Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra geoinformatiky

APLIKACE ASOCIAČNÍCH PRAVIDEL NA PROSTOROVÁ DATA

Diplomová práce

Lenka TRNOVÁ

doc. Ing. Zdena Dobešová, Ph.D.

Olomouc 2020
Geoinformatika

ANOTACE

Diplomová práce se zabývá generováním asociačních pravidel pro vybrané prostorové datové sady. Asociační pravidla jsou součástí data miningu, díky kterému lze získat nové informace z dat. Všechny tyto informace nemusí být na první pohled patrné, proto je důležité věnovat pozornost i velmi málo frekventovaným pravidlům. Díky nim je možné se dozvědět o prostorové asociaci mezi prostorovými jevy, které si nemusí být ani zdaleka podobné. Aby ale bylo možné pravidla pro prostorová data generovat, je nutné je upravit. Součástí práce je nalezení více způsobů, jak toho dosáhnout. Veškerá práce probíhá v open-source řešeních. V rámci práce je otestováno několik technických řešení, které umožňují generovat asociační pravidla. V případových studiích je nadále pracováno pouze se *SW Orange*. Zároveň je kladen důraz na vizualizaci výsledných pravidel zpátky do prostoru. Doplňujícím nástrojem jsou *Google Tabulky*, ve kterých dochází ke zjednodušení úpravy asociačních pravidel. Dále je také vyvíjený skript pro *QGIS* (*PyQGIS*), který umožňuje nahrát upravená pravidla jako podmínky pro vizualizaci v mapě. Závěr práce je obohacený o přehledný návod, jak celého postupu, od úpravy dat až po finální mapu s asociačními pravidly, dosáhnout.

KLÍČOVÁ SLOVA

Asociační pravidla; prostorová data; Apriori; kolokační vzor; data mining

Počet stran práce: 65
Počet příloh: 5 (z toho 2 volná a 2 elektronické)

ANOTATION

The Master's thesis deals with the generating association rules for selected spatial datasets. Association rules are part of data mining, which helps obtain new information from the dataset. All this information may not be obvious at first glance, so it is important to pay attention to the rules that are not frequent. Thanks to them it is possible to learn about spatial colocation between spatial elements, which may not be like each other. To generate spatial data, it needs to preprocess them. Part of the work is focused on finding more ways how to achieve it. All work is done in open-source software. Many software can generate association rules. In case studies, only software called Orange has been used. Also, work is focused on visualization of the resulting rules back into spatial. The complementary tool is Google Sheets, which simplifies the modification of association rules. Also, a script for QGIS (*PyQGIS*) has been developed to import modified rules as conditions for visualization on the map. The conclusion of the thesis is summed up with a clear step-by-step manual with the whole procedure, from data preprocessing to the final map with association rules.

KEYWORDS

Association rules; spatial data; Apriori; collocation pattern; data mining

Number of pages: 65
Number of appendixes: 5

Dílčí části diplomové práce byly realizovány v rámci projektu IGA_PrF_2020_027 „Pokročilé aplikace geoinformačních technologií pro prostorové analýzy, modelování a vizualizace jevů reálného světa“ podpořeného interní grantovou agenturou Univerzity Palackého v Olomouci.

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 Lenka Trnová

Děkuji vedoucí práce doc. Zdeně Dobešové za podněty a připomínky při vypracování práce. Dále děkuji doktorce Aleně Vondrákové za pomoc s mapovými výstupy. Zároveň děkuji rodině, a především příteli za podporu během celého studia.

UNIVERZITA PALACKÉHO V OLOMOUCI
Přírodovědecká fakulta
Akademický rok: **2018/2019**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Bc. Lenka TRNOVÁ**

Osobní číslo: **R18868**

Studijní program: **N1301 Geografie**

Studijní obor: **Geoinformatika**

Název tématu: **Aplikace asociačních pravidel na prostorová data**

Zadávající katedra: **Katedra geoinformatiky**

Zásady pro vypracování:

Cílem práce je aplikovat na vhodná data postup generování asociačních pravidel v souvislosti s prostorovými daty. Studentka vybere a navrhne tři případové studie, na kterých demonstruje postup a nasazení vhodných postupů a softwarů tak, aby byly prezentovány asociační pravidla, jak častá, tak preferenční pravidla včetně výjimek z asociačních pravidel.
Studentka vyplní údaje o všech datových sadách, které vytvořil nebo získal v rámci práce, do Metainformačního systému katedry geoinformatiky a současně vytvoří zálohu údajů ve formě validovaného XML souboru. Celá práce (text, přílohy, výstupy, zdrojová a vytvořená data, XML soubor) se odevzdá v digitální podobě na CD (DVD) a text práce s vybranými přílohami bude odevzdán ve dvou svázaných výtiscích na sekretariát katedry. O diplomové práci studentka 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 grafických prací: **dle potřeby**

Rozsah pracovní zprávy: **max. 50 stran**

Forma zpracování diplomové práce: **tištěná**

Seznam odborné literatury:

Šarmanová J.: Metody analýzy dat, VŠB-TU, Ostrava, 2012, ISBN 978-80-248-2565-6.
Tan P.N., Steinbach M., Kumar V. Introduction to Data Mining, chapter 6 Association Analysis: Basic Concepts and Algorithms
Petr. P. Metody Data Miningu (část 1, část 2). Pardubice, Univerzita Pardubice. 2014
Dao, T.H.D. (2018). Rule Learning for Spatial Data Mining. The Geographic Information Science & Technology Body of Knowledge, John P. Wilson (ed.). DOI: 10.22224/gistbok/2018.1.3
Shekhar S., Xiong H., Zhou X. Encyclopedia of GIS. Springer, Cham, 2017, ISBN: 978-3-319-17884-4
Voženílek, V. Diplomové práce z geoinformatiky. Univerzita Palackého, Olomouc, 2002, 31 s.

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

Datum zadání diplomové práce: **1. listopadu 2018**

Termín odevzdání diplomové práce: **5. května 2020**

UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA GEOINFORMATIKY
17. listopadu 50, 771 46 Olomouc

L.S.

doc. RNDr. Martin Kubala, Ph.D.
děkan

prof. RNDr. Vít Voženílek, CSc.
vedoucí katedry

V Olomouci dne 10. prosince 2018

OBSAH

SEZNAM POUŽITÝCH ZKRATEK

Zkratka	Význam
AQI	Air quality index
CSV	Comma-separated values
ČÚZK	Český úřad zeměměřický a katastrální
DM	Data mining
EEA	Evropská agentura pro životní prostředí
GIS	Geografický informační systém
HTML	Hypertext Markup Language
KP	Kreativní průmysl
MS	Microsoft Office
OSM	OpenStreetMap
POI	Point of interest
RÚIAN	Registr územní identifikace, adres a nemovitostí
SARM	Spatial Association Rule Mining
SQL	Structured Query Language
SW	Software
XLS	Excel Spreasdheet

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM MAP

ÚVOD

Asociační pravidla je jedna z metod data miningu, díky které lze nalézt mezi daty asociace, které nemusí být na první pohled zřetelné. Primárně se jedná o neprostorová data, ale v rámci této práce byla snaha hledat asociační pravidla pro prostorová data. Prostorová data nelze ale nahrát přímo do softwaru, který umožňuje generovat asociační pravidla. Prostorová data je nutné předpřipravit. V teoretické části dojde k seznámení se s asociačními pravidly, jejich děleními a zmínění dostupných prací, které se tímto tématem zabývaly. Dojde k seznámení se s dostupnými nástroji, které pracují s asociačními pravidly. Část z nich bude následně použito v praktické části, ve které se otestuje jejich funkcionalita. Na základě vytvořené testovací prostorové datové sady bude zvolený SW Orange, ve kterém se bude nadále pracovat.

V práci jsou sepsány celkem tři případové studie, ve kterých je nalezen postup, jak upravit prostorová data v QGIS tak, aby bylo možné pro ně hledat asociační pravidla. Součástí práce je také zobrazení asociačních pravidel zpět do prostoru pomocí mapových výstupů s bodovou metodou. Pro ulehčení práce byl také vytvořen pomocný Python skript, který je spustitelný v QGIS. Závěrečným výstupem práce je step-by-step návod, ve kterém je podrobné sepsán celý postup od úpravy dat až po výslednou vizualizaci výsledných asociačních pravidel.

1 CÍLE PRÁCE

Cílem diplomové práce je nalézt vhodný způsob, jak aplikovat asociační pravidla na prostorová data. Způsob bude následně aplikován na tři praktické případy. Dílčí částí bude nalezení technického řešení v open-source prostředí.

V teoretické části bude představení data miningu (DM) jako metody získávání užitečných informací z dat. Následovat bude samotný rozbor asociačních pravidel a k čemu slouží. Dílčí část rešerše bude zaměřena na analýzu dostupných prací týkajících se aplikace asociačních pravidel na prostorová data. Dále dojde k rozboru dostupných technických řešení. Z těchto řešení se následně zvolí jedno, které bude využito do praktické části.

V praktické části dojde k vytvoření uceleného postupu, jak upravit prostorová data do takové podoby, aby bylo možné vygenerovat asociační pravidla. Dále budou vytvořena testovací data, která budou sloužit k ověření funkčnosti vytvořeného postupu. Po ověření funkčního postupu se vytvoří 3 případové studie týkající se prostorových dat, pro které se budou hledat asociační pravidla. Cílem bude naleznout i taková pravidla, která nejsou na první pohled patrná, nejsou tedy frekventovaná. Na případových studiích bude demonstrován vhodný postup a nasazení softwaru (SW).

Výsledkem práce bude step-by-step návod, jak prostorová data upravit tak, aby bylo možné pro ně asociační pravidla vygenerovat. Zároveň v návodu bude uvedený postup, jak tato pravidla zobrazit zpět do prostoru. Cílem práce není nalézt co nejracionálnější asociační pravidla, ale nalézt způsob, jak prostorová data lze upravit, aby toho bylo možné docílit.

2 METODY A POSTUPY ZPRACOVÁNÍ

Nalezení způsobu, jak aplikovat asociační pravidla na prostorová data s sebou nese několik fází práce. V první řadě je potřeba nadefinovat, jaké metody budou v rámci praktické části využity. Je nutné se seznámit s použitými daty, které budou vstupovat do jednotlivých testování jako doplňující vrstvy. V neposlední řadě je potřeba zmínit programy, které budou součástí testování, nebo budou sloužit jako doplňující nástroj k tvorbě této práce.

2.1 Použité metody

Prostorové analýzy

V rámci praktické části bylo použito několik prostorových metod, z nichž nejčastější z nich byla **tvorba obalové zóny** (*bufferu)* kolem prvků. Jedná se o nástroj spadající do kategorie analýzy blízkosti. Pomocí něj lze získat informace o okolí prvku. Jedná se o oblast, která je uživatelem nadefinovaná pomocí vzdálenosti (např. v metrech). Nástroj vytvoří takovou oblast, která je od všech uzlů prvků vzdálena právě tuto vzdálenost.

Dalším velmi využívaným nástrojem je **připojení dat podle umístění** (*spatial join*). Nástroj umožňuje převést atributy jedné datové sady do druhé na základě jejich prostorové vazby. Prostorová vazba je předem definovaná. Může se jednat o protínání vrstev, dotyk vrstev na hranici, jeden prvek se nachází uvnitř prvku druhé vrstvy atd.

V rámci sjednocení dvou vrstev stejné tématiky, kdy jsou data v bodové a zároveň polygonové podobě bylo využito nástroje pro **tvorbu centroidů** *(centroids)*. Dalším krokem bylo **spojení** těchto vrstev *(merge vector layers)*.

V neposlední řadě bylo využito **kalkulátoru polí** *(field calcultor)* pro zapisování vybraných hodnot do nově vytvořených atributových sloupců.

Skriptování

Pro tvorbu skriptu bylo využito programování v jazyce Python, jež je podporovaným jazykem pro tvorbu skriptů do QGIS. Python jazyk je označován jako jeden z nejjednodušších programovacích jazyků. Jeho výhodou je jednoduchá syntaxe a velmi přehledný kód. V poslední letech se Python stává stále oblíbenějším jazykem pro programování nejrozmanitějších programů a nástrojů.

2.2 Použitá data

Zde jsou sepsány datové sady, které byly využity v rámci testování jako doplňující vrstvy pro získání informací o okolí daného prostorového prvku. Samotné primární datové sady, ke kterým se tyto doplňující vrstvy připojují, jsou představeny v rámci příslušné případové studie (Rekola, nabíjecí stanice a dětská hřiště v Olomouci).

Adresní body

Bodová vrstva vzniklá v rámci RÚIAN (*Registr územní identifikace, adres a nemovitostí*) pod záštitou ČÚZK (*Český úřad zeměměřický a katastrální*). Zapůjčeny doc. Burianem z Katedry geoinformatiky. Datová sada je za území České republiky a obsahuje celkem 62 atributů.

OSM

OpenStreetMap je projekt, jehož cílem je poskytovat volně dostupná prostorová data, jež se následně zobrazí formou topografických map. Jedná se o vektorovou databázi, kde uživatelé mohou sami data upravovat či přidávat. Tato data lze pomocí speciálních webových stránek či nástrojů stahovat a používat pro své vlastní účely. V rámci této práce bylo využito QGIS extenze QuickOSM, pomocí které byly získávány vybrané prostorové informace. Dále bylo využito balíčku OSM pro území celé České republiky, které lze stáhnout např. na portálu *Geofabrik*[1]. O konkrétních vrstvách je psáno podrobněji v jednotlivých případových studiích.

Podkladová mapa

Jako podklad pro výsledné mapy byly využity dostupné podklady v QGIS v záložce *XYZ Tiles*. Pro detailní mapy bylo použito topografické mapy *Esri Topo World*, pro celé území ČR *Esri Gray (light).*

Urban Atlas 2012

Volně dostupná datová sada, která nabízí využití území pro celou Evropu. V rámci první případové studie byla použita data pro území města Olomouce. Data byla stažena z webové stránky Copernicus, kde je umožněno data stahovat po přihlášení se do bezplatného účtu.

2.3 Použité programy

QGIS Desktop 3.8.3

Veškerá práce s prostorovými datovými sadami probíhala v QGIS Desktop 3.8.3., jež je vhodnou open source alternativou k placenému ArcGIS for Desktop. V QGIS byly provedeny prostorové analýzy a veškeré atributové úpravy. Dále byl také použit zásuvný modul *QuickOSM*, který slouží pro stahování vybraných dat z databáze OSM. V rámci QGIS došlo také ke zpětné vizualizaci výsledných asociačních pravidel pro případové studie s prostorovými daty. V rámci QGIS byl také vytvořen jednotný vizuální styl pro výsledné prezentované mapy.

Orange 3.22.0

Orange je vizuální data miningový software. Byl zvolen pro generování asociačních pravidel na základě předchozího výběru vhodných DM softwarů. Jedná se o open-source program, který byl vyvinut v roce 1996 je napsán v Python, Cython, C++ a C. Je multiplatformní a obsahuje mnoho nástrojů pro data mining. Orange je vyvíjen na University of Ljublaň ve Slovinsku. Je dostupný zdarma.

Google Tabulky

Pro úpravu tabulek bylo použito online nástroje Google Tabulky od společnosti Google. Komerční alternativou může být Excel z balíčku Microsoft Office (MS). Program mimo jiné také umožňuje tabulky uložit do požadovaného formátu CSV *(Comma-separated values).* V rámci programu bylo využito maker, která slouží k zautomatizování práce.

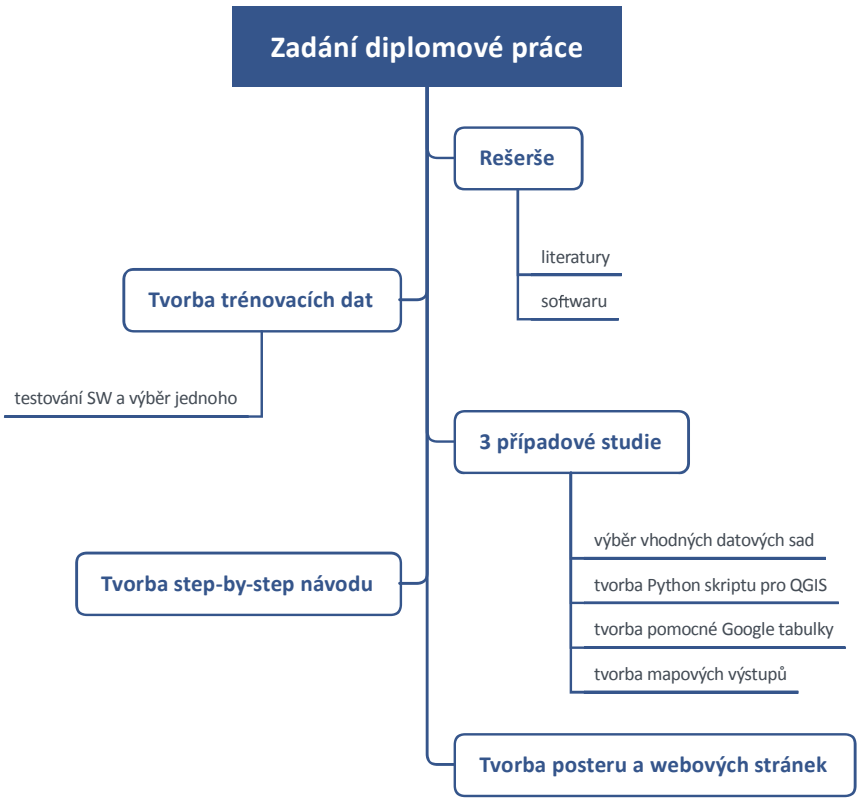
[1] https://download.geofabrik.de/europe/czech-republic.html

2.4 Postup zpracování

Obr. 1 popisuje diagram postupu diplomové práce. První krok představuje rešerši knižních zdrojů, odborných článků a elektronických studií. Tyto práce budou mimo jiné sloužit jako podklad pro nalezení vhodného SW (*software*) pro generování pravidel. Součástí rešerše bude také snaha nalézt práce týkající se přímo aplikace na prostorová data. Díky tomu bude možné získat představu o tom, jaký typ dat, resp. jaká metoda úpravy prostorových dat je použitelná. Následně bude připravena testovací datová sada, která bude sloužit pro seznámení se s dostupným technickým řešením. Ověří se jeho funkčnost a spolehlivost. Tato testovací datová sada bude zcela smyšlená a bude pouze experimentální.

Do praktické části bude následně zvoleno takové technické řešení, které bude z mnoha ohledů nejvhodnější. V tomto technickém řešení bude provedená veškerá práce, co se generování asociačních pravidel týče. Součástí práce je vytvoření tří případových studií, kde budou hledána asociační pravidla. Datové sady bude potřeba na základě nastudované literatury jistým způsobem upravit tak, aby na ně bylo možné nasadit zvolený nástroj. Úpravou se rozumí využití prostorových analýz v rámci QGIS viz *Použité metody*. Nedílnou součástí praktické části je vytvoření skriptu, který poslouží k ulehčení vizualizace jednotlivých asociačních pravidel. Také se vytvoří jednotný vizuální styl, který poslouží pro mapové výstupy zobrazující vygenerovaná asociační pravidla.

V závěru práce bude sepsán přehledný postup úpravy dat až po samotné vygenerování asociačních pravidel. Postup bude sloužit jako step-by-step návod pro aplikaci na vlastní datové sady. Návod musí být sepsán tak, aby bylo možné všechny jeho kroky zreplikovat. V poslední fázi budou vytvořeny náležitosti jako je poster a webové stránky, informující o výsledcích dosažených v této diplomové práci.



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

3 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

Data mining je metoda analýzy dat, díky které lze nahlížet na datové sady z jiné perspektivy a získávat užitečné informace, které lze využít v rozmanitých odvětvích. Asociační pravidla jsou jedna z metod data miningu, která umožňuje nalézt důležité vztahy ve velkých datových sadách, které by byly jinak pro nás skryté. Pravidlo slouží k vyjádření jisté zákonitosti, tzv. vzorce v datové sadě. Obecná forma asociačního pravidla je následující:

$$Antecedent \rightarrow Consequent\ [podpora\ \%, spolehlivost\ \%]$$

Jedná se o predikáty – *Antecedent* značí předpoklad, *Consequent* závěr. Šipka značí implikaci a samotný vztah můžeme číst: „Jestliže je splněná podmínka (předpoklad), pak platí závěr“. Podpora (*support*) říká, jak často je pravidlo aplikovatelné na daný datový set, zatímco spolehlivost (*confidence*) říká, jak četný je výskyt daných objektů, kterých se pravidlo týká. Díky hodnotě podpory lze eliminovat ta pravidla, která jsou velmi ojedinělá a tvoří spíše náhodu. Na straně předpokladu a závěru může být více podmínek než pouze jedna, ale to s sebou přináší problém. Rouse (2018) ve svém článku uvádí, že čím více položek asociační pravidlo obsahuje, tím menší význam pravidlo má.

Je důležité zvolit správně hodnoty podpory a spolehlivosti. Jayababu a kol. (2018) se ve své práci zabývali právě testováním různých hodnot těchto proměnných s jejich vlivem na výsledný počet vygenerovaných pravidel. Čím vyšší minimální podpora je, tím méně vygenerovaných asociačních pravidel je. Šarmanová (2012) obecně doporučuje nastavit primárně minimální podporu na 5 %. U spolehlivosti doporučuje hodnotu:

- 90 % pro přesná data, které nejsou subjektivně ovlivněné (lékařská data)
- 70 % pro data, kde se může projevit subjektivní hodnocení (data z dotazníků)

Tab. 1 Čtyřpolní tabulka četností

	závěr platí	závěr neplatí
předpoklad platí	a	b
předpoklad neplatí	c	d

Peter (2014) dále definuje termín pokrytí *(coverage)*, což je podmíněná pravděpodobnost předpokladu, pokud platí závěr:

$$\frac{a}{a+c}$$

Dále se také definuje hodnota *lift.* Ten říká, zda se objekty nacházejí spolu častěji, než bylo očekáváno. Lift dosahující hodnot menších jak 1 vykazuje mezi objekty zápornou korelaci. Pokud je hodnota vyšší jak 1, jsou pozitivně korelovány a výskyt těchto objektů společně je vyšší, než se předpokládalo. Hodnota rovná jedné říká, že objekty jsou vůči sobě nezávislé.

Každý z nabízených SW nástrojů vygeneruje výsledek (ať ve struktuře tabulky nebo jiného reportu), ve kterém vyjadřuje pomocí charakteristik sílu jednotlivých asociačních pravidel. Tabulka může vypadat následovně:

Tab. 2 Ukázkové vygenerované asociační pravidlo

Předpoklad	Závěr	Výskyt	Podpora [%]	Spolehlivost [%]
Voda_ANO	Suchá_NE	5	33,33	100

Předpokladem zde je, že je rostlina v blízkosti vody, závěr říká, že rostlina není suchomilná. Pravidlo má podporu 33,33 % což znamená, že 33,33 % dat z celé datové sady splňuje tuto podmínku. Spolehlivost uvádí, že ve 100 % těchto případů je pravidlo pravdivé.

3.1 Dělení asociačních pravidel

Existuje mnoho dělení asociačních pravidel podle autora daného vědeckého článku. Zde je výčet těch nejzajímavějších.

Slimani (2014) dělí pravidla do těchto skupin:

a) Jednorozměrné – objekty nebo atributy jsou vztaženy pouze k jednomu rozměru (dimenzi)
b) Mnohorozměrné – objekty nebo atributy vztaženy k více rozměrům
c) Logický datový typ *(Boolean)* – asociace mezi přítomností a absencí objektu – hodnoty 0/1
d) Kvantitativní – hodnoty mají kvantitativní povahu nebo jsou rozdělené do intervalů
e) Korelační – tento způsob může generovat velké množství pravidel, které je následně zahrnuto do statistické korelace, ze které lze následně získat důležitá pravidla

Suzuki (1998) definuje 3 typy pravidel:

a) Očekáváné z reality *(common sense rule)* – pravidlo s vysokou podporou a vysokou spolehlivostí
b) Referenční *(reference rule)* – nízká podpora a nízká spolehlivost
c) Spolehlivá výjimka *(reliable exception rule)* – nízká podpora, ale vysoká spolehlivost

Šarmanová (2012) zmiňuje tři typy pravidel:

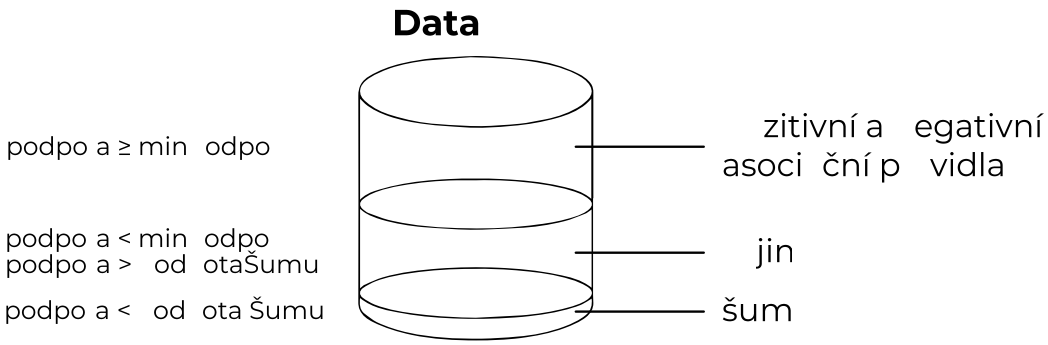
a) Klasické – asociace mezi 2 objekty
b) Transakční – rozsáhlá množina atributů
c) Agregované – podmnožina atributů

Transakční data jsou typicky data nákupního košíku. Jedná se o seznam zboží, které nakoupil zákazník v konkrétní den. Díky asociačním pravidlům lze nalézt nové informace o zboží, které se společně nakupuje. Konkrétním příklad, který Šarmanová uvádí může být, že zákazník, který si koupí jogurt, si k němu se spolehlivostí 72 % koupí i vločky. Díky těmto poznatkům mohou obchody cílit marketing na tyto dva produkty a umístit je strategicky ve svých prodejnách tak, aby si zákazník nakoupil se zbožím i jiný produkt.

Dále také zmiňuje pojem frekventované a silné pravidlo. **Frekventované pravidlo** je takové, které splňuje minimální podporu, která byla uživatelem definována. **Silné pravidlo** splňuje, jak minimální podporu, tak i minimální spolehlivost.

3.1.1 Spolehlivá výjimka

Taniar (2008) se ve své práci navazuje na definování asociačních pravidel dle Suzuki (1998) a dále uvádí, že datová sada se skládá ze 3 složek – pozitivních a negativních asociačních pravidel, spolehlivých výjimek a ze šumu. Šum je náhodný a nespolehlivý jev v datech. Složení databáze popisuje podle obrázku níže.



Obr. 2 Pravidla v databázi *(upraveno dle Yoo, 2005)*

Nejvíce ze všech se zabýval ve své práci výjimkami. Takové pravidlo není v databázi frekventované, ale oba objekty jsou vždy spolu. Právě tyto výjimky jsou důležitým krokem pro nalezení zajímavých vzorů ve velkých datových sadách. Sám ve své práci uvádí, že většina výzkumů se naopak zaměřuje pouze na frekventovaná pravidla a tyto výjimky jsou jim často naprosto cizí.

Příkladem z prostoru může být např. práce Yoo a kol. (2005), kteří zmiňují, že ekologové přišli na vzájemný výskyt (symbiózu) krokodýlů nilských společně s kulíkem nilským. Výjimka má nízkou podporu, ale naopak má vysokou spolehlivost.

$$Krokodýl\ nilský \rightarrow Kulík\ nilský$$

Postup nalezení výjimek je následující. Jako první krok je nutné nalézt v datové sadě silná asociační pravidla, jak pozitivní, tak i negativní. Silné pravidlo chápe tak, že má vysokou podporu i spolehlivost. Jedná se o asociaci, která je pozorována po dlouhou dobu a tento jev vždy nastane. Takové pravidlo lze brát jako normu, od které se budou hledat výjimky. Výjimka totiž nastane pokaždé, pokud není splněno silné pravidlo.

3.2 Algoritmy

Existuje velké množství algoritmů, které slouží pro generování asociačních pravidel. Níže je výběr nejvíce zmiňovaných ve vědeckých pracích, které byly součástí teoretické rešerše.

3.2.1 Algoritmus CARMA

Algoritmus CARMA *(Continuous Association Rule Mining Algorithm)* je využívaný pro data mining obrovských datových sad, kde dochází ke zmenšování intervalu podpory pro každou položku. Umožňuje uživateli minimální hodnotu podpory kdykoliv během prvního skenování. CARMA provádí Apriori algoritmus na nízké minimální podpoře. Je vhodný do velkých datových sad, které jsou pro Apriori neřešitelné. Jeho výhodou je rychlost a počet průchodů potřebných pro vyhledání všech kombinací položek.

3.2.2 Apriori algoritmus

Nejpoužívanějším algoritmem pro generování pravidel je Apriori. Základem algoritmu je mnohonásobné procházení transakčních dat, dokud neobjeví opakující množiny položek (*frequent itemsets*). V každém cyklu se hledá kandidátní množina položek a je vypočtena skutečná podpora těchto položek. Poté dojde k vygenerování asociačních pravidel s požadovanou četností.

3.2.3 ECLAT algoritmus

Nesiba (2019) uvádí, že tento algoritmus je obdobný Apriori algoritmu, jen se liší v rovině, ve které s daty pracuje. Zatímco Apriori pracuje v horizontální rovině, ECLAT pracuje ve vertikální. Výhodou je nižší časová náročnost, je ale vhodný spíše na menší datové sady.

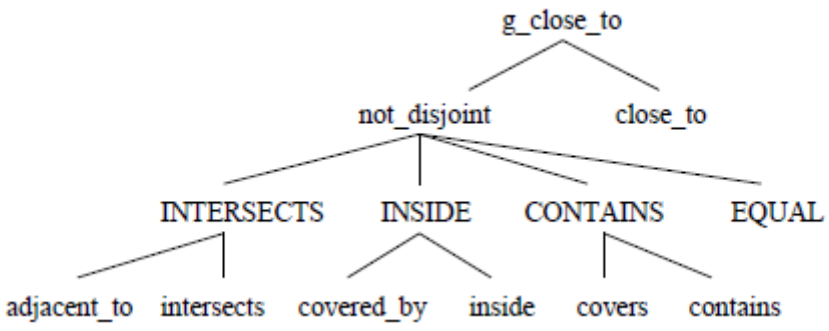
3.2.4 Algoritmus FP-Growth

Jeden z alternativních algoritmů, který může být použit pro určení nejběžnějších dat v datové sadě. Algoritmus hledá asociační pravidla pomocí hodnot parametrů podpory a spolehlivosti. Tento algoritmus využil Supiyandi a kol. (2017) ve své práci týkající se analýzy prodeje ovoce, která bude použita ke zjišťování propagační strategie pro zlepšování celkového prodeje.

3.3 Data mining pro prostorová data

Chattamvelli (2011) ve své knize definuje zkratku SARM (*Spatial Association Rule Mining*). Touto zkratkou můžeme označit takovou datovou sadu, která obsahuje alespoň jeden prostorový atribut. Dále uvádí, že často je předpoklad prostorový údaj, zatímco závěr neprostorový. Zatímco u klasických asociačních dat se ke generování pravidel využívají transakční data, pro prostorové datové sady taková data nejsou.

Prostorové datové sady jsou často obsáhlé, co se počtu záznamů týče. Pro uživatele je náročné v takovém objemu dat pracovat a hledat důležité detaily. Právě proto je důležitý SARM, který se snaží tuto práci zautomatizovat. Prostorová data v sobě obsahují vzájemné vztahy, které jsou pro asociační pravidla důležitá. Samotná asociační pravidla pro prostorová data mají pomoci nalézt jisté pravidelnosti mezi objekty ve velkých datových sadách. Prostorová asociační pravidla často využívají predikáty jako je *je_blízko*, *protíná*, *překrývá* apod. Tyto predikáty využil např. Koperski (1995) ve své práci, kde pracoval s datovou sadou velkých měst a dalších okolních objektů v Britské Kolumbii.



Obr. 3 Hierarchie topologických vztahů *(zdroj: Koperski, 1995)*

Jistým problémem může být fakt, že generování asociačních pravidel je cíleno na kategorická data, a ne na data numerická (např. vzdálenost). Mennis (2005) ve své práci uvádí jeden z možných přístupů. Vzdálenostní hodnoty se rozdělí do kategorií *blízko* a *daleko*. Důležitou proměnnou je stanovení intervalů pro dané kategorie, jelikož ve výsledku mohou ovlivnit samotná pravidla.

Hledáním asociačních pravidel pro prostorová data se zabývalo několik vědeckých prací. Např. Chen a kol. (2011) studovali asociační pravidla mezi dvěma prostorovými vrstvami – digitální model reliéfu a využití území. Z digitálního modelu využili sklon, nadmořskou výšku a orientaci, které zařadili do kategorií. Ke generování pravidel využili Apriori algoritmus.

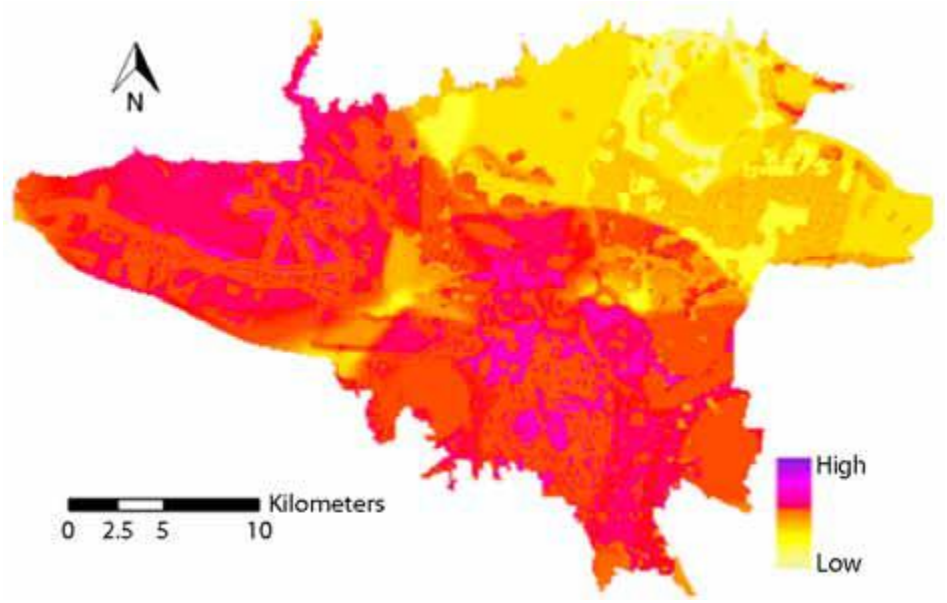
Další práce zkoumala výskyt zločinů v závislosti na umístění obchodů s alkoholem (Roncek a kol., 1991). V jejich práci došli k závěru, že výskyt obchodů s alkoholem signifikantně navyšuje počet spáchaných zločinů v dané lokalitě.

Mennis (2005) studoval město Denver z pohledu urbánního růstu. Do analýzy vstupovalo využití území a socioekonomické prvky za časové období.

Sukaesih Sitanggang (2013) ve své práci hledal asociační pravidla pro zjištění vlivu proměnných na výskyt lesních požárů. Z datových sad využil jak fyzickogeografická (využití území, řeky, silnice), tak i socioekonomická data (počet obyvatel, počet škol). Data měli uložené v databázovém prostředí PostgreSQL, které oplývá několika prostorovými predikáty. Predikát je takové sdělení, u kterého lze získat odpověď ve formě boolean odpovědi. Tedy lze na něj odpověď ano nebo ne. Pomocí *ST_Within* se na data dotazovali, zda se dané body o výskytu požárů nacházejí v konkrétním polygonu využití území. Výsledkem práce bylo vygenerování mnoha pravidel, ve kterých došli ke zjištění několika významných vazeb. Například že dochází ke zvýšenému počtu požárů v oblastech, které se nacházejí do 2,5 km od řek a silnic.

Versichele (2014) nainstaloval celkem 29 skenerů Bluetooth v blízkosti památek v městě Gent (Belgie), kde sledoval návštěvnost po dobu 15 dní. Senzory vyhledávaly blízká zařízení a ukládaly si jejich MAC adresy. Cílem práce bylo zaznamenat pohyb turisty a které památky navštíví.

Sadat a kol. (2015) zkoumali alergické astma u dětí na základě místa bydliště. K dispozici měli bodovou vrstvu s místem bydliště 1 000 dětských pacientů v Teheránu, kteří byli právě s alergickým astmatem hospitalizováni. Ve vztahu k bydlišti analyzovali vzdálenost od nejbližší silnice a parku na vliv výskytu astma. Jako další podklad jim sloužila vrstva s koncentrací znečišťujících látek ve vzduchu. Numerická data kategorizovali pomocí indexu AQI (Air quality index). Výsledné indikátory látky byly: velmi vysoké, vysoké, střední a nízké. Zároveň každá látka má jiné limitní hodnoty, proto provedli normalizaci daného indexu. Z výsledných vygenerovaných asociačních pravidel vybrali pouze ty, které se vázaly na výskyt astma. Výsledkem publikace je mapa rizik, ve kterých lze vidět místa náchylná na projevení astmatu u dětí. Jedno z vyplývajících doporučení je vyhnout se bydlení v blízkosti parků a znečištěných oblastí.



Obr. 4 Riziková mapa alergického astmatu u dětí v Teheránu *(zdroj: Sadat a kol., 2015)*

Faridi (2018) využil tří datových sad z oblasti Indie: využití území, podzemní vody, pustiny a půdní typy. Zjistil, že většina pustin jsou krajiny bez křovin a mají nemalý obsah podzemních vod pod sebou. Tyto oblasti jsou vhodné pro agrikulturu.

3.3.1 Prostorový vzor společného umístění

V pracích se často zmiňuje pojem prostorový vzor společného umístění (*spatial co-location* pattern). Nehri a kol. (2014) uvádí, že prostorový vzor se od prostorových asociačních pravidel liší technickým provedením. Tento pojem můžeme definovat jako set prostorových prvků, které jsou často pozorovány v těsné blízkosti. Sledováním této proměnné lze objevit fakta, která jsou důležitá pro rozmanitá odvětví. Ať se jedná o ekologii, veřejnou bezpečnost, dopravu nebo byznys. Co se týče byznysu, tak pro lokální obchod bude podstatná znalost svého okolí, jaké druhy obchodů, resp. konkurence se zde nacházejí.

Yue a kol. (2017) ve své práci studovali region v Číně, kde se zaměřili na využití území a na lokalizaci spáchaných zločinů pomocí výpočtu lokačního kvocientu.

Sypion-Dutkowska a kol. (2017) se také zabývali vlivem využití území na počet zločinů, konkrétně v Polsku ve Štětíně. K analýze využili nástroj *Multiple Ring Buffer* v ArcGIS for Desktop, kdy vytvořili obalové zóny kolem zájmových bodů dle předem stanovených vzdáleností. Výsledkem byla tabulka, ze které lze vyčíst, že nejvíce zločinů se spáchalo do 50 m od barů, obchodů s alkoholem, klubů a diskoték.

Hu (2018) studoval nebezpečné křižovatky pro cyklisty a chodce. Z datových sad využil data o dopravních nehodách, silniční síť a lokaci semaforů a dopravních STOP značení. Křižovatky rozdělil dle typu signalizace (semafory, STOP značení a bez značení). Výsledkem článku je zjištění, že nejvážnější nehody se staly na křižovatkách se semafory.

3.3.2 Vstupní data

Pro asociační pravidla je nutné mít jako vstupní data tabulku s potřebnými atributy. Aby bylo možné pravidla generovat, je potřeba mít v tabulce alespoň dva atributové sloupce. Ukázku lze vidět níže. Tabulka obsahuje informace o využití území a o typu zájmového bodu (*POI*).

Tab. 3 Ukázka vstupní tabulky

ID	využití území	POIs
1	průmysl	sklad
2	park	škola
3	park	hotel
4	rezidenční	hotel
5	rezidenční	obchod

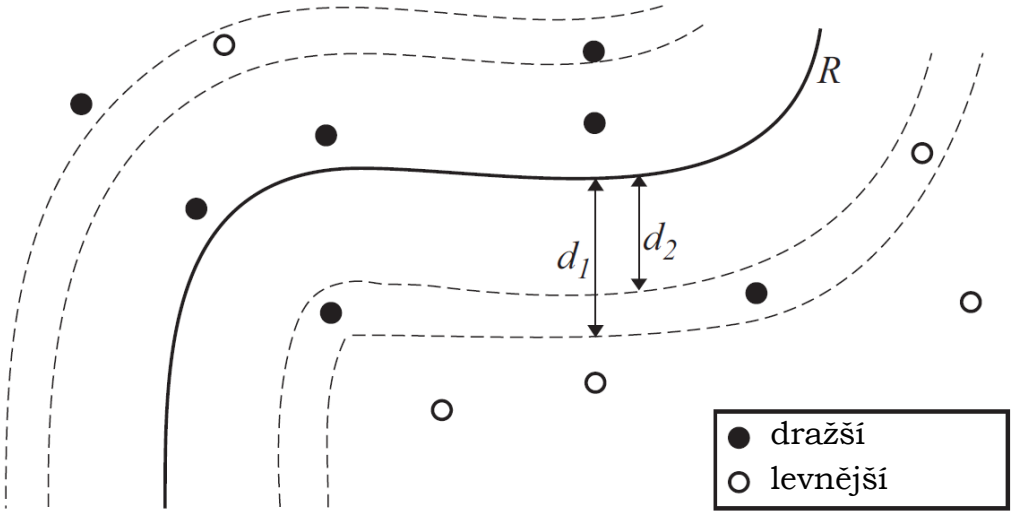
Samotný proces hledání asociačních pravidel si žádá, aby každý atribut obsahoval pouze binární hodnoty (0/1, ano/ne). Některé SW umějí zpracovat kategorická data, u jiného SW je nutné data do této podoby předpřipravit (např. u Weky). Proces převodu kategorických dat na data binární se nazývá *dichotomizace*. Během procesu se každá atributová hodnota stane novým atributovým sloupcem a pokud daný prvek splňuje

Tab. 4 Dichotomizace tabulky

ID	průmysl	park	rezidenční	sklad	škola	hotel	obchod
1	1	0	0	1	0	0	0
2	0	1	0	0	1	0	0
3	0	1	0	0	0	1	0
4	0	0	1	0	0	1	0
5	0	0	1	0	0	0	1

3.3.3 Úprava dat

Aby bylo možné prostorová data použít ke generování prostorových pravidel, je nutné je upravit v GIS (*Geografickém informačním systémovém*) prostředí. V rámci rešerše bylo ve většině prací pouze zmíněné, že důležitým krokem bylo předzpracování prostorových dat, ale již autor neuvedl, jakým způsobem. Ze všech studií lze vyzdvihnout práci Laube (2008), který pracoval s obalovou zónou. Obalová zóna rozdělí prvky na minimálně dvě kategorie. Na ty, co do ní spadají a zbytek, který leží mimo. Laube ukazuje příklad pravidel na ceně a poloze pozemku. Pokud je dům v blízkosti řeky, je dražší (*Obr. 5*). Vzdálenost od řeky je zde označena proměnnou *d*. Je nutné určit vhodnou hodnotu. Vzdálenost se určí na základě povahy prvků. Tato limitní hodnota může sloužit jako kategorizace dat – na prvky, které jsou blízko (splňují vytyčenou limitní vzdálenost) a na prvky, které jsou vzdálené (jdou nad limitní vzdálenost). Cílem praktické části této diplomové práce bude otestování vlivu hodnoty vzdálenosti obalové zóny na výsledný počet vygenerovaných pravidel.



Obr. 5 Cena domů na základě vzdálenosti *(d1, d2)* od řeky *(R)*
(Upraveno podle Laube, 2008)

V rámci generování asociačních pravidel je vhodné použít kategorická data. Numerická data lze upravit tak, že se rozdělí do kategorií. Je důležité zvolit adekvátní počet kategorií. Při nízkém počtu je možnost ztráty důležitých informací, naopak při velkém počtu kategorií dochází k náročnosti generování pravidel. Lee (2014) ve své práci uvádí způsoby, jak data takto upravit. Surová data jsou v prvním kroku agregována na plochy (v tomto případě se jedná o administrativní členění). Data určená k agregaci mohou být jak bodová, tak liniová, a dokonce i polygonová. U polygonové vrstvy dochází k agregaci na vyšší územní celky. Nejdůležitějším nástrojem v GIS prostředí pro jeho práci je operace průniku (*intersect*) vrstev. Takto upravená data jsou následně kategorizována. V jeho případě bylo použito 5 intervalů, resp. kategorií.

3.4 Software využitelný pro generování pravidel

Cílem práce je zaměřit se především na SW, který je otevřený (*open source*), ale je důležité zmínit i jeho komerční alternativy. Níže jsou zmíněny takové nástroje, které byly využity v odborných článcích z různého časového období. Výsledkem této části rešerše je mimo jiné zjištění, že ne všechny nástroje jsou stále plně funkční, resp. jejich vývoj byl již ukončen.

3.4.1 Komerční SW

Oracle Data Miner

Aggarwal (2012) jej definuje jako balíček nástrojů běžící v databázovém prostředí Oracle. Díky uložení dat v databázi je umožněné generovat data automaticky. Jedná se jednoduché grafické prostředí, ve kterém průvodce radí krok po kroku od přípravy dat až po samotné vyhodnocení dat. Právě tento nástroj využila Hůlová (2010) ve své diplomové práci, ve které zpracovávala asociační pravidla pro trestní činnost v jednotlivých krajích České republiky. Pracuje pouze s Apriori algoritmem.

SPSS Statistics

Komerční SW vyvíjený firmou IBM, který nabízí pokročilé statistické analýzy. Jednou z analýz je i generování asociačních pravidel pro prostorová data pomocí vybraných algoritmů. Pro použití prostorové analýzy je potřeba modulu IBM SPSS Association. Nástroj nabízí 14denní trial verzi, ve které je možno si vše vyzkoušet. Zároveň se na platformě YouTube nachází podrobné video[2] s ukázkou na kriminálních datech, která lokalizují trestné činy dle konkrétních kategorií. Zároveň do analýzy zahrnuje demografické údaje jako počet obyvatel, hustota obyvatel, počet domácností atd. Výsledkem jsou statistické tabulky, ale také zároveň mapy vizualizující jednotlivá pravidla. Dle barevné symbologie odlišuje bodové prvky splňující pravidlo a body, které pravidlo nesplňují. Zároveň zvýrazní administrativní prvky, které obsahují pouze prvky splňující pravidlo.

3.4.2 Open-source SW

Open-source řešení nabízí velké množství samostatných SW, které umožňují tvorbu asociačních pravidel. Open-source znamená, že zdrojový kód daného nástroje je otevřený. Zde je výčet nástrojů využívaných v odborných článcích.

EasyMiner/R

Projekt vytvořený na VŠE v Praze. Jedná se o open-source nástroj běžící ve webovém prostředí. Využívá Apriori algoritmus a balíček *arules* z R. prostorová asociační pravidla sám o sobě neumí. Data do něj musí být příslušně upravena. Nejnovější verze je 2.6, vydána 15. 7. 2019. V roce 2017 byl nástroj využit v evropském projektu OpenBudgets, ve kterém se analyzovala finanční data.

EasySDM

SW, který je popsán v odborném článku od Hamdad a kol. (2015). Jedná se o nástroj vytvořený přímo pro práci s prostorovými daty. Jedná se o desktop aplikaci, ve které se po přihlášení lze napojit do konkrétní databáze. K jeho správné funkčnosti je potřeba mít v počítači nainstalován PostgreSQL zároveň s prostorovou extenzí PostGIS. Z algoritmů využívá již několikrát zmiňovaný Apriori. V aktuální době je další vývoj nástroje ukončen.

KEEL

KEEL (Knowledge Extraction based on Evolutionary Learning) představuje jednoduché grafické prostředí, které je naprogramováno v Javě. Jedná se o další z nástrojů, který nabízí funkce DM: regrese, klasifikace, shlukování, hledání vzorů atd. Vše pro generování neprostorových asociačních pravidel. Poslední vydaná verze dohledatelná v repositáři GitHub[3] je z roku 2015, dále se vývoj nástroje zastavil.

Orange

Orange je SW vhodný pro úplné začátečníky, ale i pro experty, co si píší vlastní nové skripty v Pythonu. Pro nováčky je atraktivní jednoduché moderní vizuální prostředí. Nástroj obsahuje několik balíčků, které lze doinstalovat. Jedním z nich je právě *Associate*, který obsahuje funkci pro hledání asociačních pravidel.

---

[2] https://www.youtube.com/watch?v=lRU9PKxTd8o

[3] https://github.com/SCI2SUGR/KEEL

Vstupní data mohou být ve formátu excelovské tabulky (.XLS – *Excel Spreadsheet)*, CSV tabulky nebo dokonce SQL *(Structured Query Language)* tabulky. Podle oficiální dokumentace Orange 3[4] nástroj využívá algoritmus FP-Growth.

Výsledná vygenerovaná tabulka zobrazuje následující hodnoty:

- Supp – podpora
- Conf – spolehlivost
- Covr *(coverage)* – pokrytí (jak často se nacházejí v datové sadě objekty splňující předpoklad)
- Strg *(strength)* – síla
- Lift – jak často je pravidlo platné pro závěr
- Levr *(leverage)* – česky páka, rozdíl mezi dvěma položkami, které se objevují v transakci a dvěma položkami, které jsou samostatně

RStudio

Pro generování asociačních pravidel je nutné doinstalovat knihovnu *arules*, která obsahuje Apriori algoritmus. Tuto knihovnu využil pro svou diplomovou práci Bc. Tomáš Matonoha (2014), který analyzovat data ze studentských dotazníků.

Weka

Rangra (2014) ve svém článku uvádí, že pro generování asociačních pravidel je nejvhodnější právě Weka. Jedná se o nástroj napsaný v Javě, který obsahuje několik algoritmů a nástrojů pro analýzu a vizualizaci dat. Kromě asociačních pravidel umožňuje klasifikaci dat. V rámci práce s asociačními pravidly je nutné data předzpracovat. Používá algoritmus Apriori.

Vybraný SW bude vstupovat do praktické části, kde dojde k hledání asociačních pravidel. Po úspěšném výběru jednoho z nich budou vypracovány celkem tři případové studie. Samozřejmostí bude sepsání podrobného návodu, jak v daném SW pracovat a jak asociační pravidla vygenerovat.

[4] https://orange3-associate.readthedocs.io/en/latest/widgets/associationrules.html

4 TECHNICKÉ ŘEŠENÍ

Jak již bylo v předchozí kapitole zmíněno, existuje několik softwarů, ve kterých lze asociační pravidla vygenerovat. Žádný SW nebyl schopen generovat pravidla přímo z prostorových dat. Dostupné jsou SW, které generují pravidla pouze z neprostorových dat. Pro generování asociačních pravidel je důležité najít spolehlivý způsob, jak převést prostorová data na neprostorová a následně použít software pro generování asociačních. Cílem této dílčí praktické části je otestování těchto nástrojů. Testování proběhne tak, že budou vytvořena testovací data, která jsou velmi jednoduchá a asociační pravidla jsou na první pohled patrná. Jedině tak se ověří spolehlivost daného softwaru. Kromě spolehlivosti se také ověří plynulost nástroje, jeho náročnost na přípravu dat apod. Ze zmíněných SW v teoretické části byly vybrány pouze ty nástroje, které nadále fungují a jejich otestování bylo úspěšné. U všech nástrojů byly nastaveny hodnoty proměnných:

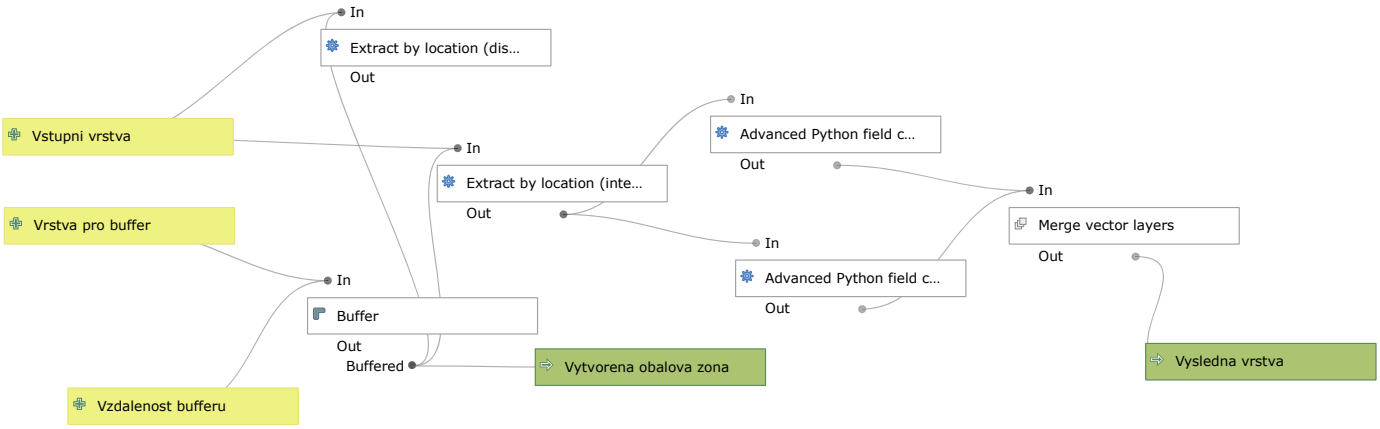
- Podpora *(support)*: 5 %
- Spolehlivost *(confidence)*: 90 %

4.1 Metody úpravy dat

Jak již bylo v předchozích kapitolách zmíněno, vybrané datové sady je nutno upravit. Vždy je tzv. primární datová sada, ke které se budou připojovat informace z dalších datových sad. Během výzkumu bylo nalezeno několik postupů, díky kterým lze prostorové datové sady použít pro generování asociačních pravidel. Metody a jejich nástroje byly ozkoušeny jednotlivě a následně byly použity pro vytvoření modelu k automatizaci práce. Jak již bylo dříve zmíněno, prostorová data je potřeba rozdělit do kategorií. Jednou z možností je tvorba obalové zóny kolem vybraného prvku. Následně se prvky rozdělí na ty, co do dané zóny spadají, a na ty co nespadají. Další možností je přiřazení hodnot na základě polohy prvku.

4.1.1 Vzdálenostní obalová zóna

Součástí práce bylo vytvoření modelu, který vytvoří obalovou zónu kolem vybrané vrstvy. Uživatel může zadat libovolnou vzdálenost v rámci dialogového okna nástroje *Buffer*. Dalším krokem v modelu je zjištění, které prvky z druhé vrstvy do této obalové zóny spadají. K tomu je využito nástroje *Extract by location.* Výsledkem modelu je nová vrstva, ve které je vytvořený nový atribut, který nabývá hodnot 0 nebo 1. Pro docílení těchto hodnot bylo využito nástroje *Advanced Python Field Calculator*. Hodnotu 1 nabývá prvek tehdy, zdali spadá do dané obalové zóny. Takovýto model je sice jednoduchý, ale přesto efektivní, navíc jeho část je použitelná do případových studií.



Obr. 6 Model vytvořený v prostředí QGIS *(zdroj: autorka)*

4.1.2 Přiřazení hodnot

Další možností je použít datovou sadu s informacemi v atributových sloupcích, které budou přiřazeny zkoumaným prvkům. Příkladem může být využití území ve formě polygonové vrstvy. Každý zkoumaný prvek (bod, linie nebo polygon) získá hodnotu podle své polohy. Například pokud se banka nachází v rezidenční části, tak v příslušném atributu bude tato informace zapsána.

Nástroj se v QGIS nazývá *Join Attributes by Location*, který lze česky přeložit jako *Připojit atributy podle umístění*. Jeho cílem je přesně jeho název – spojí atributy dvou vrstev na základě umístění. U nástroje lze vybrat geometrický predikát – zda se vrstvy mají protínat, dotýkat, překrývat, rovnat se apod. Pro tyto účely bylo zvoleno protínání (*intersects*).

4.1.3 Kategorizace dat

Možnost, jak zařadit data do předem určených intervalů je nespočet. Jako příklad lze uvést sklon povrchu. Pokud by se generovala asociační pravidla čistě pro atributovou tabulku s údajem o sklonu, tak by prakticky nevyšla řádná spolehlivá pravidla. Lepším způsobem je tato numerická data rozdělit do kategorií dle stupně sklonu (např. rovina do 2°, naopak nad 55° se jedná o stěny, …). Takto lze rozdělit i numerická data s počtem obyvatel apod.

Druhý typem kategorizace dat je vytvoření obalové zóny, u které se bude hledat, zda do ní daný prostorový prvek spadá či ne – hodnoty ano/ne (0/1). Tento postup bude využit u tvorby testovacích jednoduchých dat. Praktickým příkladem může být:

Je v okolí nabíjecí stanice čerpací stanice? → ano/ne.

4.1.4 Úprava dat v komerčním SW

Tentýž postup úpravy dat lze provést ve známé komerční alternativě ke QGIS – ArcGIS. Obě varianty (ArcGIS for Desktop, ArcGIS Pro) obsahují velmi podobnou logiku nástrojů, pouze se mohou lišit názvem. Například nástroj pro připojení atributů dle umístění se zde nazývá *Spatial Join*. Pro tvorbu obalové zóny se zde nabízí nástroj se stejným názvem – *Buffer*. Funkcionalita nástrojů je velmi obdobná, proto se postup nebude více řešit.

4.2 Testování SW

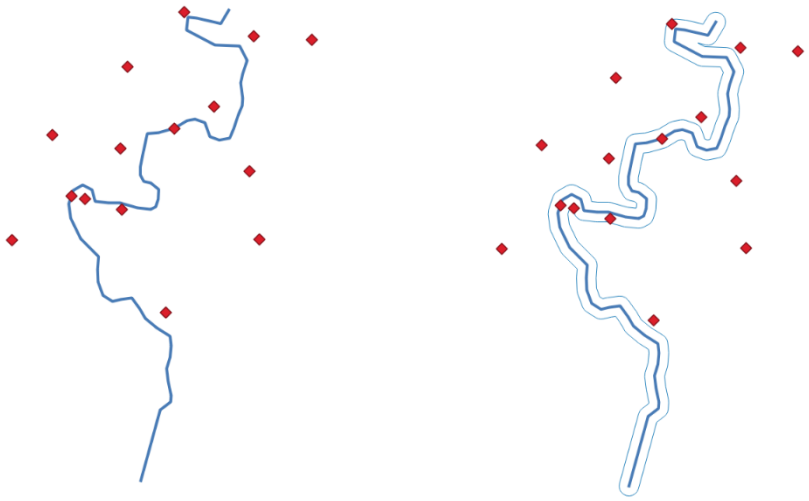
V rámci práce byly otestovány následující nástroje, které jsou shrnuty v tabulce níže. Pro otestování a seznámení se s jednotlivými nástroji byla vytvořená velmi jednoduchá datová sada, která slouží k otestování funkcionality.

Tab. 5 Srovnání testovaných SW

Nástroj	Rok vydání	Stabilní verze	Licence	Operační systém	Programovací jazyk
EasyMiner	2015	2.6 (15. 7. 2019)	Apache License verze 2.0	Multiplatformní	C, R
Orange	2009	3.22.0 (26. 6. 2019)	GNU	Multiplatformní	Python, C++, C, Cython
Rstudio	2011	1.2.5001 (29. 9. 2019)	AGPL v3	Multiplatformní	Java, C++, JavaScript
Weka	1993	3.8.3 (4. 9. 2018)	GNU	Multiplatformní	Java

4.2.1 Testovací data

Testovací sada byla vytvořena tak, aby byla snadná k interpretaci. Data byla vytvořena v prostředí QGIS a jedná se o dvě nově vytvořené a smyšlené vektorové vrstvy. Jednou z vrstev je linie řeky, druhá je bodová vrstva s rostlinami. Rostliny mají atribut, zda je rostlina suchá a nabývá hodnot 0 (ne) nebo 1 (ano). Kolem vytvořené řeky se následně modelem vytvoří obalová zóna o zadané vzdálenosti 25 m. Pokud rostlina leží v zadané oblasti, atribut *buffer* nabývá hodnoty 1, v opačném případě 0. Data byla vytvořena tak, aby rostliny, které spadají do této vzdálenosti, nebudou trpět suchem, jelikož jsou do limitní vzdálenosti od vody. Tato hodnota je zcela smyšlená a slouží pouze k potvrzení faktu, který je na první pohled zřetelný.



Obr. 7 Testovací sada, vytvořená obalová zóna kolem řeky *(zdroj: autorka)*

V tabulce níže lze vidět finální tabulku, která byla využita v rámci testování SW. Obsahuje celkem 15 záznamů, pro které jsou důležité dva atributové sloupce.

Tab. 6 Atributy určené pro generování asociačních pravidel

ID	sucha	buffer
1	0	1
2	0	1
3	1	0
4	0	1
5	1	0
6	1	0
7	1	0
8	1	0
9	1	0
10	1	0
11	1	0
12	1	0
13	1	0
14	0	1
15	0	1

4.2.2 EasyMiner/R

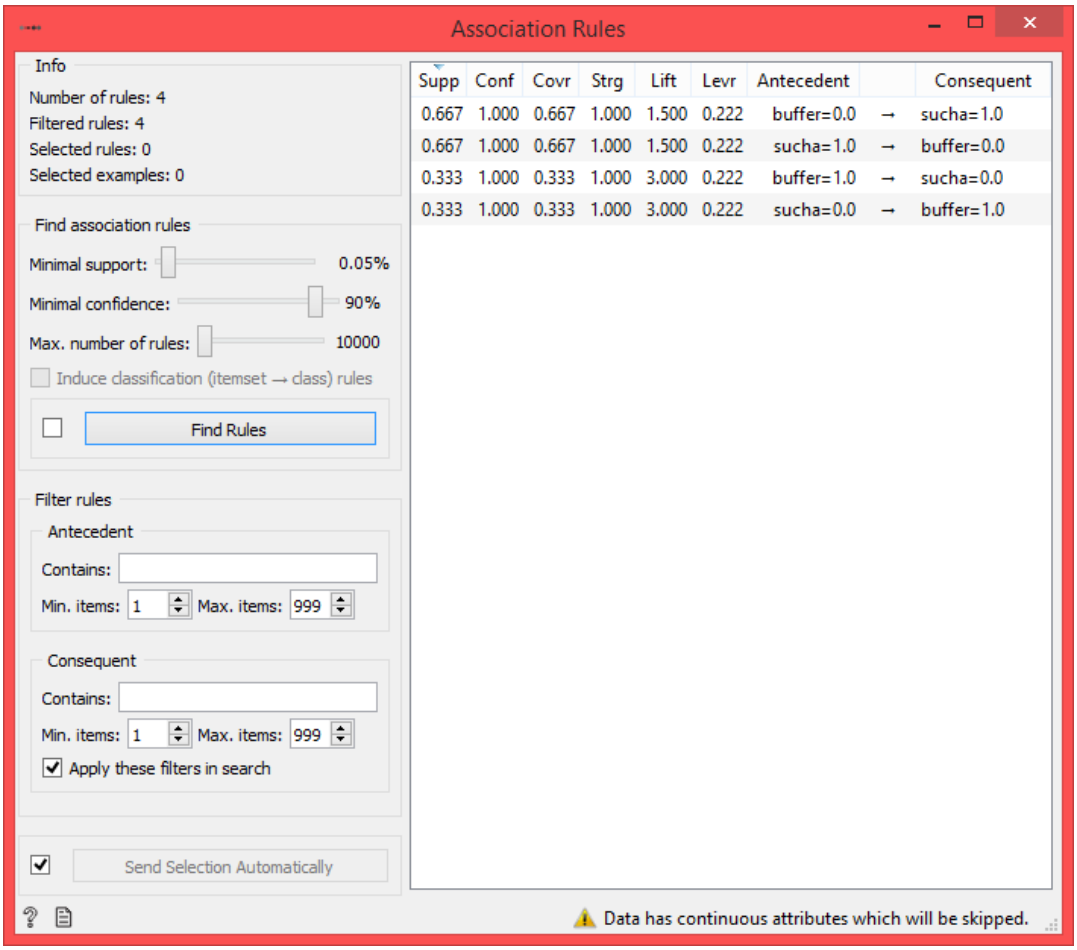
Testovaným nástrojem byl EasyMiner/R, který byl spuštěn v prohlížeči. Postup práce byl díky oficiálnímu návodu poměrně jednoduchý. Aplikace běží v cloudovém prostředí a je nutné se přihlásit do jednoho z vybraných účtů (nový účet, Facebook nebo Google účet). Po přihlášení lze nahrát soubor pouze ve formátu CSV. Pro úspěšné nahrání je potřeba datovou sadu upravit, aby si ji nástroj přečetl správně. Následně lze zkombinovat atributy tak, aby výsledná asociační pravidla byla vygenerována na základě daného výběru. Při snaze nastavit zvolené proměnné byl problém v nastavení hodnoty podpory na 5 %, která je mimo povolený interval hodnot. Nejnižší hodnota je 6,7 %. Při testovací sadě byly výsledky zmatené. Výsledné pravidlo totiž říká, že pokud rostlina leží ve zvolené oblasti (buffer = 1), tak je suchá (sucha = 1). Při snaze nahrát CSV soubor reálné datové sady vůbec nic neproběhlo. Pokus nahrát ten samý soubor v ZIP formátu také nepomohl.

Výhodou EasyMiner/R je jeho neustálý vývoj a práce v cloudovém prostředí. Není potřeba si nic stahovat ani instalovat do svého zařízení. Součástí webové stránky je také podrobný návod. Nevýhodou je jeho nespolehlivost, kdy nelze pracovat s určitým CSV souborem a nutnost vytvoření si vlastního účtu. Dále nelze nastavit minimální podporu na menší hodnotu, než je 6,7 %. V případě testování velmi málo frekventovaných pravidel je tedy tento nástroj nevhodný.

4.2.3 Orange

Práce v Orange je velmi jednoduchá. Nástroj je uživatelsky přívětivý, pracuje se v jednoduchém grafickém prostředí. Pracovní okno se skládá z nástrojů, které se nacházejí v levém panelu a hlavní plochy, ve které se tvoří diagram z vybraných nástrojů. Pro generování asociačních pravidel stačí do této plochy přetáhnout nástroj *File* (česky soubor) a nástroj *Association Rules*, který je součástí extenze, kterou je nutné doinstalovat. Poklikáním na daný nástroj se otevře dialogové okno. U souboru stačí zvolit soubor, který je potřeba načíst.

Výhodou Orange je, že není potřeba datovou sadu upravovat, bez ohledu na daný formát. Nástroj umí pracovat jak s .XLS, tak s CSV tabulkou. Atributy, které jsou pro generování pravidel nevhodné, lze pomocí atributu role označit k přeskočení (*skip*), algoritmus je neobsáhne do analýzy. Po nahrání souboru se pokliká na asociační pravidla a otevře se interaktivní okno viz *obr. 8*.



Obr. 8 Dialogové okno v prostředí Orange *(zdroj: autorka)*

V něm lze měnit hodnoty podpory a spolehlivosti a také lze měnit maximální počet vygenerovaných pravidel. Zároveň lze filtrovat pravidla pomocí předpokladu a závěru – lze nastavit konkrétní hodnotu, kterou má předpoklad, resp. závěr obsahovat. Dále je také možné limitovat počet objektů, které budou obsaženy v předpokladu/závěru. Po vybrání vhodných hodnot parametrů lze výsledna vygenerovaná pravidla exportovat do reportu. Výsledný soubor je HTML (*Hypertext Markup Language*) stránka.

Association Rules

Number of rules: 4
Selected rules: 0
Covered examples: 0

Rules

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.667	1.000	0.667	1.000	1.500	0.222	buffer=0.0	→	sucha=1.0
0.667	1.000	0.667	1.000	1.500	0.222	sucha=1.0	→	buffer=0.0
0.333	1.000	0.333	1.000	3.000	0.222	buffer=1.0	→	sucha=0.0
0.333	1.000	0.333	1.000	3.000	0.222	sucha=0.0	→	buffer=1.0

Obr. 9 Exportovaný report pravidel pro testovací data *(zdroj: autorka)*

Velkou výhodou Orange je, že data není potřeba upravovat. Pouze je nutné si zkontrolovat, které atributy jsou vhodné ke generování pravidel, popřípadě jim změnit jejich roli či datový typ. Orange je velmi graficky přívětivý, pro uživatele laika je rozhraní velmi jednoduché a na webu SW je dostupná přehledná dokumentace. SW je neustále vyvíjen a je aktualizován. Další výhodou je interaktivní dialogové okno, kde lze velmi intuitivně měnit minimální hodnoty podpory a spolehlivosti a takřka v reálném čase vidět výsledek. Zároveň je zde možnost filtrovat předpoklad či závěr na základě slovního spojení. Lze také pevně nastavit, kolik položek bude použito pro předpoklad/závěr. Předností je schopnost zpracovat datové sady o rozsáhlém počtu záznamů bez ohledu na formát – zda se jedná o CSV nebo XLS tabulku. Jedinou možnou nevýhodou je absence české lokalizace, která ale není příliš častá pro nástroje umožňující generovat asociační pravidla.

4.2.4 RStudio

Jak již bylo v teoretické části zmíněno, pro použití asociačních pravidel je nutné doinstalovat knihovnu *arules*. V RStudiu se pracuje v příkazech. Jako první je potřeba vytvořit novou proměnnou, v tomto případě *rules*, pro kterou se nadefinuje použití algoritmu Apriori a také hodnoty podpory a spolehlivosti. Dále stačí pouze napsat příkaz s funkcí *inspect*, která vypíše vygenerovaná asociační pravidla. RStudio vygenerovalo celkem 4 pravidla, která lze níže vidět:

```
rules <- Apriori(rostliny,parameter = list(supp = 0.05, conf = 0.9, tar
get = "rules"))
```

```
> inspect(rules)
    lhs                rhs             support   confidence lift count
[1] {sucha=FALSE}   => {buffer=TRUE}  0.3333333 1          3.0   5
[2] {buffer=TRUE}   => {sucha=FALSE}  0.3333333 1          3.0   5
[3] {sucha=TRUE}    => {buffer=FALSE} 0.6666667 1          1.5  10
[4] {buffer=FALSE}  => {sucha=TRUE}   0.6666667 1          1.5  10
```

Vygenerovaná pravidla se nikterak neliší od pravidel vygenerovaných v SW Orange. Předpoklad je zde označený sloupcem s názvem *lhs*, závěr sloupcem *rhs*. Další sloupec představuje podporu pravidla a jeho spolehlivost. Stejně jako u Orange je další sloupec hodnota *lift*. Předností použití RStudia je práce v příkazech, což může být v určitých situacích výhodnější. Zároveň toto může být nevýhodou, pokud uživatel upřednostňuje graficky přívětivější prostředí, ve kterém všechny nástroje lze vidět.

4.2.5 Weka

Před generováním pravidel je nutné datovou sadu upravit pomocí série nástrojů, které Weka poskytuje. Prvním krokem je úprava dat na data dichotomická. To se zajistí pomocí filtru. V záložce se zvolí *Filter-Unsupervised-Attribute-NominalToBinary*. Pokud se v datové sadě nachází velké množství nul, je potřeba takovou sadu opravit. Bez úpravy dat by výsledkem bylo nespočet negativních asociačních pravidel. Opět tedy v záložce *Filter-Unsupervised-Attribute-NumericCleaner*. V tomto kroku se přepíší hodnoty 0 na NaN. Následně se data převedou na nominální hodnoty. *Filter-Unsupervised-Attribute-NumericToNominal*. K samotným asociačním pravidlům se lze dostat v záložce *Associate*. V dialogovém oknu se nastaví zmíněné hodnoty podpory a spolehlivosti. Bohužel na testovací datové sadě s danými minimálními hodnotami podpory a spolehlivosti Weka nebyla schopná vygenerovat žádná asociační pravidla.

Během testování potenciálních datových sad bylo zjištěno, že Weka odmítá načíst datový soubor s 20 000 záznamy ve formátu XLS. Weka vrátí chybovou hlášku *„Problem setting base instances: java.lang.reflect.InvocationTargetException“*.

Nevýhodou Weky může být nutnost předúpravy dat. Vzhled nástroje je staršího rázu, oproti Orange se může jevit zastaralejší. Dalším problémem bylo, že u identických testovacích dat nevygeneroval SW ani jedno pravidlo. Pokud je testovací sada obsáhlejší, co se záznamů týče, je nutné mít data ve formátu CSV.

4.2.6 Zhodnocení a výběr jednoho SW

Weka patří k lépe řešeným technickým řešením, avšak velkou nevýhodou je nutnost úpravy dat před samotným generováním pravidel. Tato úprava lze případně zautomatizovat pomocí *Workbench*, samotné prostředí není ale příliš intuitivní. Samotný postup není náročný, ale lepší přístup je u Orange, ve kterém stačí data nahrát v souboru XLS. Výhodou je intuitivní uživatelské prostředí, které je velmi jednoduché. RStudio je také vhodným kandidátem pro testování, avšak subjektivně bylo v rámci této práce upřednostněno grafického prostředí ku psaní příkazů.

4.3 Hledání vhodných datových sad

V kapitole došlo k otestování různých technických řešení, ze kterých bylo následně vybráno jedno – Orange, které se jeví jako nejrozumnější řešení. Nezbytným krokem pro případové studie je nalezení takových datových sad, které budou vykazovat asociační pravidla. Obrovskou část praktické části představuje právě hledání dat. Datové sady se mohou na první pohled zdát jako vhodné pro analýzu, avšak až po samotném otestování se dojde k závěru, že žádná častější asociační pravidla neobsahují. Tímto se rozumí, že z celé datové sady tuto podmínku splňoval pouze jeden prvek. Z datových sad byly otestovány například *Brno Urban Grid*, který obsahuje data v čtvercové mřížce pro Brno a jeho přilehlé okolí. Dále byly testovány *Kreativní průmysly*, které byly dodány doktorem Nětkem. Jedná se o bodovou vrstvu pro Olomouc, kde jsou lokalizované kreativní průmysly.

5 PŘÍPADOVÁ STUDIE 1 – REKOLA

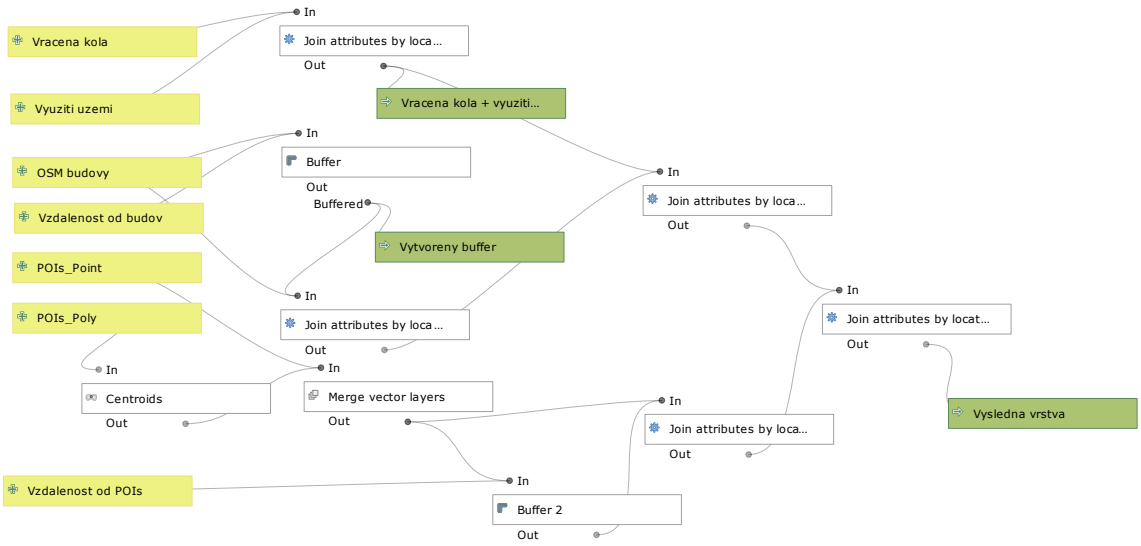
Primární vrstvou byla zvolená vrstva *Vrácená* kola. Jedná se o bodovou datovou vrstvu poskytnutou firmou Rekola pro účely diplomové práce Filipa Hrice s názvem *Analýza využívání komunitních výpůjček jízdních kol* (2018). Vrstva obsahuje záznamy o výpůjčkách jízdních kol v Olomouci za rok 2016. Pro účely této práce byly vybrány pouze záznamy týkající se vrácení jízdních kol na konkrétní místo. Záznamů o vrácení kol je celkem 3 162. Samotná vrstva neobsahuje žádný atributový sloupec, jehož informace by bylo možné pro generování asociačních pravidel použít. Jako doplňková vrstva bylo vybráno *využití území* od *Urban Atlas*. Mimo jiné byla také využita polygonová vrstva budov z *OSM*, u které byl důležitý typ budovy. Tato data byla získána stažením nejnovější datové sady OSM ze stránek Geofabrik. Jako poslední vrstva byly využity zájmové body OSM, které se zde nacházejí jak v bodové, tak i polygonové podobě.

Do dalšího postupu vstupují vrstvy (+ vybraný atribut):

- Primární vrstva
    - Vrácená Rekola (bodová vrstva) – žádný atribut
- Doplňkové vrstvy
    - Využití území od Urban Atlas (polygonová vrstva) – atribut *landuse*
    - Zájmové body OSM (bodová a polygonová vrstva) – atribut *POIS*
    - Typy budov OSM (polygonová vrstva) – atribut *Budova*

5.1 Postup úpravy dat

Prvním krokem bylo získání využití území pro každý bod s údajem o vrácení kola. K tomu bylo použito nástroje *Join attributes by location*. Jedná se o nástroj, který připojí atributy na základě umístění. Dalším nástrojem bylo využití obalové zóny. Vzhledem k tomu, že se kola nemusí vracet v bezprostřední blízkosti budovy, byla nastavená obalová vrstva 100 m od všech budov. Aby obalová zóna získala údaje o budově, od které je tato zóna tvořena, byl znovu použit nástroj *Join attributes by location*. Stejný postup byl aplikován na zájmové body. Vzhledem k tomu, že z OSM vrstvy jsou zájmové body jak bodové, tak i polygonové, bylo využito nástroje *Centroids*. Díky němu jsou polygony převedeny na body a následně jsou *zájmové body* sloučeny. Posledním krokem je znovu použití *Join attributes by location*, aby se spojily informace z obalové zóny a bodové vrstvy vrácených kol. Byla ověřena funkčnost tohoto postupu a následně byl přepracován do přehledného modelu, který je níže zaznačen na obr. 10.



Obr. 10 Vytvořený model pro případovou studii Rekola *(zdroj: autorka)*

Všechny vstupní informace (vrstvy, parametry) jsou žluté obdélníky. Jedná se o datové sady a o parametr, kterým je vzdálenost obalové zóny. Uživatel si může nastavit libovolnou hodnotu a v rámci této práce bude dále také otestován vliv velikosti obalové zóny na celkový počet vygenerovaných pravidel. Zelené obdélníky reprezentují výstupy mezikroků a výslednou vrstvu. Bíle jsou zaznačeny všechny použité nástroje, které jsou zmíněny v postupu výše. Model je velmi jednoduchý a splňuje to, co je potřeba.

Výsledkem tohoto postupu je jedna vektorová vrstva, která obsahuje všechny informace získané ze vstupních datových sad. Mezi vrstvami dochází k mnohonásobnému protnutí. Protnutí lze zde rozumět tak, že v rámci obalové zóny leží více budov/zájmových bodů než jen pouze jeden prvek. QGIS dotyky zapisuje pouze do dvou atributových sloupců. Vrstva vrácených kol obsahuje nové atributové sloupce s názvem *Budova, Budova_2, POIs, POIs_2* a *landuse*. V tab. 7 lze vidět ukázku záznamů u výsledné vrstvy.

Tab. 7 Ukázka záznamů z výsledné vrstvy pro Rekola

Landuse	Budova	Budova_2	POIS	POIS_2
Sports and leisure facilities	civic	dormitory	waste_basket	university
Continuous urban fabric (S.L. : > 80%)	civic	civic	university	library
Continuous urban fabric (S.L. : > 80%)	civic	dormitory	university	courthouse
Other roads and associated land	garage	residential	tourist_info	park
Railways and associated land	transportation	civic	tourist_info	fountain
Sports and leisure facilities	civic	residential	swimming_pool	swimming_pool
Industrial, commercial, public, military and private units	residential		stadium	stadium
Continuous urban fabric (S.L. : > 80%)	civic	civic	school	college
Industrial, commercial, public, military and private units	residential	residential	pub	mall

5.2 Asociační pravidla

Po aplikaci modelu je výsledná vrstva exportována jako CSV tabulka do SW Orange. Pravidla jsou seřazená sestupně od nejvíce frekventovaných. Výsledná pravidla lze vidět v tab. 8. Pravidla jsou seřazená sestupně podle hodnoty podpory a bylo vybráno pouze prvních 8 pravidel. Je patrné, že žádné z pravidel nelze označit jako velmi frekventované, ani jedno nedosahuje marginální hodnoty. Co se ale týče spolehlivosti, tak například třetí pravidlo lze označit jako spolehlivou výjimku. Jeho podpora je nízká – jedná se o výjimku, ale zároveň má 100 % spolehlivost → spolehlivá výjimka. Pro samotné pravidlo lze konstatovat, že ve 100 % případů vždy platí jak předpoklad, tak i závěr. Jelikož se jedná o prostorová data, je vhodné se vrátit zpět do prostoru.

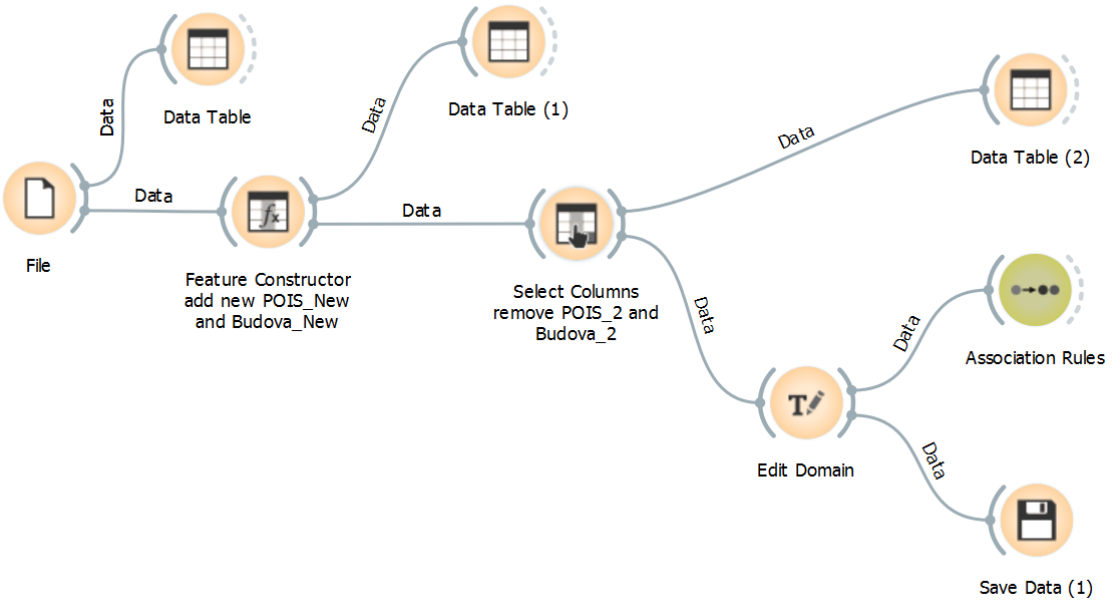
5.2.1 Duplicitní hodnoty

Během prohlížení výsledných pravidel bylo zjištěno, že pokud se v okolí vráceného kola nachází pouze jeden zájmový bod, tak je jeho hodnota zapsána do obou atributových sloupců. Výsledkem jsou pravidla, které mají stejnou hodnotu jako předpoklad, tak i závěr.

Tab. 8 Vygenerovaná asociační pravidla pro Rekola

Podpora	Spolehlivost	Předpoklad	→	Závěr
0.083	0.825	POIS=university	→	POIS_2=university
0.076	0.988	POIS=fountain	→	POIS_2=fountain
0.049	1.000	POIS=fountain, Budova_2=church	→	POIS_2=fountain
0.049	0.837	POIS_2=fountain, Budova_2=church	→	POIS=fountain
0.043	0.846	POIS=pitch	→	POIS_2=pitch
0.041	0.884	POIS_2=library, Budova_2=civic	→	landuse=Continuous urban fabric (S.L. : > 80%)
0.041	0.812	POIS_2=library, landuse=Continuous urban fabric (S.L. : > 80%)	→	Budova_2=civic
0.039	0.891	POIS=university, Budova=civic	→	POIS_2=university

Pro další práci je nutné tento problém vyřešit a tyto záznamy upravit. K tomu poslouží SW Orange pro generování asociačních pravidel a to tak, že se zde vytvoří model, který lze vidět na obr. 11. Do tohoto modelu vstupuje CSV tabulka, která byla vyexportována z QGIS. Následně se do vrstvy přidá nový sloupec *POISNew*, do kterého se vepíše hodnota *POIS_2* pod podmínkou, že se nejedná o duplikující hodnotu ze sloupce *POIS*. Dále se původní sloupec *POIS_2* smaže a nahradí se nově vytvořeným. Stejný postup je aplikován i na atributové sloupce *Budova* a *Budova_2*. V celém modelu je možné si zobrazit aktuální podobu tabulky pomocí nástroje *Data Table*. Výsledek modelu se automaticky uloží a je možné si pro takto upravená data generovat asociační pravidla.



Obr. 11 Orange model pro úpravu duplicitních hodnot *(zdroj: autorka)*

V tab. 9 lze vidět upravená asociační pravidla, která již neobsahuje duplicitní hodnoty a přinášejí mnohem zajímavější výsledky. Byla vybrána postupně ta pravidla, jejichž předpoklad/závěr neobsahoval duplicitní kombinaci stejných hodnot.

Tab. 9 Asociační pravidla po úpravě duplicitních hodnot

Podpora	Spolehlivost	Předpoklad	→	Závěr
0.034	0.991	Budova=transportation	→	Budova_2=civic
0.032	0.935	Budova=transportation	→	POIS_2=fountain
0.032	0.833	POIS_2=fountain, Budova_2=civic	→	Budova=transportation
0.032	0.943	POIS=post_box	→	Budova=residential
0.031	0.858	POIS=tourist_info	→	Budova_2=civic
0.031	0.850	POIS=tourist_info	→	POIS_2=fountain
0.031	0.800	POIS_2=fountain, Budova_2=civic	→	POIS=tourist_info
0.029	0.814	POIS=tourist_info	→	Budova=transportation

5.3 Vizualizace asociačních pravidel

Výsledná vygenerovaná asociační pravidla jsou prostorová, je vhodné je zobrazit v prostoru. Proto se opět práce vrací do QGISu, ve kterém proběhne vizualizace. Nejjednodušší způsob je použít symbologii u konkrétní vrstvy. Jeden z druhů vizualizace je založený na pravidlech (*rule-based symbology*). Každé pravidlo bude jednou podmínkou.

5.3.1 Převod pravidel z Orange do QGIS

Pravidla je potřeba upravit tak, aby každá hodnota atributu byla uvozena jednoduchými uvozovkami. Současně mezi každou podmínkou musí být vepsán logický operátor „AND“. Manuální úprava není nikterak složitá, ale pro zjednodušení práce byla vytvořena tabulka v Google Tabulkách. Jedním z důvodů byla úspora času. V rámci testování se používalo několik postupů úpravy dat, které nebyly zdaleka finálními, ale bylo je nutné vyzkoušet. Proto je automatizace úpravy pravidel na podmínky přívětivou možností, jak tento postup urychlit.

Ukázka původního pravidla:

$$Budova\ =\ transportation \rightarrow Budova\_2 =\ civic$$

Upravené pravidlo na podmínku do QGIS:

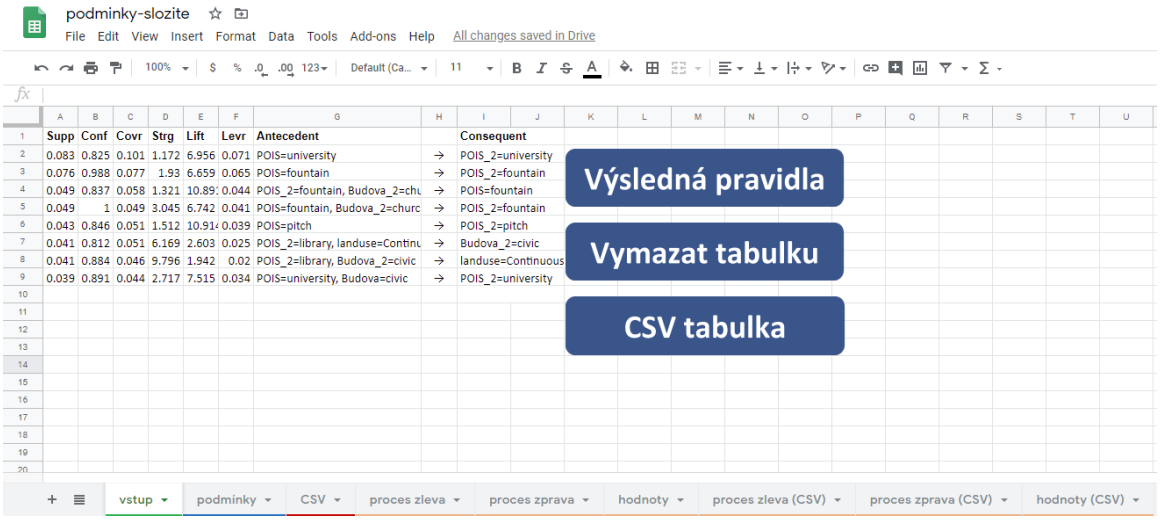
$$Budova\ = 'transportation'\ AND\ Budova\_2 =\ 'civic'$$

Tabulka obsahuje celkem 9 listů:

- *Vstup* – vložení tabulky
- *Podmínky* – vygenerované podmínky určené ke kopírování
- *CSV* – podmínky pro uložení do CSV tabulky
- *Proces zleva* – úprava předpoklad
- *Proces zprava* – úprava závěr
- *Hodnoty* – zkopírování hodnot
- *Proces zleva (CSV)* – úprava předpoklad pro CSV tabulku
- *Proces zprava (CSV)* – úprava závěr pro CSV tabulku
- *Hodnoty (CSV)* – vygenerované hodnoty pro CSV tabulku

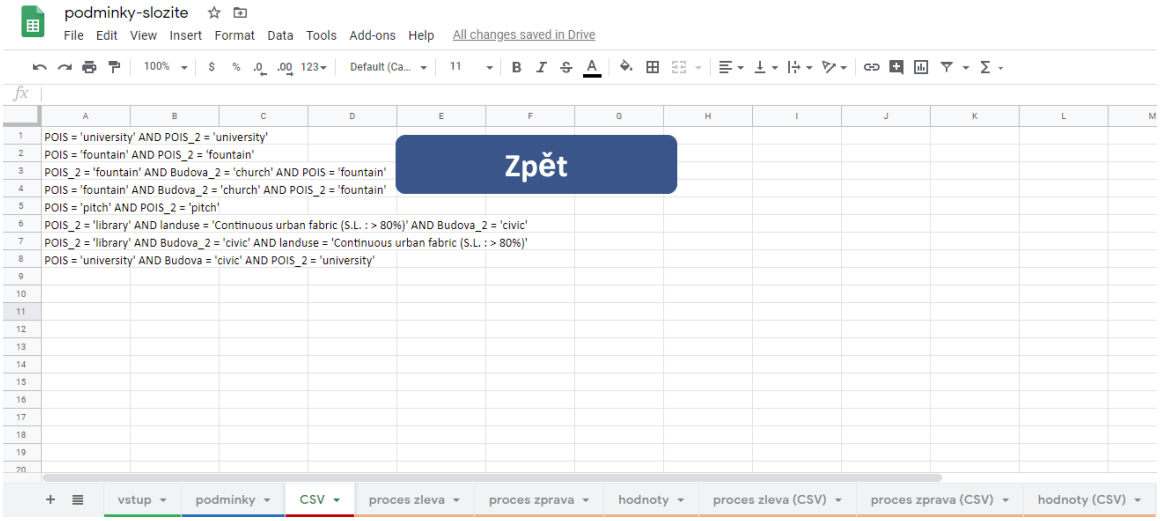
Pro uživatele je důležitý list *vstup* (označený zelenou barvu), do kterého se zkopíruje dle návodu tabulka, která byla vygenerována v rámci reportu od Orange. List je připravený tak, že se vloží pouze záznamy z tabulky bez záhlaví. Maximální počet atributů na jedné straně je čtyři. Pro více atributů tabulka již nebude fungovat. Zbytek zpracují ostatní listy – *proces zleva* a *proces zprava*. Zde jsou nadefinované vzorce, které postupně upravují pravidla do podmínek. Následně uživatel pouze klikne na příslušné tlačítko *Výsledná pravidla*. Tlačítko uživatele přenese do listu *podmínky*, ve kterém jsou všechna asociační pravidla upravená do požadovaného formátu. Pro vrácení se do úvodního listu stačí kliknout na tlačítko *Zpět*.

Každý záznam v listu je upravený a lze jej snadno zkopírovat jako podmínku do QGIS symbologie. Pokud uživatel chce zadat nová data, stačí pouze použít nadefinované tlačítko *Vymazat tabulku*, které smaže veškerý obsah. Tabulka obsahuje tlačítka, která volají makra. Tato makra byla vytvořená velmi jednoduše pomocí funkce k zaznamenávání nových maker. Aktuálně je tabulka upravená tak, že pracuje pouze s tabulkami, které mají maximální počet řádků 100. Je to z důvodu velikosti a funkcionality souboru. Pro rozšíření o více řádků stačí pouze zkopírovat jednotlivé vzorce z konkrétního sloupce. Zároveň je nutné upravit i funkce vybraných maker.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.083	0.825	0.101	1.172	6.956	0.071	POIS=university	→	POIS_2=university
0.076	0.988	0.077	1.93	6.659	0.065	POIS=fountain	→	POIS_2=fountain
0.049	0.837	0.058	1.321	10.89	0.044	POIS_2=fountain, Budova_2=chu	→	POIS=fountain
0.049	1	0.049	3.045	6.742	0.041	POIS=fountain, Budova_2=churc	→	POIS_2=fountain
0.043	0.846	0.051	1.512	10.91	0.039	POIS=pitch	→	POIS_2=pitch
0.041	0.812	0.051	6.169	2.603	0.025	POIS_2=library, landuse=Continu	→	Budova_2=civic
0.041	0.884	0.046	9.796	1.942	0.02	POIS_2=library, Budova_2=civic	→	landuse=Continuous
0.039	0.891	0.044	2.717	7.515	0.034	POIS=university, Budova=civic	→	POIS_2=university

Obr. 12 Náhled na vytvořenou Google tabulku s vloženými pravidly *(zdroj: autorka)*



Obr. 13 Upravená asociační pravidla na podmínky pro QGIS *(zdroj: autorka)*

5.3.2 Skript v QGIS

Připravená tabulka bude následně sloužit pro vytvořený skript, který je spustitelný v QGIS. Stačí pouze zvolit tlačítko *CSV tabulka*, která otevře list s názvem *CSV*. Následně je potřeba vybrat v horní liště *Soubor-Stáhnout-CSV*, díky kterému se vybraný list stáhne. Tento soubor bude následně použit pro vytvořený skript.

Skript je napsán v prostředí QGIS pomocí *Python konzole*. Celý kód je napsán v jazyce Python, v prostředí QGIS se skriptování označuje jako PyQGIS. První část skriptu definuje funkci vizualizace založené na pravidlech. Funkce má nadefinované parametry. Jedním z nich je *layer,* tedy vrstva, pro které skript poběží. Vrstva je nastavená tak, že do skriptu bude vybrána aktivní vrstva v katalogu. Dále se definuje symbol, jehož tvar je určený výchozím tvarem vrstvy. Součástí skriptu je načtení dané CSV tabulky, kterou použije pro přečtení výrazů, resp. podmínek symbologie. Je nutné dodržet podmínku absolutní cesty k souboru (nadefinování proměnné *f1*), popř. ji upravit dle svých potřeb – změnit název souboru, popř. jeho umístění. Stačí změnit část uvozenou dvojitými uvozovkami.

Skript projde každý řádek v tabulce a vytvoří pro něj samostatnou kategorii dle zadané podmínky. Zároveň každou kategorii odliší barevně dle nadefinovaného pole s hexadecimální kódy barev. Aktuálně skript obsahuje nadefinovaných 8 barev, v případě více pravidel skript neproběhne. Řešením je dopsání dalších kódu barev. Skript je nastavený tak, že pokud je již řádek v *CSV* tabulce prázdný, skript ukončí a uživatele informuje jednoduchou hláškou o provedení skriptu.

```
def rule_based_style(layer, symbol, label, expression, color):
    root_rule = renderer.rootRule()
    rule = root_rule.children()[0].clone()
    rule.setLabel(label)
    rule.setFilterExpression(expression)
    rule.symbol().setColor(QColor(color))
    root_rule.appendChild(rule)
layer = iface.activeLayer()
symbol = QgsSymbol.defaultSymbol(layer.geometryType())
renderer = QgsRuleBasedRenderer(symbol)
# lokace souboru k nacteni
f1 = open("d:/CSV-podminky.csv", "r", encoding='utf-8', errors='ignore')
# nadefinovane barvy
colours =
['#66c2a5','#fc8d62','#8da0cb','#e78ac3','#a6d854','#ffd92f','#e5c494',
'#b3b3b3']
cislo=-1
cisloRange = len(colours)
for line in f1.readlines():
    if cislo<=cisloRange:
        cislo+=1
    if '=' in line:
        rule_based_style(layer, symbol, str(line), line, colours[cislo])
# nastaveni symbologie pro ostatni prvky
rule_based_style(layer, symbol, 'nesplňuje pravidlo', 'ELSE',
'#00F0F8FF')
print('Konec skriptu')
layer.setRenderer(renderer)
layer.triggerRepaint()
iface.layerTreeView().refreshLayerSymbology(layer.id())
```

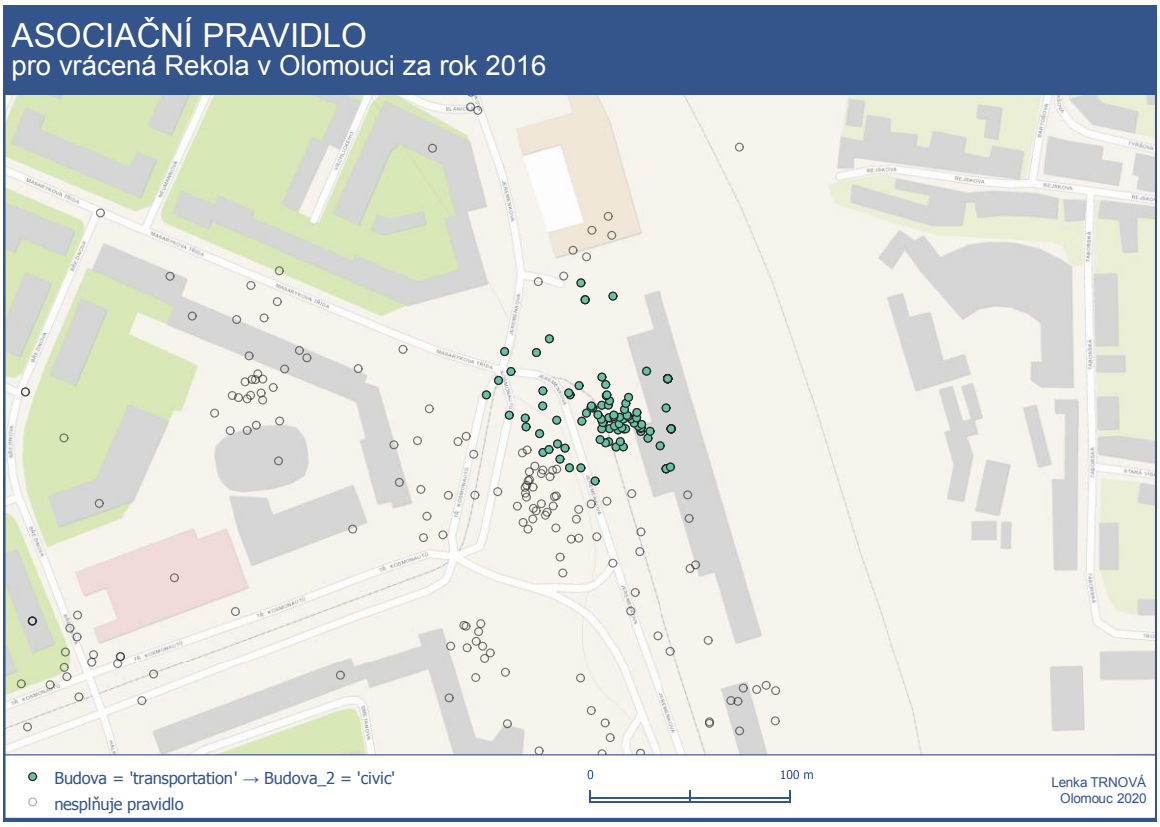
5.3.3 Výsledná mapa

Díky skriptu lze na vygenerovaná asociační pravidla nahlédnout v prostorové rovině. V rámci QGIS je možné si zobrazit, kolik prvků splňuje danou podmínku, resp. pravidlo. Došlo ke zjištění, že jeden prvek může splňovat více podmínek. Proto nelze všechna pravidla zvizualizovat do jedné mapy. V tabulce níže lze vidět počet prvků splňujících konkrétní pravidla.

Tab. 10 Počet prvků vrstvy Rekol splňující konkrétní pravidlo

ID	Pravidlo			Počet prvků splňující pravidlo
1	Budova=transportation	→	Budova_2=civic	106
2	Budova=transportation	→	POIS_2=fountain	101
3	POIS_2=fountain, Budova_2=civic	→	Budova=transportation	101
4	POIS=post_box	→	Budova=residential	101
5	POIS=tourist_info	→	Budova_2=civic	97
6	POIS=tourist_info	→	POIS_2=fountain	96
7	POIS_2=fountain, Budova_2=civic	→	POIS=tourist_info	96
8	POIS=tourist_info	→	Budova=transportation	92

Pro ukázku výsledné mapy bylo vybráno 1. asociační pravidlo. Pomocí bodové metody jsou zobrazeny prvky, které splňují a které nesplňují vybrané pravidlo. Výsledek lze vidět na mapě 1. Na mapě lze pozorovat shluk bodových prvků, které splňují pravidlo *Budova=transportation → Budova_2=civic*. Ostatní prvky, které nesplňují pravidlo, jsou zobrazeny symbolem s transparentní výplní. Body splňující toto pravidlo, jsou lokalizovány v okolí hlavního nádraží v Olomouci. Je patrné, že kola zde zákazníci využili pro další transport na své cestě.



Mapa 1 Vybrané asociační pravidlo pro vrácená Rekola v Olomouci za rok 2016

5.4 Strukturní diagram

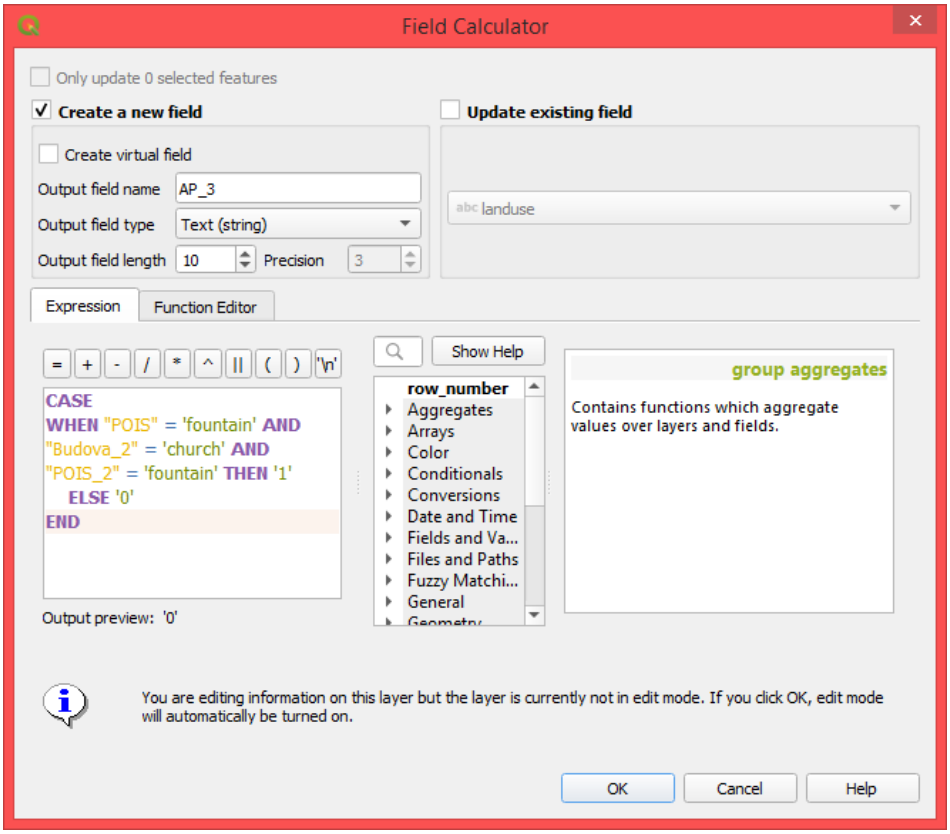
V rámci studie byly snaha nalézt i jiný způsob vizualizace dat, který bude mnohem propracovanější a bude podávat informace o více asociačních pravidlech zároveň. Jednou z možností je využití tvorby strukturního diagramu v QGIS *(Pie chart)*. Pro tuto metodu je ale nutná předpříprava dat.

Nejdůležitější částí je tvorba nových atributových sloupců. Každý sloupec představuje jedno asociační pravidlo. Záleží na uživateli, která pravidla zvolí, v tomto případě se bude pokračovat s výše vygenerovanými pravidly. Těchto pravidel je celkem 8, bude tedy vytvořeno 8 nových sloupců. Každý sloupec se pro přehlednost označí číslem pravidla. Označení sloupce tedy může vypadat takto: *AP_1*.

Každý sloupec je potřeba naplnit binárními hodnotami – nulou nebo jedničkou. K tomu se využije nástroje *Field Calculator*, který je dostupný v rámci atributové tabulky. Pomocí nástroje a příslušného příkazu se jednotlivé sloupce vyplní konkrétní hodnotou. Příkaz může vypadat např. takto:

```
CASE
        WHEN "Budova" = 'transportation' AND "Budova_2" = 'civic' THEN '1'
        ELSE '0'
END
```

Příkaz tvoří podmínka, kterou lze následně získat v připravené Google Tabulce, která byla využita i pro předchozí vizualizaci. Pouze se využije jiný list, konkrétně druhý list s názvem *„podmínky“*. Zde jsou jednotlivá pravidla upravená na podmínky, které lze jednoduše zkopírovat a použít pro příkazy. Samotný příkaz obsahuje tuto podmínku (upravené asociační pravidlo). V případě, že jej daný prvek splňuje, získá hodnotu jedna. V opačném případě se pro konkrétní prvek zapíše hodnota nula – podmínku nesplňuje.



Obr. 14 Vytvoření nového sloupce *AP_3* se zadanou podmínkou *(zdroj: autorka)*

Poslední sloupec, který je nutno vytvořit je sloupec *AP_Sum*, jehož obsah, jak již z názvu vyplývá, je součet všech nově vytvořených sloupců. Díky tomu lze získat informaci, kolik pravidel zároveň splňuje konkrétní prvek.

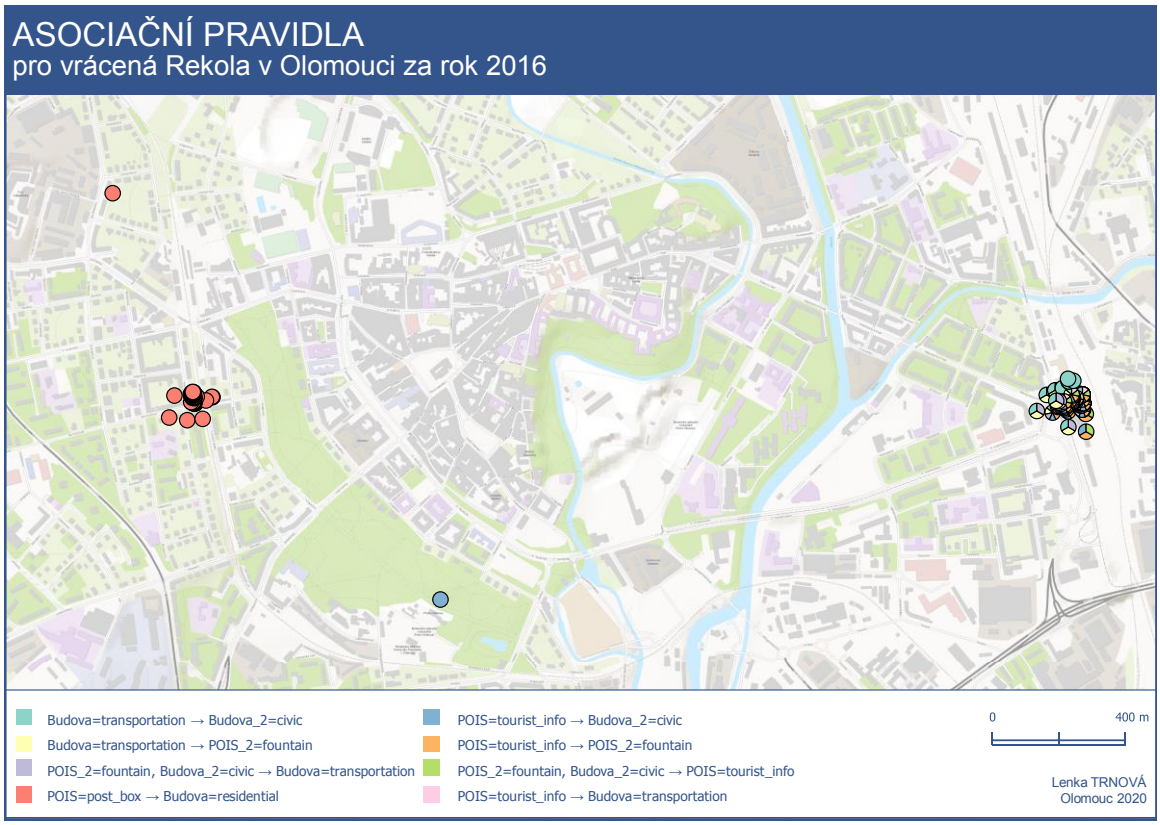
Po proběhnutí příkazů pro všechna pravidla bude tabulka vypadat následovně:

Tab. 11 Ukázka záznamů s výslednými atributy pro jednotlivá asociační pravidla

AP_1	AP_2	AP_3	AP_4	AP_5	AP_6	AP_7	AP_8	AP_sum
1	0	0	0	0	0	0	1	2
1	0	0	0	0	0	0	1	2
0	0	0	0	0	1	1	0	2
0	0	0	0	0	1	1	0	2
1	0	0	0	0	0	0	0	1
0	1	1	1	0	0	0	0	3

Do vizualizace vstupují nově vytvořené atributové sloupce kromě sloupce *AP_sum*. Ten je použit v záložce *Rendering* v sekci *Visibility* jako podmínka pro zobrazení pouze těch prvků, které splňují alespoň jedno asociační pravidlo. Na výsledné mapě tedy chybí ty prvky, které nesplňují ani jedno z výše uvedených pravidel. Podmínka je velmi jednoduchá a má podobu:

```
"AP_sum" >0
```



Mapa 2 Strukturní diagram s vybranými asociačními pravidly pro vrácená Rekola

Výsledek lze vidět na mapě 2. Opět se potvrzuje, že jeden prvek může splňovat více pravidel zároveň. Mapa obsahuje strukturní diagramy, jejichž výseče se skládají ze splněných pravidel. Pokud je symbol jedné barvy, znamená to, že splňuje pouze jedno pravidlo. Pokud je naopak kruh dělený na sedm výsečí, splňuje sedm pravidel. Kromě shluku u hlavního nádraží se jeden ze shluků nachází na Nové Ulici, konkrétně v okolí ulice Na Vozovce. Zde vrácená kola splňují čtvrté asociační pravidlo, které má předpoklad, že se vrátilo kolo v blízkosti poštovní schránky a závěr říká, že se v okolí nachází budova, která je určená k bydlení.

5.5 Velikost obalové zóny

Vzhledem k tomu, že je tato případová studie jednodušší, je vhodná pro analýzu vlivu velikosti obalové zóny. V následující podkapitole dojde k jejímu otestování. V rámci testování odlišné velikosti obalové zóny budou sledovány nejfrekventovanější asociační pravidla v rámci datové sady. V tabulce níže lze vidět prvních 8 pravidel jako v případě obalové zóny o 100 m, tentokrát ale pro obalovou zónu o velikosti 50 m. Opět byla data upravená pomocí postupu na odstranění duplicitních hodnot.

Čím větší je obalová zóna, tím více budov bude obsaženo a tím méně přesně bude určený dotyk mezi vrstvami. Proto je vhodné zvolit velikost dle povahy datové sady. U polygonové vrstvy budov, které jsou ve městě velmi hustě rozmístěné, je tedy na místě zvolit rozumnou nízkou hodnotu pro obalovou zónu. To stejné platí i pro zájmové body.

Tab. 12 Asociační pravidla pro obalovou zónu 50 m

Podpora	Spolehlivost	Předpoklad	→	Závěr
0.035	1.000	Budova=train_station	→	Budova_2=civic
0.032	0.803	POIS=post_box	→	Budova_2=residential
0.030	0.990	POIS=post_box, Budova=civic	→	Budova_2=residential
0.030	0.814	POIS=recycling	→	landuse=Continuous urban fabric (S.L. : > 80%)
0.028	0.957	POIS=biergarten	→	Budova=train_station
0.028	0.957	POIS=biergarten	→	Budova_2=civic
0.028	1.000	POIS=biergarten, Budova_2=civic	→	Budova=train_station
0.024	0.962	landuse=Continuous urban fabric (S.L. : > 80%), POIS=post_box	→	Budova=civic

Tato případová studie je první ukázkou uceleného postupu, jak pro prostorová data hledat asociační pravidla. I přesto, že mezi pravidly nebylo nalezeno ani jedno frekventované pravidlo, lze konstatovat, že postup je funkční a spolehlivý. Zároveň byl nalezen způsob, jak asociační pravidla vizualizovat, a to nejen jedním způsobem. Také je důležité si určit správně velikost obalové zóny, její velikost ovlivňuje výsledná asociační pravidla. Všechny získané poznatky budou využity pro práci v následujících dvou případových studiích.

6 PŘÍPADOVÁ STUDIE 2 – NABÍJECÍ STANICE

Snaha prosadit ekologičtější způsob dopravy automobily s sebou nese výstavbu nabíjecích stanic pro elektromobily. Cílem této případové studie je nalézt možné souvislosti při umístění stávajících nabíjecích stanic.

6.1 Data

Data nabíjecích stanic byla získána z webové stránky *EVMAPA*[5], na které se nacházejí souřadnice. Tyto souřadnice byly zadány do QGIS pomocí *Add Delimited Text Layer* a zvizualizovány do bodové vrstvy. Původní vrstva obsahuje stanice i za hranicemi ČR, proto byla data ořezána na naše území. Takto upravená vrstva obsahuje celkem 475 záznamů. Vzhledem k přenosu pouze souřadnic datová sada neobsahuje žádný atribut, který by mohl vstupovat do analýzy. Vrstva slouží pouze jako prostorový základ, ke kterému se další prostorové informace připojí.

Vrstva využití území pro celou ČR byla využitá z OSM balíčku, který je stažitelný pomocí *Geofabrik*. Další vrstvy byly získány pomocí pluginu *QuickOSM*. Konkrétně se jedná o liniovou vrstvu silnic, zájmových bodů a čerpacích stanic.

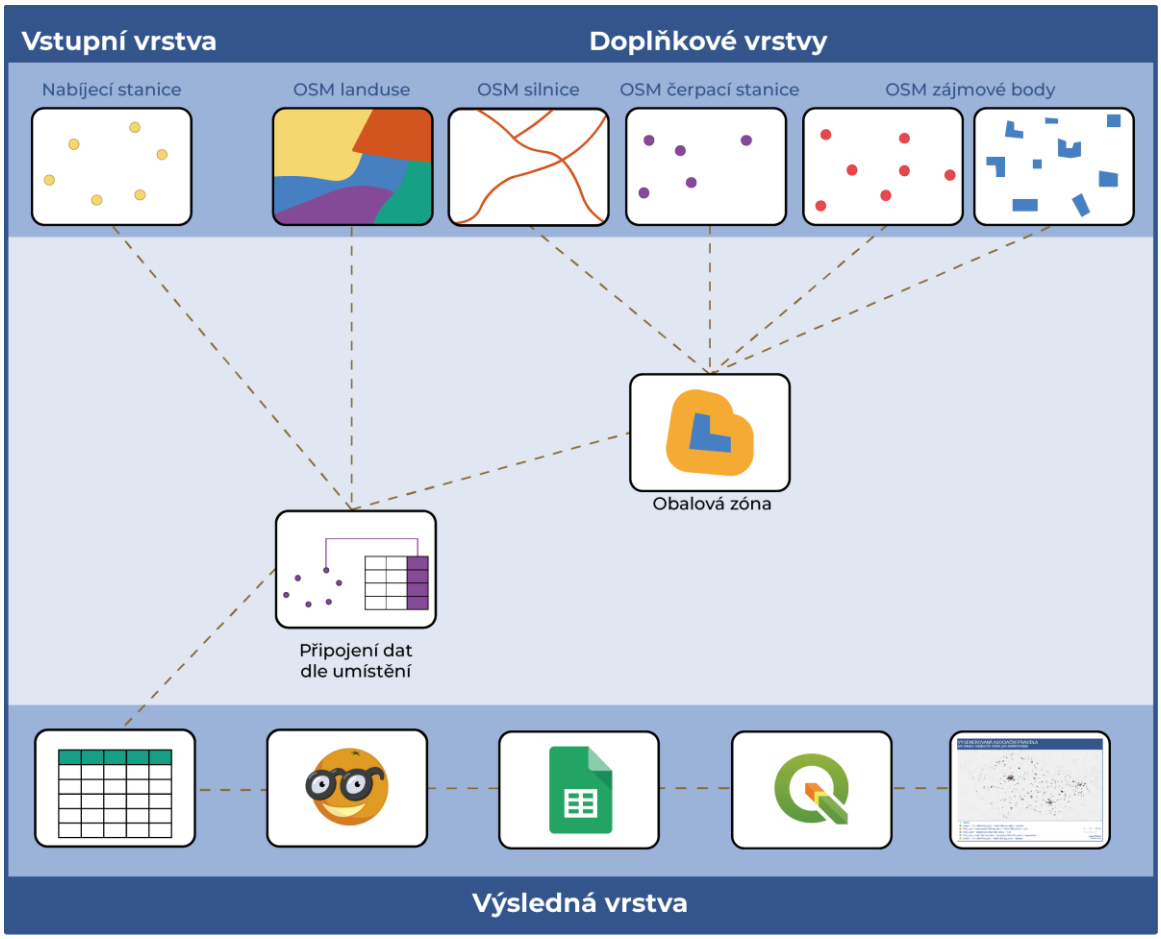
Do dalšího postupu vstupují vrstvy (+ vybraný atribut):

- Primární vrstva
  - Nabíjecí stanice pro elektromobily (bodová vrstva) – žádný atribut
- Doplňkové vrstvy
  - Silnice OSM (liniová vrstva) – atribut *Typ_silnic*
  - Využití území OSM (polygonová vrstva) – atribut *Landuse*
  - Zájmové body OSM (bodová a polygonová vrstva) – atribut *POIS*

6.2 Tvorba modelu

Jako podklad byl využit model, který byl vytvořený pro potřeby předchozí případové studie týkající se Rekol. Ten se postupně rozšiřoval, až vznikl model, který je součástí přílohy č. 1. Také byl pro jeho vznik využit počáteční model pro testovací datovou sadu, který zapíše do nového atributu hodnotu 0 nebo 1 na základě toho, zda se ve vzdálenosti 100 m nachází čerpací stanice. Na obr. 15 je možné vidět ilustrační schéma postupu. Výsledkem celého procesu je mapa, ve které lze pozorovat, které prvky splňují, jaké asociační pravidlo.

[5] https://www.evmapa.cz/



Obr. 15 Zjednodušené schéma postupu *(zdroj: autorka)*

Jako vstupní vrstva slouží datová sada s umístěním všech nabíjecích stanic. Zbylé datové sady jsou označené jako doplňkové vrstvy. Pro doplňkové vrstvy kromě *OSM využití území* byla vytvořená obalová zóna, jejíž vzdálenost lze uživatelem libovolně nastavit. Obalová zóna je vždy tvořená od doplňující vrstvy, ne od primární. V rámci měření byla pro všechny vrstvy nastavená obalová zóna 100 m. Pouze u silnic byla zvolená vzdálenost 500 m, která byla určená jako adekvátní na základě lokace nabíjecích stanic. Vrstva *Využití území* přímo vstupovala do analýzy připojení dat dle umístění, jelikož pokrývá celé zájmové území a všechny prvky se dotýkají, v datech není žádné prázdné místo.

Poté, co pro zbylé vrstvy byly vytvořené obalové zóny s danou vzdáleností, bylo použito jednotlivě připojení dat dle umístění tak, aby každá obalová zóna získala informace ze své původní vrstvy. Tyto obalové zóny obohacené o informace následně vstupovaly do další fáze připojení dat dle umístění tak, aby výsledná vrstva nabíjecích stanic obsahovala nové atributové sloupce pro každou doplňkovou vrstvu. Každý atribut tedy obsahuje informaci, zda v okolí konkrétní nabíjecí stanice se nachází prvek doplňkové vrstvy. U *využití území* je naopak uvedená informace o tom, v jakém využití území se nachází nabíjecí stanice. Takto upravená tabulka bude následně vyexportovaná ve formě CSV tabulky pro účely SW Orange, kde dojde k vygenerování asociačních pravidel dle určitých nastavení.

Tab. 13 Ukázka výsledné tabulky vytvořené pro nabíjecí stanice

Benzinka	Landuse	Typ_silnic	Typ_silnic_1	POIS	POIS_2
ano	meadow	motorway	motorway	toilet	toilet
ano	residential	trunk	primary	chemist	supermarket
ne		primary	primary	stadium	stadium
ne		secondary	primary	recycling_glass	sports_centre
ne	farm	secondary	secondary	playground	playground
ne	grass			vending_parking	vending_parking
ne	industrial	motorway	motorway		
ne	recreation_ground	motorway	motorway	fast_food	fast_food
ne	residential	secondary	secondary	town_hall	town_hall
ne	residential	motorway	motorway	tower	car_dealership
ne	residential	secondary	secondary	restaurant	sports_centre

6.2.1 Časová náročnost

Nevýhodou tohoto modelu, resp. datové sady je její časová náročnost. Kvůli velkému území a velkému počtu prvků, které do analýzy vstupují, model běžel přes 3 hodiny na stolním počítači.

6.2.2 Vzdálenost od čerpacích stanic

V rámci případové studie byla obalová zóna určená subjektivně na 100 m. Zde lze konstatovat, že čím větší obalová zóna by se nastavila, tím více čerpacích stanic by se v blízkosti nabíjecích stanic nacházelo. Otázkou ale zůstává, zda je vyšší vzdálenost správná. Řidič, vlastnící elektromobil, nevyhledává čerpací stanici kvůli palivu. Dá se předpokládat, že zde případně zastaví něco zakoupit nebo na toaletu. V tomto případě je 100 m adekvátní vzdáleností.

6.3 Generování pravidel

Po proběhnutí modelu se výsledná vrstva exportuje do *CSV* tabulky, která je následně nahrána do Orange. V této případové studii byl postup pro odstranění duplicitních hodnot přeskočen. Testování bude také zaměřeno na taková pravidla, která splňují termín spolehlivá výjimka. Pro taková pravidla je nutné nastavit minimální spolehlivost na 100 %. Pro generování pravidel byly využity následující atributové sloupce, které obsahují informace z doplňkových vrstev:

- *Benzinka* – zda je v obalové zóně čerpací stanice
- *Landuse* – využití území, ve kterém se nabíjecí stanice nachází
- *POIS* – zájmové body v obalové zóně
- *Typ_silnic* – třídy silnice, která se v okolí nachází

Tab. 14 Asociační pravidla pro nabíjecí stanice

Podpora	Spolehlivost	Předpoklad	→	Závěr
0.331	0.963	Landuse=residential	→	Benzinka=ne
0.291	0.902	typ_silnic=primary	→	Benzinka=ne
0.291	0.926	typ_silnic=secondary	→	Benzinka=ne
0.124	0.937	Landuse=residential, typ_silnic=secondary	→	Benzinka=ne
0.105	0.980	Landuse=residential, typ_silnic=primary	→	Benzinka=ne
0.084	0.930	Landuse=industrial	→	Benzinka=ne
0.067	0.914	POIS=restaurant	→	Benzinka=ne
0.032	1.000	POIS=mall	→	Benzinka=ne

Výsledkem je výběr celkem 8 prvních pravidel, které lze vidět v tabulce výše. Nejfrekventovanější pravidlo má podporu 33,1 % a spolehlivost 96,3 %. Tato pravidlo zní následovně: předpoklad říká, že nabíjecí stanice se nachází ve využití území, které je definováno jako rezidenční a zároveň v 96,3 % platí závěr, že se zde v okolí nenachází čerpací stanice. Druhé pravidlo má předpoklad, že typ silnice, která se v okolí nachází, je silnice I. třídy a v 90,2 % platí závěr, že se zde v okolí nenachází čerpací stanice. Nejméně frekventované pravidlo v rámci tohoto výběru je 8. pravidlo, které má podporu pouze 3,2 %, ale naopak spolehlivost 100 %. V následující podkapitole bude věnována pozornost právě tomuto typu pravidel.

6.3.1 Spolehlivá výjimka

Jak již bylo naznačeno, mezi pravidly se nacházejí i taková pravidla, která mají spolehlivost 100 %. Spolehlivá výjimka musí mít hodnotu spolehlivosti 100 %. V tabulce výše lze zaznamenat dvě pravidla s tou hodnotou. U těchto pravidel lze konstatovat, že platí předpoklad i závěr zároveň. Konkrétně se jedná o pravidlo říkající, že v blízkosti nabíjecí stanice u hotelu se nenachází čerpací stanice. Druhá spolehlivá výjimka říká, že v blízkosti nabíjecí stanice u nákupního centra se také nenachází čerpací stanice.

6.3.2 Filtrování vrstev

U této případové studie je nevýhodou, že se prvně zobrazují nejčastější pravidla, která jsou záporná – stanice se nenachází blízko čerpací stanice (*Bezinka = ne*). Celkem 435 záznamů totiž podmínku blízkosti benzínky do 100 m nesplňuje (z celkového počtu 475). V Orange je možnost některý z atributů přeskočit, proto v rámci tohoto testu bude atributu čerpacích stanic (*benzinka*) nastavena role k přeskočení. Výsledkem byla pravidla, která nedosahují podpory vyšší jak 0,6 %, což znamená, že takové pravidlo splňují přibližně tři prvky z celé datové sady.

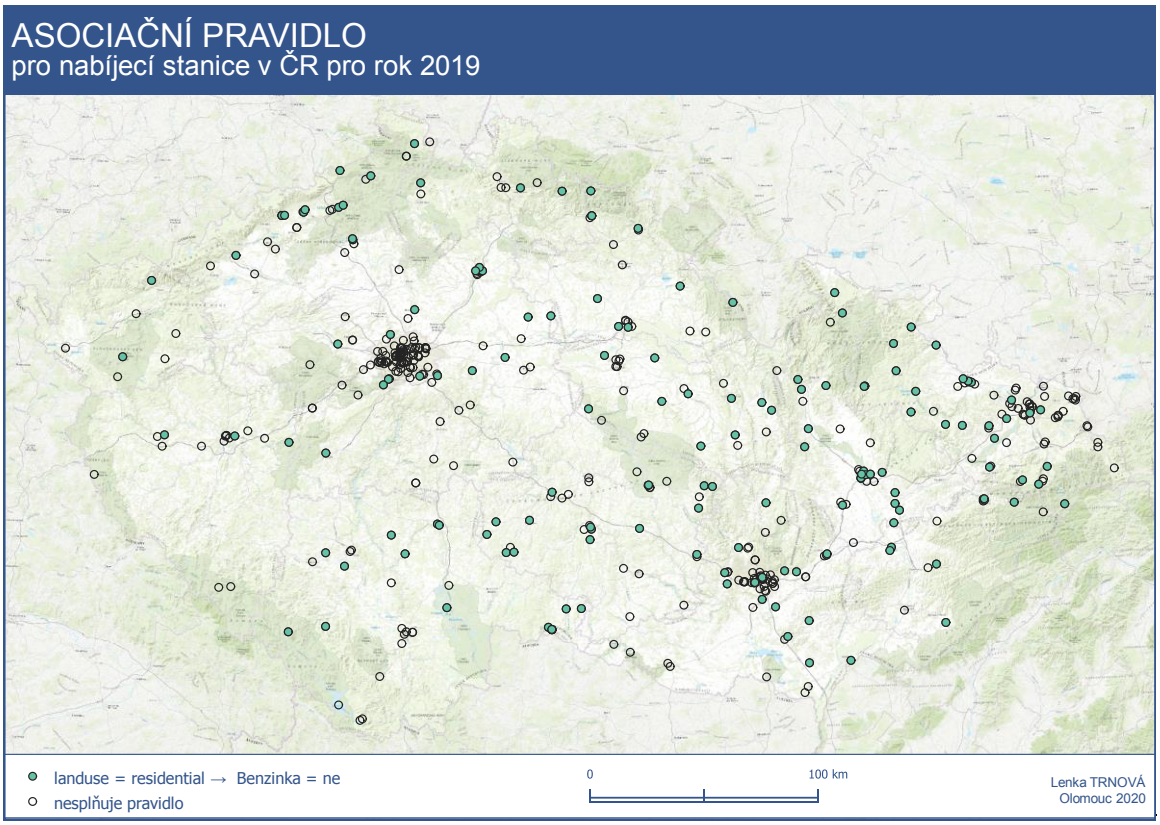
6.4 Vizualizace asociačních pravidel

Pro vizualizaci pravidel do prostoru bylo opět využito Google Tabulek pro převedení pravidel na podmínky do symbologie. Následně byl použit Python skript pro nahrání podmínek do symbologie a nastavení nadefinovaných barev. V tab. 15 lze vyčíst, kolik prvků splňuje dané asociační pravidlo.

Tab. 15 Počet prvků nabíjecích stanic splňující konkrétní pravidlo

ID	Pravidlo			Počet prvků splňující pravidlo
1	Landuse=residential	→	Benzinka=ne	157
2	typ_silnic=primary	→	Benzinka=ne	138
3	typ_silnic=secondary	→	Benzinka=ne	138
4	Landuse=residential, typ_silnic=secondary	→	Benzinka=ne	59
5	Landuse=residential, typ_silnic=primary	→	Benzinka=ne	50
6	Landuse=industrial	→	Benzinka=ne	40
7	POIS=restaurant	→	Benzinka=ne	32
8	POIS=mall	→	Benzinka=ne	15

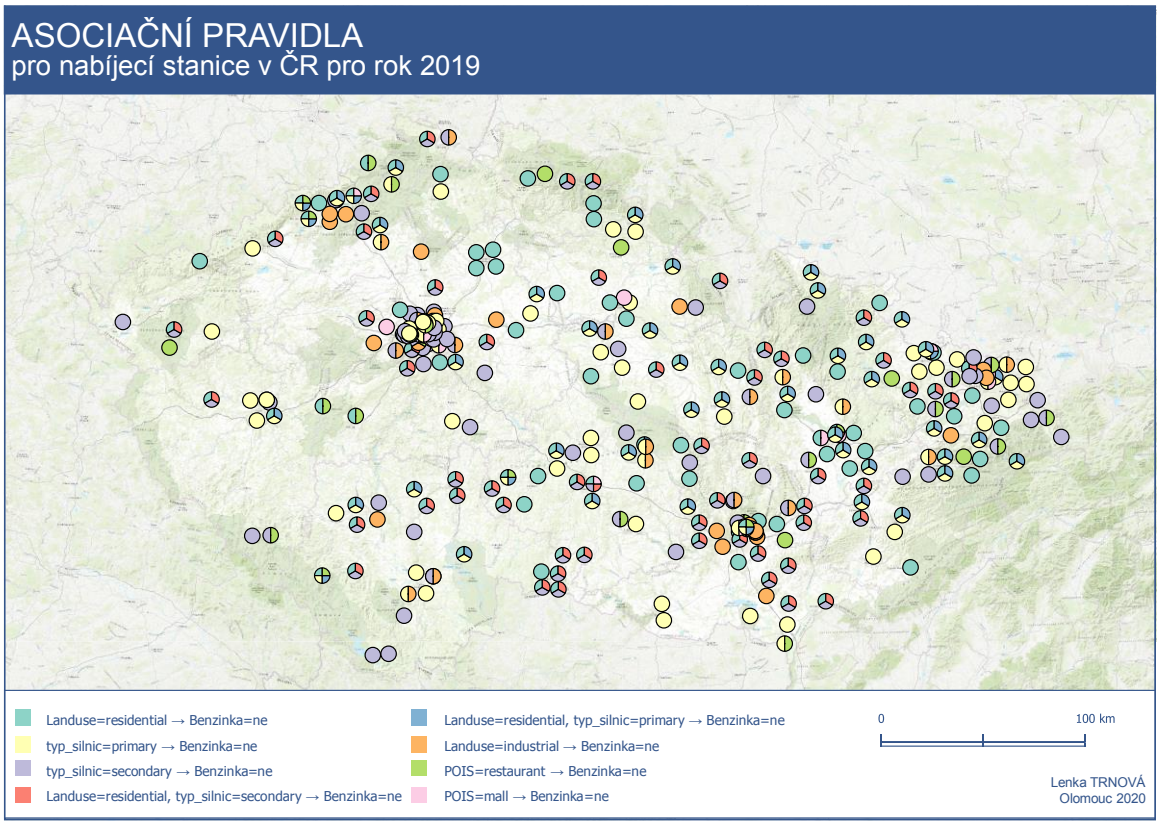
Na mapě 3 lze sledovat výskyt prvního asociačního pravidla, které se týká využití území, ve kterém se nabíjecí stanice nachází. Celkem 157 prvků z datové sady splňuje tuto podmínku, že se stanice nachází v rezidenční oblasti. Zbylé nabíjecí stanice, které toto pravidlo nesplňují, jsou zobrazeny symbolem s transparentní výplní.



Mapa 3 Asociační pravidlo pro nabíjecí stanice

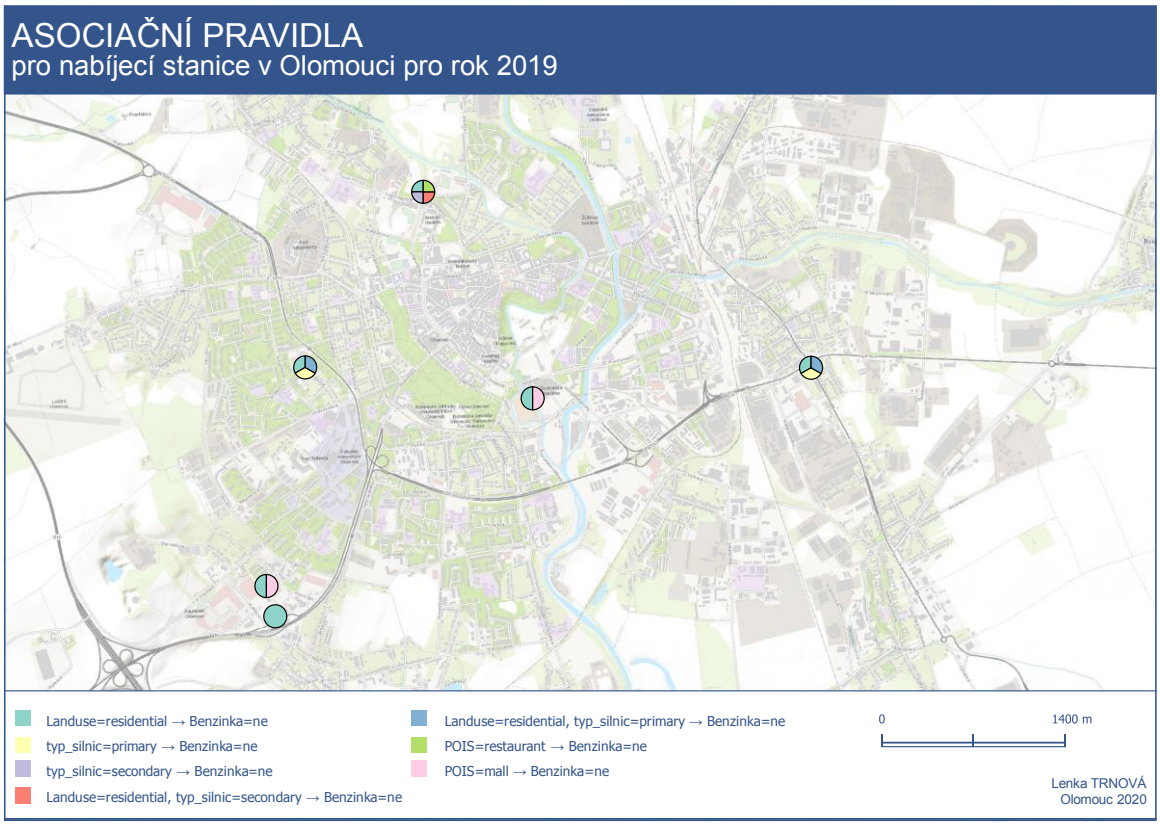
6.4.1 Strukturní diagram

Došlo k obdobné úpravě atributové tabulky jako u případové studie Rekol. Bylo přidáno celkem 8 atributových sloupců (*AP_1* až *AP_8*), jejichž hodnoty byly zapsány na základě podmínky, zda prvek dané pravidlo splňuje. Devátým sloupcem je opět suma s počtem pravidel, které prvek splňuje (*AP_sum*). Výsledkem je mapa 4. Zobrazeny jsou pouze ty prvky, které splňují alespoň jedno z vybraných asociačních pravidel. Z mapy lze pozorovat, že některé z prvků splňují i čtyři asociační pravidla zároveň.



Mapa 4 Strukturní diagram s vybranými asociačními pravidly pro nabíjecí stanice

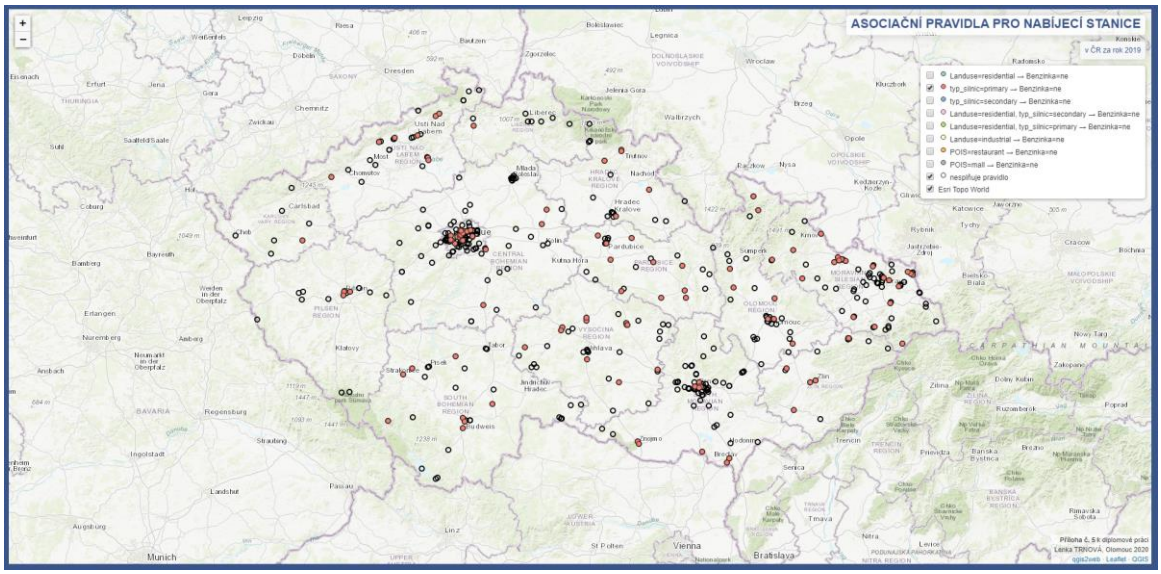
Pro zajímavost a lepší přehlednost byla vytvořená mapa 5 s detailem na město Olomouc. Jedinou zásadní změnou je, že ani jedna z celkem devíti nabíjecích stanic nesplňovala 6. pravidlo, které bylo proto z legendy odstraněno. Nabíjecí stanice jsou různorodé – splňují od jednoho pravidla až po čtyři zároveň. Tři z nich ale v mapovém poli nejsou zobrazené – nesplňují ani jedno z pravidel.



Mapa 5 Asociační pravidla pro nabíjecí stanice v Olomouci

6.4.2 Leaflet mapa

Vzhledem k tomu, že jsou data za celé území České republiky a na hodně místech se prvky překrývají, je vhodné mít interaktivní mapu, ve které se bude možno přiblížit. Jednou z možností je tvorba mapy v prostředí Leaflet, což je JavaScriptová knihovna, která je open-source. V rámci QGIS lze stáhnout zásuvný modul *qgis2web*, který umožňuje exportovat mapu buď v prostředí OpenLayers nebo zmiňovaném Leaflet. Tohoto nástroje bylo využito a následně byl kód mapy pouze poupraven pro své potřeby. Výsledkem je interaktivní mapa, jejíchž ukázku lze vidět na obrázku níže. Uživatel zde může vybírat v seznamu vrstev vybrané asociační pravidlo, které chce zvizualizovat. Zároveň je nastavená vyskakovací okno, které informuje, kolik asociačních pravidel daný prvek zároveň splňuje.



Obr. 16 Interaktivní mapa pro nabíjecí stanice v prostředí Leaflet *(zdroj: autorka)*

Cílem této případové studie bylo nalezení takového způsobu úpravy dat tak, aby bylo možné analyzovat více vstupujících datových sad. Je patrné, že i přesto, že vstupuje celkem pět doplňkových vrstev, většina nejfrekventovanějších pravidel v rámci této sady obsahuje maximálně dvě položky na jedné ze stran implikace. Narozdíl od Rekol zde byly nalezeny i taková pravidla, která lze označit za mnohem frekventovanější. První z nich má podporu přes 33 %.

Funkcionalita vytvořené Google Tabulky a skriptu v rámci první případové studie byla ověřena a je spolehlivým způsobem, jak ulehčit práci při vizualizaci pravidel zpět do prostoru. Zároveň bylo použito nové prostředí pro vizualizaci všech asociačních pravidel, a to interaktivní prostředí knihovny Leaflet.

7 PŘÍPADOVÁ STUDIE 3 – DĚTSKÁ HŘIŠTĚ

Poslední případová studie se zaměří podrobněji na generování asociačních pravidel pouze vždy pro dvě doplňkové datové sady. To znamená, že ke vstupní vrstvě se budou přidávat prostorové údaje z dalších dvou datových sad. Jak již zaznělo dříve, čím méně datových sad do testování vstupuje, tím je větší šance získat frekventovaná pravidla.

7.1 Použitá data

Jako primární datová sada byla použitá data od studenta Filipa Urbančíka, který je využije ve své bakalářské práci. Data byla pro něj poskytnutá od Mgr. Jana Dygrýna z Fakulty tělesné kultury Univerzity Palackého v Olomouci. Bodová vrstva obsahuje celkem 138 prvků lokalizujících různá dětská hřiště v Olomouci a jeho přilehlém okolí.

K lokacím dětských hřišť byly připojeny vrstvy týkajících se vybraných zájmových bodů a adresní body. Zájmové body byly opět získány pomocí *QuickOSM* a vzhledem k povaze dat byly vybrány body a polygony, které představují buď školku, školu nebo park. Klíč pro školky a školy ve filtrování záznamů je *amenity*, zatímco pro park je klíčem *leisure*.

Další vrstvou je bodová vrstva adresních bodů. Vrstva obsahuje celkem 62 atributů, pro práci je vybrán pouze atribut s názvem *BUDOBYTSL*, který uvádí počet obyvatel s uvedeným trvalým pobytem. Adresní body jsou bodová vrstva, která je velmi detailní, proto je potřeba ji agregovat na větší celky. Jedním z možných přístupů je agregace dat na mřížku.

Do dalšího postupu vstupují vrstvy (+ vybraný atribut):

- Primární vrstva
    - Dětská hřiště (bodová vrstva) – žádný atribut
- Doplňkové vrstvy
    - Adresní body (bodová vrstva) – atribut *BUDOBYTSL*
    - Vybrané zájmové body OSM (bodová a polygonová vrstva) – atribut *amenity*

7.2 Tvorba modelu

Vrstvy *vybraných zájmových bodů* byly předchystány tak, že společně byly agregovány bodové prvky a zvlášť polygonové. Pro snazší práci v dalších krocích byl také sjednocený atribut obsahující hodnotu kategorie na *amenity*. Je logické, že dětská hřiště se nacházejí velmi často v okolí těchto prvků. Vzhledem k dostupnosti prvků v bodové i polygonové podobě, došlo k úpravě prvků tak, aby byly sjednocené na bodové prvky. Toho bylo dosaženo díky nástroji *Centroids*, který jednotlivé polygony převedl na body. Dále byl postup obdobný jako u předchozích studií. Byla nastavená obalová zóna a byly připojeny atributy k dětským hřištím na základě průniku.

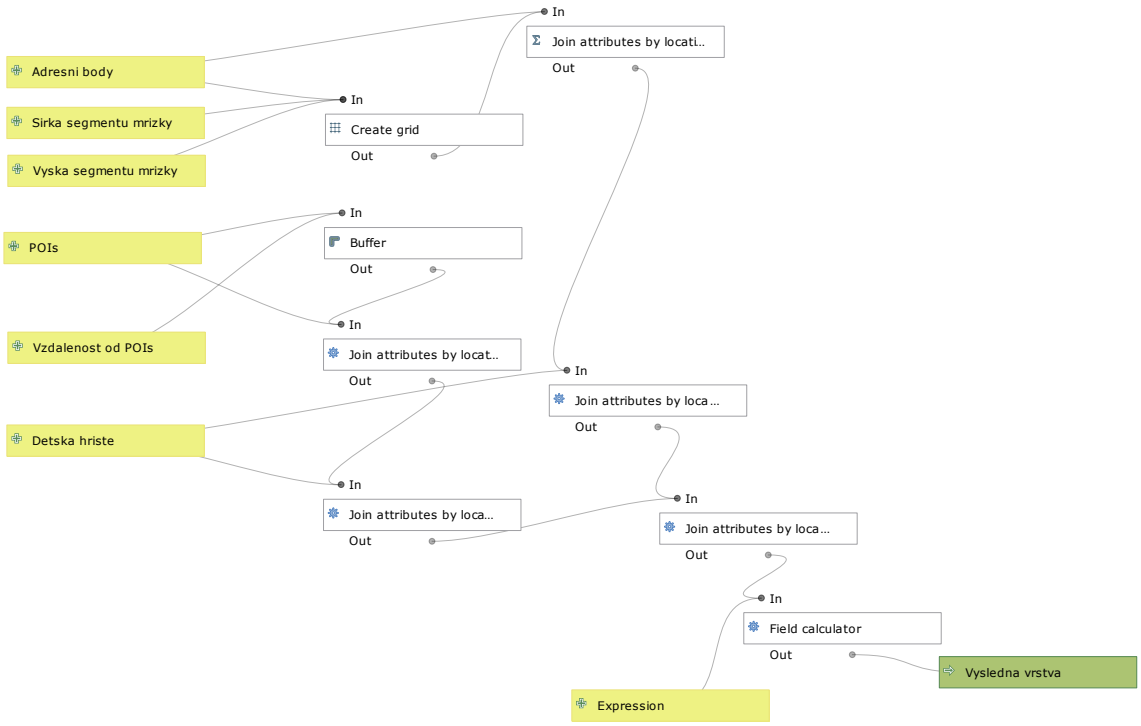
Nad adresními body byla vytvořená mřížka nástrojem *Create Grid* s volitelným rozměrem jednoho segmentu. Následně byly bodové hodnoty vztaženy k jednotlivým segmentům mřížky pomocí *Join attributes by Location (summary).* Samotné sečtené hodnoty jsou numerické, pro asociační pravidla nevhodné. Proto je nutné data zařadit do kategorií, resp. intervalů. K tomu poslouží nově vytvořený atributový sloupec v rámci modelu. Atribut se bude jmenovat *Obyv* a hodnoty byly rozděleny do 3 kategorií tak, aby každá kategorie byla zastoupená rovnoměrně. K tomu bylo využito filtrování záznamů na základě atributu.

Konkrétní podoba podmínky vypadá následovně:

```
CASE
        WHEN "BUDOBYTSL_sum" < 1300 THEN 'nizka'
        WHEN "BUDOBYTSL_sum" > 1300 AND  "BUDOBYTSL_sum" < 5000 THEN
        'stredni'
        ELSE 'vysoka'
END
```

Podmínka je nastavená tak, že pokud počet obyvatel je nižší než 1 300, pak se do nového atributu zapíše hodnota *nizka*. Pokud je počet obyvatel v intervalu od 1 300 do 5 000, tak se zapíše hodnota *stredni*, v jiném případě se zapíše hodnota *vysoka*.

Výsledný model lze vidět na obr. 17, ve kterém jsou zahrnuty všechny nástroje, které byly zmíněny výše. V případě, že je potřeba změnit typ mřížky, stačí poklepat na příslušné pole s tímto nástrojem a typ mřížky změnit. Výchozí mřížkou je čtvercová mřížka. V případě, že není potřeba nic měnit, stačí model spustit, vložit příslušné vrstvy a zadat potřebné hodnoty. Výsledkem je vrstva označená jako *Vysledna vrstva*.



Obr. 17 Použitý model pro hledání pravidel pro dětská hřiště *(zdroj: autorka)*

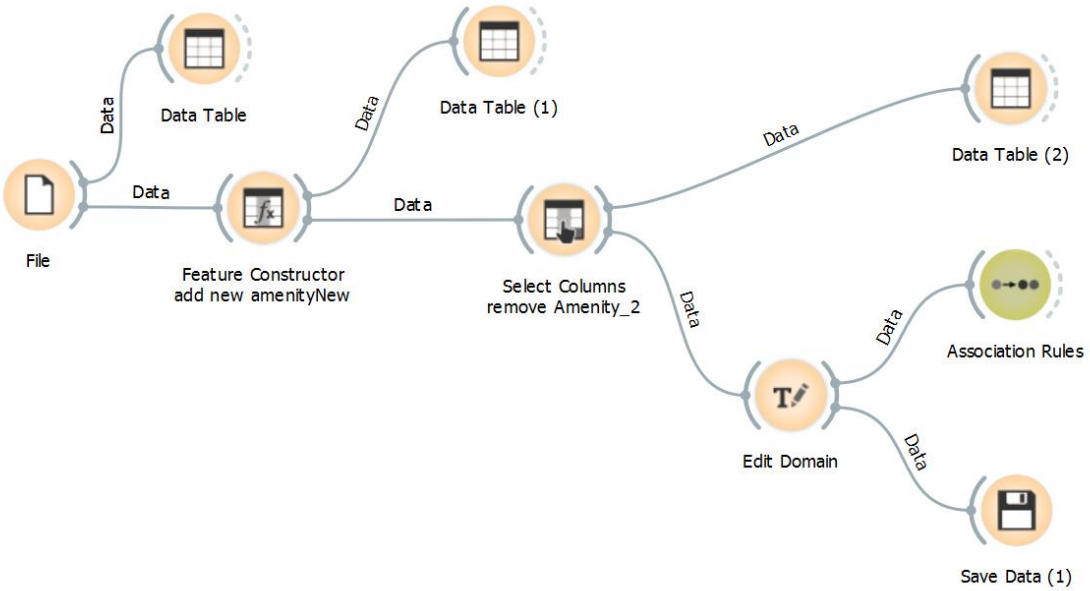
7.3 Úprava dat

Výslednou vrstvu s vybranými záznamy lze vidět v tab. 16. Problém použití dat z OSM v případě lokace zájmových bodů nastává, když školka nebo škola vlastní více jak jednu budovu. Často je v datech zaznamenána každá budova školy, což ve výsledků způsobí, že model nalezne v okolí dětských hřišť více školek či škol, i když se jedná o tutéž, akorát o jinou budovu (záznam ve sloupci *amenity_2*).

Tab. 16 Ukázka výsledných záznamů pro dětská hřiště

OBJECTID	amenity	amenity_2	BUDOBYTSL_sum	Obyv
6	kindergarten	kindergarten	4578	stredni
11	kindergarten	school	9120	wysoka
12	park	park	3203	stredni
17	school	school	10090	wysoka
22	kindergarten	kindergarten	7888	wysoka
24	park	park	5477	wysoka
25	kindergarten	kindergarten	1506	stredni
96	school	kindergarten	10090	wysoka

Pro další práci je nutné tento problém vyřešit a tyto záznamy upravit. K tomu poslouží SW Orange pro generování asociačních pravidel a to tak, že se zde vytvoří model, který lze vidět na obr. 18. Do tohoto modelu vstupuje CSV tabulka, která byla vyexportována z QGIS. Následně se do vrstvy přidá nový sloupec *amenityNew*, do kterého se vepíše hodnota *amenity_2* pod podmínkou, že se nejedná o duplikující hodnotu ze sloupce *amenity*. Dále se původní sloupec *amenity_2* smaže a nahradí se nově vytvořeným. V celém modelu je možné si zobrazit aktuální podobu tabulky pomocí nástroje *Data Table*. Výsledek modelu se automaticky uloží a je možné si pro takto upravená data generovat asociační pravidla.



Obr. 18 Workflow pro úpravu dat a generování asociačních pravidel *(zdroj: autorka)*

7.4 Asociační pravidla

Při hledání pravidel bez ohledu na nastavení min. podpory či spolehlivosti byla nalezena pouze tři asociační pravidla. Všechna pravidla mají spolehlivost 100 %, platí tedy zároveň předpoklad i závěr. První pravidlo říká, že v okolí dětských hřišť se do vzdálenosti 100 m nachází škola a zároveň se dětské hřiště nachází v segmentu mřížky, kde je hustota obyvatel označená jako vysoká, tj. hodnota je vyšší jak 5000 obyv./segment mřížky.

Tab. 17 Vygenerovaná asociační pravidla

Podpora	Spolehlivost	Předpoklad	→	Závěr
0.123	1.000	amenity=school	→	Obyv=vysoka
0.109	1.000	amenity=kindergarten	→	Obyv=vysoka
0.058	1.000	amenity=park	→	Obyv=vysoka

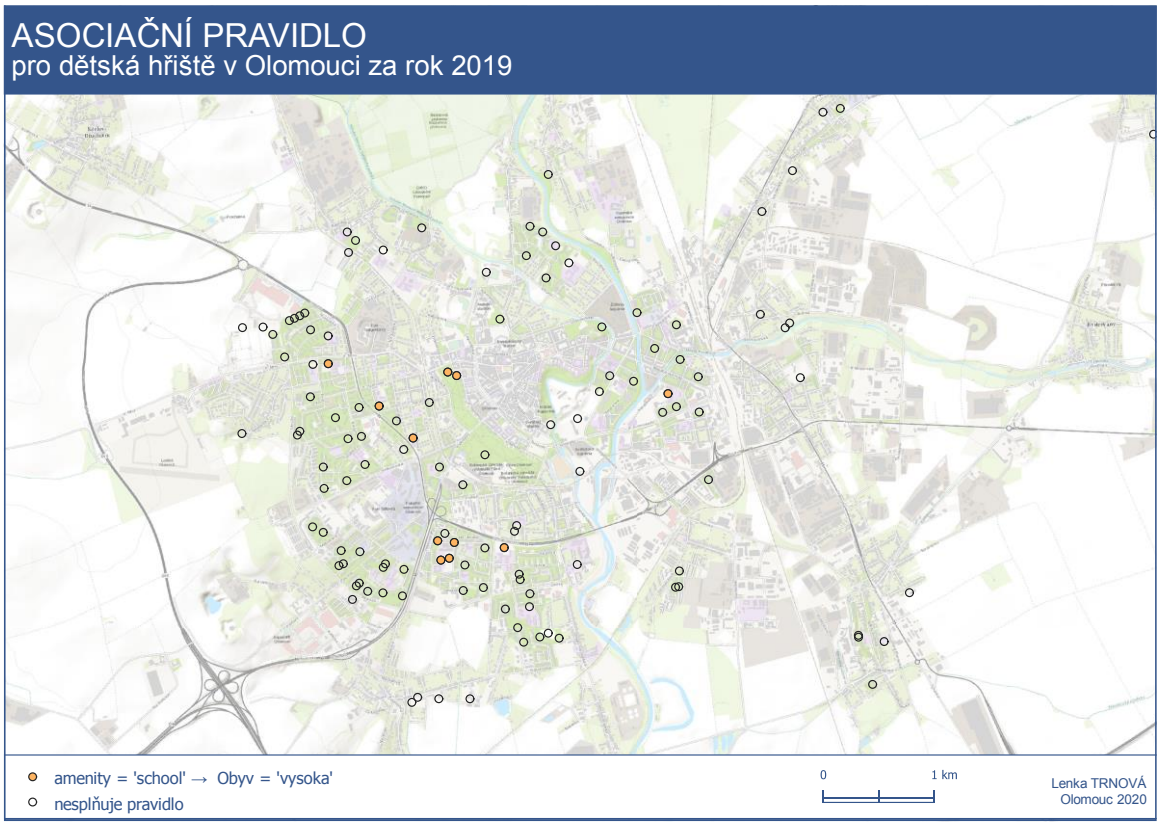
7.5 Vizualizace

Před samotnou vizualizací byla vytvořená tab. 18, sumarizující počet prvků splňující konkrétní pravidlo. Je patrné, že poslední třetí pravidlo splňuje pouze jeden prvek z celé datové sady.

Tab. 18 Počet prvků dětských hřišť splňující konkrétní pravidlo

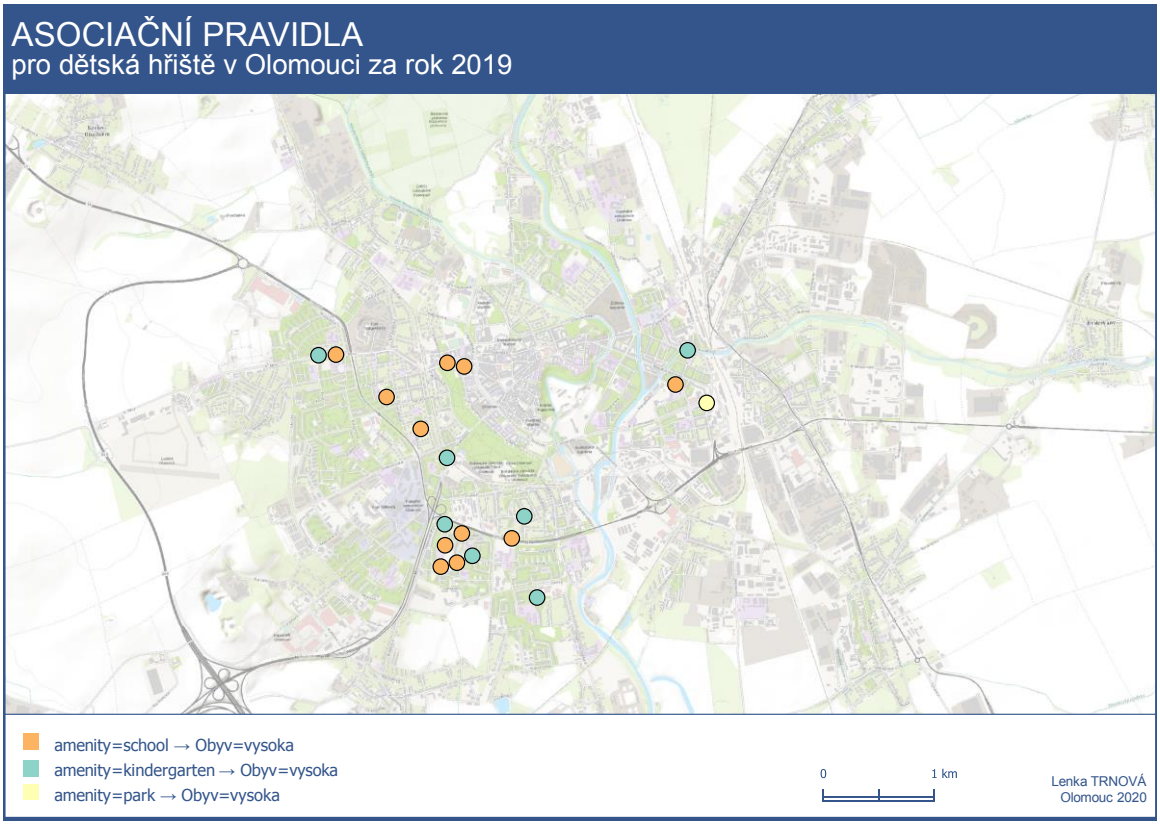
ID	Pravidlo			Počet prvků splňující pravidlo
1	amenity=school	→	Obyv=vysoka	11
2	amenity=kindergarten	→	Obyv=vysoka	7
3	amenity=park	→	Obyv=vysoka	1

Následně byla vytvořená mapa 6 s vybraným prvním asociačních pravidlem.



Mapa 6 Vybrané asociační pravidlo pro dětská hřiště v Olomouci

Zároveň byla opět snaha vytvořit mapu se strukturními diagramy. V tomto případě se ale žádný strukturní diagram neskládá z více než jednoho segmentu – ani jeden z prvků nesplňuje více jak jedno asociační pravidlo.



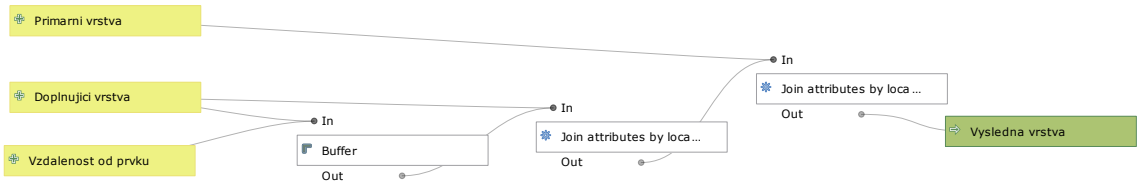
Mapa 7 Strukturní diagram s vybranými asociačními pravidly pro dětská hřiště

8 VÝSLEDKY

Diplomová práce přinesla několik výsledků, které jsou shrnuty v podkapitolách níže. Jeden z nejdůležitějších výsledků je samotný cíl práce – nalézt způsob, jak hledat asociační pravidla pro prostorová data. Toho bylo dosaženo ve všech třech případových studiích.

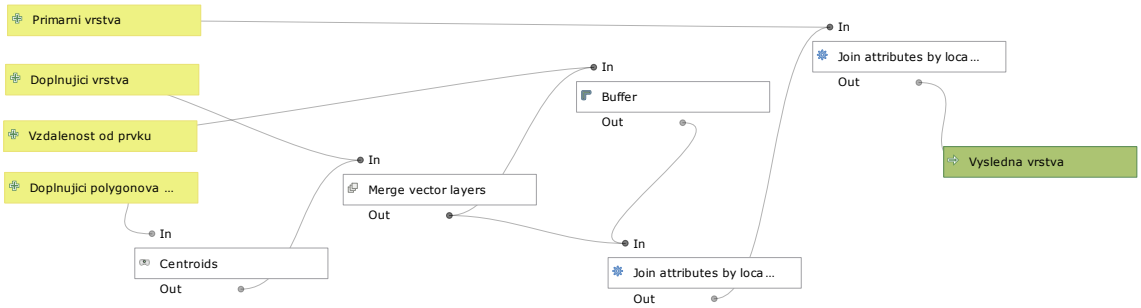
8.1 Modely do QGIS

Výsledkem jsou tři jednoduché modely, které lze opakovaně použít na datové sady dle potřeby. První model je jednodušší – vstupuje do něj pouze primární a doplňující vrstva. Obě vrstvy mohou být libovolné – bodové, liniové či polygonové. Uživatel v dialogovém okně pouze zadá velikost obalové zóny. Výsledkem je *Vysledna vrstva*, která se po skončení modelu přidá do seznamu vrstev.



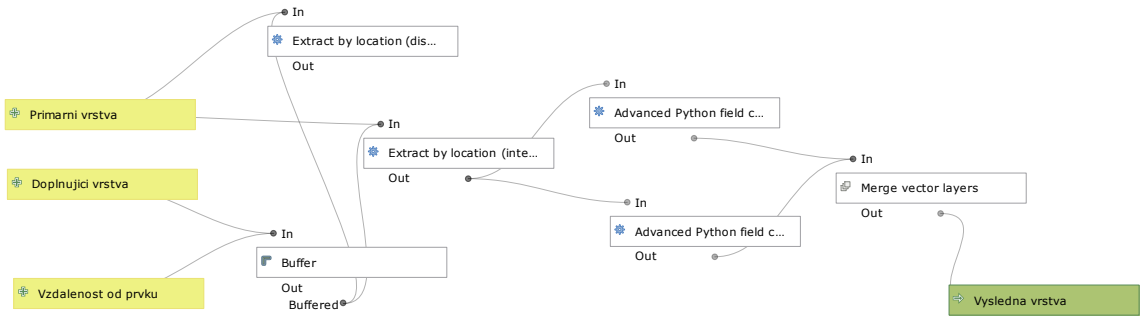
Obr. 19 Jednoduchý model *(zdroj: autorka)*

Druhý model se liší pouze v tom, že je vhodný pro vrstvu, která má dvě podoby – např. bodovou a zároveň polygonovou. Případem takové vrstvy mohou být data z OSM, jako jsou např. zájmové body. Aby nevznikaly nové atributové sloupce zvlášť pro bodovou a polygonovou vrstvu, model převede polygony na body a spojí do jedné vrstvy s bodovou. Uživatel opět pouze zadá velikost obalové zóny a výsledkem modelu je *Vysledna vrstva*.



Obr. 20 Model pro 2 vrstvy stejné tématiky *(zdroj: autorka)*

Třetí model je určený pro výsledný atributový sloupec, ve kterém budou binární hodnoty. Na základě zadané obalové zóny se bude hledat průnik mezi primární a doplňující vrstvou. Pokud se vrstvy protínají, bude do nového atributového sloupce zapsána hodnota 1, v opačném případě 0. Příkladem může být ona testovací sada, ve které se zjišťovalo, zda rostlina roste v blízkosti vody, resp. vodního toku.



Obr. 21 Model s binárními hodnotami *(zdroj: autorka)*

Velkou výhodou těchto modelů je jejich univerzálnost a jednoduchost. V případě, že ve výsledné datové sadě není nalezeno ani jedno frekventované asociační pravidlo, není problémem použít vybraný model znovu. Stačí smazat požadovaný atributový sloupec, který se nahradí jiným. Také je možností žádný z atributů nemazat, pouze jim v SW Orange přiřadit roli k přeskočení.

8.2 Faktory ovlivňující výsledná asociační pravidla

V rámci případových studiích byla snaha sledovat různé faktory, které by výsledná asociační pravidla mohla ovlivnit, popř. je doporučené si na ně dávat pozor.

8.2.1 Detailnost vrstev

Během práce na případových studiích bylo využito datových sad týkajících se využití území. V prvním pokusu s datovými sadami bylo využito Corine Land Cover, což bylo chybné z důvodu měřítka dat. Corine Land Cover má data v měřítku 1 : 100 000, které je pro tuto práci nevhodné. V konečném postupu je datová sada nahrazená využitím území z OSM a v případě Rekol vrstvou od Urban Atlas. Je tedy důležité věnovat pozornost vstupním datům a jejich detailností. V případě, že je použita nevhodná datová sada, má za následek nižší počet pravidel, resp. pravidla s menší četností.

8.2.2 Počet vstupujících datových sad

Dalším ovlivňujícím faktorem je počet vstupujících datových sad do analýzy. Je doporučeno pracovat minimálně se dvěma, resp. třemi datovými sadami. Dvě v případě, že primární datová sada v sobě nese atributovou informaci, kterou lze pro generování asociačních pravidel použít. Pokud je primární datová sada pouze použita pro lokalizaci, je nutno k ní přidat alespoň dvě datové sady, resp. dvě atributové informace. V případě více vstupujících vrstev nedochází k žádnému problému. SW Orange zvládne mezi všemi atributy nalézt ta nefrekventovanější pravidla bez ohledu na počet atributů.

8.2.3 Velikost obalové zóny

Dále byl testován vliv velikosti obalové zóny, kterou ale nelze s jistotou určit. Čím větší obalová zóna, tím pouze větší prostor, ke kterému se připojují atributové informace. V případech, kdy je obalová zóna příliš velká může dojít k problému, že se překrývají plochy např. více typů budov a ve výsledku vznikne více než pouze jeden atributový sloupec s novou informací. Velikost obalové zóny je subjektivní a musí se učit na základě daných dat.

8.3 Další výstupy

Mimo vytvořené modely určené do QGIS byly také vytvořeny další podpůrné nástroje. Mezi nimi je pomocná Google tabulka a Python skript. Závěrem každé případové studie byly také mapové výstupy zobrazující asociační pravidla v prostoru.

8.3.1 Google tabulka

Součástí práce bylo navržení Google tabulky, obsahující vzorce a makra tak, aby byla ulehčená práce a úprava asociačních pravidel na podmínky pro pravidly řízenou symbologii v QGIS. Uživatel pouze nahraje zvolená asociační pravidla a následně se přesune do konkrétního listu, kde nalezne potřebnou úpravu. V rámci step-by-step návodu je vše podrobně popsáno i s praktickou ukázkou.

8.3.2 Skript pro vizualizaci pravidel

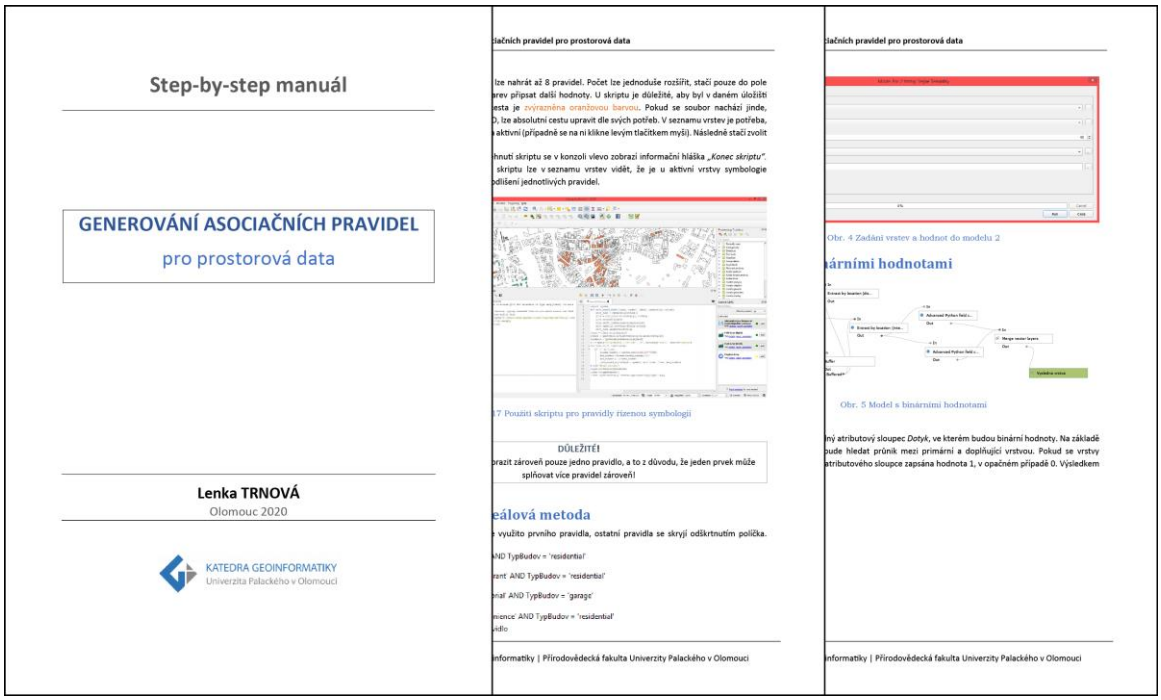
Pro ulehčení práce při tvorbě symbologie asociačních pravidel bylo využito skriptu, která byl vytvořený v rámci praktické části. Díky tomuto skriptu lze teoreticky nahrát najednou např. i 20 pravidel a skript je automaticky převede na podmínky pro jednotlivé symbologie. U skriptu je ale potřeba vzít v potaz, že aktuálně je nastaveno pole obsahující barvy na 8 barev a pokud je počet pravidel vyšší, skript se neprovede.

8.3.3 Mapové výstupy

Součástí práce je také vytvoření jednoduché výsledné mapy pomocí bodové metody. Díky ní lze vždy zobrazit asociační pravidlo společně s prvky, které jej splňují. Jako doplňující prvky zde mohou být zobrazeny ty prvky, které toto pravidlo nesplňují. V rámci přípravy návodu bylo ověřeno, že stejný postup lze použít i na areálovou metodu. V případových studiích byly nalezeny i jiné způsoby vizualizace – např. pomocí strukturních diagramů. Díky nim lze ihned v mapě pozorovat, které prvky splňují více asociačních pravidel zároveň.

8.4 Step-by-step návod

V praktické části byl nalezený způsob, jak pro prostorová data hledat asociační pravidla a zároveň bylo umožněno výsledná pravidla zobrazit zpátky do prostoru. Pro snadnou replikaci postupu byl sepsán step-by-step návod, který detailně popisuje jednotlivé kroky práce. Uživatel si stáhne *ZIP* soubor, který obsahuje vybrané adresáře, ve kterých se nacházejí všechny použité vrstvy v rámci návodu. Zároveň se zde nacházejí jak upravené vrstvy, tak i *CSV* tabulky i ukázka mapového výstupu.



Obr. 22 Ukázka step-by-step návodu

9 DISKUZE

Primárním výsledkem této práce nebylo nalezení nových informací v datové sadě. Cílem bylo nalezení způsobu, jak prostorová data upravit tak, aby byla použitelná pro generování asociačních pravidel. Lze konstatovat, že to se podařilo bez ohledu na přínos nově získaných informací. Byla snaha, aby každá ze tří případových studií byla zaměřená na jiný postup úpravy dat tak, aby pro každou z nich byl vytvořený odlišný model v prostředí QGIS.

U případové studie č. 1 bylo použito dat se stejnou tématikou, jejichž podoba byla jak bodová, tak i polygonová. Aby bylo možné data použít sjednoceně, byly polygony převedeny na body. Body představují centroidy původních polygonů a je proto důležité počítat s tím, že je ovlivněná vzdálenost od těchto prvků.

U případové studie č. 2, týkající se nabíjecích stanic, se liší markantně výsledek, co se blízkosti čerpací stanice týče. Zde je důležité si stanovit správně vzdálenost od čerpacích stanic. V rámci praktické části byla obalová zóna určena subjektivně.

Otázkou může být, zda jsou prostorová asociační pravidla vhodnější než jiné GIS metody nabízené v rámci softwaru. Výhodou generování asociačních pravidel je práce s mnoha datovými sadami bez ohledu na počet záznamů. Lze říct, že díky asociačním pravidlům lze nalézt v datech vazby, které by jinými metodami bylo snadné přehlédnout. Týká se především spolehlivých výjimek, které nejsou frekventované, ale zároveň se vždy nacházejí společně. Díky této metodě je také každá vazba podložená statistickými údaji o tom, v kolika procentech záznamů, kterých se to týká, tato vazba platí. Není pochyb, že hledání spolehlivých výjimek by bez těchto pravidel bylo takřka nemožné.

Součástí práce bylo vytvoření návodu, jak prostorová data upravit pro možnost hledání asociačních pravidel. Očekává se, že takto vytvořený návod bude sloužit primárně studentům KGI, kteří s prostorovými daty pracují. Byla snaha udělat návod co nejpodrobnější, aby bylo od začátku až do konce jasné, jak se má postupovat. Jedinou dá se říct nevýhodou je, že návod je mířen na použití QGIS jako SW pro úpravu dat a zároveň byla během samotného postupu nastavená anglická lokalizace uživatelského prostředí. Celý postup je možné alternativně zpracovat i v komerčním řešením, kterým může být ArcGIS for Desktop, popř. ArcGIS Pro.

Diplomová práce řeší nový způsob úpravy prostorových dat, lze na to dále navázat a vylepšit. Do budoucna by bylo možné práci rozšířit o použití mnohem kvalitnějších dat, co se poskytovaných informací týče. V rámci práce bylo použito takových dat, která jsou dostupná volně na internetu, popř. byla poskytnutá katedře v rámci řešení jiné bakalářské či diplomové práce. Místo Google tabulky by bylo možné napsat vlastní skript, který by dokázal stejnou funkcionalitu s makry a vzorcemi. Jedná se ale již o pokročilejší programování.

10 ZÁVĚR

Cílem diplomové práce bylo nalézt funkční způsob úpravy prostorových dat tak, aby bylo možné pro ně naleznout asociační pravidla. Asociační pravidla by měla primárně přinášet informace o vazbách mezi prvky, které nemusí být na první pohled zřetelné.

Teoretická část práce se zabývá hledáním dostupných odborných prací, které se tomuto problému věnovaly. Zároveň zde byla snaha nalézt co nejvíce programů, které umožňují generovat asociační pravidla. Takové programy umějí pracovat pouze s neprostorovými daty, proto bylo nutné hledat možnou předúpravu prostorových dat.

V praktické části byly otestovány vybrané programy pomocí vytvořených testovacích dat. Dále byly hledány vhodné datové sady a jejich kombinace tak, aby bylo možné získat co nejzajímavější asociační pravidla. Na základě vybraných datových sad byly vytvořené celkem tři případové studie, které využívají vlastní připravený model na úpravu prostorových dat.

Výsledkem takové úpravy je jedna atributová tabulka, kterou lze bez problému nahrát do vybraného SW Orange. V něm došlo ke hledání asociačních pravidel, které lze díky pomocné Google tabulce a předchystanému PyQGIS skriptu nahrát zpět do prostoru. Pravidla zde vystupují jako podmínky pro symbologii a v mapovém výstupu lze zobrazit prvky, které dané pravidlo splňují, popř. naopak nesplňují. Také byl nalezen jiný způsob vizualizace, a to tvorba strukturního diagramu. Díky němu je možné zobrazit více asociačních pravidel zaráz a zároveň lze získat informaci o tom, kolik pravidel splňuje konkrétní prvek.

Na základě nalezeného postupu byl vytvořený step-by-step návod, který sumarizuje celý nalezený postup od úpravy prostorových dat až po výsledný mapový výstup. Součástí návodu jsou celkem tři modely, které lze opakovaně využít pro konkrétní úpravu dat.

Cíl práce byl na základě tří případových studií a vytvořeného návodu splněný. Díky návodu je možné upravit vlastní vybrané datové sady a lze s nimi dále pracovat.

POUŽITÁ LITERATURA A INFORMAČNÍ ZDROJE

AGGARWAL, Niyati, Amit KUMAR, Harsh KHATTER a Vaishali AGGARWAL. Analysis the effect of data mining techniques on database. Advances in Engineering Software [online]. 2012, 47(1), 164-169 [cit. 2019-06-24]. DOI: 10.1016/j.advengsoft.2011.12.013. ISSN 09659978

ALCALÁ-FDEZ, J., L. SÁNCHEZ, S. GARCÍA, et al. KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing [online]. 2009, 13(3), 307-318 [cit. 2019-08-26]. DOI: 10.1007/s00500-008-0323-y. ISSN 1432-7643.

Association Rules — Orange3-Associate 1 documentation. [online]. Dostupné z: https://orange3-associate.readthedocs.io/en/latest/widgets/associationrules.html

Association Rules Analysis on FP-Growth Method in Predicting Sales. International Journal of Recent Trends in Engineering and Research [online]. 2017, 3(10), 58-65 [cit. 2019-09-03]. DOI: 10.23883/IJRTER.2017.3453.DHCOA. ISSN 24551457

DAO, Diep. Rule Learning for Spatial Data Mining. Geographic Information Science & Technology Body of Knowledge [online]. 2018, 2018(Q1) [cit. 2019-09-03]. DOI: 10.22224/gistbok/2018.1.3.

DEMŠAR, Janez, Blaž ZUPAN, Gregor LEBAN a Tomaz CURK. Orange: From Experimental Machine Learning to Interactive Data Mining. BOULICAUT, Jean-François, Floriana ESPOSITO, Fosca GIANNOTTI a Dino PEDRESCHI, ed. Knowledge Discovery in Databases: PKDD 2004 [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, 2004, s. 537-539 [cit. 2019-09-02]. Lecture Notes in Computer Science. DOI: 10.1007/978-3-540-30116-5_58. ISBN 978-3-540-23108-0.

FARIDI, Mainaz, Seema VERMA a Saurabh MUKHERJEE. Integration of GIS, Spatial Data Mining, and Fuzzy Logic for Agricultural Intelligence. PANT, Millie, Kanad RAY, Tarun K. SHARMA, Sanyog RAWAT a Anirban BANDYOPADHYAY, ed. Soft Computing: Theories and Applications [online]. Singapore: Springer Singapore, 2018, 2018-11-25, s. 171-183 [cit. 2019-06-24]. Advances in Intelligent Systems and Computing. DOI: 10.1007/978-981-10-5687-1_16. ISBN 978-981-10-5686-4.

HAMDAD, Leila, Amine ABDAOUI, Nabila BELATTAR a Mohamed AL CHIKHA. EasySDM - An Integrated and Easy to Use Spatial Data Mining Platform. In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management [online]. SCITEPRESS - Science and and Technology Publications, 2015, 2015, s. 394-401 [cit. 2019-09-03]. DOI: 10.5220/0005615903940401. ISBN 978-989-758-158-8

HOLSHEIMER, M. – Siebes,A. The search for knowledge in databases. Tech.Rep. CSR9406, CWI, Amsterdam, 1994

HU, Yujie, Yu ZHANG a Kyle S. SHELTON. Where are the dangerous intersections for pedestrians and cyclists: A colocation-based approach. Transportation Research Part C: Emerging Technologies [online]. 2018, 95, 431-441 [cit. 2019-10-15]. DOI: 10.1016/j.trc.2018.07.030. ISSN 0968090X.

HŮLOVÁ, Hana. Aplikace vybraných metod prostorového dolování dat v databázových systémech. Plzeň, 2010. Diplomová práce. Západočeská univerzita v Plzni. Fakulta aplikovaných věd. Vedoucí práce Ing. Karel JANEČKA, Ph.D.

CHATTAMVELLI, Rajan. Data mining algorithms. Oxford, U.K.: Alpha Science International, [2011], xvii stran, 424 různě číslovaných. ISBN 978-1-84265-684-6.

CHEN, Junming, Guangfa LIN a Zhihai YANG. Extracting spatial association rules from the maximum frequent itemsets based on Boolean matrix. In: 2011 19th International Conference on Geoinformatics [online]. IEEE, 2011, 2011, s. 1-5 [cit. 2019-06-05]. DOI: 10.1109/GeoInformatics.2011.5980870. ISBN 978-1-61284-849-5.

IBM Knowledge Center. [online]. Copyright © Copyright IBM Corp. 2017 [cit. 02.09.2019]. Dostupné z: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/product_landing.html

JAYABABU, Y., G.P.S. VARMA a A. GOVARDHAN. Incremental topological spatial association rule mining and clustering from geographical datasets using probabilistic approach. Journal of King Saud University – Computer and Information Sciences [online]. 2018, 30(4), 510-523 [cit. 2019-06-24]. DOI: 10.1016/j.jksuci.2016.12.006. ISSN 13191578.

KOPERSKI, Krzysztof a Jiawei HAN. Discovery of spatial association rules in geographic information databases. EGENHOFER, Max J. a John R. HERRING, ed. Advances in Spatial Databases [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, 1995-6-1, s. 47-66 [cit. 2019-06-24]. Lecture Notes in Computer Science. DOI: 10.1007/3-540-60159-7_4. ISBN 978-3-540-60159-3.

KUMAR, G.Kiran, P.Premchand P.PREMCHAND a T.Venu GOPAL. Mining Of Spatial Co-location Pattern from Spatial Datasets. International Journal of Computer Applications [online]. 2012, 42(21), 25-30 [cit. 2019-06-05]. DOI: 10.5120/5836-7994. ISSN 09758887.

LAUBE, Patrick, Mark de BERG a Marc VAN KREVELD. Spatial Support and Spatial Confidence for Spatial Association Rules. RUAS, Anne a Christopher GOLD, ed. Headway in Spatial Data Handling [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, 2008, s. 575-593 [cit. 2019-06-13]. Lecture Notes in Geoinformation and Cartography. DOI: 10.1007/978-3-540-68566-1_33. ISBN 978-3-540-68565-4.

LEE, Ickjai. Mining Multivariate Associations within GIS Environments. ORCHARD, Bob, Chunsheng YANG a Moonis ALI, ed. Innovations in Applied Artificial Intelligence [online].

Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, 2004, s. 1062-1071 [cit. 2019-06-13]. Lecture Notes in Computer Science. DOI: 10.1007/978-3-540-24677-0_109. ISBN 978-3-540-22007-7.

LI, Deren, Shuliang WANG a Deyi LI. Spatial Data Mining [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015 [cit. 2019-12-04]. DOI: 10.1007/978-3-662-48538-5. ISBN 978-3-662-48536-1.

MALÁ, Markéta. Asociační pravidla v dataminingových úlohách. Liberec, 2015. Bakalářská práce. Technická univerzita v Liberci. Fakulta mechatroniky, informatiky a mezioborových studií.

MALÁ, Markéta. Vzory chování ukryté v provozních datech. Liberec, 2017. Diplomová práce Technická univerzita v Liberci. Fakulta mechatroniky, informatiky a mezioborových studií.

MATONOHA, Tomáš. Analýza dat ze studentských dotazníků. Olomouc, 2013. Přírodovědecká fakulta Univerzity Palackého v Olomouci. Katedra informatiky.

MENNIS, Jeremy a Jun Wei LIU. Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS [online]. 2005, 9(1), 5-17 [cit. 2019-06-13]. DOI: 10.1111/j.1467-9671.2005.00202.x. ISSN 1361-1682.

NEHRI, Ruhi a Meghana NAGORI. Spatial Co-location Patterns Mining. International Journal of Computer Applications [online]. 2014, 93(12), 21-25 [cit. 2019-06-05]. DOI: 10.5120/16267-5994. ISSN 09758887.

NESIBA, Marek. Analýzy vztahů. Brno, 2019. Diplomová práce. Masarykova univerzita. Přírodovědecká fakulta. Vedoucí práce RNDr. Radim Navrátil, Ph.D.

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

RANGRA, K. a K. L. BANSAL. "Comparative Study of Data Mining Tools," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 6, JUNE 2014.

RONCEK, Dennis W. a PAMELA A. MAIER. Bars, Blocks, and Crime Revisited: Linking the Theory of Routine Activities to the Empiricism of Hot Spots. Criminology [online]. 1991, 29(4), 725-753 [cit. 2019-06-05]. DOI: 10.1111/j.1745-9125.1991.tb01086.x. ISSN 0011-1384.

SADAT, Yousef Kanani, Tina NIKAEIN a Farid KARIMIPOUR. FUZZY SPATIAL ASSOCIATION RULE MINING TO ANALYZE THE EFFECT OF ENVIRONMENTAL VARIABLES ON THE RISK OF ALLERGIC ASTHMA PREVALENCE. Geodesy and Cartography [online]. 2015, 41(2), 101-112 [cit. 2019-12-04]. DOI: 10.3846/20296991.2015.1051339. ISSN 2029-6991.

SENFT, Martin. Vytěžování databáze Poradny pro poruchy metabolismu. Praha, 2014. Diplomová práce. Univerzita Karlova. Filozofická fakulta. Vedoucí práce prof. RNDr. Jiří Ivánek, CSc.

SHEKHAR, Shashi. Encyclopedia of GIS. 2nd edition. New York, NY: Springer Berlin Heidelberg, 2017. ISBN 978-3-319-17885-1.

SHI, Wenzhong, Anshu ZHANG a Geoffrey I. WEBB. Mining significant crisp-fuzzy spatial association rules. International Journal of Geographical Information Science [online]. 2018, 32(6), 1247-1270 [cit. 2019-09-15]. DOI: 10.1080/13658816.2018.1434525. ISSN 1365-8816.

SLIMANI, Thabet a Amor LAZZEZ. Efficient Analysis of Pattern and Association Rule Mining Approaches. International Journal of Information Technology and Computer Science [online]. 2014, 6(3), 70-81 [cit. 2019-08-19]. DOI: 10.5815/ijitcs.2014.03.09. ISSN 20749007.

SUKAESIH SITANGGANG, Imas. Spatial Multidimensional Association Rules Mining in Forest Fire Data. Journal of Data Analysis and Information Processing [online]. 2013, 01(04), 90-96 [cit. 2019-06-06]. DOI: 10.4236/jdaip.2013.14010. ISSN 2327-7211.

SUZUKI, Einoshin a Yves KODRATOFF. Discovery of surprising exception rules based on intensity of implication. ŻYTKOW, Jan M. a Mohamed QUAFAFOU, ed. Principles of Data Mining and Knowledge Discovery [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, 1998-10-19, s. 10-18 [cit. 2019-08-20]. Lecture Notes in Computer Science. DOI: 10.1007/BFb0094800. ISBN 978-3-540-65068-3.

SYPION-DUTKOWSKA, Natalia. Land use impact on spatial distribution of crime. Journal of Psychology & Psychotherapy [online]. 2018, 08 [cit. 2019-06-05]. DOI: 10.4172/2161-0487-C2-026. ISSN 21610487. Dostupné z: https://www.omicsonline.org/conference-proceedings/forensic-psychology-dual-diagnosis-2018-tracks.digital

ŠARMANOVÁ, Jana. Metody analýzy dat: učební text. Ostrava: Vysoká škola báňská - Technická univerzita Ostrava, 2012. ISBN 978-80-248-2565-6.

TANIAR, David, Wenny RAHAYU, Vincent LEE a Olena DALY. Exception rules in association rule mining. Applied Mathematics and Computation [online]. 2008, 205(2), 735-750 [cit. 2019-08-20]. DOI: 10.1016/j.amc.2008.05.020. ISSN 00963003.

VERSICHELE, Mathias, Liesbeth DE GROOTE, Manuel CLAEYS BOUUAERT, Tijs NEUTENS, Ingrid MOERMAN a Nico VAN DE WEGHE. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. Tourism Management [online]. 2014, 44, 67-81 [cit. 2019-06-24]. DOI: 10.1016/j.tourman.2014.02.009. ISSN 02615177.

VOJÍŘ, Stanislav, Václav ZEMAN, Jaroslav KUCHAŘ a Tomáš KLIEGR. EasyMiner.eu: Web framework for interpretable machine learning based on rules and frequent itemsets. Knowledge-Based Systems[online]. 2018, 150, 111-115 [cit. 2019-07-15]. DOI: 10.1016/j.knosys.2018.03.006. ISSN 09507051.

What is association rules (in data mining)? - Definition from WhatIs.com. Business Analytics/Business Intelligence information, news and tips – SearchBusinessAnalytics [online].
Dostupné z: https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

WITTEN, Ian H., Eibe FRANK a Mark A. HALL. Algorithms. Data Mining: Practical Machine Learning Tools and Techniques [online]. Elsevier, 2011, 2011, s. 85-145 [cit. 2019-06-06]. DOI: 10.1016/B978-0-12-374856-0.00004-3. ISBN 9780123748560.

YOO, Jin Soung a Mark BOW. Mining spatial colocation patterns: a different framework. Data Mining and Knowledge Discovery [online]. 2012, 24(1), 159-194 [cit. 2019-06-05]. DOI: 10.1007/s10618-011-0223-0. ISSN 1384-5810.

YOO, Jin Soung, S. SHEKHAR a M. CELIK. A Join-Less Approach for Co-Location Pattern Mining: A Summary of Results. In: Fifth IEEE International Conference on Data Mining (ICDM'05) [online]. IEEE, 2005, s. 813-816 [cit. 2019-06-05]. DOI: 10.1109/ICDM.2005.8. ISBN 0-7695-2278-5.

YUE, Han, Xinyan ZHU, Xinyue YE a Wei GUO. The Local Colocation Patterns of Crime and Land-Use Features in Wuhan, China. ISPRS International Journal of Geo-Information [online]. 2017, 6(10) [cit. 2019-06-05]. DOI: 10.3390/ijgi6100307. ISSN 2220-9964.

PŘÍLOHY

SEZNAM PŘÍLOH

Vázané přílohy:

Příloha 1 Model vytvořený pro nabíjecí stanice

Volné přílohy:

Příloha 2 Poster

Příloha 3 CD

Elektronické přílohy:

Příloha 4 Step-by-step návod

Příloha 5 Leaflet mapa pro nabíjecí stanice

Popis struktury CD

Adresáře:

Přílohy

Poster

Step-by-step návod

Leaflet mapa pro nabíjecí stanice

Adresář přiložený k step-by-step-návodu

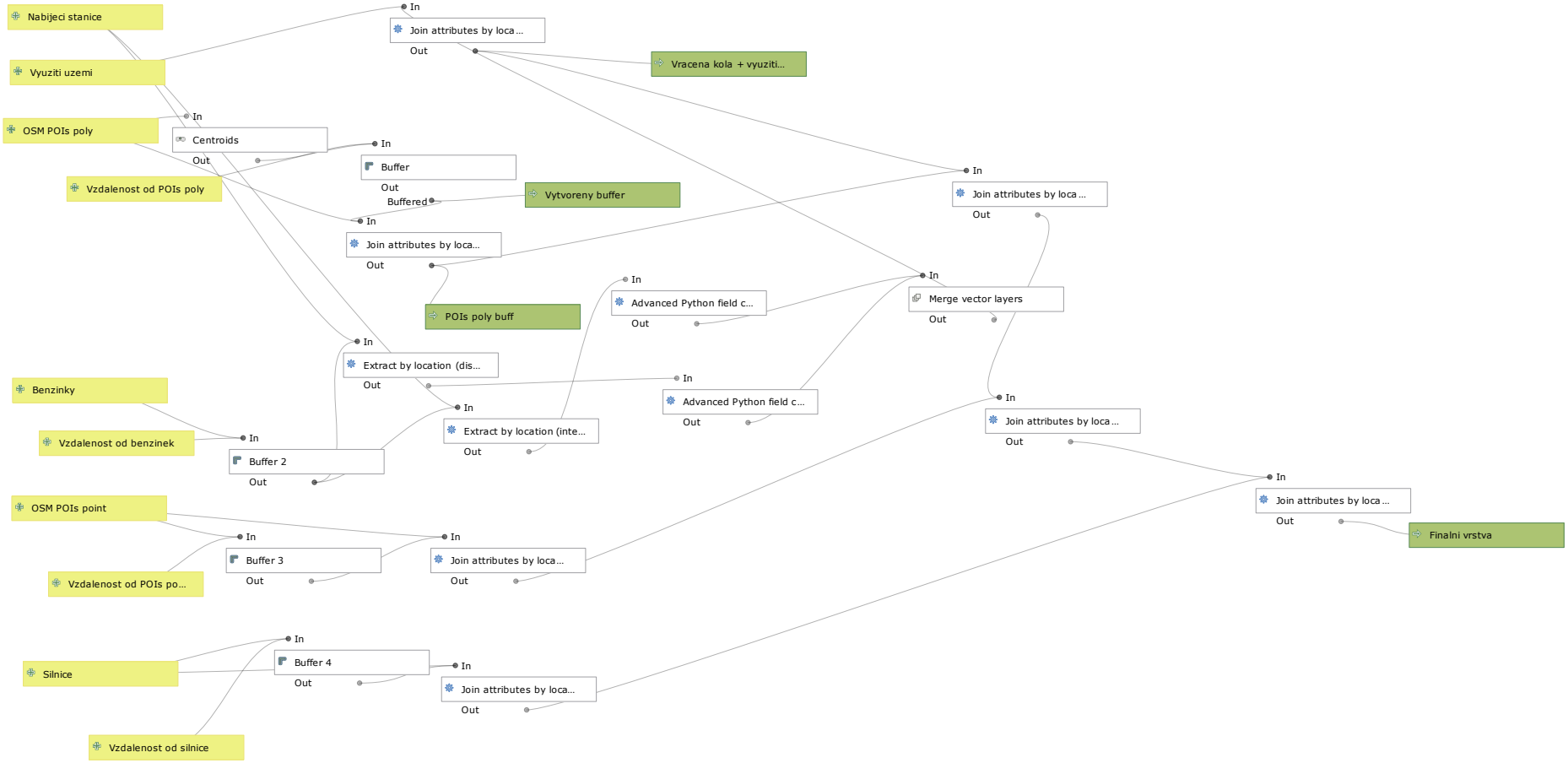
Data

- Podadresáře pro 3 případové studie

Web

PyQGIS skript

Text práce



Obr. 23 Model vytvořený pro nabíjecí stanice *(zdroj: autorka)*