Organizace projektu v R, opakovatelnost postupů, prostorová data a jejich vizualizace

Jan Caha

Interpretace geodat - 2. blok

12. 4. 2018

1. Organizace dat v R a RStudiu
2. Opakovatelnost a analýz
3. Prostorová data v R

Projekt

- snaha o dodržení logiky - 1 zadaní = 1 RStudio projekt
- RStudio projekt - složka s několika spefickými soubory a složkami
  - `.Rproj.user`
  - `.Rhistory`
  - `.RData` - tomu se snažíme vyhnout - nastavení RStudia
  - `?.Rproj` - asociační soubor pomocí něhož lze projekt přímo otevřít v RStudiu (analogie k `.mxd` souborům)
- zbytek struktury je na uživateli
- pozor na mezery v cestách, ne všechny balíky se s nimi umí rozumně vypořádat

Vhodná struktura projektu

- systém složek, který zjednodušuje a organizuje práci
- nemusí fungovat obecně, ale jako výchozí bod ho lze dopručit
- složky
  - `raw_data` - původní data, bez jakýchkoliv uživatelských zásahů
  - `produced_data` - upravená data, vytvořená v rámci řešení projektu
  - `functions` - `.R` soubory s funkcemi
  - `rmd` - `.Rmd` soubory obsahující nezkompilované zprávy
  - `reports` - složka obsahující zkompilované zprávy v ruzných formátech
  - `models` - `.R` soubory obsahující delší postupy a zpracovávací procesy

Balíček `here`

- malý a velice prostý balíček, který ale neuvěřitelně zjednodušuje proces odkazování se na soubory v projektu
- funkce `here::here()` - vysvětlení syntaxe, `...` argument
  - výsledkem je absolutní cesta k souboru
- jako argumenty funkce uvadíme složky, či soubor a to vždy relativně vůči tzv. "root" složce
- odpadá potřeba řešit umístění spouštěcího souboru a relativizovat cestu vůči němu, celkově nezávislé na lokálním umístění projektu

```
here::here("raw_data", "dataset_1.csv")
here::here("produced_data", "demografie",
           "okresy.xlsx")
```

Funkce `source`

- možnost propojení dvou a více `.R` souborů
- funkce `source` efektivně spustí odkazovaný soubor
- lze využít např. pro načítání balíků, funkcí nebo dat, které budou potřebné v různých dalších souborech
- v případě že se v odkazovaném `.R` souboru objevují české znaky - např. názvy proměnných atd. je nezbytné nastavení parametru `encoding` na hodnotu `utf8`

```
source(here::here("functions", "_initialization.R"),
       encoding = "utf8")
```

- ukázka s vytvořením inicializačního souboru

Opakovatelnost analýz

- většina standardních "GUI" nástrojů příliš na automatizaci a opakovatelnosti postupů nelpí a uživatele k nim systematicky nevede - analýzou např. v ArcGIS se většinou "proklikáte"
- co ale, pokud je nutné analýzu opakovat? (jak opakovatelný je klik myši???)
- pokud celá analýza vznikne v R jako popis postupu zpracování (zdrojový kód), pak není problém proces kdykoliv spustit znovu (třeba i na odlišných datech - vzorek dat $\times$ kompletní dataset)
- důraz na automatizaci práce a opakovatelnost postupů
- v projektech uživateli vzniká jakási "knowledge base" z níž následně čerpá
    - čím více projektů máte, tím rychleji vznikají ty následující
- literatura (Gandrud, 2014)

Dokumentace analýz

- "jakýkoliv kód, který jste půl roku neviděli, mohl dost dobře napsat někdo jiný" (Autor ???)
- jestliže při programování platí: "dokumentovat, dokumentovat, dokumentovat!", pak v data science platí: "komentovat, komentovat, komentovat!"
- potřeba popisovat postup analýzy, důležitá rozhodnutí, ale i výsledky
- lze sice využít komentáře v kódu, ale to má do ideálu daleko
- balík `knitr` a jeho závislosti a formát `R Markdown`

R Markdown

- rozšíření jazyku Markdown, který umožňuje kombinovat Markdown přímo s R kódem
- Markdown snaha o vytvoření co nejjednoduššího a intuitivního jazyka pro popis formátování dokumentů
- překlad z Markdown obvykle do HTML, a nebo jiných formátů např. Word, PDF (někdy může být třeba instalace softwaru pandoc a případně LaTeX)
- struktura - hlavička + dokument
- ukázka - soubor cv_2_01_ukazka_RMarkdown.Rmd, interaktivní práce

R Markdown - kompilace

- možná buď interaktivně nebo pomocí skriptu
- např. takto

```
render(
    input = file,
    output_format = html_document(
      toc = TRUE, toc_depth = 1,
      code_folding = code_folding
    ),
    output_file = file_name,
    output_dir = report_dir,
    envir = new.env(), encoding = "UTF-8"
  )
```

- složky `rmds` a `reports` zmíněné dříve

Nejduležitější balíky

- `sp` (Pebesma and Bivand, 2005), modernější verze `sf` (Pebesma, 2017)
- `raster` (Hijmans, 2017), ve vývoji je balík `stars`(Pebesma, 2018)
- vizualizace `tmap` (Tennekes, 2018) a vývojová verze `ggplot2` (Wickham, 2016) instalovaná z GitHubu
- velké množství balíků v CRAN Views Spatial a SpatioTemporal ale i jinde

Balík `sf`

- novinka od roku 2016
- narozdíl od původního balíku `sp` vychází funkčně z balíku `dplyr` s nímž je kompatabilní -> možnost jednoduše pracovat s daty, např. `group_by()`, `summarise()` platí i pro geometrie -> značné zjednodušení oproti `sp`
- základní závislosti na knihovnách: GDAL, GEOS, PROJ.4, liblwgeom, udunits2 (knivny v C)
- závislosti na řadě R balíků
- značná část funkcionality používá pojemnování `st_*`

Kompatabilita `sf` s `sp`

- na `sp` zavisí relativně hodně balíků - prostorové operace, statistiky a jiné.
- přechod na `sf` bude trvat a některé balíky patrně nikdy přepsané nebudou
- funkce `as_Spatial()` zajistí konverzi z formátu balíku `sf` do `sp`
- inverzně funguje funkce `st_as_sf()`
- doporučení: zpracovávat data jako `sf` a konvertovat až při potřebě speciálních funkcí
- obdobně fungují `raster` a `stars`, byť tam je vztah komplikovanější o to, že `stars` je stále ještě výrazně ve vývojové fázi

Praktický příklad

- s použitím volně dostupných dat vytvořte mapu procentuálního zisku vybrané politické strany ve volbách do PS 2017
- podklady dostupné z GitHubu
- data o volbách předchystaná - tři `.csv` soubory ve složce `raw_data`
- prostorová data - interaktivní stažení z webu ČÚZK

Stažení prostorových dat

- adresu pro stažení dát získáme z https://nkod.opendata.cz/, konkrétně pak odsud

```
temp <- tempfile()
download.file(
  "http://services.cuzk.cz/shp/stat/epsg-5514/1.zip",
   temp)

unzip(temp, exdir = here::here("raw_data"),
      junkpaths = TRUE)
```

Data o volbách

- taktéž nalezené na https://nkod.opendata.cz/
- z důvodu formátu XML, který navíc není úplně dobře zpracován, bylo zpracování provedeno v Pythonu
- 3 .csv soubory ve složce raw_data

```
library(tidyverse)

data_okresy <- read_csv(here::here("raw_data",
                                   "data_okresy.csv"))
data_strany <- read_csv(here::here("raw_data",
                                   "data_strany.csv"))
ciselnik_stran <- read_csv(here::here("raw_data",
                                      "ciselnik_strany.csv"
```

Selekce zájmové strany

- podle číselníku stran určíme ID zájmové strany a tu pak vyfiltrujeme

```
data <- data_strany %>%
  filter(id_strany == 7)
```

Načtení prostorových dat

```
library(sf)
```

```
## Linking to GEOS 3.6.1, GDAL 2.2.3, proj.4 4.9.3
```

```
okresy <- st_read(here::here("raw_data", "OKRESY_P.shp"),
                  stringsAsFactors = FALSE)
```

```
## Reading layer ‘OKRESY_P’ from data source ‘D:\R_projects
## Simple feature collection with 77 features and 5 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -904585.3 ymin: -1227296 xmax: -43
## epsg (SRID):    5514
## proj4string:    +proj=krovak +lat_0=49.5 +lon_0=24.83333
```

Propojení tabulky s prostorový daty

```
okresy <- okresy %>%
  right_join(data, by = c("LAU1_KOD" = "nuts_kod"))
```

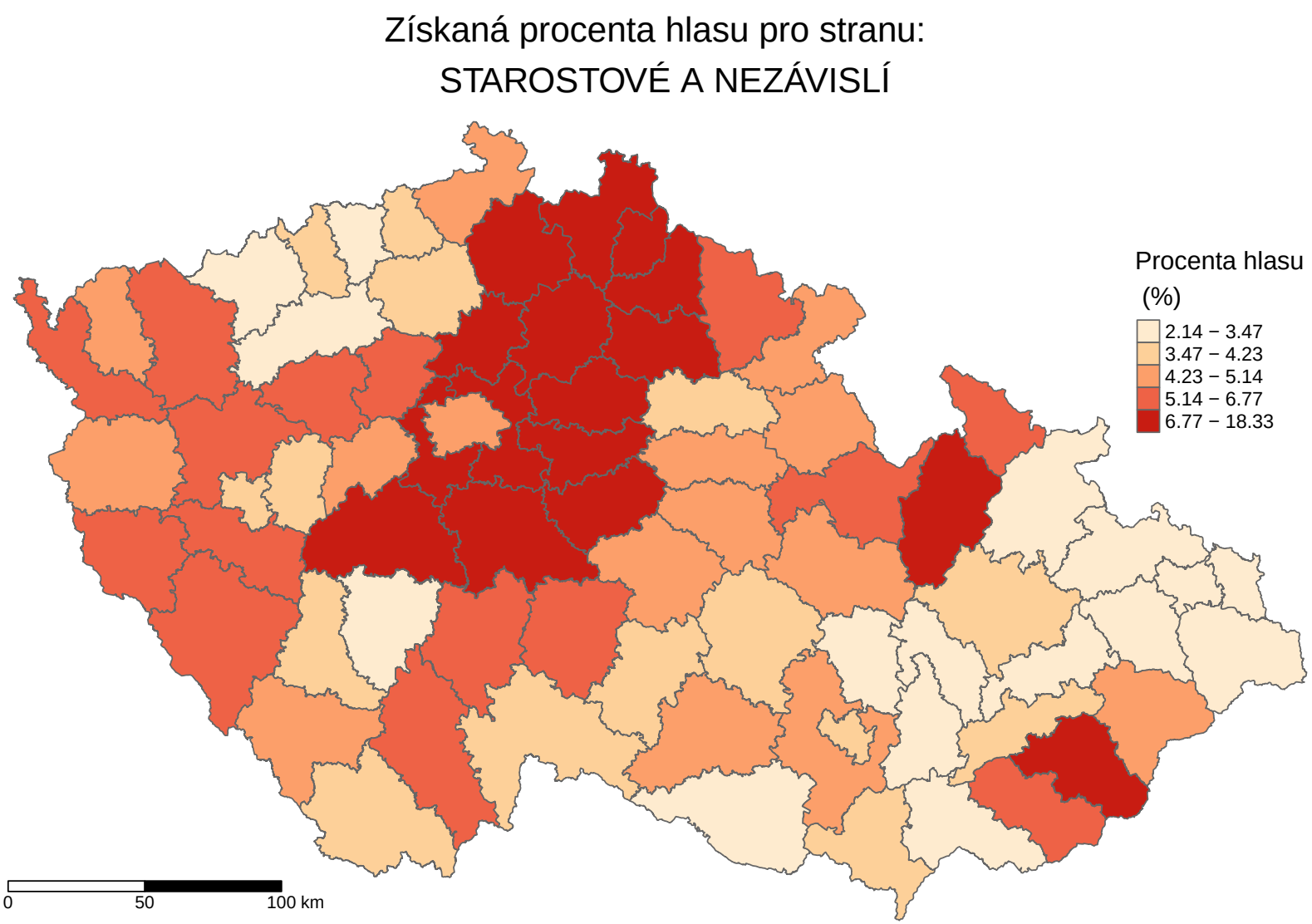
```
View(okresy)
```

Vizualizace

```
library(tmap)

tm_shape(okresy) +
  tm_polygons(col = "hlasu_procenta", n = 5 ,
              style = "quantile", palette = "OrRd",
              title = "Procenta hlasů \n (%)") +
tm_scale_bar(position = c("left", "bottom"),
             breaks = c(0, 50, 100), size = 0.75) +
tm_layout(frame = FALSE,
          legend.title.size = 1.3,
          legend.text.size = 0.8,
          legend.format = list(text.separator = "-"),
          legend.position = c(0.85, 0.6),
          main.title = "Získaná procenta hlasů pro
          stranu:\n STAROSTOVÉ A NEZÁVISLÍ",
          main.title.position = "center")
```

Vizualizace



Interaktivní vizualizace

- prosté přepnutí pomocí metody `tmap_mode` před odesláním mapy na výstup

```
tmap_mode("view")
tmap_mode("plot")
```

Parametrizace a automatizace vizualizací

- příklady `cv_2_05_prostorová_vizualizace_dat_parametrizace.R` a `cv_2_06_prostorová_vizualizace_dat_automatizace.R`
- lze takto automatozovaně nagenerovat výrazné množství vizualizací

Úkol

Vytvořte `R Markdown dokument`, který shrne volební výsledky jedné vámi zvolené strany. Výsledkem by měla být kratičká zpráva s naprosto základnimi statistikami (minimální, mediánový a maximální zisk v okresech). Histogram rozložení procentuálních hodnot a mapa.

Literatura I

Gandrud, C. (2014) *Reproducible Research with R and RStudio*. CRC Press, Boca Raton. ISBN 9781466572850.

Hijmans, R. J. (2017) *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7.
URL: *https://CRAN.R-project.org/package=raster*

Pebesma, E. (2017) *sf: Simple Features for R*. R package version 0.5-5.
URL: *https://CRAN.R-project.org/package=sf*

Pebesma, E. (2018) *stars: Scalable, Spatiotemporal Tidy Arrays for R*. R package version 0.1-1.
URL: *https://github.com/r-spatial/stars/*

Pebesma, E. and Bivand, R. (2005) Classes and methods for spatial data in R. *R-NEWS*, 5(2), pp. 9–13.

Tennekes, M. (2018) tmap: Thematic Maps in R. *Journal of Statistical Software*, 84(6), pp. 1–39. doi:10.18637/jss.v084.i06.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4.
URL: *http://ggplot2.org*

Dotazy?

Děkuji za pozornost.