2020.3.10

Longley, P. A., Goodchild, M., Maguire, D. J., & Rhind, D. W. (2005). *Geographic Information Systems and Science* (2nd ed.). John Wiley & Sons, Inc.

Macků, K. (2020). *Multidisciplinární hodnocení kvality života v Evropě na regionální úrovni*. Univerzita Palackého v Olomouci. https://doi.org/10.5507/prf.20.24458410

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, *2*.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley. https://doi.org/10.1002/0470010940

Moran, P. A. P. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society: Series B (Methodological)*, *10*(2), 243–251. https://doi.org/10.1111/j.2517-6161.1948.tb00012.x

Percival, J., & Tsutsumida, N. (2017). Geographically Weighted Partial Correlation for Spatial Analysis. *GI_Forum*, *1*, 36–43. https://doi.org/10.1553/giscience2017_01_s36

Ripley, B. D. (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(2), 172–192. https://doi.org/10.1111/j.2517-6161.1977.tb01615.x

Rossiter, D. G. (2020). *Exercise: Spatial Point Pattern Analysis*.

Rousseeuw, P. J., & Bossche, W. Van Den. (2018). Detecting Deviating Data Cells. *Technometrics*, *60*(2), 135–145. https://doi.org/10.1080/00401706.2017.1340909

Spurná, P. (2008). Prostorová autokorelace - všudypřítomný jev při analýze prostorových dat? [Spatial Autocorrelation - A Pervasive Phenomenon in the Analysis of Spatial Data?]. *Czech Sociological Review*, *44*(4), 767–788. https://doi.org/10.13060/00380288.2008.44.4.08

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. In *Economic Geography* (Vol. 46, pp. 234–240).

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addision-Wesley Publishing.

Unwin, D. J. (2009). Spatial Statistics. In *International Encyclopedia of Human Geography* (pp. 452–457). Elsevier. https://doi.org/10.1016/B978-008044910-4.00539-3

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press. https://doi.org/10.1201/9781420059496

Waller, L. A., & Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471662682

Yau, N. (2013). *Data Points - Visualization That Means Something*. John Wiley & Sons, Inc.

POKROČILÉ ZPRACOVÁNÍ GEODAT

Mgr. Karel Macků, Ph.D.

Odpovědná redaktorka Háta Kreisinger Komňacká

Předseda ediční komise PřF UP prof. RNDr. Jan Hlaváč, Ph.D.

Grafická úprava Jakub Koníček

Technická redakce Karel Macků

Publikace neprošla redakční jazykovou úpravou ve vydavatelství

Vydala Univerzita Palackého v Olomouci, Křížkovského 8, 771 47 Olomouc

Vydáno pro Katedru geoinformatiky PřF UP jako její 103. publikace

vydavatelstvi.upol.cz

1. vydání

Olomouc 2023

DOI: 10.5507/prf.23.24463209

ISBN 978-80-244-6320-9

VUP 2023-199

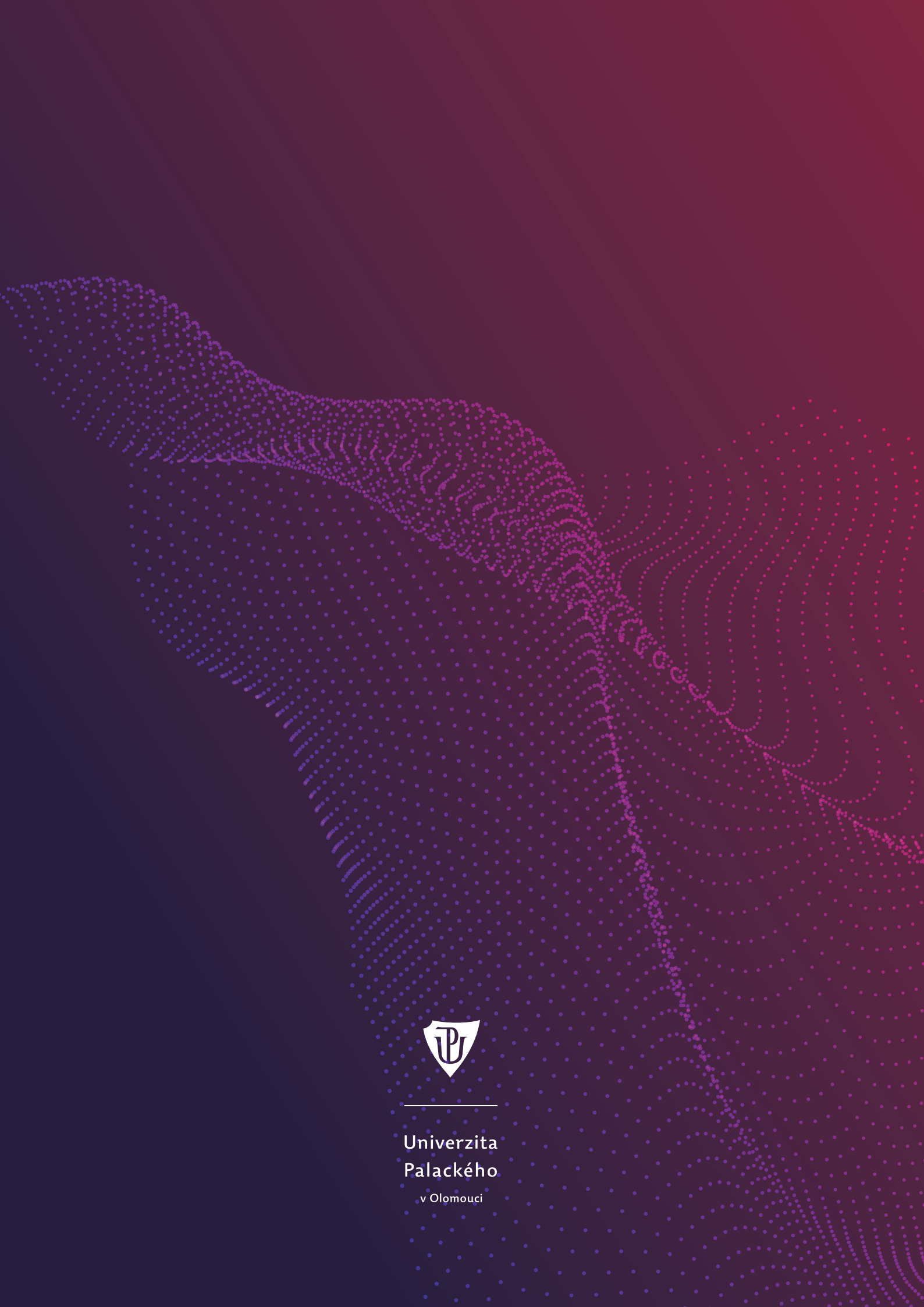
KATALOGIZACE V KNIZE - NÁRODNÍ KNIHOVNA ČR

Macků, Karel
Pokročilé zpracování geodat / Karel Macků. -- 1. vydání. -- Olomouc : Univerzita Palackého v Olomouci, 2023. -- 1 online zdroj
Nad názvem: Přírodovědecká fakulta, katedra geoinformatiky. -- Obsahuje bibliografii

ISBN 978-80-244-6320-9 (online ; pdf)

* 004.6-023.5 * 004.62 * 519.22-023.5 * 519.23/.24 * 004:91 * (075.8)
– prostorová data
– analýza dat
– prostorová statistika
– statistické metody
– geoinformatika
– učebnice vysokých škol

004.4/.6 - Programování. Software [23]
37.016 - Učební osnovy. Vyučovací předměty. Učebnice [22]



Univerzita
Palackého
v Olomouci