Správa, analýza a prezentace zdravotnických prostorových dat pomocí R

Vojtěch Cícha[1]

[1]Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci,
Tř. 17. listopadu 50, 771 46, Olomouc, Česká Republika
vojtech.cicha@gmail.com

Abstrakt. Svobodné programové prostředí R nabývá poslední dobou na popularitě i v prostředí GIS. Můžou za to jeho mnohé výhody jako například rozsáhlá funkcionalita, mocné nástroje, možnost kontroly všech detailů, opakovatelnost postupů, snadná rozšiřitelnost, efektní i efektivní grafické výstupy a v neposlední řadě také žádné pořizovací náklady. Tato práce po úvodní teoretické části prochází nad reálnými prostorovými daty běžnými geoinformatickým operacemi jako je vstup dat, transformace souřadnic, prostorové i atributové napojování dat, jejich agregace či dotazové výběry. Dále zkouší statisticko-analytický potenciál R při zkoumání neprostorových závislostí i prostorových vztahů pomocí některých analýz. Důraz je kladen na popis a ukázky různých možností vizualizace. Nejvíce se soustředí na běžné statické výstupy, dále ale také na animační či externí výstupy do programů jako např. Google Earth. Součástí práce je popsaný programový kód, který se může stát návodem při podobných prostorových aktivitách v R.

Klíčová slova: R, R project, prostorová data v R, vizualizace v R, prostorové analýzy v R

Abstract. Free software environment R is recently becoming more popular even in a GIS. This is happening thanks to many advantages as the wide variety of functions, the powerful tools, the possibility of taking control of all details, the repeatability of procedures, the easy way to extending, the effective and efficient graphic outputs and also none acquisition costs. After an introductory theoretical part, this work uses a real spatial data to go throught common geoinformatics operations like a data import, a transformation of coordinates, both spatial and atributive joins, a data aggregation or a querying. It also tests statistical-analytical potential of the R by examination non-spatial dependencies or spatial relationships using certain analysis. The emphases is on description and demonstration of visualisation capabilities. It focus the most on common static outputs, and also on animation or external outputs to other programs, for example Google Earth. The description of a programming code, which can help as manual to some similar spatial activities in R, is included in the work.

Keywords: R, Project, spatial data in R, vizualization in R, spatial analysis in R

1 Cíle práce

Cílem bakalářské práce je na základě dostupné literatury vypracovat pojednání o možnostech správy, analýzy a prezentace zdravotnických prostorových dat s využitím open source programu R a jeho Graphical User Interface (GUI). Jednotlivé části bakalářské práce budou obsahovat charakteristiku programového prostředku včetně souhrnu využitých balíků, popis metod analýzy prostorových dat a možnosti jejich vizualizace. Součástí práce bude také rešerše zabývající se využitím zdravotnických dat s prostorovou složkou a jejich analýzám. V neposlední řadě budou provedeny případové studie na datech dostupných na Katedře geoinformatiky. Formou těchto studií budou shrnuty veškeré teoreticky nastudované postupy a procesy od prvotní úpravy dat, přes jejich geokódování a prostorovou i neprostorovou analýzu, až po vizualizaci. Jako shrnutí bakalářské práce bude vytvořen i poster.

2 Představení R

“R je jazykem a prostředím pro statistické výpočty a grafiku, dostupný jako volně šiřitelný software (Free Software) při dodržení podmínek GNU General Public License nadace Free Software Foundation. (...) Mnozí uživatelé si pod R představují statistický software. Je však vhodnější si pod R představit prostředí, v němž jsou kromě jiného implementovány i statistické metody” [4]. R má svůj původ v jazyku a prostředí zvaném S, který byl vyvinut týmem kolem Johna Chamberse. Mezi jazyky R a S jsou sice rozdíly, ale velká část kódu pro S je možné využít i v prostředí R.
Jednou z největších výhod R je jeho velmi jednoduchá rozšiřitelnost pomocí balíků (package). Seznam nejdůležitějších balíků je v tabulce č. 1.

Tabulka 1. Nejdůležitější balíky využité v bakalářské práci

názvy balíků	oblast působení
ggplot2, ggmap, animation, mapplots	grafické výstupy
sp, rgdal, rgeos	práce s prostorovými daty
spatstat, spdep, splancs	prostorové analýzy
plotKML	export do prostředí Google Earth
RColorBrewer	práce s barvami

Výchozí grafické uživatelské rozhraní (Graphical User Interface - GUI) R nepůsobí na běžného uživatele příliš přívětivě. Uživatelskou základnou jsou proto velmi oblíbené GUI jako např. RStudio[2].

[2] http://www.rstudio.com/

3 Data

Teoretické poznatky v této bakalářské práci se budou v praxi testovat na dvou datových souborech. První z nich pochází z programu EPIDAT. Jedná se o celonárodní databázi povinného hlášení výskytů infekčních onemocnění. Mezi roky 1993 - 2010 bylo celkově evidováno přes 2,5 milionu výskytů nemocí [3]. Konkrétní data poskytla pro účely práce Krajská hygienická stanice (dále KHS) Olomouckého kraje. Jsou to záznamy z let 2002 - 2011 týkající se území Olomouckého kraje. Ze zhruba desítky infekčních nemocí byla vybrána diagnóza označená A08 - virové střevní infekce. Důvodem byl pro analýzy vhodný počet výskytů - jde o běžné onemocnění jehož výskyt není extrémní vzhledem k raritním ani nadměrným výskytům. Ze záznamů byly v rámci ochrany soukromí a lékařského tajemství odstraněny údaje o nakaženém jako rodné číslo, jméno, příjmení, věk a přesná adresa (myšleno číslo domu, informace o ulici a městu byly v zájmu zachování prostorového aspektu infekce zachovány). Každý záznam obsahuje dále informaci o datu hlášení (vstupu do databáze), pohlaví nemocného, kolektivu, ve kterém se infikovaný pohybuje, zaměstnání, příslušnost k etniku a další údaje, které však v této práci nejsou využity.
Druhý datový soubor rovněž poskytla KHS Olomouc, jedná se o data vzniklá na podzim roku 2012 v období metanolové aféry. Jsou to záznamy z kontrol plnění opatření Ministerstva zdravotnictví ve stravovacích objektech, opět na území Olomouckého kraje. Tento datový soubor obsahuje informace o datu kontroly, názvu a adrese kontrolovaného objektu, informace o orgánu, který kontrolu vykonal a o výsledku kontroly, tedy jestli byl nalezen přestupek pro daný den platného bezpečnostního opatření.

4 Teoretické cíle

V této části práce jsou rozebírány možnosti správy, analýzy a prezentace. Obecně se metody GIS dělí do čtyř základních funkčních kategorií: vstup, správa, analytické zpracování a prezentace výsledků. V rámci této práce však do správy budou zařazeny všechny přípravné operace s datovými sadami a objekty tak, aby byly připraveny na samotné analytické zpracování. To znamená, že kromě vstupu a výstupu dat bude řešeno geokódování a prostorové operace jako transformace souřadnic, ořez, prostorový join nebo agregace dat. V rámci analýzy pak proběhnou operace, na jejichž základě bude možno data hodnotit, odhadovat pravidelnost či naopak nahodilost jejich chování. Dojde na statistické numerické analýzy jako základní popisná charakteristika datových souborů nebo testování hypotéz, na grafické metody jako boxplot a na prostorové analýzy pro zjišťování typů prostorových procesů a hodnocení prostorového rozložení dat. Výsledky analýz i správy je potřeba nějak graficky prezentovat, to se bude řešit v rámci třetí části, kterou je vizualizace dat.

5 Případové studie

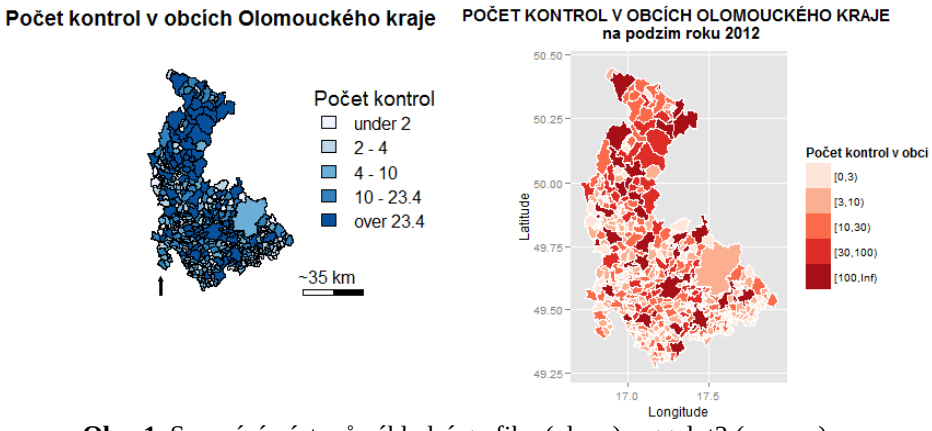
Teoretické cíle vytyčené v předchozí kapitole jsou nyní v praxi vyzkoušeny na reálných datech. Důležitou součástí této kapitoly je k bakalářské práci přiložený programový kód z R s popisem všech kroků.

5.1 Data z metanolových kontrol

Cílem této případové studie je vyzkoušet geokódování a nezbytné operace pro vyzkoušení možností vizualizace.

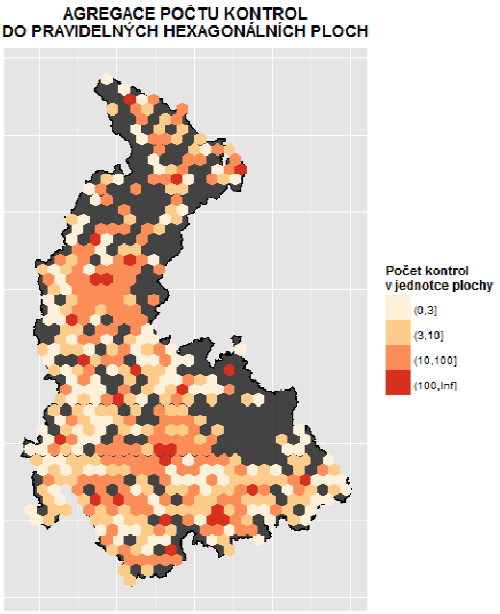
Geokódování. Z původních 13452 záznamů se pomocí webové geokódovací aplikace MAPY API společnosti mapy.cz ve výsledku geokódovalo 13359 záznamů na území Olomouckého kraje. Samotnému geokódování předcházely časově náročné manuální úpravy dat, ve kterých se adresní atributy „čistily“ od překlepů nebo údajů do adresy nepatřících. Dalším řešeným problémem byla nejednoznačnost obce. Pro absenci údaje o okresu byly často výsledky nesprávně lokalizovány po celém území ČR, někdy i v zahraničí.

Vizualizace. Geokódované záznamy jsou nyní v prostředí různě vizualizovaný. Základem vizualizace je systém balíku ggplot2. Tento způsob vizualizace oproti běžné grafice R má velkou výhodu v estetičtějších výstupech, nevýhodou je nepřímá spolupráce s prostorovými daty (objekty prostorových tříd je nutno nejprve převést do podoby běžných datasetů, které navíc obsahují prostorové souřadnice). Po této operaci jsou popsány některé možnosti vizualizace jako např. základní kartogramové metody, vizualizace s podkladem z Google Maps[3] nebo OpenStreetMap[4] nebo metoda biningu (agregace do pravidelných ploch, viz obr. 2). R nabízí i možnosti vytváření dynamických výstupů ve formátu gif či export výstupů do prostředí Google Earth[5].



Obr. 1. Srovnání výstupů základní grafiky (vlevo) a ggplot2 (vpravo)

[3] https://maps.google.com/
[4] http://www.openstreetmap.org/
[5] http://www.google.cz/intl/cs/earth/



Obr. 2. Metoda biningu pomocí balíku ggplot2

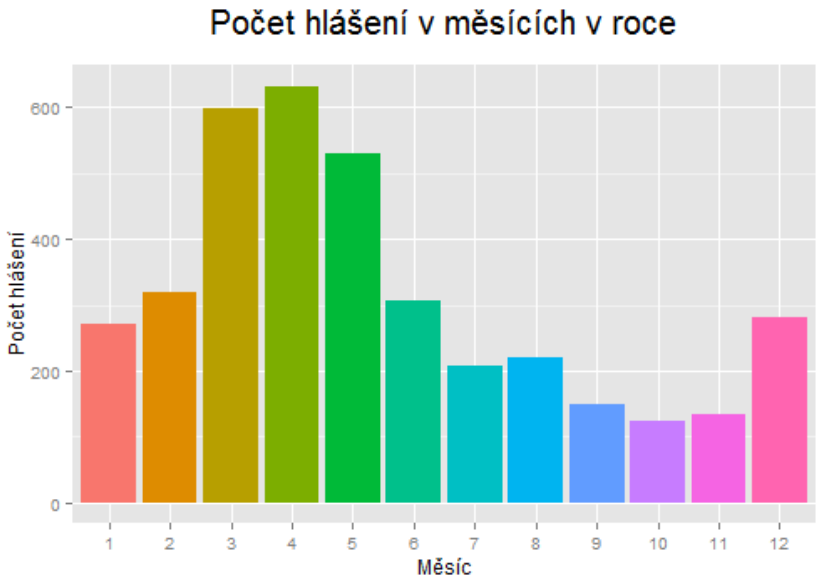
5.2 Epidemiologická data

Cílem případové studie nad epidemiologickými daty je vyzkoušení procesu geokódování a následné metody statistické a prostorové analýzy. Důraz však je kladen na samotné provedení v R, ne na vysvětlení teorie statistických metod.

Geokódování. Z původních 3836 záznamů jich 3775 bylo úspěšně lokalizováno na území Olomouckého kraje. Výhodou tohoto datasetu byl atribut okresu, který odstraňoval nejednoznačnost výsledků.

Statistické analýzy. Kromě základní popisné charakteristiky (průměr, medián, extrémy, směr. odchylka, atd.) byly pomocí neparametrického testování zjištěny statisticky významné rozdíly mezi pohlavím v závislosti na věku nakažených nebo při testování závislosti okresu a data hlášení Kruskal-Willisovým testem byl taktéž prokázán signifikantní rozdíl. Varianta tohoto testu s metodou vícenásobného porovnávání následně ukázala, že významný rozdíl je mezi všemi okresy navzájem kromě vzájemných kombinací Přerova, Prostějova a Šumperku. Chí-kvadrátový test prokázal rozdíl poměrů mezi pohlavími v jednotlivých okresech.

Grafické statistické metody. Vhodným krokem k lepší interpretaci zjištěných závislosti je jejich grafická vizualizace, např. pomocí histogramu (viz obr. 3) nebo boxplotu.

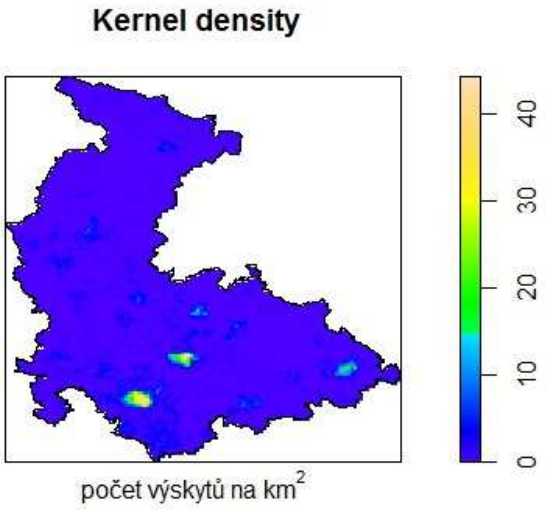


Obr. 3. Histogram počtu hlášení výskytů virové střevní infekce v měsících v roce

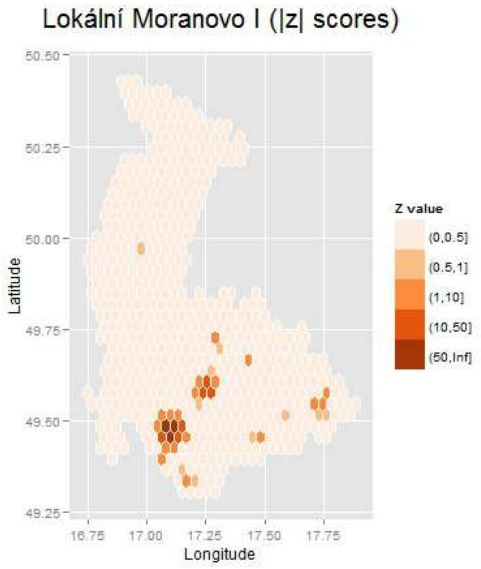
Prostorové analýzy. "Prostorové analýzy jsou souborem technik pro analýzu a modelování lokalizovaných objektů, kde výsledky analýz závisí na prostorovém uspořádání těchto objektů a jejich vlastností."[1] Pro zkoumání distribucí (prostorového rozložení) jednotlivých výskytů bodových vrstev slouží **analýzy prostorových vzorků** (point patterns). Existují čtyři základní typy těchto vzorků rozlišené právě podle způsobů distribuce [5]

- klastry (shlukování)
- normální
- náhodné
- pravidelné

Pomocí analýz prostorových vzorků se zjišťuje, zda dochází ke shlukování (častější výskyty v blížšé vzdálenosti), či jsou výskyty rozmístěny náhodně (nejsou ovlivněny žádným faktorem) nebo podle normálního rozdělení. Jeden z modelů náhodných vzorků se nazývá CSR (complete spatial randomness - úplná prostorová nahodilost, někde známé pod pojmem homogenní Poissonův proces). Nastává v situaci, kdy každý výskyt má stejnou pravděpodobnost výskytu na jakékoliv jiné pozici ve zkoumané oblasti, bez ohledu na umístění ostatních výskytů [5].
Využité analýzy: **Morisita index**, (míra prostorové agregace bodového vzoru v prostoru založenou na frekvenci výskytu jevu v pravidelné čtvercové síti – kvadrátech) [2], **G funkce** (srovnání distribuce datového souboru s distribucí náhodného procesu), **kvadrátový test** (rozdělení zkoumaného území do pravidelných čtverců a sumarizace počtu bodů datového souboru lokalizovaných uvnitř každého kvadrátu), **kernel density** (jádrové odhady hustoty bodů v území, obr. 4.), globální i lokální varianta **Moranova I** (obr.5.), **Gearyho C** nebo **Getis-Ordova I** (testování shlukování).



Obr. 4. Výstup metody kernel density



Obr. 5. Hodnoty Z score lokální analýzy Moranova I výskytů agregovaných do pravidelných hexagonálních ploch

6 Závěr

R disponuje spoustou kladných vlastností. Rychlá tvorba grafů s bohatou nabídkou možností, možnost kontroly všech detailů, snadná opakovatelnost postupů, velké množství analýz, jednoduchá rozšiřitelnost (balíky, otevřenost kódu), velká uživatelská základna a v neposlední řadě žádné pořizovací náklady. Pokud uživatel zvládne ustát negativa, jako je neinteraktivita (žádný zoom, pan nebo editace jednotlivých prvků výběrem myši), náročnost na orientaci, velká vstupní investice času, nutná znalost pozadí procesů nebo rozdílná filozofie práce a myšlení (v porovnání např. s ArcGIS, MS Excel), nabízí se mu v podobě R mocný nástroj pro práci na všech polích geoinformatiky.

Reference

1. HORÁK, Jiří. Prostorové analýzy dat. 3. vyd. VŠB-TU Ostrava, HGF, Institut geoinformatiky, 2011

2. MORISITA, M. (1959). Measuring of the dispersion of individuals and analysis of the distributional patterns. Memoir of the Faculty of Science (Series E2., pp. 215 – 235). Kyushu University.

3. PROCHÁZKA, B., Č. BENEŠ, H. ŠEBESTOVÁ. STÁTNÍ ZDRAVOTNÍ ÚSTAV. Popis systému EPIDAT. Praha, 2011. Dostupné z: http://www.szu.cz/uploads/documents/CeM/epidat/Epidat_3_popis.doc

4. R projekt v ČR [online]. 2013 [cit. 2013-05-03]. Dostupné z: http://www.r-project.cz/index.html

5. WALLER, Lance A a Carol A GOTWAY. Applied spatial statistics for public health data. Hoboken, N.J.: John Wiley, 2004, xviii, 494 p. ISBN 04-713-8771-1.