UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Hot spot analýza

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **Mgr. Kamila Fačevicová, Ph.D.**
Vypracoval(a): **Bc. Aneta Dvořáková**
Studijní program: N0541A170026 Aplikovaná matematika
Studijní obor: Aplikovaná matematika
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Aneta Dvořáková

Název práce: Hot spot analýza

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Kamila Fačevicová, Ph.D.

Rok obhajoby práce: 2024

Abstrakt: Cílem práce je popsat metody využívané při Hot spot analýze dat s jejich následnou aplikací na reálných datech. Hlavními metodami, kterými se práce zabývá, jsou prostorová autokorelace, kvadrátová analýza, metoda průměrné nejbližší vzdálenosti a prostorově vážená regrese (GWR). Každá z těchto metod je následně, s použitím softwaru `R`, aplikována na datech týkající se nehodovosti města Brna.

Klíčová slova: hot spot, prostorová autokorelace, kvadrátová analýza, metoda průměrné nejbližší vzdálenosti, GWR, Moranův index, Monte Carlo, G-statistika

Počet stran: 106

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Aneta Dvořáková

Title: Hot Spot Analysis

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Kamila Fačevicová, Ph.D.

The year of presentation: 2024

Abstract: The aim of this thesis is to describe methods used in Hot Spot Analysis with their application on real data. The main methods used in the thesis are Spatial Autocorrelation, Quadrat Analysis, Average Nearest Neighbor and Geographically Weighted Regression (GWR). Each of these methods is then applied, using the software R, on accident data of the city of Brno.

Key words: Hot Spot, Spatial Autocorrelation, Quadrat Analysis, Average Nearest Neighbor, GWR, Moran's I, Monte Carlo, Getis-Ord

Number of pages: 106

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením paní Mgr. Kamily Fačevicové, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych na tomto místě poděkovala paní Mgr. Kamile Fačevicové, Ph.D. za odborné vedení diplomové práce, cenné rady, trpělivost a spoustu času, který mi při psaní této práce věnovala.

Úvod

Hot spot analýza je způsob zpracování dat, uplatňovaný převážně v geoinformatické praxi, k nalezení *hot spotů*, které lze chápat především jako oblasti s vysokou koncentrací zkoumaných událostí, jež jsou obklopeny dalšími oblastmi stejného charakteru. Cílem této práce je tedy popsat metody, jejichž použití povede k nalezení těchto *hot spotů* a tyto teoretické znalosti uplatnit na zvolených reálných datech.

Teoretická část práce je rozdělena do dvou hlavních kapitol. První kapitola slouží k zadefinování základních pojmů potřebných pro práci s prostorovými daty, což jsou taková data, ke kterým se váže informace o geografické poloze na zemském povrchu. Druhá kapitola je věnována metodám prostorové analýzy, s jejichž využitím je možné nalézt *hot spoty*. Volba metod pro identifikaci těchto *hot spotů* závisí na typu dat, která máme k dispozici. V této práci lze nalézt metody zaměřené na bodová data - obsahující informaci pouze o lokalizaci událostí (např. místo, ve kterém došlo k nehodě), atributová bodová data - obsahující navíc informaci o statistickém znaku (např. měsíc, ve kterém došlo k nehodě), nebo na polygonová data - reprezentující územní jednotky (např. státy, kraje, městské části), do kterých jsou agregována bodová data.

Praktická část práce, obsažená ve třetí kapitole, je zaměřená na demonstraci metod prostorové analýzy uvedených v teoretické části na reálných datech. Data, která byla pro analýzu použita, se týkají nehodovosti na území města Brna, přičemž tato data, která jsou volně dostupná veřejnosti, byla sbírána od roku 2010 Policií České republiky.

Kapitola 1

Terminologie

K sepsání této kapitoly bylo čerpáno především z [22]. Zaměříme se zde na zadefinování základních pojmů prostorové statistiky, které tvoří základ při práci s prostorovými daty. Prostorová statistika pracuje s daty, které oproti těm neprostorovým obsahují navíc informaci o geografické poloze na zemském povrchu (tj. zeměpisnou šířku a délku). V práci se setkáme se dvěma typy prostorových dat – bodovými a polygonovými.

1.1. Body

Bodová data dělíme na základě dostupných informací do dvou skupin:

1. Bodová data obsahující pouze informaci o lokalizaci událostí (např. nehodovost, kriminalita, výskyt zemětřesení), případně objektů (např. nemocnice nebo policejní stanice), v podobě uspořádaných dvojic souřadnic $(a_i, b_i) \in \mathbb{R}^2$, pro $i = 1, \ldots, n_b$, kde $n_b \in \mathbb{N}$ vyjadřuje celkový počet bodů (resp. pozorování).

2. Bodová data obsahující informaci z 1. odrážky spolu s atributovými informacemi (tj. informacemi o statistických znacích) zaznamenanými

pro každý bod (a_i, b_i), $i = 1, \ldots, n_b$, v datové matici $\mathbf{X} \in \mathbb{R}^{n_b \times m}$, kde $m \in \mathbb{N}$ značí počet sledovaných atributů. Pod těmito atributy si, v kontextu této práce, lze představit například den, ve kterém došlo na daných souřadnicích (a_i, b_i) k nehodě, věkovou skupinu osob zúčastněných této nehody, druh komunikace, na které k této nehodě došlo, nebo také příčinu zavinění nehody. Dokud však nebude napsáno jinak, budeme pracovat pouze s jedním, i-tým, atributem $\mathbf{X}_i = (X_{1i}, \ldots, X_{n_b i})^T$, $i = 1, \ldots, m$, který pro jednoduchost budeme v průběhu práce značit jako $X = (X_1, \ldots, X_{n_b})$.

1.2. Polygony

Polygonem P, dle [36], obecně rozumíme uzavřený útvar v $\mathbb{R}^2$ složený z konečného počtu úseček, které jsou alespoň tři, a které se setkávají ve svých koncových bodech. Úsečkami rozumíme strany polygonu a body, ve kterém se úsečky setkávají, nazýváme vrcholy. Matematicky polygon zapisujeme jako množinu uspořádaných dvojic bodů, tedy $P = \{(a_1, b_1), \ldots, (a_v, b_v)\}$, kde $v \in \mathbb{N}$, $v \geq 3$ je počet vrcholů. V situaci, kdy máme k dispozici $n_p \in \mathbb{N}$ vzájemně se nepřekrývajících polygonů, bude i-tý polygon s počtem vrcholů v_i množina $P^i = \{(a_1^i, b_1^i), \ldots, (a_{v_i}^i, b_{v_i}^i)\}$, $i = 1, \ldots, n_p$. V prostorové analýze se polygony používají k reprezentaci územních jednotek - správních oblastí (kraje, okresy, správní obvody), států, pozemků, lesů aj., do kterých jsou často agregována bodová data. Touto agregací získáváme atributové informace, kde pro p polygonů je i-tý atribut definován stejným způsobem, jako tomu bylo v případě bodů, tj. $\mathbf{X}_i = (X_{1i}, \ldots, X_{n_p i})^T$, $i = 1, \ldots, m$, přičemž budeme zatím opět pracovat pouze s jedním atributem $X = (X_1, \ldots, X_{n_p})$.

1.2.1. Centroid polygonu

Při práci s prostorovými daty se často setkáme s pojmem *centroid polygonu* [6]. Obecně je centroid polygonu definován jako geografický střed polygonu P, kterým je myšleno jeho předpokládané těžiště. Matematicky je centroid polygonu bod $C = (C_a, C_b)$, jehož souřadnice jsou definovány následujícím způsobem

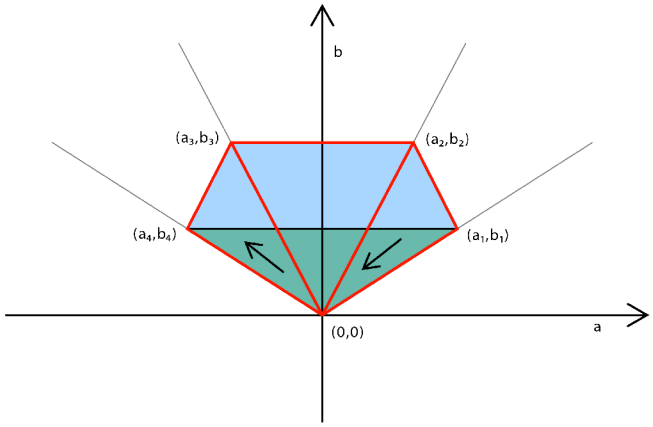
$$C_a = \frac{1}{6A} \sum_{i=1}^{v} ((a_i + a_{i+1})(a_i b_{i+1} - a_{i+1} b_i)), \tag{1.1}$$

$$C_b = \frac{1}{6A} \sum_{i=1}^{v} ((b_i + b_{i+1})(a_i b_{i+1} - a_{i+1} b_i)), \tag{1.2}$$

kde v je počet vrcholů $(a_1, b_1), (a_2, b_2), \ldots, (a_v, b_v)$ polygonu P, jejichž souřadnice jsou brány proti směru hodinových ručiček, přičemž budeme uvažovat, že $(a_{v+1}, b_{v+1}) = (a_1, b_1)$ (z toho důvodu, že polygon je uzavřený útvar) a A je plocha polygonu P, která je definována pomocí této *Shoelace formule*

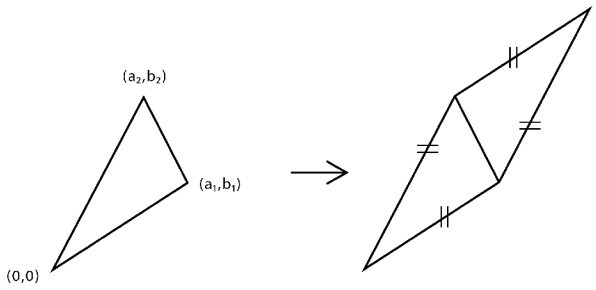
$$A = \frac{1}{2} \left| \sum_{i=1}^{v} (a_i b_{i+1} - a_{i+1} b_i) \right|. \tag{1.3}$$

Shoelace formule [26] funguje tak, že pro výpočet plochy polygonu využívá střed (0,0) souřadnicové osy. Ten poslouží jako pomocný vrchol, s jehož pomocí nejprve rozdělíme daný polygon, se souřadnicemi vrcholů číslovanými proti směru hodinových ručiček, na trojúhelníky tak, jak je zobrazeno na obrázku 1.1.



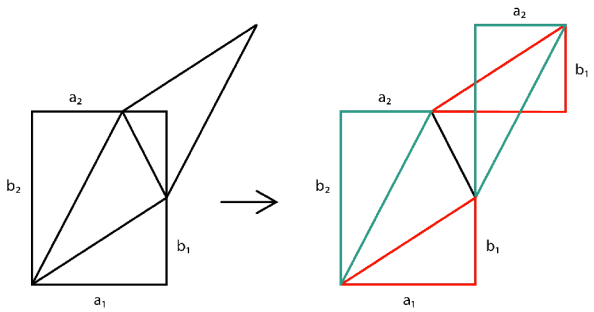
Obrázek 1.1: Rozdělení zvoleného polygonu na trojúhelníky (červeně) s využitím vrcholu (0,0)

U každého z těchto trojúhelníků spočítáme "zprostředkovaně" jeho obsah. Zprostředkovaně proto, protože z jednotlivých trojúhelníků nejprve vytvoříme rovnoběžníky způsobem znázorněným na následujícím obrázku 1.2.



Obrázek 1.2: Ukázka tvorby rovnoběžníku z vybraného trojúhelníku z obrázku 1.1

Obsah rovnoběžníků je počítán mechanismem *Solomona Golomba*, schematicky znázorněným na obrázku 1.3,



Obrázek 1.3: Princip výpočtu plochy rovnoběžníku

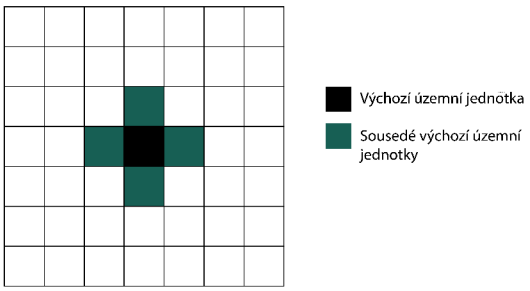
odkud je patrné, že obsah rovnoběžníku je roven rozdílu obsahů obdélníků, tedy $a_1b_2 - a_2b_1$. K získání obsahu trojúhelníku stačí vzít polovinu z tohoto rozdílu, tj. $\frac{1}{2}(a_1b_2 - a_2b_1)$. Takto spočítáme obsah každého trojúhelníku s tím, že (v tomto konkrétním případě) v posledním kroce postupujeme při výpočtu ve směru hodinových ručiček (znázorněno šipkami na obrázku 1.1), což se projeví tím, že tento obsah trojúhelníku bude záporný. Sečtením všech obsahů trojúhelníků získáme výslednou plochu polygonu, jejíž výpočet je obecně zapsán vztahem (1.3) (kde je navíc přidaná absolutní hodnota pro případ, že bychom na počátku číslovali ve směru hodinových ručiček, a získali tak záporný obsah). Souřadnice centroidu polygonu jsou následně pouze váženým průměrem centroidů trojúhelníků, kde váhy odpovídají obsahům těchto trojúhelníků v poměru k celkové ploše polygonu. K výpočtu souřadnic centroidu jednoho trojúhelníku slouží jednoduchý aritmetický průměr příslušných souřadnic vrcholů, tedy $C = (\frac{a_i+a_{i+1}+a_{i+2}}{3}, \frac{b_i+b_{i+1}+b_{i+2}}{3})$, přičemž vzhledem k tomu, že třetí vrchol má vždy souřadnice (0,0), je ve vztazích (1.1) a (1.2) uveden pouze součet $(a_i + a_{i+1})$ a $(b_i + b_{i+1})$.

1.2.2. Sousednost

Jednou z prvních věcí, kterou je potřeba při analýze vztahů mezi prostorovými polygonálními daty zjistit je, které polygony spolu sousedí. To, jakým způsobem budeme sousednost polygonů chápat, závisí na definici sousednosti, kterou použijeme. K dispozici máme definice "*rook's case*" (případ věže) a "*queen's case*" (případ královny).

Rook's case

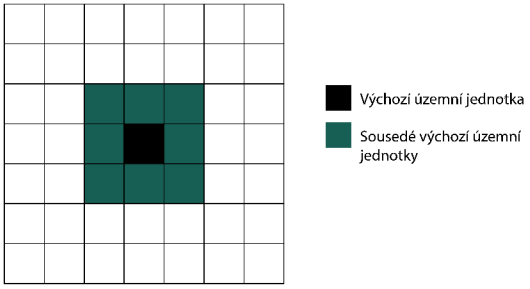
V tomto případě, jak můžeme vidět na obrázku 1.4, považujeme územní jednotky (tj. polygony) za sousedy nějaké výchozí územní jednotky, pokud spolu sdílejí hranici o délce větší než 0. To znamená, že pokud by se výchozí územní jednotka a nějaká jiná územní jednotka dotýkaly pouze v jednom jediném bodě, nelze je považovat za sousední.



Obrázek 1.4: Znázornění sousednosti typu *"rook's case"*

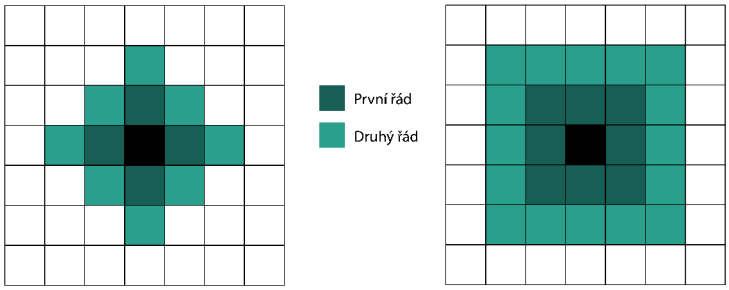
Queen's case

Oproti výše zmíněnému případu zde za sousedy výchozí územní jednotky považujeme všechny územní jednotky, které se s touto výchozí územní jednotkou dotýkají alespoň v jednom bodě. Tento případ je znázorněn na následujícím obrázku 1.5.



Obrázek 1.5: Znázornění sousednosti typu *"queen's case"*

Sousední jednotky, které jsou definované předchozími dvěma způsoby, se označují jako bezprostřední sousedé nebo sousedé 1. řádu. Dále můžeme hovořit i o sousedech 2. řádu, jejichž znázornění je možné vidět na obrázku 1.6. Ty jsou definovány jako sousedé sousedů výchozí územní jednotky. Takto by se teoreticky dalo pokračovat i pro vyšší řády, v praxi ale vyšší řády nejsou moc využitelné.



Obrázek 1.6: Znázornění sousednosti 2. řádu pro sousednost *"rook's case"* (vlevo) a *"queen's case"* (vpravo)

1.3. Vzdálenost

Na to, zda jsou polygony sousedé, se dá, kromě definice sousednosti, pohlížet čistě z hlediska vzdálenosti [2]. Vzdálenost definujeme prostřednictvím metrik, které zde budou postupně představeny, přičemž tento koncept určení sousednosti je možné použít nejen pro polygonální data, ale i pro bodová data.

1.3.1. Euklidovská metrika

Prvním typem metriky, prostřednictvím které definujeme vzdálenost, je Euklidovská metrika (přímková míra vzdálenosti). V případě polygonových dat je vzdálenost počítána mezi centroidy polygonů, tedy

$$d_{ij} = \sqrt{(C_{a_i} - C_{a_j})^2 + (C_{b_i} - C_{b_j})^2} \tag{1.4}$$

pro i-tý a j-tý centroid polygonu s příslušnými souřadnicemi (C_{a_i}, C_{b_i}) a (C_{a_j}, C_{b_j}). Pro bodová data by se výpočet Euklidovské metriky provedl analogicky.

1.3.2. Manhattanská metrika

Alternativou, která je někdy upřednostňována před Euklidovskou metrikou, je Manhattanská metrika, jelikož snižuje vliv odlehlých hodnot. Vzdálenost mezi centroidy polygonů je v tomto případě počítána takto

$$d_{ij}^m = |C_{a_i} - C_{a_j}| + |C_{b_i} - C_{b_j}| \tag{1.5}$$

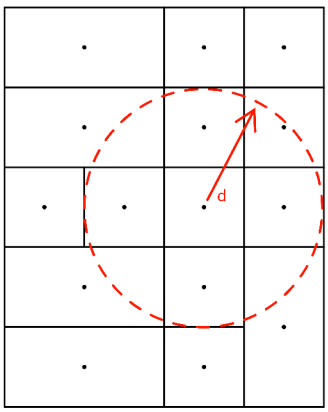
pro i-tý a j-tý centroid polygonu s příslušnými souřadnicemi (C_{a_i}, C_{b_i}) a (C_{a_j}, C_{b_j}). Stejně jako v předchozím případě by se výpočet metriky pro bo-

dová data provedl analogicky.

S takto zavedenými vzdálenostmi můžeme přistoupit k metodám, které nám pomohou k rozhodnutí, zda jsou polygony nebo body sousední. Metody, které se v prostorové analýze používají, jsou *Metoda k nejbližších sousedů* nebo *Metoda prahové vzdálenosti*.

Metoda prahové vzdálenosti

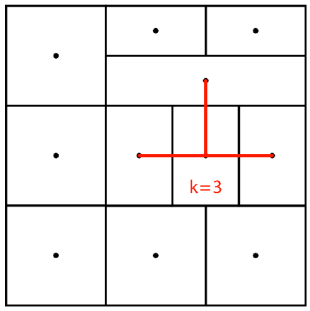
Při použití této metody, jejíž znázornění lze vidět na obrázku 1.7, jsou za sousedy i-tého polygonu považovány takové polygony, které mají od tohoto i-tého polygonu vzdálenost nejvýše d. Matematicky jde tedy o množinu polygonů $\{j \in \{1, \ldots, n_p\} : d_{ij} \leq d\}$. V případě bodových dat bychom postupovali analogicky. Zároveň by volba d měla být taková, aby měl každý polygon (resp. bod) alespoň jednoho souseda.



Obrázek 1.7: Znázornění metody prahové vzdálenosti

Metoda k nejbližších sousedů

V této metodě zvolíme za sousedy i-tého polygonu prvních $k > 0$ polygonů, které mají od tohoto i-tého polygonu nejmenší vzdálenost. Ke znázornění, pro případ $k = 3$ sousedů, je zde přiložen obrázek 1.8. Matematicky tedy za sousedy i-tého polygonu považujeme množinu polygonů $\{(j_1, \ldots, j_k) \in \{1, \ldots, n_p\} : d_{ij} < d_{ik}, \forall j \in (j_1, \ldots, j_k), \forall k \notin (j_1, \ldots, j_k)\}$. Stejně jako v předchozím případě, i zde by se u bodových dat postupovalo analogicky. Výhodou oproti předchozímu přístupu je to, že zde má každý polygon (resp. bod) stejný počet sousedů (nemůže se tedy stát, že by někdo souseda neměl). Nevýhodou při pevném počtu sousedů však je, že pokud budou polygony (resp. body) v nějakém místě nahuštěné, a v jiném daleko od sebe, tak v jedné oblasti budeme analyzovat velmi malou část z celkového zkoumaného území, a ve druhé zase mnohem větší část z celkového území.



Obrázek 1.8: Znázornění metody k nejbližších sousedů

1.4. Matice vah

Po rozhodnutí, s jakou definicí sousednosti, případně vzdálenosti, budeme pracovat, lze přistoupit k tvorbě matice vah $\mathbf{W} \in \mathbb{R}^{n \times n}$ tvaru

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

pro n pozorování, kdy je možné uvažovat buď $n = n_b$ nebo $n = n_p$ v závislosti na tom, zda pracujeme s bodovými nebo polygonovými daty (proto dále nebudeme rozlišovat mezi n_b a n_p, a budeme, pro jednoduchost, počet pozorování souhrnně označovat n).

Matice vah zachycuje prostorové vztahy mezi územními jednotkami. Každá územní jednotka je reprezentována řádkem a sloupcem matice, přičemž každá hodnota matice vyjadřuje prostorový vztah mezi dvěma příslušnými geografickými prvky. Obecně i-tý řádek matice $\mathbf{W}$ vyjadřuje, jak i-tá územní jednotka prostorově souvisí se všemi ostatními jednotkami. Počet nenulových hodnot v i-tém řádku potom vyjadřuje počet územních jednotek, se kterými i-tá územní jednotka sousedí. Matematicky budeme řádkový součet značit tímto způsobem

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

dále sloupcový součet takto

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

a celkový součet přes všechny řádky a sloupce matice budeme označovat následovně

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}. \tag{1.6}$$

Vzhledem k různým kritériím, prostřednictvím kterých definujeme vztahy sousedních územních jednotek, můžeme následně vytvořit různé matice vah.

1.4.1. Binární matice

Jedním z nejpoužívanějších typů váhových matic je binární matice. Při sestavování matice využijeme předpokladu, že územní jednotky, které spolu sdílí hranici, jsou sousední (bez ohledu na to, zda pro definici sousednosti použijeme *případ věže* nebo *případ královny*). Pokud spolu dvě územní jednotky sousedí, přiřadíme jim hodnotu 1. V opačném případě jim přiřadíme hodnotu 0. Vzhledem k tomu, že matice vyjadřuje jakým způsobem jsou dvojice územních jednotek propojeny, bývá někdy označována jako "*binární matice konektivity*".

Vlastnosti matice

Je-li matice $\mathbf{W}$ binární, potom její prvky w_{ij} obsahují hodnoty 0 nebo 1, přičemž indexy i a j odkazují na i-tou a j-tou územní jednotku. Matici lze charakterizovat pomocí následujících bodů:

- Prvky na hlavní diagonále matice jsou nulové, tj. $w_{ii} = 0, \forall i \in \{1, \ldots, n\}$ (předpokládáme, že územní jednotka nesousedí sama se sebou).
- Je symetrická, tj. $w_{ij} = w_{ji}$, $\forall i, j \in \{1, \ldots, n\}$.

Pokud bychom místo definice sousednosti z podkapitoly 1.2.2 při sestavování binární matice využili *metodu prahové vzdálenosti*, nebo *metodu k nejbližších sousedů* s využitím libovolné metriky z podkapitoly 1.3, prvky matice sestavíme následujícím způsobem:

1. Pro metodu prahové vzdálenosti:

$$w_{ij} = \begin{cases} 1 & \text{pro } d_{ij} \leq d,\ i \neq j \\ 0 & \text{pro } d_{ij} > d,\ i \neq j \\ 0 & \text{pro } \ \ i = j, \end{cases}$$

2. Pro metodu k nejbližších sousedů:

$$w_{ij} = \begin{cases} 1 & \text{pro } d_{ij} \leq d_{ij}^{(k)},\ i \neq j \\ 0 & \text{pro } d_{ij} > d_{ij}^{(k)},\ i \neq j \\ 0 & \text{pro } \ \ \ i = j, \end{cases}$$

pro $k \in \{1, \ldots, n-1\}$, kde $d_{ij}^{(k)}$ značí vzdálenost mezi jednotkami i a j, kde $i, j \in \{1, \ldots, n\}$, $i \neq j$, na k-té pozici při vzestupném seřazení vzdáleností $d_{ij}^{(1)} \leq, \ldots, \leq d_{ij}^{(n-1)}$ (pozn. kvůli množství dolních indexů je pořadí zaznačeno pomocí horního indexu).

Nevýhodou této matice je její neefektivní způsob zachycení prostorového vztahu. Touto neefektivitou je myšlena duplikace informací (horní a dolní trojúhelník matice obsahuje tutéž informaci o sousednosti) a vysoký počet nul v matici, čímž zabíráme mnoho místa informací o nesousedících územních jednotkách. Z tohoto důvodu se někdy (čistě pro informativní účely, abychom viděli, s jakými územními jednotkami, která jednotka sousedí) využívá zápis, ve kterém máme v řádcích pod sebou vypsány jednotlivé územní jednotky a ve sloupcích sousedy dané územní jednotky tak, jak vidíme v tabulce 1.1.

Katastr/Soused	Soused 1	Soused 2	Soused 3	Soused 4
Bosonohy	Jundrov	Kohoutovice	Nový Lískovec	
Líšeň	Slatina			
Bystrc	Kníničky	Žebětín		
Chrlice	Tuřany	Holásky	Brněnské Ivanovice	
Dvorska	Tuřany			

Tabulka 1.1: Infomace o sousedních územních jednotkách vybraných katastrů města Brna při volbě prahové vzdálenosti $d = 0.04$

1.4.2. Řádkově standardizovaná matice

Další často používanou váhovou maticí **W** je řádkově standardizovaná matice. Matice je tvořena stejným způsobem jako binární matice s tím rozdílem, že jednotlivé prvky matice v daném řádku jsou podělené odpovídajícím řádkovým součtem $w_{i.}$. Po této úpravě tedy dostaneme pro každou sousední jednotku následující váhu

$$w_{ij} = \frac{w_{ij_{puv}}}{w_{i.}},$$

kde $w_{ij_{puv}}$ zde značí původní váhu matice (tj. váhu před vydělením) a w_{ij} značí novou váhu po provedené úpravě.

Vlastnosti matice

Váhová matice **W**, která je řádkově standardizovaná, má tyto vlastnosti:

- Pro prvky matice platí, že $w_{ij} \in \langle 0, 1 \rangle$, $\forall i, j \in \{1, \ldots, n\}$.
- Prvky na hlavní diagonále matice jsou nulové, tj. $w_{ii} = 0, \forall i \in \{1, \ldots, n\}$.
- Není symetrická, tj. $w_{ij} \neq w_{ji}$ pro alespoň jednu dvojici $i \neq j$.

Rozdíl oproti předchozí, binární, matici je tedy především ten, že řádkově standardizovaná matice obecně není symetrická.

1.4.3. Matice vzdáleností centroidů (resp. bodů)

K zachycení prostorového vztahu je možné použít také matici vzdáleností centroidů, kdy se na tyto vzdálenosti díváme jako na váhy. Prvky w_{ij} zde tedy představují vzdálenost centroidů územních jednotek i a j, přičemž pro výpočet vzdálenosti máme na výběr mezi vztahem (1.4) pro Euklidovskou metriku a (1.5) pro Manhattanskou metriku. Pro zohlednění informace o sousednosti územních jednotek v matici vzdáleností centroidů využijeme metod z podkapitoly 1.3, kdy váhy budou konstruovány jednou z následujících možností

1. Pro metodu prahové vzdálenosti:

$$w_{ij} = \begin{cases} d_{ij} & \text{pro } d_{ij} \leq d \\ 0 & \text{pro } d_{ij} > d, \end{cases}$$

2. Pro metodu k nejbližších sousedů:

$$w_{ij} = \begin{cases} d_{ij} & \text{pro } d_{ij} \leq d_{ij}^{(k)},\ i \neq j \\ 0 & \text{pro } d_{ij} > d_{ij}^{(k)},\ i \neq j \\ 0 & \text{pro } \quad i = j. \end{cases}$$

V praxi se však často používá převrácená hodnota váhy w_{ij}, jelikož zachycuje, jak se při zvyšující se vzdálenosti síla prostorového vztahu zmenšuje. Předchozí váhy (ať už při použití metody prahové vzdálenosti nebo k nejbližších sousedů) proto budou v tomto případě upraveny prostřednictvím následujícího vztahu

$$w_{ij} = \frac{1}{d_{ij}}.$$

Vlastnosti matice

Matice vzdáleností centroidů **W** má tyto vlastnosti:

- Pro prvky matice platí, že $w_{ij} \in \langle 0, +\infty)$, $\forall i, j \in \{1, \ldots, n\}$.
- Prvky na hlavní diagonále matice jsou nulové, tj. $w_{ii} = 0, \forall i \in \{1, \ldots, n\}$.
- Je symetrická, tj. $w_{ij} = w_{ji}$, $\forall i, j \in \{1, \ldots, n\}$.

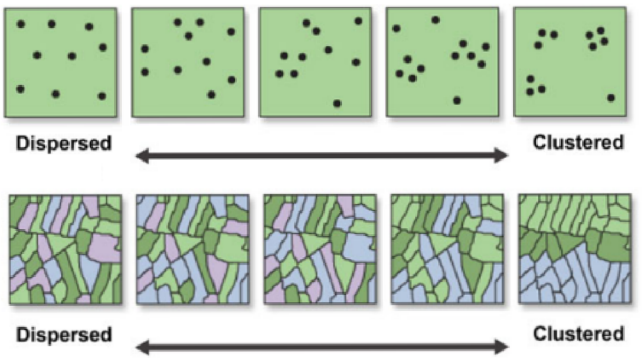
Nevýhodou využití tohoto typu matice vah však je, že tvar polygonu ovlivňuje výslednou lokaci centroidu. Polygony s neobvyklým (nekonvexním) tvarem potom mohou mít centroidy umístěné například v jiném polygonu.

Kapitola 2

Metody prostorové analýzy

2.1. Prostorová autokorelace

Obecně při popisu a analýze prostorových vzorů tvořených hodnotami zkoumaného atributu u atributových dat, resp. uspořádáním neatributových bodových dat, využíváme prostorové statistiky. Prostorové vzory (obrázek 2.1) dělíme do tří kategorií - shluklé (Clustered), rozptýlené (Dispersed) a náhodné (na obrázku uprostřed).

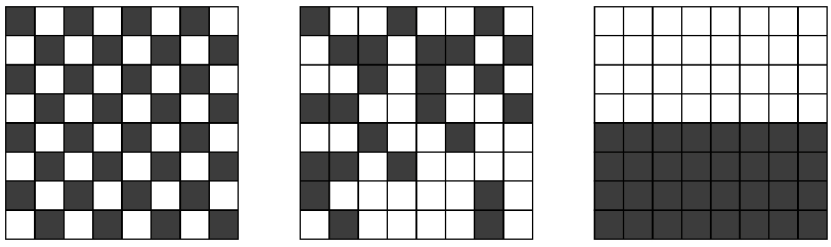


Obrázek 2.1: Typy prostorových vzorů, zdroj: [12]

Při pouhém pohledu na vzor našich dat nejsme obvykle schopni jedno-

značně rozhodnout, do jaké kategorie vzor spadá. Důvodem je, že v praxi vzor nikdy nebude extrémně shluklý/rozptýlený nebo náhodný. Zajímá nás proto, jak blízko je prostorový vzor, který máme k dispozici, k jednotlivým kategoriím vzorů. V tomto ohledu hraje u atributových dat, kterými se budeme zabývat nejdříve, významnou roli prostorová autokorelace.

Prostorová autokorelace vyjadřuje (samo)korelaci hodnot atributu X, která je způsobená prostorovým uspořádáním těchto hodnot ve zkoumaném území. Ptáme se tedy, jak moc atribut v jednom místě koreluje s hodnotami téhož atributu v jeho okolí. Na základě hodnoty prostorové autokorelace můžeme rozhodnout, do jaké kategorie prostorových vzorů, z hlediska hodnot zkoumaného atributu, spadá náš zkoumaný vzor. Je-li prostorová autokorelace pozitivní, potom je náš prostorový vzor shluklý (ve zkoumaném území se vyskytují oblasti podobných hodnot - ať už nízkých nebo vysokých). V případě negativní prostorové autokorelace jsou zkoumané hodnoty rovnoměrně rozloženy v prostoru (tzv. "rozptýlený" vzor, kdy nízké hodnoty atributu jsou obklopeny vysokými nebo naopak). Pokud prostorová autokorelace neexistuje (tj. pokud je nulová), jedná se o náhodný vzor. Vzory při daném typu autokorelace je možné, pro lepší pochopení, vidět na obrázku 2.2.



Obrázek 2.2: Prostorové vzory při daném typu autokorelace, inspirace: [37]

K výpočtu prostorové autokorelace využíváme statistiky, které nám umožňují pracovat s atributovými bodovými nebo polygonovými daty. Nejprve se podíváme na autokorelaci z globálního, a následně z lokálního hlediska. Zároveň je potřeba zmínit, že při psaní těchto podkapitol (včetně úvodu kapitoly 2.1) bylo čerpáno především z [22] a [25].

2.1.1. Globální statistiky prostorové autokorelace (asociace)

Globální statistiky lze považovat za globální míry prostorové autokorelace. Výsledkem těchto statistik je jedna hodnota, která popisuje prostorový vzor hodnot zkoumaného atributu v rámci celé zkoumané oblasti.

Globální Moranův index

Globální Moranův index je definován následujícím vztahem

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{S \sum_{i=1}^{n}(X_i - \overline{X})^2}, \tag{2.1}$$

kde n je počet pozorování, S je součet všech prvků váhové matice definovaný vztahem (1.6), hodnoty $(X_1, \ldots, X_n)$ jsou hodnoty stanoveného atributu X a

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

vyjadřuje celkový (globální) průměr hodnot atributu X ve zkoumaném území.

Váhy v Moranově indexu hrají důležitou roli při určení příspěvku každé dvojice lokalit k celkovému I. Proto pokud bychom v datech měli dvojice $X_i \gg \overline{X}$ a $X_j \ll \overline{X}$, tak se může stát, že tyto dvojice budou v hodně vzdále-

ných lokalitách, a ačkoliv je jmenovatel Moranova indexu vždy kladný, váha w_{ij} bude malá, a tím pádem příspěvek k celkovému I bude taktéž malý. Z čitatele je patrné, že pokud by hodnoty atributu blízkých územních jednotek i a j byly vyšší než globální průměr, výsledná hodnota čitatele pro tyto dvě konkrétní jednotky je vysoká kladná hodnota. Stejně tak je tomu v případě, kdy jsou jejich hodnoty nižší než globální průměr. Tyto situace vypovídají o pozitivní prostorové autokorelaci. Pokud by ale hodnota zkoumaného atributu územní jednotky i byla vyšší než globální průměr, a její blízké územní jednotky j nižší, v čitateli by byla záporná hodnota. V tomto případě by výsledek vypovídal o negativní prostorové autokorelaci. Proto pokud se v celé studované oblasti vyskytují blízko sebe častěji podobné hodnoty než rozdílné (ve smyslu porovnání s globálním průměrem), Moranův index má tendenci být kladný a naopak. Při použití metody Moranova indexu tedy porovnáváme celkovou odlišnost pozorované hodnoty v oblasti i od globálního průměru v kombinaci s odlišností hodnot v okolí oblasti i od globálního průměru (čitatel), a celkovou odlišnost pozorovaných hodnot od průměrné globální hodnoty (jmenovatel).

Moranův index je, dle článku [9], zobecněním Pearsonova dvouvýběrového korelačního koeficientu, definovaného vztahem

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}}},$$

pro výběry $(X_1, \ldots, X_n)$ a $(Y_1, \ldots, Y_n)$, na jednovýběrový autokorelační koeficient a následném zobecnění jednorozměrného výběrového autokorelačního koeficientu z časových řad, definovaného následujícím vztahem (pro případ autokorelace řádu 1, tj. sledujeme, jak spolu souvisí hodnoty náhodné veličiny Y v čase t a v čase $t-1$)

$$\rho = \frac{\sum_{t=2}^{n}(Y_t - \overline{Y})(Y_{t-1} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},$$

pro n hodnot pozorovaných v čase $t = 1, \ldots, n$, na dvourozměrný autokorelační koeficient prostorové distribuce. Spojitost mezi tímto korelačním koeficientem a Moranovým indexem spočívá v *operátoru posunu* [10], který se v časových řadách pro jeden časový posun označuje jako B, a definuje se rovností $BY_t = Y_{t-1}$. Tento operátor vyjadřuje, jak hodnota náhodné veličiny v čase t závisí na hodnotě v čase $t - 1$. V prostorové statistice je ekvivalentem tohoto operátoru váhová matice $\mathbf{W}$. Prostřednictvím této matice vyjadřujeme, jak spolu souvisí hodnota X_i atributu X v lokalitě i s hodnotami v sousedních lokalitách j (jedná se tedy o sousednost prvního řádu, která je ekvivalentem časového posunu o jedno časové období). V tomto kontextu lze, dle [1], *operátor posunu* pro všechny lokality obecně zapisovat jako $\mathbf{W}X$. Pro vztah konkrétní lokality i s ostatními lokalitami j lze potom *operátor posunu* rozepsat buď takto

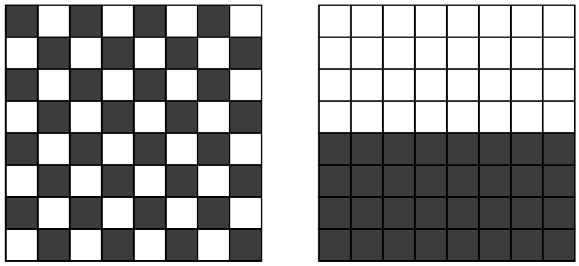
$$[\mathbf{W}X]_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{in}X_n,$$

nebo tímto způsobem

$$[\mathbf{W}X]_i = \sum_{j=1}^{n} w_{ij}X_j,$$

kde $[\mathbf{W}X]_i$ značí hodnotu $\mathbf{W}X$ v lokalitě i.

Hodnota Moranova indexu se pohybuje od -1 (negativní prostorová autokorelace, tj. rovnoměrné rozložení hodnot v prostoru) po 1 (pozitivní prostorová autokorelace, tj. výskyt shluků podobných hodnot v prostoru). Tyto vzory jsou při daných hodnotách Moranova indexu schematicky znázorněny na obrázku 2.3.



Obrázek 2.3: Prostorové vzory při daných hodnotách Moranova indexu, inspirace: [4]

Za platnosti nulové hypotézy o nepřítomnosti prostorové autokorelace (bude popsáno později) pochází hodnoty zkoumaného atributu ze stejného rozdělení a jsou nezávislé, přičemž I je asymptoticky normálně rozdělené se střední hodnotou

$$E[I] = -\frac{1}{(n-1)}, \tag{2.2}$$

kdy pro $n \to \infty$ lze přímo psát $E[I] = 0$.

Rozptyl Moranova indexu je vyjádřen tímto vztahem

$$Var[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}, \tag{2.3}$$

kde

$$S_0 = \sum_{i \neq j} w_{ij},$$

$$S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2,$$

$$S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik} \right)^2 .$$

Při dostatečně velkém množství zkoumaných oblastí/polygonů má I normální rozdělení a při porovnání testové statistiky zvané z-score, definované pomocí vztahu

$$Z_I = \frac{I - E[I]}{\sqrt{Var[I]}},$$

kde $E[I]$ je střední hodnota ze vztahu (2.2) a $Var[I]$ rozptyl ze vztahu (2.3), s kritickým oborem $W_c = (-\infty, -u_{1-\frac{\alpha}{2}}\rangle \cup \langle u_{1-\frac{\alpha}{2}}, \infty)$ normovaného normálního rozdělením při zvolené hladině α, lze rozhodnout, zda se pozorovaný vzor odchyluje od náhodného vzoru (tj. zda je rozdíl mezi pozorovaným a náhodným vzorem statisticky významný). Pokud $Z_I \in W_c$, potom rozdíl mezi pozorovaným a náhodným vzorem je statisticky významný (je zde přítomna prostorová autokorelace). V opačném případě mezi pozorovaným a náhodným vzorem není statisticky významný rozdíl (není zde přítomna prostorová autokorelace).

V praxi je předpoklad o normálně rozděleném I téměř nemožný, proto se přechází k alternativnímu přístupu ověření statistické významnosti Moranova indexu. Tento přístup využívá metodu Monte Carlo, která je založena na simulaci, namísto teoretického přístupu, který využívá předpoklady o rozdělení.

Monte Carlo

Metoda, dle zdroje [24], funguje tak, že nejprve spočítá globální Moranův index pro naše data. Následně náhodně[1] přerozdělí hodnoty zkoumaného atributu mezi lokalitami (tzn. pro každou lokalitu máme dán nějaký počet nehod, což jsou hodnoty, které v lokalitách přerozdělíme) a spočítá globální Moranův index pro tento permutovaný soubor dat. Proces tvorby permutovaného souboru dat a výpočtu Moranova indexu vzniklého souboru opakujeme několikrát (např. 99, nebo 999). Z opakovaného výpočtu Moranových indexů z permutovaných souborů dat získáme referenční rozdělení, které se použije pro výpočet p-hodnoty (a k porovnání statistického rozdílu mezi Moranovým indexem původních dat a vzniklého referenčního rozdělení). P-hodnoty se spočítají pomocí následujícího vztahu

$$p = \frac{R+1}{M+1}, \tag{2.4}$$

kde R je počet případů, kdy byl Moranův index I^* počítaný pro permutovaný datový soubor (ty, které jsme vytvořili náhodným přiřazením hodnot atributu k lokalitám) stejný nebo extrémnější než Moranův index I napočítaný na našich datech, tzn. počet případů, kdy $|I^*| \geq |I|$, a M je rovno počtu permutací (nejčastěji se volí 99, 999,...). Tato p-hodnota se potom použije k určení statistické významnosti Moranova indexu našich dat (resp. k (ne)zamítnutí nulové hypotézy zmíněné níže) při zvolené hladině významnosti α.

Při ověřování statistické významnosti hodnoty Moranova indexu testu-

[1] Vytvoříme všechny možnosti (permutace), jak uspořádat v celkové oblasti hodnoty zkoumaného atributu, tzn. v lokalitách se promíchají hodnoty atributu. Celkový počet permutací je $n!$ při souboru s n hodnotami zkoumaného atributu, a z těchto permutací náhodně vybereme dostatečně velké množství variant (tj. zvolíme počet opakování - v R parametrem *nsim*), které se použijí k tvorbě referenčního rozdělení.

jeme hypotézu

$$H_0 \colon I = 0,$$

která říká, že rozmístění hodnot atributu v prostoru představuje náhodný vzor (není zde tedy přítomna prostorová autokorelace), oproti alternativě

$$H_A \colon I \neq 0,$$

která naopak říká, že rozmístění hodnot atributu v prostoru odpovídá nenáhodnému vzoru (ať už jde o shluky nebo rovnoměrné rozložení hodnot v celkovém zkoumaném území).

Kromě výše zmíněné hypotézy je možné použít i její jednostranné verze. Jednou z jednostranných hypotéz je tato

$$H_0 \colon I \leq 0 \qquad H_A \colon I > 0, \tag{2.5}$$

kde v rámci H_0 testujeme, že zde prostorová autokorelace buď není přítomná, nebo je záporná (hodnoty jsou rovnoměrně rozložené v prostoru), oproti H_A, že prostorová autokorelace je kladná (v prostoru se vyskytují shluky podobných hodnot). Výpočet p-hodnoty v tomto případě upravíme následovně

$$p = \frac{R_v + 1}{M + 1},$$

kde místo R z původního výpočtu p-hodnoty (2.4) máme R_v, které představuje počet případů, kdy $I^* \geq I$.

Druhá jednostranná hypotéza, kterou je možno testovat, je formulována tímto způsobem

$$H_0 \colon I \geq 0 \qquad H_A \colon I < 0,$$

kde H_0 říká, že prostorová autokorelace je pozitivní, nebo není přítomná, a H_A naopak říká, že prostorová autokorelace je záporná. Zároveň je zde opět potřeba upravit vztah pro výpočet p-hodnoty, a to tímto způsobem

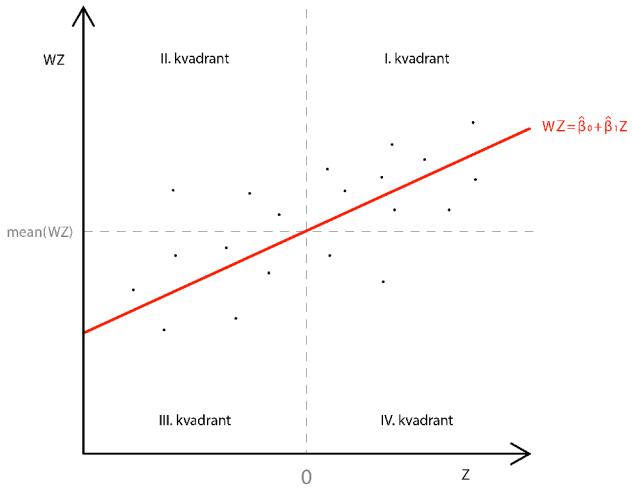
$$p = \frac{R_n + 1}{M + 1},$$

kde místo R z původního výpočtu p-hodnoty (2.4) máme R_n, které představuje počet případů, kdy $I^* \leq I$.

P-hodnota je tedy důležitým výstupem metody Monte Carlo, jelikož vysoká hodnota Moranova indexu sama o sobě nutně neurčuje, zda je na celém zkoumaném území přítomna prostorová autokorelace.

Moranův scatterplot

Součástí zkoumání Moranova indexu bývá i Moranův scatterplot k vizualizaci prostorové autokorelace a k detekci *hot spotu*. Před použitím scatterplotu data centrujeme (pro lepší rozlišení velkých a malých hodnot). Tento scatterplot následně porovnává centrovanou hodnotu zkoumaného atributu X v lokalitě i na ose x, jež budeme dále značit písmenem Z, s prostorově váženými centrovanými hodnotami téhož atributu v okolí této oblasti na ose y. Výsledný graf, jak můžeme vidět na obrázku 2.4, bývá následně rozdělen na čtyři kvadranty, kdy vertikální dělící přímka odpovídá celkovému průměru centrovaných hodnot Z a horizontální dělící přímka odpovídá celkovému průměru centrovaných hodnot Z, které jsou prostorově vážené podle vzdáleností lokalit od dané lokality i, tedy $\mathbf{W}Z$.



Obrázek 2.4: Kvadranty Moranova scatterplotu

První kvadrant se označuje jako *High-High* a odpovídá situaci, kdy máme v lokální oblasti i nadprůměrnou hodnotu zkoumané proměnné, přičemž je tato oblast obklopená nadprůměrnými hodnotami téže proměnné v blízkém okolí (bývá to označováno jako *hot spot* [25], přičemž je ale potřeba mít na paměti, že v této situaci nevíme nic o jeho statistické významnosti). Třetí kvadrant představuje opačnou situaci a označuje se jako *Low-Low*. Druhý kvadrant je definován jako *Low-High*, což znamená, že v lokální oblasti i máme podprůměrnou hodnotu proměnné obklopenou nadprůměrnými hodnotami v blízkém okolí (jde tak o *prostorový outlier*). Čtvrtý kvadrant je pak opakem druhého. Součástí grafu je i regresní přímka vyjadřující lineární vztah mezi proměnnými na osách x a y, přičemž koeficient β_1 regresní přímky v případě použití řádkově standardizované matice[2] $\mathbf{W}$ odpovídá hodnotě globálního Moranova indexu. Toto tvrzení lze odvodit z následujícího maticového zápisu Moranova indexu

[2]Při volbě jiné matice toto tvrdit nelze.

$$I = \frac{nZ^T\mathbf{W}Z}{SZ^TZ},$$

kde $\mathbf{W}$ je řádkově standardizovaná matice vah a Z je centrovaný atribut X. Dále pro řádkově standardizovanou matici vah dostáváme $S = n$, čímž se vztah zjednoduší na

$$I = \frac{Z^T\mathbf{W}Z}{Z^TZ}.$$

Podíváme-li se na odhad koeficientu β_1 regresní přímky, který lze, dle [18], obecně psát ve tvaru

$$\hat{\beta}_1 = \frac{(X - \overline{X})^T Y}{(X - \overline{X})^T (X - \overline{X})}, \tag{2.6}$$

a nahradíme-li v tomto vztahu (2.6) vektor Y (necentrovaná závislá proměnná) vektorem $\mathbf{W}Z$, a vektor X (necentrovaná nezávislá proměnná) vektorem Z, dostaneme toto vyjádření

$$\hat{\beta}_1 = \frac{(Z - \overline{Z})^T\mathbf{W}Z}{(Z - \overline{Z})^T (Z - \overline{Z})},$$

kde vzhledem k centrovaným datům platí, že $\overline{Z} = 0$. Z tohoto důvodu dostáváme

$$\hat{\beta}_1 = \frac{Z^T\mathbf{W}Z}{Z^TZ} = I.$$

Regresní přímku Moranova scatterplotu s předchozím odhadem $\hat{\beta}_1 = I$ obecně zapisujeme ve tvaru

$$\mathbf{W}Z = \hat{\beta}_0 + \hat{\beta}_1 Z,$$

ve kterém odhad $\hat{\beta}_0$ vychází z následujícího vztahu [18]

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X},$$

ve kterém místo Y uvažujeme $\mathbf{W}Z$, a místo X uvažujeme Z, přičemž vzhledem k rovnosti $\overline{Z} = 0$ tento odhad definujeme následovně

$$\hat{\beta}_0 = \overline{\mathbf{W}Z}.$$

Smysl použití Moranova scatterplotu tedy spočívá v pozorování závislosti mezi hodnotami v dané lokální oblasti a hodnotami v okolí této oblasti.

Globální G-statistika

Globální G-statistika, nebo také globální getis-ord, nabývající hodnot mezi 0 a 1, pro n pozorování, je definovaná prostřednictvím následujícího vztahu

$$G(d) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d) X_i X_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j},$$

pro $j \neq i$, přičemž zbylé značení je stejné jako bylo u Moranova indexu, a kde

$$w_{ij}(d) = \begin{cases} 1 & \text{pro } d_{ij} \leq d \\ 0 & \text{pro } d_{ij} > d \end{cases}$$

s tím, že $\mathbf{W}$ je binární váhová matice s prvky $w_{ij}(d)$ závisejícími na prahové vzdálenosti d z kapitoly 1.3, a značení d_{ij} odpovídá vzdálenosti mezi lokalitami i a j, pro jejíž výpočet lze volit mezi Euklidovskou nebo Manhattanskou metrikou.

Globální G-statistika nabývá vysokých hodnot v situaci, kdy se ve zkoumaném území vyskytují blízko sebe lokality s vysokými hodnotami zkouma-

ného atributu (statistika tak indikuje prostorovou asociaci vysokých, resp. nadprůměrných hodnot, a pokud je její hodnota statisticky významně vyšší než její očekávaná hodnota, jedná se o *hot spot*) a nízkých hodnot v situaci, kdy se blízko sebe vyskytují lokality s nízkými hodnotami atributu (statistika indikuje prostorovou asociaci nízkých, resp. podprůměrných hodnot, kdy pokud je její hodnota statisticky významně nižší než její očekávaná hodnota, jedná o *cold spot*). Oproti globálnímu Moranovu indexu ale nedokáže identifikovat negativní prostorovou autokorelaci (lokality s odlišnými hodnotami oproti hodnotám svých sousedů).

Hodnota G-statistiky se, jak již bylo řečeno, porovnává s její očekávanou hodnotou, která je dána vztahem

$$E[G(d)] = \frac{S(d)}{n(n-1)},$$

kde

$$S(d) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d).$$

Očekávaná hodnota $E[G(d)]$ představuje hodnotu G-statistiky v situaci, kdy ve zkoumané oblasti není přítomná významná prostorová asociace hodnot. Test statistické významnosti rozdílu G-statistiky a její očekávané hodnoty vychází z normálního rozdělení, a pokud zkoumaná proměnná není normálně rozdělena, je G-statistika asymptoticky normální při dostatečně velkém množství sousedů ve vzdálenosti d (uvádí se alespoň 8). O tom, zda je rozdíl mezi $G(d)$ a $E[G(d)]$ statisticky významný, rozhodujeme opět na základě z-score definovaném předpisem

$$Z_{G(d)} = \frac{G(d) - E[G(d)]}{\sqrt{Var[G(d)]}},$$

kde

$$Var[G(d)] = E[G(d)^2] - (E[G(d)])^2,$$

a

$$E[G(d)^2] = \frac{[B_0 m_2^2 + B_1 m_4 + B_2 m_1^2 + m_2 + B_3 m_1 m_3 + B_4 m_1^4]}{(m_1^2 - m_2)^2 n(n-1)(n-2)(n-3)},$$

kde

$$B_0 = (n^2 - 3n + 3)S_1(d) - nS_2(d) + 3S^2(d),$$

$$B_1 = -[(n^2 - n)S_1(d) - 2nS_2(d) + 3S^2(d)],$$

$$B_2 = -[2nS_1(d) - (n+3)S_2(d) + 6S^2(d)],$$

$$B_4 = S_1(d) - S_2(d) + S^2(d),$$

$$S_1(d) = \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}(w_{ij}(d) + w_{ji}(d))^2,\ j \neq i,$$

$$S_2(d) = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij}(d) + \sum_{j=1}^{n} w_{ji}(d)\right)^2,\ j \neq i,$$

s parametrem m_j získaným vztahem

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j,\ j = 1, 2, 3, 4,$$

přičemž pro $j = 1$ představuje m_j průměr hodnot X_i přes všechny územní jednotky $i = 1, \ldots, n$.

Kladná hodnota z-score potom vypovídá o vysoké kladné prostorové asociaci hodnot (tj. *hot spot*) a záporná hodnota z-score o nízké kladné prostorové asociaci hodnot (tj. *cold spot*).

2.1.2. Lokální statistiky prostorové autokorelace (asociace)

V předchozí podkapitole jsme rozebírali globální statistiky prostorové autokorelace. Takové statistiky předpokládají homogennost[3] prostorového procesu. Míra prostorové autokorelace se ale může lišit v různých lokalitách celkového zkoumaného území, a prostorový proces tak může být heterogenní. Pro popis prostorové heterogenity se proto spoléháme na míry, které detekují prostorovou autokorelaci v lokálním měřítku. Těmito měrami jsou LISA[4], pod kterou spadá lokální Moranův index s lokálním Gearyho poměrem (který ale nemá využití při detekci *hot spotu*, proto zde bude věnován prostor pouze Moranově indexu), a lokální G-statistika.

Lokální Moranův index

Lokální verze Moranova indexu je dána vztahem

$$I_i = \frac{n(X_i - \overline{X}) \sum_{j=1}^{n} w_{ij}(X_j - \overline{X})}{\sum_{j=1}^{n}(X_j - \overline{X})^2},$$

pro $i \neq j$, přičemž vztah mezi globálním a lokálním Moranem je dán následovně

$$I = \frac{\sum_{i=1}^{n} I_i}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}},$$

tj. globální Moranův index lze rozložit do n lokálních oblastí. Po zamítnutí nulové hypotézy u globálního Moranova indexu počítáme statistiku lokálního Moranova indexu zvlášť pro každou lokální oblast celkového zkouma-

[3]Homogenností v prostoru se myslí, že velikost prostorové autokorelace je rovnoměrná v celé zkoumané oblasti. Opakem je prostorová heterogenita, kdy je možné v jedné části zkoumané oblasti najít pozitivní prostorovou autokorelaci a v jiné negativní.

[4]LISA je zkratkou pro "Lokální indikátor prostorové asociace".

ného území. Chceme zjistit, které lokální oblasti jsou si podobné (co se týče hodnoty zkoumaného atributu) nebo odlišné od lokálních oblastí v jejich sousedství, případně chceme nalézt prostorově odlehlé lokální oblasti, kde se hodnota zkoumaného atributu v takovéto oblasti velmi liší od svého okolí. Statistika může nabývat libovolné hodnoty z množiny reálných čísel a stejně jako v globálním případě, srovnáváme každou hodnotu Moranova indexu s očekávanou hodnotou definovanou vztahem

$$E[I_i] = -\frac{\sum_{j=1}^{n} w_{ij}}{(n-1)}.$$

Pro každou lokalitu i určíme statistickou významnost rozdílu mezi I_i a $E[I_i]$ prostřednictvím z-score, stejně jako tomu bylo v globálním případě, s rozdílem, že rozptyl indexů I_i je tentokrát definován vztahem

$$Var[I_i] = w_{i.}^{(2)} \frac{(n - m_4/m_2^2)}{(n-1)} + 2w_{i(kh)} \frac{(2m_4/m_2^2 - n)}{(n-1)(n-2)} - \frac{w_{i.}^2}{(n-1)^2},$$

kde

$$w_{i.}^2 = \left(\sum_{j=1}^{n} w_{ij} \right)^2,$$

$$w_{i.}^{(2)} = \sum_{j=1}^{n} w_{ij}^2, \ i \neq j,$$

a

$$2w_{i(kh)} = \sum_{k \neq i} \sum_{h \neq i} w_{ik} w_{ih}.$$

Opět ale nelze předpokládat, že se výsledné statistiky řídí normálním rozdělením, a proto se při testování významnosti rozdílu hodnoty lokálního Moranova indexu od očekávané hodnoty přistupuje k *podmíněné permutaci* [3]. Ta spočívá v tom, že je zafixována hodnota X_i atributu X v dané lokalitě

i, a zbylých $n-1$ hodnot je náhodně permutováno v sousedních lokalitách k získání referenčního rozdělení pro porovnání s danou lokální statistikou I_i. Výsledkem jsou p-hodnoty pro každou lokální oblast i, které lze použít k posouzení významnosti rozdílu mezi indexy a očekávanými hodnotami odpovídajících lokalit.

V případě lokálního Moranova indexu testujeme hypotézu

H_0: *Neexistuje prostorová autokorelace mezi hodnotami v lokalitě i a hodnotami v jejím okolí*

oproti alternativě

H_A: *Existuje prostorová autokorelace mezi hodnotami v lokalitě i a hodnotami v jejím okolí.*

Vzhledem k problému mnohonásobného testování hypotéz je ale potřeba upravit získané p-hodnoty. Bez úpravy by došlo ke zvýšení pravděpodobnosti chyby 1. druhu (tj. zamítneme H_0 a přitom H_0 platí). Mezi nejznámější korekční metodu úpravy p-hodnot spadá *Bonferroniho metoda*. Touto metodou dojde k omezení p-hodnoty tak, že se zvolená hladina významnosti α vydělí počtem testovaných hypotéz n. Tedy $\frac{\alpha}{n}$ je mezní hodnota pro určení významnosti. Další častou volbou je, dle zdroje [40], *FDR* korekce p-hodnoty. Při použití *FDR* metody nejprve vzestupně seřadíme p-hodnoty získané při testování hypotéz v n lokalitách (popř. atributových bodech), tedy $p_{(1)} \leq \cdots \leq p_{(n)}$, a najdeme největší index i, pro který je $p_{(i)} \leq \frac{i\alpha}{n}$, kde α je zvolená hladina významnosti. Následně zamítneme všechny nulové hypotézy, pro něž jsou p-hodnoty menší nebo shodné s $p_{(i)}$.

Lokální G-statistika

Lokální G-statistika je definovaná vztahem

$$G_i(d) = \frac{\sum_{j:j\neq i} w_{ij}(d)X_j}{\sum_{j:j\neq i} X_j}, \tag{2.7}$$

kde definice $w_{ij}(d)$ je stejná jako u globální verze G-statistiky a X_i s X_j jsou opět hodnoty atributu X v lokalitách i a j.

Statistika je počítána pro každou lokální oblast, aby bylo zřejmé, jak je hodnota v dané lokální oblasti i spjata s hodnotami svého okolí definovaném prahovou vzdáleností d (ať už Euklidovskou nebo Manhattanskou). Za sousedy lokální oblasti i považujeme oblasti, které jsou od i vzdáleny nejvýše d. Výše napsaná verze statistiky nebere v úvahu hodnotu zkoumané proměnné v lokální oblasti i (jinak řečeno, nepovažuje za souseda sebe samu). Jde tedy o poměr váženého průměru hodnot v okolí oblasti i k součtu všech hodnot bez započtení hodnoty X_i v lokalitě i. Jiná verze, která při výpočtu zahrnuje i hodnotu X_i, a proto v tomto případě bude výjimečně $w_{ii} = 1$, je dána vztahem

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)X_j}{\sum_{j=1}^{n} X_j},$$

přičemž ji lze interpretovat jako poměr váženého průměru v okolí oblasti i k celkovému součtu hodnot.

Pokud jde o interpretaci, hodnota G-statistiky, která je větší než její očekávaná hodnota daná vztahem

$$E[G_i^*(d)] = E[G_i(d)] = \frac{\sum_{j=1}^{n} w_{ij}(d)}{(n-1)},$$

naznačuje, že jde o *High-High* oblast, resp. *hot spot*. Pokud je hodnota menší

než očekávaná hodnota, jedná se o *Low-Low* oblast, resp. *cold spot*. Na rozdíl od lokálního Moranova indexu tento přístup ale nezohledňuje prostorově odlehlé hodnoty. Definitivní rozhodnutí o přítomnosti *hot spotu*, resp. *cold spotu* je ale potřeba opět podpořit p-hodnotou, kdy nás zajímá, zda je rozdíl mezi lokální hodnotou G-statistiky a její očekávanou hodnotou statisticky významný. Tento statisticky významný rozdíl je, tak jako v globálním případě, testován prostřednictvím z-score, ve kterém je rozptyl lokální G-statistiky definován takto

$$Var[G_i(d)] = E[G_i^2(d)] - (E[G_i(d)])^2,$$

kde pro $j \neq i$ definujeme

$$E[G_i^2(d)] = \frac{1}{(\sum_{j=1}^{n} X_j)^2}\left[\frac{S_i(d)(n-1-S_i(d))\sum_{j=1}^{n} X_j^2}{(n-1)(n-2)}\right] + \frac{S_i(d)(S_i(d)-1)}{(n-1)(n-2)},$$

kde

$$S_i(d) = \sum_{j=1}^{n} w_{ij}(d),$$

přičemž pro $Var[G_i^*(d)]$ je vztah stejný, jen bude platit i pro $j = i$.

V případě lokální G-statistiky testujeme hypotézu

H_0: *Neexistuje žádná souvislost mezi hodnotami v okolí lokality i vymezeném vzdáleností d*

oproti alternativě

H_A: *Existuje souvislost mezi hodnotami v okolí lokality i vymezeném vzdáleností d*

Stejně jako u Moranova indexu není vhodné předpokládat, že se výsledné statistiky pro jednotlivé lokální oblasti řídí normálním rozdělením. Proto je i zde možnost použití *podmíněné permutace*. Opět je následně potřeba provést korekci p-hodnot.

2.2. Základní metody analýzy bodů

V rámci této kapitoly, k jejíž sepsání bylo čerpáno především z [22] a [11], se nejprve podíváme na metody prostorové analýzy bodů, které zohledňují pouze umístění bodů v prostoru, nikoliv hodnoty atributů těchto bodů. Tyto metody lze chápat jako *hot spot* analýzu v situaci, kdy pracujeme s body, přičemž nemáme k dispozici žádné vymezené části zkoumané oblasti (ve smyslu např. městských částí, okresů, krajů apod.). Na závěr kapitoly uvedeme, jakým způsobem provést *hot spot* analýzu na bodech, u nichž zohledňujeme hodnotu nějakého atributu X.

2.2.1. Kvadrátová analýza

Kvadrátová analýza je jednou z metod pro detekci prostorových vzorů na základě uspořádání bodů v prostoru (jinak řečeno, na základě bodové distribuce) v situaci, kdy nemáme k dispozici žádná prostorová ohraničení uvnitř zkoumané oblasti (tj. např. městské části). Metoda hodnotí bodovou distribuci na základě toho, jak se mění hustota rozložení bodů v prostoru. Tuto hustotu obvykle porovnává s teoreticky vytvořeným náhodným vzorem, aby se zjistilo, zda je dané bodové rozložení v prostoru více shluklé, nebo více rovnoměrně rozložené oproti náhodnému vzoru.

Než začneme zkoumat prostorový vzor bodů, je potřeba vytvořit pravidelnou čtvercovou mřížku k překrytí studované oblasti (mřížka svým způsobem nahradí chybějící ohraničení uvnitř zkoumané oblasti) abychom, stejně jako tomu bylo u polygonů, spočítali počty bodů v jednotlivých čtvercích. Otázkou však je, kolik čtverců k tvorbě mřížky zvolit. Podle Greig-Smith experimentu

je optimální počet čtverců k tvorbě mřížky určen vztahem

$$Os = \frac{2A}{n}, \tag{2.8}$$

kde Os vyjadřuje obsah mřížky (tj. celkový počet čtverců, jež tvoří výslednou mřížku), n je počet bodů v distribuci a A je plocha studované oblasti. Po vytvoření mřížky a zjištění rozložení bodů ve čtvercích lze přejít k metodě, pomocí které lze zjistit, zda je rozložení zkoumaných bodů náhodné či nikoliv.

Nejjednodušším teoretickým modelem představujícím náhodný vzor, který lze použít k porovnání s naší bodovou distribucí, je úplná prostorová náhodnost označována zkratkou *CSR* (*Complete Spatial Randomness*). *CSR* předpokládá, že události mají stejnou pravděpodobnost výskytu kdekoliv v rámci studované oblasti, nezávisle na umístění jiných událostí, což je reprezentováno Poissonovým procesem, jehož základem je Poissonovo rozdělení. Při použití Poissonova rozdělení je pravděpodobnost, že ve čtverci leží k bodů definována následovně

$$p(k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

kde e je Eulerovo číslo a λ je parametr Poissonova rozdělení. Jelikož střední hodnota Poissonova rozdělení je λ, lze ji v kontextu kvadrátové analýzy chápat jako průměrný, resp. očekávaný, počet bodů ve čtverci, čehož bude využito při testování hypotézy o *CSR*.

Při zjišťování, zda náš bodový vzor odpovídá náhodnému vzoru testujeme hypotézu

$$H_0\text{: } je\ CSR$$

oproti alternativě

H_A: *není CSR.*

Nulovou hypotézu testujeme pomocí *Testu dobré shody*, který se opírá o skutečnost, že v rámci *CSR* je očekávaný počet pozorování v každé stejně velké oblasti stejný. Označme tedy m počet čtverců stejné velikosti, n celkový počet pozorovaných bodů a n_i počet bodů v i-tém čtverci. Očekávaný počet bodů v každém čtverci spočítáme jako $n^* = n/m$ (n^* tedy odpovídá parametru λ Poissonova rozdělení) a testová statistika je dána tímto vztahem

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n^*)^2}{n^*}.$$

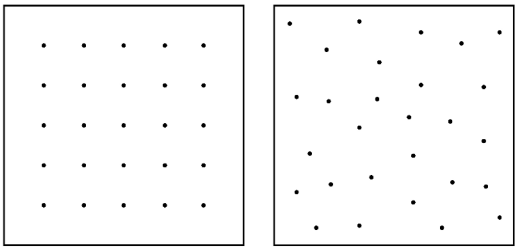
Za platnosti nulové hypotézy má testová statistika χ^2_{m-1} rozdělení, kde $m-1$ značí stupně volnosti. Kritický obor definujeme jako
$W_c = (0, \chi^2_{m-1,(1-\frac{\alpha}{2})}\rangle \cup \langle \chi^2_{m-1,\frac{\alpha}{2}}, \infty)$, kdy pokud $\chi^2 \in W_c$, potom zamítáme nulovou hypotézu, která říká, že náš vzor odpovídá náhodnému vzoru. Nevýhodou metody je, že lze použít pouze v případě, že v každém čtverci je očekávaný počet bodů alespoň 5. V opačném případě je nutné buď změnit velikost mřížky, nebo použít metodu Monte Carlo, která nageneruje několik permutovaných bodových rozložení (resp. souborů) za platnosti nulové hypotézy a spočítá pro každé bodové rozložení χ^2 statistiku. Výsledné rozložení χ^2 statistiky potom porovná se statistikou spočítanou na původních nepermutovaných datech. Nulovou hypotézu následně (ne)zamítne na základě p-hodnoty spočítanou způsobem, který byl definován v kapitole 2.1 pomocí metody Monte Carlo.

Mimo testování hypotézy o náhodném rozložení bodů můžeme chtít přímo vědět, zda je rozložení bodů v prostoru shluklé nebo rovnoměrně rozložené. V případě, že chceme zjistit, zda se body shlukují v prostoru, budeme testovat hypotézu

H_0: *vzor je CSR nebo rovnoměrně rozložený*

H_A: *vzor je shluklý*,

kde rozdíl mezi *CSR* a rovnoměrně rozloženým vzorem je pro lepší pochopení znázorněn na následujícím obrázku 2.5,



Obrázek 2.5: Rozdíl mezi a) rovnoměrně rozloženým vzorem a b) *CSR*

přičemž H_0 zamítáme, pokud $\chi^2 \in \langle \chi^2_{m-1,(1-\alpha)}, \infty)$, kde α je zvolená hladina významnosti. Pro metodu Monte Carlo bychom rozhodli stejným principem, jako bylo popsáno v kapitole 2.1.

Pokud chceme zjistit, zda je prostorový vzor rovnoměrně rozložený, budeme testovat

H_0: *vzor je CSR nebo shluklý*

H_A: *vzor je rovnoměrně rozložený.*

přičemž H_0 zde zamítáme pro $\chi^2 \in (0, \chi^2_{m-1,\alpha}\rangle$. Pro metodu Monte Carlo bychom opět rozhodli způsobem z kapitoly 2.1.

Touto metodou jsme schopni rozhodnout o typu rozmístění bodů z globálního hlediska. V případě, že dojdeme k závěru, že bodová data tvoří shluklý vzor, můžeme přejít k hledání *hot spotů* v lokálních oblastech. K tomu je možné využít metody prostorové autokorelace, které byly zmíněny dříve, jelikož v tomto případě máme místo polygonů k dispozici kvadráty.

2.2.2. Metoda průměrné nejbližší vzdálenosti

Metoda průměrné nejbližší vzdálenosti k identifikaci příslušného bodového vzoru je založena na porovnávání pozorovaných průměrných vzdáleností mezi nejbližšími sousedními body a těmi z náhodného vzoru. Pokud je pozorovaná průměrná vzdálenost větší než ta v případě náhodného vzoru, lze říct, že vzor pozorovaných bodů je více rozptýlený než náhodný. V opačném případě, kdy je pozorovaná průměrná vzdálenost menší než u náhodného vzoru, je náš bodový vzor více shluklý. Při testování, o jaký vzor se v datech jedná, používáme statistiku R (od slova *randomness*) definovanou jako poměr pozorované průměrné vzdálenosti mezi nejbližšími sousedními body a očekávané průměrné vzdálenosti mezi nejbližšími sousedními body za přepokladu *CSR*, tedy

$$R = \frac{r_{obs}}{r_{exp}},$$

ve kterém

$$r_{obs} = \frac{1}{n}\sum_{i=1}^{n} \min_{j \neq i}(d_{ij}),$$

$$r_{exp} = \frac{1}{2\sqrt{n/Pl}},$$

kde n je celkový počet pozorování (resp. bodů), d_{ij} je vzdálenost mezi bodem i a bodem j, a Pl představuje plochu minimálního ohraničujícího pravoúhelníku všech bodů (tj. nejmenší obdélníková nebo čtvercová plocha taková, že obsahuje všechna pozorování). Tato plocha je, dle [8], dána jako součin rozdílů souřadnic $(a_{max} - a_{min}) \cdot (b_{max} - b_{min})$, kde indexem max (resp. min) značíme největší (resp. nejmenší) hodnotu odpovídající souřadnice. Při výpočtu statistiky mohou nastat tři situace:

1. $R > 1$ - výsledek vypovídá o bodovém vzoru, který je více rozptýlený než náhodný.

2. $R < 1$ - pozorovaný bodový vzor je více shluklý než náhodný.

3. $R = 1$ - pozorovaný bodový vzor odpovídá náhodnému vzoru.

Pouze na základě R hodnoty ale nejsme schopni rozhodnout, zda rozdíl mezi pozorovanou a očekávanou hodnotou je statisticky významný. Vypočítanou R hodnotu je potřeba podpořit statistickým testem. Hodnoty zde porovnáme užitím standardizovaného z-score Z_R, daného poměrem

$$Z_R = \frac{r_{obs} - r_{exp}}{SE_{exp}},$$

kde SE_{exp} je směrodatná odchylka průměrné vzdálenosti k nejbližšímu sousedu za předpokladu úplné prostorové náhodnosti (CSR) definovaná jako

$$SE_{exp} = \frac{0.26136}{\sqrt{n^2/Pl}},$$

kdy tento výraz byl, dle [28], odvozen (použitím gama funkce $\Gamma()$) z rozdělení vzdáleností nejbližších sousedů za předpokladu CSR, s využitím výpočtu pravděpodobností, že se v kruhové vzdálenosti o poloměru r od zvoleného bodu nevyskytuje žádný jiný bod a v intervalu mezi r a $r + dr$, kde $d \in \mathbb{R}^+$, se vyskytuje právě jeden bod.

Při zvolené hladině $\alpha = 0.05$ má Z_R statistika z tabulek normálního rozdělení hodnotu 1.96. Pokud je $Z_R > 1.96$ nebo $Z_R < -1.96$ řekneme, že rozdíl mezi pozorovanou a očekávanou hodnotou je statisticky signifikantní. Je-li $-1.96 < Z_R < 1.96$ řekneme, že se pozorovaný vzor statisticky neliší od náhodného vzoru. Testujeme tedy

$$H_0 \colon R = 1,$$

která říká, že *pozorovaný vzor odpovídá náhodnému vzoru*, oproti alternativě

$$H_A \colon R \neq 1,$$

neboli, *pozorovaný vzor neodpovídá náhodnému vzoru.*

Znaménko u Z_R vypovídá o tom, zda je pozorovaný vzor shluklý nebo rozptýlený. Je zde tedy možné testovat i jednostranné alternativy hypotéz, při kterých je hodnota Z_R, při hladině významnosti $\alpha = 0.05$, porovnávána s hodnotou 1.645. V praxi ale testování statistické významnosti za předpokladu normálně rozdělených dat opět není vhodné, proto se i zde využívá Monte Carlo přístup.

Při využití Monte Carla v metodě průměrné nejbližší vzdálenosti sousedů se postupuje tak, že si na pravoúhelníku ohraničujícím všechny body nagenerujeme náhodný vzor tvořený n nezávislými stejně rozdělenými body a z tohoto souboru spočítáme průměrnou vzdálenost mezi nejbližšími sousedními body. Tento proces opakujeme několikrát (např. 599 krát), abychom získali referenční rozdělení průměrných vzdáleností za platnosti nulové hypotézy o náhodném vzoru. Na základě p-hodnoty spočtené vztahem již uvedeným v části Monte Carlo rozhodneme o tom, zda je rozdíl mezi pozorovanou průměrnou vzdáleností mezi nejbližšími sousedními body a průměrnou vzdáleností mezi nejbližšími sousedními body při náhodném vzoru statisticky signifikantní.

Při zamítnutí nulové hypotézy o náhodném vzoru lze přistoupit k vyobrazení *hot spotu.* V kontextu analýzy bodů bez informace (zde konkrétně co bod, to jedna autonehoda, žádný další atribut zde nevystupuje) v situaci,

kdy nemáme žádná další dělení zkoumané oblasti, do kterých bychom body zahrnuli (tj. kraje, okresy apod.), se dá *hot spot* do jisté míry chápat jako oblast s vysokou intenzitou nějaké události (zde nehodovost, uplatňuje se však nejvíce při zkoumání kriminality). To znamená, že k vyobrazení zde využijeme jádrových odhadů hustoty, známé pod zkratkou *KDE* (*Kernel Density Estimation*), která je detailně popsána například v [35]. Nevýhodou přístupu s použitím *KDE*, jak upozorňují články [23] a [19], je, že metoda umí shluky (resp. *hot spoty*) pouze vykreslit, už nám ale neřekne, zda jsou statisticky významné. K určení signifikantnosti shluků je zapotřebí body agregovat do hexagonální, případně čtvercové, sítě a použít metody prostorové autokorelace na polygonech.

2.2.3. Analýza atributových bodových dat

Při hledání *hot spotů* u atributových bodových dat je ze statistik prostorové autokorelace nejčastěji využívána lokální G-statistika. Postupujeme prakticky stejně, jako tomu bylo u polygonů. Pro nalezení lokálních *hot spotů* napočítáme pro každý bod G-statistiku, kterou srovnáme s její očekávanou hodnotou. Jak již bylo dříve řečeno, to, zda je mezi nimi statisticky významný rozdíl určíme prostřednictvím *podmíněné permutace*, neboť nelze předpokládat, že statistika pochází z normálního rozdělení. Jakmile z *podmíněné permutace* získáme potřebné p-hodnoty, je možné výstup (tedy *hot spoty*, resp. *cold spoty*) zakreslit dvěma možnými způsoby.

1. způsob zakreslení

Jedním ze způsobů zakreslení výstupu je prostřednictvím barevně odlišených bodů. Body, pro něž je hodnota lokální G-statistiky signifikantně vyšší než její očekávaná hodnota, nazveme *hot spot* a zaznačíme je červeně.

V opačném případě je nazveme jako *cold spot* a zaznačíme modrou barvou.

2. způsob zakreslení

Druhým způsobem zakreslení výstupu je použití interpolace, konkrétně inverzní vážené vzdálenosti známé pod zkratkou *IWD* (*Inverse Distance Weighting*). Interpolace se provádí na pravidelné husté čtvercové mřížce (rastru), vytvořené na základě rovnoměrného umístění velkého počtu nenapozorovaných bodů (ideální je zvolit alespoň 10 000 bodů k dosažení co možná nejhladšího grafického výstupu), pomocí které rozdělíme celkovou zkoumanou oblast (např. území města Brna). Cílem interpolace je odhadnout hodnoty požadovaného atributu X (zde např. právě (ne)signifikantní hodnoty G-statistiky) v nových bodech o souřadnicích $\{(a_s, b_s) \in \mathbb{R}^2 : \nexists\ (a_i, b_i) \in \mathbb{R}^2 : (a_i, b_i) = (a_s, b_s), i = 1, \ldots, n\}$. Odhad hodnoty atributu X v novém bodě s, o souřadnicích (a_s, b_s), metodou *IWD* zapisujeme $\hat{X}_s$, a definujeme, dle [5], jako vážený průměr blízkých pozorování, tedy

$$\hat{X}_s = \frac{\sum_{i=1}^{n} w_{is} X_i}{\sum_{i=1}^{n} w_{is}},$$

kde X_i je hodnota atributu v napozorovaném bodě i o souřadnicích (a_i, b_i) a

$$w_{is} = \frac{1}{d_{is}},$$

s Euklidovskou vzdáleností d_{is} mezi body i a s. Pokud by se při praktickém výpočtu (v softwaru) stalo, že se souřadnice bodu i a s shodují, bude jako odhad hodnoty atributu vrácena hodnota bodu i, ve kterém je tato hodnota známá (resp. pozorována). Výsledkem tohoto způsobu vykreslení je mapa husté čtvercové sítě vybarvená pomocí barevné škály pohybující se od modré (pro *cold spoty*) až po červenou (pro *hot spoty*).

2.3. Prostorově vážená regrese

V této kapitole, pro jejíž vypracování bylo čerpáno z [15] a [14], dojde k mírnému přeznačení oproti tomu, které bylo zavedeno v kapitole 1. Budeme potřebovat jeden z m atributů jakožto závislou proměnnou. Tento atribut budeme značit jako $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^{n\times 1}$, kde $n \in \mathbb{N}$ představuje počet pozorování. Zbylých $m-1$ atributů, které máme k dispozici, a mohou být použité jako nezávislé proměnné, shrneme do datové matice $\mathbf{X} \in \mathbb{R}^{n\times m}$ s prvky

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1(m-1)} \\ 1 & x_{21} & \cdots & x_{2(m-1)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(m-1)} \end{bmatrix},$$

kde první sloupec slouží k zohlednění absolutního členu regrese.

Model prostorově vážené regrese, známé pod zkratkou *GWR* (*Geographically Weighted Regression*), vychází z globálního regresního modelu, pro i-té pozorování, definovaného vztahem

$$Y_i = \beta_0 + \sum_{j=1}^{m-1} \beta_j x_{ij} + \epsilon_i, \ i = 1, \ldots, n, \tag{2.9}$$

který rozšiřuje tím, že umožňuje odhadovat místo globálních regresních parametrů lokální. Regresní model prostorově vážené regrese pro i-té pozorování má tvar

$$Y_i = \beta_0(a_i, b_i) + \sum_{j=1}^{m-1} \beta_j(a_i, b_i) x_{ij} + \epsilon_i, \ i = 1, \ldots, n, \tag{2.10}$$

kde Y_i je hodnota závislé proměnné, x_{ij} představuje hodnotu j-té nezávislé

proměnné v i-tém bodě, $\epsilon_i \sim N(0, \sigma^2)$ jsou chybové členy, (a_i, b_i) značí souřadnice i-tého bodu v prostoru a $\beta_j(a_i, b_i)$ je realizací spojité funkce $\beta_j(a, b)$ v i-tém bodě. Rovnice (2.10) je tedy speciálním případem rovnice (2.9), u které se předpokládá, že parametry $\beta_0, \ldots, \beta_{m-1}$ jsou prostorově invariantní. *GWR* tím pádem pracuje s předpokladem, že sledovaný jev se v různých lokalitách zkoumané oblasti chová různě, což se projeví tím, že se parametry $\beta_0(a_i, b_i), \ldots, \beta_{m-1}(a_i, b_i)$, $i = 1, \ldots, n$, mohou v prostoru měnit. Tento předpoklad se označuje jako *prostorová nestacionarita*.

2.3.1. Odhad parametrů modelu

U metody *GWR* kalibrujeme regresní model pro každý bod zvlášť. Tím dostáváme pro každý bod odhady lokálních regresních koeficientů. Tyto koeficienty odhadujeme pomocí *Metody vážených nejmenších čtverců*. V metodě *GWR* jsou pozorování, jež budeme dále nazývat datové body, vážena na základě jejich blízkosti k regresnímu bodu, kterým označujeme bod, ve kterém provádíme odhad regresních parametrů. Datovým bodům, které jsou regresnímu bodu blíž, přiřazujeme větší váhu než těm, které jsou od tohoto bodu dál. Zajímavostí metody je, že odhadnout lokální parametry modelu lze pro jakýkoli bod v prostoru bez ohledu na to, zda jde o bod, ve kterém byla pozorovaná naše data (rezidua modelu lze však počítat pouze pro napozorovaná data, proto dále nebudeme pracovat s variantou odhadu regresních parametrů pro libovolný bod v prostoru). Odhad lokálních regresních parametrů je dle *Metody vážených nejmenších čtverců* definován vztahem

$$\hat{\boldsymbol{\beta}}(a_i, b_i) = (\mathbf{X}^T \mathbf{W}(a_i, b_i) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(a_i, b_i) \mathbf{Y}$$

kde $\mathbf{Y}$ je vektor hodnot závislé proměnné, $\mathbf{X}$ je matice hodnot vysvětlujících proměnných, jejíž první sloupec obsahuje prvky rovny jedné, typu $n \times m$, $\mathbf{W}(a_i, b_i)$ je matice vah řádu n, jejíž nediagonální prvky jsou nulové a diagonální prvky představují prostorovou váhu každého z n datových bodů určenou na základě vzdálenosti daného datového bodu od regresního bodu a $\hat{\boldsymbol{\beta}}$ je odhad parametrů $\boldsymbol{\beta}$.

Odhad všech parametrů $\boldsymbol{\beta}$

Podíváme-li se na klasickou regresi více vysvětlujících proměnných maticově definovanou jako

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{2.11}$$

s odhadem regresních parametrů, které jsou v prostoru konstantní, tj.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

vidíme, že ekvivalentem maticového zápisu (2.11) je v případě *GWR* zápis

$$\mathbf{Y} = (\boldsymbol{\beta} \otimes \mathbf{X})\mathbf{1} + \epsilon,$$

kde symbol $\otimes$ je operátor logického násobení, kdy je každý prvek $\boldsymbol{\beta}$ násoben odpovídajícím prvkem matice $\mathbf{X}$. Pro n pozorování a $m-1$ vysvětlujících proměnných jsou $\boldsymbol{\beta}$ a $\mathbf{X}$ řádu $n \times m$ a $\mathbf{1}$ značí jednotkový vektor řádu $m \times 1$. Matice $\boldsymbol{\beta}$ se tak skládá z n sad lokálních parametrů a má tuto strukturu

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(a_1, b_1) & \beta_1(a_1, b_1) & \cdots & \beta_{m-1}(a_1, b_1) \\ \beta_0(a_2, b_2) & \beta_1(a_2, b_2) & \cdots & \beta_{m-1}(a_2, b_2) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(a_n, b_n) & \beta_1(a_n, b_n) & \cdots & \beta_{m-1}(a_n, b_n) \end{bmatrix}, \tag{2.12}$$

přičemž parametry v každém řádku matice $\boldsymbol{\beta}$ odhadujeme vztahem

$$\hat{\boldsymbol{\beta}}(i) = (\mathbf{X}^T\mathbf{W}(i)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(i)\mathbf{Y}, \tag{2.13}$$

kde i reprezentuje řádek v matici (2.12) a $\mathbf{W}(i)$ je matice prostorových vah řádu n tvaru

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}.$$

kde w_{in} je váha přidělená n-tému datovému bodu při odhadu parametrů pro i-tý regresní bod. Odhad (2.13) je váženým odhadem metody nejmenších čtverců, kdy místo konstantní váhové matice pracujeme s maticí, jejíž váhy se mění v závislosti na vzdálenosti datových bodů od regresního bodu, přičemž tuto váhovou matici je nutno počítat pro každý regresní bod zvlášť. Výsledkem kalibrace modelu *GWR* je tedy soubor odhadů lokálních regresních koeficientů pro každý regresní bod.

Rezidua modelu

Kromě odhadů regresních parametrů nás může zajímat reziduální[5] součet čtverců modelu, definovaný vztahem

$$S_e = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2,$$

kde $\hat{Y}_i$ značí vyrovnané hodnoty získané prostřednictvím vztahu

[5]Reziduum = odhad chybového členu

$$\hat{Y}_i = \hat{\beta}_0(a_i, b_i) + \sum_{j=1}^{m-1} \hat{\beta}_j(a_i, b_i) x_{ij},\ i = 1, \ldots, n,$$

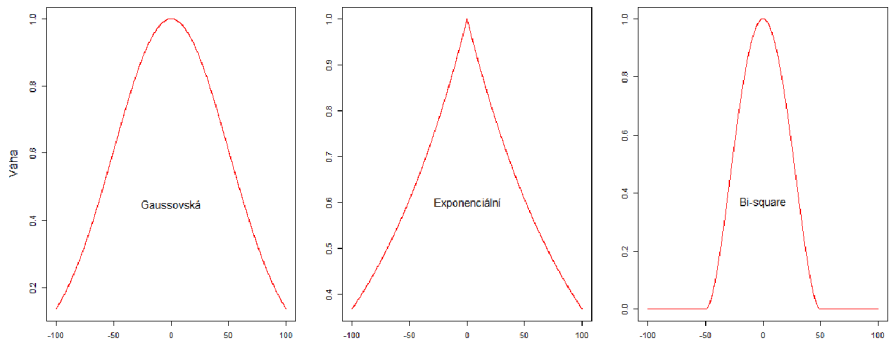
a

$$e_i,\ i = 1, \ldots, n$$

jsou rezidua, přičemž na těchto reziduích bude v praktické části provedena *hot spot* analýza způsobem, jakým se provádí u atributových bodových dat.

2.3.2. Tvorba váhové matice

Při odhadech parametrů jsme využívali matici vah $\mathbf{W}(i)$ (resp. $\mathbf{W}(a_i, b_i)$), která je sestavená na základě blízkosti regresního bodu k datovým bodům v jeho okolí. Okolí regresního bodu definujeme pomocí jádrových funkcí, které vyjadřují, jakým způsobem s rostoucí vzdáleností mezi regresním bodem a sousedními datovými body klesá vliv těchto datových bodů. Klesající vliv datových bodů vyjadřujeme klesající hodnotou váhy, přičemž v regresním bodě je hodnota váhy rovna jedné. Mezi nejčastěji používané typy jádrových funkcí při výpočtu vah patří gaussovská, exponenciální a bi-square funkce, jejichž grafy vidíme na obrázku 2.6.



Obrázek 2.6: Jádrové funkce s volbou šířky pásma jádra $b = 50$

Gaussovská jádrová funkce

Váhu w_{ij} určíme jako spojitou funkci euklidovské vzdálenosti d_{ij} mezi libovolným i-tým regresním bodem, pro nějž odhadujeme parametry, a j-tým bodem v prostoru, v němž jsou data pozorována. Vztah pro výpočet vah je dán následovně

$$w_{ij} = \exp[-\frac{1}{2}(d_{ij}/b)^2],$$

kde b představuje šířku pásma jádra (*bandwidth*). Pro $i = j$, tzn. pro regresní bod, který je zároveň bodem prostoru, ve kterém jsou pozorována data, bude váha rovna jedné a váha ostatních dat bude klesat podle Gaussovy křivky s rostoucí vzdáleností mezi i-tým a j-tým bodem.

Bi-square jádrová funkce

Alternativou k výše zvolenému výpočtu váhy je použití *bi-square* funkce, prostřednictvím které definujeme váhu vztahem

$$w_{ij} = \begin{cases} [1 - (d_{ij}/b)^2]^2 & \text{pro } d_{ij} < b \\ 0 & \text{pro } d_{ij} \geq b, \end{cases}$$

která poskytuje spojitou váhovou funkci až do vzdálenosti b od regresního bodu a nulovou váhu jakémukoli datovému bodu ve vzdálenosti větší než b.

Exponenciální jádrová funkce

Dalším možným způsobem určení vah je použití exponenciální jádrové funkce definované vztahem

$$w_{ij} = \exp[-\frac{1}{2}(|d_{ij}|/b)].$$

Výše zmíněné možnosti určení vah označujeme jako fixní jádra. Fixní jádra jsou charakteristická svou prostorovou neměnností spolu s neměnností velikosti šířky pásma jádra (tzn. že nereagují na měnící se hustotu bodů v prostoru). Opakem jsou adaptivní jádra, která jsou charakteristická tím, že v oblastech s vysokou hustotou bodů mají menší šířku pásma jádra a v oblastech s nízkou hustotou bodů větší šířku pásma jádra.

Adaptivní jádrové funkce

Jedním ze způsobů určení vah je seřazení datových bodů z hlediska jejich vzdálenosti od každého regresního bodu tak, že R_{ij} bude značit pořadí j-tého datového bodu od i-tého regresního bodu z hlediska jejich vzdálenosti. Nejbližší datový bod k regresnímu bodu tak bude mít váhu jedna a zbylé váhy se snižují s rostoucím pořadím podle nějaké spojité funkce, jako je např.

$$w_{ij} = \exp(-R_{ij}/b),$$

kde b opět značí šířku pásma jádra. Touto volbou se automaticky sníží šířka pásma v oblastech s velkým množstvím datových bodů, jelikož vzdálenost např. k desátému datovému bodu bude mnohem menší, než kdyby se regresní bod nacházel v oblasti s nižší hustotu bodů.

Jiným možným způsobem získání vah je užití principu k nejbližších sousedů z podkapitoly 1.3.2 v kombinaci s bi-square jádrovou funkcí pro regresní bod i. Ten definujeme takto

$$w_{ij} = \begin{cases} [1 - (d_{ij}/b)^2]^2 & \text{pro } d_{ij} \leq d_{ij}^{(k)},\ i \neq j \\ 0 & \text{pro } d_{ij} > d_{ij}^{(k)},\ i \neq j \\ 1 & \text{pro } \quad i = j. \end{cases}$$

Na výsledky získané metodou *GWR* má největší vliv volba šířky pásma b, s jejíž pomocí jsou počítány váhy. Šířku pásma lze považovat za vyhlazovací parametr, kdy čím větší šířku pásma zvolíme, tím většího dosáhneme vyhlazení. Pokud zvolíme šířku pásma příliš velkou, dostaneme příliš vyhlazený model, jehož výstupem budou lokální parametry s velmi podobnými hodnotami napříč celou studovanou oblastí (tzn. model nebude reflektovat lokální vlivy). Naopak pokud zvolíme příliš malou šířku pásma, dostaneme natolik odlišné lokální parametry, že nebude možné pozorovat žádné lokální vzorce/zákonitosti. Z tohoto důvodu je potřeba šířku pásma vhodně zvolit.

2.3.3. Volba šířky pásma jádra

1. Křížová validace (Cross-Validation)

Dle metody křížové validace je optimální šířkou pásma taková, která minimalizuje

$$CV = \sum_{i=1}^{n} [Y_i - \hat{Y}_{\neq i}(b)]^2,$$

kde n je počet datových bodů, Y_i je hodnota závislé proměnné v i-tém (regresním) bodě, $\hat{Y}_{\neq i}(b)$ značí, že predikce hodnoty Y_i byla provedena pomocí datových bodů v okolí regresního bodu bez tohoto bodu, jinak by šířka pásma b byla velmi malá a CV by bylo rovno nule (váhy všech bodů kromě regresního bodu by byly zanedbatelné).

2. Akaikeho informační kritérium

Akaikeho informační kritérium používáme při porovnání více modelů mezi sebou, kdy jako "lepší" model vybereme ten, pro nějž je hodnota tohoto kritéria menší. Kritérium definujeme pro metodu GWR následovně

$$AIC_c = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left\{\frac{n + \mathrm{tr}(\mathbf{H})}{n - 2 - \mathrm{tr}(\mathbf{H})}\right\},$$

kde n je počet datových bodů, $\mathrm{tr}(\mathbf{H})$ představuje stopu (součet prvků na diagonále) "hat" matice $\mathbf{H}$, kterou definujeme takto

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{W}(a_i, b_i)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(a_i, b_i),$$

a $\hat{\sigma}$ je odhad směrodatné odchylky chybového členu, jehož rozptyl je definován vztahem

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^T(\mathbf{I} - \mathbf{H})^{\mathbf{T}}(\mathbf{I} - \mathbf{H})\mathbf{Y}}{n}$$

s jednotkovou maticí $\mathbf{I}$.

Poznámka: V případě, že do regresního modelu zahrnujeme kvalitativní vysvětlující proměnné (např. druh komunikace, na které došlo k nehodě - dálnice, komunikace místní apod., nebo stav vozovky v místě nehody - povrch suchý/mokrý), hovoříme o tzv. *umělých proměnných*. Má-li daná kvalitativní

proměnná k různých kategorií, je potřeba vytvořit $k-1$ regresorů tvořených nulami (subjekt do kategorie patří) a jedničkami (subjekt do kategorie nepatří), které představují indikátory jednotlivých kategorií kategoriálních proměnných. Vynechanou kategorii tvořenou nulami potom označujeme jako *referenční*. Tímto procesem tak vlastně vytvoříme k podmodelů regresního modelu, kdy pro každou kategorii dostáváme jeden model. Dále je možné zahrnout do modelu *interakce*, které ukazují, jak se kombinuje vliv vysvětlujících proměnných na vysvětlovanou proměnnou (o vztahu mezi vysvětlujícími proměnnými však interakce nic neříkají).

Kapitola 3

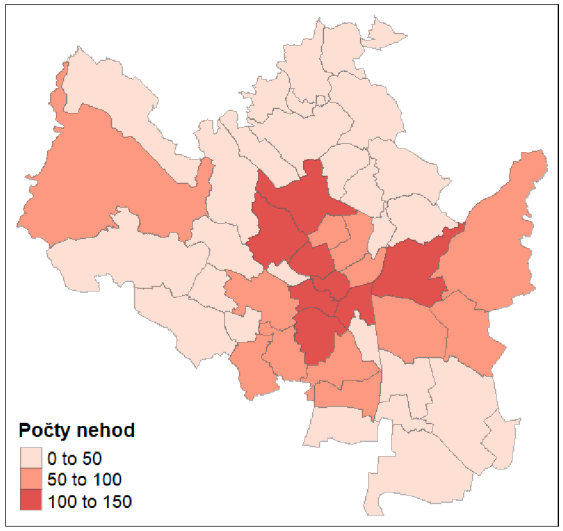
Hot spot analýza nehodovosti města Brna

Pro *hot spot* analýzu byla použita volně dostupná data nehodovosti na území města Brna ze stránky [20]. Datová sada obsahuje celkem 69 741 záznamů nehod s 56 atributy postupně sesbíranými od roku 2010 Policií České republiky. *Hot spot* analýza byla provedena ve statistickém softwaru `R`, přičemž jako podklad pro práci s prostorovými daty posloužily knihy [7] a [21] spolu s internetovými zdroji [38] a [16].

3.1. Hot spot analýza polygonových dat

Nejprve se podíváme, jakým způsobem nalézt *hot spoty* na polygonových datech. Polygony zde zastupuje celkem 48 katastrálních území města Brna získaných ze stránky [17]. V první řadě bylo potřeba data vyfiltrovat takovým způsobem, aby se každá nehoda započítala pouze jednou, neboť záznamy nehod jsou vedené způsobem "co jeden účastník nehody, to jeden záznam". Dále byl pro každý katastr spočítán celkový počet nehod ve vybraném roce, zde

konkrétně v roce 2015. Tím jsme získali atributovou informaci - tj. atribut X představuje počet nehod v daném katastru v roce 2015. Výsledné rozložení počtu nehod v jednotlivých polygonech můžeme vidět na následujícím obrázku 3.1.



Obrázek 3.1: Mapa absolutních četností nehod v jednotlivých katastrech města Brna

Situace na obrázku 3.1 představuje výchozí situaci před zahájením *hot spot* analýzy. Na první pohled lze vidět, že k nejvíce nehodám docházelo především v centru města Brna. Abychom mohli některou z oblastí označit za *hot spot*, je potřeba nejprve provést globální test prostorové autokorelace. Tím zjistíme, zda se v našich datech vůbec vyskytuje nějaký prostorový vzor/shluk.

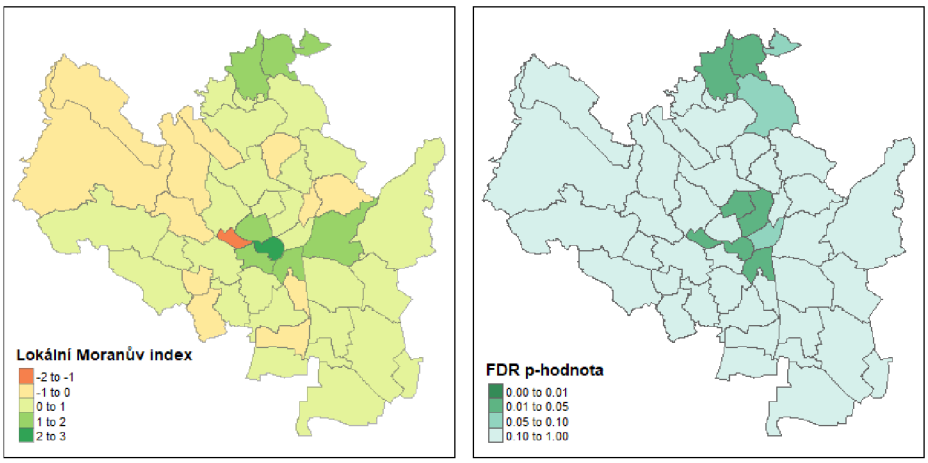
Při testování globální prostorové autokorelace byla zvolena sousednost *queen's case* (vzhledem k tomu, že *queen's case* a *rook's case* se lišil pouze dvěma sousedy, tak zde na volbě prakticky nezáleží) s řádkově standardizovanou maticí vah $\mathbf{W}$. Jako indikátor globální prostorové autokorelace byl použit globální Moranův index. Výpočet byl proveden funkcí `moran.mc` z balíčku

`spdep` [31]. Funkce, kromě výpočtu hodnoty globálního Moranova indexu, testuje jednostrannou hypotézu (2.5) užitím metody Monte Carlo na zvolené hladině významnosti $\alpha = 0.05$, přičemž bylo zvoleno 999 permutací. Výstupem simulace je hodnota Moranova indexu $I = 0.35952$ s p-hodnotou $= 0.001$. Jelikož je p-hodnota $< \alpha$, lze zamítnout nulovou hypotézu ve prospěch alternativy, která říká, že se ve zkoumaném prostoru vyskytují shluky podobných hodnot.

Po zamítnutí nulové hypotézy je možné přejít k prozkoumání prostorové autokorelace z lokálního hlediska s cílem nalézt *hot spoty* nejprve užitím lokálního Moranova indexu, a poté lokální G-statistiky.

3.1.1. Lokání Moranův index

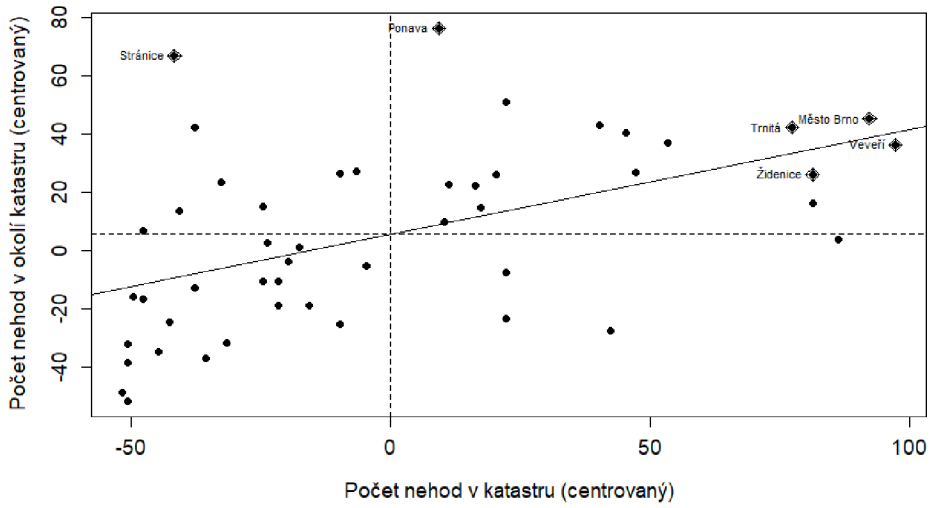
Výpočet hodnoty lokálního Moranova indexu pro každý katastr (při stejné volbě váhové matice i typu sousednosti jako v globálním případě) byl proveden užitím funkce `localmoran_perm` z balíčku `spdep`. Touto funkcí byla současně, v rámci každého katastru, testována oboustranná hypotéza
$H_0\colon I = 0$ oproti alternativě $H_A\colon I \neq 0$ metodou *podmíněné permutace* na hladině významnosti $\alpha = 0.05$, přičemž bylo opět voleno 999 permutací. P-hodnoty k určení statistické významnosti rozdílu mezi očekávanou hodnotou indexu a hodnotou indexu napočítanou z dat, jež jsou taktéž výstupem funkce, byly upraveny pomocí *FDR* korekce použitím funkce `p.adjust` z balíčku `stats` [33]. Vizualizované hodnoty lokálního Moranova indexu pro každý katastr spolu s p-hodnotami upravenými pomocí *FDR* korekce, jež jsou výstupem funkce `localmoran_perm`, jsou zobrazeny na obrázku 3.2.



Obrázek 3.2: Hodnoty Moranova index v daných katastrech (vlevo) spolu s p-hodnotami k určení statistické významnosti (vpravo)

Na levém obrázku vypovídá červeně zbarvená hodnota Moranova indexu o tom, že v daném katastru, konkrétně se jedná o Stránice, docházelo k odlišnému počtu nehod oproti sousedním katastrům (nelze odtud však říct, zda v tomto katastru docházelo k většímu nebo menšímu počtu nehod oproti jeho sousedům). Zároveň, jak je vidět z pravého obrázku, z hlediska p-hodnoty, jež je menší než hladina významnosti $\alpha = 0.05$, se hodnota Moranova indexu tohoto katastru významně liší od očekávané hodnoty Moranova indexu. O tmavě zeleně zbarvených katastrech naopak mluvíme jako o katastrech se srovnatelnými počty nehod v porovnání se sousedními katastry (opět ale nelze rozlišit, zda jde o katastr s vysokým počtem nehod, který je obklopený katastry s vysokými počty nehod nebo naopak). Ne všechny tmavě zelené katastry jsou však vyhodnoceny jako statisticky významné.

Pro nalezení *hot spotů* si částečně pomůžeme Moranovým scatterplotem, který je možné vykreslit funkcí `moran.plot` z balíčku `spdep`. Výstup funkce můžeme vidět na obrázku 3.3.

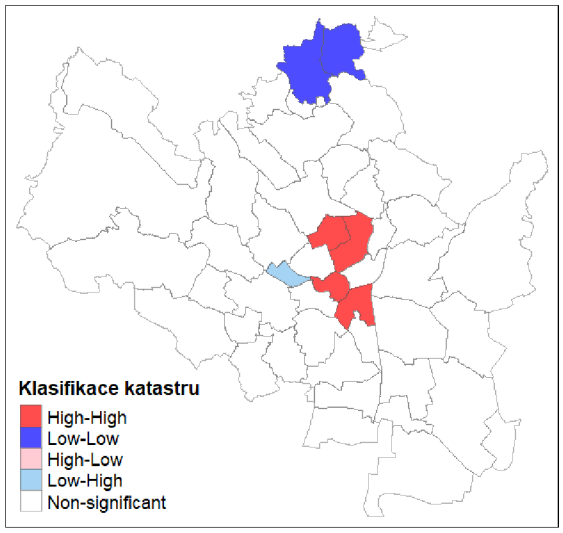


Obrázek 3.3: Moranův scatterplot rozdělený na čtyři kvadranty

Scatterplot nám poskytuje prvotní představu o tom, zda je možné v datech najít nejen *hot spoty*, případně *cold spoty*, ale i odlehlé hodnoty (opět je třeba upozornit na to, že neříká nic o jejich statistické významnosti). Nejvíce katastrů vidíme v prvním kvadrantu (obsahující *hot spoty*) a třetím kvadrantu (obsahující *cold spoty*). Za zmínku stojí první kvadrant, který označujeme jako *High-High*. Vidíme zde 5 názvů katastrů města Brna, které mají největší vliv na výsledný sklon regresní přímky, a které lze současně označit jako *hot spoty*, neboť se jedná o katastry, které jsou charakterizovány vysokým počtem nehod, přičemž jsou zároveň obklopeny katastry stejného charakteru. Dále stojí za zmínku druhý kvadrant, který označujeme jako *Low-High*. Katastr *Stránice*, který má taktéž značný vliv na sklon regresní přímky, zde charakterizujeme jako odlehlou hodnotu ve smyslu katastru s níz-

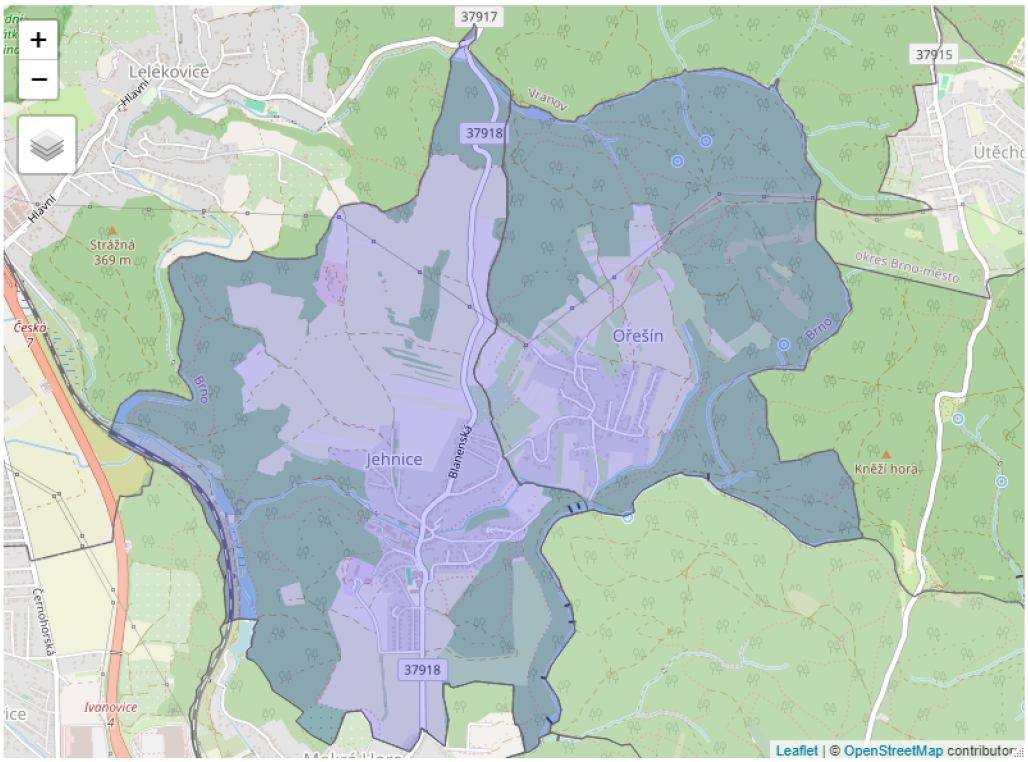
kým počtem nehod, jež je obklopený katastry s vysokým počtem nehod.

Kombinací výstupu Moranova scatterplotu a p-hodnot z obrázku 3.2 dostáváme na obrázku 3.4 nejen požadované statisticky významné *hot spoty* nehod města Brna, které byly našim hlavním cílem analýzy, ale i dva *cold spoty* a jednu odlehlou hodnotu typu *Low-High*.



Obrázek 3.4: (Nejen) *hot spoty* nehod v katastrech nalezené užitím Moranova indexu

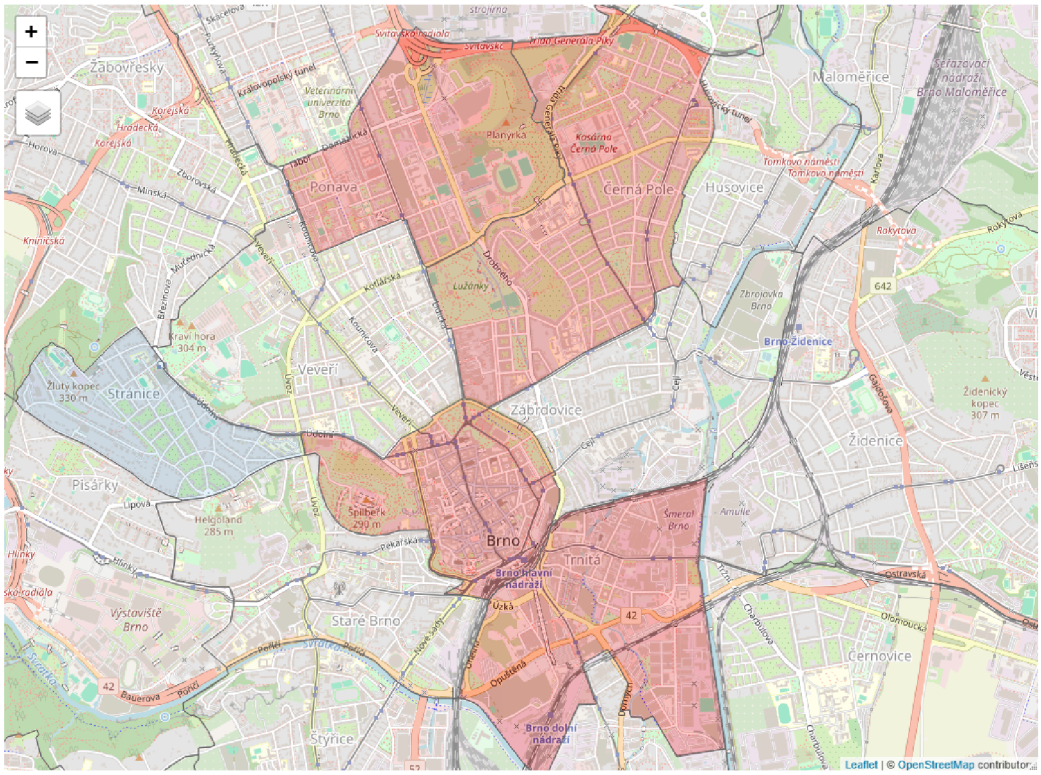
Jako statisticky významné *cold spoty* byly na obrázku 3.4 označeny katastry Ořešín a Jehnice. Jde o katastry, jak můžeme vidět z detailnější mapy 3.5, nacházející se z velké části v zalesněném území Soběšické vrchoviny. Vzhledem k této informaci a k tomu, že se v Jehnicích nachází pouze jediné silniční spojení s Brnem, přičemž do Ořešína vede jediná příjezdová cesta, a to právě z Jehnic, dává smysl, že zde dochází k významně nižší nehodovosti.



Obrázek 3.5: Mapa s vyznačenými *cold spoty* nehod katastrů

Katastr Stránice byl označen klasifikací *Low-High*. Jedná se převážně o zastavěnou/obydlenou oblast s velkým množstvím jednosměrných ulic a absencí tramvajové dopravy (která má jinak značný podíl na počtu nehod v jiných katastrech). Jelikož jde právě primárně o obydlenou oblast, přes kterou nevede žádné významné silniční spojení, které by ztížilo dopravu, je toto možným důvodem, proč byl katastr takto klasifikován. Zbylé vyznačené katastry, jež označujeme jako *hot spoty*, jsou Trnitá, Černá pole, Ponava a město Brno. Město Brno je typické svou hustou tramvajovou sítí a dennodenním velkým provozem. Nachází se zde spousta světelných křižovatek, různých, někdy i nepřehledných, odboček, které mnohdy kříží cestu tramvajím i chodcům. Všechny tyto aspekty nasvědčují důvodu pro vyšší nehodovost v tomto katastru. Co se týče katastru Trnitá, který leží vedle katastru město Brno,

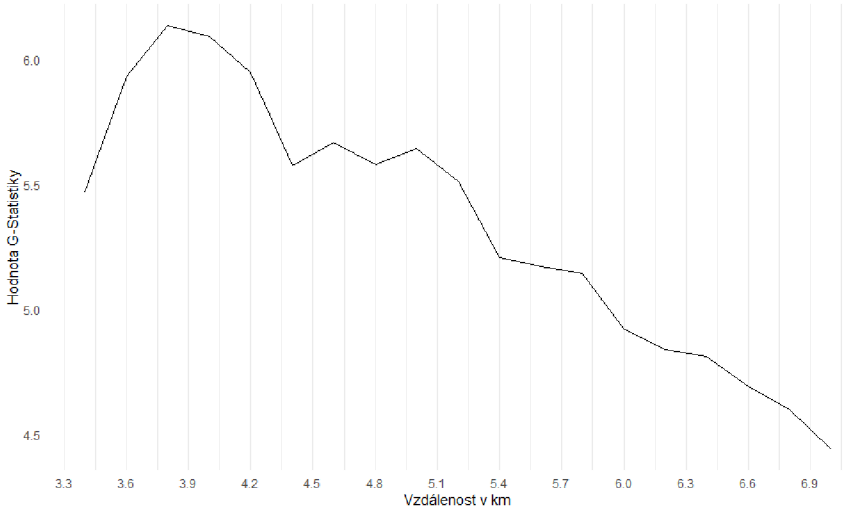
je zde zajímavostí, že jím vede dvouproudová silnice (dálnice), která může být důvodem větší nehodovosti vlivem například nepřizpůsobení rychlosti. Zbylé dva katastry - Ponava a Černá pole jsou typické výskytem nákupních center, hotelů, nebo univerzitou (konkrétně v katastru Černá pole), kde se dá očekávat větší koncentrace řidičů, a může zde proto docházet k více nehodám, byť může jít jen o menší "srážky" například na parkovištích. Co mají ale katastry označené za *hot spot* společné je to, že se zde vyskytuje hustá tramvajová síť a větší množství křižovatek, ulic a odboček, ve kterých zvláště při velkém provozu může docházet k nehodám. Pro detailnější náhled je zde opět přiložena mapa 3.6 s vyznačenými klasifikacemi.



Obrázek 3.6: Mapa s vyznačenými *hot spoty* nehod katastrů spolu s *Low-High* oblastí

3.1.2. Lokální G-statistika

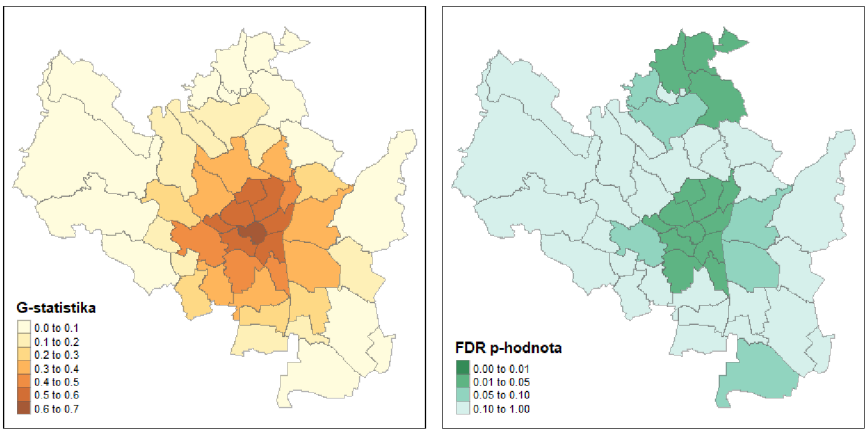
Pro výpočet lokální G-statistiky byla použita verze (2.7) s binární maticí vah **W**, přičemž vzhledem k tomu, jak je statistika konstruovaná, bylo potřeba zvolit vhodnou prahovou (zde konkrétně Euklidovskou) vzdálenost d, po kterou budou katastry považovány za sousedy katastru, pro nějž je statistika počítána. Jednou z možností jak zvolit vhodnou prahovou vzdálenost je vzít takovou vzdálenost, při které je hodnota z-score globální G-statistiky největší. Jelikož nelze předpokládat, že se G-statistika řídí normálním rozdělením, byla vzata taková vzdálenost, při které je přímo hodnota G-statistiky největší. Proces výběru vzdálenosti spočíval v nagenerování několika hodnot G-statistiky pro dané vzdálenosti d, jehož grafický výstup je zobrazen na následujícím obrázku 3.7.



Obrázek 3.7: Volba prahové vzdálenosti určená na základě nejvyšší hodnoty G-statistiky

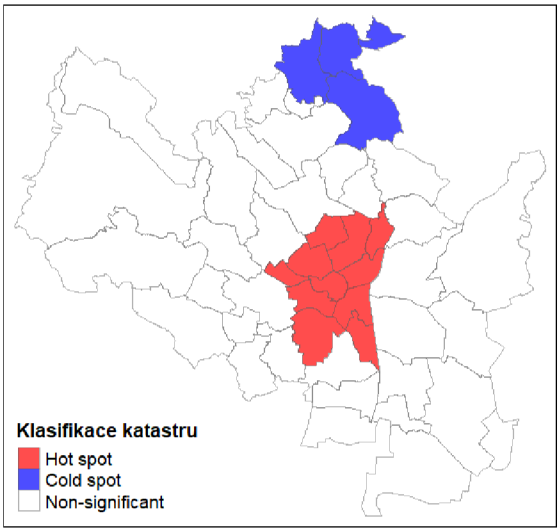
Z obrázku 3.7 je vidět, že nejvyšší hodnoty globální G-statistiky bylo dosaženo při vzdálenosti zhruba 3.8 km (přesná hodnota je 3.8719 km). Pro tuto

prahovou vzdálenost byla pro každý katastr, metodou *podmíněné permutace*, spočítána hodnota G-statistiky spolu s p-hodnotami k následnému určení statistické významnosti rozdílu G-statistiky a očekávané hodnoty G-statistiky užitím funkce `localG_perm`, taktéž z balíčku `spdep`. Výsledek *podmíněné permutace* je vykreslen na obrázku 3.8.



Obrázek 3.8: Hodnoty G-statistiky v daných katastrech (vlevo) spolu s p-hodnotami k určení statistické významnosti (vpravo)

Ačkoliv se nabízí vysoké hodnoty G-statistiky na obrázku 3.8 interpretovat jako potenciální *hot spoty* nebo *cold spoty* nehod a hodnoty blízké nule jako katastry, ve kterých je prostorová autokorelace nulová, je potřeba tyto hodnoty porovnat s očekávanými hodnotami. Důvodem je, že pouze srovnání s očekávanou hodnotou nám ukáže, zda je hodnota G-statiky opravdu vysoká nebo nízká. K tomu, zda je rozdíl mezi G-statistikami a jejich příslušnými očekávanými hodnotami statisticky významný byly použity p-hodnoty z obrázku 3.8, kdy pro p-hodnoty < 0.05 byly katastry klasifikovány jako *hot spoty*, případně *cold spoty*, což je vidět na následujícím obrázku 3.9.

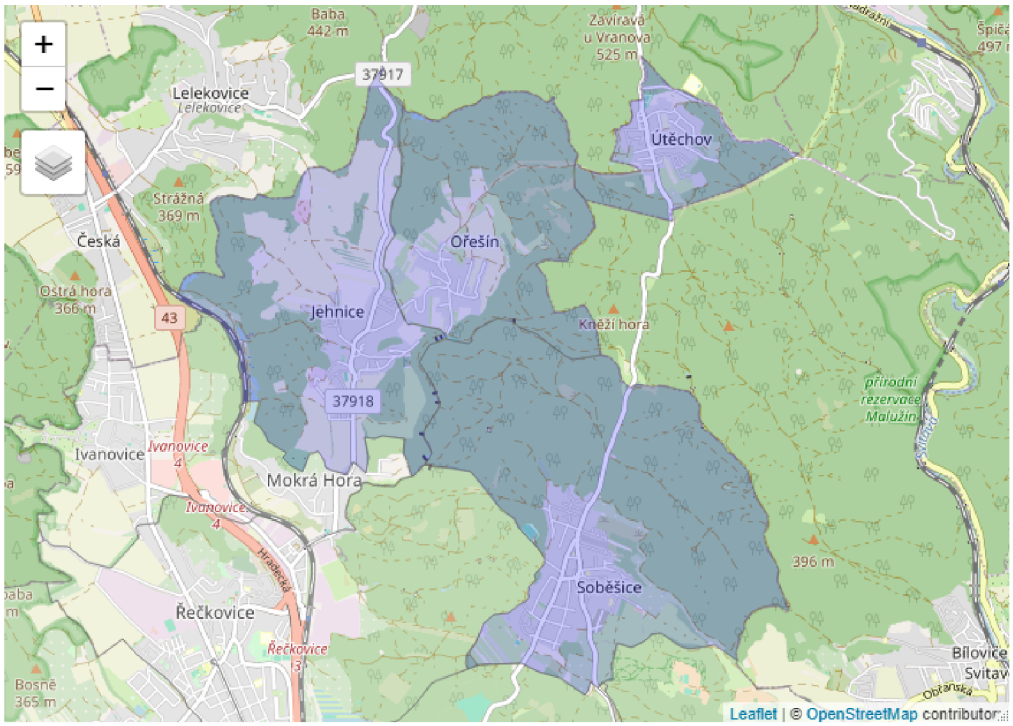


Obrázek 3.9: (Nejen) *hot spoty* nehod v katastrech nalezené užitím G-statistiky

Jak vidíme na obrázku 3.9, lokální G-statistika klasifikovala jako *hot spot* nehod mnohem více katastrů, než lokální Moranův index. Důvodem je především způsob výpočtu lokální G-statistiky, neboť byla použita verze, která při výpočtu G-statistiky pro daný katastr nezohledňuje počet nehod v tomto katastru, jako je tomu u Moranově indexu, ale jen počty nehod v jeho sousedních katastrech. Zároveň je G-statistika hodně ovlivněná volbou prahové vzdálenosti. Dalším důvodem je to, že G-statistika neumí rozeznat odlehlé hodnoty v rámci *hot spotů*, resp. *cold spotů*.

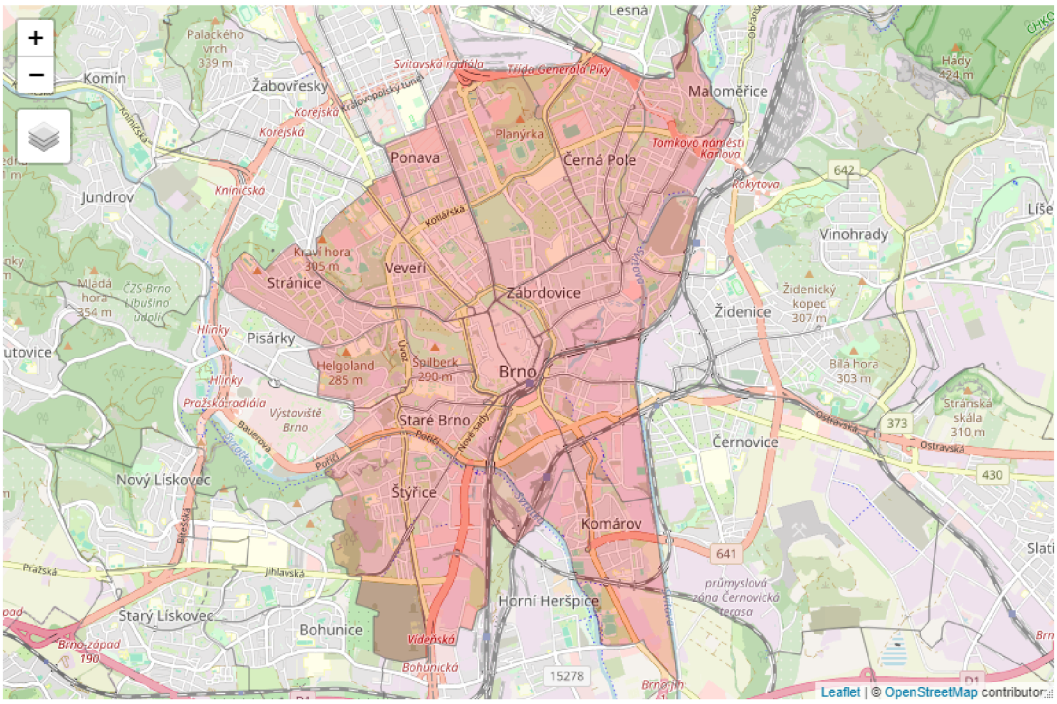
Podíváme se ještě na tyto výsledky detailněji. Na mapě 3.10 máme zobrazeny *cold spoty*, kde oproti případu, kdy byl využit lokální Moranův index, přibyly katastry Soběšice a Útěchov. Oba katastry se nacházejí na okraji Brna v převážně zalesněné oblasti (Soběšické vrchoviny, stejně jako u Jehnic a Ořešína), přičemž jediné, a především přehledné, silniční spojení mezi centrem Brna a Útěchovem vede právě přes katastr Soběšice, ze kterého dále vedou dvě cesty (taktéž nijak komplikované) směrem k centru města Brna.

Dalo by se tedy očekávat, že tudy projíždí spíše obyvatelné těchto katastrů, což by mohlo vést k nižší koncentraci řidičů oproti centru města Brna, a tím pádem by zde nehod mohlo být méně.



Obrázek 3.10: Detailnější pohled na *cold spoty* nehod katastrů

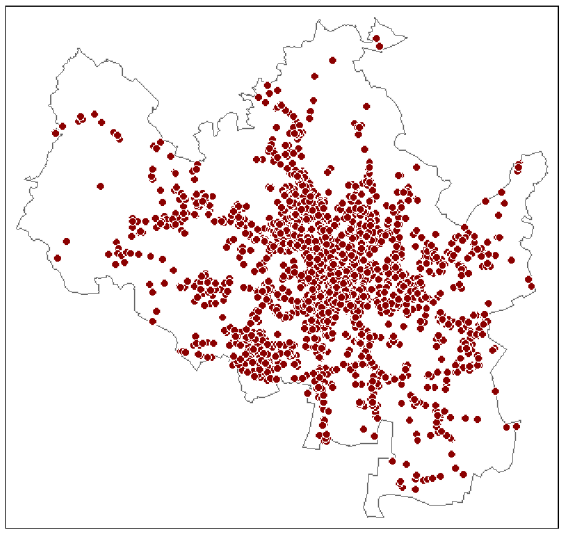
Pokud jde o *hot spoty*, jež jsou zobrazeny na mapě 3.11, přibyly nám zde navíc katastry Komárov, Štýřice, Zábrdovice, Husovice, Veveří a Staré Brno. Tyto katastry mají převážně společné znaky s katastry identifikovanými jako *hot spoty* při použití lokálního Moranova indexu. Zajímavé jsou zde dva katastry - Štýřice a Komárov. Vedou jimi dvě dálnice, ze Štýřic vede dálnice k Rakousku a z Komárova na Slovensko. Vzhledem k možnosti rychlejší jízdy se zde dá předpokládat, že by zde mohlo docházet nejen k většímu množství nehod, ale i k větším finančním škodám způsobených těmito nehodami.



Obrázek 3.11: Detailnější pohled na *hot spoty* nehod katastrů

3.2. Hot spot analýza bodových dat

Předpokladem pro provedení *hot spot* analýzy v této části je absence hranic katastrů města Brna a jakékoliv atributové informace, tj. máme k dispozici pouze souřadnice nehod, k nimž došlo v roce 2015. Výchozí situaci za tohoto předpokladu vidíme na obrázku 3.12.



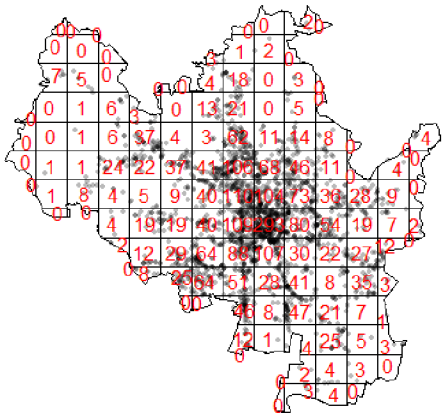
Obrázek 3.12: Bodová data nehodovosti na území města Brna

S touto situací si lze poradit dvěma způsoby - užitím kvadrátové analýzy nebo metody nejbližší vzdálenosti.

3.2.1. Kvadrátová analýza

Při analýze bodových dat, jež obsahují pouze informaci o lokalizaci události (zde nehodovosti) je potřeba nejprve otestovat, zda bodový vzor odpovídá shluklému vzoru, aby pro nás mělo vůbec smysl provádět *hot spot* analýzu. Před provedením testu byla vytvořena čtvercová mřížka k proložení oblasti města Brna. Aby tuto mřížku bylo možné vytvořit, bylo nutné data převést do planárního souřadnicového systému (postup v `R` skriptu). Dále

byl určen "optimální" počet čtverců v mřížce vztahem (2.8) a spočítán počet nehod v každém čtverci funkcí `quadratcount` z balíčku `spatstat` [30]. Výsledek je zobrazen na obrázku 3.13.



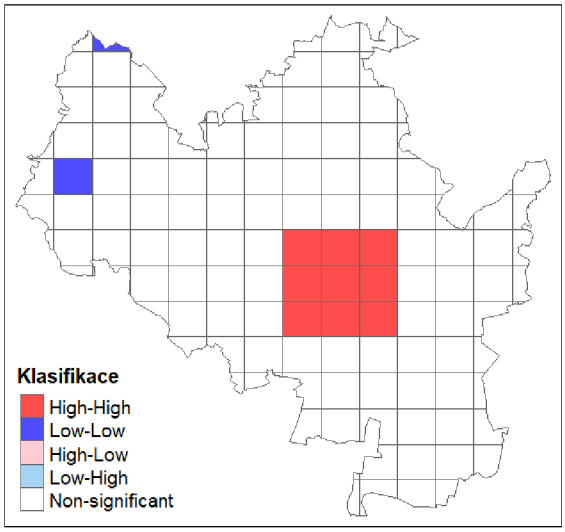
Obrázek 3.13: Počty nehod ve čtvercích jejichž počet je určený vztahem (2.8)

Jak je z obrázku 3.13 vidět, vyskytují se zde čtverce s nulovým počtem nehod. Z tohoto důvodu byl pro test shluklého bodového vzoru zvolena Monte Carlo verze *Testu dobré shody*, jež je možno v `R` provést funkcí `quadrat.test` s parametry `method = "MonteCarlo"` a `alternative = "clustered"`, taktéž z balíčku `spatstat`. Výsledkem testu je hodnota $\chi^2 = 6053.2$ s p-hodnotou = 0.001. Na hladině $\alpha = 0.05$ zamítáme nulovou hypotézu o náhodném (resp. náhodném nebo rovnoměrně rozloženém) vzoru ve prospěch alternativy.

Pro nalezení *hot spotů* v lokálních oblastech území města Brna se postupuje stejně jako u polygonových dat, jelikož zde roli polygonů zastupují čtverce.

Lokální Moranův index

Pro výpočet Moranova indexu byla opět zvolena řádkově standardizovaná matice **W** spolu se sousedností *queen's case*. Vzhledem ke stejnému principu hledání *hot spotů*, jako u polygonálních dat, zde uvedeme pouze výslednou vizualizaci *hot spotů* v dané čtvercové mřížce. Grafický výstup metody tedy vidíme na obrázku 3.14.

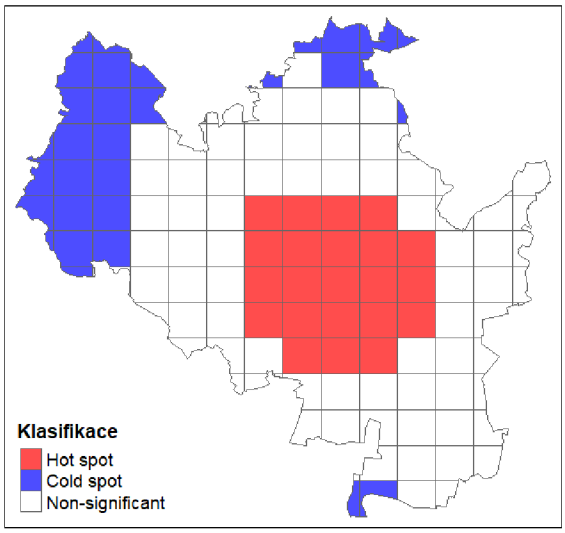


Obrázek 3.14: (Nejen) *hot spoty* nehod nalezené na čtvercové síti užitím lokálního Moranova indexu

Jak je na obrázku 3.14 vidět, použití přístupu kvadrátů s následnou aplikací Moranova indexu nenašlo žádné odlehlé hodnoty, jako tomu bylo při použití katastrů jakožto polygonů (což je ale samozřejmě dáno počáteční volbou čtvercové mřížky). Oblast *hot spotů* nehodovosti však metoda vyhodnotila obdobně jako v předchozím přístupu, tedy v centru města Brna. Odlišné je ale vyhodnocení *cold spotů*, které bude dáno tím, že čtverce, vzhledem k tvaru a velikosti, obsahují jiné počty nehod než katastry.

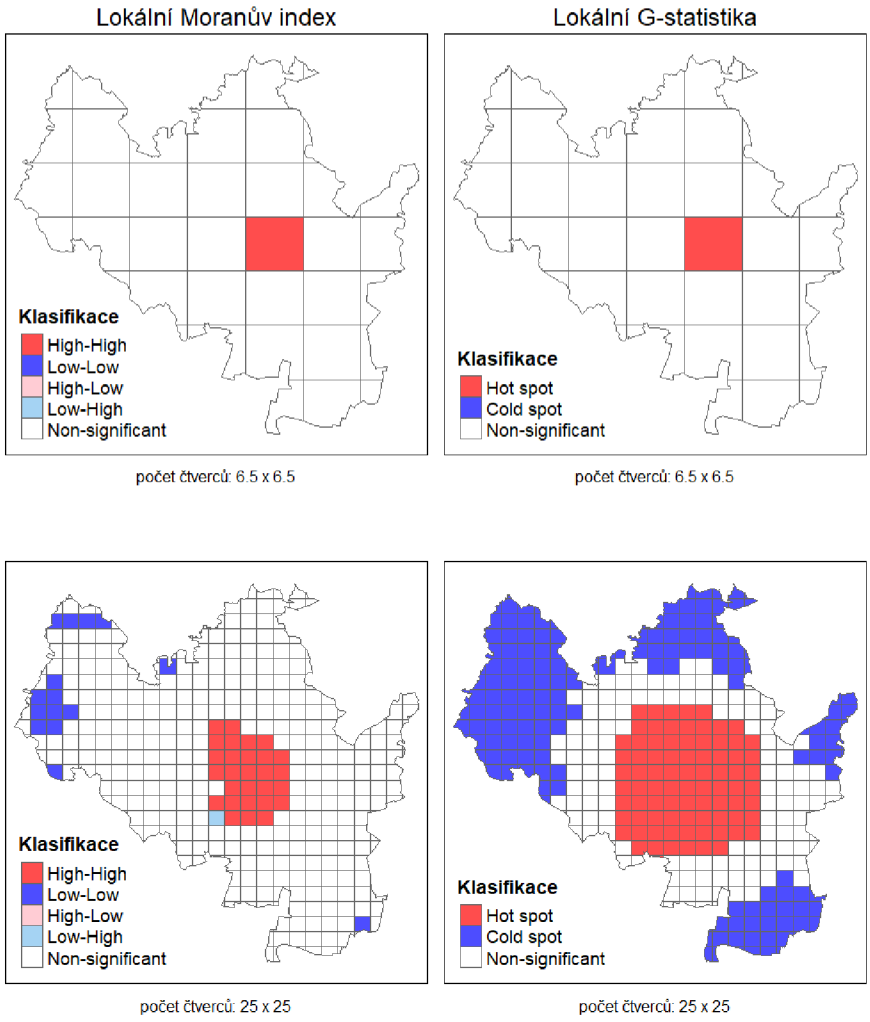
Lokální G-statistika

V případě lokální G-statistiky byla zvolena binární váhová matice **W** s volbou prahové vzdálenosti $d = 3.6$ km, jež byla určena stejným způsobem jako v případě polygonálních dat. Výsledné *hot spoty* nehod města Brna vidíme na obrázku 3.15.



Obrázek 3.15: (Nejen) *hot spoty* nehod nalezené na čtvercové síti užitím lokální G-statistiky

Použitím lokální G-statistiky bylo opět identifikováno mnohem více *hot spotů* i *cold spotů* nehodovosti než u lokálního Moranova indexu. Ačkoliv jsou *hot spoty* opět identifikované v centru města Brna, zasahují napravo částečně i mimo centrum. Zároveň je na levé straně vidět velké množství *cold spotů*, které v katastrech jakožto polygonech nalezeny nebyly. Proto se ze zajímavosti podíváme ještě na obrázek 3.16 s různými volbami počtu čtverců použitých k tvorbě mřížky, abychom viděli, jak moc jsou výsledné *hot spoty* a *cold spoty* těmito volbami ovlivněné.



Obrázek 3.16: (Nejen) *hot spoty* nehod při různých volbách mřížky

Při volbě počtu čtverců 6.5×6.5 ve směru osy x a y dostáváme pro lokální Moranův index stejný *hot spot* jako u lokální G-statistiky. Naopak při volbě počtu čtverců 25×25 jsou vidět značné rozdíly ve výsledném zobrazení *hot spotů*. Lokální Moranův index při této volbě našel jednu odlehlou hodnotu v podobném místě jako u polygonových dat. Lokální G-statistika však při této volbě označila za *hot spoty*, a především *cold spoty* velkou část území města Brna.

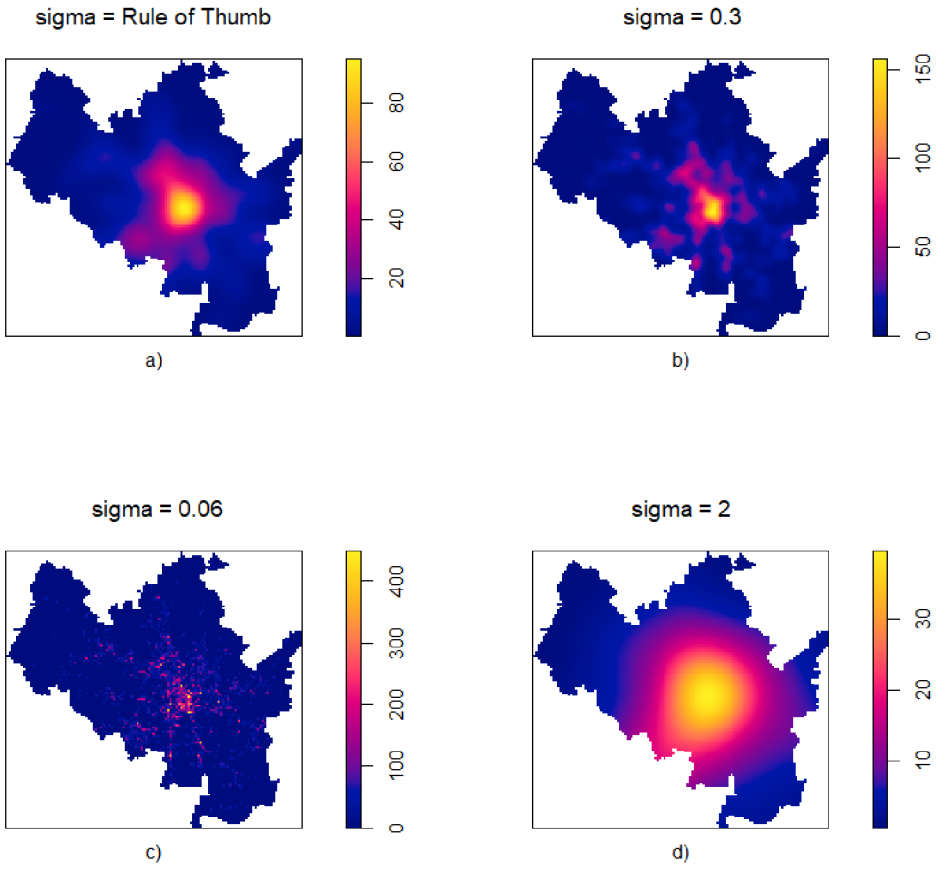
3.2.2. Metoda průměrné nejbližší vzdálenosti

Test (ne)náhodnosti bodového vzoru na hladině významnosti $\alpha = 0.05$ byl proveden metodou Monte Carlo, jejímž cílem bylo nagenerovat rozdělení průměrné nejbližší vzdálenosti sousedů za platnosti náhodného rozložení bodů (nehod) v prostoru k porovnání se skutečnou průměrnou vzdáleností sousedů. Před simulací bylo potřeba napočítat průměrnou vzdálenost nejbližších sousedů pro naše data, k čemuž posloužila funkce `nndist` z balíčku `spatstat`. Průměrná vzdálenost má v našem případě hodnotu $r_{obs} = 0.079$. Pro informativní účely byl zjištěn i index průměrné nejbližší vzdálenosti souseda $R = 0.4187$ (vidíme, že hodnota je menší než jedna, proto by pozorovaný vzor mohl odpovídat shluklému vzoru), pro jehož výpočet bylo potřeba napočítat očekávanou průměrnou vzdálenost, která je $r_{exp} = 0.1886$. Pro výpočet očekávané průměrné vzdálenosti bylo nutno určit plochu minimálního ohraničujícího pravoúhelníku, k čemuž posloužila složená funkce `area.owin(bounding.box.xy())`, taktéž z knihovny `spatstat`. V rámci simulace byla průměrná nejbližší vzdálenost za platnosti H_0 počítána pomocí funkce náhodně umisťující body v prostoru, kterou je funkce `rpoint`, již je možno nalézt ve stejném balíčku jako funkce předchozí. Výsledkem simulace s volbou 999 permutací je p-hodnota = 0.001, proto na hladině významnosti $\alpha = 0.05$ zamítáme H_0 o náhodném prostorovém vzoru.

Jak již bylo zmíněno v teoretické části, v případě dat týkající se kriminality nebo nehodovosti je možné do jisté míry (tj. nedozvíme se nic o statistické významnosti) využít k zobrazení *hot spotů* jádrové odhady hustoty. To, které části města Brna interpretujeme jako *hot spoty* závisí především na volbě vyhlazovacího parametru, jelikož volba jádrové funkce nemá výrazný vliv na výsledný odhad hustoty.

Pro zobrazení míst, které by mohly být *hot spoty* byla použita Gaussovská

jádrová funkce s různými volbami parametru míry vyhlazení, aby bylo zřejmé, jak může volba tohoto parametru ovlivnit výslednou interpretaci *hot spotu*. Výsledné (potenciální) *hot spoty* vidíme na obrázku 3.17.



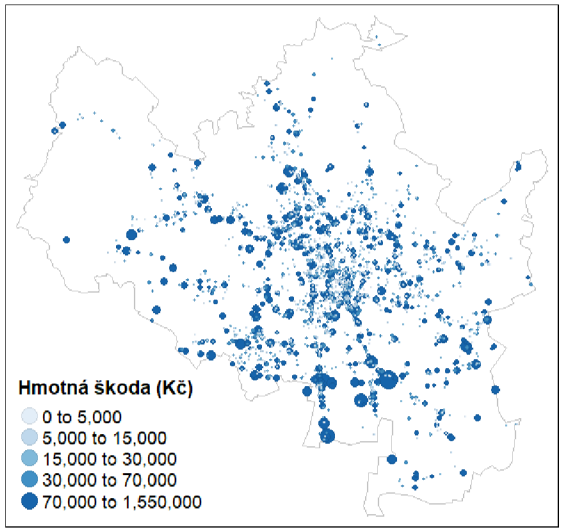
Obrázek 3.17: Potenciální *hot spoty* nalezené pomocí *KDE* aplikované na body z obrázku 3.12

Vyhlazovací parametr na obrázku 3.17 při volbě a) byl určen pomocí pravidla *Rule of thumb* z knihy [7], které bylo vytvořeno k získání optimálního vyhlazovacího parametru jádrové funkce. V porovnání s možnostmi c) a d) varianta a) nejlépe odpovídá oblasti *hot spotu* nehod (žlutě zbar-

veno) tak, jako jej určily i předchozí metody. Stejně tak volba parametru $sigma = 0.3$ ještě relativně dobře odhaduje oblast *hot spotu* nehod města Brna. Co se týče variant c) a d), dostáváme velmi neurčité výsledky. V případě c) je volba parametru tak malá, že prakticky nelze nazvat žádnou oblast jako *hot spot*. S variantou d) se naopak dostáváme k opačnému problému, tj. *hot spotem* bychom prakticky nazvaly většinu území města Brna, což nám nedává žádnou relevantní informaci o nehodovosti. Toto vyobrazení *hot spotu* je však obecně pouze informativní, proto je doporučováno jej používat zároveň s metodami prostorové autokorelace (hlavně z důvodu určení statistické významnosti *hot spotu*) způsobem, jaký byl představen v předchozích částech analýzy.

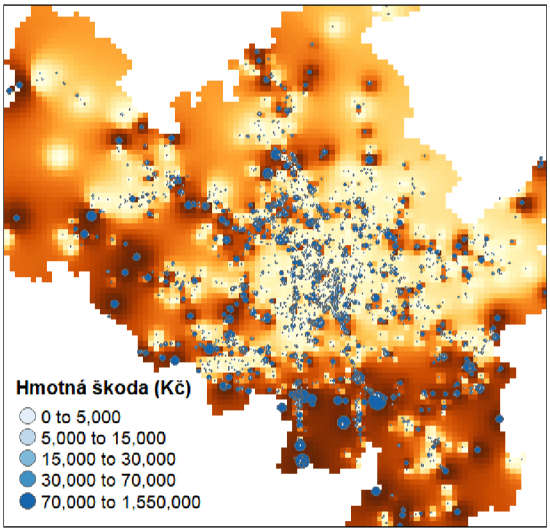
3.3. Hot spot analýza atributových bodových dat

Oproti předchozí části analýzy budeme pracovat s předpokladem, že kromě souřadnic nehod máme k dispozici i informaci o nějakém atributu. Atributovou informací zde budeme rozumět výši hmotné škody v Kč, jež byla způsobená nehodou v roce 2015. Výchozí situaci můžeme vidět na obrázku 3.18.



Obrázek 3.18: Výše hmotné škody způsobená nehodou na území města Brna

Pro lepší viditelnost rozložení výše hmotné škody na území města Brna lze data vizualizovat pomocí interpolace *IDW*. Výchozí situace vizualizovaná užitím interpolace spolu s body z předchozího obrázku 3.18 je demonstrována na obrázku 3.19.

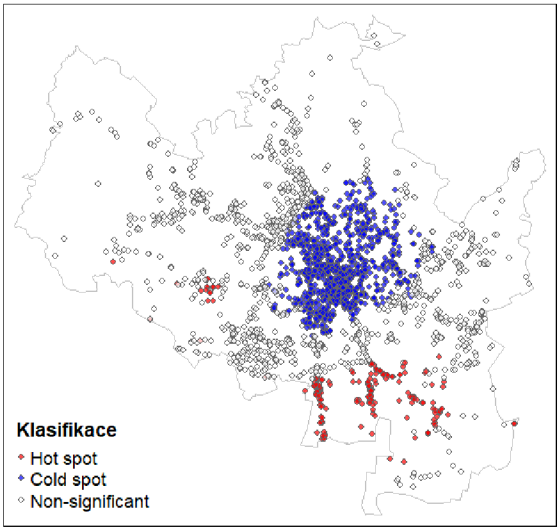


Obrázek 3.19: Výše hmotné škody způsobená nehodou na území města Brna zobrazená užitím interpolace

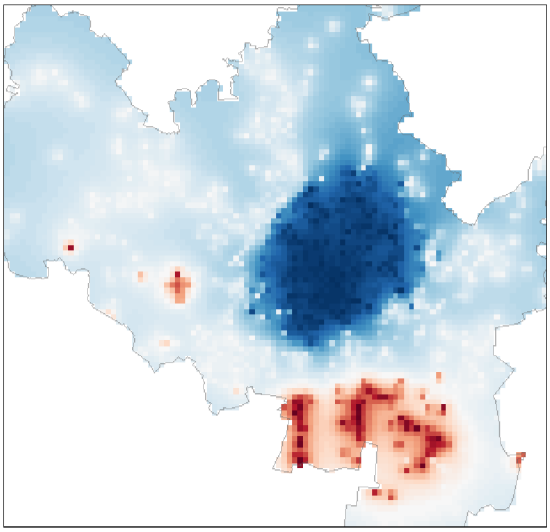
Z obou typů vizualizací je patrné, že k nákladným nehodám docházelo spíše mimo centrum města Brna.

Podíváme se, jak *hot spoty* výše hmotné škody vyhodnotí lokální G-statistika, která je pro tyto účely používána nejvíce. Postup je zde stejný jako u polygonových dat, proto zde uvedeme pouze nalezené *hot spoty* užitím G-statistiky vizualizované dvěma různými způsoby.

Prvním způsobem je vykreslení *hot spotů* pomocí barevně odlišených bodů, kdy G-statistika identifikovala tyto *hot spoty* mimo centrum města Brna, což je možné vidět na obrázku 3.20. Druhým způsobem vizualizace je použití interpolace *IDW*, kterou vidíme na obrázku 3.21. Zároveň je z obrázků vidět, že ačkoliv k nejvíce nehodám docházelo v centru města Brna, škody způsobené nehodou byly vyšší právě mimo centrum.



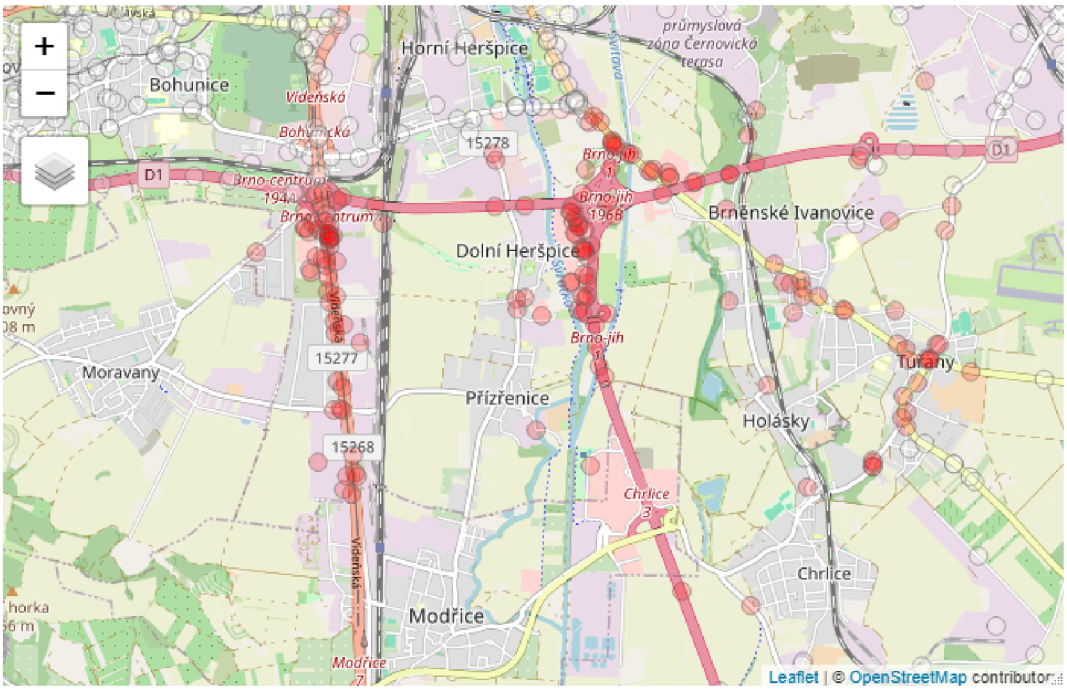
Obrázek 3.20: (Nejen) *hot spoty* výše hmotné škody nalezené lokální G-statistikou



Obrázek 3.21: Interpolovaný výstup *hot spotů* (červeně), *cold spotů* (modře) a nesignifikantních oblastí (bíle)

Současně je potřeba dodat, že k interpolaci *IDW* zvoleného počtu nenapozorovaných bodů v `R`, jež byla použita k vytvoření grafického výstupu na obrázku 3.21, bylo potřeba nejprve nenapozorované body nagenerovat pravidelným umístěním těchto bodů v husté čtvercové mřížce pokrývající území města Brna. Pro tyto účely slouží funkce `spsample` z knihovny `sp` [29], do které je potřeba zadat parametr `type = "regular"` s počtem bodů `n = "pocet_bodu"`, které budeme chtít interpolovat (čím větší `n` zvolíme, tím hladší grafický výstup dostaneme). Pro interpolaci byla následně použita funkce `iwd` z knihovny `spatstat`, ve tvaru `idw(pozorovani ~ 1, data = nase_data, newdata = mrizka)`.

Podívejme se ještě na nalezené *hot spoty* na mapě. První oblast, na kterou se zaměříme (mapa 3.22, odpovídá spodní části obrázku 3.20) bude oblast spadající do katastrů Dolní a Horní Heršpice, Přízřenice a Tuřany.

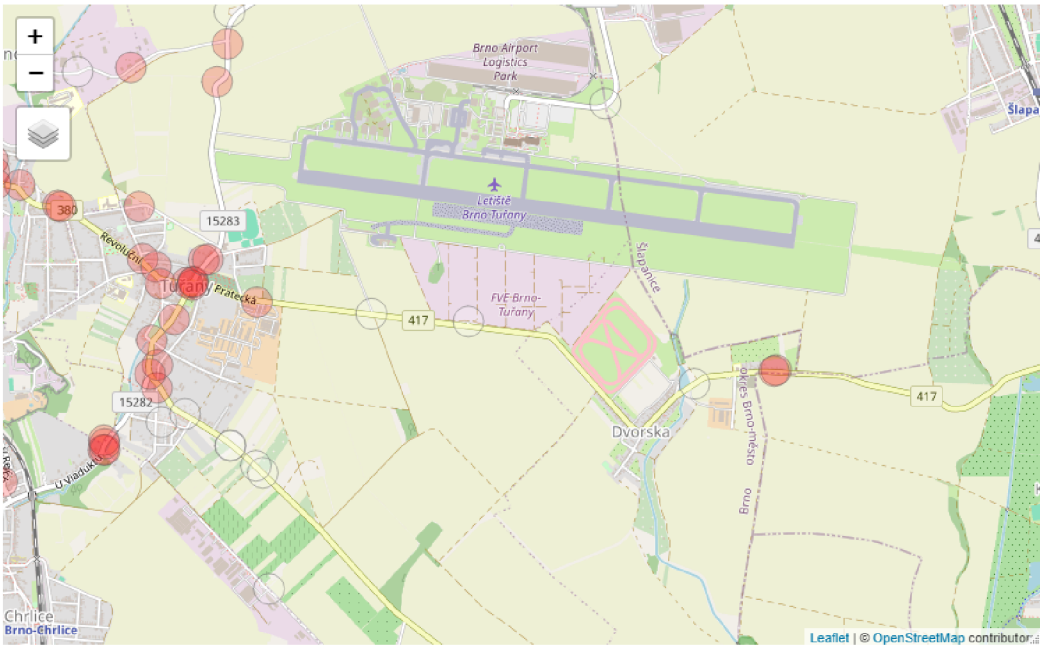


Obrázek 3.22: Detailní pohled na *hot spoty* nehod na jižní části Brna

Dvě hlavní oblasti na mapě 3.22, ve kterých došlo k nákladným nehodám jsou v místě křížení dálnic vedoucích k Rakousku a Slovensku s dálnicí D1. Důvodem, proč byly v těchto místech nehody nákladnější, by mohly být hromadné nehody způsobené ať už vlivem nepozornosti (např. nedobržděním při srážce vozidel přede mnou), nepřizpůsobením jízdy (např. nedáním přednosti při odbočení) nebo kombinací obojího. Z obou dálnic (směr Rakousko a Slovensko) dále vede poměrně rovná cesta, což může pobízet i k rychlejšímu způsobu jízdy a nebezpečnému předjíždění, které následně vede k nehodě. Z dat bylo zjištěno, že ke spoustě nehodám opravdu docházelo vlivem nepozornosti nebo rychlé jízdy řidičů, přičemž hromadné nehody jsou charakteristické spíše pro dálnici označenou jako Vídeňská (na mapě 3.22 vlevo). Zajímavou informací o nehodách je i to, že v oblasti Dolních Heršpic poměrně často docházelo k nehodám, od kterých řidiči ujeli. K nákladným nehodám dále docházelo na silnici vedoucí kolem katastrů Horní a Dolní Heršpice směrem ke katastru Tuřany. První část cesty, vedoucí ze severu k Brněnským Ivanovicím, je přehledná cesta pouze s jednou odbočkou k benzince a Makru. Nabízí se tedy, že výši škody způsobenou nehodou mohla zapříčinit skutečnost, že mezi řidiče, kteří na této trase projíždí a odbočují k Makru, patří převážně řidiči dražších vozidel typu dodávky, teoreticky nákladní automobily, s cílem natankovat nebo nakoupit v Makru. Makro totiž obecně slouží spíše pro velkoobjemové nákupy s cílem zásobování supermarketů, hypermarketů apod. Na základě informací obsažených v datech byly ve skutečnosti nehody v této oblasti způsobovány buď z důvodu nepřiměřené rychlosti, nepřizpůsobení se dopravně technickému stavu vozovky (např. klesání, zatáčka,...), nesprávným couváním, nebo dokonce samovolným rozjetím osobních, případně nákladních, automobilů. V centru Tuřan docházelo k nákladným nehodám hlavně v místě křížení silnic, přičemž je na této trase přítomno větší množ-

ství přechodů pro chodce a zastávek autobusů. Proto k nákladným nehodám mohlo docházet buď proto, že mezi účastníky patřily i autobusy, nebo opět z důvodu nepozornosti vedoucí k hromadné nehodě, přičemž vzhledem k poměrně frekventované trase mezi účastníky nehody mohly patřit i dodávky. Po bližším prozkoumání dat bylo o nehodách v této oblasti zjištěno, že jejich hlavní příčinou bylo nevěnování se řízení vozidla, nebo nepřizpůsobení rychlosti. Zároveň zde docházelo nejen k nehodám osobních automobilů (a to i hromadných, např. zde byla nehoda 3 osobních automobilů), ale i motocyklů, přičemž nebylo výjimkou, pokud některý z účastníků nehody ujel.

Na jihu Brna ještě zůstaneme, neboť zde byla identifikována ještě dvě místa jako *hot spot*, konkrétně na pravém cípu z obrázku 3.20, které se nacházejí napravo od Tuřan a jsou detailně vidět na mapě 3.23.

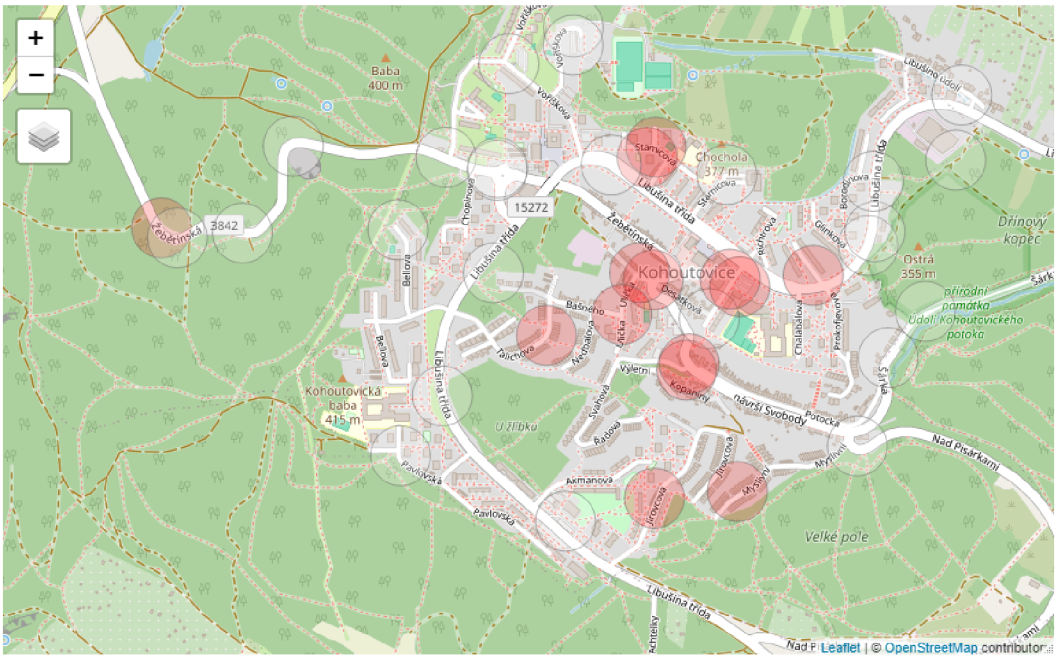


Obrázek 3.23: Detailní pohled na *hot spoty* nehod v cípu z obrázku 3.20

K nehodám zde došlo na přehledném místě silnice v blízkosti autobusové zastávky a polní cesty vedoucí podél pole. Mohli bychom tedy očekávat,

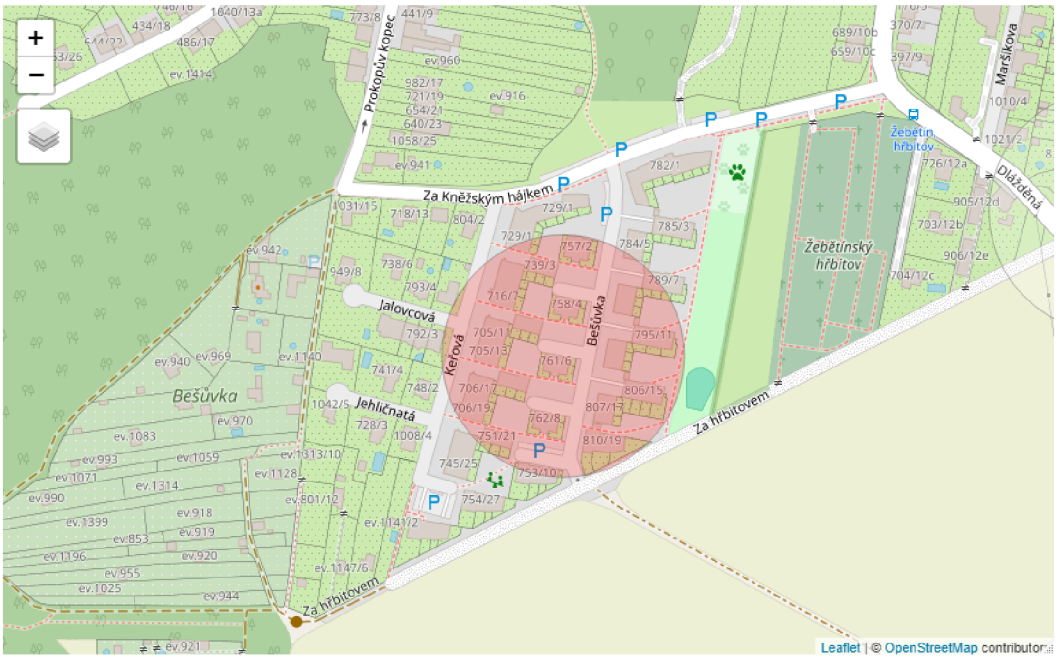
že výše škod těchto dvou nehod byla způsobená tím, že zde došlo ke střetům automobilů s traktorem nebo kombajnem (oba typy vozidel spadají do dražší kategorie). Z dat však bylo zjištěno, že tento předpoklad není správný, neboť jedna z nehod, nezaviněná řidičem, byla způsobená srážkou osobního automobilu bez přívěsu se zvířetem (dle dat z kategorie "lesní zvěří, domácím zvířectvem") v osm hodin ráno. Druhá nehoda, která se stala o půlnoci, se týkala srážky osobního automobilu bez přívěsu s nákladním automobilem vlivem nevěnování se řízení.

Předposlední zajímavá oblast *hot spotů*, znázorněná na mapě 3.24, se nachází převážně v katastru Kohoutovice. Při detailnějším zkoumání, o jaké oblasti se jedná, bylo zjištěno, že ve většině případů byla vysoká škodovost způsobená nehodou, jež se odehrála na parkovištích, resp. v jejich blízkosti. Proto se zde jako možné vysvětlení vysoké škody nabízí to, že se nejspíše jednalo o nehody vozidel dražší kategorie. Z dat vyplynulo, že tyto nehody byly často způsobeny vlivem nevěnování se řízení, kdy některé z nehod, které se nestaly přímo na parkovišti, byly způsobeny srážkou osobních automobilů bez přívěsu s nákladními automobily (z kategorie "nákladní automobil včetně multikáry, autojeřábu, cisterny atd."). Výjimkou, co se týče místa nehody, je *hot spot* v oblasti vyznačené jako Žebětínská (v katastru Žebětín). Zde se nehoda odehrála v místě prudké zatáčky, přičemž v jednom směru je silnice dvouproudová. Tato nehoda byla ve skutečnosti způsobená srážkou osobního automobilu s pevnou překážkou vlivem nepřizpůsobení rychlosti dopravně technickému stavu vozovky (zatáčka, klesání, stoupání, šířka vozovky apod.).



Obrázek 3.24: Detailní pohled na *hot spoty* nehod v katastru Kohoutovice

Poslední *hot spot*, jež tu zmíníme, je *hot spot* nacházející v katastru Žebětín, ve směru na západ od katastru Kohoutovice z obrázku 3.21. Jedná se, jak vidíme na mapě 3.25, o jednu nehodu na strmé ulici Bešůvka, kde se nachází bytové družstvo. Vzhledem k typu domů, které se zde nachází, a tomu, že podél ulice jsou vozidla parkována ve směru z kopce dolů, mohlo dojít k vysoké škodě vlivem nezajištění drahého typu vozidla, které mohlo sjet z kopce a v průběhu narazit do jiného vozidla. Ve skutečnosti se však jednalo o nehodu dvou osobních automobilů bez přívěsu, která byla způsobená vlivem nevěnování se řízení vozidla.



Obrázek 3.25: Detailní pohled na *hot spot* nehody na západě Brna

3.4. Hot spot analýza na reziduích prostorově vážené regrese

V této části byla použita data týkající se pouze nehod automobilů s jinými automobily v roce 2015. V prostorově vážené regresi byl v roli závislé proměnné použit atribut výše hmotné škody (v Kč) způsobené autonehodou. Jako nezávislé proměnné byly zvoleny dva kategoriální atributy - stav vozovky (suchý povrch, mokrý povrch apod.) a druh komunikace (místní, účelová apod.), a jeden kvantitativní atribut - průměrná délka dne. Tento atribut byl přidán dodatečně, kdy pro každý měsíc byl (s využitím informací z [39]) spočítán průměr hodin trvání dne (tj. doba mezi tím, kdy slunce vyšlo a zapadlo) prvního a posledního dne v měsíci.

Model prostorově vážené regrese v `R` voláme funkcí `gwr` z knihovny `spgwr` [32]. Náš model byl zadán ve tvaru

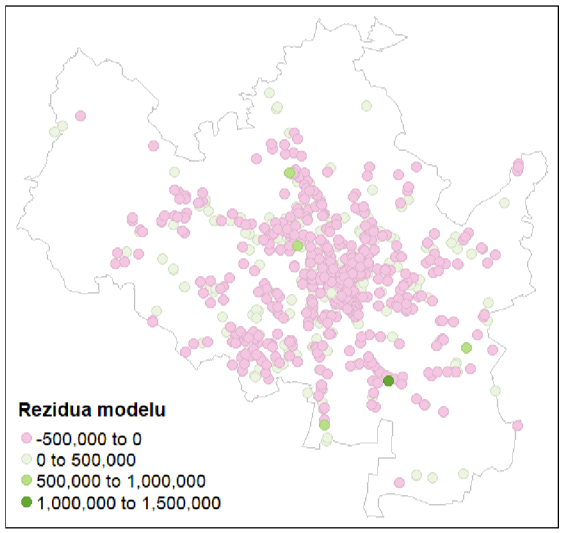
```
gwr.fit <- gwr(HMOTNA_SKODA ~ relevel(STAV_VOZOVKY,
           ref = "povrch mokrý") + relevel(DRUH_KOMUN,
           ref = "dálnice") + prum_delka_dne, data = data,
           adapt = band, hatmatrix = T),
```

kde bylo potřeba přidat funkci `relevel` k nastavení referenční kategorie, vzhledem k přítomnosti kategoriálních atributů. Šířka pásma adaptivního jádra `adapt = band` byla určena funkcí `gwr.sel` z knihovny `spgwr` ve tvaru

```
band <- gwr.sel(HMOTNA_SKODA ~ relevel(STAV_VOZOVKY, ref =
          "povrch mokrý") + relevel(DRUH_KOMUN, ref = "dálnice")
          + prum_delka_dne, data = data, adapt = T),
```

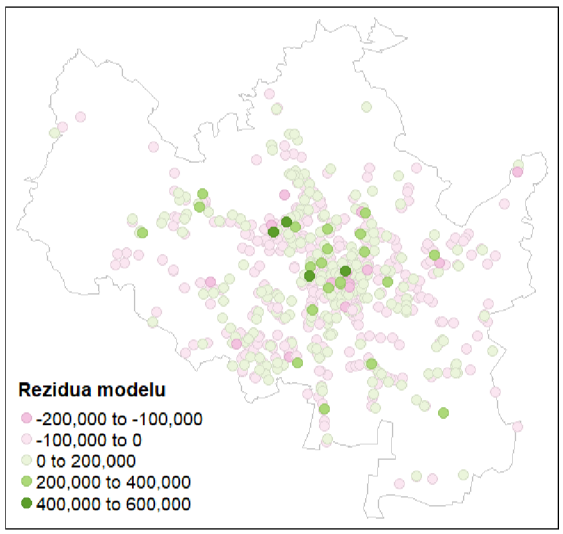
ve které bylo parametrem `adapt = T` zvoleno adaptivní jádro, jelikož hustota nehod je v různých oblastech Brna různá (na okrajích je nehod méně než v centru). Tato funkce zároveň používá k určení optimálního vyhlazovacího parametru metodu křížové validace, kterou nebylo potřeba nastavovat, jelikož jde o výchozí nastavení funkce.

Následně byla vykreslena rezidua tohoto modelu, která můžeme vidět na následujícím obrázku 3.26.



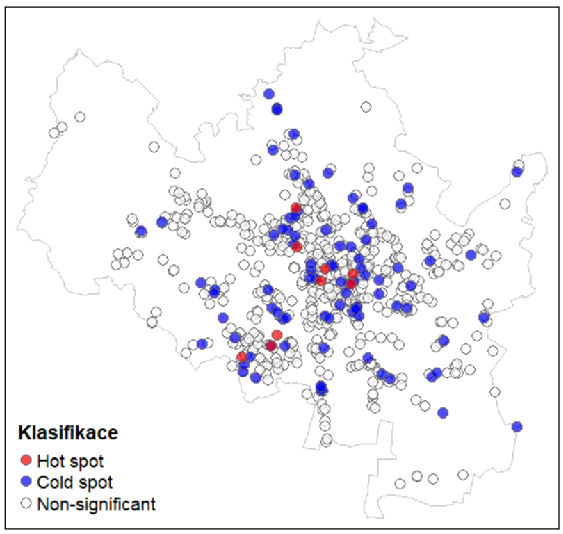
Obrázek 3.26: Rezidua modelu prostorově vážené regrese

Vidíme zde jednu odlehlou hodnotu v rozmezí $1,000,000$ a $1,500,000$ Kč, která značně ovlivňuje model, a proto byla odebrána. Nová rezidua napočítaná z modelu neobsahujícím odlehlou hodnotu, a která byla použita na *hot spot* analýzu, vidíme na obrázku 3.27, přičemž k nalezení optimální šířky pásma jádra bylo použito Akaikeho informační kritérium (bez této volby zde nebyly nalezeny žádné *hot spoty* reziduí), které bylo ve funkci `gwr.sel` nastaveno parametrem `method = "aic"`.



Obrázek 3.27: Rezidua modelu prostorově vážené regrese bez odlehlé hodnoty

Pro nalezení *hot spotů* reziduí z obrázku 3.27 byla, jako v předchozích případech, použita lokální G-statistika. Vzhledem k záporným hodnotám nebylo možné určit prahovou vzdálenost pro tvorbu váhové matice pomocí globální G-statistiky, a proto byla využita metoda k nejbližších sousedů, kdy byla zvolena taková vzdálenost, aby každý z bodů měl alespoň jednoho souseda. Výsledné *hot spoty* je možné vidět na obrázku 3.28 na následující straně.



Obrázek 3.28: *Hot spoty* reziduí modelu prostorově vážené regrese

Z tohoto obrázku 3.28 lze vidět, že výsledek není tak intuitivní, jako tomu bylo v předchozích částech analýzy. Vzhledem k tomu, že při různých volbách parametrů (jako je např. způsob určení vhodné šířky pásma jádra) bylo dosahováno pouze dvou výstupů - výstup, kdy žádné body nebyly identifikovány jako *hot spoty* ani *cold spoty*, nebo výstup, který je značně neintuitivní, se jako vysvětlení takového výsledku jeví, že mezi rezidui bude, v kontextu *hot spot* analýzy, slabý nebo žádný vztah. V každém případě, pokud bychom tento výstup chtěli interpretovat, řekli bychom, že nalezené *hot spoty* představují oblasti, v jejichž okolí model statisticky významně podhodnocuje výši hmotné škody (v Kč) způsobenou autonehodou, tedy, výše hmotné škody určená modelem zohledňujícím vliv stavu vozovky, denního světla a druhu komunikace na výši škody, byla v okolí dané oblasti nižší, než jaká byla ve skutečnosti. Důvodem vzniku/identifikace takovýchto oblastí mohou být faktory, které v modelu nejsou zahrnuty, přičemž toto zjištění nám pomáhá při vytváření, a následném optimalizování modelů používaných k predikcím, kdy na základě množství nalezených *hot spotů* se můžeme rozhodnout, zda do modelu při-

dáme i jiné faktory, či nikoliv (resp. zda po těchto faktorech hodláme pátrat). Celkově je ale potřeba brát v potaz to, že v místech identifikovaných jako *hot spoty* bude model výši škody podhodnocovat. U *cold spotů* by potom interpretace byla opačná.

Závěr

Cílem této diplomové práce bylo seznámit čtenáře s metodami využívanými při *hot spot* analýze dat, spolu s jejich následnou aplikací na reálných datech.

V první kapitole byly popsány pojmy, které jsou nutné pro pochopení metod prostorové analýzy dat, jako jsou například typy prostorových dat, vzdáleností, nebo váhových matic.

V rámci druhé kapitoly byly rozebrány metody prostorové analýzy, jež slouží k nalezení *hot spotů*, a to konkrétně metoda prostorové autokorelace (využívaná pro atributová bodová a polygonová data), kvadrátová analýza s metodou průměrné nejbližší vzdálenosti (pro neatributová bodová data) a prostorově vážená regrese, ve které se *hot spot* analýza provádí na reziduích modelu prostorově vážené regrese, na která lze, v kontextu této práce, pohlížet jako na atributová bodová data.

Třetí kapitola je věnována uplatnění teoretických poznatků na datech týkající se nehodovosti na území města Brna. Nejprve byla provedena *hot spot* analýza na polygonových datech s atributovou informací o počtu nehod v jednotlivých polygonech ve zvoleném roce 2015, při které byly *hot spoty* nehod nalezeny převážně v centru města Brna. Následovala analýza neatributových bodových dat, v rámci které jsme měli k dispozici informaci pouze o místě nehody, přičemž jsme dospěli ke stejnému závěru jako v předchozím případě. Další část *hot spot* analýzy se týkala atributových bodových dat, kdy roli atributu zastupovala výše hmotné škody (v Kč) způsobená nehodou v roce 2015. V tomto případě byly *hot spoty* výše hmotné škody (v Kč) nalezeny na okrajích území města Brna, nikoliv v centru. Poslední část *hot spot* analýzy byla provedena na reziduích z modelu prostorově vážené regrese. V roli závislé proměnné zde byl použit atribut výše hmotné škody způsobené

nehodou (v Kč) v roce 2015, v roli nezávislých proměnných byly použity dva kategoriální atributy - stav vozovky a druh komunikace, a jeden kvantitativní atribut - průměrná délka dne. Ačkoliv bylo nalezeno pár *hot spotů* reziduí, oproti předchozím částem analýzy výsledek tolik neodpovídal tomu, co bychom očekávali, což mohlo být způsobeno tím, že mezi rezidui není, v kontextu *hot spot* analýzy, žádný vztah, nebo je mezi nimi tento vztah pouze slabý.

Téma diplomové práce bylo pro mě osobně velkým přínosem především co se týče rozšíření si obzorů v oblasti analýzy dat, jelikož se jedná o téma, které pro mě bylo donedávna neznámé. Současně jsem v rámci práce získala možnost vyzkoušet knihovnu `tmap` [34] pro tvorbu map, a rozšířit tak své schopnosti grafické reprezentace dat.

Literatura

[1] Anselin, L.: *Applications of Spatial Weights.* [online], [cit. 2024-03-10]. Dostupné z:

`https://geodacenter.github.io/workbook/4d_weights_applications/lab4d.html`.

[2] Anselin, L.: *Distance-Band Spatial Weights.* [online], [cit. 2023-10-30]. Dostupné z:

`https://geodacenter.github.io/workbook/4b_dist_weights/lab4b.html#fn1`.

[3] Anselin, L.: *Local Spatial Autocorrelation (1).* [online], [cit. 2023-11-15]. Dostupné z:

`https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html`.

[4] Barreca, A., Curto, R., A., Rolando, D.: *Hotspot Analysis Using ArcGIS.* [online], [cit. 2023-10-31]. Dostupné z:

`https://glenbambrick.com/tag/spatial-autocorrelation/`.

[5] Bivand, S., R., Pebesma, E., Gómez-Rubio, V.: *Applied Spatial Data Analysis with R (second edition).* Springer Science & Business Media, 2013.

[6] Bourke, P.: *Polygons and meshes.* [online], [cit. 2023-11-14]. Dostupné z:

`https://paulbourke.net/geometry/polygonmesh/`.

[7] Brundson, C., Comber, L.: *An Introduction to R for Spatial Analysis and Mapping.* SAGE, 2019.

[8] Chen, D., Getis, A.: *POINT PATTERN ANALYSIS (PPA).* [online], [cit. 2024-02-01]. Dostupné z:

`https://www.nku.edu/~longa/geomed/ppa/doc/html/ppa.html`.

[9] Chen, Y.: *New Approaches for Calculating Moran's Index of Spatial Autocorrelation.* [online], [cit. 2024-02-03]. Dostupné z:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068336.

[10] Cipra, T.: *Analýza časových řad s aplikacemi v ekonomii.* SNTL, Alfa, Praha, 1986.

[11] *Complete spatial randomness.* [online], [cit. 2023-11-16]. Dostupné z:

https://www.paulamoraga.com/book-spatial/complete-spatial-randomness.html.

[12] Ebert, E., C., Prufer, K., M., Macri, M., J., Winterhalder, B., Kennett, D., j.: *TERMINAL LONG COUNT DATES AND THE DISINTEGRATION OF CLASSIC PERIOD MAYA POLITIES.* [online], [cit. 2023-11-16]. Dostupné z:

https://escholarship.org/content/qt2312p0h9/qt2312p0h9_noSplash_757c62e46cad27ceac9f610f220a3fda.pdf?t=o06jrk.

[13] *Expected value of global Moran index.* [online], [cit. 2023-11-16]. Dostupné z:

https://stats.stackexchange.com/questions/567411/expectation-and-variance-of-morans-i-under-the-null.

[14] Fišerová, E.: *Lineární statistické modely (2. vydání).* Univerzita Palackého v Olomouci, Olomouc, 2015. Skripta.

[15] Fotheringham, S., A., Brunsdon, C., Charlton, M.: *Geographically Weighted Regression - the Analysis of Spatially Varying Relationships.* John Wiley & Sons Ltd, London, 2002.

[16] Gimond, M.: *Point pattern analysis in R.* [online], [cit. 2024-03-24]. Dostupné z:

https://mgimond.github.io/Spatial/point-pattern-analysis-in-r.html.

[17] Hrčková, L.: *Katastr nemovitostí.* [online], [cit. 2024-03-24]. Dostupné z:

https://gis.brno.cz/ost/edas/public/eb94d941-e313-41ac-865f-e8534fd1d0db.

[18] Hron, K., Kunderová, P., Vencálek, O.: *Základy počtu pravděpodobnosti a metod matematické statistiky (3. přepracované vydání).* Univerzita Palackého v Olomouci, Olomouc, 2018. Skripta.

[19] Kalinic, M., Krisp, M., J.: *Kernel Density Estimation (KDE) vs. Hot-Spot Analysis - Detecting Criminal Hot Spots in the City of San Francisco.* [online], [cit. 2024-03-20]. Dostupné z:

`https://www.researchgate.net/publication/325825793_Kernel_Density_Estimation_KDE_vs_Hot-Spot_Analysis_-_Detecting_Criminal_Hot_Spots_in_the_City_of_San_Francisco`.

[20] Komínek, J.: *Dopravní nehody / Traffic accidents.* [online], [cit. 2024-03-02]. Dostupné z:

`https://data.brno.cz/datasets/mestobrno::dopravn%C3%AD-nehody-traffic-accidents/about`.

[21] Lansley, G., Cheshire, J.: *An Introduction to Spatial Data Analysis and Visualisation in R.* CDRC Learning Resources, 2016.

[22] Lee, J., Wong, D. W. S.: *Statitical Analysis with ArcView GIS.* John Wiley, New York, 2001.

[23] Li, Y., Zhang, L., Yan, J., Wang, P., Hu, N., Cheng, W., Fu, B.: *Mapping the hotspots and coldspots of ecosystem services in conservation priority setting.* [online], [cit. 2024-03-20]. Dostupné z:

`https://www.researchgate.net/publication/308338921_Mapping_the_hotspots_and_coldspots_of_ecosystem_services_in_conservation_priority_setting#pf12`.

[24] Long, J.: *Global Spatial Autocorrelation.* [online], [cit. 2023-10-18]. Dostupné z:

`https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html`.

[25] Macků, K.: *Pokročilé zpracování geodat.* Univerzita Palackého v Olomouci, Olomouc, 2023.

[26] Mathologer: *Gauss's magic shoelace area formula and its calculus companion.* [online], [cit. 2024-02-10]. Dostupné z:

`https://www.youtube.com/watch?v=0KjG8Pg6LGk&t=328s`.

[27] McKenzie, H.: *How to calculate spatial hotspots.* [online], [cit. 2023-10-17]. Dostupné z:

`https://carto.com/blog/spatial-hotspot-tools`.

[28] NNI: *Nearest neighbour methods.* [online], [cit. 2024-3-17]. Dostupné z:

`https://www.spatialanalysisonline.com/HTML/nearest_neighbor_methods.htm`.

[29] Package 'sp': *Classes and Methods for Spatial Data.* [online], [cit. 2024-04-02]. Dostupné z:

`http://cran.nexr.com/web/packages/sp/sp.pdf`.

[30] Package 'spatstat': *Spatial Point Pattern Analysis, Model-Fitting, Simulation, Tests.* [online], [cit. 2024-04-02]. Dostupné z:

`https://cran.r-project.org/web/packages/spatstat/spatstat.pdf`.

[31] Package 'spdep': *Spatial Dependence: Weighting Schemes, Statistics.* [online], [cit. 2024-04-02]. Dostupné z:

`https://cran.r-project.org/web/packages/spdep/spdep.pdf`.

[32] Package 'spgwr': *Geographically Weighted Regression.* [online], [cit. 2024-04-02]. Dostupné z:

`https://cran.r-project.org/web/packages/spgwr/spgwr.pdf`.

[33] Package 'stats': *The R Stats Package.* [online], [cit. 2024-04-02]. Dostupné z:

`https://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html`.

[34] Package 'tmap': *Thematic Maps.* [online], [cit. 2024-04-02]. Dostupné z:

`https://cran.r-project.org/web/packages/tmap/tmap.pdf`.

[35] Scott, W., D.: *Multivariate density estimation: Theory, Practise, and Visualization (second edition).* Willey, New York, 1992.

[36] Silva, W., J., F., Souza, M., C., R., Cysneiros, F., J. A.: *Polygonal data analysis: A new framework in symbolic data analysis.* [online], [cit. 2024-3-10]. Dostupné z:

`https://www.sciencedirect.com/science/article/pii/S0950705118304052`.

[37] *Spatial autocorrelatione.* [online], [cit. 2023-10-31]. Dostupné z:

`https://manuals.pqstat.pl/en:przestrzenpl:autocorpl`.

[38] Spatial Statistics for Data Science: Theory and Practice with R: *Spatial autocorrelation* [online], [cit. 2024-03-15]. Dostupné z:

`https://www.paulamoraga.com/book-spatial/spatial-autocorrelation.html`.

[39] Sunrise & Sunset: *Sun in European district.* [online], [cit. 2024-2-17]. Dostupné z:

`https://www.timeanddate.com/sun/@11550319?month=3&year=2015`.

[40] Winkler, A., M., Taylor, P., A., Nichols, T., E., Rorden, C.: *False Discovery Rate and Localizing Power.* [online], [cit. 2023-11-14]. Dostupné z:

`https://arxiv.org/html/2401.03554v1`.

[41] Wolf, L., J.: *Confounded Local Inference: Extending Local Moran Statistics to Handle Confounding.* School of Geographical Sciences, University of Bristol, 2023.