UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Prostorová analýza vybraných mortalitních dat v České republice

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Eva Fišerová, Ph.D.**
Vypracoval(a): **Adam Čech**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2021

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Adam Čech

Název práce: Prostorová analýza vybraných mortalitních dat v České republice

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2021

Abstrakt: V této bakalářské práci se zaměříme na popis prostorových charakteristik, mezi které patří prostorová autokorelace a prostorová nestacionarita, pro které si uvedeme metody jejich určení. Podrobně si představíme metodu Moranova I kritéria, která umožňuje zkoumat prostorovou autokorelaci, a v případě prostorové nestacionarity si blíže představíme metodu prostorově vážené regrese, uvedené metody budou aplikované na vybraná mortalitní data.

Klíčová slova: prostorová autokorelace, prostorová nestacionarita, prostorová vážená regrese, Moranovo I kritérium

Počet stran: 77

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Adam Čech

Title: Spatial analysis of selected mortal data in the Czech Republic

Type of thesis: Bachelor thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2021

Abstract: In this bachelor's thesis we will focus on the description of spatial characteristics, which include spatial autocorrelation and spatial nonstationarity, for which we present methods for their determination. We will introduce in detail the method of Moran I criterion, which will allow us to examine the spatial autocorrelation and in the case of spatial nonstationarity we will introduce in more detail the method of geographically weighted regression. The methodology will be applied to selected mortality data.

Key words: prostorová autokorelace, prostorová nestacionarita, prostorová vážená regrese, Moranovo I kritérium

Number of pages: 77

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením paní doc. RNDr. Evy Fišerové, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Úvod **7**

1 Charakter a analýza prostorových dat **8**
1.1 Prostorová autokorelace 8
1.1.1 Moranovo I kritérium 10
1.1.2 Matice vah 20
1.2 Prostorová stacionarita a nestacionarita 31
1.2.1 Prostorově vážená regrese 33
1.2.2 Umělé proměnné 41

2 Popis dat **45**
2.1 charakteristiky zkoumaných jevů 49

3 Prostorová analýza mortalitních dat **58**
3.1 Prostorová autokorelace 59
3.2 Prostorová nestacionarita 63

Závěr **71**

Literatura **72**

Poděkování

Touto cestou bych rád poděkoval vedoucí bakalářské práce doc. RNDr. Evě Fišerové, Ph.D za metodické vedení a konstruktivní připomínky při koncipování práce. Děkuji také Mgr. Michalu Lehnertovi, Ph.D. a Mgr. Davidu Fiedorovi, Ph.D. za jejich připomínky a poskytnutí dat k analýze.

Úvod

Standardní analytické metody umožňují zkoumat a popisovat zákonitosti světa kolem nás, např. je možno zkoumat výšku potomka v závislosti na výšce rodičů. Pokud si však uvědomíme, že každý náhodný jev se vyskytuje v určitý čas na daném místě *geografického prostoru*, můžeme standardní analytické metody rozšířit o zmíněné aspekty. V této práci si blíže popíšeme pouze metodiku umožňující zahrnut prostorový faktor do analýzy. Proto si podrobně rozepíšeme základní prostorové charakteristiky, mezi které patří prostorová autokorelace a prostorová nestacionarita. Pro uvedené charakteristiky si popíšeme i metody k jejich určení, v případě prostorové autokorelace si podrobně popíšeme metodu Moranova I kritéria, která umožní zkoumat prostorovou autokorelaci. U prostorové nestacionarity si uvedeme metodu prostorově vážené regrese, pomocí které budeme schopní říct, zda je zvolený model vhodný pro celé zkoumané území. Následně budou uvedené metody aplikované na prostorová data, které obsahují informace ohledně mortality, kvality ovzduší a klimatických podmínek na území České republiky.

Kapitola 1

Charakter a analýza prostorových dat

Prostorová data lze odlišit od standardních (resp. neprostorových) dat pomocí toho, že mimo obsahu informace ohledně libovolných *faktorů* (např. počet obyvatel v daném městě, rozloha města atd.) obsahují i informaci ohledně *polohy v mapovém díle*, tj. *zeměpisnou šířku* a *délku*, pomocí kterých lze výskyt statistických znaků zaznamenat, jako bod v mapovém díle. S uvedenou interpretací se však v praxi často nesetkáme a to z důvodu, že jsou zkoumané statistické znaky zaznamenávány v určitých úrovních státní správy (např. obce, kraje atd.), tzn. že provádíme agregování prostorového faktoru. V následujícím textu budou úrovně státní správy značeny, jako *prostorové jednotky*.

Z uvedeného rozdílu mezi jednotlivými typy dat je zřejmé, že důvodem analýzy prostorových dat je zjistit, zda poloha (resp. prostorový faktor) ovlivňuje hodnotu zkoumaného statistického znaku. Abychom však byly schopni zkoumat prostorový vliv na zkoumaná data je nutné si v první řadě zadefinovat prostorové charakteristiky, které je možné zkoumat a následně pomocí jejich definic upravit standardní statistické metody, které nám umožní danou charakteristiku zkoumat.

1.1. Prostorová autokorelace

Mezi základní prostorové charakteristiky je řazena *prostorová závislost* (resp. *prostorová autokorelace*), která vyjadřuje korelaci jedné a té samé náhodné veli-

činy (X) v závislosti na prostorovém umístění. Pro lepší představu je uveden příklad, ve kterém je analyzována závislost mezi průměrnou výškou mužské populace (cm) a krajem České republiky, ve kterém daný muž žije. Při zkoumaní prostorového vlivu na výšku mužské populace se využívá předpokladu, který se nazývá *Toblerův první zákon geografie* a má následující znění „*všechno souvisí se vším, ale blízké věci spolu souvisejí více než-li věci vzdálené*“.

Abychom mohli uvedenou myšlenku převést na uvedený příklad, je nutné si zvolit *výchozí prostorovou jednotku* od které budeme zkoumán prostorový vliv tzn. určit prostorově blízké (resp. sousedící) prostorové jednotky. Pro uvedený příklad můžeme např. za výchozí prostorovou jednotku zvolit Moravskoslezský kraj a za sousedící jednotky označit Olomoucký a Zlínský kraj (obrázek 1.1).

Obrázek 1.1: Prostorové rozložení průměrné výšky mužské populace.

Ve své podstatě budeme hledat závislosti mezi průměrnou výškou v Moravskoslezský kraji (tj. výchozí prostorová jednotka) vzhledem k průměrné výšce mužské populace žijící v Olomoucký a Zlínský kraj (tj. sousedící prostorové jednotky). Z mapového díla, které je uvedeno na obrázek 1.1 je patrný rozdíl ve výšce mužské populace žijící na západě a východě České republiky. Proto lze tvrdit, že na

území České republiky se vyskytuje trend shlukování mužské populace s přibližně stejnou průměrnou výškou. Uvedená charakterizace nám tedy říká, že v určitých krajích se vyskytují statistické znaky, které buď zvyšují (resp. snižují) průměrnou výšku mužské populace, např. zvýšené škodlivé látky obsažené ve vodě, množství stresu a další možné veličiny. Aby však uvedené závěry nebyly pouze subjektivní pohled, je nutné uvedenou hypotézu ověřit pomocí vyčíslení míry seskupení.

Prostorová analýza je soubor specifických statistických metod, které umožňují zkoumat prostorové charakteristiky mezi které patři . Uvedené charakteristiky si blíže popíšeme a následně si ukážeme metody pro jejich určení. Aby byly uvedené prostorové charakteristiky dobře pochopeny budou, uvedené metody aplikovaný na vybraná mortalitní data. Pro výpočet míry seskupení se obvykle využívá *Moranovo I kritérium, Gearyho C kritérium* a *obecná G statistika*. Z uvedených metod bude blíže popsáno Moranovo I kritérium, které představuje nejčastější používaný ukazatel.

1.1.1. Moranovo I kritérium

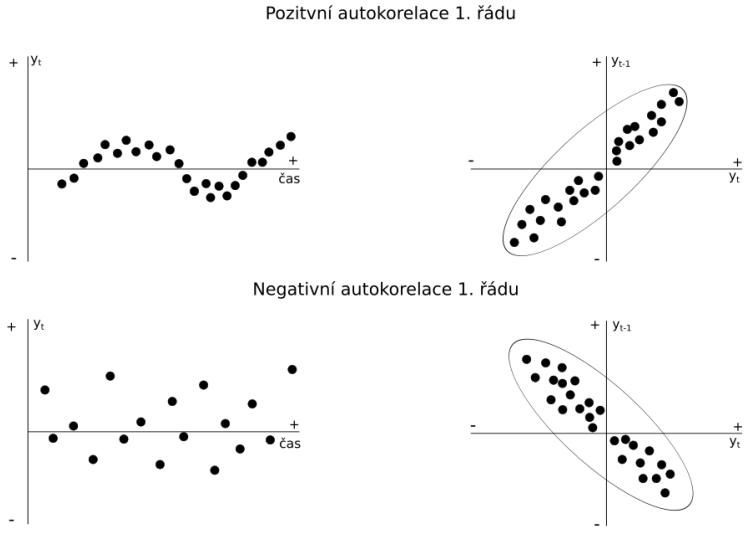
Protože při prostorové autokorelaci je zkoumaná zavilost jednotlivých pozorování náhodné veličiny X mezi sebou, je vhodné si nejprve uvést princip *časové autokorelace*, která se vyskytuje např. v autoregresních modelech časových řadách. Uvedený princip, umožňuje vyjádřit míru závislosti mezi jednotlivými pozorováními náhodné veličiny Y v čase t (y_t) a pozorováními posunutými o danou časovou periodu zpět, např. dnešní hodnota směnného kurzu se odvíjí od hodnoty včerejšího směnného kurzu. Tento vztah můžeme zapsat pomocí operátoru zpoždění prvního řádu ve tvaru $Ly_t = y_{t-1}$, který vyjadřuje posun o jednu časovou periodu. Protože se závislost může projevit až za delší časovou periodu, lze operátor zpoždění zobecnit na *k-tý* řád $L^k y_t = y_{t\text{-}k}$, $k = 1, ..., n\text{-}1$, kde n značí počet pozorování. Uvedenou problematiku lze formálně zapsat pomocí autokorelační funkce *k-tého* řádu

$$cor\,(Y_t, Y_{t\text{-}k}) = \frac{cov\,(Y_t, Y_{t\text{-}k})}{\sqrt{var\,(Y_t)\,var\,(Y_{t\text{-}k})}},\ k = 0, 1, ..., n\text{-}1.$$

Vzhledem k tomu, že hodnoty které potřebujeme pro výpočet závislosti mezi jednotlivými pozorováními jsou neznámé, bude pro odhad závislosti využitá výběrová *autokorelační funkce*

$$g_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}, \; n \in \mathbb{N}, \; k = 1, 2, ..., n\text{-}1 \; ,$$

ve které proměnná $\bar{y}$ znázorňuje výběrový průměr časové řady. Výsledná hodnota autokorelační funkce vyjadřuje sílu lineární závislosti, a může nabývat hodnot z intervalu $\langle$-1, 1$\rangle$, jako v případě Pearsonova korelačního koeficientu. Pokud je výsledná hodnota záporná, jedná se o negativní autokorelaci, která odpovídá situaci, že mezi jednotlivými pozorováními převažuje klesající lineární závislost. V opačném případě se jedná o pozitivní autokorelaci, která vyjadřuje převažující rostoucí lineární závislost (obrázek 1.2). Pokud hodnota autokorelační funkce je rovná nule, pak se mezi zkoumanými daty nevyskytuje žádná lineární závislost daného řádu.



Obrázek 1.2: Možné výsledky autokorelační funkce.

Při analýze prostorových dat však nebude zkoumaná autokorelace časového, ale prostorového faktoru. Proto je nutné uvedenou autokorelační funkci modifi-

kovat, aby bylo možno zkoumat prostorové shluky, tzn. lineární závislost mezi jednotlivými prostorovými jednotkami. Problém je však v tom, že časový faktor je pouze jednorozměrný ($\mathbb{R} \to \mathbb{R}$) a proto umožňuje posun (resp. zpoždění) pouze v jednom směru. V případě prostorového faktoru ($\mathbb{R}^2 \to \mathbb{R}$) je možné posun (resp. zpoždění) provést ve směru všech světových stran (obrázek 1.3), a proto je nutné specifikovat, kterým směrem se má prostorové zpoždění provést. K tomuto účelu je využit princip *Toblerova prvního zákona geografie*, který je vyjádřen pomocí proměnné w_{ij}, určující míru (resp. váhu) podobnosti mezi *i-tou* a *j-tou* prostorovou jednotkou. Samotné prostorové zpoždění vyjádříme pomocí vzorce $Lx_i = \sum_{j=1}^{n} w_{ij}x_j$, které lze zapsat i pro všechna pozorování (resp. prostorové jednotky) prostřednictvím maticového zápisu $L\mathbf{x} = \mathbf{W}\mathbf{x}$, kde vektor $\mathbf{x} = (x_1, x_2, ..., x_n)'$ představuje jednotlivá prostorová pozorování a $\mathbf{W}$ značí *matici vah*

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix}.$$

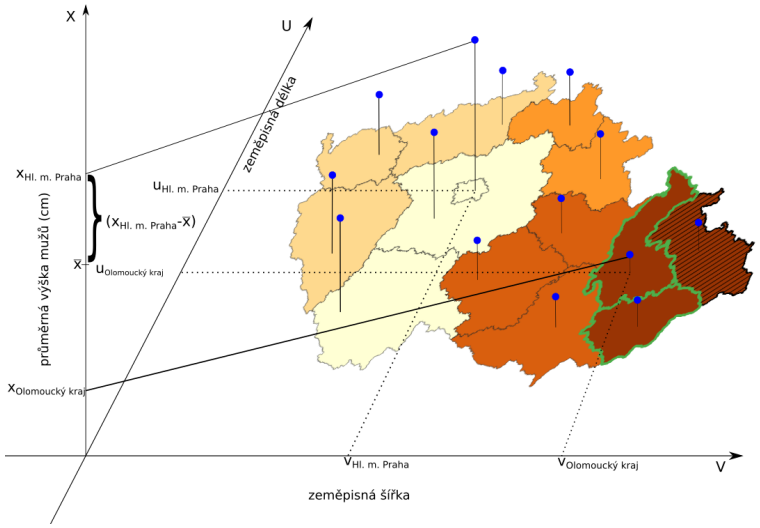
Aby však mohla být matice $\mathbf{W}$ označena za *matici vah*, musí splňovat následující podmínky

1. Matice vah je *symetrická* ($w_{ij} = w_{ji}$), protože vzdálenost od *i-té* k *j-té* prostorové jednotce je vždy stejná.

2. Matice vah obsahuje na *hlavní diagonále samé nuly* ($Stopa\,(\mathbf{W}) = 0$), a to z důvodu, že samotná prostorová jednotka se prostorově neovlivňuje.

3. V případě, že jsou hodnoty vah standardizované, musí být splněna *normalizační podmínka*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = 1.$$



Obrázek 1.3: Interpretace prostorových zpoždění.

U prostorové autokorelace je možné za určitých podmínek uvažovat operátory zpoždění vyšších řádů, stejně jako v případě časové autokorelace. Aby bylo možné uvažovat operátory vyšších řádů, je nutno vhodně zvolit typ sousedství, která budou blíže popsány v podkapitole *1.1.2* a proto problematiku vyšších řádu přenecháme na uvedenou podkapitolu.

Prostorovou autokorelaci můžeme vyjádřit ve tvaru *Pearsonova korelačního koeficientu*

$$I = cor(X, WX) = \frac{cov(X, WX)}{\sqrt{var(X)var(WX)}}.$$

I v tomto případ nejsou hodnoty k výpočtu prostorové autokorelace známé, a proto budou odhadnuty ze zkoumaných dat. Například mějme n pozorování, u kterých je měřen libovolný statistický znak X, konkrétně se může jednat o počet obyvatel vyskytující se v jednotlivých městech České republiky. Vzhledem k tomu, že jednotlivá města mají rozdílnou rozlohu, musí se hodnoty náhodné veličiny standardizovat, k tomu bude využit *výběrový průměr* a *výběrový rozptyl*. Aby bylo možno odhadnout hodnotu výběrového rozptylu, je nutné prvně vypočítat *výběrového průměru*, pomocí následujícího vzorce

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Výslednou hodnotu následně využijeme pro odhad *výběrového rozptylu*, který vypočteme následovně

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x} \text{ - } \mathbf{1_n}\overline{x})^2 = \frac{1}{n}(\mathbf{x} \text{ - } \mathbf{1_n}\overline{x})'(\mathbf{x} \text{ - } \mathbf{1_n}\overline{x}) = \frac{1}{n}\mathbf{u'u}, \text{ kde } \mathbf{1_n} = (1_1, ..., 1_n)'.$$

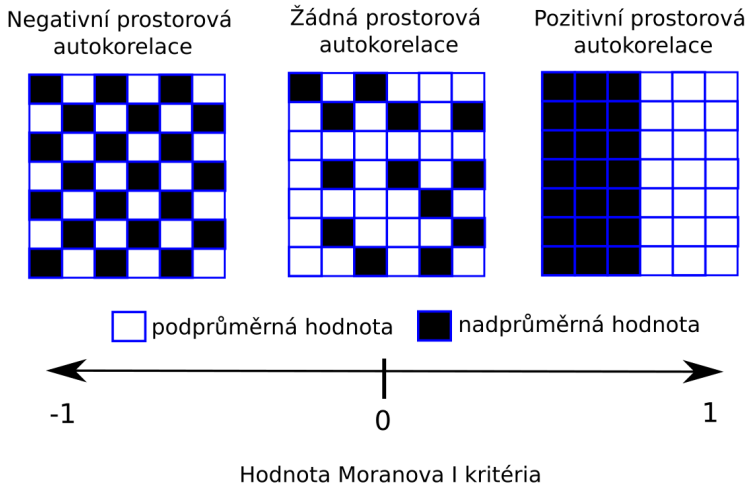
Pro standardizaci náhodné veličiny X se využije základní možná standardizace značená, jako *z-skóre*

$$\mathbf{z} = \frac{\mathbf{x} \text{ - } \mathbf{1}\overline{x}}{\sigma} = \frac{\mathbf{u}}{\sigma}.$$

Nyní jsou veškeré potřebné hodnoty odhadnuty, a proto můžeme vyjádřit odhad prostorové autokorelace pomocí *Moranová I kritéria*

$$I = \mathbf{z'Wz} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i \text{ - } \overline{x})(x_j \text{ - } \overline{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n}(x_i \text{ - } \overline{x})^2}, \ i \neq j.$$

Z uvedeného vzorce je zřejmé, že Moranovo I kritérium vyjadřuje prostorovou autokorelaci pro celé zkoumané území a výsledná hodnota i v tomto případě nabývá hodnoty z intervalu $\langle$-1, 1$\rangle$. Pokud výsledná hodnota nabývá záporného znaménka, pak prostorový charakter náhodné veličiny označujeme, jako *negativní prostorovou autokorelace*, tzn. že na zkoumaném území převažují prostorové shluky, ve kterých výchozí prostorová jednotka má nadprůměrnou (resp. podprůměrnou) hodnotou a je obklopena prostorovými jednotkami s podprůměrnou (resp. nadprůměrnou) hodnotou náhodné veličiny. Uvedená charakterizace je uvedena v levé části obrázku 1.4, která vizualizuje negativní prostorovou autokorelaci pokud hodnota Moranova I kritéria je rovná -1.



Obrázek 1.4: Grafické znázornění možnosti prostorové autokorelace.

V případě, že se na zkoumaném území nevyskytuje žádná prostorová závislost, pak hodnota globálního kriteria je rovná nule. Tato situace je vyobrazena uprostřed obrázku 1.4 znázorňující náhodné rozdělení zkoumaného statistického znaku. Poslední situace, která může nastat se označuje jako *pozitivní prostorová autokorelace* a nastává v případě, kdy výsledná hodnota Moranova I kritéria je kladná. V tomto případě na zkoumaném území převládají prostorové shluky, ve kterých mají výchozí prostorové jednotky nadprůměrnou (resp. podprůměrnou) hodnotou a jsou obklopeny prostorovými jednotkami s nadprůměrnou (resp. podprůměrnou) hodnotou zkoumané náhodné veličiny. V pravé části obrázku 1.4, lze pozorovat pozitivní prostorovou autokorelaci za předpokladu, že hodnota Moranova I kritéria je rovná 1.

Vzhledem k náhodné povaze Moranova I kritéria je nezbytné vypočtenou hodnotu statisticky ověřit. Testována je nulová hypotéza $H_0 : I = 0$, která říká „*na zkoumaném území se nevyskytuje prostorová autokorelace*“. Oproti tomu alternativní hypotéza $H_A : I \neq 0$ se vyjádří následovně „*na zkoumaném území se vyskytuje prostorová autokorelace*“. Uvedené hypotézy se mohou testovat dvěma způsoby, buď pomocí simulace *Monte Carlo* (permutační test) a nebo pomocí *z-testu* (statistický test). Abychom však mohli pro otestování uvedených hypo-

též využít *z-test* je nezbytně nutné, aby hodnota prostorové autokorelace pocházela z normálního rozdělení. Ovšem uvedený předpoklad není v praxi dosažitelný (viz [2]), a proto si blíže popíšeme pouze simulaci *Monte Carlo*. Při simulaci Monte Carlo pro ověření signifikantnosti prostorové autokorelace postupujeme následovně:

1. Náhodně přiřadíme každé prostorové jednotce jednu hodnotu zkoumané náhodné veličiny.

2. Pro vytvořený náhodný datový soubor se vypočte Moranovo I kritérium.

3. Pozorovaná hodnota Moranovo I kriteria se porovná s hodnotou vypočtenou v druhém kroku.

Uvedené tři kroky se opakují podle zvoleného počtu permutací (M), který se volí podle osobního uvážení. Ovšem pomocí principu, který je znám pod názvem „*zákon velkých čísel*“ lze tvrdit, že čím více permutací bude provedeno tím bude odhad rozdělení náhodné veličiny přesnější. Pro určení signifikantnosti pozorovaného Moranova I kritéria se využívá p-hodnota, která se vypočte pomocí vzorce

$$\text{p-hodnota} = \frac{M_{ex}}{M},$$

kde M_{ex} vyjadřuje počet simulovaných Moranových I hodnot, které jsou extrémnější než pozorovaná Moranova I kritéria, tj. hodnoty vyskytující se na chvostech odhadnutého rozdělení. Výslednou p-hodnotu lze interpretovat, jako pravděpodobnost mylného zamítnutí nulové hypotézy, které nastává právě tehdy, když je p-hodnota menší, jak zvolená hladina významnosti α. Jak již bylo zmíněno, výsledná hodnota Moranova I kritéria popisuje prostorovou autokorelaci pro celé zkoumané území, a proto lze Moranovo I kritérium označit, jako *globální* charakteristiku. Nedostatek uvedené charakteristiky je v tom, že neumožní pozorovat rozmístění prostorových shluků na zkoumaném území.

Proto si uvedeme metodu *lokálního indikátoru prostorové asociace* (LISA), která ve své podstatě rozloží Moranovo I kritérium do jednotlivých prostorových jednotek zkoumaného území. Princip uvedené metody je v tom, že pro každou prostorovou jednotku je vypočtena hodnota Moranovo I kritérium pouze z hodnot sousedících prostorových jednotek tzn. z vytvořené podmnožiny prostorových jednotek. Nutné je však podotknout, že při výpočtu je fixovaná *globální hodnota výběrového průměru* a *globální hodnota směrodatné odchylky*. Uvedená metoda se nazývá *lokální Moranovo I kritérium* a vypočte se následovně

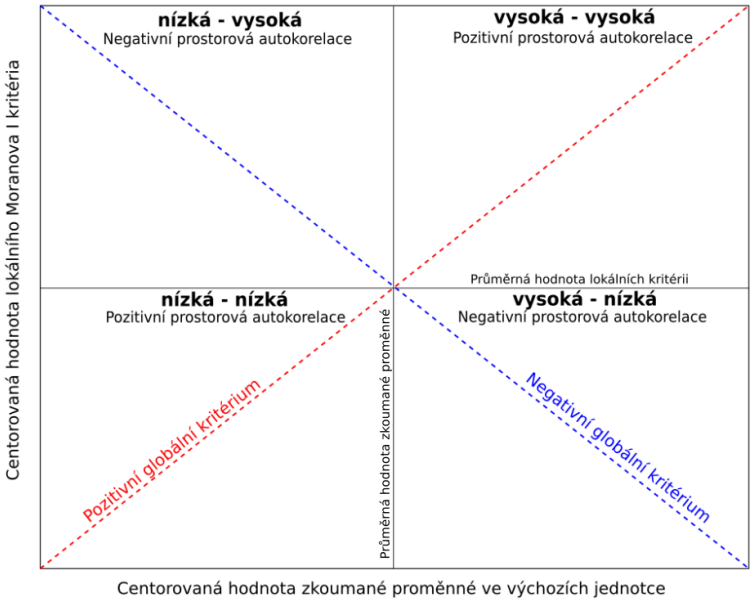
$$I_i = z_i \sum_{j=1}^{n} w_{ij} z_j = \frac{n\,(x_i \text{ - } \overline{x}) \sum\limits_{j=1}^{n} w_{ij}\,(x_j \text{ - } \overline{x})}{\sum\limits_{l=1}^{n} (x_l \text{ - } \overline{x})^2}, \; i \neq j.$$

Lokální Moranovo I kritérium je úměrné hodnotě globálního kritéria [2], neboť platí

$$\frac{\sum\limits_{i=1}^{n} I_i}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}} = \frac{n \sum\limits_{i=1}^{n} \left((x_i \text{ - } \overline{x}) \sum\limits_{j=1}^{n} w_{ij}\,(x_j \text{ - } \overline{x}) \right)}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij} \sum\limits_{i=1}^{n} (x_i \text{ - } \overline{x})^2} = I$$

Výsledné hodnoty lokálního kritéria, mohou nabývat jakékoliv hodnoty z množiny reálných čísel ($I_i \in \mathbb{R}$), protože výběrový průměr a výběrová směrodatná odchylka je převzata z globálního kritéria. Jelikož *lokální Moranovo I kritérium* vyjadřuje pouze prostorový shluk v okolí výchozí jednotky, nelze výslednou hodnotu interpretovat jako v případě globálního kritéria. Aby bylo možno využít interpretaci, která byla uvedena pro globální kritérium, je nutné výsledné hodnoty lokálního kritéria zobrazit do *Moranova diagramu* (obrázek 1.5). V diagramu jsou na horizontální ose zaznamenány centrované hodnoty *zkoumané proměnné* ve výchozích jednotkách a na vertikální ose centrované hodnoty *lokálního Moranova I kriteria*. Centrování se provádí z důvodu rozlišitelnosti vysokých (resp. nízkých) hodnot, u kterých je možno určit, zda se jedná o odlehlé pozorování,

a to pomocí *Cookovy vzdálenosti*. Aby bylo možné vypočítat *Cookovu vzdálenost* (D), je nutné hodnoty uvedené v Moranově diagramu proložit regresní přímkou, jak můžeme vidět na obrázku 1.5, kde jsou znázorněny pomocí modré a červené přímky.



Obrázek 1.5: Moranův diagram.

Matematický zápis pro regresní přímky vyjádřeme, jako $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, kde proměnná y_i znázorňuje hodnotu i-*te* vysvětlované proměnné, tj. centrovaná hodnota lokálního Moranova I kritéria, dále proměnná x_i znázorňuje hodnotu i-*te* vysvětlující proměnné, tj. centrovaná hodnota zkoumané proměnné ve výchozí jednotce a ε_i vyjadřuje náhodnou chybu. Obecně platí, že regresní koeficient β_0 znázorňuje konstantní posun regresní přímky, tzn. kde přímka protne x-*ovou* osu. Protože se v Moranově diagramu vyskytují pouze centrované hodnoty zkoumaných proměnných, bude vždy platiti, že regresní přímka protne osu x v počátku, který je tvořen průnikem přímek znázorňující průměrné hodnoty. Regresní koeficient β_1 obecně vyjadřuje směrnici regresní přímky, ale v případě Moranova diagramu bude zároveň platit, že hodnota směrnice je rovná hodnotě globálního Moranova I kritéria, což je dáno výše uvedeným vztahem mezi lokálním a globál-

ním kritériem. Pro odhad regresních koeficientů využijeme metodu nejmenších čtverců, kterou si blíže popíšeme v podkapitole 1.2., v tuto chvíli bude dostačující o uvedené metodě pouze znát to, že minimalizuje hodnotu náhodné chyby. Protože skutečnou náhodnou chybu nelze pozorovat, využívá se k jejímu pozorování reziduum (resp. odhad náhodné chyby), které pro *i-te* pozorování vypočte následovně $e_i = y_i - \hat{y}_i$, kde hodnota proměnné $\hat{y}_i$ vyjadřuje odhad hodnoty *i-te* vysvětlované proměnné. Pro výpočet Cookovy vzdálenosti budeme potřebovat znát rozptyl (resp. odhad rozptylu) vysvětlované proměnné, který je zároveň rovný rozptylu náhodné chyby a výpočet provedeme pomocí následujícího vzorce

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n-2}.$$

V tuto chvíli již známe veškeré proměnné, které využijeme při určení Cookovy vzdálenosti a výpočet pro *i-te* pozorování určíme pomocí tohoto vzorce

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_{(-i)j} - \hat{y}_j\right)^2}{2\hat{\sigma}^2}.$$

Jediný parametr, který je nutné dodatečně objasnit je $\hat{y}_{(-i)j}$. Uvedený parametr vyjadřuje, že při odhadu regresních koeficientů nebylo využito *i-te* pozorování a proto při porovnání odhadnutých hodnot vysvětlované proměnné, dokážeme určit vliv *i-te* proměnné na výslednou hodnotu regresní přímky. V případě, že je Cookova vzdálenost větší než jedna, pak se *i-te* pozorování označeno za odlehlé pozorování.

Hlavní nedostatek Moranova digramu je v tom, že neumožňuje zachytit prostorový faktor uvedených shluků. Uvedený problém lze vyřešit přidáním kategoriální proměnné, která bude zaznamenávat, ve kterém kvadrantu Moranova digramu se prostorová jednotka vyskytuje. Následně již stačí hodnoty vytvořené kategoriální proměnné interpretovat pomocí mapového díla, které bude znázorňovat, jak prostorové rozmístění shluků, tak i autokorelační hodnotu v jednotlivých prostorových jednotkách. Vzhledem k rozdílné charakterizaci lokálního a globál-

ního kriteria je nutné při testování signifikantnosti lokálního Moranova I kritéria pozměnit tvar nulové hypotézy, která je ve tvaru „*v okolí i-té prostorové jednotky se nevyskytuje prostorová autokorelace*“ a alternativní hypotéze „*v okolí i-té prostorové jednotky se vyskytuje prostorová autokorelace*“. Aby bylo možné pro ověření statistické významnosti lokálního Moranova I kritéria využít simulaci *Monte Carlo*, je nutné provést určité modifikace.

První úprava se týká zafixování hodnoty náhodné veličiny výchozí jednotky, která odpovídá proměnné x_i ve vzorci lokálního Moranova I kritéria. Důvodem je vzájemný vztah mezi proměnnou I_i a hodnotou náhodné veličiny ve výchozí jednotce. Uvedená změna bude mít pouze dopad na počet hodnot, které lze náhodně přiřadit k sousedícím jednotkám (tj. n - 1 možností). Další úprava se týká zohlednění *problému mnohonásobného testování hypotéz*, které vyjadřují problém s rostoucí pravděpodobnosti získání falešně pozitivního výsledku, vůči narůstajícímu počtu testovaných hypotéz. K tomuto účelu využijeme korekční proceduru, která pro zamítnuti (resp. nezamítnuti) nulové hypotézy bere v úvahu i celkový počet provedených testů. Nejjednodušší korekční procedura je *Bonferroniho procedura*, která zamítá nulovou hypotézu pravě tehdy, když je p-hodnota menší nebo rovna hodnotě $\frac{\alpha}{v}$, kde hodnota α vyjadřuje zvolenou hladinu významnosti testu, která se nejčastěji rovná hodnotě 0,05 (resp. 0,01) a proměnná v znázorňuje počet provedených testů. Uvedená metoda je však velmi konzervativní, a proto se pro zohlednění problému mnohonásobného testování využívá např. *Benjaminova-Hochbergova procedura*, která zamítá nulovou hypotézu v případě, že je p-hodnota testovaných hypotéz menší než hodnotě $\frac{i}{v}\alpha$, kde i značí pozici vzestupně seřazených p-hodnot.

Při vypracování uvedených kapitol byly využity tyto zdroje: [1], [4], [6], [9], [13], [14], [15], [19], [21].

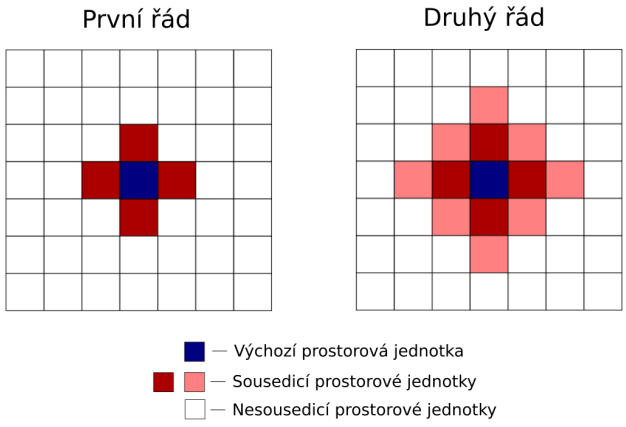
1.1.2. Matice vah

Vzhledem k tomu, že matice vah **W** určuje, které prostorové jednotky spolu sousedí (resp. nesousedí), lze ji označit za nejdůležitější parametr při výpočtu

prostorové autokorelace, proto si ukážeme podrobný postup pro výpočet matice vah, u kterého se řešit dva základní problémy. První z nich je určení metody, která umožní rozhodnout, zda prostorové jednotky i a j jsou si prostorově blízké ($w_{ij} > 0$) a druhý problém se týká číselného ohodnocení vah, které vyjadřuje prostorovou podobnost $i - t$ a $j - t$ prostorové jednotky.

Pro definování prostorově blízkých jednotek se využívá dvou základních metod, které se nazývají *sousedství* a *vzdálenosti*. V případě metody založených na *sousedství* jsou za sousedy považovány prostorové jednotky, které přímo sdílejí alespoň část společné geografické hranice o nenulové délce. K určování sousedů se nejčastěji využívá dvou principů, které jsou inspirovány pohyby šachových figurek po šachovém poli.

1. *Metoda sousedství věže:* Prostorová jednotka i sousedí s prostorovou jednotkou j, pokud spolu sdílejí část společné hranice, ale nikoliv jen jeden bod. Metodu lze specifikovat pomocí řádů, které umožňující zachytit širší okolí sousedství (obrázek 1.6).

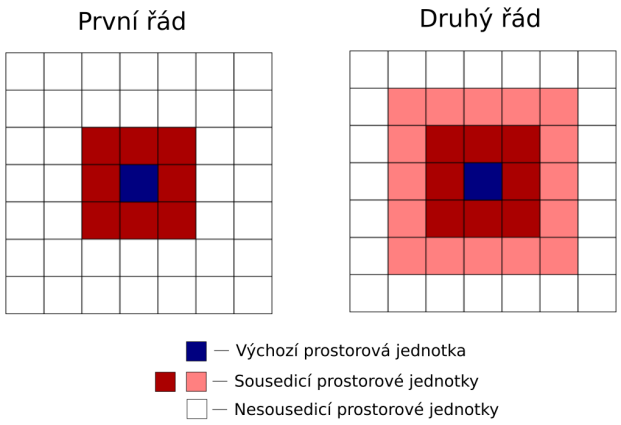


Obrázek 1.6: Princip metody *sousedství věže*.

Pro první řád jsou za sousedy *i-té* (resp. výchozí) prostorové jednotky označeny ty prostorové jednotky, které přímo sdílejí hranici s *i-tou* jednotkou.

V případě druhého řádu jsou za sousedy *i-té* jednotky považovány prostorové jednotky, které sdílejí hranici se sousedy definované v prvním řádu (obrázek 1.6). Výši řádu lze libovolně zvětšovat, ovšem maximální počet sousedů je omezen počtem jednotek na zkoumaném území. Z uvedené charakteristiky by mělo být zřejmé, že pokud je pro definování sousedství využita metoda sousedství, pak lze uvažovat operátory zpoždění vyššího řádu, jako v případě časové autokorelace.

2. *Metoda sousedství královny:* Prostorovou jednotku i a j označíme za sousedy, pokud sdílejí alespoň jeden bod společné hranice (obrázek 1.7). I v tomto případě lze uvažovat sousedy vyšších řádů.

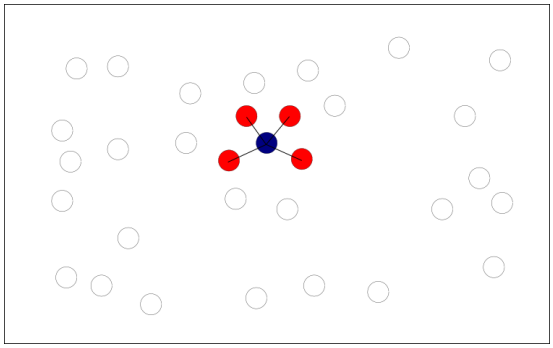


Obrázek 1.7: Princip metoda *sousedství královny*.

Druhou možností pro definování sousedství je pomocí *vzdálenosti* mezi zkoumanými prostorovými jednotkami. Aby bylo však možné měřit vzdálenost, je nutné určit neměnné body označující pozici v geografickém prostoru jednotlivých prostorových jednotek. Nejčastěji se volí geografický[1] a nebo populační střed[2] v dané jednotce. Hlavním rozhodujícím kritériem, pro určení středu prostorové jednotky je charakter zkoumané náhodné veličiny. V tomto případě se pro určení sousedství využívá opět dvou základních principů.

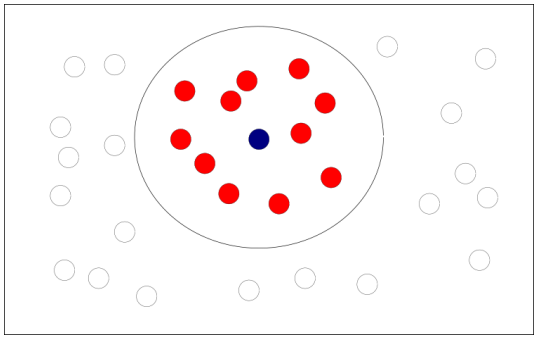
[1] Jedná se o geometrický střed zkoumaného území.
[2] Střed je určen místem, kde se vyskytuje největší hustota zalidnění.

1. *Metoda k-nejbližších sousedů:* Princip metody je založen na srovnávání vzdálenosti od *i-té* prostorové jednotky k zbylým zkoumaným prostorovým jednotkám. Za sousedící prostorové jednotky je označeno prvních k $(k > 0)$ jednotek s nejmenší vzdáleností od výchozí jednotky (obrázek 1.8).



Obrázek 1.8: Princip *k-nejbližších sousedů*, pro $k = 4$.

2. *Metoda prahové vzdálenosti:* Prostorová jednotka j je označena za souseda prostorové jednotky i, pokud vzdálenost mezi těmito jednotkami je menší, nebo rovna zvolené prahové vzdálenosti (d). Uvedenou metodu si lze představit jako kruh s poloměrem rovným prahové vzdálenosti a se středem v *i-té* prostorové jednotce. Veškeré prostorové jednotky, které náleží do vytvořeného kruhu jsou označeny za sousedy *i-té* jednotky (obrázek 1.9).



Obrázek 1.9: Princip *prahové vzdálenosti*.

V tuto chvíli již můžeme přejít k dalšímu kroku definování matice vah, a to k číselnému vyjádření vztahu mezi sousedy, tj. přiřazovat číselné hodnoty k proměnné w_{ij}. Pro určení prostorových vah se však nevyužívají libovolné funkce, ale

takové, které zohledňující princip, který je uveden v *Toblerově prvním zákonu geografie*. Matici vah lze rozdělit na dva typy, a to podle toho, jakých hodnot proměnná w_{ij} může nabývat, buď se jedná o *binomickou* a nebo *spojitou* matici vah. V případě *binomické váhové matice* mohou proměnné (resp. váhy) w_{ij} nabývat pouze dvou hodnot 0 a 1, tzn. že *i-tá* a *j-tá* prostorová jednotka spolu nesousedí právě tehdy, když $w_{ij} = 0$ a naopak v případě, že $w_{ij} = 1$ spolu sousedí. Binomickou matici vah je možné dále rozdělit podle metod, které jsou využité při definování sousedství. Pokud je využitá metoda *sousedství*, lze pro určení vah binomické matice využít pouze jediný princip.

1. *Váhy prostorové souvislosti:* Hodnota váhy je přiřazena podle toho, zda prostorové jednotky sdílí část společné hranice. Pokud množinu hraničních bodů *i-té* (resp. *j-té*) prostorové jednotky označíme pomocí proměnné M_i (resp. M_j), pak vyjádření hodnot vah v případě *metody sousedství královny*, lze vyjádřit následovně

$$w_{ij} = \begin{cases} 1 & M_i \cap M_j \neq \emptyset \\ 0 & M_i \cap M_j = \emptyset \end{cases}.$$

   Uvedená podmínka však není moc striktní, protože postačí pouze jediný hraniční bod (obrázek 1.7) a zkoumané prostorové jednotky jsou označeny za sousedy. Z uvedeného důvodu bude požadováno, aby byla sdílená část hranice o určité délce. Pokud délku hranice mezi prostorovými jednotkami i a j označíme jako l_{ij}, pak váhy sousedství vyjádříme pomocí následujícího vzorce

$$w_{ij} = \begin{cases} 1 & l_{ij} > 0 \\ 0 & l_{ij} = 0 \end{cases}.$$

Pokud pro určení sousedství byla využita metoda založena na *vzdálenosti*, pak lze prvky binomické matice vah určit pomocí dvou metod, které si blíže představíme.

1. *Váhy k-nejbližších sousedů*: Pro určení vah je nutné změřit vzdálenosti od *i-té* jednotky k zbylým prostorovým jednotkám ($r_{ij} > 0$) a určit hodnotu k ($k = 1, ..., n$-1), která vyjadřuje počet nejbližších sousedů. Veškeré naměřené vzdálenosti se seřadí od nejmenší po největší hodnotu ($r_{ij(1)} \leq r_{ij(2)} \leq ... \leq r_{ij(n-1)}$, $i, j = 1, ..., n$, $j \neq i$) a prvních k hodnot vytváří množinu *k-nejbližších sousedů* $\left(N_k^i = (r_{ij(1)}, r_{ij(2)}, ..., r_{ij(k)}\right)$. Číselné vyjádření prostorových vah vyjádříme pomocí následujícího vzorce

$$w_{ij} = \begin{cases} 1 & r_{ij} \in N_k^i \\ 0 & jinak \end{cases}.$$

Výše uvedená metoda se označuje, jako *standardní forma vah k-nejbližších sousedů*. K uvedené metodě existuje i alternativa, která se nazývá *symetrická forma vah k-nejbližších sousedů*. Tato metoda označí *i-tou* a *j-tou* prostorovou jednotku za sousedy právě tehdy, když *j-tá* jednotka náleží do N_k^i a zároveň *i-tá* jednotka náleží do N_k^j. Formálně lze uvedenou metodu zapsat následovně

$$w_{ij} = \begin{cases} 1 & j \in N_k^i \wedge i \in N_k^j \\ 0 & jinak \end{cases}.$$

2. *Váhy prahové vzdálenosti:* Při definování sousedství pomocí prahové vzdálenosti je nezbytné určit prahovou vzdálenost d, vyjadřující hranici prostorového vlivu na *i-tou* (resp. výchozí) prostorovou jednotku. V tomto případě bude jednotka j označená za souseda *i-té* prostorové jednotky, pokud vzdálenost mezi uvedenými jednotkami (r_{ij}) je menší nebo rovna zvolené prahové vzdálenosti. Uvedenou charakterizaci lze vyjádřit pomocí následujícího vzorce

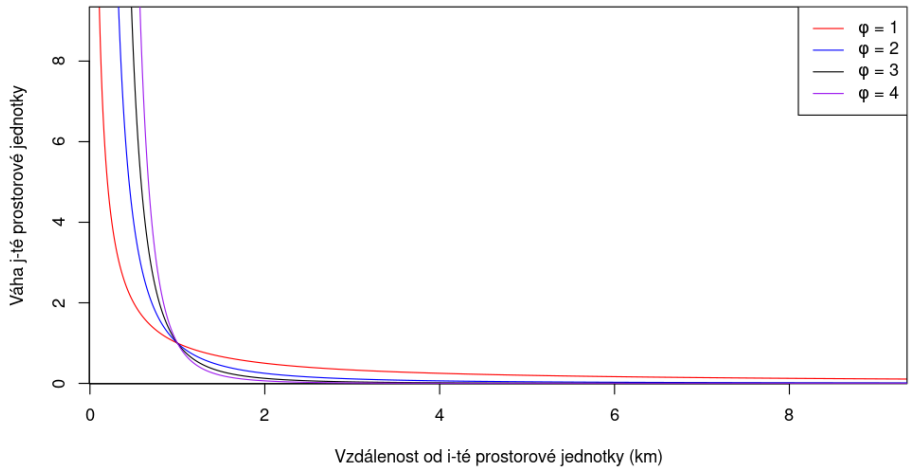
$$w_{ij} = \begin{cases} 1 & 0 < r_{ij} \leq d \\ 0 & r_{ij} > d \end{cases}.$$

Z uvedených metod je zřejmé, že výsledné hodnoty *binomické matice vah* neumožňují zohlednit princip prvního zákonu geografie. Proto si uvedeme alternativu váhové matice, která se nazývá *spojitá matice vah*, u které je možné hodnotu vah vyjádřit pomocí kladného reálného čísla ($w_{ij} \in \mathbb{R}^+$). Spojitou matici vah je možné dělit podle jednotlivých typu sousedství, stejně jako v případě binomické matice vah. Pokud je pro určení sousedství využitá metoda založena na principu *vzdálenosti*, pak *váhy spojité matice* lze vypočítat pomocí třech základních funkcí.

1. *Mocninné prostorové váhy:* Aby bylo možno vypočítat hodnotu vah, je nutno znát vzdálenost mezi jednotlivými prostorovými jednotkami (r_{ij}), které následně dosadíme do námi zvolené rovnice, která nám umožní vypočítat váhy prostorových jednotek. V tomto případě se jedná o mocninnou funkci ve tvaru

$$w_{ij} = r_{ij}^{-\phi},$$

kde exponent ϕ ($\phi \in \mathbb{N}^+$) určuje rychlost poklesu vah s přibývající vzdálenosti od výchozí jednotky, při řešení konkrétního problému se nejčastěji volí $\phi = 1$, nebo $\phi = 2$ (obrázek 1.10).



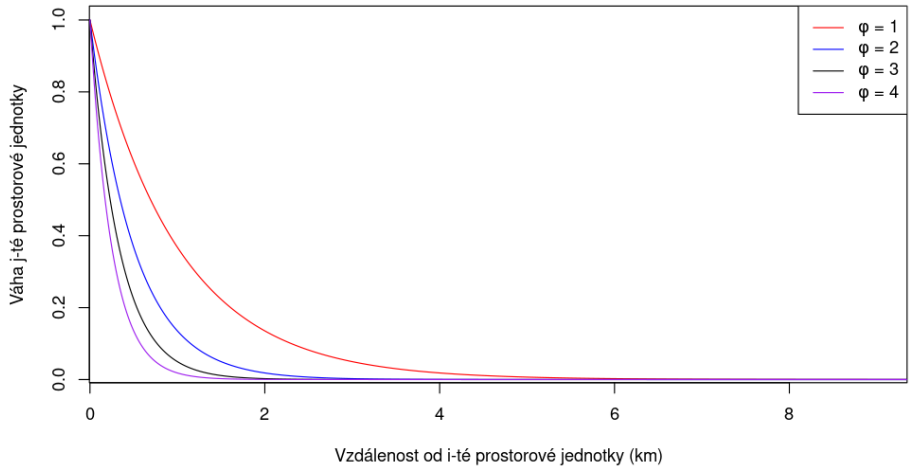
Obrázek 1.10: *Mocninné prostorové váhy* s rozdílnými ϕ.

2. *Exponenciální prostorová váha:* Pro určení hodnoty vah je v tomto případě

využívaná exponenciální funkce

$$w_{ij} = exp(\text{-}\phi r_{ij}).$$

Protože vzdálenost mezi prostorovými jednotkami ($r_{ij} \geq 0$) nemůže nabývat záporných hodnot, realizuje se exponenciální funkce pouze v prvním kvadrantu souřadnicového systému. Hodnota ϕ opět určuje rychlost poklesu vah s přibývající vzdálenosti od výchozí prostorové jednotky (obrázek 1.11).

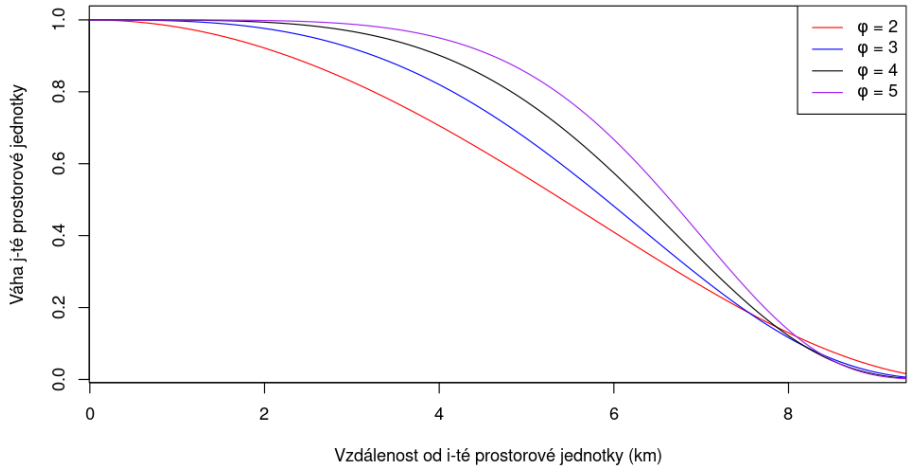


Obrázek 1.11: *Exponenciální prostorové váhy* s rozdílnými α.

3. *Dvojitě umocněné prostorové váhy:* V tomto případě pro výpočet vah je potřeba znát parametr prahové vzdálenosti d, dále vzdálenost mezi jednotlivými prostorovými jednotkami r_{ij} a vhodně zvolit hodnotu ϕ. Veškeré uvedené parametry následně dosadíme do vzorce, který nám určí hodnoty vah

$$w_{ij} = \begin{cases} [1 \text{ - } (r_{ij}/d)^{\phi}]^{\phi} & 0 \leq r_{ij} < d \\ 0 & r_{ij} \geq d \end{cases}.$$

Kvůli mírnějšímu poklesu vah s přibývající vzdálenosti se tato metoda často využívá pro definování prostorových vah (obrázek 1.12).



Obrázek 1.12: *Dvojitě umocněná prostorové váhy* s rozdílnými α.

Pokud byla pro určení prostorově blízkých jednotek využita metoda *sousedství*, pak se pro výpočet vah využívá dvou principů.

1. *Sdílené hraniční váhy:* U této metody je nutné znát proměnou l_i, která charakterizuje celkovou délku hranice *i-té* prostorové jednotky. Proměnnou l_i lze vyjádřit jako součet délky geografických hranic přímo sousedících jednotek s *i-tou* prostorovou jednotku

$$l_i = \sum_{j=1}^{n} l_{ij}.$$

   Prostorové váhy se následně vypočtou, jako podíl délky společné hranice *j-té* jednotky s *i-tou* jednotkou vůči celkové délce hranice *i-té* jednotky

$$w_{ij} = \frac{l_{ij}}{l_i}.$$

2. *Metoda kombinace vzdálenosti a hraniční váhy:* Metoda pro výpočet vah kombinuje dvě již zmíněné metody, konkrétně se jedná o metodu *mocninné prostorové váhy* a *sdílené hraniční váhy*. Spojením uvedených metod je získaná prostorová váha zahrnující jak délku společné hranice, tak i vzdálenost od zkoumané prostorové jednotky. Formální zápis pro výpočet hodnot vah vyjádříme následovně

$$w_{ij} = \frac{l_{ij} r_{ij}^{-\alpha}}{\sum_{j=1, j \neq i}^{n} l_{ij} {r_{ij}}^{-\alpha}}.$$

Proto, aby byla odstraněna závislost na cizích faktorech, např. nerovnoměrný počet sousedů, což je charakteristické v případě prostorových jednotek umístěných u státních hranic, je nutné hodnoty vah standardizovat. Pro standardizaci se využívají dvě základní metody, a to *řádková* a *skalární*. V případě *řádkové normalizace* se požaduje, aby součet vah v jednotlivých řádcích byl roven jedné

$$\sum_{j=1}^{n} w_{ij} = 1, \ i = 1, .., n.$$

Pokud pro výpočet vah využijeme metodu *mocninných prostorových vah*, pak se *řádková standardizaci* provede takto

$$w_{ij} = \frac{r_{ij}^{-\phi}}{\sum_{j=1, j \neq i}^{n} r_{ij}^{-\phi}}.$$

V případě metody *exponenciální prostorových vah* se *řádková standardizaci* vypočte následovně

$$w_{ij} = \frac{exp(\text{-}\phi r_{ij})}{\sum_{j=1, j \neq i}^{n} exp(\text{-}\phi r_{ij})}.$$

Uvedený princip lze aplikovat na veškeré metody, které jsou uvedeny při výpočtu prostorových vah. Je však nutno podotknout, že veškeré řádky jsou standardizovány zvlášť, a proto může dojít k porušení symetrie *matice vah*. Například si představme, že na zkoumaném území existují tři prostorové jednotky *i, j* a *p*, pro které platí, že *j-tá* jednotka sousedí s *i-tou* a *p-tou* jednotkou, ale prostorové jednotky *i* a *p* spolu nesousedí. Pomocí *binomické matice vah* lze uvedené sousedství

vyjádřit následovně

$$\mathbf{W} = \begin{pmatrix} 0 & w_{ij} & w_{ip} \\ w_{ji} & 0 & w_{jp} \\ w_{pi} & w_{pj} & 0 \end{pmatrix} = \begin{pmatrix} 0\ 1\ 0 \\ 1\ 0\ 1 \\ 0\ 1\ 0 \end{pmatrix}.$$

Pak výsledný tvar řádkové standardizace matice $\mathbf{W}$ je následující

$$\mathbf{W_s} = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}.$$

Z uvedené řádkové standardizované matice je zřejmé, že prostorová závislost mezi *i-tou* a *j-tou* jednotkou je větší než v opačném pořadí (tj. porušení symetrie). Uvedený problém lze vyřešit tím, že místo řádkové standardizace bude standardizace provedena pomocí *skaláru*

$$\gamma = \frac{1}{max(w_{ij})},$$

kde parametr $max(w_{ij})$ znázorňuje maximální hodnotu vyskytující se v matici vah. Skalární standardizace spočívá v tom, že veškeré hodnoty matice vah jsou vynásobeny stejnou hodnotou (resp. skalárem). Popsanou standardizaci lze vyjádřit ve tvaru $\gamma * \mathbf{W}$ a při aplikaci skalární standardizace na uvedený příklad bude výsledek následující

$$\gamma * \mathbf{W} = \begin{pmatrix} 0\ 1\ 0 \\ 1\ 0\ 1 \\ 0\ 1\ 0 \end{pmatrix}.$$

Z uvedené standardizované matice vidíme, že nedošlo k porušení symetrie, ale zároveň hodnoty matice vah zůstaly nepozměněny, protože maximální hodnota matice $\mathbf{W}$ je rovná jedné. Pro vyřešení uvedeného problému, lze vyzkoušet i alternativní přístup, který maximální hodnoty matice $\mathbf{W}$ nahradí maximální hodnotou z množiny *vlastních čísel* (λ) matice $\mathbf{W}$. Upravenou hodnotu skaláru vyjádříme, jako

$$\gamma = \frac{1}{max\left(\lambda_1, ..., \lambda_n\right)}, \text{ kde } n \in \mathbb{N}^+.$$

Ovšem i v tomto případě se maximální hodnota vlastních čísel rovná jedné ($max\,(\lambda_1, ..., \lambda_n) = 1$). Z uvedené charakteristiky vyplývá, že binomickou matici vah lze standardizovat pouze *řádkovou* metodou za cenu porušení symetrie.

Při vypracování této kapitoly byly využit tento zdroj: [17]

1.2. Prostorová stacionarita a nestacionarita

Prostorovou stacionaritu lze chápat jako podmínku, která vyjadřuje neměnnost pravděpodobnostní funkce zkoumaného statistického znaku v prostoru. Uvedený předpoklad je velmi silný a lze ho uvažovat pouze v případě zkoumání určitých fyzikálních procesů, např. gravitační síla je stejná, jak v Praze, tak i v Tokiu. Při zkoumat sociálních procesů však předpoklad prostorové stacionarity je nutné ověřit, protože lidské chování je nahodilé a nelze jej vyjádřit pomocí jednoho konkrétního vzorce.

Pro lepší pochopení prostorové nestacionarity si uvedeme konkrétní příklad. Uvažujme, že na zkoumaném geografickém území je analyzován vztah mezi *cenou bytu* (Y) vzhledem ke *vzdálenost od centra města* (x). Pokud je pro analýzu uvedeného vztahu předpokládaná prostorová stacionarita, tzn. že ve všech částech zkoumaného prostorou bude *vzdálenost od centra města* zvyšovat cenu bytu, potom lze uvedenou závislost popsat pomocí *globálního modelu*, který představuje lineární regresní funkci

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j,\ \ j = 1, ..., n,\ \ n \in \mathbb{N},$$

kde β_0, β_1 jsou regresní koeficienty a ε je náhodná chyba. Obecně model přepíšeme do maticového zápisu

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Vzhledem k tomu, že regresní model může obsahovat více vysvětlujících proměnných, budeme kvůli obecnosti uvažovat f vysvětlujících proměnných. Uvedené parametry regresního modelu lze rozepsat následovně

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1f} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nf} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_f \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

kde hodnota f vyjadřuje počet regresních koeficientů ($\boldsymbol{\beta}$). Pro výše uvedený model se hodnota proměnné f rovná jedné, a to z důvodu, že konstantní parametr β_0 se do proměnné f nezapočítává. Aby byly odhady neznámých parametrů vektoru $\boldsymbol{\beta}$ nejlepšími nestrannými odhady, musí být splněny následující předpoklady

1. Pro všechna pozorování platí, že střední hodnota náhodných chyb je nulová $\left(E\left(\boldsymbol{\varepsilon}\right) = \mathbf{1_n}0, \text{ kde } \boldsymbol{\varepsilon} = \left(\varepsilon_1, ..., \varepsilon_n\right)', \ \mathbf{1_n} = \left(1_1, ..., 1_n\right)'\right)$.

2. Náhodné chyby jsou navzájem nekorelované ($\forall \ i \neq j : cov\left(\varepsilon_i, \varepsilon_j\right) = 0$) a charakterizují se homoskedasticitou, tj. konstantní a konečný rozptyl ($var\left(\boldsymbol{\varepsilon}\right) = \sigma^2 \boldsymbol{I}$). Uvedenou strukturu variační matici náhodných chyb (Σ) lze zapsat následovně

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}.$$

3. Hodnost matice $\boldsymbol{X}$ odpovídá počtu lineárně nezávislých regresorů ($h\left(\boldsymbol{X}\right) = f < n$, kde hodnota f vyjadřuje počet koeficientů vektorů $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_f\right)'$). V případě, že zkoumaný model obsahuje i parametr β_0, pak se hodnost matice $\boldsymbol{X}$ zvětší o hodnotu 1 ($h\left(\boldsymbol{X}\right) = f + 1 < n$, protože $\boldsymbol{\beta} = \left(\beta_0, \beta_1, ..., \beta_f\right)$).

V případě, že jsou splněny uvedené podmínky, je možné pro odhad regresních koeficientů (β) využít *metodu nejmenších čtverců* (MNČ), která umožní vyjádřit vztah mezi zkoumanými proměnnými. Pro výše uvedený příklad lze globální model použít pouze v případě, že vzdálenost od centra města je relativní faktor pro veškeré zkoumané prostorové jednotky. V případě, že v jiných prostorových

jednotkách bude cena bytu záviset na hustotě zalidnění a nikoliv na vzdálenosti od centra, bude uvedená prostorová jednotka označena za nestacionární.

Proto, abychom mohli zkoumat prostorovou nestacionaritu, uvedeme si metodu prostorově vážené regrese (geographically weighted regression), kterou si blíže představíme v následující kapitole.

1.2.1. Prostorově vážená regrese

Základní myšlenkou prostorové vážené regrese je vytvořit regresní model pro každou prostorovou jednotku zvlášť. K tomuto účelu využijeme globální model, u kterého však pro odhad regresních koeficientů uvažujme pouze m ($m < n$) prostorově nejbližších pozorování od výchozí prostorové jednotky. Lokální model *i-té* prostorové jednotky budeme značit

$$Y_j^i = \beta_0^i + \beta_1^i x_j^i + \varepsilon_j^i,\ j = 1, ..., m,$$

a pro jednoduchost uvažujme model s jednu vysvětlující proměnnou.

Vzhledem k tomu, že pomocí prostorově vážená regrese, je zkoumána závislost statistického znaků u které je zároveň zohledněn i prostorový faktor, je nutné si uvést metodu, která nám umožní při odhadu regresních parametrů zohlednit hodnoty matice vah. K tomuto účelu lze využít *metodu vážených nejmenších čtverců*, která minimalizuje součet druhých mocnin reziduí, které jsou vynásobeny váhou prostorově blízkých jednotek $\left(w_j^i\right)$

$$S_w\left(\beta_0^i, \beta_1^i\right) = \sum_{j=1}^{m} w_j^i \left(y_j^i \text{ - } \beta_0^i \text{ - } \beta_1^i x_j^i\right)^2.$$

Aby bylo možné uvedený součet minimalizovat, je nutno vypočítat první parciální derivace funkce S_w

$$\frac{\partial S_w\left(\beta_0^i, \beta_1^i\right)}{\partial \beta_0^i} = \text{ - } 2 \sum_{j=1}^{m} w_j^i \left(y_j^i \text{ - } \beta_0^i \text{ - } \beta_1^i x_j^i\right),$$

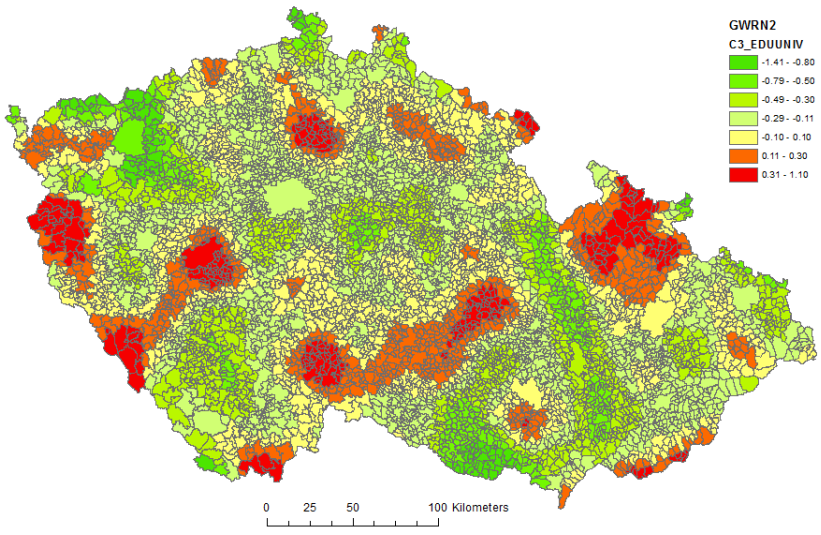
$$\frac{\partial S_w\left(\beta_0^i, \beta_1^i\right)}{\partial \beta_1^i} = \text{ - } 2 \sum_{j=1}^{n} w_j^i \left(y_j^i \text{ - } \beta_0^i \text{ - } \beta_1^i x_j^i\right) x_j^i.$$

Následně první parciální derivace položíme rovny nule a vyjádříme odhad regresních koeficientů β_0^i, β_1^i

$$\hat{\beta_0}^i = \frac{\sum_{j=1}^{m} w_j^i y_j^i - \hat{\beta_1}^i \sum_{j=1}^{n} w_j^i x_j^i}{\sum_{j=1}^{m} w_j^i},$$

$$\hat{\beta_1}^i = \frac{\sum_{j=1}^{n} w_j^i \sum_{j=1}^{m} w_j^i y_j^i x_j^i - \sum_{j=1}^{n} w_j^i x_j^i \sum_{j=1}^{m} w_j^i y_j^i}{\sum_{j=1}^{n} w_j^i \sum_{j=1}^{m} w_j^i x_j^{i^2} - \left(\sum_{j=1}^{m} w_j^i x_j^i\right)^2}.$$

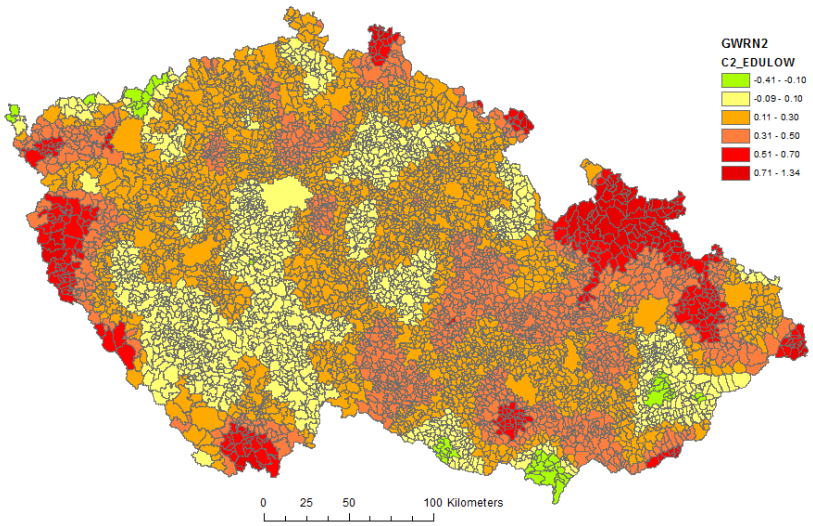
Uvedeným postupem bychom odhadli regresní koeficienty pro veškeré prostorové jednotky na zkoumaném území. V případě, že se na zkoumaném území zřetelně vyskytují oblasti s negativním a zároveň s pozitivním vlivem na vysvětlovanou proměnou, pak se na zkoumaném území vyskytuje prostorová nestacionarita (obrázek 1.13).



Obrázek 1.13: Lokální regresní koeficienty vlivu podílu osob s vysokoškolským vzděláním na míru nezaměstnanosti v obcích ČR 3/2011 [11].

V opačném případě je vliv faktoru označen za prostorově stacionární, po-

kud se na zkoumaném území vyskytuje pouze negativní resp. pozitivní regresní koeficienty. Ovšem může nastat situace, že na zkoumaném území se vyskytují negativní (resp. pozitivní) regresní koeficienty, a i přesto se zkoumaný faktor označí za stacionární (obrázek 1.14). Důvod je ten, že se přímo nejedná o prostorová nestacionaritu, ale o rozdílnou sílu vztahu.



Obrázek 1.14: Lokální regresní koeficienty vliv podílu osob s nízkým vzděláním na míru nezaměstnanosti v obcích ČR, 3/2011 [11].

Proto si uvedeme testování stacionarity pomocí metody Monte Carlo, která bude testovat nulovou hypotézu ve tvaru $H_0 : \beta_p^1 = \beta_p^2 = ... = \beta_p^n$ a alternativu vyjádříme jako $H_a : \exists\, i, j = 1, .., n : \beta_p^i \neq \beta_p^j, i \neq j$. I v tomto případě budeme postupovat podle určitých bodů

1. Odhadneme jednotlivé regresní koeficienty pomocí prostorové vážené regrese a následně pro jednotlivé koeficienty vypočteme prostorový rozptyl

$$var\left(\hat{\beta}_p\right) = \frac{1}{n\text{ - }1}\sum_{i=1}^{n}\left(\hat{\beta}_p{}^{i}\text{ - }E\left(\hat{\beta}_p\right)\right)^2,$$

kde $E\left(\hat{\beta}_p\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_p{}^{i}$ značí střední hodnotu odhadnutého koeficientu v prostoru.

2. Náhodně rozdělíme hodnoty vysvětlujících proměnných po jednotlivých prostorových jednotkách.

3. Vypočteme regresní koeficienty a rozptyly pro náhodně vytvořený soubor v kroku 2.

4. Body 2 a 3 opakujeme podle počtu zvolených permutací.

5. Vypočteme p-hodnotu jako podíl počtu rozptylů které jsou větší než rozptyl vypočtený v prvním kroku.

Jestliže je p-hodnota menší jak zvolena hladina významnosti, je nulová hypotéza zamítnuta a přikláníme se k alternativní hypotéze. Samotné výsledky prostorové vážené regrese neslouží k popisu závislosti mezi jednotlivými prostorovými jednotkami, ale k námětu pro další hlubší analýzu. Důvodem je, že prostorově vážená regrese umožňuje spojovat prostorové jednotky se stejnou charakterizací, která následně umožní odhalit jiné prostorové závislosti.

Pokud bude závislá proměnná záviset na více než jednom regresoru, je vhodné převést uvedený model do maticového zápisu, z důvodu jednoduššího výpočtu. Proto uvedeme lokální model s f vysvětlujícími proměnnými $(X_1, ..., X_f)$, který vyjádříme pomocí maticového zápisu

$$\mathbf{Y^i} = \mathbf{X^i}\boldsymbol{\beta}^i + \boldsymbol{\varepsilon}^i,$$

kde jednotlivé proměnné jsou následující

$$\mathbf{Y}^i = \begin{pmatrix} y_1^i \\ \vdots \\ y_m^i \end{pmatrix}, \ \mathbf{X}^i = \begin{pmatrix} 1 & x_{11}^i & \dots & x_{1f}^i \\ \vdots & \vdots & & \vdots \\ 1 & x_{m1}^i & \dots & x_{mf}^i \end{pmatrix}, \ \boldsymbol{\beta}^i = \begin{pmatrix} \beta_0^i \\ \beta_1^i \\ \vdots \\ \beta_f^i \end{pmatrix}, \ \boldsymbol{\varepsilon}^i = \begin{pmatrix} \varepsilon_1^i \\ \vdots \\ \varepsilon_m^i \end{pmatrix}.$$

Váhy pro *i-tou* prostorovou jednotku můžeme zapsat do matice

$$\mathbf{W}^i = \begin{pmatrix} w_1^i & 0 & \cdots & 0 \\ 0 & w_2^i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_m^i \end{pmatrix}.$$

Při hledaní odhadu regresních koeficientů je využita *metoda zobecněných nejmenších čtverců* (MZNČ), která pro veškeré realizace $\left(\mathbf{y^i}\right)$ náhodného vektorů $\mathbf{Y^i}$, $var(\mathbf{Y^i}) = \Sigma^i$, minimalizuje kvadratickou formu

$$min\left(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i\right)'\left(\boldsymbol{\Sigma}^i\right)^{-1}\left(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i\right),$$

Varianční matice náhodného vektorů $\mathbf{Y^i}$ je do kvadratické formy přidaná z důvodu, že metoda *zobecněných nejmenších čtverců minimalizuje* tzv. *Mahalanobisovou vzdálenost* vektorů $(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i)$. *Mahalanobisova vzdálenost* se od *Eukleidovské vzdálenosti* liší pouze tím, že hodnota vektorů $(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i)$ je vynásobena inverzní variační matici náhodného vektorů $\mathbf{Y^i}$, a proto ji lze definovat, jako

$$D\left(\mathbf{y}^i\right) = \sqrt{\left(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i\right)'\left(\boldsymbol{\Sigma}^i\right)^{-1}\left(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i\right)}.$$

Jestliže jsou prvky varianční matice $\boldsymbol{\Sigma}^i$ mimo diagonálu nulové, tj. $cov\left(\varepsilon_i, \varepsilon_j\right) = 0,$ kde $i \neq j$, je pro odhad regresních koeficientů využit speciální případ MZNČ, který se nazývá *metoda vážených nejmenších čtverců* (MVNČ). Uvedené metody se liší v tom, že u metody MVNČ je místo variační matice $\boldsymbol{\Sigma^i}$ využitá matice $\boldsymbol{A}^i$, která na hlavní diagonále má převracené hodnoty rozptylů a mimo diagonálu jsou nulové hodnoty $\left(\boldsymbol{A}^i = \left(\boldsymbol{\Sigma}^i\right)^{-1}\right)$. Protože mezi MZNČ a MNČ existuje vzájemná ekvivalence, lze pro odhad regresních koeficientů v případě MZNČ využít stejného postupu, jako u MNČ, která zachovává vlastnosti odhadu tj. *nejlepší nestíraný lineárním odhad* (viz. Gaussova-Markovova věta[3]).

Proto budeme postupovat následovně, prvně si vyjádříme kvadratickou formu, kterou budeme minimalizovat za účelem odhadnutí regresních koeficientů

$$H\left(\mathbf{y}^i, \boldsymbol{\beta}^i\right) = \left(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i\right)'\left(\boldsymbol{\Sigma}^i\right)^{-1}\left(\mathbf{y}^i \text{ - } \mathbf{X}^i\boldsymbol{\beta}^i\right).$$

Následně uvedenou kvadratickou formu zderivujeme a dostaneme rovnici, která je ve tvaru vyjádřená pomocí následující rovnice

[3]Lineární odhady neznámých parametrů v lineárních regresních modelech optimální ve smyslu principu nejmenších čtverců jsou nejlepší nestranné lineární odhady. [7]

$$\frac{\partial H\left(\mathbf{y}^i, \boldsymbol{\beta}^i\right)}{\partial \boldsymbol{\beta}^i} = 2\left(\mathbf{X}^i\right)'\left(\boldsymbol{\Sigma}^i\right)^{-1}\mathbf{X}^i\boldsymbol{\beta}^i - 2\left(\mathbf{X}^i\right)'\left(\boldsymbol{\Sigma}^i\right)^{-1}\mathbf{y}^i.$$

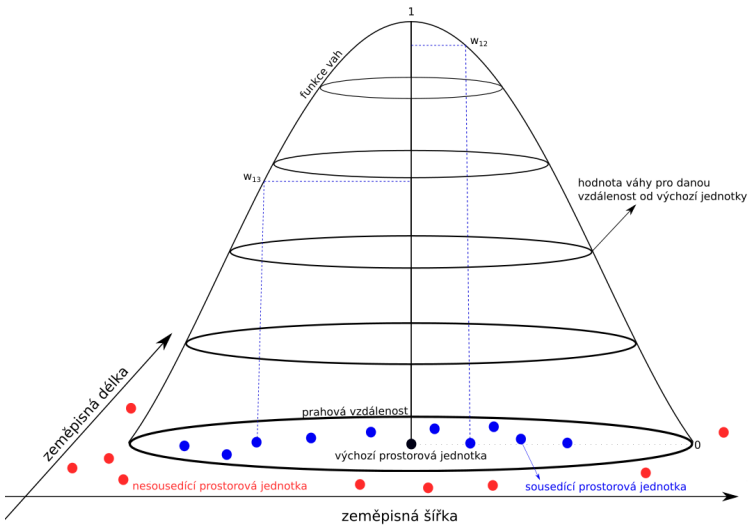
Výslednou derivaci položíme rovno nule a vyjádříme vektor $\boldsymbol{\beta}^i$, který budeme značit jako odhad

$$\hat{\boldsymbol{\beta}}^i = \left(\mathbf{X}^{iT}\left(\boldsymbol{\Sigma}^i\right)^{-1}\mathbf{X}^i\right)^{-1}\mathbf{X}^{iT}\left(\boldsymbol{\Sigma}^i\right)^{-1}\mathbf{y}^i.$$

V případě prostorově vážené regrese využijeme stejného principu jako u MVNČ, ovšem s tím rozdílem, že místo $\boldsymbol{A}^i = \left(\boldsymbol{\Sigma}^i\right)^{-1}$ budeme uvažovat $\boldsymbol{A}^i = \mathbf{W}^i$. Pomocí uvedené úpravy budeme schopni zohlednit prostorový faktor, při odhadu regresních parametrů, které vypočteme následovně

$$\hat{\boldsymbol{\beta}}^i = \left(\mathbf{X}^{iT}\mathbf{W}^i\mathbf{X}^i\right)^{-1}\mathbf{X}^{iT}\mathbf{W}^i\mathbf{Y}^i.$$

Tímto způsobem odhadneme regresní parametry pro každou prostorovou jednotku zvlášť. Abychom mohli vypočítat odhady regresních parametrů pro *i-tou* jednotku, musíme určit matici prostorových vah $\mathbf{W}^i$ resp. *prostorové jádro*, které určuje prostorové váhy v tři dimenzionálním prostoru (obrázek 1.15).



Obrázek 1.15: Interpretace prostorového jádra.

Protože jsou prostorové váhy konstantní ve všech směrech světových stran, využívají se pro definování sousedství metody založené na *vzdálenosti*, ze kterých se nejvíce využívá metoda *prahové vzdálenosti*. Uvedená metoda se často využívá, protože za zvolenou prahovou vzdáleností jsou veškerým pozorováním přiřazeny nulové váhy (obrázek 1.15). Ovšem pro samotný výpočet váhy se nejvíce využívá funkce, které se nazývá Gaussovo jádro a matematicky se vyjádří následovně

$$w_{ij} = exp\left(-\frac{1}{2}\left(\frac{r_{ij}}{d}\right)^2\right).$$

V uvedeném vzorci parametr r_{ij} vyjadřuje vzdálenost od *i-te* k *j-te* prostorové jednotky a parametr d určuje prahovou vzdálenost. V případě, že z uvedeného vzorce odebereme druhou mocninu, pak získáme vzorce pro *exponenciální* jádro. Pro výpočet vah lze využít i metodu, kterou jsem si uvedli v podkapitole *matice vah* a přesně se jedná o metodu *dvojitě umocněných prstových vah*.

Samotné hodnoty vah však nejsou v tomto případě tím nejdůležitějším faktorem při odhadu regresních koeficientů, ale volba prahové vzdálenosti. Pokud je zvolená prahová vzdálenost příliš velká, pak se lokální model velmi podobá globálnímu a dochází k zkreslení prostorového faktoru. V opačném případě, je-li prahová oblast příliš malá, tak nastane situace, že pro odhad koeficientů nebude dostatek pozorování, což bude mít za následek nepřesnost modelu. Nejvhodnější hodnotu prahové vzdálenosti lze zjistit pomocí porovnávání modelů s různými prahovými vzdálenostmi, které jsou voleny podle osobního uvážení. Aby se však předešlo situaci, že neustálým zvětšováním prahové vzdálenosti se model bude vylepšovat, je určená maximální prahová vzdálenost pomocí metody *zlatého řezu*.

Proto, aby bylo objektivně rozhodnuto, který z modelů (resp. prahová vzdálenost) je nejvhodnější, využívají se dvě základní metody, a to *adjustované Akaikeho informační kritérium* (AIC_c) a nebo *kritérium křížové validace* (KV). Pro výpočet AIC_c se využívá následující vzorec

$$AIC_c = m\left(2ln\left(\hat{\sigma}\right) + ln\left(2\pi\right) + \left(\frac{m + stopa\left(\mathbf{S}\right)}{m\ -\ 2\ -\ stopa\left(\mathbf{S}\right)}\right)\right),$$

kde proměnná m vyjadřuje počet pozorování a matice $\mathbf{S} = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}$ se nazývá *hat-matrix*. Odhad směrodatné odchylky provedeme pomocí vzorce

$$\hat{\sigma} = \sqrt{\frac{1}{m - q}\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)'},$$

kde proměnná q vyjadřuje počet odhadovaných regresních koeficientů. Za nejvhodnější model označíme ten, který má nejmenší hodnotu AIC_c a to z důvodu, že *adjustované Akaikeho informační kritérium* slouží k posouzení schopnosti modelu vysvětlit variabilitu v datech.

Při využití *křížové validace* se opět porovnává hodnota, která je vypočtena při určitém prahové vzdálenosti. V tomto případě se jedná o reziduální součet čtverců, který vypočteme následovně (d)

$$KV = \sum_{j=1}^{m}\left(y_j - \hat{y}_j\left(d\right)\right)^2, \; j \neq i.$$

Hodnotu *i-té* (resp. výchozí) prostorové jednotky při odhadu regresních koeficientů nevyužíváme, protože chceme přecházet situaci, že by prahová vzdálenost byla zvolena blízko nuly. Při porovnávání modelu považujeme za nejvhodnější ten, který má minimální hodnotu reziduálního součtu čtverců.

Uvedené metody se využijí u každé prostorové jednoty zvlášť, aby byla co nejlépe definována prahová vzdálenost, pomocí které se následně vyjádří prostorové jádro. Uvedený princip označujeme jako *metodu adaptivních prostorových jader*. Existuje i alternativní přístup, který se značí jako *metoda fixních prostorových jader*, která pro odhad lokálního modelu pokaždé využije stejnou prahovou vzdálenost (resp. prostorové jádro). Hodnotu prahové vzdálenosti fixního jádra lze zvolit jako průměrnou hodnotu prahových vzdáleností vypočtených pomocí *metody adaptivních prostorových jader*. V případě, že na zkoumaném území převládá rovnoměrné rozložení dat, pak se pro odhad lokálních modelů využívá fixní prostorové jádro, v opačném případě se využije metoda adaptivních prostorových jader.

Při vypracování uvedených kapitol byly využity tyto zdroje: [2], [5], [7], [9], [8], [10], [11], [13], [18].

1.2.2. Umělé proměnné

Umělé proměnné slouží k tomu, abychom do lineárního modelu mohli zařadit i kategoriální proměnné, které se dělí na dva typy proměnných *dichotomické* a *vícekategoriální*. *Dichotomická* proměnná nabývá pouze dvou kategorií (např. muž a žena), oproti tomu *vícekategoriální* proměnná nabývá, více než dvou kategorií (např. nejvyšší dokončené vzdělaní). Princip metody umělých proměnných popíšeme na konkrétním příkladě. Předpokládejme, že je zkoumaná závislost mezi *výší mzdy* (Y) vzhledem k *délce praxe* (x) a *pohlaví* (a) pomocí n pozorování. Abychom byli schopni zohlednit v modelu kategoriální proměnou, je nutno do regresního modelu přidat *umělou proměnnou* (a_i). Výsledný model je ve tvaru

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 a_i + \varepsilon_i, \;\; i = 1, .., n,$$

kde hodnota proměnné a_i, může nabývat pouze dvou hodnot

$$a_i = \begin{cases} 0 \text{ žena} \\ 1 \text{ muž} \end{cases}.$$

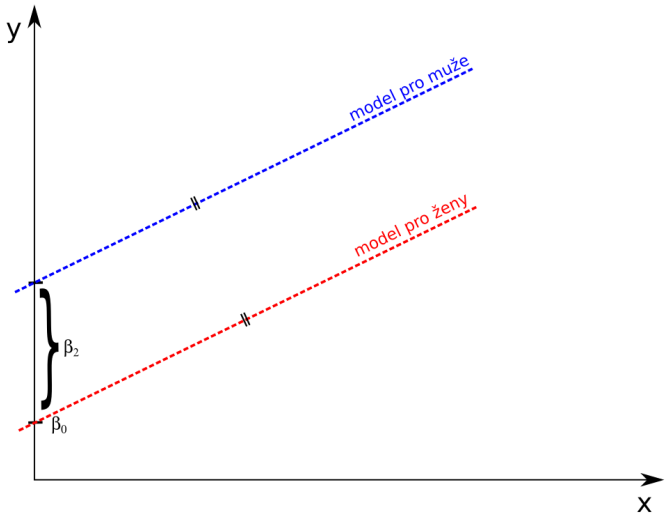
Pokud je *i-té* pozorování ženského pohlaví ($a_i = 0$), pak je výsledný model ve tvaru

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Následně pokud je *i-té* pozorování mužského pohlaví ($a_i = 1$) je model určující výši platu následující

$$Y_i = \beta_0 + \beta_2 + \beta_1 x_i + \varepsilon_i.$$

Ve výsledku tedy dostaneme dva modely, které se liší o konstantu určenou hodnotou koeficientu β_2 (obrázek 1.16).



Obrázek 1.16: Interpretace modelu s dichotomickou proměnnou.

Metoda umělých proměnných se využívá z důvodu, že při využití separovaných modelů

$$Y_h = \beta_0 + \beta_1 x_h + \varepsilon_h \text{ model pro ženy},$$

$$Y_j = \beta_2 + \beta_1 x_j + \varepsilon_j \text{ model pro muže},$$

je pro odhad regresních koeficientů využita pouze část pozorování a nikoliv celý statistický soubor ($h, j < n$ a zároveň $h + j = n$). Při interpretování výsledných modelů, pomocí obrázku 1.16 dospějeme k závěru, že existují dvě možnosti interpretace. Buď můžeme říci, že „muži mají v průměru o β_3 větší plat než ženy“ a nebo „ženy mají v průměru o β_3 menší plat než muži“. Proto se určuje tzv. *referenční skupina*, vůči které se bude provádět porovnávání s ostatními skupinami kategoriální proměnné. Mezi další důvody, proč volíme referenční skupinu je ten, že umožňuje zachovat nezávislost sloupců v *matici plánu* ($\mathbf{X}$) i v případě, kdy je v modelu zachován konstantní regresní koeficient.

Pro uvedený příklad je kategorie žen označena, jako referenční, a to z důvodu, že pokud do sdruženého modelu dosadíme za umělou proměnou nulovou hodnotu dostaneme separovaný model pro ženy. Nedostatkem uvedeného sdruženého modelu je předpoklad, že směrnice regresní funkce jsou pro obě zkoumané skupiny totožné, tzn. že koeficient vyjadřující průměrnou změnu platu (β_1), při jednot-

kové změně délky praxe je stejný pro obě skupiny. Proto se do modelu přidá tzv. interakce ($\beta_3 x_i a_i$)

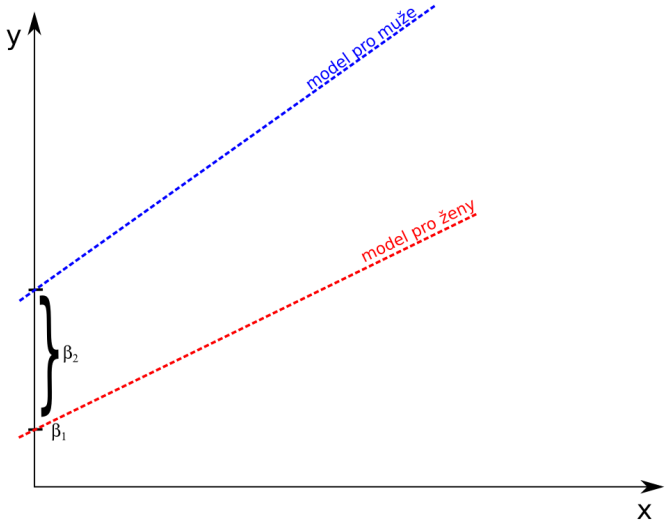
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 a_i + \beta_3 x_i a_i + \varepsilon_i,$$

která umožňuje zohlednit rozdílnou směrnici jednotlivých skupin (obrázek 1.17). Proto výsledný model pro muže (tj. $a_i = 1$) vyjádříme pomocí následujícího vzorce

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_i + \varepsilon_i,$$

a pro ženy (tj. $a_i = 0$) se výsledný model zapíše takto

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$



Obrázek 1.17: Interpretace modelu s interakcí.

Nyní již stačí odhadnout regresní koeficienty pomocí vhodné metody, např. metody nejmenších čtverců.

Problematiku vícekategoriální proměnné si opět uvedeme na příkladě. Uvažujme, že je zkoumaná závislost mezi *výši mzdy* vzhledem *délce praxe* a *nejvyššímu dokončenému vzdělání*, které dělíme do tři skupin tj. textitzákladní, středoškolské a vysokoškolské. Postup pro zohlednění vícekategoriální proměnné je

totožný s *dichotomickou proměnnou*, jen s tím rozdílem, že místo jedné interakce a jedné umělé proměnné budou přidány dvě

$$b_i = \begin{cases} 0 & \text{jiné vzdělání} \\ 1 & \text{základní vzdělání} \end{cases}, \quad s_i = \begin{cases} 0 & \text{jiné vzdělání} \\ 1 & \text{středoškolské vzdělání} \end{cases}.$$

Z uvedené charakteristiky vyplývá, že počet umělých proměnných je vždy roven počtu kategorií zmenšený o jedna, protože je nutné označit jednu z kategorií, jako referenční skupinu. Pro uvedený příklad je skupina vysokoškolsky vzdělaných osob zvolená jako referenční, a proto je sdružený model ve tvaru

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 b_i + \beta_3 s_i + \beta_4 b_i x_i + \beta_5 s_i x_i + \varepsilon_i.$$

Pokud umělé proměnné b_i a s_i položíme rovny nule, pak získáme separovaný model pro skupinu *vysokoškolsky vzdělaných* osob (resp. referenční skupiny)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Separovaný model *středoškolsky vzdělaných* osob získáme pokud, umělá proměnná b_i je rovná nule a s_i jedné

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)\, x_i + \varepsilon_i.$$

V opačném případě získáme separovaný model pro osoby se *základním vzděláním*

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)\, x_i + \varepsilon_i.$$

I pro tento případ využijeme k odhadu regresních koeficientů námi zvolenou metodu, např. metodu nejmenších čtverců.

Kapitola 2

Popis dat

Datová sada (ukázka tabulka 2.1), na kterou budou aplikovány výše uvedené metody prostorové analýzy, se skládá ze tří samostatných datových sad. Pro lepší přehled budou datové sady označeny jako **A, B** a **C**.

kód POÚ	počet úmrtí na 100 tisíc obyvatel (infarkt myokardu)	počet úmrtí na 100 tisíc obyvatel (rakovinu tlustého střeva a konečníku)	benzo[a]pyren $[ng.\ m^{-3}]$	prachové částice $< 10\mu m$ $[\mu g.\ m^{-3}]$	prachové částice $< 2,5\mu m$ $[\mu g.\ m^{-3}]$	klima
10000	33,52	29,39	1,05	23,64	17,50	teplé
21011	41,25	34,77	0,61	18,97	13,98	teplé
21013	36,44	44,21	0,71	20,18	14,97	teplé

Tabulka 2.1: Částečná ukázka datové sady.

Datová sada **A** obsahuje informace ohledně počtu úmrtí způsobené *infarktem myokardu* a *rakovinným onemocněním tlustého střeva a konečníku* na úrovní *obcí s pověřeným obecním úřadem*[1] (dále jen POÚ). Počty úmrtí jsou znázorněny jako absolutní četnosti úmrtí za jeden kalendářní rok v rozmezí let 2014 až 2018. Aby bylo možné správně přiřadit data pro jednotlivé POÚ, vyskytuje se v každé datové sadě tzv. kód POÚ. Pro lepší představu jsou do ukázky datové sady **A** (tabulka 2.2) přidány i názvy POÚ.

název POÚ	kód POÚ	2014	2015	2016	2017	2018
Hlavní město Praha	10000	453	435	423	425	399
Benešov	21011	21	14	19	16	19
Týnec nad Sázavou	21013	3	4	0	7	5

Tabulka 2.2: Ukázka datové sady četností úmrtí na *infarkt myokardu*.

[1]Na území České republiky se vyskytuje 393 POÚ, mezi které patří i čtyři vojenské újezdy (Březina, Boletice, Libava, Hradiště).

Při porovnání absolutních počtů úmrtí *(APU)* z tabulky 2.2 dospějeme k závěru, že největší počet úmrtí na *infarkt myokardu* se nachází v Hlavním městě Praha a nejmenší v Týnci nad Sázavou. Protože v jednotlivých POÚ nežije stejný počet obyvatel *(PO)*, nelze uvedený závěr brát jako platný. Proto musíme absolutní počty úmrtí standardizovat podle vzorce

$$RPU_{ij} = \frac{APU_{ij}}{PO_{ij}}\ 100\ 000,\ \ i = 1, .., 393,\ \ j = 2014, ..., 2018.$$

Výsledná hodnota se značí jako relativní počet úmrtí *(RPU)*, která se v geografické terminologii interpretuje následovně „počet úmrtí na 1 obyvatele“. Aby uvedená interpretace byla srozumitelnější, násobí se *RPU* zvolenou konstantou, která je v nejčastějších případech rovna hodnotě 100 000 a výsledná hodnota se interpretuje jako „počet úmrtí na 100 000 obyvatel“. Pokud pro srovnání POÚ využijeme relativního počtu úmrtí viz tabulka 2.3 dospějeme k naprosto odlišnému závěru než v případě srovnání pomocí absolutního počtu úmrtí.

název POÚ	absolutní počet úmrtí	relativní počet úmrtí na 100 000 obyvatel
Hlavní město Praha	399	30,67
Benešov	19	43,31
Týnec nad Sázavou	5	47,41

Tabulka 2.3: Rozdíl mezi absolutní a relativní hodnotou počtu úmrtí na *infarkt myokardu* za rok 2018.

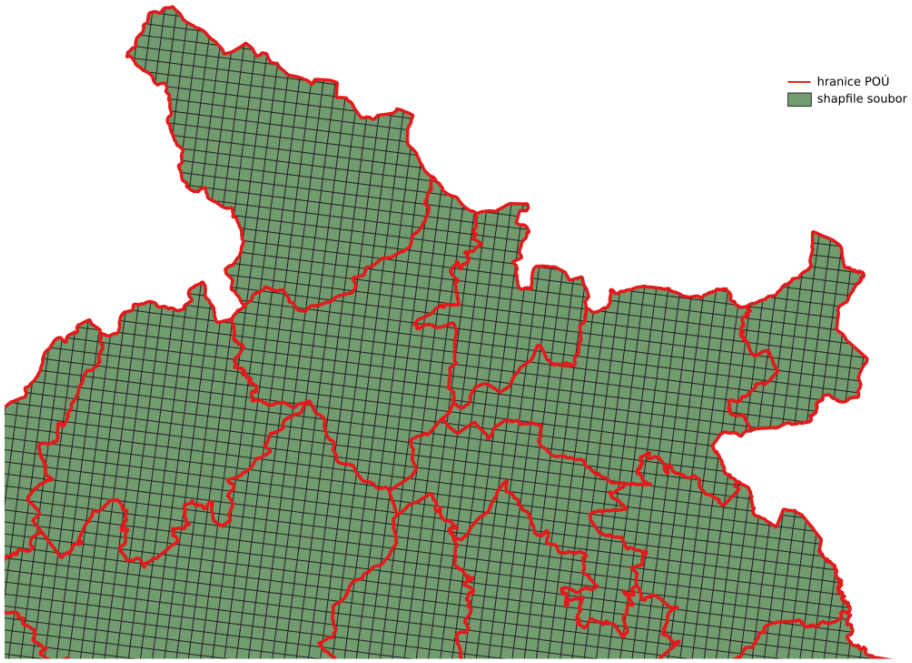
Protože budou prostorové metody aplikovány na datovou sadu, která se skládá ze tří samostatných datových sad **A, B** a **C**, je nutné sjednotit interpretaci zkoumaných statistických znaků. Pro jednotnou interpretaci bude využit aritmetický průměr *(AP)*, protože se jedná o nestranný a zároveň nejlepší odhad střední hodnoty. V případě datové sady **A**, bude aritmetický průměr vypočten pomocí vzorce

$$AP_i = \frac{\sum_{j=2014}^{2018} RPU_{ij}}{5},\ \ i = 1, .., 393.$$

Výsledná hodnota se interpretuje, jako „průměrný počet úmrtí za období 2014-2018 na 100 000 obyvatel“.

Datová sada **B** poskytuje informace ohledně kvality ovzduší na území České republiky. Uvedená datová sada je poskytovaná Českým hydrometeorologickým ústavem ve formě shapefile[2], který převedeme do standardní datové sady, např. excelové tabulky. V tomto případě je shapefile znázorněn jako čtvercová síť rozkládající se po celém území České republiky, přičemž jednotlivá čtvercová pole mají rozměr $1 \times 1\ km^2$ (obrázek 2.1). Pro každé čtvercové pole je zaznamenána průměrná hodnota měřených prvků a látek v ovzduší za pětileté období od roku 2014 do roku 2018. Na obrázku 2.1 vidíme, že jsou jednotlivá POÚ tvořena určitou podmnožinou čtverců určující kvalitu ovzduší v dílčích částech POÚ.

Aby bylo možné charakterizovat celé území POÚ, je zapotřebí vypočítat aritmetický průměr z dílčích čtvercových polí, k čemuž byl využit program *QGIS*[3]. Výslednou hodnotu interpretujeme, jako průměrnou hodnotu koncentrace zkoumaných látek v ovzduší na $1\ km^2$ za období 2014-2018.



Obrázek 2.1: Částečná vizualizace shapefile souboru pro jednotlivé POÚ.

[2]Datový formát využívaný pro ukládání vektorových prostorových dat pro geografické informační systémy.

[3]Volně dostupný geografický informační systém pro analýzu a interpretaci prostorových dat.

Český hydrometeorologický ústav měří koncentraci mnoha látek, ale v této práci budou blíže popsány pouze tři látky, které jsou svou vysokou koncentrací označovány za hlavní problém znečištění ovzduší na území České republiky. Mezi tyto látky patří *benzo[a]pyren* (BaP), který v roce 2018 překročil roční imisní limit 1 $ng.m^{-3}$ na 56 % stanic, na kterých se tato látka měří (22 z 39 stanic). Hlavní příčina vzniku benzo[a]pyrenu je spalování organických látek, proto lze za hlavní zdroj označit *výfukové plyny, kouř vzniklý při spalování uhlí a další spalování organických látek*. Podle Českého hydrometeorologického ústavu je za největší emisní zdroj benzo[a]pyrenu považováno lokální vytápění domácností, které za rok 2018 mělo 98,8% podíl na celkovém znečištění ovzduší. Důvodem měření koncentrace benzo[a]pyrenu v ovzduší je jeho negativní dopad na lidský organismus. Při pravidelném vdechování vyšší koncentrace se zvyšuje riziko rakovinného onemocnění, poškození DNA atd. [20].

Mezi další hlavní problémy znečištění ovzduší se řadí poletavé prachové částice menší než 10 μm (PM 10). V tomto případě je roční imisní limit stanoven na hodnotu 40 $\mu g.\ m^{-3}$, která v roce 2018 byla překročena pouze na 0,04 % území České republiky. Při zohlednění uvedené charakteristiky bychom dospěli k závěru, že se nejedná o tak závažný problém, jako v případě BaP. Důvod, proč je PM 10 označován za hlavní problém znečištění ovzduší, je ten, že v roce 2018 byla dosažena pouze na 51,6 % území České republiky ideální roční koncentrace méně než 20 $\mu g.\ m^{-3}$. Prachové částice mohou vznikat dvěma způsoby, a to buď *lidskou činností*, nebo *přírodními jevy*. Mezi prachové částice, které vznikají *lidskou činností* řadíme např. těžbu uhlí, štěrku, nebo spalování fosilních paliv. Pokud je zdroj prachových částic např. výbuch sopky, lesní požáry atd., je zdroj označen jako *přírodní jev*.

V České republice za období 2014-2018 lze za hlavní zdroj označit lokální vytápění domácností, které v roce 2018 mělo 73,9% podíl na celkovém znečištění. Hlavním důvodem měření koncentrace prachových částic je negativní vliv na lidský organismus, způsobující onemocnění horních cest dýchacích.

Poslední látku, kterou si blíže představíme, jsou poletavé prachové částice

menší než $2,5\ \mu m$ (PM 2,5). Protože jsou tyto částice zahrnuty i v hodnotě určující koncentraci PM 10, nebude překvapivé, že PM 2,5 také patří mezi hlavní problémy znečištění ovzduší na území České republiky. V případě PM 2,5 je situace horší, protože v roce 2018 byla ideální koncentrace méně jak 10 $\mu g.\ m^{-3}$ dosažená pouze na 7,0 % území České republiky. Vznik a hlavní zdroje prachových částic menší než $2,5\ \mu m$ jsou stejné, jako v případě prachových částic menších než $10\ \mu m$. Vzhledem k tomu, že jsou PM 2,5 menší než červené krvinky v lidském těle (tj. $7,5\ \mu m$), je jejich negativní dopad na lidský organismus větší než v případě PM 10. Zvýšená inhalaci prachových částic menší než $2,5\ \mu m$ může způsobit onemocnění kardiovaskulárního systému a dolních cest dýchacích.

Datová sada **C** obsahuje informace ohledně převažujících klimatických podmínkách v jednotlivých POÚ. Pro určení převažujícího lokálního klimatu na území jednotlivých obcí bylo využito mapového díla *klimatické oblasti (1901-2000)* [12]. V uvedeném mapovém díle se nachází celkově 13 klimatických oblastí, které jsou dělené podle teplotních a srážkových poměrů na daném území (např. velmi teplé, mírně teplé na srážky chudé a další). Veškeré klimatické oblasti jsou určeny podle klasifikace, která pro určení typu klimatu využívá 14 různých meteorologických charakteristik, mezi které patří *množství srážek*, *průměrná teplota vzduchu ve vybraných měsících, počet letních, ledových a mrazových dnů* atd. Veškeré uvedené charakteristiky lze nalézt v knize *Klimatické oblasti Československa* [16].

Při analýze bude využito pouze 5 ze 14 možných klimatických oblastí, mezi které patří *velmi chladná*, *chladná*, *mírně teplá*, *teplá* a *velmi teplá* klimatická oblast. Důvod uvedené redukce je dán tím, že zastoupení klimatických oblastí, které jsou charakterizovány jak z hlediska teplotního, tak srážkového úhrnu např. mírně teplé na srážky chudé, jsou na úrovní POÚ zastoupené v malém počtu tj. 53 z 393 POÚ.

2.1. charakteristiky zkoumaných jevů

Pro vytvoření ucelené představy budou pro jednotlivé statistického znaku uvedeny číselné charakteristiky a mapová díla. Mezi základní číselné charakteristiky

řadíme výběrový průměr ($\overline{x}$), medián ($\tilde{x}$), výběrovou směrodatnou odchylku (SD) a variační koeficient (CV), který je vyjádřen v procentech. Veškeré číselné charakteristiky jsou z důvodu lepší interpretace vypočteny na úrovni krajů. Dříve než přejdeme na interpretaci samotných charakteristik, je nutné si uvést problematiku výpočtu uvedených charakteristik v případě Hlavního města Prahy. Problém spočívá v tom, že na úrovní POÚ, ORP[4], krajů atd. je Hlavní město Praha definováno, jako jeden stejný prostorový celek. Proto nelze pro Hlavní město Prahu vypočítat výběrovou směrodatnou odchylku a variační koeficient a pro výběrový průměr a medián platí, že se jedná o přesnou realizaci zkoumané proměnné.

První charakteristika se zaměří na proměnou znázorňující počet úmrtí na *infarkt myokardu*. Největší hodnota výběrového průměru se nachází v *Jihočeském*, *Karlovarském*, *Libereckém* a *Jihomoravském* kraji. Naopak nejnižší průměrná hodnota se vyskytuje v *Hlavním městě Praze*, *Plzeňském*, *Královéhradeckém* a *Olomouckém* kraji (tabulka 2.4).

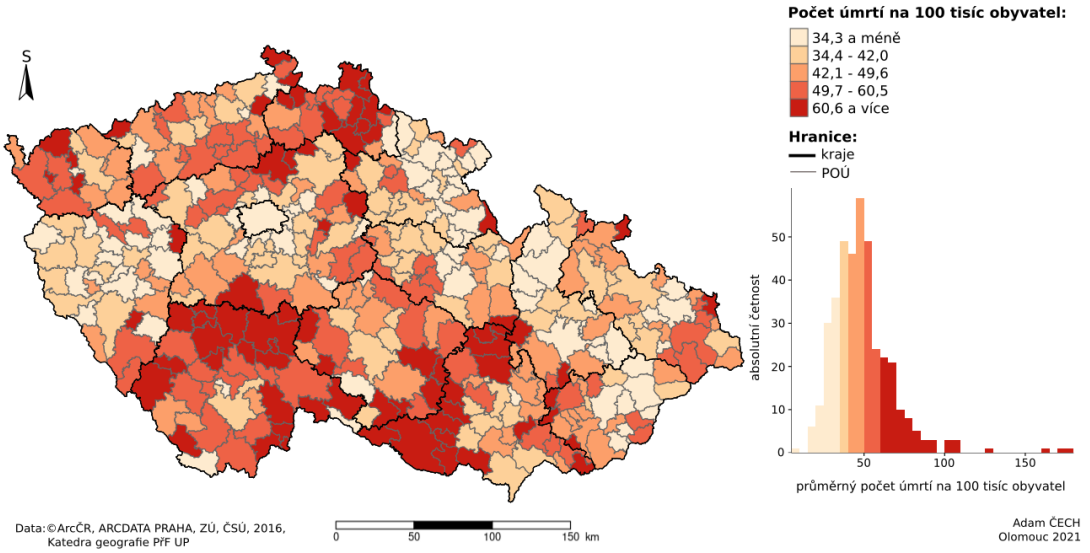
	infarkt myokardu				*rakovina tlustého střeva a konečníku*			
kraje	$\overline{x}$	$\tilde{x}$	SD	CV	$\overline{x}$	$\tilde{x}$	SD	CV
Hl. m. Praha (PHA)	33,52	33,52	-	-	29,39	29,39	-	-
Středočeský (STC)	45,50	42,40	17,69	38,89	33,84	34,77	10,51	31,04
Jihočeský (JHC)	61,41	62,28	17,61	28,67	37,85	35,51	12,85	33,96
Plzeňský (PLK)	36,91	35,26	12,18	33,01	36,77	35,95	10,71	29,12
Karlovarský (KVK)	54,90	48,87	18,43	33,56	40,44	39,11	10,83	26,78
Ústecký (ULK)	49,58	50,17	11,85	23,89	35,58	36,29	7,87	22,13
Liberecký (LBK)	62,84	60,64	15,87	25,25	37,75	33,88	11,41	30,21
Královéhradecký (KHK)	38,87	37,29	11,53	29,66	33,72	32,67	8,74	25,93
Pardubický (PAK)	42,92	40,87	10,70	24,92	31,76	30,91	7,91	24,92
Kraj Vysočina (VYS)	53,33	49,97	15,35	28,79	31,42	31,09	12,29	39,10
Jihomoravský (JHM)	68,31	53,47	37,01	54,17	33,85	34,09	8,32	24,59
Olomoucký (OLK)	38,44	37,03	14,57	37,91	39,02	36,05	8,74	22,39
Zlínský (ZLK)	44,38	44,21	13,73	30,94	35,34	35,83	7,55	21,35
Moravskoslezský (MSK)	43,63	44,54	10,75	24,64	33,61	33,34	7,76	23,08
Česká republika (ČR)	49,04	44,54	20,05	24,64	35,12	34,78	10,00	28,47

Tabulka 2.4: Číselné charakteristiky průměrné úmrtnosti na 100 000 obyvatel pro zkoumané nemoci v krajích a pro celou Českou republiku za období 2014-2018.

Vzhledem k tomu, že výběrový průměr je náchylný na odlehlá pozorování, uvedeme si robustní odhad zkoumaných veličin. K tomuto účelu využijeme hodnotou mediánu, tj. 50% kvantil, pomocí kterého budeme schopni určit, zda dochází k vychýlení hodnot v určitém směru. Při porovnání hodnot z tabulky 2.4 je

[4]Obce s rozšířenou působností.

zřejmé, že v *Karlovarském* a *Jihomoravském* kraji se vyskytují pozorování s vysokým počtem úmrtí. Další důležitý ukazatel je výběrová směrodatná odchylka, která vyjadřuje průměrnou míru vychýlení dat od průměru. V případě *infarktu myokardu* se od ostatních krajů výrazně odlišuje *Jihomoravský* kraj, ve kterém se vyskytují POÚ s vyšším (resp. nižším) počtem úmrtí (obrázku 2.2).

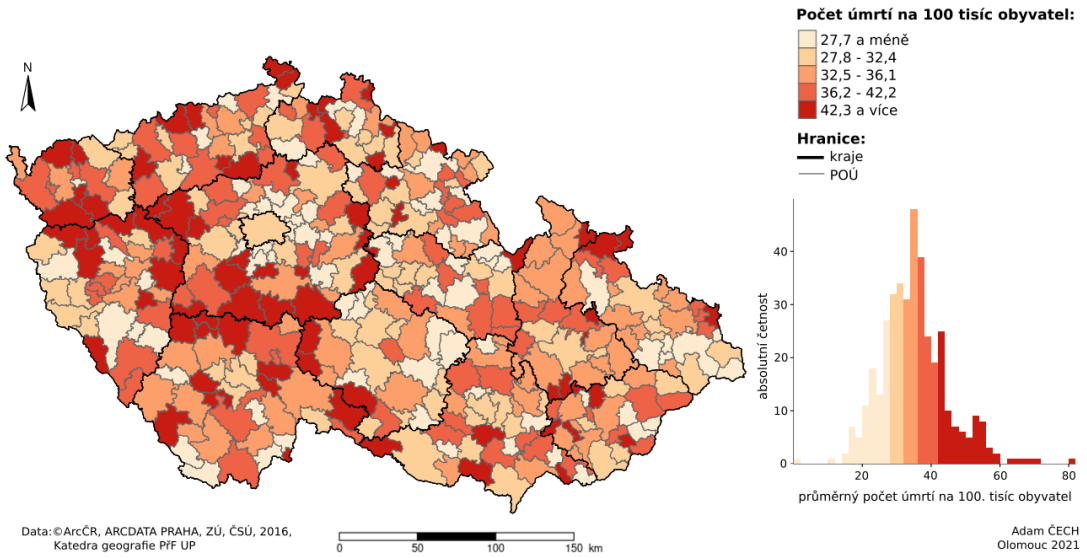


Obrázek 2.2: Průměrný počet úmrtí na *infarkt myokardu* na úrovní jednotlivých krajů.

Aby bylo možné porovnávat hodnoty výběrových směrodatných odchylek, mezi jednotlivými kraji, je nutné vypočítat variační koeficient, který zohledňuje rozdílné výběrové průměry v jednotlivých krajích. Pokud porovnáme hodnoty výběrových směrodatných odchylek v *Plzeňském* a *Karlovarském* kraji, dospějeme jednoznačně k závěru, že v *Karlovarském* kraji je větší rozptyl úmrtnosti než v *Plzeňském* kraji. Pokud pro porovnání vyběrové směrodatné odchylky využijeme variační koeficient, pak zjistíme, že v *Plzeňském* a *Karlovarském* kraji je variabilita úmrtnosti velmi podobná.

Mezi kraji s vysokým počtem úmrtí na *rakovinné onemocnění tlustého střeva a konečníku* patří *Jihočeský*, *Karlovarský*, *Liberecký* a *Olomoucký* kraj. Naopak nejmenší průměrný počet úmrtí se vyskytuje v krajích *Hlavní město Praha*, *Par-*

dubický a *Kraj Vysočina* (tabulka 2.4). Z hodnot mediánů lze pozorovat, že nedochází k tak velkému vychýlení od výběrového průměru jako v případě infarktu myokardu. Mírné vychýlení se vyskytuje v *Olomouckém* a *Libereckém* kraji, protože se na území uvedených krajů vyskytují POÚ s nadprůměrným počtem úmrtí pro daný kraj (obrázek 2.3).



Obrázek 2.3: Průměrný počet úmrtí na *rakovinné onemocnění* na úrovní jednotlivých krajů.

Výsledné hodnoty směrodatné odchylky poukazují na malý rozptyl počtu úmrtí v jednotlivých krajích. Pokud však pro srovnání jednotlivých krajů využijeme variační koeficient, pak dospějeme k závěru, že rozptyly počtu úmrtí jsou velmi podobné. Při porovnání jednotlivých nemocí mezi sebou je zřejmé, že obyvatelé České republiky za období 2014-2018 umírali více na *infarkt myokardu* než na *rakovinné onemocnění tlustého střeva a konečníku*, což je dáno velmi odlišným průběhem onemocnění. Ovšem při srovnání variačního koeficientu dospěje k závěru, že variabilita úmrtnosti je větší v případě rakovinného onemocnění.

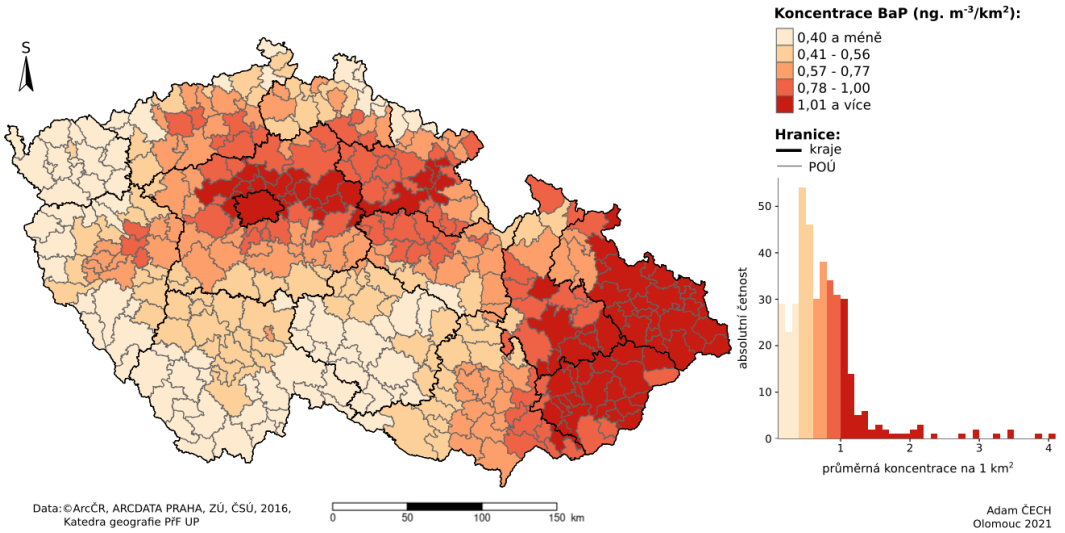
Dále si pomocí číselných charakteristik blíže popíšeme míru znečištění ovzduší *benzo[a]pyrenem*. Největší znečištění se za období 2014-2018 vyskytovalo v *Hlavním městě Praze*, *Olomouckém*, *Zlínském* a *Moravskoslezském* kraji. Nejvíce však

zaujme průměrná hodnota v *Moravskoslezském* kraji, která se velmi odlišuje od ostatních krajů. Při srovnání průměrné hodnoty s mediánem dospějeme k závěru, že v *Moravskoslezském* kraji se vyskytují POÚ s vysokou koncentrací *BaP*. Naopak nejmenší průměrná koncentrace se vyskytovala v *Jihočeském*, *Plzeňském*, *Karlovarském* kraji a *Kraji Vysočina* (tabulka 2.5). Nezajímavější je na uvedených krajích jejich prostorový faktor, a to z důvodu, že se jedná o sousedící kraje.

	Benzo(a)pyren				*Prachové částice* $< 10\ \mu m$				*Prachové částice* $< 2,5\ \mu m$			
kraje	$\overline{x}$	$\tilde{x}$	SD	CV	$\overline{x}$	$\tilde{x}$	SD	CV	$\overline{x}$	$\tilde{x}$	SD	CV
PHA	1,06	1,06	-	-	23,65	23,65	-	-	17,50	17,50	-	-
STC	0,83	0,83	0,23	27,39	20,95	21,09	2,17	10,35	15,69	15,63	1,76	11,19
JHC	0,30	0,31	0,16	54,16	16,03	16,78	2,45	15,29	11,82	12,39	1,87	15,83
PLK	0,46	0,49	0,20	43,86	17,29	17,62	2,63	15,23	12,87	13,16	2,10	16,28
KVK	0,21	0,19	0,09	43,93	14,95	15,13	1,61	10,78	11,09	11,05	1,22	11,04
ULK	0,58	0,57	0,21	35,83	19,49	19,90	2,71	13,90	14,34	14,29	1,88	13,10
LBK	0,55	0,56	0,18	32,33	17,35	17,85	1,93	11,11	13,12	13,44	1,59	12,16
KHK	0,82	0,89	0,23	27,40	20,29	21,10	2,91	14,34	15,45	16,10	2,31	14,97
PAK	0,70	0,70	0,17	23,50	19,94	19,59	1,91	9,59	15,13	14,84	1,53	10,09
VYS	0,37	0,37	0,09	23,62	17,63	17,74	0,84	4,78	13,18	13,28	0,84	6,40
JHM	0,65	0,64	0,18	28,30	21,44	21,71	1,72	8,02	16,74	17,06	1,51	9,02
OLK	0,93	0,88	0,27	29,31	21,10	21,47	3,52	16,69	16,23	16,34	2,94	18,10
ZLK	1,14	1,12	0,17	14,72	22,74	22,78	2,27	9,99	17,77	17,82	1,84	10,34
MSK	2,01	1,89	0,96	47,96	26,64	26,15	6,82	25,59	20,78	20,55	5,52	26,59
ČR	0,75	0,64	0,54	72,03	19,88	19,44	4,06	20,43	15,07	14,63	3,36	22,24

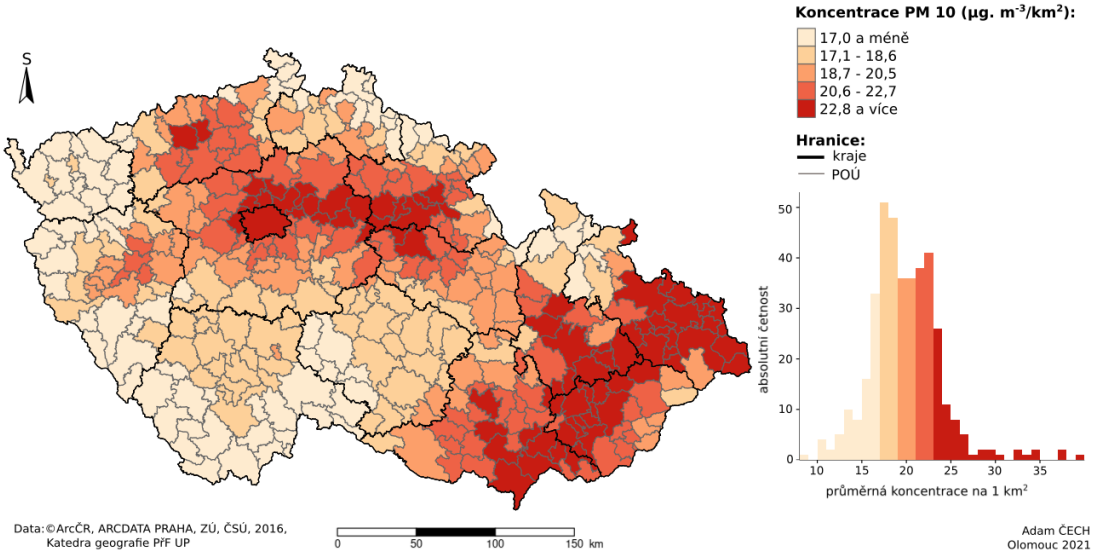
Tabulka 2.5: Číselné charakteristiky průměrné koncentrace látek Benzo(a)pyren [$ng.\ m^{-3}$] a prachových částic [$\mu g.\ m^{-3}$] na 1 km^2 pro zkoumané látky na území České republiky za období 2014-2018.

Největší hodnota směrodatných odchylek se vyskytuje v *Moravskoslezském*, *Olomouckém*, *Královéhradeckém* a *Plzeňském* kraji. Pokud ale k porovnání hodnot směrodatné odchylky využijeme variačního koeficientu, zjistíme, že v *Plzeňském* a *Karlovarském* kraji je variabilita koncentrace *benzo[a]pyrenem* srovnatelná s *Moravskoslezským* krajem. V případě *Olomouckého* kraje je hodnota variačního koeficientu oproti uvedeným krajům nižší, dokonce je srovnatelná s *Jihomoravským* krajem. Největší hodnota variačního koeficientu se nachází v *Jihočeském* kraji, protože se na území uvedeného kraje vyskytují oblasti POÚ s vyšší a zároveň nižší koncentrací *benzo[a]pyrenem* (obrázku 2.4).

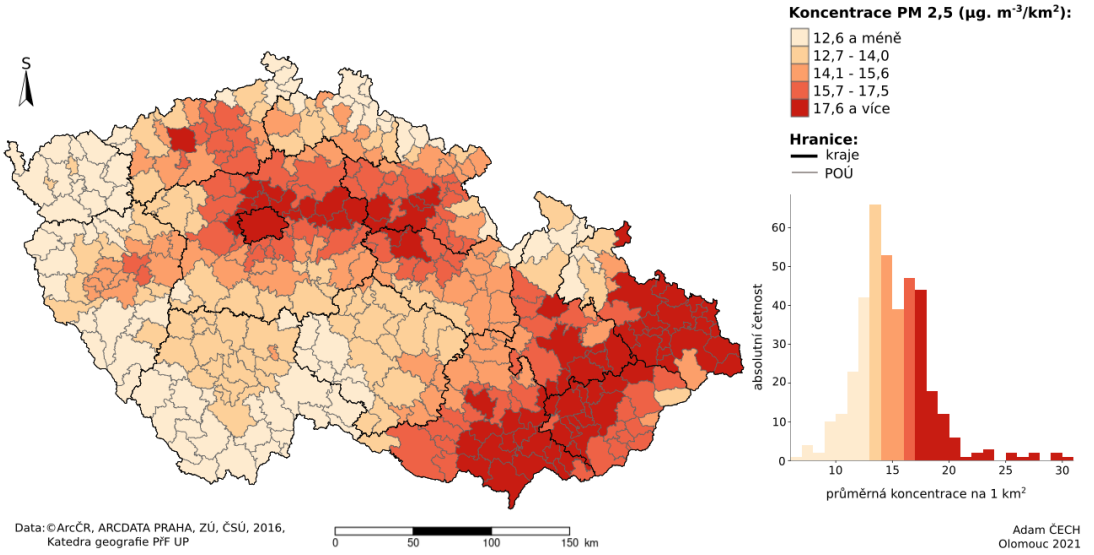


Obrázek 2.4: Průměrné množství *benzo[a]pyren* $[\mu g.m^{-3}]$ pro jednotlivé kraje České republiky za období 2014-2018.

Protože do prachových částic menších jak 10 μm patří i prachové částice menší jak 2,5 μm, bude uvedena číselná charakterizace, vztahující se na obě zkoumané skupiny zároveň. Vysoká hodnota výběrového průměru se vyskytuje v *Hlavním městě Praze*, *Jihomoravském*, *Zlínském* a *Moravskoslezském* kraji. Naopak nejmenší koncentrace se vyskytuje v *Jihočeském*, *Plzeňském* a *Karlovarském* kraji. Protože se hodnoty výběrového průměru skoro neodlišují od hodnot mediánů, lze tvrdit, že se v jednotlivých krajích nevyskytují odlehlá pozorování, která by vychýlila výběrový průměr. Největší variabilita se nachází v *Královéhradeckém*, *Olomouckém* a *Moravskoslezském* kraji, ale při srovnání variačního koeficientu je zřejmé, že v *Moravskoslezském* kraji se vyskytují oblasti jak s nízkou, tak s vysokou koncentrací prachových částic. Naopak v *Jihomoravském* kraji a *Kraji Vysočina* se vyskytuje nejmenší variabilita oproti ostatním krajům tzn. že se v uvedených krajích nevyskytují POÚ s velmi odlišnou koncentrací prachových částic (obrázek 2.5 a 2.6).



Obrázek 2.5: Průměrné množství prachových částic menší jak 10 μm [$\mu g.m^{-3}$] pro jednotlivé kraje České republiky za období 2014-2018.

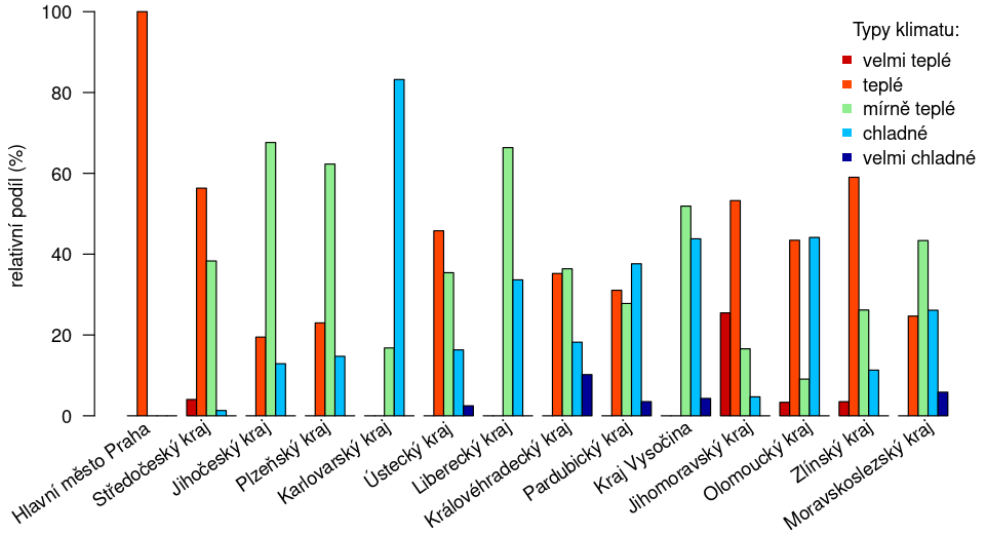


Obrázek 2.6: Průměrné množství prachových částic menší jak 2,5 μm [$\mu g.m^{-3}$] pro jednotlivé kraje České republiky za období 2014-2018.

Při charakterizaci jednotlivých látek na úrovní České republiky lze pozorovat, že v případě *benzo[a]pyrenu* se výběrový průměr velmi přibližuje k ročnímu imisnímu limitu (1 $ng.m^{-3}$). Srovnáme-li výběrový průměru s mediánem, je zřejmé,

že na území České republiky se nacházejí oblasti s velmi nadprůměrnou hodnotou, což potvrzuje i histogram uvedený na obrázku 2.4. Pro prachové částice platí, že na území České republiky se nacházejí oblasti s velmi nadprůměrnou koncentrací, ovšem ne tak extrémní jako v případě *benzo[a]pyrenu*. Uvedenou charakterizaci je pomocí porovnání mediánu a výběrového průměru v tabulce 2.5. Při porovnání jednotlivých variačních koeficientů dospějeme k závěru, že variabilita koncentrace mezi jednotlivými prachovými částicemi je srovnatelná, ale v případě *benzo[a]pyrenu* je diametrálně odlišná.

Poslední číselná charakteristika je zaměřena na převažující lokální klima na území jednotlivých POÚ. Aby bylo možné porovnávat jednotlivé kraje mezi sebou jsou jednotlivé kategorie interpretovány, jako podíl rozlohy v jednotlivých krajích (obrázek 2.7).



Obrázek 2.7: Klimatické podmínky v jednotlivých krajích České republiky.

Na území České republiky se nejčastěji vyskytuje teplé a mírné klima, naopak nejmenší zastoupení mají klimatické podmínky značené jako velmi teplé (resp. chladné). Ovšem najdou se i takové kraje např. *Karlovarský*, ve kterém je významně zastoupeno chladné klima. V případě *Jihomoravského* kraje pozorujeme větší zastoupení velmi teplého klimatu, vůči ostatním krajům, přesně se jedná o 25 % z celkového území. Naopak největší míra zastoupení velmi chladného kli-

matu se vyskytuje v *Královéhradeckém* kraji, přibližně se jedná o 15 % území. Z obrázku 2.7 je možné usoudit, že v každém kraji kromě Hlavního města Prahy se vyskytují velmi rozmanité typy lokálního klimatu.

Při vypracování uvedených kapitol byly využity tyto zdroje: [3], [12], [16], [20].

Kapitola 3

Prostorová analýza mortalitních dat

V této kapitole budou zkoumáno, zda má geografický prostoro vliv na hodnoty výše uvedených mortalitních dat. Dříve, než budeme aplikovat metody prostorové analýzy, je nutné si zadefinovat matici vah. Vzhledem k tomu, že zkoumaná onemocnění nejsou virového, ani bakteriálního původu, tj. nedochází k přenosu z člověka na člověka, je pro určení sousedství využita metoda *sousedství královny* a hodnoty vah stanoveny pomocí metody *prostorových souvislostí*. Metoda *sousedství královny* v průměru každé POÚ přiřadí šest sousedů, ale nastává i situace, že POÚ, které jsou umístěny v blízkosti státních hranic, je přiřazen pouze jediný soused (tabulka 3.1). Přesně se jedná o Aš, Rumburk, Javorník, Nové Město pod Smrkem a Osoblahu.

minimum	maximum	výběrový průměrná hodnota	medián	výběrová směrodatná odchylka
1,0	11,0	5,5	5,0	2,0

Tabulka 3.1: Číselná charakterizace rozložení sousedství, určené pomocí metody *sousedství královny*.

Aby bylo možné při analýze zohlednit i rozdílný počet sousedů, je nutné *binomickou matici vah* standardizovat, a to pomocí *řádkové* standardizace i za předpokladu porušení symetrie. Další ztížení, které je nutno zohlednit, se tyká vojenských újezdů, které se řadí mezi 393 zkoumaných POÚ. Problematika vojenských újezdů spočívá v tom, že se na jejich území nevyskytuje zkoumané statistické znaky, tj. počet úmrtí na dané onemocnění. Jedna z možností, jak uvedený problem vyřešit, je odebrání vojenských újezdů z datové sady, což by však mělo

za následek zkreslení prostorového vlivu u sousedících jednotek. Proto je každému vojenskému újezdu přiřazena průměrná hodnota zkoumaného statistického znaku pro celou Českou republiku, která minimálně ovlivní výsledky prostorové analýzy.

Veškeré výsledky prostorové analýzy a mapová díla jsou vypočteny (resp. vytvořeny) pomocí programu *PyCharm Community 2021.1* umožňující programování v jazyku **R** (verzi 3.6.3). Veškeré kódy a manuál k tvorbě mapových děl pomocí jazyku **R** se nachází na přiloženém CD.

3.1. Prostorová autokorelace

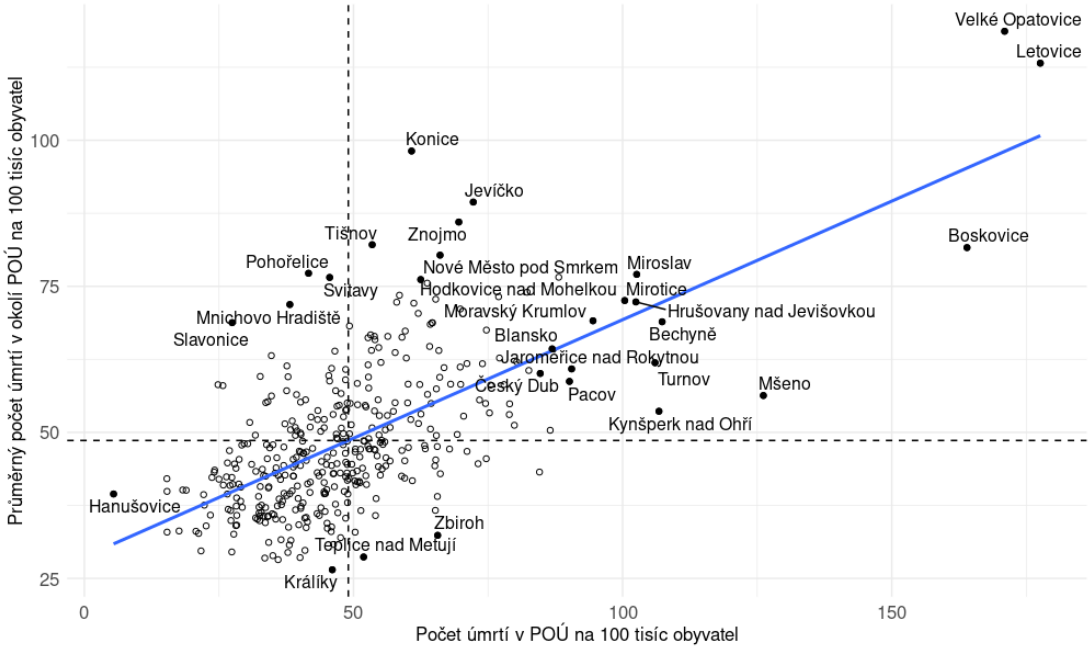
Pro ověření prostorových shluků (resp. prostorové autokorelace) je využito *globální* a *lokální* Moranovo I kritérium. Při výpočtu *globálního Moranova I kriteria* lze vyvodit závěr, že u obou typů onemocnění dochází k pozitivní prostorové autokorelaci (tabulka 3.2). Aby bylo zřejmé, že uvedená tvrzení nejsou pouhou náhodnou, je jejich signifikantnost ověřena pomocí metody *Monte Carlo*. Pro zjištění rozdělení zkoumaných veličin bylo využito jeden tisíc permutací a hladina významnosti byla zvolena na hodnotu 0,05. Protože je p-hodnota v obou případech menší než zvolená hladina významnosti, je nulová hypotéza odmítnuta a přikláníme se k alternativní hypotéze, tzn. že se na zkoumaném území vyskytuje prostorová autokorelace.

onemocnění	Moranovo I kritérium	p-hodnota
infarkt myokardu	0,41	0,001
rakovina tlustého střeva a konečníku	0,13	0,001

Tabulka 3.2: *Globální Moranovo I kritérium* a jeho statistická významnost.

Při srovnání výsledných hodnot pozorujeme, že úmrtí na *infarkt myokardu* má větší tendenci k prostorovým shlukům, než *rakovinné onemocnění tlustého střeva a konečníku*. Pomocí uvedené charakteristiky však nejsme schopni lokalizovat jednotlivé prostorové shluky, a proto vypočítáme *lokální Moranovo I kritérium*. Protože samotná hodnota lokálního kriteria nám neumožní rozeznat jednotlivé typy shluků, využijeme pro interpretaci Moranův diagram (obrázek 3.1 a 3.3).

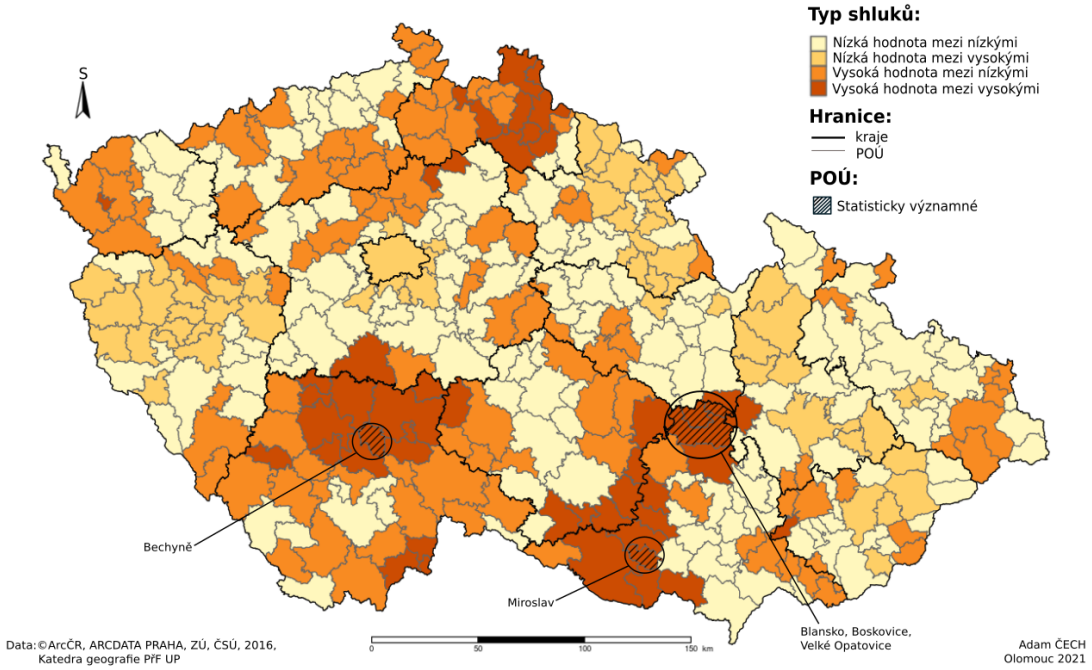
V případě infarktu myokardu je většina hodnot umístěna v prvním a třetím kvadrantu Moranova diagramu. Pokud zároveň data proložíme regresní funkcí, dostaneme vyjádření pro globální Moranovo I kritérium. Pozorování, která mají největší vliv na výslednou hodnotu jsou pro lepší orientaci označeny názvem POÚ. V porovnání vůči ostatním pozorováním, nejvíce zaujmou prostorové shluky v *Boskovicích, Letovicích* a *Opatovicích*, které jsou výrazně nadprůměrné než u ostatních POÚ (obrázek 3.1). Přesto se však nejedná o odlehlá pozorování protože Cookova vzdálenost je menší než jedna.



Obrázek 3.1: Moranův diagram pro *infarkt myokardu*.

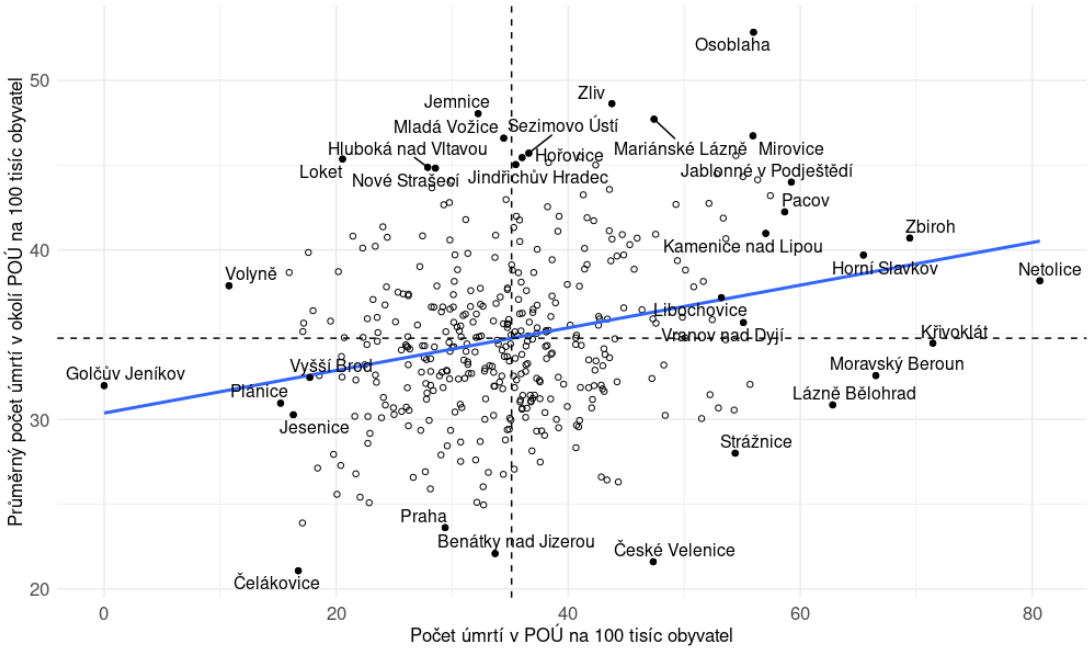
Nedostatek Moranova I digramu je v tom, že neumožňuje sledovat umístění jednotlivých prostorových shluků. Proto pomocí mapového díla (obrázek 3.2) vyjádříme, v jakém kvadrantu Moranova diagramu se POÚ vyskytuje. Pomocí mapového díla pozorujeme, že na území České republiky výrazně převažují lokální shluky nízkých hodnot, obklopeny nízkými hodnotami. Při zaměření se na výskyt vysokých hodnot, které jsou obklopeny vysokými hodnotami zjistíme, že největší zastoupení je v *Jihočeském, Jihomoravském* a *Libereckém* kraji. Ovšem nejdůležitějším faktorem pro uvedenou skupinu je výskyt statisticky signifikantních

POÚ, mezi které patří *Bechyně, Blansko, Boskovice, Miroslav* a *Velké Opatovice*. Vzhledem k tomu, že Blansko, Boskovice a Velké Opatovice spolu přímo sousedí, je možné na základě uvedené analýzy říci, že se v těchto POÚ vyskytují určité faktory, které zvyšují úmrtí na infarkt myokardu vůči okolním POÚ.



Obrázek 3.2: Prostorové shluky *infarktu myokardu*.

U rakovinného onemocnění tlustého střeva a konečníku pozorujeme, že rozptyl prostorových jednotek v jednotlivých kvadrantech Moranova diagramu je větší než u infarktu myokardu (obrázek 3.3). Uvedená charakteristika se projeví při výpočtu globálního kriteria, což potvrzuje i hodnota uvedená v tabulce 3.2. Rozdíl mezi globální hodnotou prostorové autokorelace lze pozorovat pomocí srovnání sklonu regresních funkcí, které jsou uvedeny v jednotlivých Moranových diagramech (obrázek 3.1 a 3.3). Mezi nejvíce ovlivňující okolní prostorové jednotky patří např. *Osoblaha, Netolice, České Velenice, Čelákovice* a další POÚ, u kterých je uveden název, ale ani v jednom případě se nejedná o odlehlé pozorování, protože ani v jednom případě Cookova vzdálenost nepřesáhne hodnotu větší jak jedna.

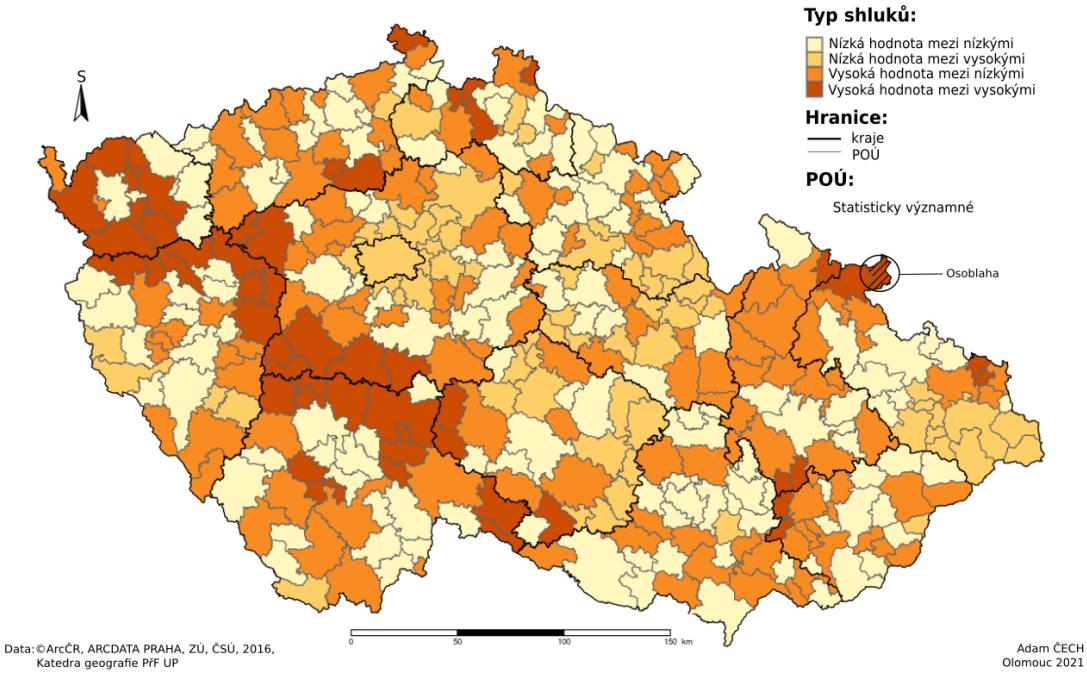


Obrázek 3.3: Moranův diagram pro *rakovinné onemocnění tlustého střeva a konečníku*.

Při vizualizaci hodnot z Moranova diagramu, pomocí mapového díla (obrázek 3.4) je hned zřejmé, proč je hodnota globálního Moranova I kritéria menší, než v případě infarktu myokardu. Přesto se na území České republiky nacházejí oblasti charakterizující se určitými typy shluků. Například v Karlovarském kraji vidíme převažující výskyt shluků vysokých hodnot mezi vysokými hodnotami, tento typ shluku se následně line okolo hranic *Středočeského, Plzeňského* a *Jihočeského* kraje. Uvedený fakt však není relativní a to z důvodu, že jediná signifikantní POÚ se vyskytuje v *Moravskoslezském kraji*. V jediné obci *Osoblaha* bylo statisticky prokázáno, že se na daném území vyskytuje zvýšený počet úmrtí na rakovinné onemocnění než v okolních POÚ.

Pro ověření signifikantnosti v obou případech byla využita metoda *Monte Carlo*, kterou jsme si uvedli při popisu *lokálního Moranova I kritéria*. Aby bylo vhodně odhadnuto rozdělení zkoumané náhodné veličiny bylo použito jeden tisíc permutací a pro snížení míry falešně pozitivních signifikantních shluků byla využita *Benjaminova-Hochbergova* korekce pro mnohonásobné porovnávání na

celkové hladině významnosti 0,05.



Obrázek 3.4: Prostorové shluky *rakovinného onemocnění tlustého střeva a konečníku.*

3.2. Prostorová nestacionarita

Pomocí prostorové autokorelace jsme zjistili, že na území České republiky se vyskytují prostorové shluky, u kterých jsou signifikantně zvýšeny počty úmrtí. Proto nás intuitivně napadne otázka „které faktory způsobují zvýšený počet úmrtí a zda jsou prostorově stacionární“. K tomuto účelu budeme na uvedená data aplikovat standardní regresní analýzu a *váženou prostorovou regresi*, která nám umožní zkoumat prostorovou nestacionaritu vybraných faktorů, mezi které patří kvalita ovzduší a převažující klimatické podmínky vyskytujících se na území POÚ.

Vzhledem k tomu, že mezi počtem úmrtí a vybranými faktory neexistuje žádná významná závislost, která by mohla být dále zkoumaná z prostorového hlediska, bude metodika zkoumání prostorové stacionarity ukázaná na modelu, který vyjadřuje závislost koncentrace prachových částic menších jak 10 μm (vysvětlovaná

proměnná) vůči koncentraci benzo[a]pyrenu a klimatickým podmínkám (vysvětlující proměnné). Výsledný model je ve tvaru

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 a_i + \beta_3 b_i + \beta_4 c_i + \beta_5 d_i + \beta_6 x_i a_i + \beta_7 x_i b_i + \beta_8 x_i c_i + \beta_9 x_i d_i + \varepsilon_i,$$

kde umělé proměnné vyjadřující jednotlivé klimatické podmínky vyjádříme následovně

$$a_i = \begin{cases} 0 & \text{jiné klima} \\ 1 & \text{velmi chladné} \end{cases}, \quad b_i = \begin{cases} 0 & \text{jiné klima} \\ 1 & \text{chladné} \end{cases}$$

$$c_i = \begin{cases} 0 & \text{jiné klima} \\ 1 & \text{teplé} \end{cases}, \quad d_i = \begin{cases} 0 & \text{jiné klima} \\ 1 & \text{velmi teplé} \end{cases}.$$

Protože se ve zkoumaném modelu vyskytují interakce a umělé proměnné, je nutné zvolit referenční skupinu, která v tomto případě je rovná *mírnému klimatu*, a proto je výsledný model pro uvedenou skupinu ve tvaru $\beta_0 + \beta_1 x_i$. Důvodem zvolení uvedené referenční skupiny je v tom, že se jedná o přechodnou klimatickou oblast, a proto umožní vhodně zkoumat rozdíl mezi teplým a chladným klimatem. V tabulce 3.3 jsou znázorněny odhady regresních koeficientů pro jednotlivé umělé proměnné $(\beta_0, \beta_2, \beta_3, \beta_4, \beta_5)$, které vyjadřující posun regresní přímky vzhledem k mírnému klimatu a odhady regresních koeficientů pro interakce $(\beta_1, \beta_6, \beta_7, \beta_8, \beta_9)$, vyjadřující rozdíl ve směrnici regresní přímky.

	umělá proměnná		*interakce*	
klima	regresní koeficient	p-hodnota (t-test)	regresní koeficient	p-hodnota (t-test)
velmi chladné	-3,55	0,001	0,78	0,501
chladné	-0,60	0,027	-0,65	0,102
mírné	14,73	0,001	6,60	0,001
teplé	1,91	0,001	-0,78	0,003
velmi teplé	4,62	0,001	-2,12	0,048

Tabulka 3.3: Číselná charakterizace*globálního modelu* pro *prachové částice* menší než 10 μm v závislosti na *benzo[a]pyrenu*.

Při zaměření se na mírné klima z tabulky 3.3 vidíme, že pokud bude koncentrace *benzo[a]pyrenu* nulová, pak se na území mírného klimatu vyskytuje koncentrace *prachových částic* menší než 10 μm o průměrné hodnotě 14,73 $\mu g.\ m^{-3}$

na 1 km^2, naopak při jednotkové změně koncentrace *benzo[a]pyrenu* se v mírném klimatu zvýší koncentrace *prachových částic* o 6,60 $\mu g.\ m^{-3}$ na 1 km^2. V obou uvedených případech je p-hodnota menší, jak zvolená hladina významnosti 0,05, a proto se budeme přiklánět k alternativní hypotéze, která říká, že odhad regresního koeficientu je různý od nuly. U teplého a velmi teplého kraje pozorujeme, že při nulové koncentraci *benzo[a]pyrenu* se průměrná hodnota zvýší vzhledem k mírnému klimatu, naopak v chladném a velmi chladném klimatu dochází ke snížení koncentrace. Oproti tomu u jednotkové změny koncentrace se statistická signifikantnost projevuje pouze u mírného, teplého a velmi teplého klimatu. V případě teplých klimatech dochází ke snížení koncentrace *prachových částic* při jednotkové změně *benzo[a]pyrenu* vůči mírnému klimatu. Abychom však zjistili skutečnou hodnotu pro interakce pro dané klima, musíme danou hodnotu přičíst k odhadu regresního koeficientu v případě mírného klimatu, tzn. že při jednotkové změně koncentrace *benzo[a]pyrenu* se v mírném klimatu zvýší koncentrace *prachových částic* o 5,82 $\mu g.\ m^{-3}$ na 1 km^2. V případě všech uvedených charakteristik je p-hodnota menší, jak hladina významnost, a proto je můžeme označit za statisticky signifikantní.

Abychom však měli přesnější představu o vytvořeném modelu uvedeme si dvě charakteristiky, mezi které patří adjustovaný koeficient determinace, který vyjadřuje podíl variability vysvětlené modelem a p-hodnotu pro Shapiro-Wilkův test normality, pomocí kterého budeme testovat, zda hodnoty rezidui (resp. odhady náhodných chyb) pocházejí z normálního rozdělení (tabulka 3.4).

charakteristika	hodnota
adjustovaný koeficient determinace	0,93
p-hodnota Shapiro-Wilkův test	0,14

Tabulka 3.4: Číselná charakterizace uvedeného *globálního* modelu.

Pokud hodnotu adjustovaného koeficientu vynásobíme stem, pak dostaneme vyjádření míry vysvětlené variabilit pomocí modelu v procentech a proto lze říci, že uvedený model popisuje 93 % variability v datech. Uvedená charakteristika nám tedy říká, že vytvořený model vhodně popisuje závislost mezí koncentraci prachových částic menší jak 10 μm vůči koncentraci benzo[a]pyrenu a klimatic-

kým podmínkách. V případě Shapiro-Wilková testu normality se p-hodnota realizuje hodnotou větší jak zvolena hladina významnosti (tj. 0,05), a proto lze říci, že rezidua uvedeného modelu se řídí normálním rozdělením, tzn. že nejdůležitější podmínka pro odhad regresních koeficientů pomocí metody nejmenších čtverců je splněná.

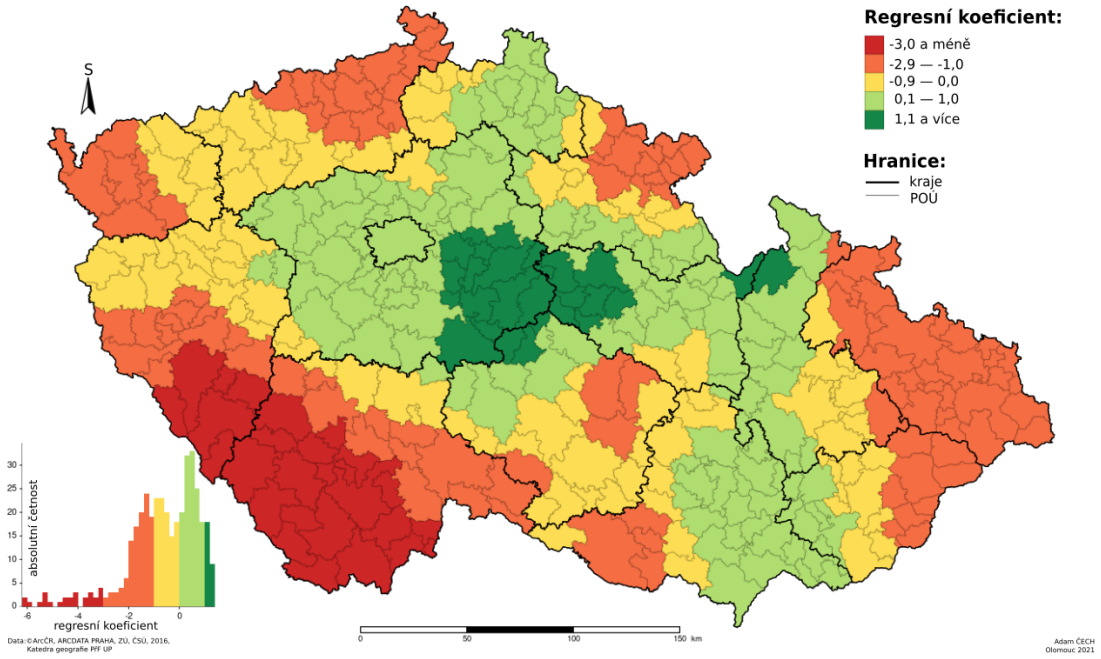
Abychom však mohli říci, že uvedené závislosti platí pro každou část České republiky, je nutné globální model otestovat na předpoklad prostorové stacionarity, k čemu využijeme *prostorově váženou regresi* a metodu Monte Carlo, kterou jsem si uvedli v teoretické části popisující prostorovou nestacionaritu. V tabulce 3.5 jsou znázorněny číselné charakteristiky (tzn. minimum (x_{min}), medián ($\bar{x}$) a maximum (x_{max})), které umožní si vytvořit představu o variabilitě regresních koeficientů na území České republiky. Pro metodu Monte Carlo je v tabulce 3.5 zaznamenána p-hodnota, která umožní rozhodnout o prostorové nestacionaritě (resp. stacionaritě).

	umělá proměnná				*interakce*			
klima	x_{min}	$\bar{x}$	x_{max}	p-hodnota	x_{min}	$\bar{x}$	x_{max}	p-hodnota
velmi chladné	-6,28	-2,23	7,43	0,578	-21,11	-1,48	14,68	0,568
chladné	-6,02	-1,08	0,58	0,127	-1,96	0,46	23,61	0,568
mírné	11,37	13,73	17,58	0,001	3,58	8,50	13,56	0,001
teplé	-0,61	1,24	6,14	0,003	-6,12	-0,49	1,30	0,026
velmi teplé	-1,94	2,44	8,29	0,300	-10,68	-0,58	1,87	0,430

Tabulka 3.5: Číselná charakterizace *prostorově vážené regrese* uvedeného *globálního* modelu.

Vzhledem k tomu, že se u regresních koeficientů globálního modelu vyjadřující interakce pro chladné a velmi chladné klima neprokázala statistická signifikantnost, budou blíže interpretovány pouze mírné, teplé a velmi teplé klima. V případě mírného klimatu vidíme, že interakce a umělé proměnné se na území České republiky vyskytují pouze s kladnými regresními koeficienty (tzn. prostorovou stacionaritu), a proto je výsledná p-hodnota metody Monte Carla v tomto případě irelevantní. Proto lze říci, že závislost, kterou jsem určili pro kombinace koncentrace benzo[a]pyrenu s mírným klimatem pomocí globálního modelu je možné aplikovat na všechna místa České republiky, tzn. že při jednotkové změněn koncentrace benzo[a]pyrenu v jakémkoliv POÚ se hodnota koncentrace prachových

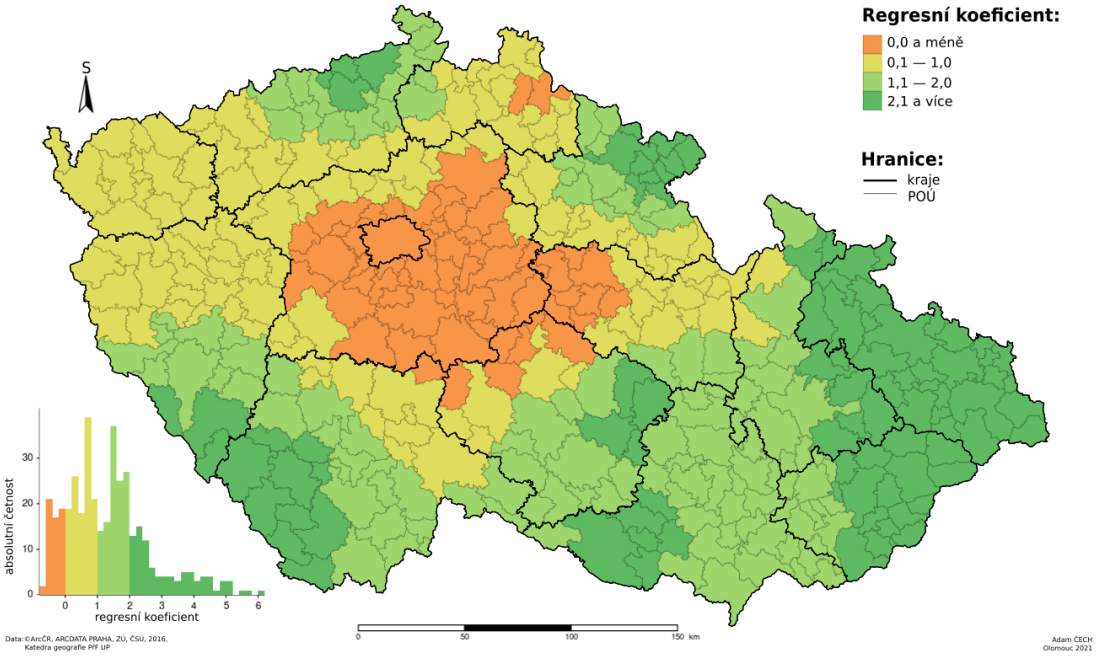
částic zvětší o 6,60 $\mu g.\ m^{-3}$. Oproti tomu v případě teplého klimatu se prostorová stacionarita neprokázala, jak můžeme vidět i na obrázku 3.5 znázorňující variabilitu regresního koeficientu vyjadřujícího směrnici regresní přímky.



Obrázek 3.5: Prostorová nestacionarita *globálního* modelu, vyjádřena pomocí hodnot interakce regresních koeficientů pro teplém klima, v kombinaci s *benzo[a]pyrenu*.

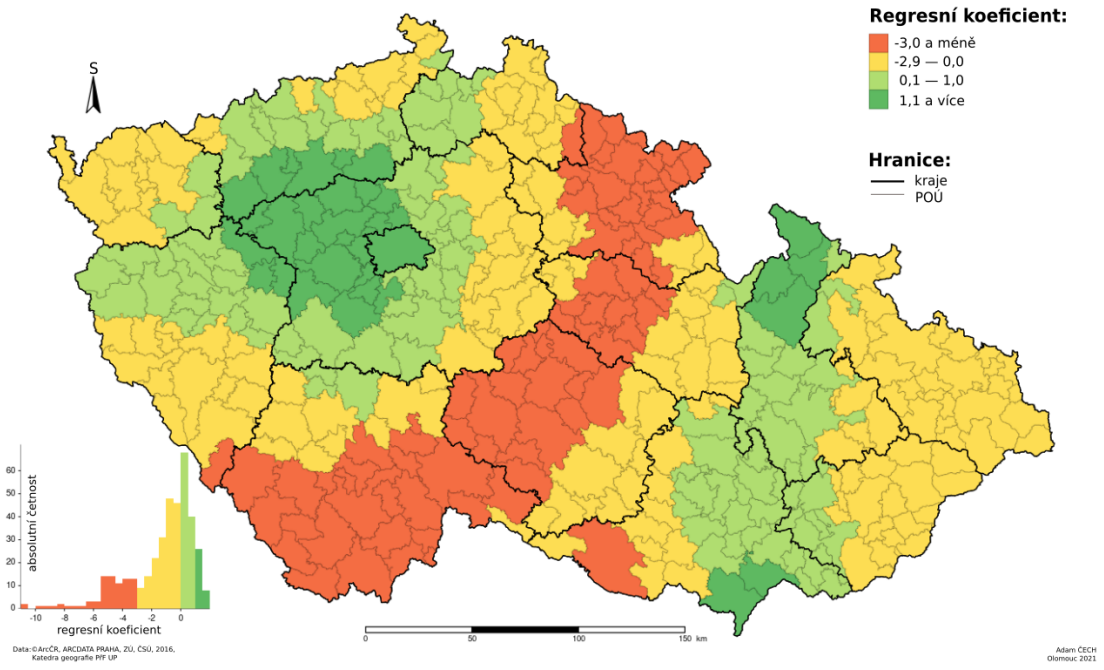
Uvedená charakteristika poukazuje na to, že koncentrace benzo[a]pyrenu s kombinací teplého klimatu není faktorem, který by signifikantně ovlivňoval míru koncentrace prachových částic tzn. že pro vysvětlení míry koncentrace prachových částic existuje jiné odůvodnění (resp. faktory). Pokud si však položíme jednoduchou otázku, „proč se v určitých částech zkoumané oblasti vyskytují rozdílné hodnoty regresních koeficientů“, pak můžeme prostorovou nestacionaritu využít v náš prospěch. Pomocí podrobné analýzy POÚ, ve kterých se vyskytují podobná hodnota regresních koeficientů, umožní odhalit souvislosti, které nejsou hned zřejmé. V případě kombinace benzo[a]pyrenu a teplého klimatu, je na místě podrobně zanalyzovat POÚ, podle toho, jak jsou rozděleny na obrázku 3.5. Prostorová nestacionarita se u teplého klimatu projevila i u regresních koeficientů pro

umělé proměnné (obrázek 3.6). V tomto případě by bylo vhodné podrobně zanalyzovat Hlavní město Prahu a POÚ blízkém okolí, jak lze pozorovat na obrázku 3.6.

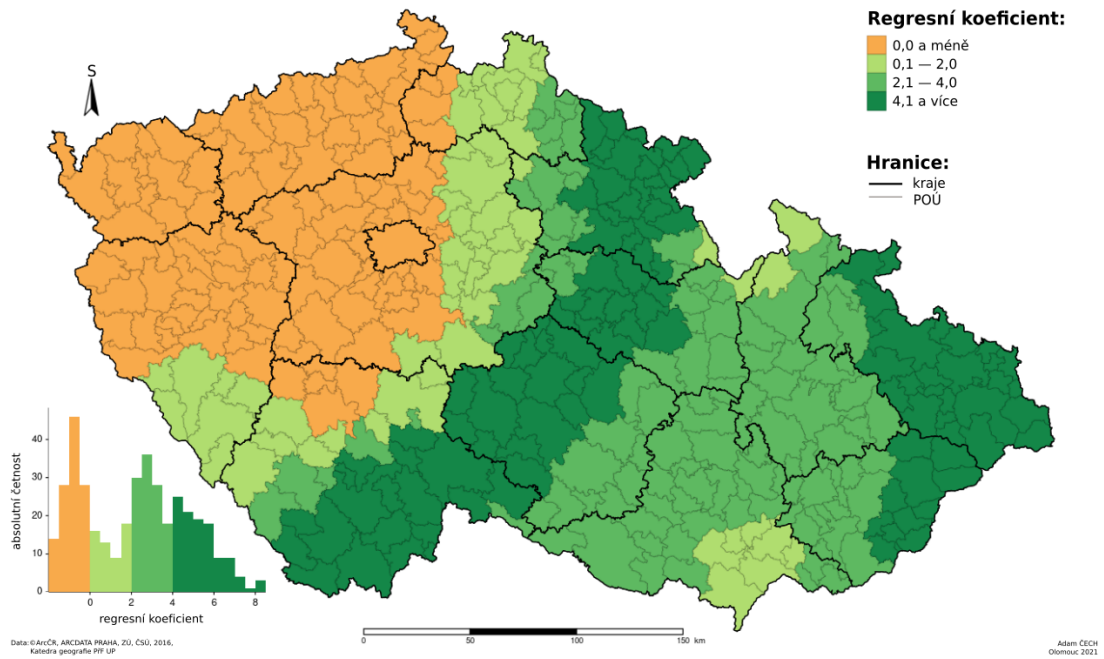


Obrázek 3.6: Prostorová nestacionarita *globálního* modelu, vyjádřena pomocí hodnot umělé proměnné regresních koeficientů pro teplém klima, v kombinaci s *benzo[a]pyrenu*.

V případě velmi teplého klimatu se prostorová stacionarita prokázala, jak u regresních koeficientů pro iteraci, tak i zároveň pro umělé proměnné, přestože se na území České republiky vyskytují POÚ, jak s negativními tak pozitivními regresními koeficienty (obrázek 3.7 a 3.8). Vzhledem k uvedené charakterizaci lze tvrdit, že při jednotkové změně koncentrace benzo[a]pyrenu v kombinaci s velmi teplým klimatem se koncentrace prachových částic sníží o 4,48 $\mu g.\ m^{-3}$ vzhledem k mírnému klimatu. V případě umělé proměnné lze tvrdit, že při nulové koncentraci benzo[a]pyrenu a při velmi teplém klimatu bude koncentrace prachových částic větší o 4,62 $\mu g.\ m^{-3}$ oproti mírnému klimatu. Obě uvedené charakteristiky lze aplikovat u každé POÚ vyskytující se na území České republiky a to z důvodu, že je u obou uvedených charakteristik splněn předpoklad stacionarity.



Obrázek 3.7: Prostorová stacionarita interakce, pro velmi teplém klima v kombinaci s *benzo[a]pyrenu*.



Obrázek 3.8: Prostorová stacionarita umělé proměnné, pro velmi teplém klima v kombinaci s *benzo[a]pyrenu*.

Při odhadu regresních koeficientů pomocí metody *prostorově vážené regrese*, bylo využito *fixní* prostorové jádro, u kterého byla prahová vzdálenost zvolena pomocí metody křížové validace a pro výpočet vah bylo využito *Gaussovo* jádro.

Závěr

V této práci jsme si podrobně přiblížili dvě základní prostorové charakteristiky, které umožňují blíže analyzovat prostorový faktor zkoumaných statistických znaků. Hlavním úkolem prostorové autokorelace a prostorové nestacionarity není hned vytvářet pevná stanoviska ohledně zkoumaných statistických znaků, ale poukázat na oblasti, které se odchylují od průměru, a tím nám dávají určitý nadhled pro odhalení netušených závislostí.

Při zkoumaní prostorové autokorelace na vybraná mortalitní data jsem zjistil, že v případě infarktu myokardu se na území České republiky vyskytují určité prostorové shluky, které by bylo vhodné blíže analyzovat, a tím získat možné odhalení konkretních důvodů zvyšující počet úmrtí. V případě aplikace prostorově vážené regrese, byly odhaleny příčiny, které ovlivňují koncentraci prachových částic menší než 10 μm. V případě mírného a velmi teplého klimatu byla odhalena silná lineární závislost mezi koncentrací benzo[a]pyrenu a prachovými částicemi, což je dáno tím, že zdroj obou zkoumaných látek je skoro totožný. Ovšem rozdílnost hodnot mezi velmi teplým a mírným klimatem není tak zřejmá a odůvodnění bychom mohli hledat v rozdílných fyzikálních vlastnostech zkoumaných látek. Z uvedeného shrnutí je zřejmé, že přidání prostorového faktoru do analýzy umožňuje nalézt závislosti a faktory, které nejsou na první pohled zřejmé.

Literatura

[1] Anselin, L. (1995). *Local Indicators of Spatial Association—LISA*. Columbus: Geographical Analysis 27, č.1, s.93-115.

[2] Anselin, L. et al. (2001). *A companion to theoretical econometrics*. New Jersey: Blackwell Publishing.

[3] Arsenović, D., Lehnert, M., Fiedor, D., Šimáček, P., Středová, H., Středa, T., Savić, S. (2019). *Heat-waves and mortality in Czech cities: A case study for the summers of 2015 and 2016*. Novi Sad: Geographica Pannonica vol. 23, br. 3, str. 162-172.

[4] Bellefon M.P., Floch J.M., Audric, S., Durieux, E., et al. (2018). *Handbook of Spatial Analysis, Theory and practical application with R*. Paris: Insee.

[5] Brunsdon, Ch., Fotheringham A.S., Charlton M.E. (1996). *Geographically Weighted Regression: A method for Exploring Spatial Nonstationarity*. Columbus: Geographical Analysis 28, č.4, s.281-298.

[6] Cliff, A.D., Ord. J.K. (1973). *Spatial autocorrelation*. London: Pion.

[7] Fišerová, E., (2015).*Lineární statistické modely*, 2. vydání. skripta PřF UP. Olomouc: Vydavatelství Univerzity Palackého.

[8] Fotheringham, S.A., Brunsdon, Ch., Charlton, M., (2002). *Geographically Weighted Regression – the Analysis of Spatially Varying Relationships*. London: John Wiley & Sons.

[9] Fu, W.J., Jiang P.K., Zhou G.M., Zhao K. L. (2014). *Using Moran's I and GIS to study the spatial pattern of forest litter carbon density in a subtropical region of southeastern China*. European Geosciences Union: Biogeosciences.

[10] Guy, L., Cheshire, J. (2016). *An Introduction to Spatial Data Analysis and Visualisation in R*. London: University College London.

[11] Horák, J., Orlíková, L., Joaquin, O.A., Svoboda, R. (2020). *Prostorové regresní modelování s příklady*. Ostrava: GIS Ostrava 2020.

[12] Hrnčiarová, T., Mackovčin, P., Zvara, I., et al. (2009). *Atlas krajiny České republiky: Landscape atlas of the Czech Republic*. Praha: Ministerstvo životního prostředí České republiky.

[13] Marek, L. (2015). *Prostorové a vícerozměrné statistické analýzy epidemiologických dat*. Olomouc: Vydavatelství Univerzity Palackého.

[14] Mhbubul, A. (2020). *Spatial Autocorrelation: Neighbors Affecting Neighbors*. Kanada: Towards data science [online]. Dostupné z: https://towardsdatascience.com/spatial-autocorrelation-neighbors-affecting-neighbors-ed4fab8a4aac.

[15] Mitchell, A. (2005). *How Spatial Autocorrelation (Global Moran's I) works*. West Redlands: ESRI Press[online]. Dostupné z: https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm.

[16] Quitto, E. (1971). *Klimatické oblasti Československa-Climatic regions of Czechoslovakia*. Brno: Geografický ústav ČSAV.

[17] Smith, T.E. (2014). *Spatial Weight Matrices*. Philadelphia: University of Pennsylvania.

[18] Spurná, P. (2008). *Geografická vážená regrese: Metoda analýzy prostorové nestacionarity geografických jevů*. Praha: Sociologický časopis 113, č. 2, s. 21–35.

[19] Spurná, P. (2008). *Prostorová autokorelace – všudypřítomný jev při analýze prostorových dat?*. Praha: Sociologický časopis 44, č. 4, s. 767–787.

[20] Šrám, R.J., Dostal, M., Líbalová, H., Rossner, P., Rossnerová, A., Švecová V.,Topinka, J., Bartonová, A. (2013). *The European Hot Spot of B[a]P and PM2.5 Exposure—The Ostrava Region, Czech Republic: Health Research Results*. Praha: Institute of Experimental Medicine AS CR.

[21] Yanguang, Ch. (2013). *New Approaches for Calculating Moran's Index of Spatial Autocorrelation*. San Francisco: Public Library of Science.

Seznam tabulek

2.1 Částečná ukázka datové sady. 45
2.2 Ukázka datové sady četností úmrtí na *infarkt myokardu*. 45
2.3 Rozdíl mezi absolutní a relativní hodnotou počtu úmrtí na *infarkt myokardu* za rok 2018. 46
2.4 Číselné charakteristiky průměrné úmrtnosti na 100 000 obyvatel pro zkoumané nemoci v krajích a pro celou Českou republiku za období 2014-2018. 50
2.5 Číselné charakteristiky průměrné koncentrace látek Benzo(a)pyren [$ng.\ m^{-3}$] a prachových částic [$\mu g.\ m^{-3}$] na 1 km^2 pro zkoumané látky na území České republiky za období 2014-2018. 53

3.1 Číselná charakterizace rozložení sousedství, určené pomocí metody *sousedství královny*. 58
3.2 *Globální Moranovo I kritérium* a jeho statistická významnost. . . 59
3.3 Číselná charakterizace*globálního modelu* pro *prachové částice* menší než 10 μm v závislosti na *benzo[a]pyrenu*. 64
3.4 Číselná charakterizace uvedeného *globálního* modelu. 65
3.5 Číselná charakterizace *prostorově vážené regrese* uvedeného *globálního* modelu. 66

Seznam obrázků

1.1 Prostorové rozložení průměrné výšky mužské populace. 9
1.2 Možné výsledky autokorelační funkce. 11
1.3 Interpretace prostorových zpoždění. 13
1.4 Grafické znázornění možnosti prostorové autokorelace. 15
1.5 Moranův diagram. 18
1.6 Princip metody *sousedství věže*. 21
1.7 Princip metoda *sousedství královny*. 22
1.8 Princip *k-nejbližších sousedů*, pro $k = 4$. 23
1.9 Princip *prahové vzdálenosti*. 23
1.10 *Mocninné prostorové váhy* s rozdílnými ϕ. 26
1.11 *Exponenciální prostorové váhy* s rozdílnými α. 27
1.12 *Dvojitě umocněná prostorové váhy* s rozdílnými α. 28
1.13 Lokální regresní koeficienty vlivu podílu osob s vysokoškolským vzděláním na míru nezaměstnanosti v obcích ČR 3/2011 [11]. . . 34
1.14 Lokální regresní koeficienty vliv podílu osob s nízkým vzděláním na míru nezaměstnanosti v obcích ČR, 3/2011 [11]. 35
1.15 Interpretace prostorového jádra. 38
1.16 Interpretace modelu s dichotomickou proměnnou. 42
1.17 Interpretace modelu s interakcí. 43

2.1 Částečná vizualizace shapefile souboru pro jednotlivé POÚ. 47
2.2 Průměrný počet úmrtí na *infarkt myokardu* na úrovní jednotlivých krajů. 51
2.3 Průměrný počet úmrtí na *rakovinné onemocnění* na úrovní jednotlivých krajů. 52
2.4 Průměrné množství *benzo[a]pyren* $[\mu g.m^{-3}]$ pro jednotlivé kraje České republiky za období 2014-2018. 54
2.5 Průměrné množství prachových částic menší jak 10 μm $[\mu g.m^{-3}]$ pro jednotlivé kraje České republiky za období 2014-2018. 55
2.6 Průměrné množství prachových částic menší jak 2,5 μm $[\mu g.m^{-3}]$ pro jednotlivé kraje České republiky za období 2014-2018. 55
2.7 Klimatické podmínky v jednotlivých krajích České republiky. . . . 56

3.1 Moranův diagram pro *infarkt myokardu*. 60
3.2 Prostorové shluky *infarktu myokardu*. 61
3.3 Moranův diagram pro *rakovinné onemocnění tlustého střeva a konečníku*. 62
3.4 Prostorové shluky *rakovinného onemocnění tlustého střeva a konečníku*. 63
3.5 Prostorová nestacionarita *globálního* modelu, vyjádřena pomocí hodnot interakce regresních koeficientů pro teplém klima, v kombinaci s *benzo[a]pyrenu*. 67
3.6 Prostorová nestacionarita *globálního* modelu, vyjádřena pomocí hodnot umělé proměnné regresních koeficientů pro teplém klima, v kombinaci
s *benzo[a]pyrenu*. 68
3.7 Prostorová stacionarita interakce, pro velmi teplém klima v kombinaci s *benzo[a]pyrenu*. 69
3.8 Prostorová stacionarita umělé proměnné, pro velmi teplém klima v kombinaci s *benzo[a]pyrenu*. 69