Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra geoinformatiky

APLIKACE VYHLEDÁVANÍ KOLOKAČNÍCH VZORŮ NA PROSTOROVÁ DATA

Diplomová práce

Simona BUČKOVÁ

doc. Ing. Zdena Dobešová, Ph.D.

Olomouc 2022
Geoinformatika a kartografie

ANOTÁCIA

Diplomová práca sa zaoberá vyhľadávaním kolokačných vzorov v priestorových dátach. Kolokačný vzor je jednou z úloh data miningu, pomáha získavať nové poznatky v dátach, ktoré nie sú na prvý pohľad zrejmé. Úloha je obsiahnutá v nástroji Colocation Analysis v programe ArcGIS Pro. Vstupnou vrstvou do nástroja je povolená len bodová reprezentácia, avšak je rozdiel, ktorá kategória je označená ako skúmaná, a ktorá je označená ako susedná. Dôležitou súčasťou celého procesu je nastavenie parametrov, ktoré výrazne ovplyvňuje výsledok nástroja. Súčasťou práce sú tri prípadové štúdie s použitím rôznych dát, a to nie len charakterom, ale aj geografickou polohou. Vďaka ich odlišnosti boli v práci otestované viaceré nastavenia nástroja a spôsoby tvorby vzťahu medzi skúmanými kategóriami. Okrem GIS softwaru bol využitý aj data miningový program Orange pre tvorbu Python scriptu v prostredí editoru PyScripter. Skrip bol nasadený do prostredia ArcGIS Pro a jeho cieľom bolo uľahčenie generovania kolokačných vzorov a zníženie časovej náročnosti prvej prípadovej štúdie. V závere práce bol spísaný praktický návod vychádzajúci z tretej štúdie o pirátskych útokoch v okolí Arabského polostrova. Posledným krokom bola tvorba posteru a webovej stránky dokumentujúcej priebeh a výsledky diplomovej práce.

KĽÚČOVÉ SLOVÁ

Kolokačný vzor; priestorové dáta; data mining; analýza

Počet strán práce: 68
Počet príloh: 4 (z toho 3 voľné a 1 elektronická)

ANOTATION

The Master's thesis deals with the search for colocation patterns in spatial data. The colocation pattern is one of the tasks of data mining, it helps to gain new knowledge in data that is not obvious at first view. The task is included in the Colocation Analysis tool in the ArcGIS Pro. Input layer of the tool is allowed only the point representation, however there is the difference which category is marked as examined and which is marked as neighbouring. An important part of the whole process is the setting of parameters which boldly affects the result of the tool. The diploma includes three case studies using different data, not only in character but also in geographical location. Due to their differences, several tool settings, and ways of creating a relationship between the examined categories. In addition to GIS software, the data mining program Orange was used to create a Python script in the PyScripter editor environment. The script was deployed in the ArcGIS Pro environment and its goal was to facilitate the generation of colocation patterns and reduce the time demandingness for the first case study. At the end of the diploma was written a practical guide based on the third study on pirate attacks around the Arabian Peninsula. The last step was the creation of a poster and website documenting the process and results of the thesis.

KEYWORDS

Colocation pattern; spatial data; data mining; analysis

Number of pages: 68
Number of appendixes: 4

Prohlašuji, že

- diplomovou práci včetně příloh, jsem vypracovala samostatně a uvedla jsem všechny použité podklady a literaturu.

- jsem si vědoma, že na moji diplomovou práci se plně vztahuje zákon č.121/2000 Sb. - autorský zákon, zejména § 35 – využití díla v rámci občanských a náboženských obřadů, v rámci školních představení a využití díla školního a § 60 – školní dílo,

- beru na vědomí, že Univerzita Palackého v Olomouci (dále UP Olomouc) má právo nevýdělečně, ke své vnitřní potřebě, diplomovou práci užívat (§ 35 odst. 3),

- souhlasím, aby jeden výtisk diplomové práce byl uložen v Knihovně UP k prezenčnímu nahlédnutí,

- souhlasím, že údaje o mé diplomové práci budou zveřejněny ve Studijním informačním systému UP,

- v případě zájmu UP Olomouc uzavřu licenční smlouvu s oprávněním užít výsledky a výstupy mé diplomové práce v rozsahu § 12 odst. 4 autorského zákona,

- použít výsledky a výstupy mé diplomové práce nebo poskytnout licenci k jejímu využití mohu jen se souhlasem UP Olomouc, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly UP Olomouc na vytvoření díla vynaloženy (až do jejich skutečné výše).

V Olomouci dne Simona Bučková

Ďakujem vedúcej práce doc. Zdeně Dobešové za všetky cenné rady a za pomoc počas písania diplomovej práce. Taktiež chcem poďakovať mojej rodine a blízkym za podporu.

UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Akademický rok: 2020/2021

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Simona BUČKOVÁ**
Osobní číslo:	**R200035**
Studijní program:	**N0532A330009 Geoinformatika a kartografie**
Studijní obor:	**Geoinformatika a kartografie**
Téma práce:	**Aplikace vyhledávání kolokačních vzorů na prostorová data**
Zadávající katedra:	**Katedra geoinformatiky**

Zásady pro vypracování

Cílem práce je aplikovat na vhodná data postup vyhledávání kolokačních vzorů v prostorových datech. Student vybere a navrhne tři případové studie, na kterých demonstruje postup a nasazení nástroje ArcGIS Colocation Analysis. Práce bude prezentovat kolokační vzory jevů v území s testováním vhodných parametrů. Jedna vybraná případová studie bude zdokumentována formou stručného návodu včetně vysvětlení metody colocation patterns.

Celá práce (text, přílohy, výstupy, zdrojová a vytvořená data) se odevzdá v digitální podobě na paměťovém nosiči (CD, DVD, SD karta, flash disk). Text práce s vybranými přílohami bude odevzdán ve dvou svázaných výtiscích na sekretariát katedry. O diplomové práci student vytvoří webovou stránku v souladu s pravidly dostupnými na stránkách katedry. Práce bude zpracována podle zásad dle Voženílek (2002) a závazné šablony pro diplomové práce na KGI. Povinnou přílohou práce bude poster formátu A2.

Rozsah pracovní zprávy:	**max. 50 stran**
Rozsah grafických prací:	**dle potřeby**
Forma zpracování diplomové práce:	**tištěná**
Jazyk zpracování:	**Slovenština**

Seznam doporučené literatury:

BARUA S., SANDER J. Statistically Significant Co-location Pattern Mining. In: Shekhar S., Xiong H., Zhou X. (eds) Encyclopedia of GIS. Springer, Cham. 2017, https://doi.org/10.1007/978-3-319-17885-1_1552

HU W. Co-location Pattern Discovery. In: Shekhar S., Xiong H., Zhou X. (eds) Encyclopedia of GIS. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-17885-1_150

MAITI S., SUBRAMANYAM R.B.V. Mining co-location patterns from distributed spatial data. Journal of King SaudUniversity– Computer and Information Sciences, 2018, https://doi.org/10.1016/j.jksuci.2018.08.010

MAMOULIS, N. Co-location Pattern. In: Shekhar S., Xiong H., Zhou X. (eds) Encyclopedia of GIS. Springer, Cham. 2017, https://doi.org/10.1007/978-3-319-17885-1_149

PAVEL, P. Metody Data Miningu, část 2. Pardubice: Univerzita Pardubice, Fakulta ekonomicko-správní, 2014

TRNOVÁ, L. Aplikace asociačních pravidel na prostorová data, Univerzita Palackého v Olomouci, Katedra geoinformatiky PřF, diplomová práce, 2020

VOŽENÍLEK, V. Diplomové práce z geoinformatiky. Olomouc, Univerzita Palackého v Olomouci, 2002

Vedoucí diplomové práce: **doc. Ing. Zdena Dobešová, Ph.D.**
Katedra geoinformatiky

Datum zadání diplomové práce: **9. listopadu 2020**
Termín odevzdání diplomové práce: **6. května 2022**

L.S.

doc. RNDr. Martin Kubala, Ph.D.
děkan

prof. RNDr. Vít Voženílek, CSc.
vedoucí katedry

V Olomouci dne 14. prosince 2020

OBSAH

ZOZNAM POUŽITÝCH SKRATIEK

Skratka	Význam
aprx	ArcGIS projekt
CSV	Comma-separated values
GIS	geografický informačný systém
LCLQ	lokálny kolokačný kvocient
NACE	Európska klasifikácia ekonomických činností
NUTS	Nomenklatúra územných štatistických jednotiek
PI	Participation Index
SHP	Shapefile
swm	matica priestorových dát
UTM	Universal Transverse Mercator
WGS	World Geodetic System

ÚVOD

Kolokačný vzor patrí medzi úlohy data miningu, vďaka nemu je možné zisťovať medzi dátami vzťahy, ktoré nemusia byť na prvý pohľad zrejmé. V procese získavania nových poznatkov sa pracuje s priestorovými dátami, ich samotná poloha je v procese kľúčová. Teória kolokačného vzoru je obsiahnutá v nástroji Colocation Analysis v prostredí ArcGIS Pro. Výhodou tohoto nástroja je fakt, že vstupné dáta nie je potrebné zdĺhavo predpripravovať, avšak malá úprava dát bola vykonaná.

V teoretickej časti práce je vysvetlený základný princíp, možný vývoj tejto oblasti a samotný nástroj. Práca ďalej obsahuje tri prípadové štúdie, na ktorých bol nástroj nasadený a na základe charakteru dát boli otestované jeho možnosti. Súčasťou diplomovej práce sú aj mapové výstupy pre jednotlivé štúdie a na uľahčenie práce bol vytvorený pomocný Python script, ktorý bol nasadený na prvú štúdiu. Skript je možné spustiť len pre software ArcGIS Pro, z dôvodu absencie nástroja v iných programoch. Ďalším, no nie posledným výstupom je návod, v ktorom je podrobne spísaný postup práce od ľahkej úpravy dát až po konečnú vizualizáciu. Konečným výstupom je poster zhrňujúci obsah diplomovej práce vo formáte A2 a webová stránka.

1 CIELE PRÁCE

Cieľom diplomovej práce je zoznámiť sa s úlohou dolovania dát, a tým je kolokačý vzor. Následne po preštudovaní dostupnej literatúry nasadiť nástroj Colocation Analysis na tri rôzne typy dát. Vďaka ich rozdielnosti budú otestované široké možnosti nástroja a možnosti konštruovania kolokačných vzorov.

V teoretickej časti bude predstavená podstata kolokačného vzoru ako data miningová úloha získavania nových, zaujímavých znalostí z dát. Následne budú spomenuté oblasti využitia možné rozšírenia a zlepšenia v budúcnosti. Ďalej bude teoretická časť obsahovať rozbor samotného nástroja, jeho parametrov a nastavení. Tie budú kľúčové v praktickej časti práce.

V praktickej časti dôjde k nasadeniu spomínaného nástroja na tri rôzne dáta, či už pochádzajúce z Česka, alebo zo zahraničných krajín. V rámci prvej štúdie bude vypracovaný aj Python skrip na uľahčenie práce pri generovaní výstupných kolokačných vzorov. Rozdielnosť jednotlivých štúdií nebude len v ich zemepisných súradniciach, ale aj parametrizácii a zameraní nástroja kolokačnej analýzy. Každá z prípadových štúdií bude obsahovať aj mapové výstupy zachytávajúce významné či zaujímavé výsledky.

Výsledkom práce bude návod vychádzajúci z jednej zo spracovaných štúdií. Jeho obsahom bude drobná predpríprava dát, nasadenie nástroja a konkrétnych hodnôt parametrov. V neposlednej časti bude návod obsahovať vyhodnotenie výsledných kolokačných vzorov a myšlienky, ako môže užívateľ výsledky interpretovať. Hlavným grafickým výstupom bude poster zachytávajúci postup práce a výsledky jednej z vypracovaných štúdií. Cieľom práce nie je poskytnúť konkrétne postupy a hodnoty pri použití kolokačného nástroja, ale ujasniť jeho princíp a možnosť využitia.

2 METÓDY A POSTUPY SPRACOVANIA

Nájdenie vhodných hodnôt parametrov a nastavení nástroja Colocation Analysis so sebou nesie niekoľko fáz. V prvom rade ide o uvedomenie charakteru dát, atribútového naplnenia, priestorového rozsahu, existencie časového aspektu a podobne. Tieto poznatky sú kľúčové pri úprave dát a nastavení konkrétnych parametrov a metód kolokačného nástroja Colocation Analysis. Súčasťou tejto kapitoly je aj zmienka použitých programov, ktoré napomohli ku získaniu dát či kolokačných vzorov.

Použité metódy

Majoritným nástrojom v tejto práci je **kolokačná analýza** *(Colocation Analysis)*, ktorá je predmetom skúmania. Jej podrobný popis je v rešeršnej kapitole. V rámci práce bolo využívaných viacero priestorových metód či nástrojov, predovšetkým pri predpríprave dát. Veľmi častou bola dvojica nástrojov **výberu prvkov podľa polohy a atribútov** *(Select By Location, Select By Attributes)*. Nástroje fungujú tak, že dochádza k subvýberu z množiny dát (v tomto prípade bodov) spĺňujúcich podmienku založenú na polohe, alebo hodnote daného atribútu.

Pre vyšetrenie okrajových častí skúmaného územia bola aplikovaná tvorba **obalovej zóny** *(Buffer)*. Nástroj spadá do kategórie analýzy blízkosti a s jeho pomocou je možné vytvoriť zónu v definovanej vzdialenosti od vstupného prvku. Výsledkom je polygónová vrstva, ktorá je od všetkých vstupných prvkov vzdialená o vopred definovanú vzdialenosť.

V rámci prvej štúdie bol skúmaný vplyv duplicitných bodov na výsledný kolokačný vzor. Vzhľadom na to, že išlo o cielený výber ponechaných bodov, bola v prvom kroku zostavená štvorcová sieť o veľkosti 30 x 30 m pomocou nástroja **generovať grid z oblasti** *(Generate Grid From Area)*. Následne bol spočítaný počet prvkov pre jednotlivé štvorce siete, bol použitý nástroj **sumarizácie medzi** *(Summarize Within)*. Po priblížení na konkrétnu oblasť bolo možné pohľadom spočítať počet bodov, a tým aj odhaliť duplicitné body nachádzajúce sa nad sebou. Tie boli potom jednoducho odstránené označením v atribútovej tabuľke a následným odstránením záznamu. Pre predstavu o vzdialenosti medzi skúmanými záznamami bol častokrát aplikovaný nástroj **merania** *(Measure)*, ktorý okrem vzdialenosti dokáže vypočítať aj uhol medzi prvkami.

Dátová sada, na ktorej bola postavená druhá štúdia obsahovala súradnice, ktoré bolo potrebné pre ďalšiu prácu zobraziť. Preto jedným z prvých nástrojov bolo zobrazenie súradníc **xy v tabuľke do bodov** *(XY Table To Point)*. Každopádne bolo nutné previesť súradnice do vhodného súradnicového systému, na to poslúžil nástroj **konverzovania súradnicového zápisu** *(Convert Coordinate Notation)*. Podstatou tejto štúdie bolo preskúmať časový aspekt v kolokačnej analýze a jej vplyv na výsledný kolokačný vzor. Vstupné dáta obsahovali dátum, avšak stĺpec bol uložený ako text namiesto dátumu. Preto pomocou nástroja **konvertovať časové pole** *(Convert Time Field)* bol vytvorený nový stĺpec so správnym dátovým typom (dátum) a skopírovaný dátum cez **vypočítanie poľa** *(Calculate Field)*. Ďalej bolo potrebné vybrať záznamy v určitom časovom okne. Na to je v programe ArcGIS Pro dostupný nástroj **výber vrstvy podľa dátumu a času** *(Select Layer By Date and Time)*. Spadá pod sadu nástrojov pre analýzu kriminality a bezpečnosti a vyberie záznamy podľa zvoleného času či dátumu.

V rámci poslednej prípadovej štúdie boli využívané už spomenuté nástroje. Bolo nutné zmeniť súradnicový systém a vytvoriť ďalšie pole. Doposiaľ nevyužitým nástrojom je **generovanie priestorovej matice váh** *(Generate Spatial Weights Matrix)*, ktorá má

niekoľko možností tvorby priestorového či časového vzťahu medzi skúmanými prvkami analýzy. Nástroj vytvorí súbor váh, ktorý následne vstupuje do kolokačnej analýzy.

Skriptovanie

V rámci prvej štúdie bol pre jej rýchlejší priebeh vytvorený skript vychádzajúci z pôvodného scriptu Colocation Analysis. Kód bol písaný v prostredí PyScripter, ktorý sa vyznačuje svojou jednoduchosťou a prehľadnosťou. Python je veľmi populárny jazyk čo sa týka programovania rozličných programov či nástrojov. Obľúbený je predovšetkým kvôli jeho jednoduchej syntaxi.

Použité dáta

Táto čas obsahuje základné informácie o doplňujúcich dátových sadách, ktoré boli v rámci diplomovej práce využité. Primárne dátové sady, na ktorých sú založené vypracované štúdie sú uvedené v rámci príslušnej prípadovej štúdie.

Časti obce

Polygónová vrstva pochádza z digitálnej vektorovej geodatabázy ArcČR® vo verzii 3.3. Databázu poskytuje spoločnosť ARCDATA PRAHA s.r.o. Vrstva obsahuje časti jednotlivých obcí Českej republiky. Pre potreby práce bola využitá len časť obce Brna, konkrétne Brno – město (ARCDATA PRAHA, 2021).

OSM

Celým názvom OpenStreetMap je projekt, ktorý je založený na voľnom poskytovaní priestorových dát, ktoré sú v konečnej forme zobrazené ako topografické mapy. Dostupnú vektorovú databázu môže obohacovať a editovať každý užívateľ so zriadeným účtom. K dátam sa je možné dopracovať viacerými spôsobmi. Pre účely tejto práce bola využitá extenzia QuickOSM dostupná v programe QGIS. Podrobnejšie informácie ku konkrétnym dátovým vrstvám sú dostupné v rámci prípadovej štúdie.

Policajné oblasti

Mesto Filadelfia disponuje katalógom OpenDataPhilly, kde je dostupných veľké množstvo priestorových dát. Jedným z nich sú policajné oblasti, ktoré boli využité v druhej prípadovej štúdií pri interpretácií a popise výsledkov. Mesto je delené na šesť policajných oblastí podľa svetových strán, tie sa ešte delia na menšie časti. Označujú sa číselným kódom (OpenDataPhilly, 2014).

Podkladová mapa

Vo výsledných mapách štúdií boli použité dostupné topografické mapy v programe ArcGIS Pro. Pre Českú republiku, respektíve jej časť bola využitá Light Gray Canvas. V druhej štúdií o Filadelfii tvorila podklad World Topografic Map a v tretej taktiež World Topografic Map. Podkladovým mapám bola nastavená priehľadnosť pre zvýraznenie kolokačného vzoru, ktorý je dominantným prvkom mapy.

Použité programy

ArcGIS Pro 2.8.3

Najvyužívanejší GIS program v rámci celej diplomovej práce. Jeho tvorcom je Esri, slúži na skúmanie, vizualizáciu, analýzu, tvorbu máp a zdieľanie práce. Najpodstatnejšou súčasťou je nástroj Colocation Analysis, na ktorom je samotná práca postavená. Okrem toho v programe vznikali aj výsledné mapové výstupy jednotlivých prípadových štúdii. Netreba opomenúť aj drobnú predprípravu dát (Esri, 2022).

QGIS Desktop 3.16

Významná alternatíva k platenému ArcGIS Pro. Bol vytvorený ako projekt Open Source Geospatial Foundation (OSGeo) a je možné ho používať na rôznych operačných systémoch, ako je Linux, Unix, Mac OSX Windows ci Android. Program bol použitý na získanie dát z OSM, konktrétne zásuvný modul QickOSM. Stačí zadať kľúč a jeho hodnotu a dáta sa nahrajú medzi vrstvy projektu. Tie je možné následne vyexportovať, QGIS ponúka veľké množstvo formátov, a nahrať do iného prostredia (QGIS, 2022).

Orange 3.30.2

Voľne dostupný data miningový software s intuitívnym ovládaním. Jeho počiatky siahajú až do roku 1996, napísaný je v Python, Cython, C++ a C. Bol použitý ako zdroj dát kriminality vo Filadelfii. Dataset je voľne dostupný a je možné ho exportovať vo formáte CSV *(Comma – separated values)* a nahrať do prostredia ArcGIS Pro (Orange, 2022).

PyScripter

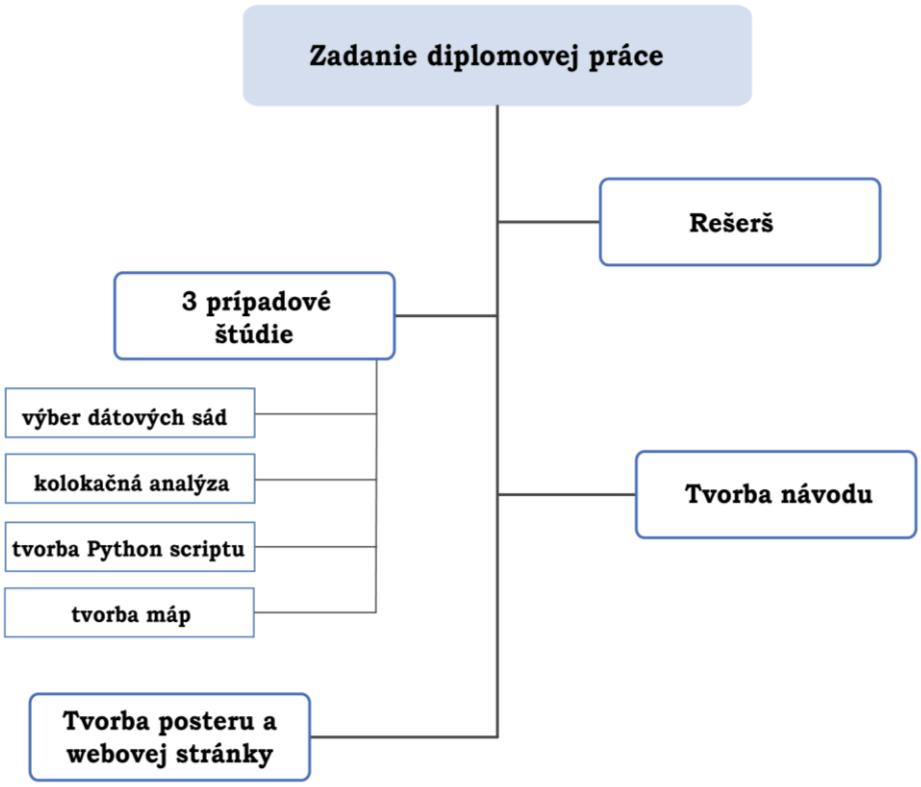
Open source prostredie vytvorené a cieľom konkurencieschopnosti s obdobnými komerčnými programami. Využitie našiel pri tvorbe Python scriptu v rámci prvej prípadovej štúdie. Program je veľmi jednoduchý s intuitívnym ovládaním (SOURCEFORGE, 2022).

Postup spracovania

Diagram na obrázku 1 popisuje postup pri diplomovej práci. Prvým a nedeliteľným krokom je rešerš literatúry, prevažne odborných článkov a webových stránok vrátane dokumentácie nástroja. Rešerš neslúži len na zoznámenie sa s tematikou práce, ale aj s možnosťami jej rozšírenia a úpravou vstupných dát pri tvorbe kolokačných vzorov. Vďaka študovaniu dostupnej literatúry je možné vybrať vhodné dátové sady, v dostupných článkoch sa častokrát spomínajú konkrétne príklady, na ktorých bol princíp použitý.

Praktickou časťou sú myslené tri prípadové štúdie, na ktorých bude nasadený nástroj Colocation Analysis. Dôležitým krokom v rámci procesu je nastavenie parametrov a metód, z ktorých je následne kolokačný vzor vytvorený. Pred samotným spustením nástroja je ešte potrebné vstupné dáta mierne poupraviť s pomocou priestorových analýz či nástrojov viď Použité metódy. Súčasťou prvej štúdie je aj vytvorenie skriptu v jazyku Python, ktorý má za úlohu uľahčiť a časovo skrátiť generovanie výsledných kolokačných vrstiev. Okrem toho bude vytvorený jednotný mapový štýl pre vizualizáciu dôležitých či zaujímavých výsledkov.

V závere bude napísaný návod, bude vychádzať z jednej z vytvorených prípadových štúdií. Oproti klasickému návodu či popisu nástroja bude vychádzať z konkrétneho príkladu, dáta bude možno stiahnuť a postup zopakovať. Posledným krokom je tvorba posteru a webovej stránky, informujúcich ako o priebehu, tak aj výsledkoch diplomovej práce.



Obr. 1 Vývojový diagram postupu práce (zdroj: autorka)

3 SÚČASNÝ STAV RIEŠENEJ PROBLEMATIKY

Data Mining, inak nazývaný dolovanie dát predstavuje pojem zastrešujúci širokú škálu techník v rade odvetví. V jednoduchosti je možné povedať, že umožňuje automaticky objavovať v dátach strategické informácie pomocou špeciálnych algoritmov. Predmetom práce je štúdium kolokačného vzoru, je to jedna z metód dolovania dát, ktorá umožňuje nájsť nové a zaujímavé priestorové vzťahy v dátových sadách (Petr, 2014).

Kolokačný vzor a získavanie pravidiel sú súčasťou procesu dolovania priestorových dát. Základnými rozdielmi medzi klasickým dolovaním a dolovaním priestorových dát je ich povaha, štatistické základy, výstupné vzory a výpočtový proces. Úspech v tejto oblasti je najmä zameraný na kategóriu výstupných vzorov, špeciálne na prediktívne modely, odľahlé hodnoty, pravidlá priestorového spoločného umiestnenia a klastre (Shekhar a kol. 2003). Prezentovaný výskum rozpoznávania priestorových vzorov zameraný na spoločné umiestnenie je najčastejšie označovaný ako objavovanie priestorového kolokačného vzorca a pravidiel spoločného umiestnenia.

Prvým definovaným slovom je booleovský priestorový prvok, je to typ geografického prvku. Môžu byť prítomné na rôznych miestach v dvojrozmernom (alebo vyššom) metrickom priestore, alebo úplne chýbajú (Shekhar a kol. 2003). Viaceré príklady booleovských prvkov sú kategorizované, ako napríklad druhy rastlín, živočíšne druhy, rakovinové bunky, zločiny, typy podnikaní a podobne. Ďalším definovaným termínom sú vzory a pravidlá spoločného umiestnenia. Vzory priestorového spoločného umiestnenia predstavujú podmnožiny (booleovských) prvkov, ktorých inštancie sa častokrát nachádzajú v tesnej geografickej blízkosti (Shekhar a kol. 2003). Typickým príkladom sú symbiotické druhy ako je napríklad krokodíl nílsky a egyptský kulík. Príklady kolokačných vzorov a rôznych domén využitia sú uvedené v Tab. 1. Z nej je možné vyčítať, že kolokačný vzor je distribuovaný v rôznych oblastiach, čo dokazuje jeho veľkú užitočnosť a rozšírenosť (Wei, 2017).

Tab. 1 Príklady aplikácie kolokačného vzoru

Doména	**Príklad**	**Príklad kolokačného vzoru**
Ekológia	Druhy	Krokodíl Nílsky, Kulík Egyptský
Veda o Zemi	Klimatické a rušivé udalosti	Požiar, teplo, sucho, blesky
Ekonómia	Typy priemyslu	Dodávatelia, výrobcovia, konzultanti
Epidemológia	Druhy ochorení, enviromentálne udalosti	Západonílska choroba, uhynutí vtáci, komári
Služby založené na polohe	Typy požiadavkov na služby	Odťah, polícia, záchranka
Počasie	Fronty, zrážky	Studená fronta, teplá fronta, sneženie
Doprava	Doručovacie služby	US Poštový Servis, UPS, doručovanie novín

3.1 História

Kolokačný vzor bol v roku 2004 definovaný Huangom ako podmnožina booleovských priestorových prvkov, ktorých inštancie sa často nachádzajú v priestorovej blízkosti. Príkladom vzoru spoločného umiestnenia je krab pustevník a morská sasanka, ktoré bývajú často pospolu z dôvodu ich symbiotického vzťahu (Gusmao, Daly, 2010).

Algoritmy kolokačného miningu sú inšpirované konceptom miningu asociačných pravidiel (Agrawal, Srikant, 1994), avšak prirodzene neobsahujú pojem transakcia (Huang, 2004). Pre transakcionizáciu priestorových dát Shekhar a Huang (2001) diskutujú tri modely, a to prvkový centrický model, centrický model s oknom a centrický model zameraný na udalosť.

Prvkový centrický model bol navrhnutý Koperskim a Hanom v roku 1995. Druhý spomínaný typ modelu – centrický model s oknom sa zväčša používa v priestorovej štatistike pre exploratórnu analýzu priestorových dát. V súčasnosti najmodernejší prístup je použitie modelu zameraného na udalosti. Ten poskytuje metodológiu založenú na susednom grafe pre generovanie ekvivalentu transakcií s priestorových dát. Dáta sú odvodené z blízkeho susedstva inštancií prvkov. Susedstvo je definované na priestorovom vzťahu ako je napríklad metrický vzťah (euklidovská vzdialenosť).

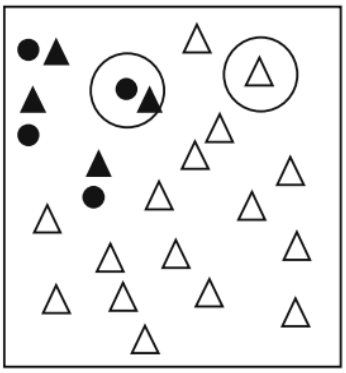
V priestorovej štatistike boli vyvinuté teórie a metódy pre modelovanie priestorových distribúcií a exploratórnu analýzu priestorových dát. Tieto metódy analyzujú medzibodové interakcie bodových dát a modelujú interakcie pre konkrétne prípady. Kolokačný mining je podobný problému s hľadaním asociácií či interakcii vo viactypovom bodovom priestorom procese. Práve tento problém je známy ako efekt druhého rádu (Ripley, 1976). Je výsledkom priestorovej závislosti a predstavuje tendenciu susedných hodnôt na seba nadväzovať. Túto závislosť je možné vyšetriť použitím Rypleyovou K – funkciou, F – funkciou či G – funkciou, ktoré dokážu sumarizovať vzor a detekovať prípadne zhlukovanie.

Väčšina prístupov pre nájdenie kolokačných vzorov, ktoré boli navrhnuté pre dolovanie priestorových dát vyžaduje zadanie prahovej hodnoty, typicky hodnoty PI *(Participation Index)*. Práve voľba správnej prahovej hodnoty je veľmi dôležitá, pri použití malej prahovej hodnoty dochádza k zvýrazneniu nezmyselných vzorov. Naopak pri príliš vysokej hodnote sa práve podstatné vzory strácajú. Ďalšou nevýhodou doterajších prístupov je, že sa používa jedna prahová hodnota pre rôzne veľkosti vzorov. Preto by táto hodnota nemala byť nastavovaná globálne, ale stanovené na základe celkového počtu inštancií zahrnutých do interakcie.

Ďalším problémom je priestorová autokorelácia a veľký počet prvkov, ktoré môžu zavádzať stávajúci prístup dolovania vzoru. Hodnota prevalencie PI nemusí nutne znamenať pozitívny vzťah (asociáciu) medzi prvkami. Nie je neobvyklá situácia, kedy podmnožina vykazuje veľmi vysokú hodnotu miery prevalencie, avšak ide len o náhodnú prítomnosť priestorovej autokorelácie. Taktiež je možné, že bude hodnota PI bude nízka i keď je vo vzore prítomná pozitívna interakcia, a to z dôvodu, že jeden z prvkov má nízky pomer participácie. Práve tieto situácie demonštrujú nasledujúce tri príklady, ktoré ilustrujú potrebu odlišného prístupu na základe štatistického testu (Barua, 2017).

3.2 Príklad 1

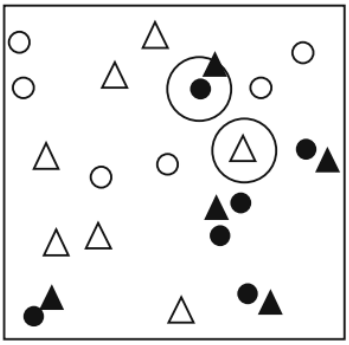
V dátach sú prítomné dva druhy prvkov O (koliesko) a T (trojuholník) (napríklad dva živočíšne druhy) so skutočnou priestorovou asociáciou, takže sa nachádzajú v priestore blízko seba. Ďalej v tomto príklade existuje len pár prvkov O, ale prvkov T je veľké množstvo (viď Obr. 2). To znamená, že pomerne veľká časť prvkov T nebude mať suseda O. V tomto prípade bude index účasti – PI bude pomerne nízky (PI = 4/20 = 0,2). Pri použití typickej prahovej hodnoty nebude kolokačný vzor vyhodnotený ako významný.



Obr. 2 Prvý príklad kolokačného vzoru (zdroj: Barua, 2017)

3.3 Príklad 2

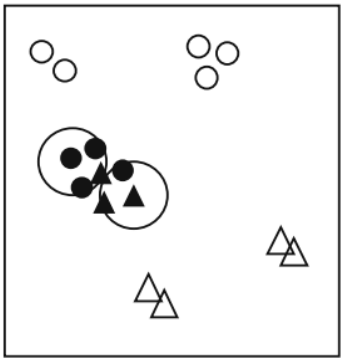
Opäť sa v dátach objavujú dva typy prvkov, kolieska a trojuholníky, ktoré sú na sebe nezávislé, avšak v priestore hojne rozmiestnené (viď Obr. 3). Výsledná hodnota PI bude pri použití typickej prahovej hodnoty vysoká (PI = 5/12 = 0,42), i keď sa v dátach žiaden kolokačný vzor nenachádza.



Obr. 3 Druhý príklad kolokačného vzoru (zdroj: Barua, 2017)

3.4 Príklad 3

Dva druhy prvkov, koliesko a trojuholník, ktoré sú v priestore vo veľkom množstve. Body sú na sebe nezávislé, ale autokorelované, čo znamená, že majú tendenciu sa zhlukovať (viď Obr. 4). Ak sa zhluk O a zhluk T náhodne stretnú, tak hodnota prevalencie bude vysoká (PI = 3/7 = 0,43). Vo výsledku bude v dátach vyhodnotený falošný, významný kolokačný vzor (Barua, 2017).



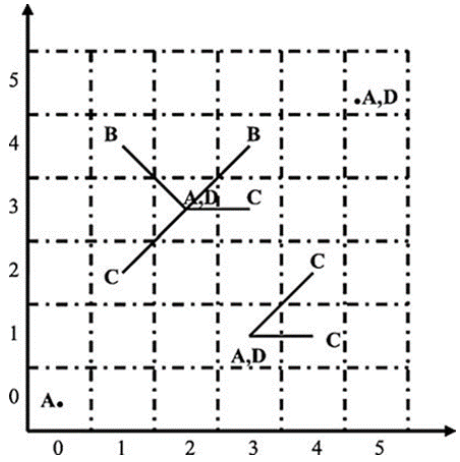
Obr. 4 Tretí príklad kolokačného vzoru (zdroj: Barua, 2017)

3.5 Metódy

Podľa istej kategorizácie (Wei, 2017) existujú tri metódy na zistenie vzoru spoločného umiestnenia v dátach. Ich rozdiel spočíva na zameraní vyhľadávania. Spomínané metódy sa nazývajú: referenčný model zameraný na prvky *(reference feature-centric model)*, model zameraný na okno *(the window-centric model)* a model zameraný na udalosti *(the event-centric model)* (Shekhar, Huang, 2001).

Referenčný model zameraný na prvky je relevantný pre aplikačné domény zamerané na špecifické booleovské prvky ako je napríklad rakovina. Cieľom vedcov je nájsť vzory spoločného umiestnenia s látkami ako je azbest a podobne. Model využíva koncept susedského vzťahu na materializáciu transakcií. Napríklad ak existujú dva typy prvkov A a B. Pokiaľ je A relevantný prvok, to znamená, že B je blízko A, ak je B susedom. Pre výpočet je možné využiť viacero vzdialeností ako je euklidovská či manhattanská. No a potom s použitím definovanej vzdialenosti môžeme prvky prehlásiť za susedov. V prípade, že vyšetrujeme prvky A, tak do úvahy vstupujú všetky ostatné prvky obklopujúce A a budú použité ako transakcie.

Druhou metódou je model zameraný na okno, alebo inak nazývaný ako model delenia dát. Podstatou je definovanie okna správnej veľkosti, následne vyčísli všetky možné okná ako transakcie. Miestne kolokačné vzory sú ohraničené hranicou okien, avšak nie problémom vzor naprieč viacerými oknami. Každé okno je samotnou transakciou, proces vyhodnotí, ktoré prvky sa objavujú spoločne najčastejšie (Shekhar, Huang, 2001). Obrázok päť ukazuje spracovanie jednotlivých okien. V tomto lokálnom modely bol vzťah A a C ignorovaný, pretože sa nenachádzajú v jednom okne.



Obr. 5 Kolokačný vzor, transakcie sú definované okolo prvkov A, relevantné s B a C (zdroj: Shekhar, Huang, 2001)

Treťou metódou je model zameraný na udalosti, týka sa najmä ekologických domén, kde vedci skúmajú špecifické javy ako El Niño, sucho a podobne. Cieľom modelu je nájsť podmnožiny priestorových prvkov, ktoré sa pravdepodobne vyskytujú v susedstve konkrétne udalosti. Pokiaľ je napríklad A susedom B, tak to platí aj naopak, B je susedom A.

Okrem predstavených modelov existujú aj iné modely a vďaka ich rôznorodosti sú k dispozícii aj rôzne algoritmy. Prístupy pre hľadanie pravidiel spoločného umiestenia je možné v jednoduchosti kategorizovať na priestorovú štatistiku a prístupy dolovania dát. Priestorová štatistika a prístupy na nej založené používajú priestorovú koreláciu pre popis vzťahu medzi rôznymi typmi priestorových prvkov. Miery priestorovej korelácie zahrnujú funkciu cross – K so simuláciou Monte Carlo, strednú vzdialenosť, najbližšieho suseda či modely priestorovej regresie. Výpočet korelácie môže byť pre

rôzne priestorové vzory nákladný (Huang a kol. 2004). Druhá kategória – prístupy pre dolovanie dát sa ešte delí do 2 kategórií, a to prístup mapového prekrytia založený na klastroch a asociačné pravidlá. Prvý prístup považuje každý atribút za mapovú vrstvu a priestorový klaster ako kandidáta pre dolovanie asociácií. Asociačné prístupy je možné opäť rozdeliť do 2 kategórií: prístupy založené na transakciách a prístupy založené na vzdialenosti. Transakčné prístupy si kladú za cieľ definovať transakcie v priestore, tie môžu byť definované referenčne orientovaným modelom. Tento prístup má však niekoľko zásadných nedostatkov. Pri výpočte môže dôjsť k duplicitnému počítaniu asociácií, v prípade, že nie je špecifikovaný referenčný prvok (Wei, 2017).

3.6 Kľúčové aplikácie

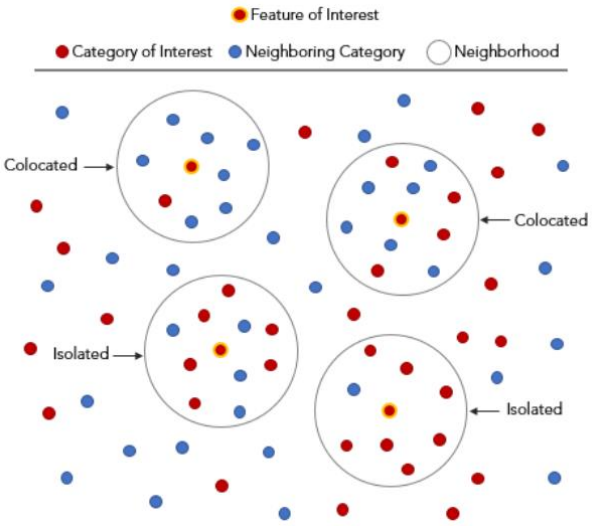
Ťažba vzorov spoločného umiestnenia je určite výhodná pre vedeckú analýzu dát (Salmenkivi, 2004; Yang a kol. 2005). V sčítaní obyvateľov môžu kolokačné vzory naznačovať rysy, ktoré sa často objavujú v takomto type priestorových dát. Napríklad obyvatelia s vyšším platom zvyknú žiť v oblastiach s nízkym znečistením. Ďalším možným príkladom analýzy je vzťah kontaminovaných vodných nádrží a ochorení v ich blízkom okolí. Biológovia dokážu prostredníctvom kolokačných vzorov identifikovať zaujímavé kombinácie, ktoré sa objavujú v zložkách proteínových, alebo chemických štruktúr. Zaujímavou oblasťou aplikácie týchto vzorov je v marketingu pre podporu rozhodovania. Spoločnosť zaoberajúca sa poskytovaním rôznych typov služieb ako sú informácie o počasí, cestovných lístkoch a podobne. Zákazníci môžu požiadavky odosielať z rôznych miest, spoločnosť môže takto získať informáciu o tom, ako sú služby rozmiestnené v priestore, a kde je vhodné poskytovať konkrétnu službu (Mamoulis, 2017).

3.7 Budúci smer

Objavovanie kolokačných vzorov a dolovanie pravidiel je veľmi dôležité, systémy extrahujú doposiaľ neznáme, no zaujímavé priestorové vzory či vzťahy v dátach (Mamoulis, 2017). Tieto metódy majú veľký potenciál pre uplatnenie sa v rôznych vedeckých oblastiach. Súčasné prístupy a algoritmy sa stále vyvíjajú aj vďaka novým štúdiám v rôznych oblastiach. Preto je možné očakávať zlepšenie ich efektivity či rozšírenie priestorových dát na úsečky a polygóny (Huang a kol. 2004). Druhý smer vývoja je vhodné zamerať na použitie iných než booleovských priestorových prvkov. Bolo by vhodné algoritmus rozšíriť tak, aby dokázal pracovať aj s kategorickými prvkami (Huang a kol. 2004). Tretím možným smerom pokroku dolovania kolokačných vzorov a pravidiel je predstava o vzore ako dekolokačného vzoru, alebo koincidenčného (Xiong a kol. 2004).

3.8 Kolokačná analýza (Colocation analysis) v ArcGIS Pro

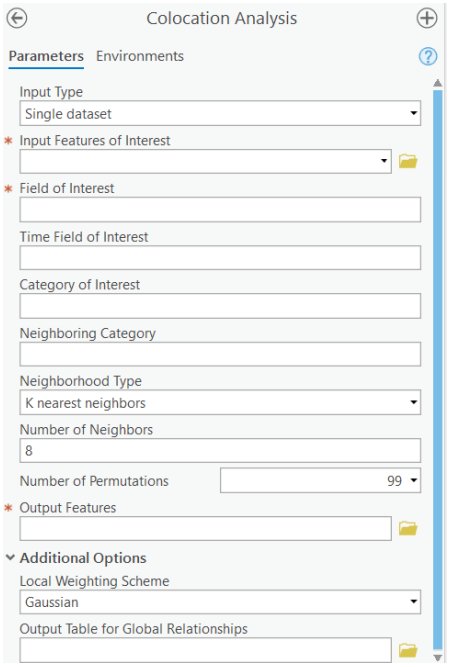
Meria lokálne vzorce priestorových kolokácií medzi dvomi kategóriami reprezentovanými bodmi, pomocou štatistiky kolokačného kvocientu. V jednoduchosti je princíp zhrnutý na Obr. 6. Nástroj akceptuje len bodovú reprezentáciu. Kategória, ktorá bude analyzovaná môže byť obsiahnutá v jednom alebo v dvoch separátnych datasetoch. Je možné použiť dataset s množstvom kategórií (napríklad typy reštaurácii), ale len jedna kategória záujmu bude využitá. To isté sa udeje aj druhým datasetom, bude použitá len jedna kategória z dostupných.



Obr. 6 Princíp kolokačnej analýzy (zdroj: https://pro.arcgis.com/)

Nástroj určí pre každý prvok záujmu či prvky susednej kategórie sú viac či menej prítomné v porovnaní s celkovou priestorovou distribúciou vstupných bodov. Napríklad pre každý bod kategórie A, ak je výsledná hodnota lokálneho kolokačného kvocientu (LCLQ) rovná jednej, tak to znamená, že je pravdepodobné, že kategória B je sused. Hodnota väčšia ako jeden značí, že je viac pravdepodobné, že A bude mať za suseda B. Hodnota LCLQ menšia ako jeden značí malú pravdepodobnosť, že kategória A bude mať za suseda bod z kategórie B.

Kolokačný vzťah tejto analýzy nie je symetrický. Vypočítaná hodnota kolokačného kvocientu bude iná v prípade porovnania kategórií A a B a iná pri dvojici B a A. Taktiež je výsledok iný v prítomnosti kategórie C v susedstve, preto je niekedy lepšie zahrnúť len kategórie A a B. Netreba však zabúdať na to, že tvorbou danej kategórie sa stráca informácia o vzťahoch k iným kategóriám. Je dôležité vybrať také kategórie, ktoré nie sú ovplyvnené výskytom inej. Rozhranie nástroja pri prvotnom otvorení demonštrované na Obr. 7 (ArcGIS Pro, 2022a).



Obr. 7 Rozhranie nástroja Colocation Analysis (zdroj: autorka)

Priestorový vzťah môže byť definovaný použitím vzdialenostného pásma *(Distance band)* K najbližších susedov *(K nearest neighbors)*, alebo pomocou matice váh *(Spatial weights)*, ktorá sa zostavuje pred samotnou analýzou a je definovaná niekoľkými typmi susedstva. Váhy je možné vytvoriť pomocou nástroja generovať priestorovú maticu váh a nutné je zadať obidve kategórie záujmu. Priestorový vzťah je možné zostaviť niekoľkými spôsobmi:

- **inverzná vzdialenosť** – vplyv jedného prvku na druhý klesá so vzdialenosťou,
- **fixná vzdialenosť** – všetky prvky v špecifikovanej vzdialenosti budú do analýzy zahrnuté,
- **k najbližších susedov** – do analýzy bude zahrnutý špecifikovaný počet najbližších prvkov,
- **susedstvo len hrán** – polygónové prvky zdieľajúce hranu budú považované za susedov,
- **susedstvo hrán** – polygónové prvky zdieľajú hranu a uzol budú považované za susedov,
- **delaunay triangulácia** – z centroidov bude vytvorená sieť neprekrývajúcich sa trojuholníkov, tie ktoré zdieľajú hrany budú považované za susedné,
- **časopriestorové okno** – prvky vo zvolenej vzdialenosti a čase sú susedia,
- **konvertovaná tabuľka** – priestorové vzťahy sú definované v tabuľke (ArcGIS Pro, 2022b).

Dáta je možné analyzovať použitím časopriestorového okna so špecifikáciou časového poľa záujmu, časového poľa susedných kategórií a parametrami typu dočasného vzťahu. Pomocou časopriestorových okien je možné kontrolovať, ktoré prvky budú do analýzy zahrnuté. Prvky, ktoré sú pri sebe blízko časovo aj priestorovo budú analyzované spoločne, pretože vzťahy medzi prvkami sú hodnotené z priestorového aj časového hľadiska. Taktiež je možné špecifikovať či nástroj bude hľadať prvky pred alebo za cieľovým prvkom, alebo je možné vytvoriť časové rozpätie, behom ktorého bude nástroj hľadať prvky pred a po analyzovaní cieľového prvku.

Globálny kolokačný kvocient je možné vypočítať zadaním cesty pre parameter výslednej tabuľky pre globálne vzťahy *(Output Table for Global Relationships)*. Táto tabuľka obsahuje kolokačné kvocienty, takže je možné analyzovať priestorové asociácie medzi všetkými kategóriami datasetu. Tabuľka umožňuje preskúmať iné vzťahy v dátach, je možné nájsť ďalšie silne kolokačné zaradené kategórie. Pokiaľ sa behom analýzy objavia iné silné kolokačné kategórie, je možné analýzu rozšíriť preskúmaním povahy tohto vzťahu, a to opätovným spustením nástroja s týmito novými kategóriami, alebo odobraním týchto kategórii a spustením analýzy, ak má užívateľ dojem, že kategórie skresľujú výsledok.

Výstupom tohto nástroja je vrstva zobrazujúca každý vstupný prvok záujmu. Prvky sú kategorizované do piatich typov (typ LCLQ):

- významne kolokované,
- nevýznamne kolokované,
- významne izolované,
- nevýznamne izolované,
- nedefinované.

Nástroj pridáva do výstupnej atribútovej tabuľky pole zahrňujúce lokálny kolokačný kvocient, p-hodnotu, LCLQ bin použitý pre symbolizáciu a typ LCLQ. Pri parametrizácií je možné zaškrknúť aj voliteľnú výslednú tabuľku pre globálne vzťahy, ktorá bude

obsahovať globálne kolokačné kvocienty medzi všetkými kategóriami v parametri pole záujmu a všetkými kategóriami v parametri pole obsahujúce susednú kategóriu.

Kolokačná analýza podporuje paralelné spracovanie, vo východzom stave používa 50 % dostupných procesorov. Počet procesorov je možné zvýšiť alebo znížiť pomocou prostredia paralelného procesného faktoru *(Parallel Processing Factor)*.

Parameter počet permutácií sa používa k výpočtu p-hodnôt. Voľbou počtu permutácií je stanovená rovnováha medzi presnosťou a dlhšou dobou spracovania. Východzia hodnota je 99, pre lepší výsledok analýzy sa odporúča zvýšiť počet permutácií (ArcGIS Pro, 2022a).

3.8.1 Prehľad parametrov

1) Typ vstupu *(Input Type)*

Určuje či hodnoty prvkov záujmu *(In Features of Interest)* pochádzajú z rovnakého datasetu so zadanými kategóriami alebo z rôznych dátových sad, s ktorými sa bude zaobchádzať ako s ich vlastnými kategóriami. Napríklad jedna dátová sada so všetkými bodmi reprezentujúcimi gepardy a druhá dátová sada, v ktorej všetky body predstavujú gazely. Dostupné sú tri možnosti vstupu:

- jedná dátová sada *(Single dataset)* – analyzované kategórie pochádzajú z polí totožnej dátovej sady,
- dve dátové sady *(Two datasets)* – analyzované kategórie pochádzajú z polí dvoch dátových sad,
- dátové sady bez kategórií *(Datasets without categories)* – dva samostatné datasety reprezentujúce dve analyzované kategórie.

2) Vstupné prvky záujmu *(Input Features of Interest)*

Trieda obsahuje body s reprezentovanými kategóriami, ktoré budú v procese analyzované.

3) Výstupné prvky *(Output Features)*

Trieda obsahuje všetky vstupné prvky záujmu so stĺpcami obsahujúcimi skóre lokálneho kolokačného kvocientu, symbológiu a p-hodnoty.

4) Pole záujmu *(Field of Interest)*

Pole obsahuje kategóriu alebo kategórie, ktoré budú analyzované. Voľba parametru nie je povinná.

5) Časové pole záujmu *(Time field of Interest)*

Dátumové pole s voliteľnou časovou známkou pre jednotlivé prvky. Slúži pre analyzovanie prvkov za použitia časopriestorového okna. Prvky, ktoré sa nachádzajú v čase a priestore blízko seba budú považované za susedné a budú analyzované spolu.

6) Kategória záujmu *(Category of Interest)*

Základná kategória pre analýzu. Nástroj identifikuje pre jednotlivé kategórie záujmu hodnoty, stupeň, do ktorej základnej kategórie sú priťahované alebo kolokované s hodnotou parametru susednej kategórie.

7) Vstupné susedné prvky *(Input Neighoring Features)*

Trieda obsahuje body s kategóriami ktoré budú porovnávané.

8) Pole obsahujúce susednú kategóriu *(Field Containing Neighboring Category)*

Pole s parametrom vstupných susedných prvkov obsahujúca kategóriu, ktorá bude porovnávaná.

9) Časové pole susedných prvkov *(Time Field of Neighboring Features)*

Dátumové pole s časovou značkou pre jednotlivé prvky na analýzu bodov používajúce časopriestorové okno. Prvky blízko seba v priestore a čase budú považované za susedov a budú analyzované spolu.

10) Susedná kategória *(Neighboring Category)*

Susedná kategória pre analýzu kolokačného vzoru. Nástroj identifikuje stupeň do akej miery je hodnota parametru kategória zájmu priťahovaná alebo izolovaná od hodnoty susednej kategórie.

11) Typ susedstva *(Neighborhood Type)*

Špecifikuje definovanie priestorového vzťahu medzi prvkami. Dostupné sú tri možnosti definovania susedstva vstupných prvkov, a to:

- Vzdialenostné pásmo *(Distance band)*

Každý prvok bude analyzovaný v kontexte susedných prvkov. Susedné prvky vo vnútri zadanej kritickej vzdialenosti určené parametrom majú váhu jeden a majú vplyv na výpočet pre cieľový prvok. Susedné prvky mimo kritickej vzdialenosti majú váhu nula a nemajú žiaden vplyv na výpočet cieľového prvku.

- K najbližších susedov *(K nearest neighbors)*

Najbližší k prvok bude zahrnutý do analýzy ako sused. Počet susedov je špecifikovaný parametrom počet susedov. Vplyv suseda v analýze je vážený na základe vzdialenosti od najvzdialenejšieho suseda. Toto je východzie nastavenie.

- Priestorové váhy so súboru *(Get spatial weights from file)*

Ak je jedna dátová sada použitá ako vstupný typ, priestorové vzťahy budú definované špecifikovaným priestorovým súborom – maticou. Vplyv suseda v analýze je vážený na základe vzdialenosti od najvzdialenejšieho suseda. Cesta k súboru priestorových váh je špecifikovaná parametrom súbor matice váh.

12) Počet susedov *(Number of Neighbors)*

Počet susedov okolo každého prvku, ktorý bude použitý pre testovanie lokálnych vzťahov medzi kategóriami. Ak nie je zadaná žiadna hodnota, bude použitá východzia (osem susedných prvkov). Zadaná hodnota musí byť dostatočne veľká, aby detekovala vzťahy medzi prvkami, ale nie príliš malá, aby dokázala identifikovať miestne vzory.

13) Vzdialenostné pásmo *(Distance band)*

Veľkosť susedstva je konštantná, alebo pevne daná vzdialenosť pre každý prvok. Všetky prvky v tejto vzdialenosti budú použité na testovanie lokálnych vzťahov medzi kategóriami. Pokiaľ nie je uvedené žiadna hodnota, bude použitá priemerná vzdialenosť.

14) Súbor matice váh *(Weight Matrix File)*

Cesta k súboru obsahujúca váhy, ktoré definujú priestorový a potencionálne časový vzťah medzi prvkami.

15) Časový typ vzťahu *(Temporal Relationship Type)*

Špecifikuje ako budú definované časové vzťahy medzi prvkami. Dostupné sú tri časové typy, a to:

- Predtým *(Before)*

Časové okno sa predĺži späť v čase pre každú z hodnôt vstupných prvkov záujmu. Susedné prvky musia mať dátumovú alebo časovú značku, ktorá sa objaví pred dátumovou, alebo časovou značkou záujmového prvku v analýze. Toto je východzie nastavenie.

- Potom (After)

Časové okno sa predĺži vpred v čase pre každú z hodnôt vstupných prvkov záujmu. Susedné prvky musia mať dátumovú, alebo časovú značku, ktorá sa objaví po dátumovej alebo časovej značke záujmového prvku v analýze.

- Predtým aj potom (Span)

Časové okno sa predĺži tam aj späť pre každú hodnotu vstupného prvku záujmu. Susedné prvky, ktoré majú dátumovú, alebo časovú značku a ktoré sa vyskytujú medzi hodnotou intervalu časového kroku pred, alebo po dátumovej, alebo časovej značke záujmového prvku analýzy. Napríklad, ak je parameter interval časového kroku *(Time Step Interval)* nastavený na jeden týždeň, okno bude nastavené na jeden týždeň pred a jeden týždeň za cieľovým prvkom.

16) Interval časového kroku *(Time Step Interval)*

Celé číslo a merná jednotka reprezentujúca počet časových jednotiek tvoriacich časové okno.

17) Počet permutácií *(Number of Permutations)*

Počet permutácií, ktoré budú použité k vytvoreniu referenčnej distribúcie. Voľbou počtu permutácii je rovnováha medzi presnosťou a zvýšením procesného času. Na výber je preferencia rýchlosti a precíznosť. Viac robustné a precíznejšie výsledky potrebujú viac času na výpočet. K dispozícií sú štyri hodnoty:

- 99

Kolokačná analýza použije 99 permutácii. S týmito permutáciami je najmenšia možná pseudo p-hodnota je 0,02 a všetky ostatné pseudo p-hodnoty budú násobené touto hodnotou. Toto je východzie nastavenie.

- 199

Analýza využije 199 permutácii. Pri tejto hodnote bude najmenšia možná pseudo p-hodnota 0,01 a všetky ostatné pseudo p-hodnoty budú násobené touto hodnotou.

- 499

Nástroj použije 499 permutácii. Najmenšia možná pseudo p-hodnota je 0,002 a všetky ostatné pseudo p-hodnoty budú násobené touto hodnotou.

- 9999

Analýza aplikuje 9999 permutácii. S 9999 permutáciami je najmenšia možná pseudo p-hodnota 0,0002 a všetky ostatné pseudo p-hodnoty budú násobené touto hodnotou.

18) Lokálna vážená schéma *(Local Weighting Scheme)*

Špecifikuje typ kernel jadra, ktoré bude použité k poskytnutiu priestorového váženia. Kernel jadro definuje ako každý prvok súvisí s ostatnými prvkami v jeho susedstve. Dostupné typy jadra:

- Bikvadratické *(Bisquare)*

Prvok bude vážený na základe vzdialenosti k najvzdialenejšiemu susedovi, alebo na základe okraja vzdialenostného pásma. Každému prvku mimo špecifikované susedstvo bude priradená váha nula.

- Gausovské *(Gaussian)*

Prvok bude vážený na základe vzdialenosti k najvzdialenejšiemu susedovi alebo na základe okraja vzdialenostného pásma, avšak o trochu rýchlejšie než voľba bikvadratická. Každému prvku mimo špecifikované susedstvo bude priradená váha nula. Toto je východzie nastavenie.

- Žiadne *(None)*

Nebude použitá žiadna váhová schéma, všetkým prvkom v susedstve bude priradená váha jeden a budú prispievať rovnocenne. Všetky ostatné prvky mimo susedstvo budú mať hodnotu váhu nula.

19) **Výstupná tabuľka globálnych vzťahov** *(Output Table for Global Relationships)*

Tabuľka obsahujúca globálne kolokačné koeficienty medzi všetkými kategóriami parametra pole záujmu a všetkými kategóriami parametra poľa obsahujúceho susednú kategóriu. Táto tabuľka môže byť nápomocná pri určení lokálnych kategórií pri analýze.

Ak sa ako hodnota parametru použije dataset bez kategórií, globálny kolokačný kvocient bude vypočítaný pre každý dataset a medzi každým datasetom (ArcGIS Pro, 2022a).

4 PRÍPADOVÁ ŠTÚDIA 1 – LEKÁRNE

Základnou vrstvou vstupujúcou do analýzy je vrstva subjektov zdravotníctva, ktorá bola poskytnutá pre účely bakalárskej práce Michalovi Kupkovi s názvom Analýza prostorového vzoru subjektu pusobících v oblasti zdravotnictví. Vrstva pochádza z Databázy Amadeus, ktorá obsahuje informácie socioekonomického charakteru o zdravotných subjektoch, poskytol ju doc. Ing František Dařena, Ph.D.

Použité vrstvy:

- Primárne
    - Subjekty zdravotníctva – bodová vrstva
    - Lekárne – bodová vrstva
- Doplnkové
    - Administratívne členenie – polygónová vrstva

Databáza Amadeus detailne mapuje podnikateľské subjekty vyskytujúce sa na území Európy. Bola vytvorená firmou Bureau van Dijk, ktorá taktiež zodpovedá za správu databázy. Obsahom je viac než 21 miliónov záznamov firiem pochádzajúcich z rôznych odvetví. Spomínaná firma zbiera dáta z väčšiny štátov Európy vrátane Českej republiky. Databáza je vhodná pre výskum európskych firiem, ich trendu v podnikaní alebo na štatistické porovnanie. Jej súčasťou sú viaceré atribúty ako Company *(Spoločnosť)*, City *(Mesto)*, Address *(Adresa)* a mnohé iné s celkovým počtom záznamov 21 077 na území celej Českej republiky. Základným a rozhodujúcim atribútom je NACE *(Európska klasifikácia ekonomických činností)* obsahujúca číselný kód s typom subjektu. Kódy zdravotníckej starostlivosti začínajú číslicami 86xx, bližšie informácie sú uvedené nižšie (viď Tab. 2). Pre analýzu boli vybrané subjekty obsahujúce kód začínajúci číslami 86xx (Kupka, 2019).

Tab. 2 Význam číselných kódov typov subjeku

Kód	**Význam**
46.46	Veľkoobchod s farmaceutickými výrobkami
47.73	Maloobchod s farmaceutickými prípravkami
47.74	Maloobchod s zdravotníckymi a ortopedickými výrobkami
86.10	Ústavná zdravotná starostlivosť – lekárska starostlivosť ako diagnostika, ošetrenie a liečba vo všeobecných a špecializovaných nemocniciach
86.20	Ambulantná a zubná zdravotná starostlivosť – konzultácie a ošetrenia u všeobecných či špecializovaných lekárov vrátane zubných lekárov
86.21	Všeobecná ambulantná zdravotnícka starostlivosť
862.2	Špecializovaná ambulantná zdravotnícka starostlivosť
85.23	Zubná starostlivosť
86.90	Ostatné činnosti súvisiace so zdravotnou starostlivosťou – činnosti vykonávané mimo nemocníc

Ďalšou vstupnou vrstvou je dátová vrstva Mapa přístupnosti – budovy pochádzajúca z Magistrátu mesta Brna v rámci dátového portálu data.Brno. Ten poskytuje dáta verejnosti, ktoré môže ktokoľvek využiť pre výskum, projekty, kontrolu či len informovanie sa o aktuálnom stave. Vrstva zahrňuje významné inštitúcie ako sú zdravotnícke zariadenia, kultúrne objekty, úrady, obchodné domy, reštaurácie a podobne. Atribútová zložka je naozaj pestrá, okrem základných vlastností ako je názov či typ budovy obsahuje aj konkrétny dátum aktualizácie či odkaz na webovú stránku s celkovým počtom 266 záznamov (data.Brno, 2021).

Doplnkovou ale nedeliteľnou súčasťou prípadovej štúdie je digitálna vektorová geodatabáza ArcČR® vo verzii 3.3 od spoločnosti ARCDATA PRAHA s.r.o. Z nej bola využitá časť administratívneho členenia s administratívnym vymedzením jednotlivých celkov Českej republiky začínajúc od NUTS *(Nomenklatúra územných štatistických jednotiek)* nula pre štát a končiac základnými sídelnými jednotkami. Pre analýzu bola zvolená časť obce Brno – město, ktorá pokrýva centrálnu a historickú časť tohto krajského mesta.

4.1 Postup úpravy dát

Prvým krokom bolo vybranie len tých zdravotníckych subjektov a budov, ktoré sa nachádzajú na území mesta Brna a časti obce Brno – město. Na to bol použitý nástroj výber podľa polohy (*Select By Location)* v prostredí ArcGIS Pro. Ako podporná vrstva bola použitá vrstva mestskej časti. Následne bola nutné zo zdravotníckych subjektov vybrať len tie, ktoré obsahujú v atribúte NACE hodnoty začínajúce 86xx. Tento krok bol uskutočnený pomocou nástroja výber podľa atribútov *(Select By Attributes)* s podmienkou kód NACE je väčší než 4774, výsledkom je 62 subjektov. Podobným postupom bola upravená vrstva budov, avšak s typom budovy „lekáreň“.

Rovnakým spôsobom došlo k vytvoreniu vrstvy zdravotníckych subjektov a lekární, avšak vrátane bodov v obalovej zóne so vzdialenosťou 100 metrov od administratívnej hranice Brno – město. Obalová zóna vznikla pomocou nástroja obalová zóna *(Buffer)* a vstupnej vrstvy hranice mestskej časti.

4.2 Kolokačná analýza

Z teoretickej časti práce je zrejmé, že nástroj kolokačná analýza *(Colocation Analysis)* meria lokálne vzorce priestorových kolokácií medzi dvoma kategóriami pomocou štatistiky kolokačného kvocientu. Vstupom môže byť len bodová vrstva, avšak je tu viacero možností. V tomto prípade bola zvolená varianta dátových sád bez kategórií, keďže samotná dátová vrstva predstavuje jednu kategóriu. Vstupnú vrstvu záujmu predstavovali lekárne a susediace prvky tvorila vrstva zdravotníckych subjektov. Časový aspekt nebol vyšetrovaný, pre analýzu priestorového vzťahu boli použité dve metódy, a to K najbližších susedov (*K nearest neighbors*) a vzdialenostné pásmo (*Distance band*). Ich hodnoty boli nastavované postupne a následne analyzované. Nástroj obsahuje aj ďalšie nastavenia ako počet permutácií (99, 199, 499, 9999) či typ funkcie (*Bisquare, Gaussian, None*).

Vrstva zdravotníckych subjektov obsahovala 62 záznamov, avšak poloha niektorých záznamov bola zhodná. Preto bola vytvorená nová vrstva bez duplicitných bodov pre porovnanie výsledkov. Pre jej zostrojenie bola nad vrstvou mestskej časti vytvorená sieť pomocou nástroja generovanie gridu z oblasti *(Generate Grid From Area)* s veľkosťou bunky 30 x 30 metrov. Následne bola spočítaná četnosť výskytu subjektov v jednotlivých bunkách prostredníctvom nástroja Zhrnutie vo vnútri *(Summarize*

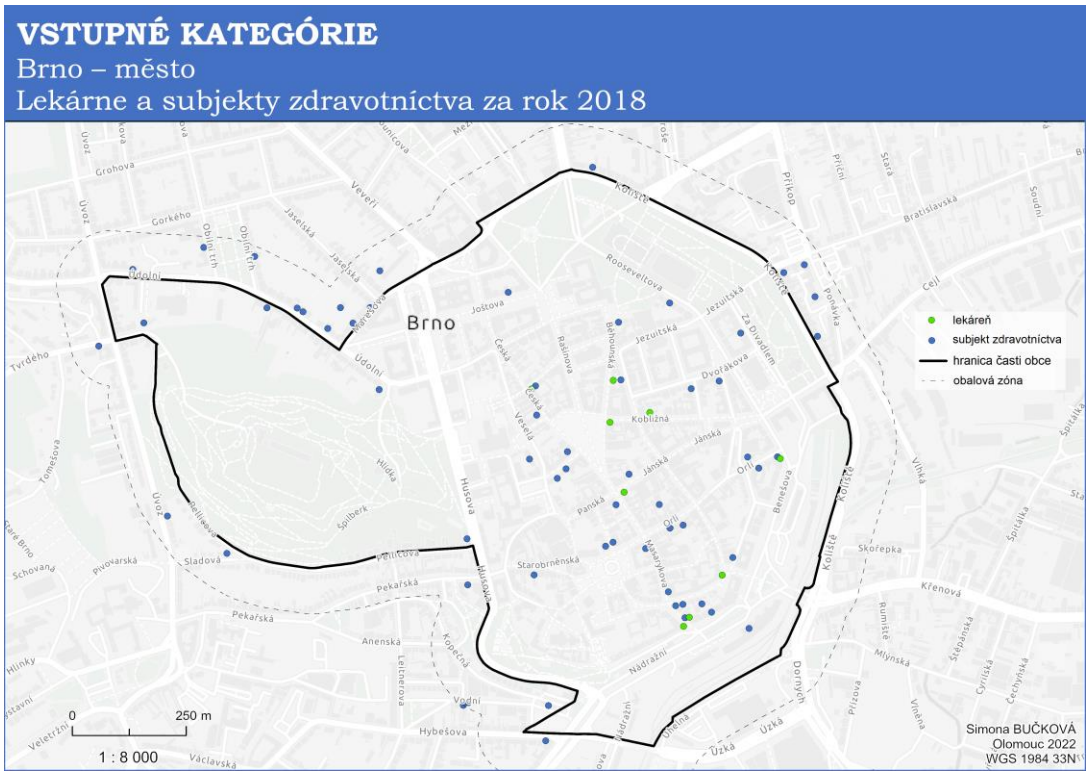
Within). Výsledok poslúžil na identifikáciu oblastí, kde je viacero záznamov na totožnom mieste. Výsledné kolokačné vzory budú vo viacerých variantách s použitím metód K najbližších susedov a vzdialenostného pásma. Varianty budú nasledujúce:

- 1A K najbližších susedov,
- 1B Vzdialenostné pásmo,
- 2A obalová zóna, K najbližších susedov,
- 2B obalová zóna, Vzdialenostné pásmo,
- 3A duplicitné záznamy, K najbližších susedov,
- 3B duplicitné záznamy, Vzdialenostné pásmo,
- 4A duplicitné záznamy a obalová zóna, K najbližších susedov,
- 4B duplicitné záznamy a obalová zóna, Vzdialenostné pásmo subjekty v rámci časti obce Brno – město.

V obalovej zóne so vzdialenosťou 100 m sa nachádza 49 subjektov, po vybraní tých s kódom 86xx pomocou výberu podľa atribútov *(Select By Attributes)* ich ostalo 44. Vrstva obsahovala taktiež duplicitné záznamy, čo znamená, že ich poloha bola totožná. Práve to bolo námetom na vytvorenie variant bez duplicít a variant s ich ponechaním. Cieľom je preskúmanie vplyvu týchto duplicít na výsledný kolokačný vzor.

Javy či objekty sa častokrát nevyskytujú v priestore pravidelne, práve naopak. Častokrát vytvárajú zhluky, čo je možné pozorovať aj v tomto prípade (viď Mapa 1). Ordinácie lekárov sú umiestnené blízko seba a v ich dochádzkovej vzdialenosti sú častokrát umiestnené lekárne pre potreby pacientov. Každopádne nie je to pravidlo, lekárne sa zvyknú umiestňovať do nákupných centier, alebo centra mesta, kde je veľká koncentrácia ľudí. Na základe tejto priestorovej distribúcie záznamov subjektov zdravotníctva a lekární vznikla hypotéza H_0:

Lekárne obklopené zdravotníckymi subjektami budú vytvárať kolokačné vzory.



Mapa 1 Prehľad distribúcie vstupných kategórií

Vzhľadom na varianty úpravy dát vznikla ďalšia hypotéza, ktorá predpokladá, že duplicitné záznamy výrazne ovplyvnia kolokačné vzory. V miestach lekární s výrazným počtom subjektov, vrátane duplicitných záznamov vzniknú významné kolokačné vzory, respektíve kolokačný nástroj vyhodnotí tieto lekárne ako významné kolokované. Znenie hypotézy H_1 je nasledovné:

Lekárne obklopené zdravotníckymi subjektami vrátane duplicít budú označené ako významne kolokované.

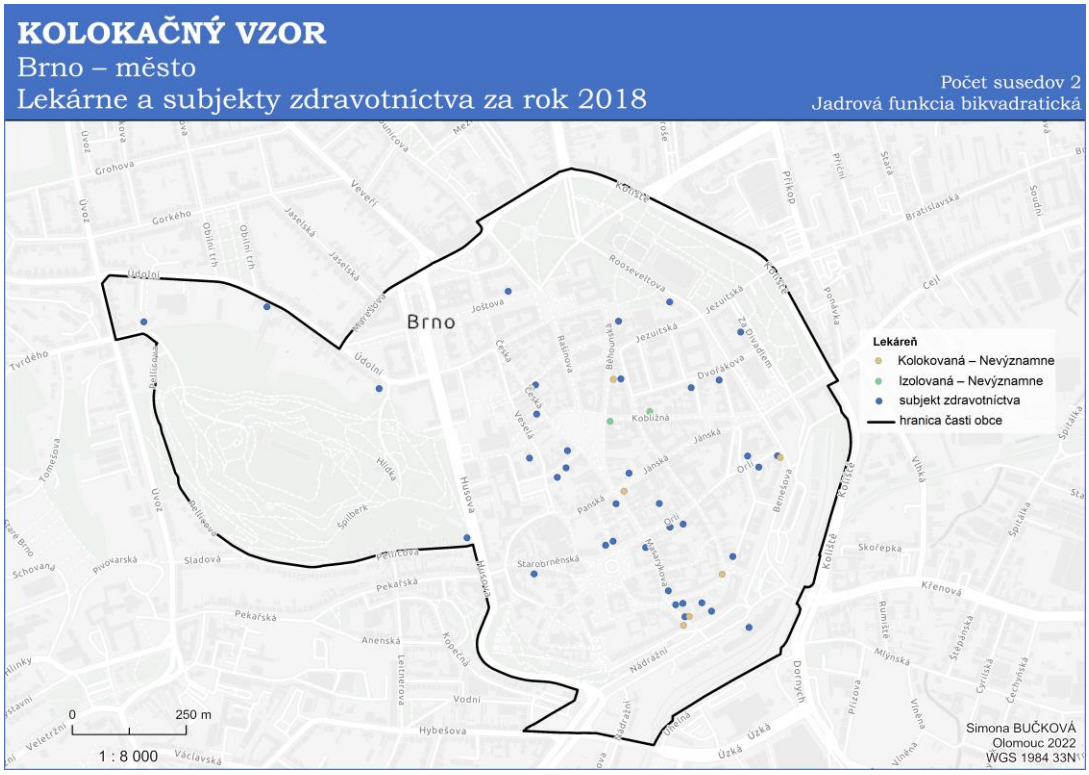
4.2.1 1A K najbližších susedov

Voľba hodnoty K (*Number of Neighbors*) je veľmi dôležitá, pretože určuje či algoritmus bude schopný nájsť významné vzory. V prípade, že zadaná hodnota je príliš nízka, významné vzory nemusia byť zahrnuté, naopak pri príliš vysokej hodnote môže algoritmus vyhodnotiť vzory ako významné i keď budú v skutočnosti falošné. Správny odhad hodnoty nie je jednoduchý a mení sa v závislosti na charaktere dát, avšak nástroj ponúka vopred nastavenú hodnotu, a to osem susedných prvkov. Vzhľadom na počet bodov je možné zhodnotiť, že určite nie je vhodné zadať príliš nízku, ale ani príliš vysokú hodnotu. Preto bol zvolený interval dvoch až piatich najbližších susedných prvkov. Následne boli postupne pridávané susediace prvky a bola sledovaná priradená hodnota s konkrétnou symbológiou a hodnotami lokálneho kolokačného kvocientu. Nástroj automaticky vytvorí novú bodovú vrstvu s piatimi triedami na základe hodnoty kolokačného kvocientu. Vrstva obsahuje atribútovú tabuľku s jeho hodnotami a príslušnú p-hodnotou. Doplňujúcim nastavením je počet permutácii, ktorý so vzrastajúcou hodnotou vypočíta presnejší výsledok. Vzhľadom na pomerne malé vstupné dátové vrstvy bola zvolená najvyššia hodnota, a to 9999 permutácií. V poslednej možnosti voľby typu lokálnej váhovej schémy (*Local Weighting Scheme*) nástroj definuje ako každý prvok súvisí so susednými prvkami. V tejto analýze preto boli testované všetky ponúkané typy jadra.

Prvotná analýza zahrňovala len dva susedné prvky čiže subjekty zdravotníctva. Je to pomerne nízka hodnota, preto sa neočakávalo, že budú objavené významné kolokačné vzory. Avšak, väčšina lekární bola vyhodnotená ako nevýznamne kolokovaná. V tomto prípade sa hodnota lokálneho kolokačného kvocientu pohybuje okolo jeden. Ak by bola lekáreň vyhodnotená ako významne kolokovaná s jej definovaným okolím, hodnota kvocientu by bola blízka či rovná nule. Objavili sa aj izolované lekárne, čo sa dalo očakávať, a to preto, že lekárne na Kobližnej ulici nemajú vo svojej blízkosti žiadny subjekt zdravotníctva. Jej LCLQ sa pohyboval okolo hodnoty tri. Zaujímavosťou je fakt, že i keď spomínané lekárne viditeľne nemajú v okolí ordináciu, neboli vyhodnotené ako izolované. Čo sa týka voľby jadrovej funkcie, podobné výsledky mali gaussovská a bikvadratická. Najviac sa odlišovala funkcia žiadna, ktorá označila až päť lekární za izolovaných. Dôvodom je jej algoritmus, ktorý neznižuje váhu pri vzrastajúcej vzdialenosti.

Zvýšením počtu susedných prvkov na tri klesla pozitivita kolokačného vzoru. Rovnaký výsledok sa zachoval len pri gaussovskom jadre, v ostatných počet kolokovaných lekární klesol. Každopádne vyšetrenie okolia s dvoma prvkami je pomerne malé, no práve pri tomto nastavení sú výsledky najpriaznivejšie (viď. Mapa 2). Po zvýšení počtu susedných prvkov z troch na štyroch nenastala skoro žiadna zmena. Výnimkou je lekáreň na nádraží pri gaussovskom type jadrovej funkcie. Významne kolokovaná nebola žiadna lekáreň, takmer polovica bola vyhodnotená ako izolovaných. Z toho je možné usúdiť, že medzi lekárňami a subjektami zdravotníctva je kolokačný vzor, avšak nie úplne taký aký by užívateľ očakával. Vyhodnotenie či je záznam

kolokovaný, alebo nie sa dá čiastočne predpokladať, osamotené lekárne sú vo väčšine prípadov označované ako izolované.



Mapa 2 Kolokačný vzor pre lekárne a subjekty zdravotníctva za rok 2018

Posledným nastavením bola hodnota piatich susedných prvkov. Vzhľadom na to, že v predchádzajúcom nastavení nenastala veľká zmena, len pribudli nevýznamne izolované lekárne, veľká zmena predpokladaná nebola. Predpoklad sa potvrdil, žiadna lekáreň nebola identifikovaná ako významne kolokovaná, práve naopak. V gaussovskom jadre sa lekáreň na Běhounskej ulici zmenila na nevýznamne izolovanú. Jadrová funkcia bikvadratická vyhodnotila kolokačné vzory lekární rovnako a funkcia žiadna pridala ďalšie dve lekárne k nevýznamne izolovaným.

Výsledky tejto varianty ukázali, že úlohu hrá aj typ jadra. Výsledky pri použití dostupných jadier boli častokrát v skúmaných variantách odlišné pri rovnakom počte susedných prvkov. Našli sa však aj zhodné výsledky, a to pri troch až piatich susedných subjektoch. Celkovo sa však základná hypotéza potvrdila, lekárne, ktoré majú v blízkosti subjekty zdravotníctva boli nástrojom vyhodnotené ako kolokované, i keď nie významne. Najlepší výsledok ponúka nastavenie s dvoma susednými prvkami.

4.2.2 1B Vzdialenostné pásmo

Vzdialenostné pásmo je oproti predchádzajúcej metóde susedstva prostejšie. Každá lekáreň je analyzovaná len v kontexte susedných prvkov so zadanou hodnotou vzdialenosti. Subjekty zdravotníctva nachádzajúce sa vo zvolenej vzdialenosti dostanú váhu jeden, ostatné nula. Opäť je však otázkou aká vzdialenosť je vhodná, aby došlo k objaveniu skutočných kolokačných vzorov. V prvom rade sa hodnota odvíja od charakteru vstupných dát (počet, veľkosť vyšetrovanej oblasti, relevantná vzdialenosť). Preto bola hodnota vzdialenostného pásma postupne zvyšovaná. Počiatočná hodnota 20 m vznikla z idey, že sa lekárne častokrát nachádzajú v bezprostrednej blízkosti ordinácie, niekedy dokonca v jednej budove či komplexe. Približné vzdialeností lekární

a okolitých subjektov boli už vopred zmerané pomocou nástroja meranie *(Measure)*, takže sa dali výsledky analýzy predpokladať. Prekvapivý výsledok nastal už hneď pri prvom nastavení, keď bola lekáreň na Českej ulici označená ako nedefinovaná, i keď v jej blízkosti (do 20 m) sa ordinácia nachádza. Rovnako bola označená aj lekáreň na Masarykovej a Běhounské ulici. Pomerne veľké množstvo lekární bolo identifikovaných ako nedefinované, s čím sa predpokladalo, až na spomínané výnimky. Postupne však nedefinovaných bodov ubúdalo a v posledných nastaveniach (80–100 m) boli označené len skutočne izolované lekárne. Tieto lekárne sa nachádzajú na Kobližnej ulici. Jedinou jadrovou funkciou, ktorá nevarovala o bodoch bez akéhokoľvek suseda je funkcia žiadna. Pri postupnom pridávaní na hodnote vzdialenosti primárne ubúdali nedefinované lekárne. Už pri vzdialenosti 40 m sa objavili izolované lekárne, a to na Masarykovej ulici. Čo je však zvláštne je, že ostali nevýznamne izolované aj pri väčších vzdialenostiach hoci mali ordinácie viditeľne v okolí. Na tento fakt netreba zabúdať pri interpretácii, nástroj Colocation Analysis pracuje so vzdialenostným nie len na základe vzdialenosti, ale aj četnosti susedných bodov. Maximálna zmena zaradenia lekární nastala pri 80 m, následne sa výsledný kolokačný vzor už len opakoval, samozrejme za použitia zhodnej jadrovej funkcie. Najlepšie výsledky boli dosiahnuté pri vzdialenosti 80 m, a to šesť kolokovaných lekární s poslednou jadrovou funkciou. Čo sa týka zhodnosti pri jadrových funkciách, vzor sa opakoval pri bikvadratickej a žiadnej.

Pri porovnaní s rovnakým nastavením, ale s prítomnosťou duplicitných záznamov je na prvý pohľad vo vzoroch viditeľný rozdiel. Priebeh bol miernejší, bolo identifikovaných viac nevýznamne kolokovaných lekární. Avšak rozdiel v zaradení nie je až tak výrazný, i keď do analýzy vstúpilo omnoho viac subjektov zdravotníctva. Vyzerá to tak, že vo vzťahu lekární a okolitých subjektov je istý trend a prítomnosť duplicitných bodov ho nijak nepretransformovala len zviditeľnila. V tomto prípade 1B a 3B nebola alternatívna hypotéza potvrdená, užívateľ neurobí drastickú chybu ako pri vymazaní duplicitných záznamov, tak i pri ich použití, avšak lepšie výsledky sú vo variante 3B.

4.2.3 2A Obalová zóna, K najbližších susedov

Dôvodom vytvorenia varianty s obalovou zónou sú skreslené výsledky v oblasti hranice mestskej časti. Na jej vytvorenie bol využitý nástroj obalová zóna s hodnotou vzdialenosti 100 m. Hodnotu obalovej zóny je potrebné voliť opatrne, opäť záleží od charakteru dát. Vzhľadom na pomerne veľký polomer mestskej časti bola vybraná taká hodnota, aby obsiahla primeraný počet subjektov zdravotníctva. V tejto obalovej zóne sa nachádza 44 zdravotníckych subjektov a štyri záznamy s lekárňou, ktoré však neboli vyšetrované. Postup kolokačnej analýzy bol rovnaký ako v predchádzajúcej variante, opäť bol postupne zvyšovaný počet najbližších susedných prvkov. Dalo sa očakávať, že už izolované lekárne sa prehĺbia do významne izolovaných, kvôli navýšeniu počtu záznamov susednej kategórie, a tým pádom navýšeniu celkového počtu vyšetrovaných prvkov. Faktom, však je, že analyzované lekárne sa nenachádzajú bezprostredne na hranici mestskej časti, takže by mohla byť táto varianta zbytočná. Každopádne cieľom bolo analyzovať čo najviac variácií vstupných dát a nastavení. Predpoklad sa potvrdil, nevýznamne kolokované lekárne boli častokrát premiestnené do kategórie významne izolovaných. Žiadna z lekární nebola významne kolokovaná. Je teda možné konštatovať, že ak nie sú analyzované záznamy bezprostredne na hranici, alebo v malej vzdialenosti, neodporúča sa používať prvky v obalovej zóne.

4.2.4 2B Obalová zóna, Vzdial. pásmo

Analýza započala s nastavením vzdialenostného pásma s hodnotou 20 m a následne bola hodnota zvyšovaná o desať metrov. Maximálna vzdialenosť pre analýzu opäť 100 m. Výsledok je podobný variante 1B, avšak počet nedefinovaných lekární je už od počiatku miernejší. Na začiatku sa mierne zvýšili nevýznamne kolokované lekárne. Od vzdialenostného pásma s hodnotu 70 metrov sa už vzor lekární nemení, ale dosahuje najvyšší počet kolokovaných bodov, a to päť. Po tejto hodnote už nezáleží ani na type jadrovej funkcie. V porovnaní s variantou 1B je táto varianta menej premenlivá v zmysle rozdelenia na nedefinované, izolované a kolokované body. To znamená, že prítomnosť subjektov zdravotníctva za hranicou Brno – město prispieva do analýzy pozitívne z hľadiska nedefinovaných lekární. Na druhej strane výsledok je veľmi podobný a opäť body v obalovej zóne nemajú veľký vplyv. Pri porovnaní s variantou 4B je pri 2B o niečo menej lekární kolokovaných. V rámci jadrových funkcií nenastávajú až také zmeny ako vo variante s duplicitnými záznamami. Viacmennej sú výsledky totožné, až na výnimku so vzdialenosťou 40 metrov a gaussovskou funkciou.

4.2.5 3A Duplicitné záznamy, K najbližších susedov

Už na prvý pohľad je vidieť rozdielnosť výsledkov v porovnaní s tým istým nastavením nástroja, ale bez duplicít (1A). Prvá analýza obsahuje dvoch K najbližších susedov. Lekárne bez blízkej prítomnosti ordinácií sú vyhodnotené ako izolované. Už je zrejmé, že prvé dve jadrové funkcie sú si podobné, absentujú nedefinované a významne izolované lekárne pri poslednej funkcii. Z teoretickej časti práce bolo zistené, že je to spôsobené nastavením, respektíve nenastavením vzdialenostných váh. Preto je potrebné myslieť na tento fakt pri interpretácií výsledkov, poprípade parametrizácií nástroja.

Pri zvýšení počtu prvkov a zvolení gaussovského jadra nastala primárna zmena dvoch lekární na Kobližnej ulici, nástroj ich vyhodnotil vzhľadom na definované okolie (tri susedné prvky) ako významne izolované s výrazne nízkou hodnotou LCLQ (0,340929 a 0,21805). Čo sa týka zmeny v dôsledku zvýšenia počtu prvkov, nedošlo k identifikovaniu významne kolokovaných lekární. Ďalšiu zmenu v porovnaní s 1A je možné spozorovať na lekárňach na Masarykovej ulici. Z izolovanej sa stala nevýznamne kolokovaná. Pri použití je možné pozorovať zmenu v okolí nádraží, lekárne sú nedefinované, čo znamená, že prvky nemali v stanovenom okolí požadovaný počet prvkov. No a v poslednom nastavení (žiadna), kedy nie je použitá váha prvkov je prekvapivo výsledok podobný gaussovskej funkci. Výnimkou sú významne izolované prvky, ktoré sú pozmenené na nevýznamne izolované.

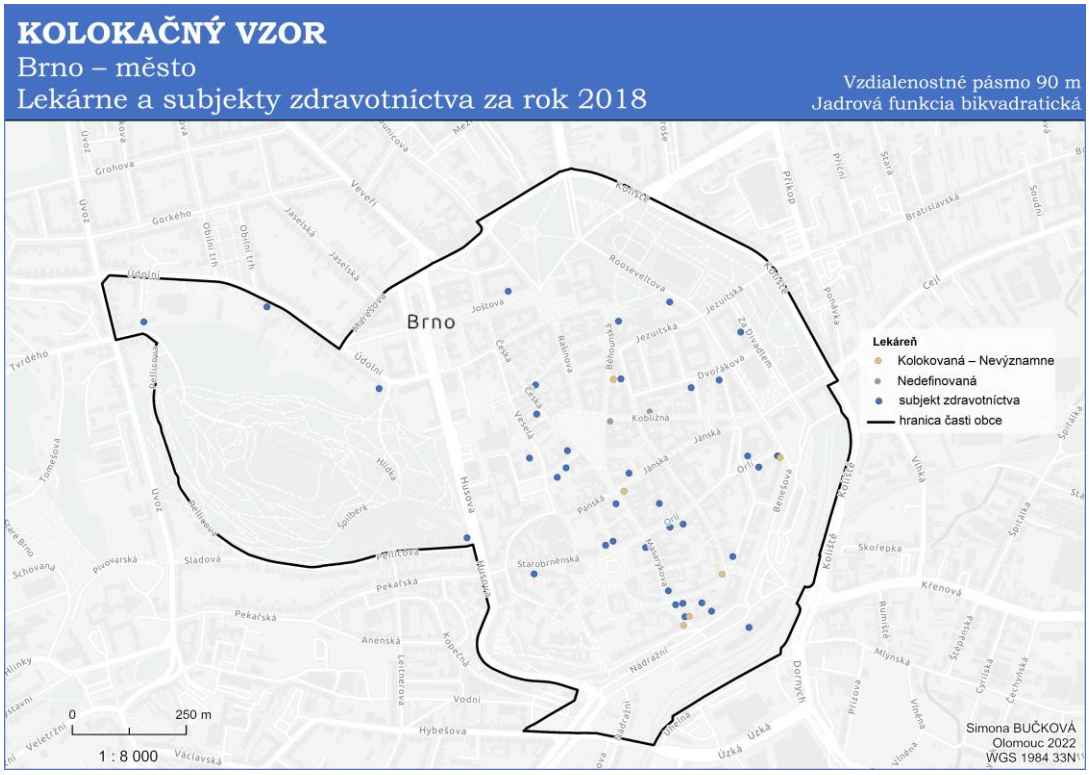
So zvyšovaním počtu susedných prvkov sa už neočakávajú lepšie výsledky. Keď bola hodnota počtu susedov zvýšená z tri na štyri susedné prvky, vzory nijak neprekvapili. Lekárne, ktoré nemali v okolí subjekty zdravotníctva boli pri funkciách s váženou vzdialenosťou (gaussovská, bikvadratická) zmenené o jeden stupeň, čiže z nevýznamne kolokované na nevýznamne izolované a nevýznamne izolované na významne izolované. Pri použití tretej funkcie sa výsledne zaradenie prvkov takmer nezmenilo. Čo sa týka lekární, ktoré boli v predchádzajúcom nastavení (tri susedné prvky) nevýznamne kolované, tam zmena nenastala. To znamená, že medzi skúmanými kategóriami nie je taký vzťah, aby boli lekárne vyhodnotené ako významne kolokované. Posledným nastavením bolo navýšenie susedov na hodnotu päť. Nenastala žiadna zmena v porovnaní s predchádzajúcim nastavením, to značí, že maximálne výsledky je možné pozorovať pri štyroch najbližších prvkoch. Zvýšenie tejto hodnoty nemá zmysel a v prípade extrémne veľkej hodnoty by mohol výsledok viesť ku falošným kolokačným

vzorom. Najviac kolokovaných lekární nástroj identifikoval pri dvoch až troch subjektoch, a to sedem pri gaussovskom jadre.

Druhá hypotéza sa nepotvrdila. Žiadna lekáreň nebola označená za významne kolokovanú, a to ani za prítomnosti duplicít. Zmena zaradenia lekární nadišla len v premene na významne a nevýznamne izolované. Najviac izolovaných lekární bolo získaných z 1A, čiže varianty, kde boli duplicitné subjekty odstránené. Avšak boli označené len ako nevýznamne izolované. Na druhú stranu pri analýze aj s duplicitnými subjektami vznikli aj významne izolované a dokonca aj nedefinované lekárne. Pri porovnaní metódy K najbližších susedných prvkov, ani pri postupnom zvyšovaní jej hodnoty neboli lekárne označené za významne kolokované, pribudli len opäť izolované. No a z hľadiska jadrových funkcií vo väčšine prípadov bolo výsledné zaradenie lekární veľmi podobné pri gassovskom a bikvadratickom jadre. Jadrová funkcia typu žiadna zväčša označila viac záznamov za izolované. Dôvodom bude fakt, že pri tomto jadre záznamy nie sú priestorovo vážené. Záver znie, že ak užívateľa len rozdelenie dát na kolokované a izolované, tak sa odporúča duplicity odstrániť. Ak však sú predmetom analýzy aj nedefinované body, je vhodnejšie duplicitné záznamy ponechať.

4.2.6 3B Duplicitné záznamy, Vzdial. pásmo

Cieľom ponechania duplicitných bodov bolo overiť druhá hypotézu, ktorá predpokladá významné kolokačné vzory z dôvodu ich prítomnosti. Krátke porovnanie je zhrnuté v sekcii 1B, kde bola zistená len mierna zmena oproti variante bez duplicitných subjektov zdravotníctva. V podstate prítomnosť duplicitných bodov spôsobila väčšiu variabilitu vo výsledkoch analýzy. Bolo identifikovaných menej nevýznamne izolovaných bodov, avšak otázkou je čo užívateľ od analýzy očakáva. Niekoho môžu zaujímať len kolované body, iného užívateľa aj tie izolované. Z celkového hľadiska bolo identifikovaných viac kolokovaných lekární, avšak maximálne o dve. Najviac lekární bolo označených za kolokované pri vzdialenosti 90 m, a to sedem (viď Mapa 3).



Mapa 3 Kolokačný vzor pre lekárne a subjekty zdravotníctva za rok 2018

4.2.7 4A Duplicitné záznamy a obalová zóna, K najbližších susedov

V tomto nastavení šlo predovšetkým o porovnanie varianty, kde neboli duplicitné záznamy v pozorovanej mestskej časti (2A). Z predchádzajúcich zistení je zrejmé, že ak je v tejto analýze oveľa viac záznamov, tak vzniká predpoklad navýšenia izolovaných lekární. K celkovému zvýšeniu nedošlo, jedine k zmene z nevýznamne izolovanej lekárne na významne izolovanú. Tento trend je možné pozorovať vo viditeľne osamotených lekárňach. Na druhú stranu nečakaným faktom je, že došlo k zlepšeniu, čo sa týka zvýšenia počtu nevýznamne kolokovaných lekární. Každopádne ich maximálny limit prevýšený nebol, najviac ich bolo sedem pri dvoch susedných prvkoch. Pri porovnaní s variantou 3A, zo začiatku boli výsledky úplne totožné. Zmena nastala až pri troch K najbližších susedoch a poslednej jadrovej funkcii avšak nič prevratné, nevýznamne izolované lekárne sa zmenili na významne izolované. Pri ďalšom navyšovaní sa výsledky opäť zhodovali. Z tohto dôvodu je možné prehlásiť, že prítomnosť obalovej zóny nemá takmer žiaden vplyv na výsledné zaradenie lekární. Druhá hypotéza je zamietnutá, významne kolokované lekárne identifikované neboli.

4.2.8 4B Duplicitné záznamy a obalová zóna, Vzdial. pásmo

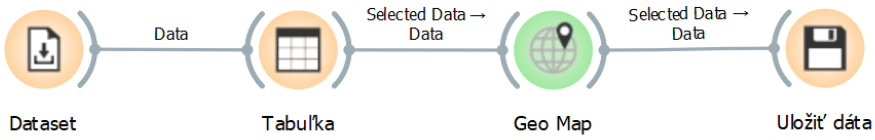
Duplicitné záznamy už boli vyšetrované niekoľkokrát, úlohou bolo potvrdiť, alebo zamietnuť druhú hypotézu. Z výsledkov analýzy je možné pozorovať, že duplicitné záznamy mierne ovplyvňujú výsledok, zvyšujú počet kolokovaných lekární. Nie sú to ale prevratné zmeny, nastávajú prevažne v kolísaní počtu nevýznamne izolovaných a kolokovaných bodov o jeden bod. Efekt obalovej zóny nie je pozitívny. Vzhľadom na navýšenie počtu bodov v susednej kategórii, došlo k zníženiu nevýznamne kolokovaných lekární.

5 PRÍPADOVÁ ŠTÚDIA 2 – FILADELFIA

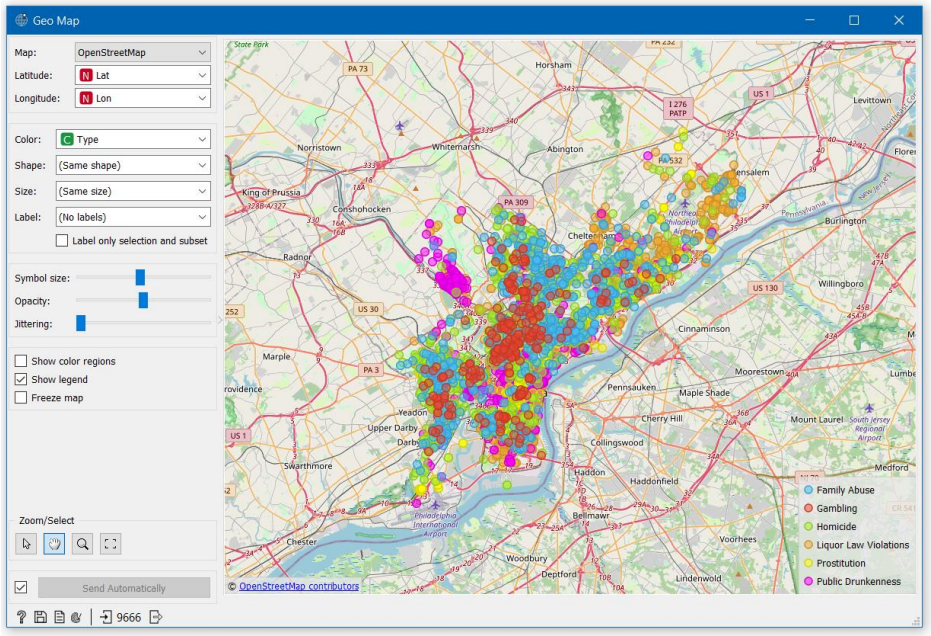
V tejto štúdii bude použitý dataset kriminálnych činov na území Filadelfie, USA. Dáta sú aktuálne k roku 2016 a boli získané pomocou softwaru Orange. Dátová sada Philadephia Crime obsahuje 9666 činov v rozmedzí rokov 2006 až 2016. Atribútová tabuľka obsahuje hodnoty zemepisnej šírky a dĺžky, dátum a typ kriminálneho činu: domáce násilie, gambling, vražda, porušenie zákona o pití alkoholu na verejnosti, prostitúcia a opilstvo na verejnosti.

5.1 Postup získania datasetu

Ako už bolo zmienené dáta boli získané v prostredí Orange, ale poskytuje ich aj stránka mesta s názvom OpenDataPhilly (2022), kde sú voľne stiahnuteľné. Prvým krokom vo workflow (viď Obr. 8) bolo použitie widgetu dataset *(Datasets)*, kde bol zvolený požadovaný dátový zdroj. Proces mohol byť skončený jednoduchým widgetom uložiť dáta *(Save Data)*, ale pre prvotné zhliadnutie dát boli zahrnuté ďalšie dva widgety, a to dátová tabuľka *(Data Table)* a geo mapa *(Geo Map)*, ktorý dokáže dáta zobraziť (viď Obr. 9). Dáta boli uložené vo formáte CSV a následne nahrané do prostredia ArcGIS Pro prostredníctvom nástroja xy tabuľka do bodov *(XY Table To Point)*.



Obr. 8 Workflow pre získanie datasetu (zdroj: autorka)



Obr. 9 Prvotné zobrazenie dát cez widget Geo Map (zdroj: autorka)

Použité vrstvy:

- Primárne
    - Kriminálne činy – bodová vrstva
    - Bary a puby – bodová vrstva
- Doplnkové
    - Policajné oblasti – polygónová vrstva

Vrstvy barov a pubov na území mesta Filadelfie boli získané pomocou nástroja QuickOSM v prostredí programu QGIS. Ako už názov nástroja napovedá, dáta pochádzajú z OpenStreetMap. Pre ich získanie je potrebný názov kľúča a hodnota, čo je v tomto prípade „*amenity*“ a „*bar*“ či „*pub*“. Posledným krokom bol export dát do formátu SHP *(Shapefile)* a nahranie do softwaru ArcGIS Pro.

Pre popis a bližšie určenie výskytu kriminality boli použité policajné oblasti, kde každá oblasť spadá pod konkrétneho policajného kapitána. Mesto je rozdelené na šesť častí podľa svetových strán, ktoré sa ďalej delia a majú svoj kód (OpenDataPhilly, 2014).

5.2 Postup úpravy dát

Dátová vrstva je automaticky nahraná do projektu v súradnicovom systéme WGS 1984 *(World Geodetic System)*, no ak by bol nástroj kolokačná analýza spustený s týmto súradnicovom systémom, bude hlásiť varovanie kvôli stupňom, ktoré systém používa. Preto je potrebné zmeniť súradnicový systém na UTM *(Universal Transverse Mercator)*, Filadelfia sa nachádza v pásme 18N. Na to je možné využiť nástroj konvertovať súradnice *(Convert Coordinate Notation)* a potom sa budú body zobrazovať správne. Druhým krokom je tvorba nového stĺpca v atribútovej tabuľke. Vrstva obsahuje dátum, ale nie je v správnom časovom formáte. Nástroj konvertovať časové pole *(Convert Time Field)* síce neprebehne, ale vytvorí stĺpec vo správnom formáte. Potom už stačí len v atribútovej tabuľke nastaviť hodnoty z pôvodného dátumového stĺpca pomocou vypočítať pole *(Calculate Field)*.

V tejto štúdii bolo hlavným cieľom preskúmať časovú zložku kriminálnych činov. K tomu je kľúčový atribút „*Datetime*“, ktorý je avšak pôvodne uvedený ako textová hodnota. Pre následné použitie času v kolokačnej analýze bolo nutné vytvoriť nový atribút so správnym dátovým typom, a to dátum, postup je spomínaný vyššie. Ďalej bolo nutné zamyslieť sa nad vhodným časovým krokom pre skúmanie kriminálnych činov v rámci variant tejto štúdie. V prípade skúmania kolokačného vzoru medzi činmi, barmi a pubmi, čas nehrá dominantnú rolu. Časový údaj by mohol byť teoreticky nápomocný pri výbere podnikov, ktoré v danom období existovali a boli v prevádzke, avšak takéto dáta neboli dostupné. Čas ale môže byť nápomocný pri vyhodnocovaní sily vzťahu medzi jednotlivými kriminálnymi činmi a či sú činy z hľadiska času kolokované.

Dáta o kriminálnych činoch sú v rozmedzí desiatich rokov. V procese skúmania kolokačných vzorov bol použitý dataset ako taký za všetky roky, ale taktiež bol rozdelený na konkrétne obdobia. Dôvodom bolo skúmanie rôznych veľkostí časových okien a výsekov vstupných dát. Záznamy s konkrétnym časovým obdobím boli vybrané pomocou výber vrstvy podľa dátumu a času *(Select Layer By Date And Time)* a následne z nich bola vytvorená samostatná vrstva vo formáte SHP.

5.3 Kolokácia vrážd a domáceho násilia

Prvou variantnou analýzou kriminálnych činov je preskúmanie existencie vzťahu medzi vraždami a záznamami o domácom násilí. Podľa prieskumu od Centers for Disease Control and Prevention (Petrosky a kol. 2017) je viac než 55 % vrážd spojených s partnerským násilím a 11,2 % obetí vrážd zažívali mesiac pred svojím úmrtím rôzne formy násilia. V dátach síce nie je uvedená žiadna podrobnosť kriminálneho činu, avšak nie je vylúčené, že tento vzťah sa v dátach nevyskytuje. Čo sa týka výberu vhodného časového obdobia, v prvom kroku boli vybrané všetky záznamy za jednotlivé roky. Prehľad o počte záznamov za vraždu a domáce násilie si je možné prehliadnuť v Tab. 3.

Tab. 3 Vývoj počtu záznamov pre vybrané kategórie

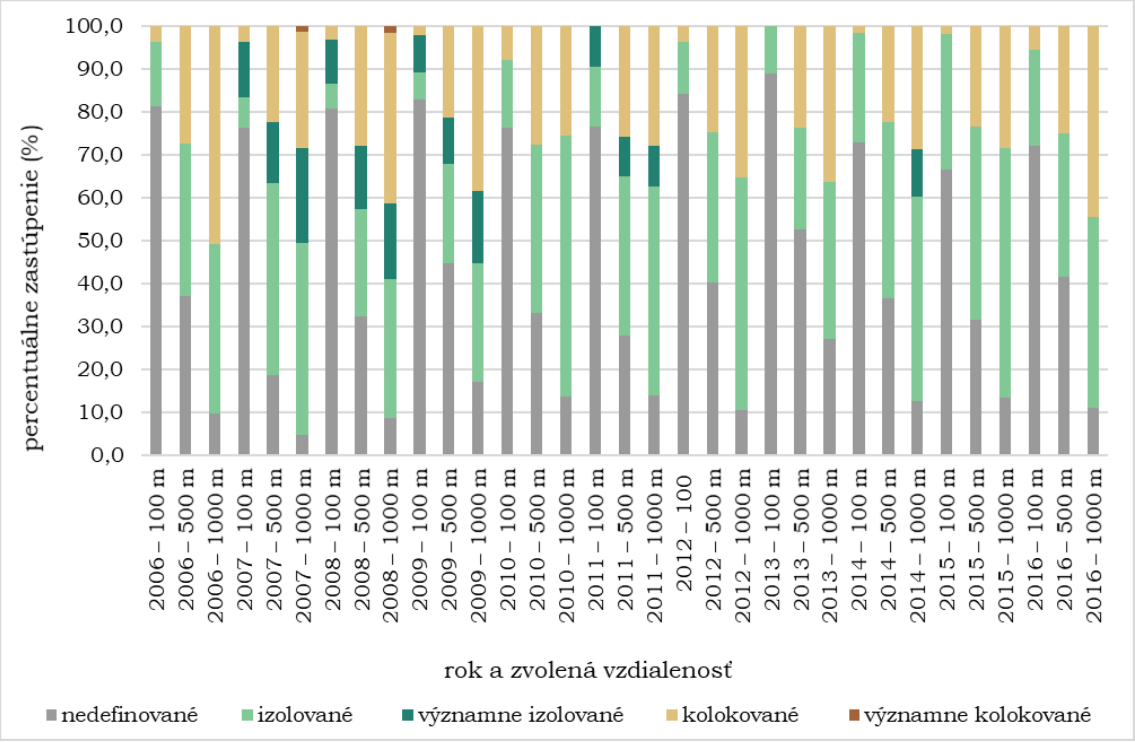
rok	**domáce násilie**	**vraždy**	**spolu**
2006	81	140	221
2007	85	120	205
2008	68	110	178
2009	47	114	161
2010	51	97	148
2011	43	119	162
2012	57	107	164
2013	55	71	126
2014	63	75	138
2015	60	96	156
2016	36	79	115

Je zrejmé, že nie všetky vraždy súvisia s domácim násilím, avšak práve touto analýzou je možné odhaliť skryté vzťahy. V tomto prípade je dôležitý parameter interval časového okna *(Time Step Interval)*, kde sa zadáva veľkosť časového okna, ktoré bude podmienkou pri skúmaní kolokačného vzťahu. Je možné zadať akúkoľvek hodnotu od sekundy až po rok. Dataset je rozdelený po rokoch a momentálnym cieľom je skúmať vzťahy za jednotlivé roky, no taktiež preskúmať menšie časové obdobie. Preto bude parameter nastavený najprv na jeden rok a potom na jeden mesiac. Druhým dôležitým parametrom je typ časového vzťahu *(Temporal Relationship Type)*, obsahuje tri možnosti, a to predtým *(before)*, potom *(after)* a časové okno obsahujúce predtým aj potom *(span)*. Neopomenuteľnou nutnosťou je aj voľba metódy, na základe ktorej bude kolokačný vzťah zostavený. V tejto variante bude využitá len metóda vzdialenostného pásma. Otázkou však je aká vzdialenosť je relevantná a naopak pri akej vzdialenosti už kolokačný vzor môže byť falošne pozitívny. V teoretickom scenári, ak by v domácnosti dochádzalo k násiliu je pravdepodobné, že by zabitie nastalo na tom istom mieste. To by znamenalo, že vzdialenosť kriminálnych činov by bola minimálna, respektíve nulová. Samozrejme to nie je pravidlom, nemusí dôjsť k vražde na totožnom mieste, a preto bude použitých niekoľko vzdialeností, a to 100 m, 500 m a 1000 m. Takto vznikne pre každý dostupný rok šesť vrstiev kolokačných vzorov.

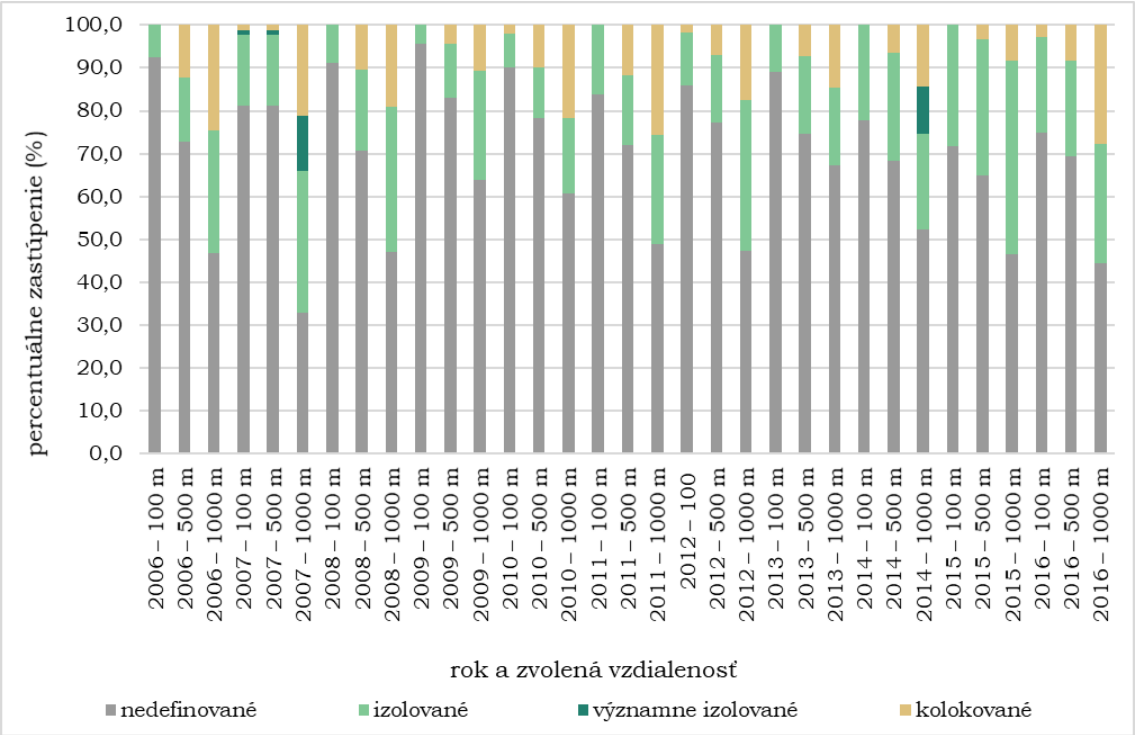
Pri prvotnom pohľade na samotné dáta je zrejmé, že počet vrážd je v každom roku vyšší než záznamov o domácom násilí. Celkovo za všetky roky bolo na území Filadelfie zaznamenaných 646 hlásení o domácom násilí a 1128 záznamov vrážd. Oba trestné činy majú v čase prevažne klesajúcu tendenciu. Mierny nárast vrážd nastal v roku 2011, prípadov s domácim násilím naopak ubudlo. Naopak v roku 2012 stúpol počet

násilí a klesli záznamy s vraždou. Čo sa týka rozdelenia kriminality medzi jednotlivé mesiace, je tam istá tendencia častejšieho výskytu v letných mesiacoch. Výraznejšie je táto tendencia viditeľná pri vraždách, v oboch zločinoch je najmenší výskyt v zimných mesiacoch.

Ako už bolo spomenuté vyššie, bolo využitých niekoľko hodnôt vzdialenostného pásma a časového intervalu. Nie je prekvapením, že so zväčšujúcou sa vzdialenosťou pribúda počet kolokovaných bodov (viď Obr. 10). Inak to nie je ani pri časovom okne, pri použití jedného mesiaca z výsledkov vymizli významne kolokované záznamy domáceho násilia (viď Obr. 11). Vo všeobecnosti je možné povedať, že čím je väčšia vzdialenosť alebo časové okno, tým väčšia šanca je, že nástroj vyhodnotí body záujmu ako kolokované. Avšak vo väčšine sú len kolokované nevýznamne. Len v dvoch konkrétnych nastaveniach bol jeden bol vyhodnotený ako významne kolokovaný s hodnotami lokálneho kolokačného kvocientu 1,7 a 1,5. Jedná sa o roky 2007 a 2008 so vzdialenostným pásmom 1000 metrov a časovým oknom jedného roka. Miesto incidentu nie je totožné, vzdušnou čiarou sú od seba vzdialené približne 2 533 metrov. V oboch prípadoch je v okolí spomínaných záznamov pomerne veľký počet záznamov s vraždami.

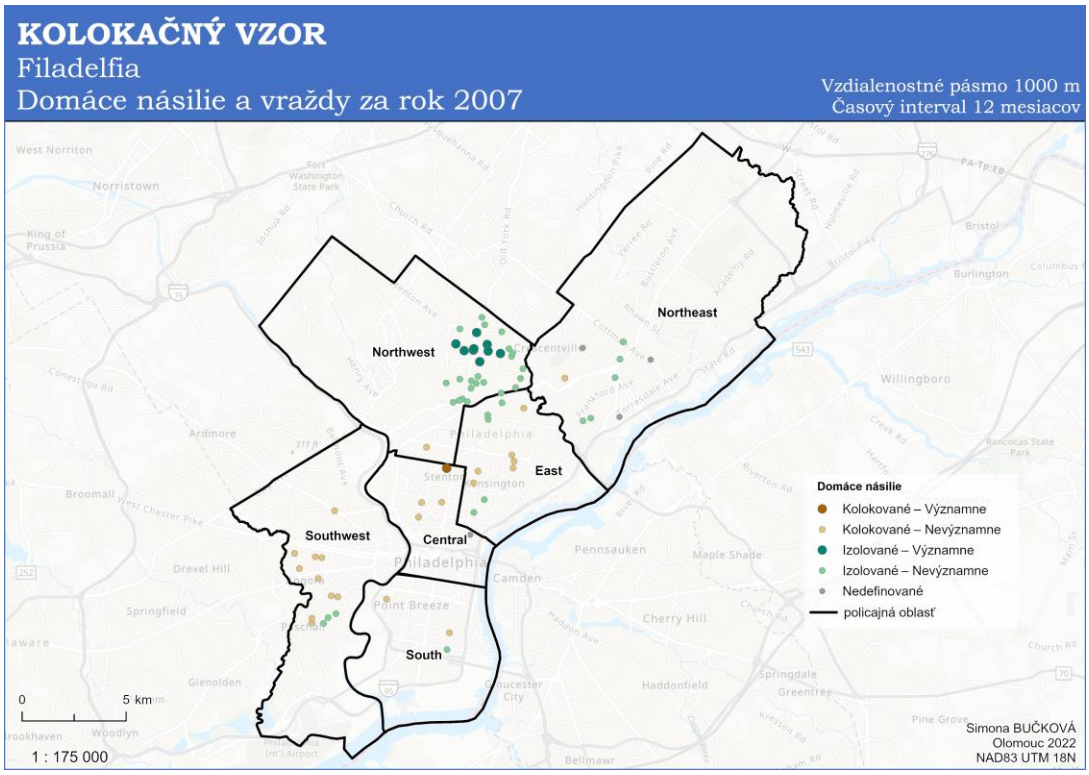


Obr. 10 Percentuálne zaradenie kategórii kolokačných vzorov kriminálnych činov 2006–2016 s časovým oknom jedného roka (zdroj: autorka)

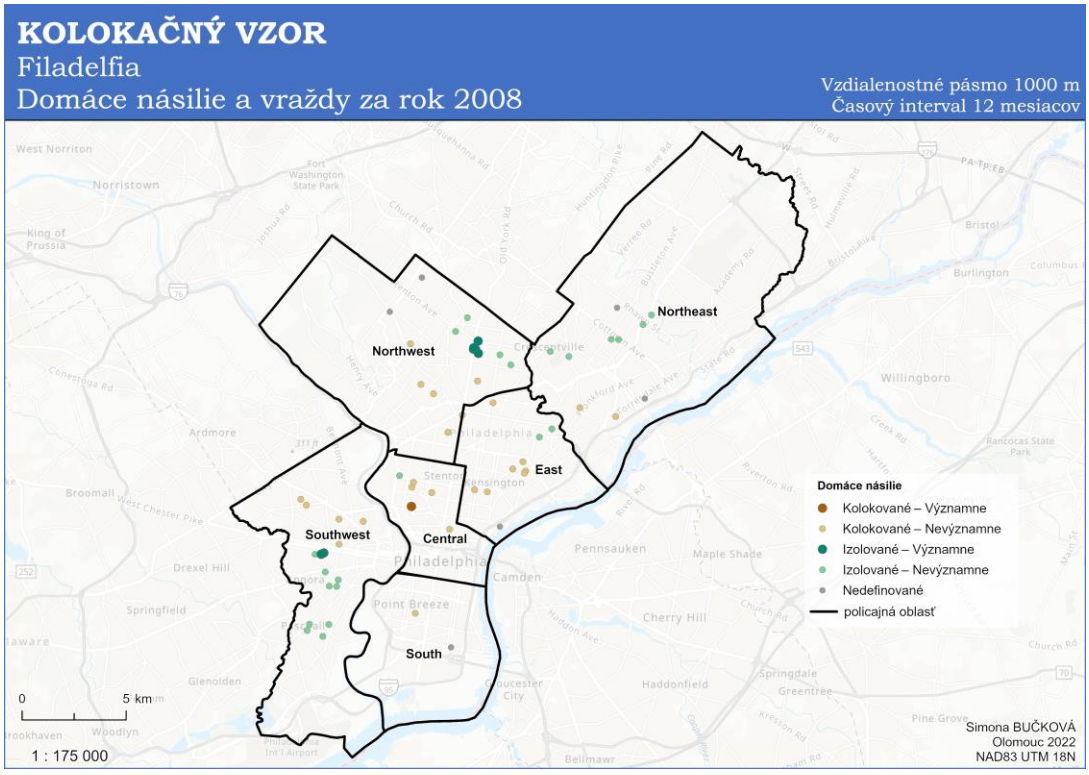


Obr. 11 Percentuálne zaradenie kategórii kolokačných vzorov kriminálnych činov 2006–2016 s časovým oknom jedného mesiaca (zdroj: autorka)

Prvý významne kolokovaný incident sa stal 28. 9. 2007 o 4:30 ráno v Centrálnej oblasti (viď Mapa 4). Všetky blízke okolité záznamy vraždy sa udiali maximálne tri mesiace pred a po incidente. Druhý incident nastal 19. 11. 2008 20:25 miestneho času, taktiež v Centrálnej oblasti (viď Mapa 5). Opäť takmer všetky okolité vraždy boli zaznamenané približne v intervale troch mesiacov. Nie je to však podmienkou, aby sa určitý bod domáceho násilia stal významne kolokovaný. Pravdepodobne to bude súhrou viacerých skutočností ako je počet okolitých bodov susednej kategórie, časový aspekt a podobne.



Mapa 4 Kolokačný vzor pre domáce násilie a vraždy za rok 2007



Mapa 5 Kolokačný vzor pre domáce násilie a vraždy za rok 2008

Vo všeobecnosti je možné konštatovať, že medzi domácim násilím a vraždami existuje istý vzťah. Hovorí o tom aj štúdia, na ktorej je táto varianta založená, avšak aj tá poukazuje na pomerne malé percento vrážd medzi prípadmi domáceho násilia. Vo väčších vzdialenostiach množstvo kolokovaných bodov stúpa, avšak opäť je vzťah

označený len ako nevýznamne kolokovaný. Percentuálne počet nevýznamne kolokovaných bodov dosahuje maximálne 50 %, avšak vo väčšine sa pohybuje okolo 30 %. Rovnako stúpa počet kolokovaných bodov pri použití väčšieho časového okna. Otázkou však je aká vzdialenosť a časový interval je relevantný pre konkrétnu analýzu, dáta a podobne. Táto varianta sa opiera o výsledky štúdie v Spojených štátoch amerických z rokov 2003–2014, kedy jedna z desiatich žien zažívala domáce násilie pred svojou vraždou. Každopádne časový aspekt v článku zmienený nebol, a preto boli vybrané dve výrazne odlišné časové okná. Z výsledkov je zrejmé, že jeden mesiac medzi domácim násilím a vraždou nehrá veľkú rolu pri zločinoch za jeden rok. Významné hodnoty lokálneho kolokačného kvocientu sa začali vyskytovať až pri použití časového okna s hodnotou piatich a deviatich mesiacov. Z toho vyplýva, že aj pri rovnakom nastavení vzdialenostného pásma sa mení dĺžka časového okna, v ktorej sú body významne kolokované. Sila kolokačného vzoru je výrazne závislá na charaktere dát, avšak pri študovaní dát sa častokrát v rámci jednej témy či jedného dátového súboru používa rovnaký interval, hodnota a podobne. Dôvodom je ich porovnanie, preto ak by boli dáta vyšetrované s použitím šiestich mesiacov, výsledky by boli opäť trochu iné a významne kolokovaný bod by bol zachytený len v roku 2007.

Úplne rozdielne a výrazne zaujímavejšie výsledky boli získané pri použití jednej vstupnej dátovej vrstvy s obsahom všetkých dostupných rokov. Pri časovom okne jedného roka a vzdialenostného pásma 1000 metrov je 30 % záznamov nevýznamne kolokovaných a 23 % významne kolokovaných. To je výrazný rozdiel oproti výsledkom za jednotlivé roky. Netreba však opomenúť, že výstupy po rokoch sú ochudobnené, čo sa týka času, ale aj priestoru. Opäť záleží na užívateľovi, čo je jeho cieľom, aké časové či priestorové rozmedzie chce užívateľ skúmať.

Vo výsledku za všetky roky sa pri nastavení jedného mesiaca signifikantne znížil počet významne kolokovaných bodov, celá výstupná vrstva obsahuje len 12 bodov z 646. Okrem počtu bodov v jednotlivých kolokačných kategóriách je možné pozorovať aj priestorovú distribúciu. Počet bodov domáceho násilia klesá smerom k hraniciam mesta. Taktiež tieto záznamy zaradené do rovnakej kategórie majú tendenciu sa zhlukovať. To znamená, že existuje aj vzťah medzi záznamami samotnými a je medzi nimi priestorová závislosť. Taktiež je možné konštatovať, že vo Filadelfii sú isté oblasti, ktoré sa vyznačujú priestorovou súvislosťou medzi domácim násilím a vraždami. Typicky sú v mestách oblasti, ktoré sú označované ako menej bezpečné. Napríklad z dôvodu zhlukovania sa sociálne slabších jedincov ako sú bezdomovci, opilci a podobne. Taktiež existujú oblasti vyššej kriminality a pocitu menšej bezpečnosti či strachu. Práve tieto miesta môže nástroj odhaliť a vyhodnotiť silu vzťahu medzi rôznymi typmi kriminálnych činov.

5.4 Kolokácia kriminálnych činov s barmi a pubmi

Nie je tajomstvom, že miesta, kde sa holduje alkoholu sú miestami sporov, bitiek a dokonca aj vážnych kriminálnych činov. Elizabeth Groff (2011) vo svojom článku spomína, že existuje spojenie medzi prítomnosťou miest ako sú bary, školy, parky či reštaurácie a zvýšenou kriminalitou. Vo svojej štúdii pre analýzu kriminálnych činov využíva euklidovskú obalovú zónu a obalovú zónu s použitím uličnej siete. S použitím uličnej siete stanovuje vzdialenosť na základe dĺžky blokov v danom meste. V závere navrhuje používať obalovú zónu s dĺžkou práve jedného bloku v študovanej oblasti.

Na základe tohto článku bola v analýze použitá metóda vzdialenostného pásma so vzdialenosťou práve jedného bloku a jeho násobkov. Podľa digitálneho vydavateľa

Reference (2020) sa dĺžka jedného bloku vo Filadelfii pohybuje od 400 (121,9 m) do 500 stôp (152,4 m). Pomocou nástroja meranie *(Measure)* bola overená táto hodnota na blokoch podkladovej mapy. Dĺžka bloku nie je rovnaká v celom meste, pohybuje sa okolo 145 m. Ako konečná vzdialenosť bola zvolená stredná hodnota, čo je 450 stôp a v prepočte 137 metrov.

Podľa E. Groff (2011) sú najlepšie výsledky analýzy kriminality zachytené pri vzdialenosti 122 metrov, čo je dĺžka jedného bloku. Na základe toho vznikla hypotéza:

Najväčší počet kolokovaných kriminálych činov je pri vzdialenosti jedného bloku.

Z predošlej štúdie je už zrejmé, že s pribúdajúcou vzdialenosťou počet kolokovaných záznamov stúpa, aj keď pri určitej hodnote sa výsledky ustália. Preto je možné očakávať zvyšujúcu sa tendenciu kolokovaných bodov aj v tomto prípade. Z tohto dôvodu bola vytvorená druhá hypotéza, ktorá predpokladá vyšší počet významne kolokovaných záznamov práve v okolí barov a pubov. Jej znenie je nasledovné:

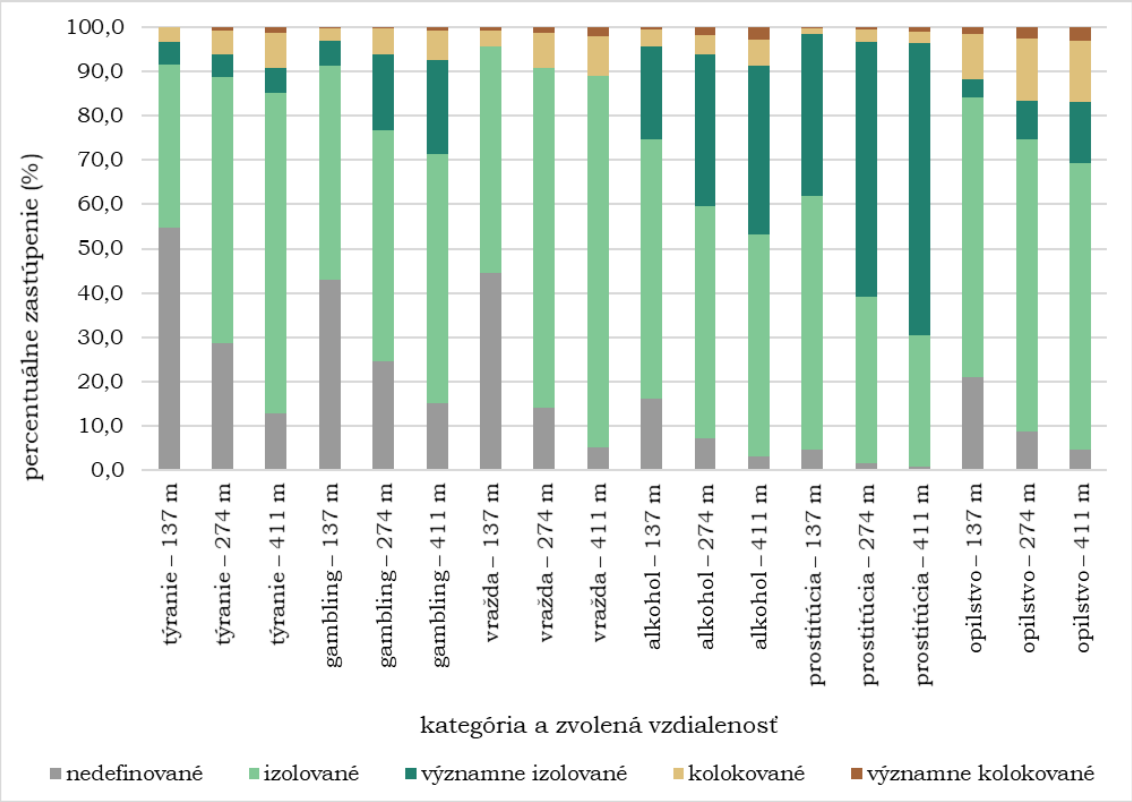
Najväčší počet kolokovaných kriminálych činov je v okolí barov a pubov.

Čo sa týka časovej zložky vstupného datasetu, tentokrát nebude rozdelený na obdobia, vstupom bude celá dátová sada. Susednú kategóriu predstavuje jeden spoločný dataset pre bary a puby, dáta sú aktuálne k roku 2022. V tomto prípade nebude použitá voľba veľkosti časového okna, nie je cieľom skúmať vzťahy medzi jednotlivými kriminálnymi činmi, ale vzťah medzi prevádzkami a kriminálnymi činmi.

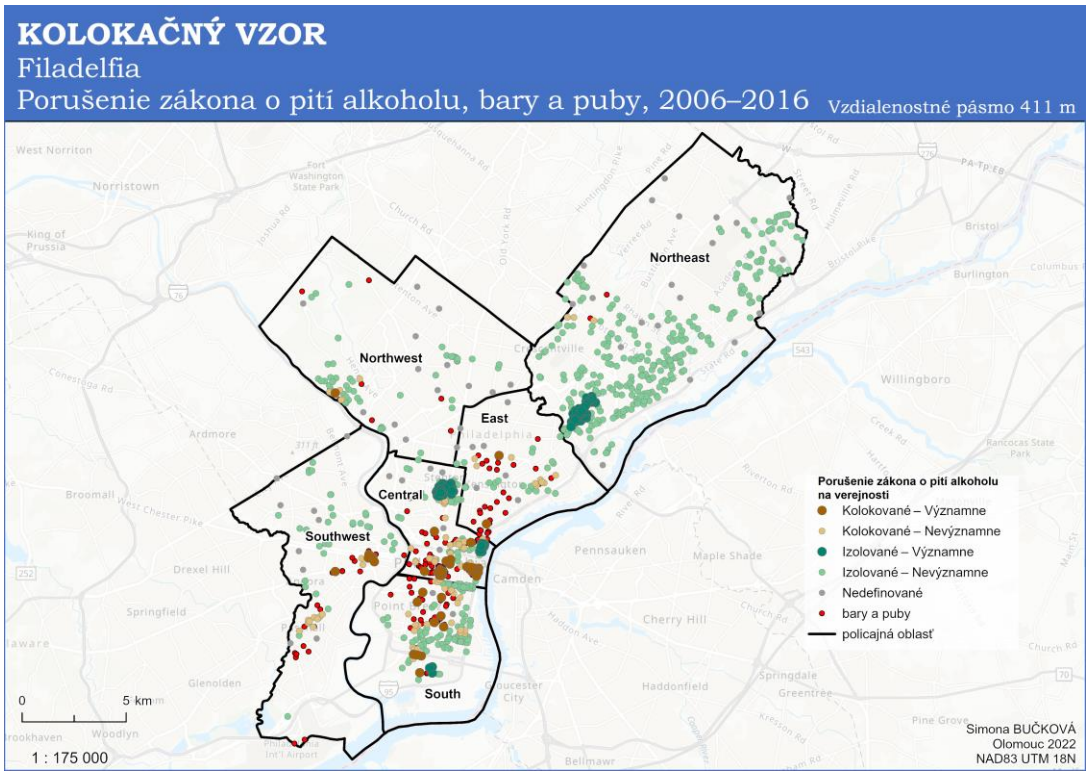
V dátovej sade je dostupných šesť kriminálnych činov, každý z nich je špecifický svojou priestorovou rozloženosťou. Veľmi výraznú priestorovú podmienenosť majú záznamy prostitúcie, body opticky vytvárajú línie pozdĺž komunikácií. Viditeľnú priestorovú a logickú súvislosť majú aj záznamy o porušení zákona o pití alkoholu na verejnosti a opilstvo na verejnosti. V ich okolí sa častokrát vyskytuje aj prostitúcia, ktorá býva spojená s pitím alkoholu, a tým pádom aj s porušením zákona o pití na verejnosti. Gambling nemá na prvý pohľad priestorový vzťah s ostatnými typmi kriminálnych činov, avšak má väčší výskyt v Severnej a Spodnej Severnej oblasti. Čo sa týka vrážd, tie sú viacmennej rovnomerne priestorovo rozdelené, predovšetkým v centrálnej, severnej, západnej a severozápadnej oblasti. Občas sú pozorovateľné malé zhluky, ale nie tak výrazne ako napríklad u zločinov spojených s alkoholom.

Výsledkom je 18 kolokačných vzorov so zväčšujúcou sa hodnotou vzdialenostného pásma *(Distance band)*, a to 137, 274 a 411 metrov. Výsledky percentuálneho zaradenia kriminálnych činov do jednotlivých kategórií sú uvedené v Obr. 12. Hypotéza sa nepotvrdila ako už bolo predom predpokladané. Nástroj so zväčšujúcou sa vzdialenosťou zaradil viac bodov ako kolokovaných či už významne alebo nevýznamne. Preto nie je možné usúdiť, že práve vzdialenosť jedného bloku je najvhodnejšia pre vyhodnocovanie sily vzťahu medzi kriminálnymi činmi, barmi a pubmi. Vhodnou voľbou bola alternatívna hypotéza, z výsledkov je očividné, že v okolí podnikov sú lokalizované práve kolokované body. V miestach, kde sa bary nenachádzajú boli body označené za izolované. Niektoré kriminálne činy sú však výraznejšie viac kolokované, čo je logické. Zločiny spojené s alkoholom budú viac asociované s miestami kde sa podáva, než napríklad domáce násilie či vraždy. Najviac významne kolokovaných bodov vykazovalo opilstvo na verejnosti a porušenie zákona o pití alkoholu na verejnosti. Na druhej strane jedná sa len o 2,7 % (viď Mapa. 6) a 2,9 % (viď Mapa. 7) zo všetkých vstupných bodov danej kategórie. Nevýznamne kolokovaných potom bolo o čosi viac, 6 % zo záznamami o porušení zákona a 14 % o opilstve. Každopádne drvivá väčšina bodov bola zaradená

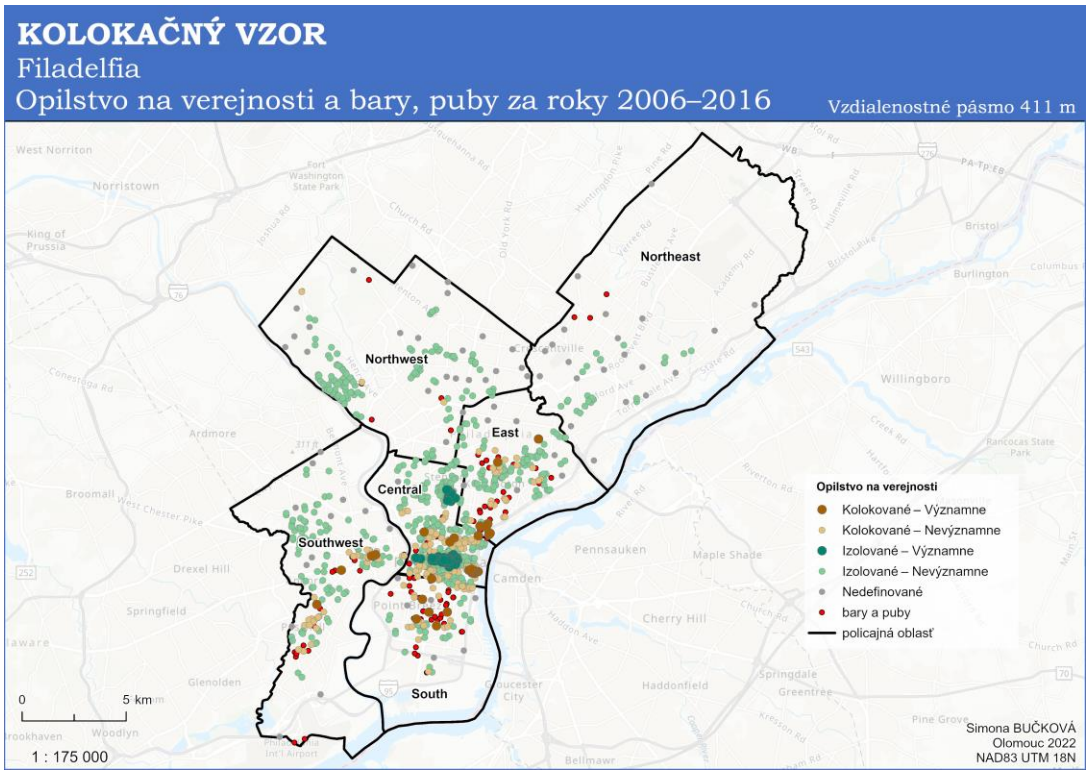
medzi izolované, čo by mohlo znamenať, že je tam ešte iná priestorová závislosť ako napríklad obytné zóny sociálne slabších a podobne.



Obr. 12 Prehľad o percentuálnom zaradení kategórii kriminálnych činov 2006–2016 (zdroj: autorka)



Mapa 6 Kolokačný vzor pre porušenie zákona o pití alkoholu na verejnosti za roky 2006–2016 porušení zákona o pití alkoholu na verejnosti a opilstvo na verejnosti



Mapa 7 Kolokačný vzor pre opilstvo na verejnosti za roky 2006–2016

5.5 Kolokácia kriminality a alkoholu

S. E. M van Sleeuwen a kolektív (2020) skúmali časovú konzistenciu kriminálnych vzorov. Kládli si otázku, kedy zločinec pácha zločin. Autori začali s hypotézou, že páchateľ koná trestný čin v obdobnom čase a týždni. Typickým príkladom je vlámanie a krádež v rezidenčnej oblasti počas dňa, keď sú ľudia v práci. Taktiež očakávali silnejší časový vzťah zločinov podobného typu, ktoré boli spáchané v kratšom časovom intervale. Hypotézu autori overovali na viac než 28 000 záznamoch z obdobia ž1996–2009. Výsledkom bolo jej potvrdenie a zistenie, že práve do jedného mesiaca je vzťah kriminálnych činov najsilnejší.

Z týchto zistení je možné zostaviť novú hypotézu nasadenú na dostupnú dátovú sadu. Zločiny, ktoré sú si podobné sa budú diať v podobnom čase a mohli by mať medzi sebou silný kolokačný vzťah. Takýmito zločinmi z dostupných v dátovej sade sú porušenie zákona o pití alkoholu na verejnosti a opilstvo na verejnosti, alebo gambling a opilstvo na verejnosti. Všetky tieto zločiny sa dejú prevažne večer, to znamená, že medzi nimi existuje súvislosť – vzťah. Konečná hypotéza znie nasledovne:

Podobné, alebo súvisiace zločiny sa dejú v obdobnom čase a majú signifikantný kolokačný vzťah pri časovom intervale jedného mesiaca.

Dvojice zločinov boli skúmané v jednej dátovej sade za všetky roky, nie potrebné dáta rozdeľovať, keďže kľúčová je voľba časového okna. Dátová sada bola skúmaná v rámci jedného, troch a šiestich mesiacov práve z dôvodu znenia hypotézy. V prípade, že skúmané kategórie kriminálnych činov majú medzi sebou kolokačný vzťah, ich vzťah bude silnejší v rámci jedného mesiaca než dlhšieho časového obdobia.

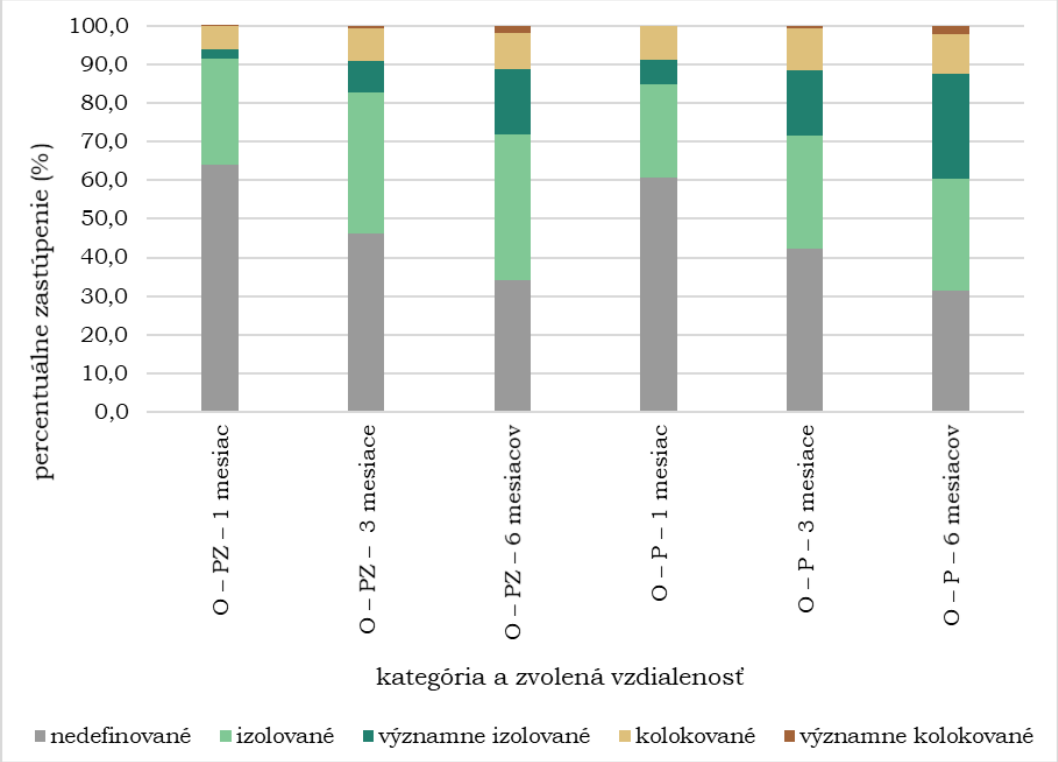
Z pôvodného datasetu kriminality boli extrahované záznamy podľa atribútu kriminálneho činu s pomocou nástroja výber podľa atribútov *(Select By Attributes)*. Takto vznikla jedna vstupná vrstva obsahujúca obidve kategórie pre všetky dostupné

roky, čo znamená, že väčšina rokov nebude okradnutá o hraničné mesiace, teda na roky 2006 a 2016. Pre túto variantu boli vybrané kategórie porušenie zákona o pití alkoholu na verejnosti, prostitúcia a opilstvo na verejnosti. Následne boli zostavené dvojice:

- 1. dvojica kriminálnych činov
  - opilstvo na verejnosti
  - porušenie zákona o pití alkoholu na verejnosti
- 2. dvojica kriminálnych činov
  - opilstvo na verejnosti
  - prostitúcia

Potom už bola menená len hodnota časového okna, vzdialenosť bola nastavená pevne na 411 m, ktorá je odvodená z prechádzajúcej varianty, kde bolo najviac kolokovaných bodov pri vzdialenostnom pásme 411 m. Cieľom bolo preskúmanie časového aspektu podobných zločinov, preto nebolo vzdialenostné pásmo bližšie skúmané.

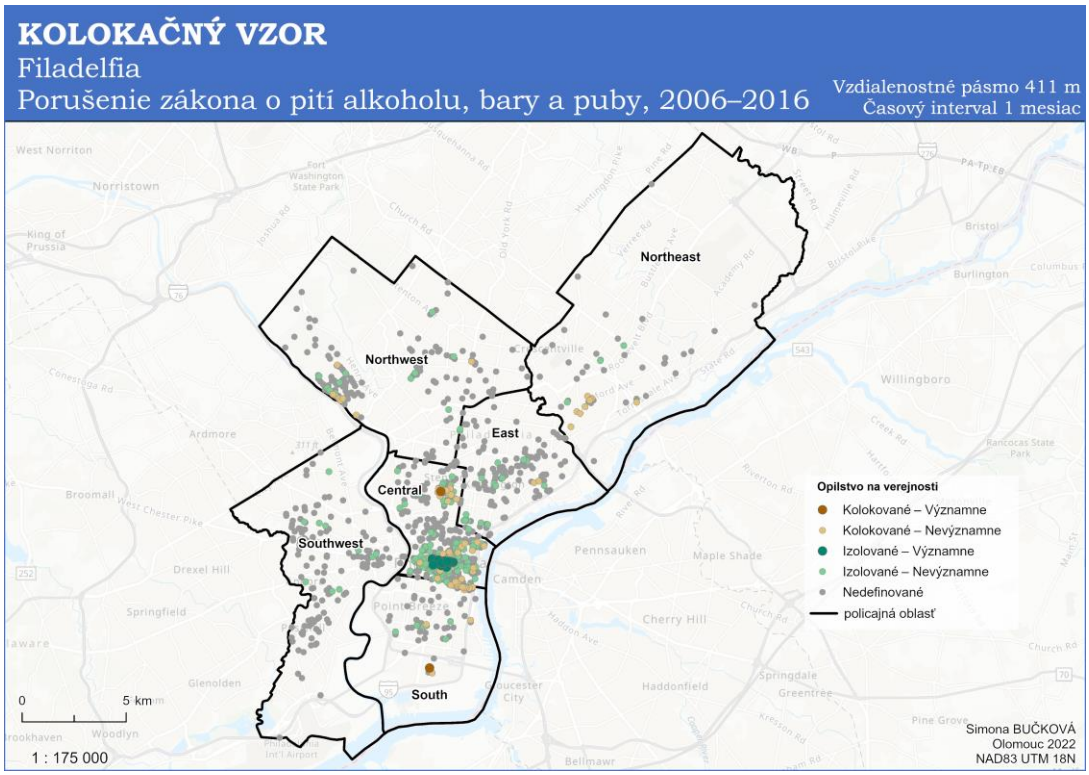
V zhodnotení celkového zaradenia zločinov je stále veľké množstvo nedefinovaných a izolovaných viď Obr. 13. Nástroj body zaradí medzi nedefinované a izolované, ak nemajú vo svojom okolí body susednej kategórie, alebo sú príliš vzdialené. V tomto prípade hrá veľkú rolu aj čas udalostí, body sú zaradené aj na základe splnenia podmienky časového intervalu s veľkosťou jedného mesiaca. Preto je možné pozorovať aj oblasti, kde sú obe kategórie blízko seba, avšak ich časové horizonty sú príliš rozdielne. Nie je teda možné povedať, že tieto kategórie spolu úzko súvisia. Určite medzi nimi vzťah existuje, avšak nie je významný pri časovom okne jedného mesiaca. Aj pri pomerne veľkom vzdialenostnom pásme je viac než 60 % záznamov nedefinovaných a ďalších 25 % izolovaných. So zväčšujúcim časovým oknom síce pribúda kolokovaných záznamov, avšak nie výrazne, drvivá väčšina bodov nie je označených za kolokovaných.



Obr. 13 Percentuálne zaradenie záznamov opilstva na verejnosti z rokov 2006–2016 (zdroj: autorka)

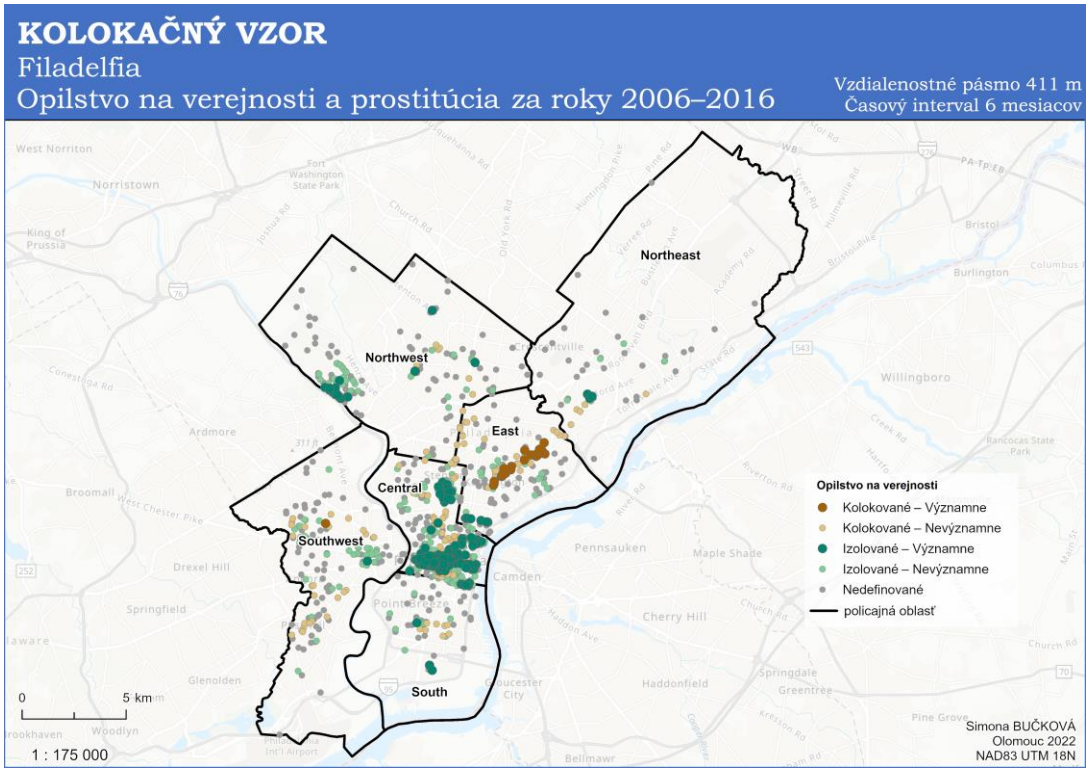
Vysvetlivky: O – opilstvo, PZ – Porušenie zákona, P - prostitúcia

V prvej dvojici kriminálnych činov boli pri nastavení jedného mesiaca nájdené len dva významne kolokované záznamy o opilstve na verejnosti. Prvý čin nastal 21. 11. 2010, 3:53 ráno (viď Mapa 8), zaujímavé je, že v jeho bezprostrednom okolí nie sú záznamy susednej kategórie v časovom horizonte jedného mesiaca. Druhý kolokovaný bod sa stal 20. 5. 2011, 20:19 a taktiež nesplňuje podmienku časového okna. Avšak nad ním, s rovnakými súradnicami sa nachádza záznam kategórie záujmu – opilstvo na verejnosti. To by ale nemalo mať na výsledný vzor vplyv, keďže sa vyhodnocuje vzťah medzi dvojicou kriminálnych činov. Pri troch mesiacoch bolo lokalizovaných desať významne kolokovaných bodov, dva pochádzajú už z predošlého nastavenia. S predĺžením okna na šesť mesiac počet nájdených záznamov stúpol na 27. Opäť základ tvorili už označené body a body, ktoré boli v predposlednom nastavení nevýznamne kolokované.



Mapa 8 Kolokačný vzor pre opilstvo na verejnosti a porušenie zákona o pití alkoholu na verejnosti za roky 2006–2016

Druhá dvojica zločinov získala o niečo významnejšie výsledky, no pri použití jedného mesiaca neboli označené žiadne záznamy za signifikantné. Jedenásť bodov bolo lokalizovaných s časovým krokom troch mesiacov, všetky sa nachádzajú na ulici Kensington Ave, opticky tvoria líniu typickú pre záznamy reprezentujúce prostitúciu (viď Mapa 9). Po zvýšení intervalu na šesť mesiacov sa počet zvýšil až na 32 významne kolokovaných záznamov, opäť ako pri prvej dvojici obsahujú body z predchádzajúceho nastavenia. Až na jeden bod sa všetky nachádzajú na spomínanej ulici, osamotený záznam sa nachádza v západnej časti mesta, avšak nad ním je „duplicitný“ bod s rovnakým dátumom a časom udalosti. I keď len 2,1 % zo všetkých záznamov je onačených za významne kolovaných, je zrejmé, že ulica Kensington Ave má isté spojenie s opilstvom a prostitúciou. Podľa mnohých amerických denníkov je táto ulica preslávená drogami, prostitúciou a vraždami, takže vyhodnotená sila vzťahu nebude náhodná.



Mapa 9 Kolokačný vzor pre opilstvo na verejnosti a prostitúciu za roky 2006–2016

Zdá sa, že významne kolokované záznamy nie sú tak časté, ako by užívateľ mohol očakávať, preto je nutné hypotézu zamietnuť. Samozrejme, výsledný kolokačný vzor je závislý na vstupných dátach a parametrizácii nástroja. V prípade skonštruovaných variant je veľká časť záznamov izolovaných a nedefinovaných, ktoré nesplnili podmienku priestoru či času. Každopádne nie len významne kolokované záznamy sú podstatné, častokrát je ich hodnota lokálneho kolokačného kvocientu vysoká a blízka hodnotám tých významných. Taktiež aj nevýznamne kolokované záznamy poskytujú istú informáciu, rovnako ako tie izolované. Užívateľ sa na kolokačný vzor môže dívať z celkového hľadiska a distribúcie kolokačných kategórií, alebo skúmať detaily a jednotlivé záznamy v priestore, poprípade v čase.

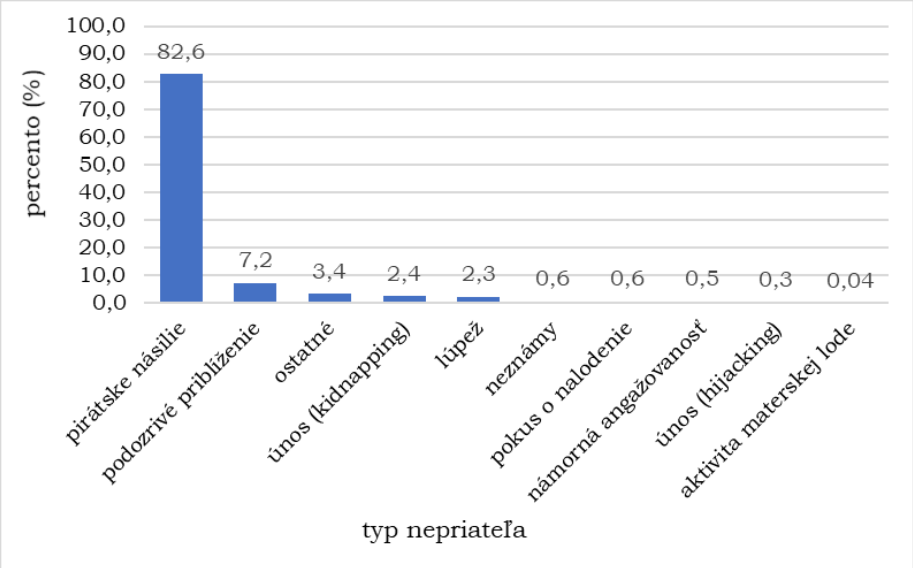
6 PRÍPADOVÁ ŠTÚDIA 3 – PIRÁTSTVO

Základným dátovým zdrojom tejto prípadovej štúdie je geodatabáza Anti-shipping Activity Messages *(ASAM)*, čiže záznamy reprezentujúce správy o aktivitách smerované proti námorníctvu. Tieto dáta sú publikované týždenne, poskytovateľom je Office of Naval Intelligence *(ONI)*. Súčasťou sú záznamy o pirátskych útokoch, nepriateľských útokoch proti komerčnej celosvetovej doprave. V reporte sú zahrnuté aj nedávne vylepšenia slúžiace ako prevencia proti pirátstvu a agresorom. ONI získava dáta kombináciou rozsiahlej námorníckej expertízy so špičkovou operačnou inteligenciou a analytickými vymoženosťami (Office of Naval Intelligence, 2022).

Na stránke Maritime Safety Information (2022) je možné voľne stiahnuť dáta ASAM ako shapefile, alebo ako geodazabázu. Okrem toho je dostupný aj formulár, kde je možné dáta bližšie špecifikovať a stiahnuť napríklad vo formáte CSV. Pre potreby štúdie bola využitá geodatabáza, ktorá obsahuje informácie o geometrii, dátume udalosti, subregióne, detail o nepriateľovi a obeti, popis udalosti, typ nepriateľa a obete a navigačnú oblasť. Čo sa týka časovej zložky, dáta sú dostupné od 1. 5. 1978 až do 8. 3. 2022, kedy bol pridaný posledný incident v čase tvorby štúdie.

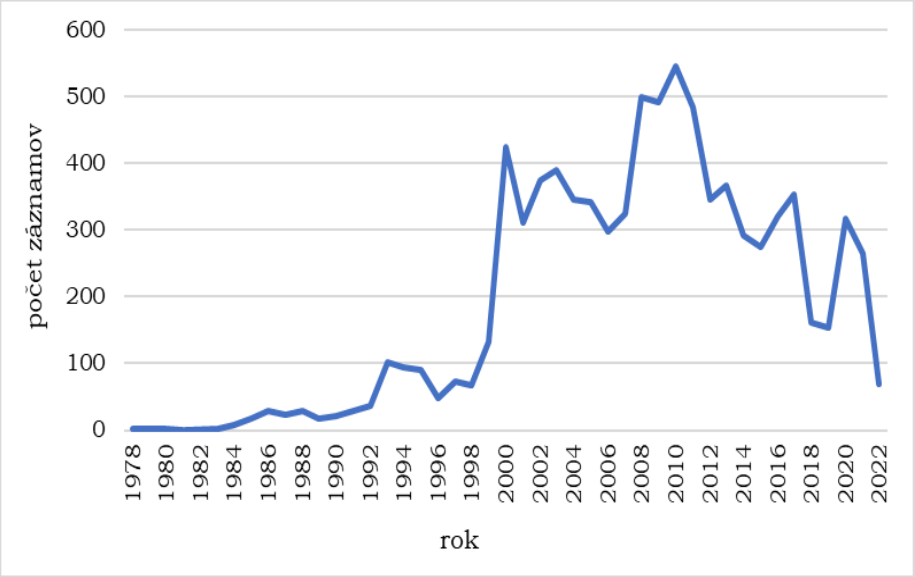
Podstatou tejto prípadovej štúdie je preskúmať posledné možné nastavenie typu susedstva, a to je externý súbor matice váh vo formáte .swm *(spatial weights matrix)*. Ten je možné vytvoriť pomocou nástroja generovať priestorovú maticu váh *(Generate Spatial Weights Matrix)*. Vstupom musia byť všetky body kolokačnej analýzy, čiže ako body záujmu, tak aj susedné body. Spôsobov ako definovať priestorový vzťah je viacero, bližšie informácie sú dostupné v rešeršnej časti práce.

Každopádne ešte pred samotnou analýzou je však vhodné pozrieť sa na celkovú distribúciu dát, ako z priestorového hľadiska, tak aj z časového. Záznamy týchto správ sa v podstate rozprestierajú po celej Zemi. Vo väčšine sa nachádzajú v oblastiach blízko pobrežia, majoritné oblasti sú Ázia a Afrika, čo samozrejme nie je tajomstvom, že v týchto miestach dochádza k častým nelegálnym záležitostiam. O podstatne menej záznamov je možné pozorovať v pobrežnej časti Ameriky a Európy. No a najmenej aktivít je v austrálskych vodách. Celkový počet aktivít pre jednotlivé typy útokov je možné vidieť na Obr. 14. Viac než 80 % záznamov patrí medzi pirátske útoky, za ním nasleduje kategória podozrivého priblíženia no jej četnosť je naozaj nízka. Ďalšie kategórie sú viacmennej zanedbateľné.

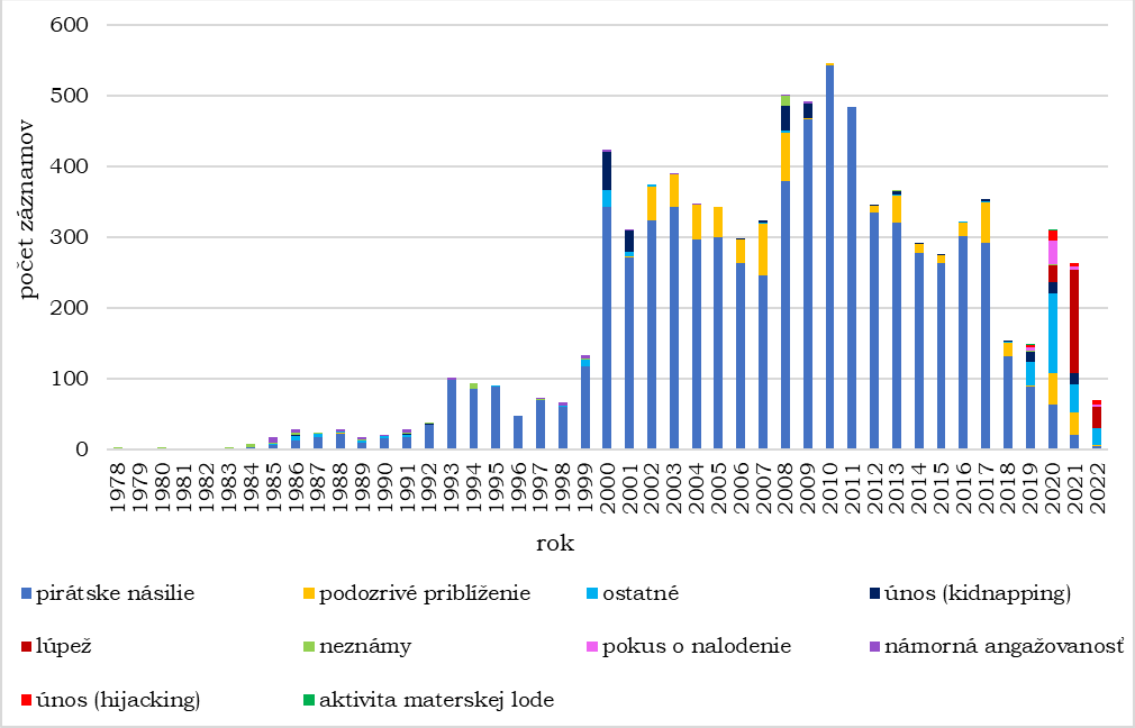


Obr. 14 Percentuálne zastúpenie typov nepriateľov za roky 1978-2022 (zdroj: autorka)

Čo sa týka časového hľadiska záznamov, priebeh ilustruje Obr. 15, kde je možné vidieť, že po roku 1999 dochádza k výraznému nárastu nepriateľských aktivít. Vrchol dosahuje v roku 2010, po ktorom dochádza k poklesu. Neskôr majú aktivity nejakú stúpajúcu tendenciu, ale nedosahujú už hodnoty, ako v roku 2010. Lepší priebeh je zhrnutý na Obr. 16, kde sú zahrnuté aj četnosti jednotlivých útokov. Ako je možné vidieť, vo väčšine dominuje pirátske násilie, no to sa za posledné roky vytráca. Naopak sa začína objavovať lúpežný typ útoku. Niektoré typy útokov sú zase veľmi málo badateľné, napríklad námorná angažovanosť či aktivita materskej lode.

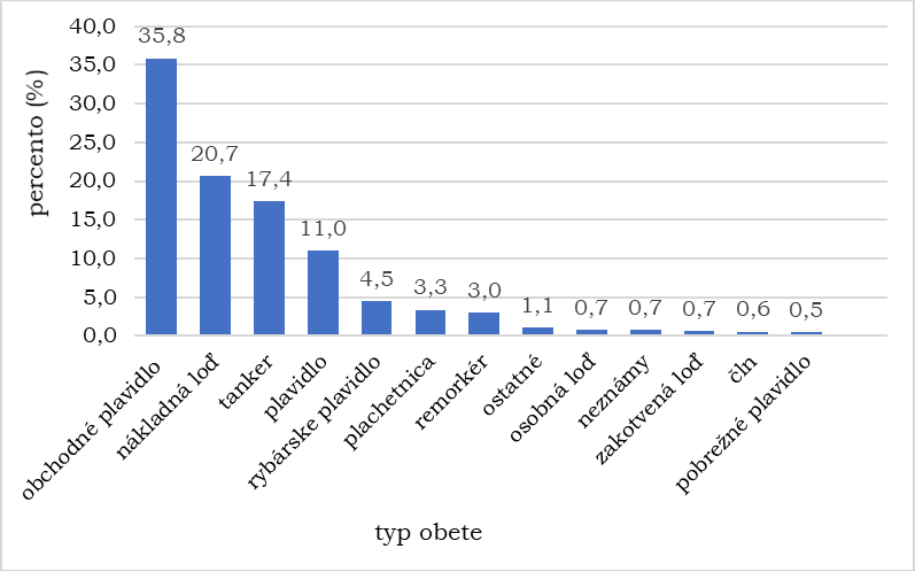


Obr. 15 Časový priebeh záznamov ASAM za roky 1978-2022 (zdroj: autorka)

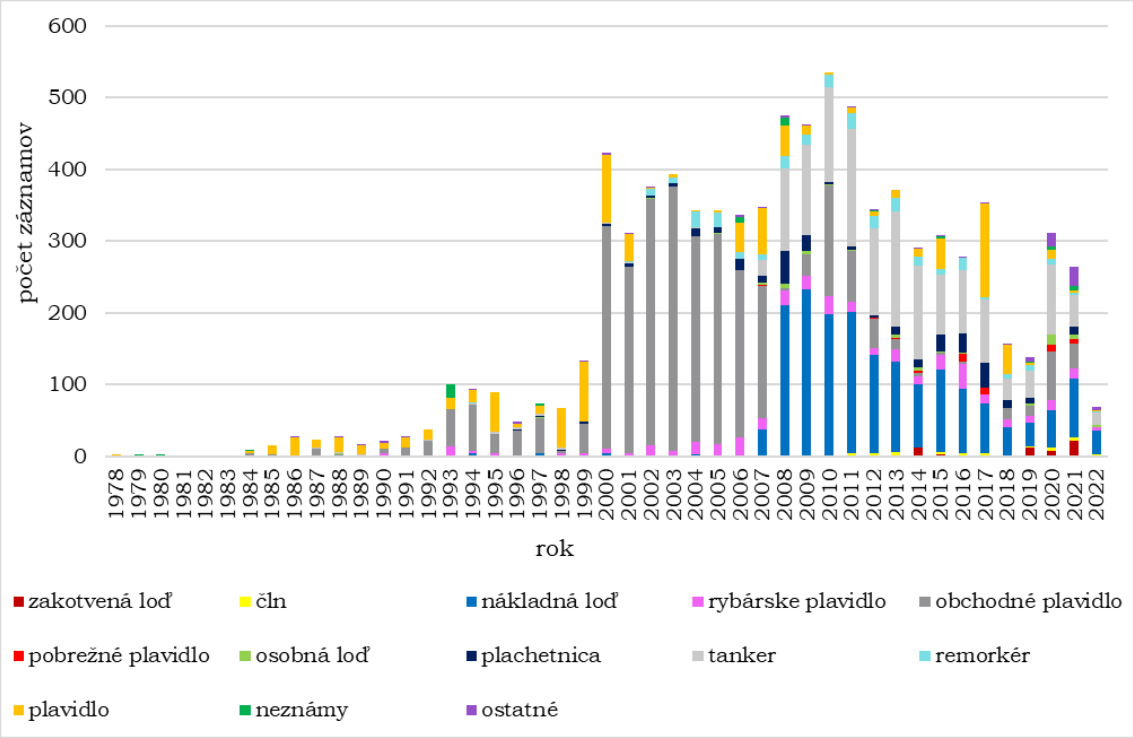


Obr. 16 Vývoj počtu záznamov pre jednotlivé kategórie nepriateľov za roky 1978-2022 (zdroj: autorka)

Z pohľadu typov plavidiel, respektíve z listu obetí (Victim List) je najčastejšie napadané obchodné plavidlo, čo je logické, pretože sa na ňom nachádza tovar či peniaze, ktoré sú pre útočníkov lákavé. Na druhom mieste je nákladná loď, opäť lákadlo a krátko za ním je tanker. Následne počty záznamov pre ďalšie kategórie rapídne klesajú (viď Obr. 17). Pri pohľade na vývoj v čase pomerne dlhý čas boli napadané hlavne obchodné plavidlá, každopádne rokom 2008 sa výrazne zmenila situácia (viď Obr. 18). Začali pribúdať záznamy s tankermi a nákladnou loďou. Dôvodom môže byť menej obchodných plavidiel na mori, ale aj zmena ekonomickej situácie či podobne.



Obr. 17 Percentuálne zastúpenie typov obetí za roky 1978-2022 (zdroj: autorka)



Obr. 18 Vývoj počtu záznamov pre jednotlivé kategórie obetí za roky 1978-2022 (zdroj: autorka)

Dáta sú pomerne objemne, preto bola zvolená len určitá časť, a to okolie Arabského polostrova. Ako už bolo zmienené vyššie, dostupné sú subregióny a navigačné oblasti, vďaka ktorým je možné vybrať určitú oblasť. Arabský polostrov spadá navigačnej oblasti IX. Pomocou nástroja výberu podľa atribútov *(Select by Attributes)* bola vybraná želaná oblasť. Z 8 572 záznamov ostalo len 1 254 s časovými záznamami od roku 1984 do 2022. Dostupných je osem typov nepriateľských útokov viď Tab. 4:

Tab. 4 Dostupné typy útokov a ich počet

typ útoku	**počet záznamov**
pirátske násilie	878
podozrivé priblíženie	248
únos (kidnapping)	70
ostatné	22
námorný angažmán	16
neznámy	11
únos (hijacking)	4
pokus o nalodenie	1

Nástroj kolokačná analýza požaduje vybrať dve kategórie, prvky záujmu a susedné prvky. Preto je dôležité vybrať, ktoré typy nepriateľských útokov budú analyzované. Problémom však je, že viac než polovicu bodov tvoria záznamy s pirátskym násilím. V prípade, že v kategórii susedných prvkov je výrazne menší počet záznamov, nápoveda varuje, že nástroj kolokáciu a lokálny kolokačný kvocient vyhodnotí ako nízky. Preto je vhodné, ak sú počtom bodov vstupné kategórie vyrovnané. Z tohto dôvodu bol preskúmaný počet záznamov pre jednotlivé obete útokov ako možnosť vybrania vyrovnanejšieho počtu bodov. Kategórie a ich zastúpenie je zhrnuté v Tab. 5.

Tab. 5 Dostupné typy obetí a ich počet

typ obete	**počet záznamov**
nákladná loď	354
obchodné plavidlo	343
tanker	287
plavidlo	170
predávajúce plavidlo	37
rybárske plavidlo	20
ostatné	13
remorkér	12
neznámy	7
osobná loď	6
pobrežné plavidlo	0

6.1 Kolokácia nákladnej lode a obchodného plavidla

V tabuľke je možné pozorovať dve kategórie, ktoré sú veľmi vyrovnané, a to nákladná loď a obchodné plavidlo. To znamená, že ak medzi útokmi na lode, respektíve plavidlá existuje vzťah, ich kolokačný kvocient nebude zbytočne nízky z dôvodu nevyrovnanosti počtu záznamov. Taktiež nie je vylúčené, že útoky medzi sebou nemajú priestorový či časový súvis, to práve overí kolokačná analýza. Záznamy s nákladnou loďou sú rozprestreté okolo viacerých kontinentov, avšak drží sa trend s najväčšou koncentráciou v afrických a ázijských vodách. Záznamy obchodného plavidla majú veľmi podobnú priestorovú distribúciu, vo väčšine sa nachádzajú v blízkosti nákladných lodí a majú najväčší podiel zo všetkých plavidiel, respektíve typov obetí.

Ako už bolo spomenuté vyššie, je viacero metód určovania vzťahu v rámci tvorby externého .swm súboru. V tejto variante boli využité takmer všetky, ktoré boli dostupné, metóda externej tabuľky *(Convert table)* nemala uplatnenie. V nie všetkých zaostávajúcich metódach bolo možné nastaviť medznú vzdialenosť *(Distance treshold)*. Avšak, kde to bolo možné, bola použitá východzia vzdialenosť, ktorú nástroj generovania priestorovej matice váh vypočíta na základe vstupných dát a činila 209 811,193 metrov. Jednotka v akej je udaná medzná vzdialenosť záleží od súradnicového systému. Okrem východzej vzdialenosť boli testované aj hodnoty desať a 50 kilometrov. Ďalším nastavením je počet susedných prvkov, opäť nie je dostupný pri všetkých metódach. S meniacou sa vzdialenosťou bol testovaný počet susedov desať a 50. Dôvodom týchto hodnôt je skúsenosť z predchádzajúcich štúdií a taktiež z priemernej hodnoty, ktorú nástroj po spustení poskytuje. Záznamov je pomerne veľké množstvo, aj kvôli veľkému časovému rozpätiu, priemerný počet susedných prvkov je 95,64. Netreba však zabúdať na to, že je to len priemerná hodnota a neberie do úvahy ďalšie skutočnosti ako je čas či vzdialenosť susedstva.

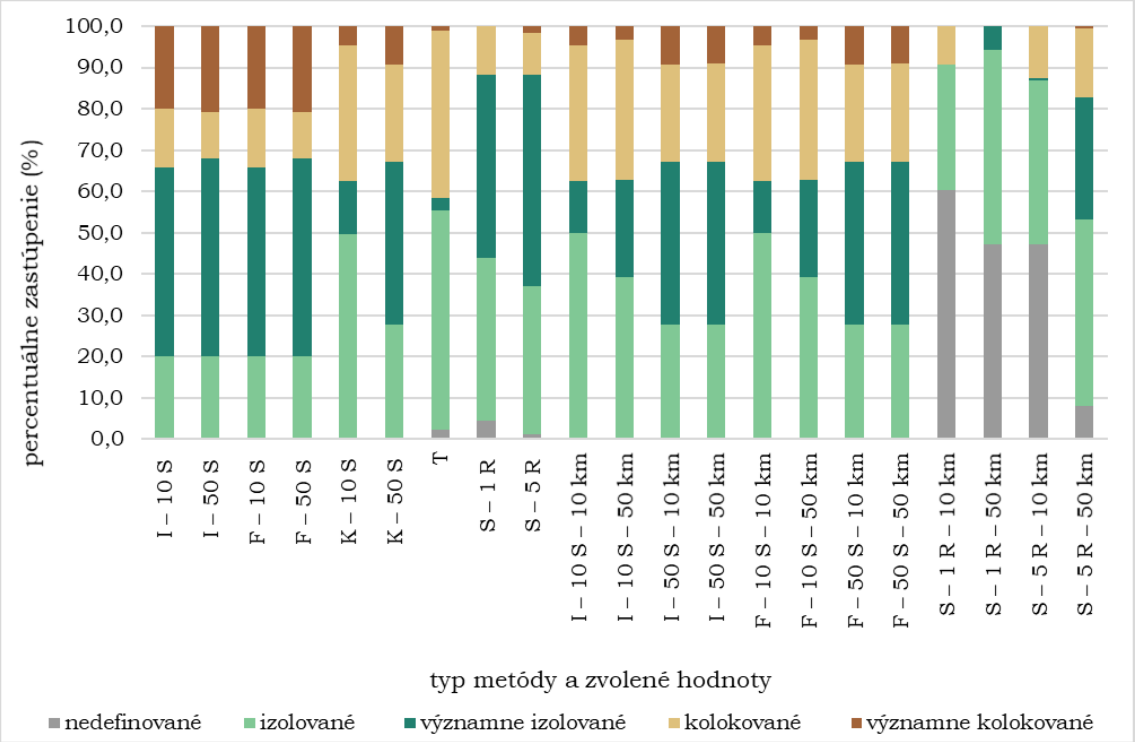
Predposledným parametrom je vzdialenostná metóda. Nástroj ponúka dve, a to euklidovskú a manhattanskú vzdialenosť. Euklidovská vytvára pomyselné rovné čiary medzi bodmi, kdežto manhattanská využíva pravé uhly a je pomenovaná na základe tvaru ulíc v meste Manhattan. Vzhľadom na to, že útoky sa dejú na mori, ktoré je viacmennej homogénne, bez prekážok, bola použitá len euklidovská vzdialenosť. Posledným možným nastavením je možnosť štandardizácie priestorových váh. Možnosť je doporučená, a preto bola zaškrtnutá.

Zo spomínaného nástroja sú získané len binárne súbory, ďalších krokom je ich vloženie do nástroja kolokačná analýza. Po vybraní metódy susedstva ako externý súbor s priestorovými váhami sa možnosti nastavenia výrazne znížia. V podstate je dostupný len typ jadra (gausovské, bikvadratické a žiadne) a časové okno, ktoré po nastavení hodnoty zmení typ susedstva na vzdialenostné pásmo. To znamená, že celá parametrizácia je závislá na externom súbore, preto bol aj nástroj na jeho generovanie detailnejšie preskúmaný. Avšak čo je možné porovnať je výsledok z tejto poslednej metódy a výsledok pri použití čisto vzdialenostného pásma, počtu susedných prvkov a časového okna.

6.1.1 Výsledky

Pri pohľade na percentuálne zastúpenie kategórií v jednotlivých nastaveniach (viď Obr. 19) je možné vidieť isté rozdiely a tendencie metód. Inverzná a fixná vzdialenosť má vo východzej vzdialenosti totožné výsledky pri použití rovnakej hodnoty počtu susedných prvkov. S najväčšou pravdepodobnosťou to bude spôsobené neskutočne veľkou medznou vzdialenosťou. Pri jej neskoršom nastavení s rovnakým počtom

susedných prvkov sa znížil ako počet významne kolokovaných, tak aj významne izolovaných bodov. Na druhej strane určite má väčší zmysel vyšetrovať signifikantnosť na vzdialenosť 50 kilometrov než viac než 200. Práve pri takejto veľkej hodnote si užívateľ koleduje o falošne pozitívne výsledky.

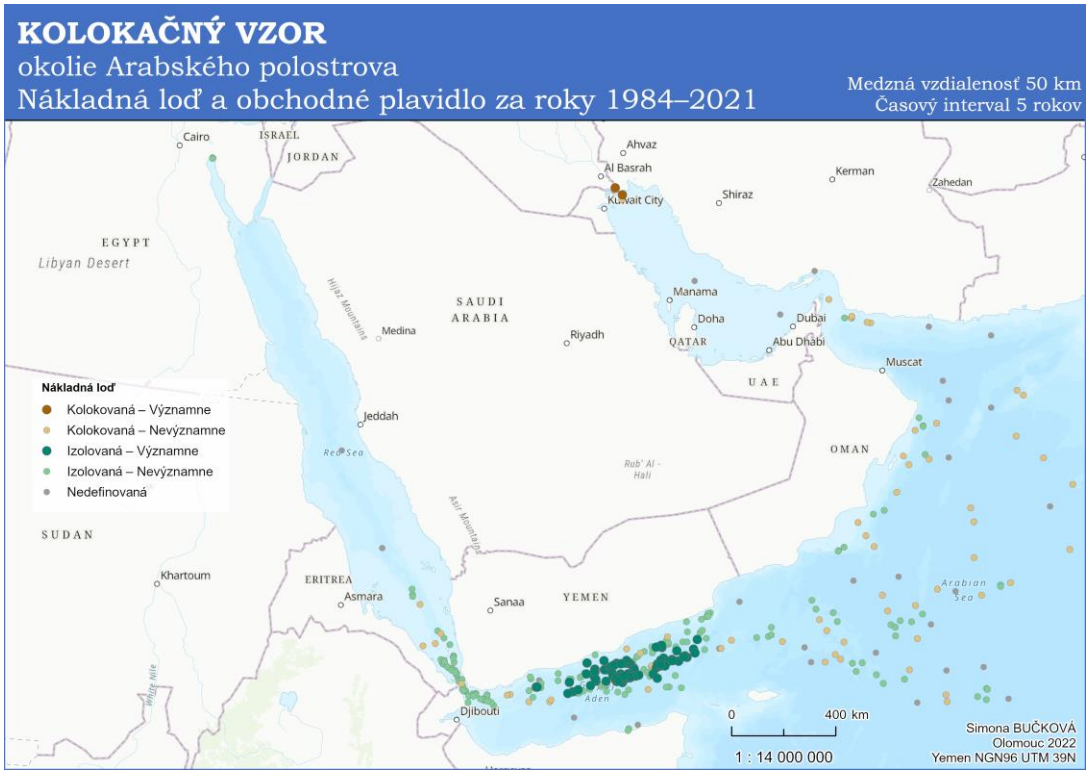


Obr. 19 Prehľad o percentuálnom zastúpení kategórií v jednotlivých metódach (zdroj: autorka)

Trochu odlišné výsledky boli pozorované pri metóde K najbližších susedov. Pre túto analýzu boli zvolené dve výrazne odlišné hodnoty, a to úzky interval desiatich najbližších prvkov a široký interval 50 prvkov. Percentuálne zaradenie je porovnateľné s nastavením v prípade inverznej a fixnej vzdialenosti vrátane medznej vzdialenosti. Výrazne rozdielne výsledky sú pozorovateľné pri metóde delaunay triangulácií, kedy je susedstvo hodnotené na základe dotyku hrán a uzlov skonštruovaných trojuholníkov. Nápadne sa znížil počet význame izolovaných a významne kolokovaných záznamov, no na druhej strane je viac než 40 % záznamov označených ako kolokovaných. Netreba však opomenúť, že doterajšie metódy nepracovali s časom. Jediná metóda, ktorá okrem priestoru používa vo výpočte aj čas je časopriestorové okno. Opäť boli použité dve hodnoty, a to jeden a päť rokov. Výsledky sa medzi sebou až tak nelíšia, avšak prvýkrát sa objavili nedefinované body, ktoré nespĺňajú podmienku časového rozpätia a východzej medznej vzdialenosti. Práve čas pridáva kolokačnému vzťahu ďalší rozmer, a tým vzor viac odpovedá skutočnosti.

V ďalšom nastavení bola upravená medzná hodnota na reálnejšie čísla, ktoré sú viac relevantné pre vyšetrenie vzťahu medzi typmi plavidiel. Vzhľadom na výrazne menšiu vzdialenosť sa znížil počet významne kolokovaných nákladných lodí, ale vzrástol počet nevýznamne kolokovaných. Zaujímavý je fakt, že pri metóde inverznej vzdialenosti nemá vzdialenosť až takú veľkú rolu ako počet susedných prvkov. Opäť nápadne odlišné výsledky je možné pozorovať pri použití časopriestorového okna. Takmer polovica záznamov nemá vzťah s obchodnými plavidlami v rozmedzí do piatich rokov a v okolí do

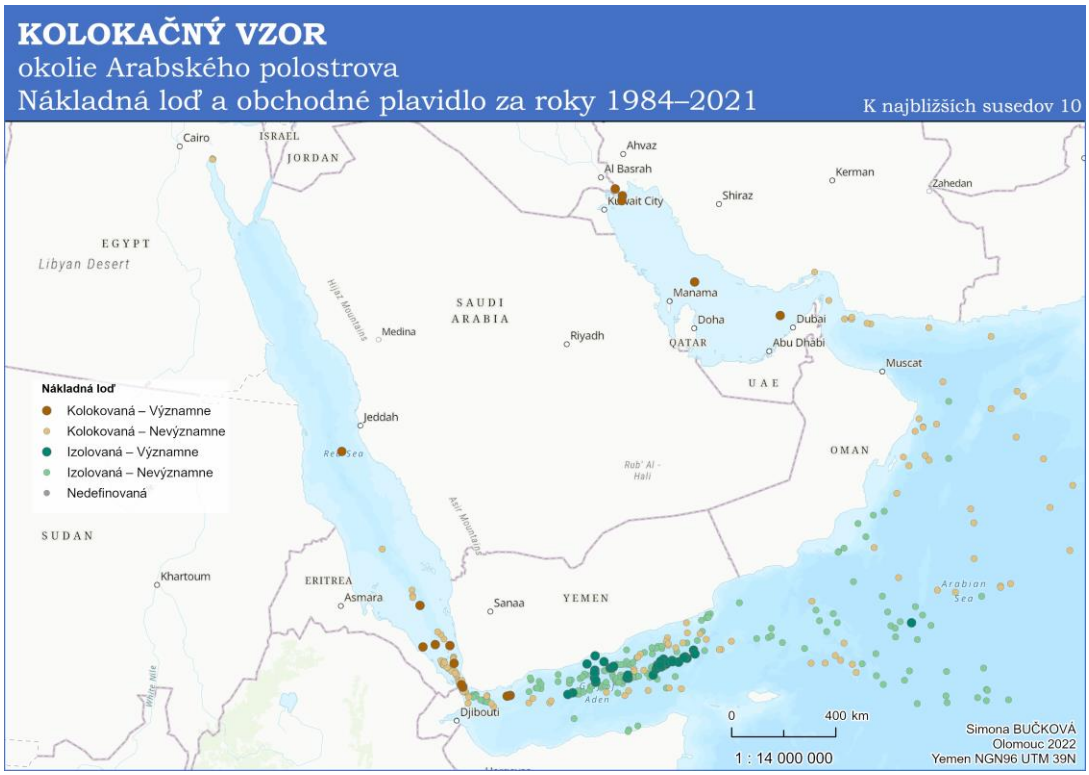
10 kilometrov. Až pri nastavení medznej vzdialenosti na 50 kilometrov a časového intervalu piatich rokov sa znížil počet nedefinovaných bodov na minimum (viď Mapa 10). Na druhej strane len o trochu viac než je štvrtina nákladných lodí bolo označených za kolokovaných. Preto je možné konštatovať, že vzhľadom na čas nie je medzi skúmanými kategóriami plavidiel kolokačný vzor až tak významný.



Mapa 10 Kolokačný vzor pre nákladnú loď a obchodné plavidlo za roky 1984–2021

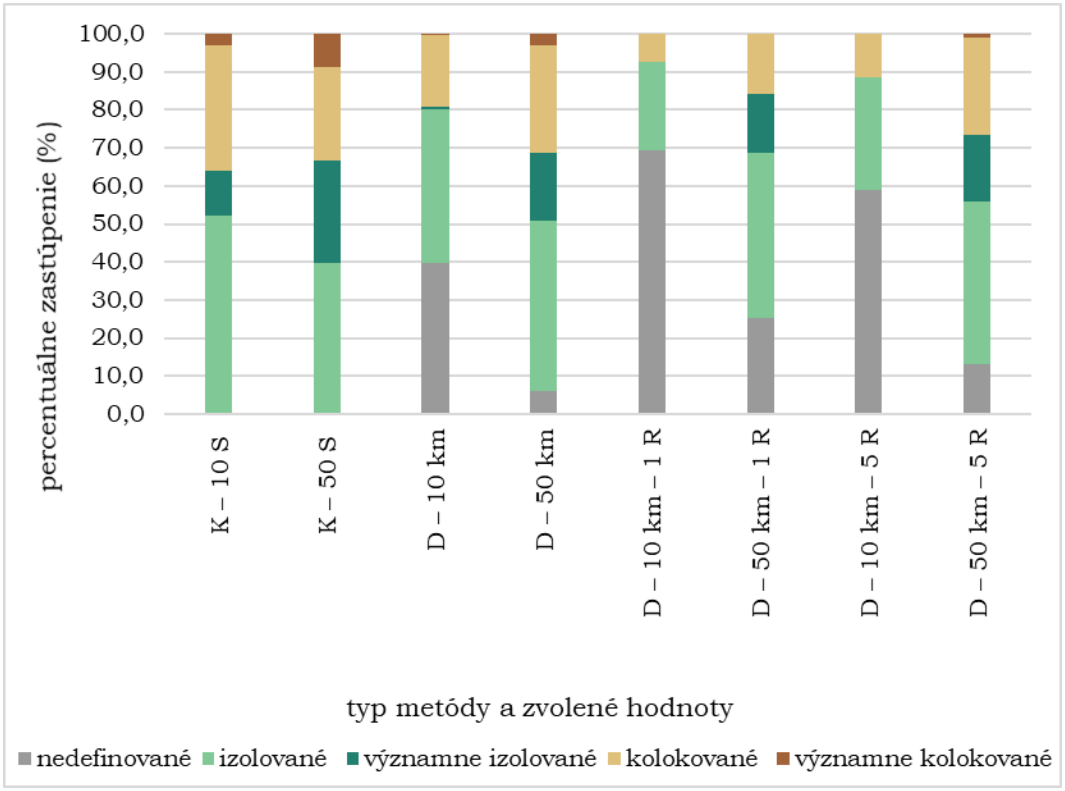
Čo sa týka priestorovej distribúcie jednotlivých kolokačných kategórií, existuje výrazný vzor s tendenčným rozdelením výrazne kolokovaných a výrazne izolovaných záznamov. Kolokované body sa vyskytujú v drvivej väčšine v zálivoch, a to v Perzskom a Ománskom. Druhou veľkou oblasťou je Červené more, respektíve prieliv Báb el-Mandeb, ktorý oddeľuje Afriku od Arabského polostrova. Dôvodom môže byť jednoduchšie napadnutie lode, kvôli menšiemu otvorenému priestoru medzi pevninami. Taktiež nie je záhadou, že množstvo pirátskych útokov pochádza z pobrežných krajín Afriky ako je napríklad Somálsko. Na druhej strane, významne izolované body, kde nákladná loď a obchodné plavidlo nie sú asociované sa výrazne zhlukuje do Adenského zálivu. No a nevýznamné záznamy, či už kolokované, alebo izolované sú rozprestreté v oblasti Arabského mora. Pri pohľade na fixnú a inverznú metódu bolo zistené, že nielenže mala rovnaké percentuálne zastúpenie kategórií, ale taktiež aj totožnú priestorovú distribúciu. To znamená, že pri týchto vstupných dátach nemajú tieto metódy priestorového vzťahu opodstatnenie v zmysle odlišného výpočtu danej metódy. Prvé odlišnosti prichádzajú s ďalšou metódou v poradí, a to K najbližších susedov, kedy nevýznamne kolokované body sa objavujú práve v oblasti Adenského zálivu, kde doteraz boli len body izolované (viď Mapa 11). Zaujímavé však je, že so zvýšeným počtom susedných prvkov klesol počet kolokovaných lodí, čo doposiaľ pozorované nebolo. Prvé nedefinované záznamy sa objavujú s metódou delaunay triangulácie, ktorá síce má podobné tendencie v distribúcii, ale identifikovala aj významne kolokovaný bod

v Adenskom zálive. Tento konkrétny bod už v ďalších nastaveniach označený nebol, čo znamená, že išlo len o princíp zvolenej metódy. No a najmenej kolokovaných lodí bolo identifikovaných pri použití časového okna, kde sa používa ako vzdialenosť, tak aj čas. Nedefinované body, ktoré nespĺňali podmienku jedného a piatich rokov so vzdialenosťou desať a 50 kilometrov sa objavujú v miestach izolovaných bodov, čiže oblasť Arabského mora a Adenského zálivu. Navyše zmena nastala aj v Perzskom zálive, ktorého záznamy zmenili zaradenie z kategórie významne kolokovaných na nedefinované.

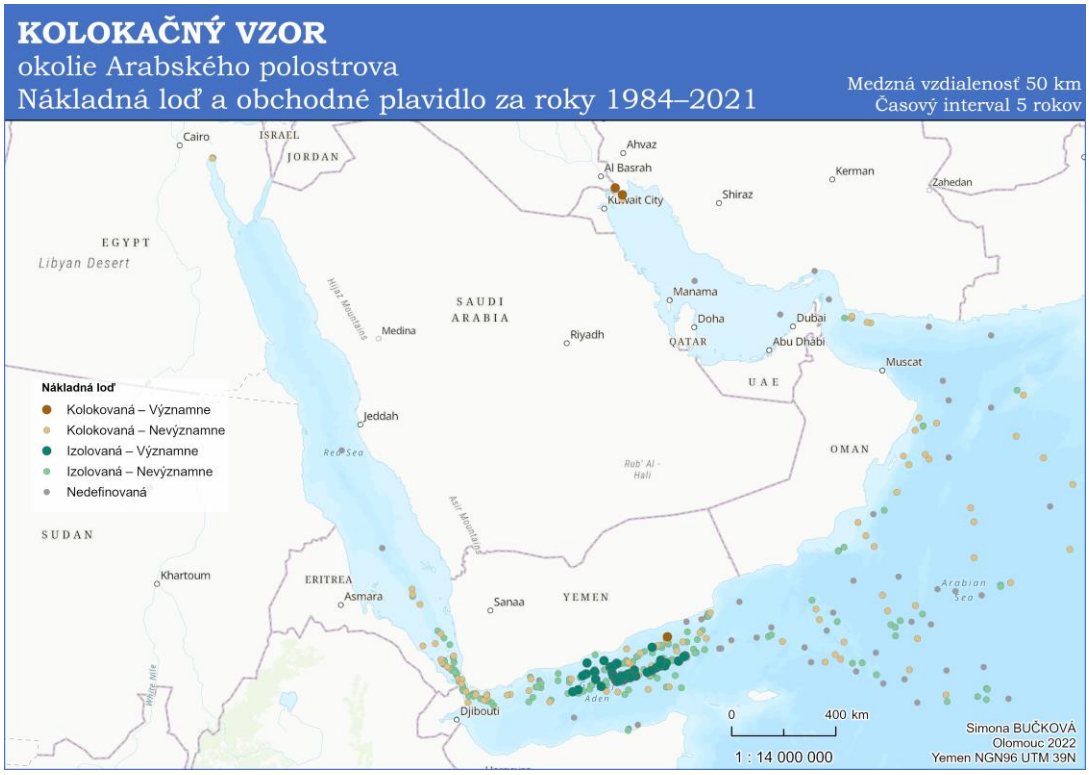


Mapa 11 Kolokačný vzor pre nákladnú loď a obchodné plavidlo za roky 1984–2021

Pre porovnanie boli skonštruované aj kolokačné vzory bez použitia externého súboru. K dispozícii boli len dve možnosti, a to K najbližších susedov a vzdialenostné pásmo. Nastavenie nástroja bolo obdobné a podobne vyšli aj výsledky pri desiatich a 50 susedných prvkoch (viď Obr. 20). Obdobnú tendenciu malo aj vzdialenostné pásmo, s menšími rozdielmi, avšak zachoval sa pomerne odlišný výsledok pri nastavení vzdialenosti na 50 kilometrov a časového intervalu päť rokov. Odlišnosť tohto nastavenia spočíva vo výraznom znížení nedefinovaných záznamov a náraste výskytu nevýznamne kolokovaných bodov práve v oblasti Adenského zálivu a Arabského mora (viď Mapa 12).



Obr. 20 Prehľad o percentuálnom zastúpení kategórií v jednotlivých metódach (zdroj: autorka)



Mapa 12 Kolokačný vzor pre nákladnú loď a obchodné plavidlo za roky 1984–2021

Výhodou využitia externého súboru matice váh spočíva v jeho možnostiach zostavenia. Nástroj generovať priestorovú maticu váh ponúka väčšie spektrum metód,

ktorých úlohou je zostaviť silu vzťahu. Ďalším plusom je fakt, že do výpočtu vstupujú všetky prvky, nie len body záujmu. Čo kolokačná analýza opomína je využitie počtu susedných prvkov pri metóde vzdialenostného pásma. Práve toto rieši možnosť inverznej či fixnej vzdialenosti, kde je okrem medznej hodnoty možné zadať aj hodnotu určujúcu veľkosť skúmaného na základe počtu okolitých bodov. V neposlednom rade je dostupná štandardizácia výsledkov, ktorá sa bežne pri úlohách data miningu používa.

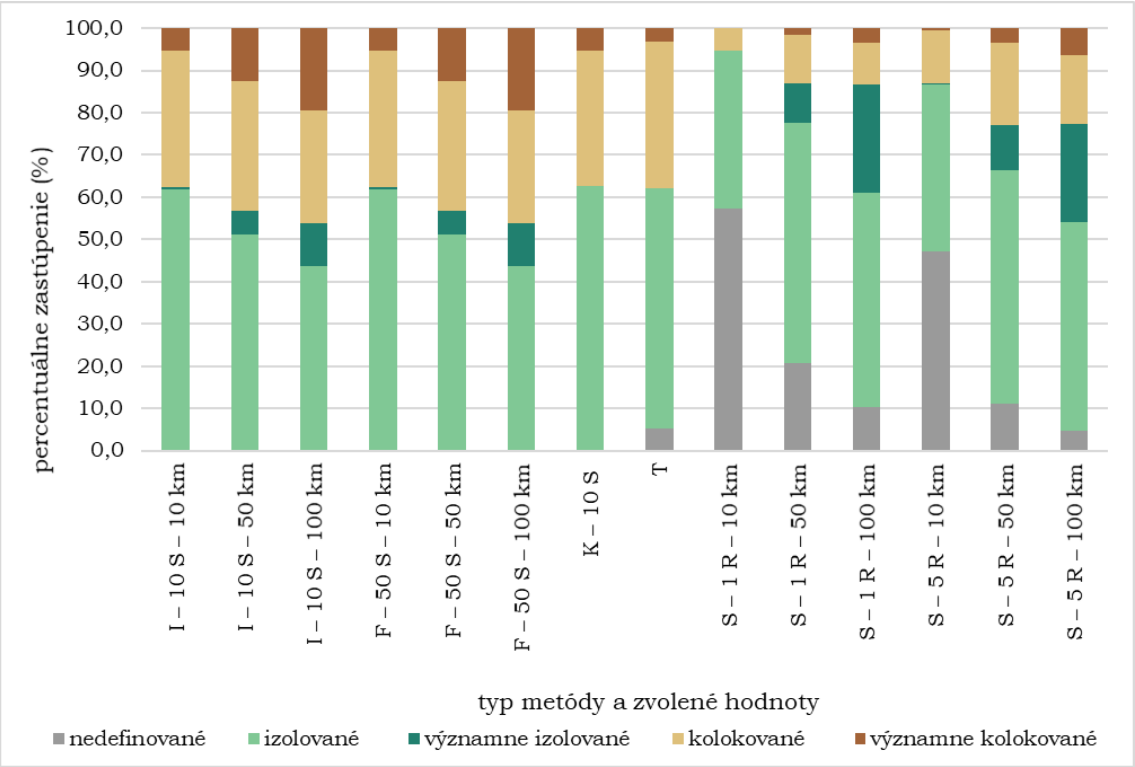
6.2 Kolokácia pirátskeho násilia a podozrivého priblíženia

Ďalšou variantou je preskúmanie sily vzťahu medzi záznamami o pirátskom násilí a podozrivom priblížení. Aktivity pirátskeho násilia tvoria majoritnú kategóriu v celom súbore dát. Najväčší výskyt je možné pozorovať v okolí Afriky a Ázie. Susedná kategória podozrivé priblíženie určite koreluje s kategóriou záujmu, pirátskeho násilia. Významné zhluky je možné pozorovať v Guinejskom zálive, v okolí Arabského polostrova a Sumatry.

Nie je vylúčené, že plavidlo, ktorého približovanie bolo nahlásené ako podozrivé nemohlo mať v pláne pirátsky útok. Každopádne nápoveda kolokačnej analýzy varuje pred nevyrovnanosťou v počte záznamov vstupných kategórií. V takom prípade by mal výsledný lokálny kolokačný kvocient nízku hodnotu. Je pravdou, že počet záznamov o pirátskom násilí výrazne prevyšuje počet záznamov podozrivého priblíženia, no pre zaujímavosť bolo zostrojených pár kolokačných vzorov. Výsledky nenesú spomínané varovanie, majú obdobné výsledky ako predchádzajúca varianta s veľmi vyrovnaným počtom záznamov. Preto boli aplikované poznatky z predchádzajúcej varianty a zostavené ďalšie vzory. Opäť bol použitý externý .swm súbor s rovnakými metódami konceptualizácie priestorových vzťahov. Zmena nastala len v niektorých hodnotách, občas bola vypustená možnosť 50 susedných prvkov pridaná medzná vzdialenosť 100 kilometrov. Následne boli vyhotovené kolokačné vzory klasicky pomocou primárneho nástroja Colocation Analysis.

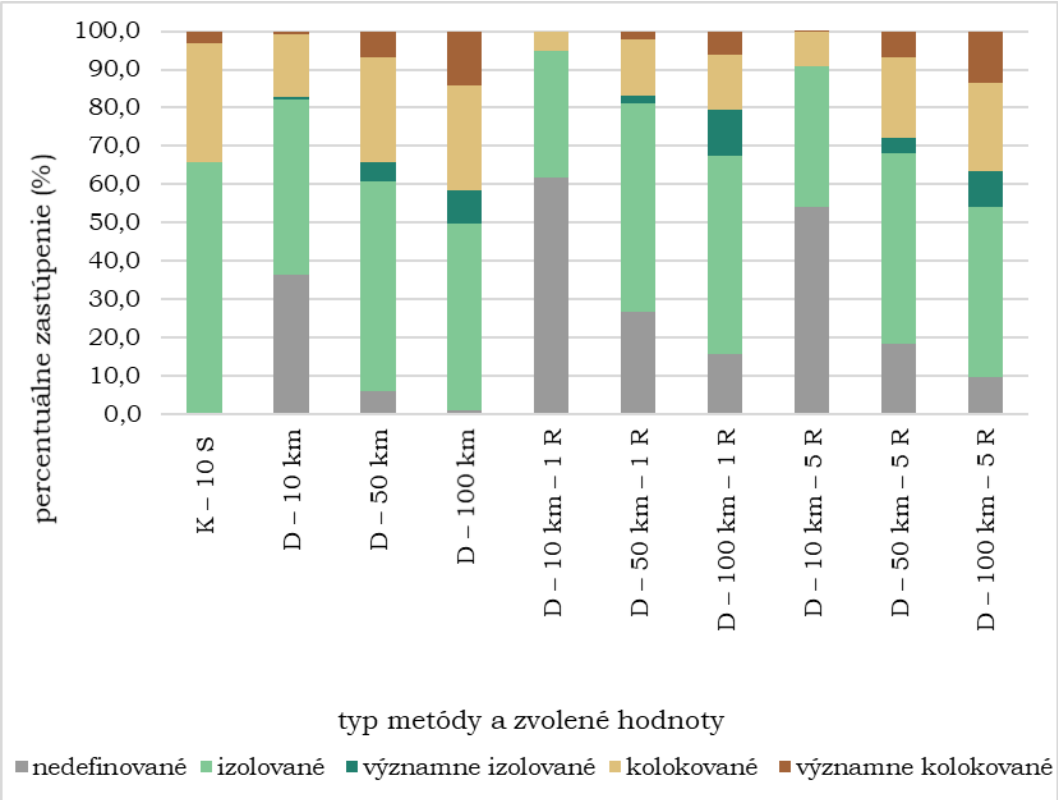
6.2.1 Výsledky

Ako v prechádzajúcej variante, ani tu neexistuje rozdiel medzi výsledkami inverznej a fixnej vzdialenosti. S pribúdajúcou vzdialenosťou sa zvyšuje počet kolokovaných záznamov, či už významných, alebo nevýznamných. No a logicky klesá počet izolovaných bodov, nedefinované záznamy nie sú prítomné (viď Obr. 21). V porovnaní s prvou variantou a totožným nastavením je viditeľné nižšie percento významne izolovaných útokov, čo by mohlo napovedať o väčšom prepojení medzi útokmi než medzi typmi plavidiel. Samozrejme je potrebné brať v úvahu odlišnosť dát a zvolených kategórií. Prvé nedefinované body sa znova objavujú až pri použití časopriestorovej metódy, avšak je možné vidieť, že nie len čas hrá rolu, ale aj medzná vzdialenosť medzi skúmanými prvkami. Dôkazom je znižujúci sa počet nedefinovaných prvkov. Na druhej strane množstvo výrazne kolokovaných bodov je pomerné nízke, pri metóde časopriestorového okna dosahuje maximálne 6,5 %.



Obr. 21 Prehľad o percentuálnom zastúpení kategórií v jednotlivých metódach (zdroj: autorka)

Vzory po použití K najbližších susedov z externého súboru a z možnosti v nástroji Colocation Analysis sú si veľmi podobné, no viac než polovica bodov je nevýznamne izolovaných. Podobné sú si aj výsledky časopriestorovej metódy a vzdialenostného pásma s časovým oknom. Vždy sa objavujú nedefinované body, pretože podmienka je pomerne prísnejšia než pri použití nastavenia bez časového aspektu. Každopádne viac kolokovaných záznamov je prítomných v klasickom vzdialenostnom pásme (viď Obr. 22). Môže to byť spôsobené tým, že nástroj primárne hodnotí vzťahy podľa vzdialenosti a časová stopa je len ako voliteľný doplnok. Celkovo je možné zhodnotiť, že čas hrá pomerne veľkú rolu vo výslednom vzore a sile vzťahu vstupných prvkov. Nepodstatnou súčasťou je aj zvolená vzdialenosť, so zvyšujúcou sa hodnotou sa zvyšuje aj pozitívny výsledok vo forme kolokovaných záznamov. Na druhej strane je tu otázka aká vzdialenosť je relevantná, to už vie sám odborník, ktorý sa napríklad týmto atakom venuje a analyzuje ich. V tejto problematike nebolo jednoduché dohľadať relevantné vzdialenosti, ktoré by boli predtým otestované, preto hodnoty parametrov vychádzali zo skúseností s predchádzajúcimi variantami a prípadovými štúdiami.

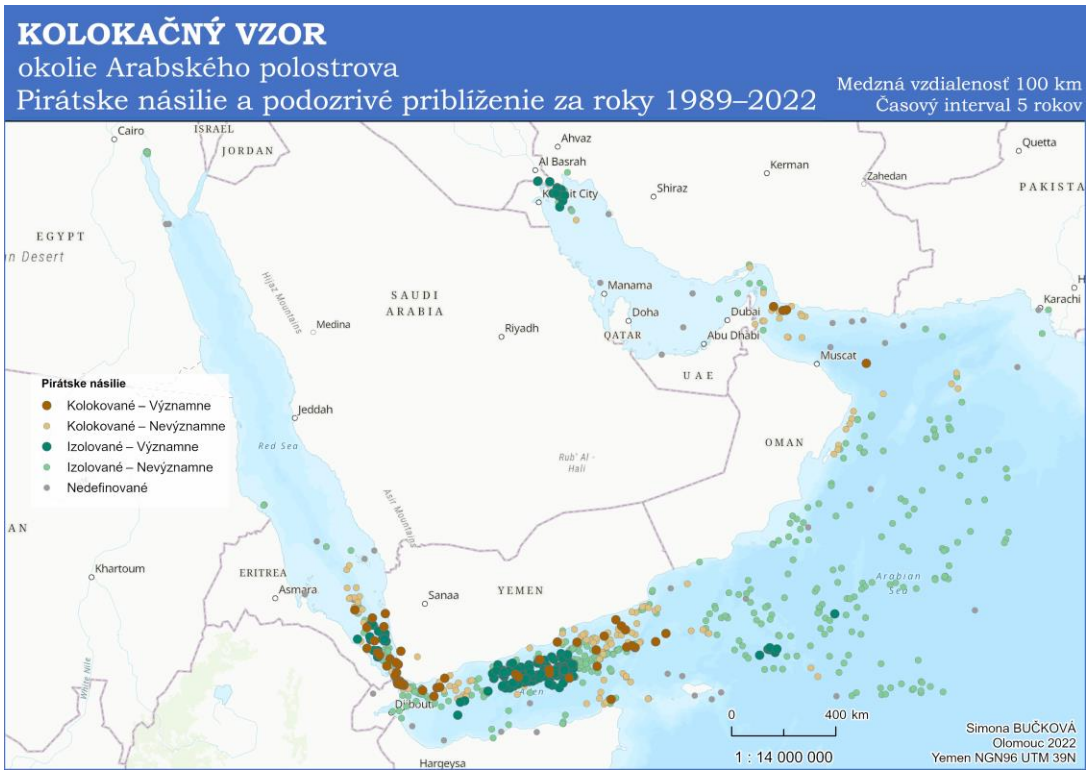


Obr. 22 Prehľad o percentuálnom zastúpení kategórií v jednotlivých metódach (zdroj: autorka)

Čo sa týka priestorovej distribúcie zvolených kategórií, vo väčšine sa nachádzajú v tesnej blízkosti, hlavne v zálivoch. Vo vodách Arabského mora však záznamy o podozrivom priblížení až na pár záznamov absentujú. Preto je možné očakávať, že body pirátskeho násilia budú izolované. Už z predchádzajúcich informácií je známe, že vzor pre inverznú a fixnú vzdialenosť je totožný. S zväčšujúcou sa vzdialenosťou pribúda množstvo kolokovaných bodov. Zhlukujú sa predovšetkým v okolí prielivu Báb el-Mandeb, Adenského a Ománskeho zálivu. Pomerne veľký zhluk významne izolovaných útokov je v oblasti Perzského zálivu a Adenského zálivu, aj keď sa v jeho okolí vyskytuje pomerne veľké množstvo bodov susednej kategórie.

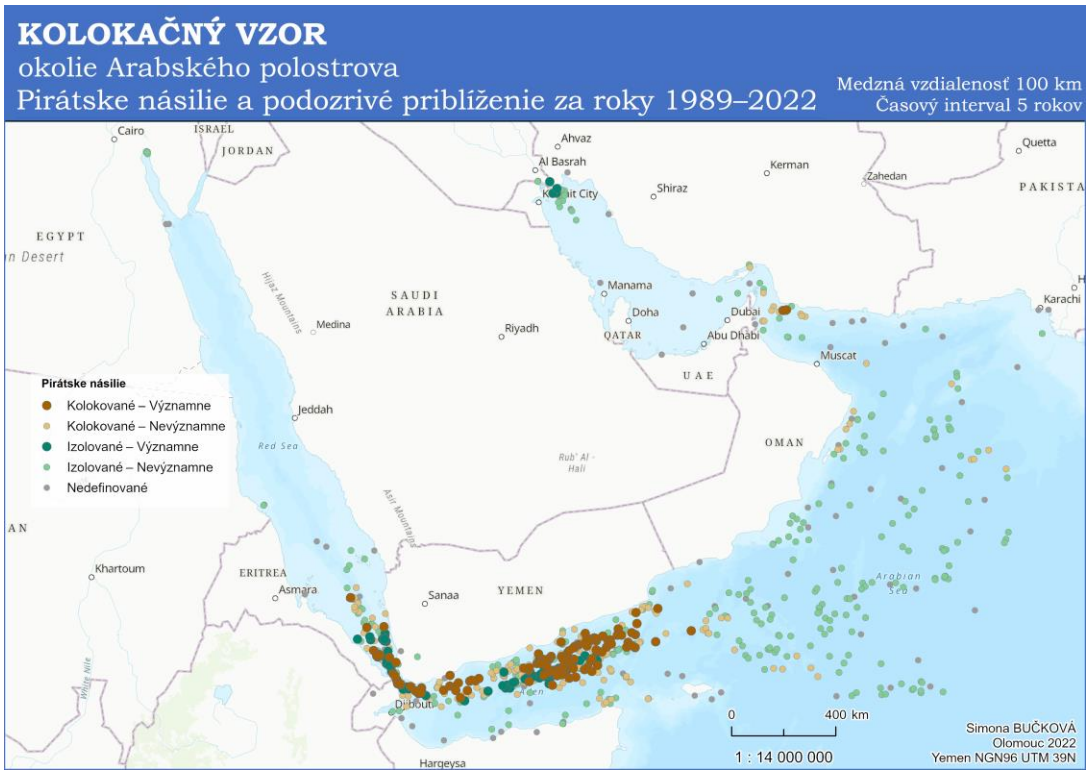
V podstate miernejší priebeh má kolokačný vzor s desiatimi najbližšími susednými prvkami. Drží sa kategoriálneho rozdelenia z predchádzajúceho nastavenia, avšak v oblastiach významne kolokovaných bodov prevládajú tie nevýznamné a podobne. Delaunay triangulácia je taktiež dosť podobná, ale v mieste medzi Ománskym zálivom a Arabským morom identifikovala záznamy ako nedefinované. V predošlých nastaveniach tieto body nevýznamne kolokované a izolované. Netreba však opomenúť ako táto metóda určuje priestorové vzťahy, čím sa výrazne líši od ostatných metód.

Z časového hľadiska je už zrejmé, že množstvo záznamov nesplňuje zadané podmienky. Nedefinované body sa vyskytujú viacmennej v celej študovanej oblasti, avšak pri zvyšovaní medznej vzdialenosti ich výrazne ubúda. Trend významných bodov je stále viditeľný najmä pri časovom okne piatich rokov a medznej vzdialenosti 100 kilometrov (viď Mapa 13). Je teda možné prehlásiť, že v spomínaných oblastiach sú zvolené kategórie v istom vzťahu, a to aj s použitím časovej zložky.



Mapa 13 Kolokačný vzor pre pirátske násilie a podozrivé priblíženie za roky 1989–2022

Čo určite ešte stojí za povšimnutie je použitie rovnakých hodnôt v primárnom nástroji, ale s voľbou vzdialenostného pásma pre analýzu skúmaných vzťahov. Najviditeľnejšiu zmenu je možné pozorovať pri najväčších hodnotách. Metóda časopriestorového okna označila väčšinu útokov v Adenskom zálive ako izolované, no vzdialenostné pásmo spomínané body vyhodnotilo ako významne kolokované (viď Mapa 14). Nápoveda nástroja neinformuje o priebehu algoritmu pri jednotlivých nastaveniach, avšak je možné, že vzdialenostné pásmo vyhodnocuje vzťahy primárne na základe priestoru, a preto sú výsledky odlišné.



Mapa 14 Kolokačný vzor pre pirátske násilie a podozrivé priblíženie za roky 1989–2022

7 VÝSLEDKY

Diplomová práca priniesla viacero výsledkov, tie sú zhrnuté v nasledujúcich podkapitolách. Primárnym cieľom práce bolo vybrať vhodné dáta a na tie aplikovať postup vyhľadávania kolokačných vzorov v rámci nástroja Colocation Analysis v programe ArcGIS Pro. Toho bolo docielené vytvorením troch prípadových štúdii.

7.1 Prvá prípadová štúdia

V prvej prípadovej štúdii bolo podstatou predovšetkým bližšie preskúmanie nástroja, pochopenie jeho spôsobu zostavovania kolokačných vzorov. Práve preto bol vybraný pomerne malý záber študovaného územia a nízky počet vstupných bodov v skúmaných kategóriách. Bolo zostavených celkom osem variant dát v časti obce Brno – město pre možné odporúčania budúcim používateľom nástroja.

V procese prípravy dát boli zostavené dve hypotézy. Prvá predpokladala, že ak sú v blízkom okolí lekární subjekty zdravotníctva, skúmané kategórie vytvoria kolokačný vzor, čiže lekárne budú kolokované. Hypotéza sa potvrdila, lekárne s viditeľným okolím subjektov boli označované ako kolokované, no treba podotknúť, že kolokácia bola nevýznamná. Druhá hypotéza mala za predpoklad, že miesta s duplicitnými záznamami výrazne ovplyvnia sledované lekárne, a preto budú identifikované ako významne kolokované. Hypotéza bola zamietnutá, v žiadnom z nastavení sa takáto lekáreň neobjavila. Čo je ale možné prehlásiť pri variante s týmito záznamami je, že výsledok sa mierne zlepšil v zmysle počtu kolovaných lekární pri jednotlivých nastaveniach či už počtu susedných prvkov, alebo veľkosti vzdialenostného pásma. Každopádne maximálny počet takýchto lekární bolo sedem z celkového počtu deviatich lekární.

Okrem prítomnosti duplicít bola preskúmaný aj hraničný efekt, ktorý nastáva po vytrhnutí časti územia a na hraniach dochádza ku skresľovaniu výsledkov. Tento problém bol odstránený zostrojením obalovej zóny okolo časti obce Brno – město s veľkosťou 100 metrov. Následne boli do analýzy pridané subjekty zdravotníctva obsiahnuté v tejto zóne. Výsledok však nemal priaznivý efekt. Vzhľadom na to, že sa navýšil celkový počet skúmaných prvkov, došlo u niektorých nastavení k zhoršeniu v zmysle prehĺbenia izolovaných prvkov. Jedným z dôvodov nezlepšenia výsledného vzoru je určite fakt, že v miestach pridaných subjektov neboli lekárne spadajúce do časti obce. Netreba však zabúdať, že závery sa týkajú konkrétnych dát a pri iných vstupných dátach môže práve obalová zóna napomôcť.

Čo sa týka metódy zostavenie priestorového vzťahu, nie je možné povedať, ktorá je vhodnejšia. Viacmennej to záleží na užívateľovi, akú otázku si kladie pri zostavovaní hypotéz. Metóda vzdialenostného pásma je exaktnejšia, užívateľ si dokáže jednoducho odmerať vzdialenosť medzi dvojicou skúmaných prvkov a zvážiť, aká vzdialenosť je ešte relevantná. Na druhej strane, ak by užívateľa zaujímala významnosť vzťahu skúmanej kategórie s okolím napríklad dvoch prvkov susednej, tak by bola táto metóda určite vhodnejšia. Samozrejme sa ale nevylučuje, možnosť využiť všetky dostupné metódy.

7.2 Druhá prípadová štúdia

Podstatou druhej štúdie bolo preskúmanie možnosti pridania časového aspektu do vyhľadávania kolokačných vzorov. Boli zostavené tri podštúdie na území mesta Filadelfia, ktoré sa zaoberali časovou zložkou vrátane metódy vzdialenostného pásma. Ďalšie metódy v tejto štúdii vyšetrované neboli.

V prvej variante boli proti sebe postavené dva kriminálne činy, a to vraždy a domáce násilie. Podľa istej štúdie by medzi kategóriami mohol existovať kolokačný vzťah a cieľom tejto varianty to bolo overiť. Prvotne boli dáta rozdelené na roky, no bola skúmaná aj dátová vrstva so všetkými dostupnými rokmi. Časový interval bol nastavený na jeden rok a jeden mesiac so vzdialenosťami 100, 500 a 1000 metrov. Z výsledných vzorov je zrejmé, že so zväčšujúcou sa vzdialenosťou a väčším časovým intervalom je viac záznamov domáceho násilia kolokovaných. Každopádne len veľmi malé percento bolo označené za významne kolokované

Druhá varianta síce neskúmala časový aspekt, no veľmi zaujímavé bolo preskúmanie vzťahu medzi kriminálnymi činmi, barmi a pubmi. Na základe článku bola zostavená hypotéza o vhodnej veľkosti vzdialenostného pásma, a to v dĺžke jedného bloku. Podľa článku by práve v tejto vzdialenosti mal byť vzťah kriminálnych činov a podnikov najvýraznejší. A preto vznikla aj druhá hypotéza, ktorá predpokladá významný kolokačný vzor zločinov a podnikov. Pre vzdialenostné pásmo boli vybrané až tri hodnoty, takže pre dostupných šesť zločinov vzniklo 18 kolokačných vrstiev. Prvá hypotéza bola zamietnutá, pretože so stúpajúcou vzdialenosťou počet kolokovaných kriminálnych činov pribúdalo. Druhá hypotéza sa naopak potvrdila, najviac kolokovaných záznamov získali zločiny spojené s alkoholom.

Posledná varianta tejto štúdie spočívala v zostavení dvojíc zločinov a pozorovaní časopriestorového vzťahu medzi nimi. Podľa jednej štúdie majú najsilnejší vzťah zločiny uskutočnené v rámci jedného mesiaca. Každopádne z prvej varianty je už zrejmé, že so zväčšujúcim sa intervalom, počet kolokovaných záznamov stúpa. Preto okrem jedného mesiaca boli zostavené aj ďalšie intervaly, a to tri a šesť mesiacov. Hodnota vzdialenosti bola prebraná z druhej varianty, a to 411 metrov. Pri prvej skúmanej dvojici opilstva na verejnosti a porušení zákona o pití alkoholu na verejnosti bolo pri časovom intervale jedného mesiaca kolokovaných len 6 % záznamov. S predĺžením intervalu sa množstvo bodov zvyšovalo no maximálne na 11,1 %. Výrazne inak tomu nebolo ani pri druhej dvojici opilstva na verejnosti a prostitúcie. Maximálne bolo kolokovaných (významne aj nevýznamne) 12,5 %. Veľká časť záznamov ostala nedefinovaná, čo znamená, že nesplnili časovú a priestorovú podmienku. Na základe týchto skutočností bola hypotéza zamietnutá a je možné prehlásiť, že čas je výrazným faktorom pri zostavovaní kolokačného vzoru. Taktiež bude vo väčšine prípadov dochádzať k zvyšovaniu počtu kolokovaných bodov pri väčšom časovom intervale a vzdialenostnom pásme.

7.3 Tretia prípadová štúdia

Poslednou nepreskúmanou metódou tvorby priestorového vzťahu je zostrojenie externého súboru. Podstatou poslednej prípadovej štúdie je práve jej nasadenie a preskúmanie jej možností. Základnými dátami sú správy smerované proti námorníctvu. Z databázy ASAM bola vybraná len časť, a to navigačná oblasť IX, ktorá sa rozprestiera v okolí Arabského polostrova. Boli zostrojené dve varianty, v prvej boli vybrané kategórie typu obete a v druhej typy nepriateľov.

7.3.1 Nákladná loď a obchodné plavidlo

Primárne sa pracovalo s treťou možnosťou voľby priestorového vzťahu, no pre porovnanie boli zostrojené aj vzory s metódami K najbližších susedov a vzdialenostného pásma. Výsledné vzory fixnej a inverznej vzdialenosti mali zhodné výsledky. Čo sa týka percentuálneho zastúpenia jednotlivých kolokačných kategórií, množstvo kolokovaných bodov dosahovalo maximálne 30 %, z toho významne kolokovaných bolo najviac 21 %.

V prípade použitia časopriestorového okna dochádza k poklesu kolokovaných lodí, predovšetkým tých významných a ku identifikovaniu nedefinovaných záznamov. To sa stáva pomerne často, dôvodom je prísna podmienka. Pri pohľade na priestorovú distribúciu, kolokované lode sa vyskytujú predovšetkým Perzskom a Ománskom zálive. Veľké množstvo záznamov je aj v oblasti prielivu Báb el-Mandeb. Naopak izolované lode sa nachádzajú v miestach, kde je priestor viac otvorený. Takými miestami je Arabské more a Adenský záliv.

7.3.2 Pirátske násilie a podozrivé priblíženie

V druhej variante boli kategóriou záujmu záznamy s pirátskym násilím. Susednou kategóriou boli body podozrivého priblíženia. Opäť boli skúmané dostupné metódy a porovnané s prvými dvoma metódami v nástroji Colocation Analysis. Ukázalo sa, že so vzrastajúcou vzdialenosťou sa zvyšuje počet kolokovaných bodov, rovnako ako s počtom susedných prvkov. Maximálny podiel kolokovaných bodov bol získaný až pri pomerne veľkej vzdialenosti, kedy bola kolokovaná takmer polovica záznamov pirátskeho násilia. V ostatných nastaveniach sa ich podiel pohyboval medzi 20 až 30 %. Pri pohľade na priestorovú distribúciu, udržiava sa istý trend rovnako ako v prvej variante. Výnimkou je nastavenie vzdialenostného pásma na 100 metrov s časovým intervalom piatich rokov. V tomto kolokačnom vzore sa výnimočne v Adenskom zálive objavil vyšší počet významne kolokovaných záznamov. Každopádne je to iba jeden vzor a vzdialenosť pre určenie vzťahu je pomerne vysoká. Užívateľ by mal takú medznú hodnotu zvážiť.

7.4 Ďalšie výstupy

Okrem zostavených troch štúdií boli vytvorené aj ďalšie výstupy. Týmito výstupmi sú skript, návod a mapy v rámci jednotlivých štúdii. Samozrejmosťou a povinnou súčasťou je aj webová stránka a poster dokumentujúci priebeh diplomovej práce a jej výsledky.

7.4.1 Skript

Pre uľahčenie práce generovania kolokačných vzorov bol využitý Python skript, ktorý bol vytvorený v praktickej časti prvej prípadovej štúdie. Jeho tvorba prebiehala v editore PyScripter, ktorý je veľmi intuitívny. Podstatou skriptu je zostrojenie troch výstupných kolokačných vzorov v jednom cykle spustenia nástroja s tromi dostupnými typmi jadrových funkcií. Na konci ešte skript pridá k názvu výstupu písmeno G, B, alebo N podľa typu funkcie. V skripte bola využitá knižnica ArcPy.

7.4.2 Manuál

Jedným z výstupov diplomovej práce je aj manuál vychádzajúci z tretej prípadovej štúdie. Dôvodom voľby zostrojenia manuálu na základe výsledkov tretej štúdie je dostupnosť dát a prítomnosť časového aspektu v vstupných kategóriách. Dostupný je ZIP súbor obsahujúci súbory a zložky so vstupnými a výstupnými vrstvami, grafy a obrázky. Nedeliteľnou súčasťou je aj projekt vo formáte .aprx *(ArcGIS project)* pre program ArcGIS Pro. Manuál je voľnou prílohou číslo jeden a je spolu s dátami k stiahnutiu z webovej stránky diplomovej práce. Bol využitý pri praktickej výuke v rámci cvičenia predmetu KGI/PROPY Programování v GIS v apríli 2022. Manuál použilo deväť študentov a pri praktickej práci výsledky úloh odpovedali výsledkom v manuály. Neboli zistené žiadne nedostatky v texte manuálu a v pripravených dátach.

7.4.3 Mapové výstupy

V rámci štúdií boli pre zaujímavé kolokačné vzory zostavené mapy podľa vopred vytvorenej šablóny. V niektorých prípadoch obsahovali len skúmanú kategóriu z dôvodu prehľadnosti mapy, inokedy ako prvky záujmu, tak aj susedné prvky. Vzhľad legendy vygenerovanej z nástroja Colocation Analysis bol ponechaný, bola len preložená do slovenčiny. Dôvodom je prehľadnosť a možná budúca nadväznosť témy.

8 DISKUSIA

Primárnym cieľom diplomovej práce bolo preskúmať nový nástroj kolokačnej analýzy dostupný v programe ArcGIS Pro. Podstatou nebolo poskytnúť konkrétne hodnoty parametrov nástroja, ale otestovať ho na rôznych typoch dát ako tematicky, tak aj geograficky. Samotný proces nastavovania a voľby vstupných dát nevyžaduje žiadnu predprípravu dát. Body môžu byť obsiahnuté v jednej, alebo v dvoch dátových sadách, na výsledok to nemá vplyv. Drobnou komplikáciou môže byť aktivácia časového okna, ktoré vyžaduje, aby stĺpec s časom bol v korektnom dátovom formáte, a to dátum. Neopomenuteľnou súčasťou parametrizácie nástroja je voľba vhodnej metódy tvorby kolokačných vzorov a následnej hodnoty. Práve preto boli jednotlivé štúdie postavené tak, aby boli využité všetky možné metódy a odlišné hodnoty.

V prvej prípadovej štúdii boli využité dáta subjektov zdravotníctva a lekární. Vrstva subjektov obsahovala duplicitné body, čo znamená, že na konkrétnych súradniciach bolo v niektorých prípadoch aj viac než päť bodov. Tento fakt mohol výrazne ovplyvniť výsledky, a preto bola zostavená varianta, kedy boli body odstránené a následne bol porovnaný výsledok. Na druhej strane, podstatnejším výsledkom je varianta so skutočnými dátami a nie umelo upravenými, ktoré neodpovedajú skutočnosti. Dalo by sa namietať, že dáta nie sú v časovom súlade, subjekty sú aktuálne k roku 2018, lekárne sú novšie, z roku 2021. Je viac než pravdepodobné, že niektoré subjekty zanikli, zmenili adresu, či dokonca pribudli, nebolo možné získať novšie dáta či rozmiestnenie subjektov z roku 2021. Čo sa týka nastavení parametrov, boli použité pomerne široké hodnoty, avšak neexistuje doporučený postup, ktorý by určil vhodnú hodnotu či interval. Preto môže nastať jav, ak užívateľ zvolí príliš malú, alebo naopak veľkú hodnotu, získa falošne pozitívny výsledok a vôbec si to nemusí uvedomiť. Ďalšiu výzvou sú skreslené výsledky na hraniciach vymedzenej oblasti, tento problém bol riešený obalovou zónou, no opäť sa núka otázka vhodnej vzdialenosti. Paradoxom je, že v oblasti hraníc nebolo veľké množstvo bodov, a tým pádom výsledok nebol významne ovplyvnený hraničným efektom.

Druhá štúdia bola zameraná na časový aspekt. Z dostupných prieskumov bolo možné extrahovať isté časové intervaly či vzdialenosti, no výsledky sa nezhodovali s predstavami obsiahnutými v hypotézach. Vo väčšine prípadov dochádzalo k zvyšovaniu počtu kolokovaných záznamov pri väčšom zábere či už času, alebo priestoru. Tým pádom je takmer vylúčené, aby výsledky boli priaznivejšie pri užšom okolí. To však ale môže užívateľa zavádzať a bude zvyšovať hodnotu za cieľom zaujímavejších výsledkov. Nikde však nie je napísané, kde je maximálna hodnota, tá sa mení s počtom prvkov v skúmanej kategórii a nie je jednoduché ju určiť.

Tretia a posledná prípadová štúdia bola výrazne odlišná. Základný rozdiel je v umiestnení skúmaných záznamov, tie sú totiž na mori. Vzdialenosti medzi hraničnými bodmi je viac ako 2000 km, samozrejme by nemalo zmysel voliť takúto vysokú hodnotu. Okrem tejto odlišnosti bola štúdia iná aj doteraz nepreskúmaným spôsobom definovania priestorového vzťahu. Nástroj na jeho zostrojenie ponúka viacero možností, no neinformuje užívateľa o vhodnosti nasadenia konkrétnej metódy. V niektorých prípadoch sú kolokačné vzory výrazne odlišné a užívateľ nemusí vedieť určiť, ktoré sú vhodné a ktoré naopak nie.

9 ZÁVER

Cieľom diplomovej práce bolo nájsť tri vhodné dátové zdroje, nasadiť nástroj Colocation Analysis a vytvoriť tri prípadové štúdie. Spomínaný nástroj vytvorí kolokačný vzor, ktorý dokáže identifikovať vzťahy medzi dátami, ktoré nemusia byť na prvý pohľad viditeľné. Ďalšou úlohou v rámci práce bolo vybrať jednu štúdiu, a na základne získaných poznatkov vytvoriť manuál pre budúcich užívateľov. Poslednými produktami bol poster a webová stránka poskytujúca informácie o priebehu a výsledkoch diplomovej práce.

Teoretická časť práce sa zaoberá súhrnom získaných informácii o problematike kolokačného vzoru. Informuje o histórii, kto a kedy začal termín používať. Aké sú typické príklady výskytu kolokácií a demonštruje možnosti na troch jednoduchých príkladoch. Taktiež diskutuje o možnom vývoji a možnostiach zlepšenia jeho identifikácie. Nedeliteľnou súčasťou rešeršnej časti je teoretický popis nástroja a jeho parametrov.

V praktickej časti bolo prvým krokom nájsť vhodné dáta pre vstup do analýzy. Cieľom bolo využitie všetkých možností parametrizácie nástroja, preto bola voľba vstupných dát veľmi dôležitá. Na základe obsahu dátových zdrojov boli skonštruované tri prípadové štúdie, ktoré sú špecifické z hľadiska témy, priestorového rozsahu a konštrukcie kolokačných vzťahov.

Výsledkom sú poznatky o možnostiach nástroja, pochopenie jeho schopností a interpretácia výsledných generovaných kolokačných vzorov. Boli vytvorené nespočetné varianty vstupných dát a skúmané ich rozdiely. Dokonca pre prvú prípadovú štúdiu bol napísaný Python skript pre rýchlejšie generovanie výsledkov. V tejto štúdii šlo predovšetkým o preskúmanie možností prvých dvoch metód konštrukcie priestorových vzťahov. Generované kolokačné vzory boli analyzované podrobne z hľadiska zaradenia prvkov záujmu do kategórií kolokačnej analýzy. Taktiež boli vyšetrované možnosti jadrovej funkcie a ich vplyv na výsledné zaradenie. Ukázalo sa, že existencia duplicitných záznamov ako jedna z variant vstupných dát bola pozitívnym príspevkom analýzy. Naopak prítomnosť obalovej zóny výsledné kolokačné vzory nepodporila. Čo sa týka skúmaných metód (K najbližších susedov, Vzdialenostné pásmo, Priestorové váhy zo súboru), nie je možné povedať, ktorá metóda je vhodnejšia. Užívateľ by si mal sám zvoliť, na základe dostupných poznatkov, na základe čoho chce priestorový vzťah určiť. To súvisí aj s vhodnými hodnotami parametrov, záleží od charakteru dát.

V druhej prípadovej štúdii bolo podstatou preskúmanie časového aspektu pri tvorbe kolokačného vzoru. Boli vytvorené tri varianty, kde sa pracovalo len s metódou vzdialenostného pásma a časovou zložkou. V dvoch variantách boli vzťahy skúmané medzi jednotlivými kategóriami, avšak druhá varianta v poradí mala za úlohu analyzovať vzťah kriminálnych činov a barov. V každej zo štúdií boli na základe nájdených prieskumov vytvorené hypotézy, ktoré mali za úlohu isté predpoklady overiť. Zistením bolo, že so zvyšovaním hodnoty vzdialenosti a časového intervalu nástroj generuje vzory s vyšším počtom kolokovaných bodov. Preto boli hypotézy zamietnuté, avšak je nutné si uvedomiť, že hodnoty parametrov nie je správne stále navyšovať. Práve v takom prípade môže dôjsť ku generovaniu náhodných a falošných vzorov.

Posledná štúdia je špecifická spôsobom tvorby priestorových a časových vzťahov. Doteraz nebola využitá posledná voľba ich tvorby, a to matica priestorových váh. Tú je možné vytvoriť v samostatnom nástroji a následne vložiť do Colocation Analysis. Vo väčšine boli výsledky podobné klasickým metódam. Dokonca ponúkali aj rovnaké

metódy, ako je napríklad K najbližších susedov. Ich výsledné vzory ale neboli totožné. Výhodou tejto tretej možnosti je množstvo ponúknutých metód. Na druhej strane nástroj užívateľa neinformuje, ktorá metóda je vhodná pre istý typ dát. Tým pádom, ak sú výsledky odlišné, užívateľ nedokáže určiť, ktorý vzor je vhodný pre zamietnutie či prijatie skúmanej hypotézy či predpokladu.

Z tretej štúdie bol vytvorený aj manuál, ktorý dôkladne popisuje predpracovanie dát a tvorbu kolokačných vzorov na základe všetkých dostupných metód. Veľká časť je ale venovaná práve spomínanej tretej možnosti, matici priestorových váh. Súčasťou návodu sú ako upravené, tak aj pôvodné dáta, grafy, obrázky a projekt programu ArcGIS Pro.

Cieľ diplomovej práce bol splnený, boli vytvorené tri prípadové štúdie a manuál pre budúcich používateľov nástroja. Okrem toho bol spísaný aj skript, ktorý v pláne nebol.

POUŽITÁ LITERATÚRA A INFORMAČNÉ ZDROJE

AGRAWAL, Rakesh; SRIKANNT, Ramakrishnan. Fast Algorithms for Mining Association Rules in Large Databases. In BOCCA, Jorge; JARKE, Matthias; ZANIOLO, Carlo. *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco : Morgan Kaufmann Publishers Inc., 1994. ISBN 1558601538.

ARCDATA PRAHA [online]. 2021 [cit. 2021-09-02]. ArcČR® 4.0. Dostupné z WWW: <https://www.arcdata.cz/produkty/geograficka-data/arccr-4-0>.

ArcGIS Pro [online]. 2022a [cit. 2022-02-17]. Colocation Analysis (Spatial Statistics). Dostupné z WWW: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/colocationanalysis.htm>.

ArcGIS Pro [online]. 2022b [cit. 2022-02-17]. Generate Spatial Weights Matrix (Spatial Statistics). Dostupné z WWW: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/generate-spatial-weights-matrix.htm#GUID-E48ADC9D-8D7F-40B2-BB7C-1284B323FC4D>.

BARUA, Sajib; SANDER, Jörg. Statistically Significant Co-location Pattern Mining, In: Shekhar, Shashi; Xiong, Hui; Zhou, Xun (eds) *Encyclopedia of GIS*. Cham : Springer, Cham, 2017. Elektronické publikovanie, s. 2204 - 2212. https://doi.org/10.1007/978-3-319-17885-1_1552.

data.Brno [online]. 2021 [cit. 2021-09-02]. Mapa přístupnosti - Budovy / Accessibility map – Buildings. Dostupné z WWW: <https://data.brno.cz/datasets/mestobrno::mapa-p%C5%99%C3%ADstupnosti-budovy-accessibility-map-buildings-1/about >.

Esri [online]. 2022 [cit. 2022-02-20]. ArcGIS Pro. Dostupné z WWW: <https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview/>.

GROFF, Elizabeth. Exploring 'near': Characterizing the Spatial Extent of Drinking Place Influence on Crime. *Australian & New Zealand Journal of Criminology. 2011,* 44, 2, s. 156 - 79. https://doi.org/10.1177/0004865811405253.

GUSMÃO, Luciana; DALY, Marymegan. Evolution of sea anemones (Cnidaria: Actiniaria: Hormathiidae) symbiotic with hermit crabs. *Molecular Phylogenetics Evolution*. 2010, 56, 3, s. 868 - 877. doi: 10.1016/j.ympev.2010.05.001.

HUANG, Yan; SHEKHAR, Shashi; XIONG, Hui. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering*. 2004, 16, 12, s. 1472 - 1485. doi: 10.1109/TKDE.2004.90.

KOPERSKI, Krzysztof; HAN, Juawei Discovery of spatial association rules in geographic information databases. In: Egenhofer, Max; Herring, John (eds) *Advances in Spatial and Temporal Databases*. USA : Springer, Berlin, Heidelberg, 1995. Elektronické publikovanie, s. 47 - 66. ISBN 978-3-540-49536-9.

KUPKA, Michal. *Bakalárska práca: Michal Kupka* [online]. 2019 [cit. 2021-09-01]. Analýza prostorového vzoru subjektů působících v oblasti zdravotnictví. Dostupné z WWW: <https://www.geoinformatics.upol.cz/dprace/bakalarske/kupka19/>.

MAMOULIS, Nikos. Co-location Pattern, Algorithms, In: Shekhar, Shashi; Xiong, Hui (eds) *Encyclopedia of GIS*. Cham: Springer, Cham, 2017. Elektronické publikovanie, s. 254 - 260. https://doi.org/10.1007/978-3-319-17885-1_149.

MARITIME SAFETY INFORMATION [online]. 2022 [cit. 2022-03-18]. WORLDWIDE THREATS TO SHIPPING REPORTS. Dostupné z WWW: <https://msi.nga.mil/Piracy>.

Office of Naval Intelligence [online]. 2022 [cit. 2022-03-18]. Shipping Threat Reports. Dostupné z WWW: <https://www.oni.navy.mil/news/shipping-threat-reports/>.

OpenDataPhilly [online]. 2014 [cit. 2022-02-02]. Police Districts. Dostupné z WWW: <https://www.opendataphilly.org/dataset/police-districts>.

OpenDataPhilly [online]. 2022 [cit. 2022-02-02]. Crime Incidents. Dostupné z WWW: <https://www.opendataphilly.org/dataset/crime-incidents>.

Orage [online]. 2022 [cit. 2022-02-20]. *Data Mining Fruitful and Fun*. Dostupné z WWW: <https://orangedatamining.com/>.

PETR, Pavel. Metody Data Miningu. Pardubice : Univerzita Pardubice, 2014. 84 s. ISBN 978-80-7395-872-5. 2014.

PETROSKY, Emiko; BLAIR, Janet; BETZ, Carter; FOWLER, Katherine; JACK, Shane; LYONS, Bridget. Racial and Ethnic Differences in Homicides of Adult Women and the Role of Intimate Partner Violence — United States, 2003–2014. *MMWR and Morbidity and Mortality Weekly Report*. 2017, 66, 28, s. 741–746. doi:10.15585/mmwr.mm6628a1.

QGIS [online]. 2022 [cit. 2022-02-20]. *Discover QGIS*. Dostupné z WWW: <https://qgis.org/en/site/about/index.html>.

Reference [online]. 2020 [cit. 2022-02-22]. How Many Philadelphia City Blocks Are In a Mile?. Dostupné z WWW: <https://www.reference.com/geography/many-philadelphia-city-blocks-mile-b86508b161734922>.

RIPLEY, Brian. The second-order analysis of stationary point processes. *Journal of Applied Probability,* 1976, 13, 2, s. 255 - 266. doi:10.2307/3212829.

SALMENKIVI, Marko. Fast Evaluating attraction in spatial point patterns with an application in the field of cultural history. In RASTOGI, Rajeev; MORIK, Katharina; BRAMER, Max; WU, Xinddong. *ICDM '04: Fourth IEEE International Conference on Data Mining*. Brighton : Institute of Electrical and Electronics Engineers, 2004. ISBN 0-7695-2142-8.

SHEKHAR, Shashi; HUANG, Yan. Discovering Spatial Co-location Patterns: A Summary of Results. In: Jensen, Christian; Schneider, Markus; Seeger, Bernhard; Tsotras, Vassilis (eds) *Advances in Spatial and Temporal Databases*. USA : Springer, Berlin, Heidelberg, 2001. Elektronické publikovanie, s. 236 - 256. ISBN 978-3-540-47724-2.

SHEKHAR, Shashi; ZHANG, Pusheng; HUANG, Yan; VATSAVAI, R, R., Trends in spatial data mining ,In: Kargupta, Hillol; Joshi, Anupam, Sivakumar, Krishnamoorthy, Yesha, Yelena Data mining: Next generation challenges and future directions.Minneapolis : AAAI Press, 2003. Elektronické publikovanie, s. 357 - 380. ISBN 978-0262612036.

SOURCEFORGE [online]. 2022 [cit. 2022-02-20]. *Summary*. Dostupné z WWW: <https://sourceforge.net/projects/pyscripter/>.

VAN SLEEUWEN, Sabine E. M.; STEENBEEK, Wouter; RUITER, Stijn. When Do Offenders Commit Crime? An Analysis of Temporal Consistency in Individual Offending Patterns. *Journal of Quantitative Criminology*. 2021, 37, 4, s. 863 - 889. ISSN 0748-4518.

Wei, Hu. Co-location Patter Discovery, In: Shekhar, Shashi; Xiong, Hui (eds) *Encyclopedia of GIS*. Cham: Springer, Cham, 2017. Elektronické publikovanie, s. 247 - 260. https://doi.org/10.1007/978-3-319-17885-1_150.

XIONG, Hui; SHEKHAR, Sashi; HUANG, Yan; KUMAR, Vipin; MA, Xiaobin; YOC, Jin Soung. A framework for discovering co-location patterns in data sets with extended spatial objects. In BERRY, Michael; DAYAL, Umeshwar; KAMATH, Chandrika; SKILLICORN, David. *Proceedings of the 2004 SIAM international conference on data mining*. Florida : Society for Industrial and Applied Mathematics. 2004. https://doi.org/10.1137/1.9781611972740.8.

YANG, Hui; PARTHASARATHY, Srinivasan; MEHTA, Sameep. Mining Spatial Object Patterns in Scientific Data. *IJCAI '05: Proceedings of the 19th international joint conference on Artificial intelligence.* Edinburgh : Morgan Kaufmann Publishers Inc., 2005.

PRÍLOHY

ZOZNAM PRÍLOH

Voľné prílohy

Príloha 1 Manuál

Príloha 2 Poster

Príloha 3 CD

Elektronické prílohy

Príloha 4 Manuál

Popis štruktúry DVD

- Adresáre:
  - Prílohy
    - Manuál
    - Poster
    - Adresár k manuálu
  - Data
  - Web
  - Python skript
  - Text práce