

R BALÍČKY URČENÉ K PRÁCI S GEODATY A JEJICH VYUŽITÍ V PROJEKTU PERUN

ONDŘEJ LEDVINKA (CZ)

Abstrakt. Příspěvek představuje současné možnosti zpracování geodat funkcemi vysokoúrovňového programovacího jazyka R. V krátkosti upozorňuje na funkce navržené pro manipulaci, tvorbu a analýzu vektorových geodat. Stěžejní část se však věnuje práci s rastrovými geodaty, neboť ta tvoří podstatu projektu PERUN, ať již jde o odhad budoucího vývoje klimatologických charakteristik prostřednictvím scénářů modelu ALADIN-CLIMATE/CZ Českého hydrometeorologického ústavu, či o dovozování dalších veličin potřebných pro hydrology zabývající se historickým a budoucím vývojem vodních zdrojů na území Česka. Na příkladech geodat, se kterými v rámci projektu autor pracuje, jsou vybrány některé tipy, které se mohou hodit začátečníkům i pokročilým uživatelům jazyka R.

Klíčová slova. Vektorová geodata, rastrová geodata, tidyverse, funkcionální programování.

R PACKAGES FOR WORKING WITH GEODATA AND THEIR USE IN THE PERUN PROJECT

Abstract. The paper presents current possibilities of geodata processing by functions of the high-level programming language R. In short, it draws attention to functions designed for manipulation, creation and analysis of vector geodata. However, the substantial part is devoted to work with raster geodata, as they form the essence of the PERUN project, whether it is to estimate the future development of climatological characteristics through scenarios of the Czech Hydrometeorological Institute's ALADIN-CLIMATE/CZ model, or to derive other quantities needed for hydrologists dealing with the historical and possible future development of water resources in the territory of Czechia. Some tips are selected on examples of geodata that the author works with within the project, which may be useful for beginners and advanced users of the R language.

Keywords. Vector geodata, raster geodata, tidyverse, functional programming.

1. Úvod

Vysokoúrovňový programovací jazyk R byl původně navržen pro účely statistických výpočtů a analýz a zároveň pro kreslení zásadních rysů výsledků či vystižení podstatných vlastností a chování dat. Postupně si tento jazyk našel mnohé oblasti svých aplikací, o čemž svědčí i úctyhodný počet přídatných balíčků, kterých je v současnosti něco přes 22 tisíc, pokud tedy jde jen o ty zveřejněné na zrcadlech CRAN (Comprehensive R Archive Network) a nepočítají se jiné repozitáře. Jeli-kož empirická data přirozeně mívají vztah k místu sběru, bylo jen otázkou času, kdy se mezi balíčky objeví nástroje připomínající geografické informační systémy

(GIS) s možnostmi tzv. geoprocessingu. Spojí-li se potenciál takto promyšleného zacházení s geoprostorovými daty (geodaty) s nabídkou stavět si vlastní skripty především z důvodů automatizace nebo aplikace stejné funkce na velké množství jednotek najednou (či naopak více funkcí na jednotku) prostřednictvím tzv. for cyklů, vzniká velmi šikovný pomocník, usnadňující život nejednomu analytikovi, který by se jinak utápěl v současném světě nepřehledného kvanta dostupných a stále se otevírajících dat. To vše R poskytuje a je tedy vhodné se tomuto tématu seriózně věnovat.

Vzhledem k tomu, že i podmnožina balíčků určených pro práci s geodaty v R je poměrně bohatá a sběr informací pro tzv. CRAN Task Views je dobrovolný (geodatům nejbližší zřejmě ty na [1] nebo [8]), není možné postihnout na limitovaném počtu stránek příspěvku celý tento svět a je nutné se zabývat jen určitou částí. Vhodný odrazový můstek přitom představuje český projekt PERUN, který byl spuštěn v červenci 2020, má končit v prosinci 2026 a jehož stěžejní dílčí cíle se týkají tvorby budoucích klimatických scénářů, jejich korekcí a aplikací (např. v klimatologii a hydrologii). Následuje tedy přehled základních balíčků pro práci s vektorovými a rastrovými geodaty v R, přičemž zpracování rastrových geodat je upřednostněno z důvodu dominance těchto dat v projektu PERUN, na kterých mohou být demonstrovány ukázky.

Autor si klade za cíl ukázat několik tipů, které pomohou urychlit proces učení zájemců. Zejména půjde o tipy pro práci s geodaty v R, které mohou i po přečtení patřičných návodů k funkcím začátečníkům snadno uniknout. Čtenář nechť naopak neočekává přehled funkcí určených pro kreslení statických a dynamických map prostřednictvím R s využitím načtených geodat. Tomuto tématu se věnuje jiný příspěvek.

2. Vektorová geodata v současném R

Poté, co se jeden z neplodnějších programátorů R balíčků zaměřených na práci s geodaty Roger Bivand rozhodl, že v roce 2023 odejde do důchodu a tímto skončí další vývoj a podpora balíčků `rgdal`, `rgeos` a `maptools` (viz např. [9]), je nepochybně v současnosti nejdůležitější sadou funkcí pro tvorbu, načítání (i z databází), manipulaci a ukládání vektorových geodat balíček `sf` (z angl. *simple features*; [11, 10]). Takto lze pracovat s atributovými tabulkami připomínajícími klasické datové rámce (*data frames*), které jsou navíc obohacené o sloupec s geometrií, jež je v něm pro každý řádek (tj. *feature*) uchována formou tzv. *well-known textu* (WKT) za využití faktu, že sloupcem datového rámce může být i seznam. Balíček `sf` nabádá při skriptování k aplikacím přístupu *tidyverse* [13, 14], takže lze využívat především jeho slovesa i při manipulaci s geometrií. Samozřejmostí je podpora zápisu s tzv. pipe operátory (od R verze 4.1.0 hlavně nativní pipe `|>` namísto `pipu %>%` vypůjčeného z balíčku `magrittr`) a zejména vylepšeného datového rámce známého pod pojmem *tibble*, který naskýtá lepší kontrolu nad třídami

sloupců tabulky, ale také mnohem lepší tisk části dat pro nahlédnutí do konzole (včetně zvýraznění chybějících hodnot či záporných čísel odlišnou barvou). Velkou výhodou je však podpora funkcionálního programování s vektorizovanými funkcemi, jež do značné míry nahrazuje for cykly.

Balíček **sf** rozeznává hlavně tzv. velkou sedmičku typů geometrií, ale možností je mnohem více. Balíček zakládá na závislostech a integrování několika externích knihoven. Načítání, ukládání geodat a i některé operace s nimi probíhají přes knihovnu GDAL, vztahy mezi geometriemi a dotazy nad nimi (včetně výpočtu měr v rovině) jsou řešeny přes knihovnu GEOS. Podpora souřadnicových referenčních systémů (CRS) je umožněna propojením s knihovnou PROJ, přičemž zde je upřednostněno odkazování se na CRS pomocí WKT řetězců, což nevylučuje odkazy typu AUTORITA:KÓD. Avšak tzv. PROJ4 řetězce jsou považovány za zastaralé, protože je doporučováno se jim již vyhýbat. Propojení s Google knihovnou S2 podporuje výpočty na sféře. V neposlední řadě stojí za zmínku, že vývojáři balíčku **sf** si potrpí na upřesňování jednotek, u nichž ale existuje možnost je zahazovat. Podobně je to i s geometrií, po jejímž zahození vzniká obyčejný datový rámec s pouhými atributy. Poněkud zrádné mohou být zkratky používané v nápovědách k funkcím: **sfg** značí třídu *simple feature geometry* a **sfc** pro změnu *simple feature column*, tedy geometrický sloupec, ke kterému již lze přidávat atributy. Setkáme se též se zkratkou **agr**, která může připomínat jakousi agregaci, ale její význam zní *attribute-geometry relationship*, což prozrazuje, že se dbá na vztahy mezi atributy a geometrií, a to právě kvůli různým operacím s geometrií, jako je např. řezání polygonů čarami.

Vektorová geodata mají v projektu PERUN podružný význam. Dalo by se říci, že slouží spíše ke kontrolám vykreslováním jednoduchých map a k omezování se na území Česka a jeho okolí. V hydrologických analýzách však nacházejí svá uplatnění v kombinaci s rastrovými geodaty, kdy je např. nutné z hodnot rastrových buněk získat hodnoty agregované přes plošné jednotky, jako jsou povodí nad vodoměrnými stanicemi. Jinou aplikací mohou být prostorové interpolace geostatistickými funkcemi, které však autor příspěvku většinou přenechává jiným řešitelům projektu PERUN. Výjimkou byla tvorba týdenních časových řad gridů vodní hodnoty sněhu na území Česka pro zimní období mezi lety 1980 a 2023 [2].

3. Rastrová geodata v současném R

Stran zpracování rastrových geodat dnes v R všemu bez debat vévodí balíček **terra** [4], který od roku 2020 nahrazuje svého předchůdce s názvem **raster**. Mnoho funkcí balíčku **terra** je ve skutečnosti implementováno v jazyku C++, a proto jsou mnohem rychlejší než jejich protějšky pocházející z balíčku **raster**. Jsou zde velmi dobře podporovány anonymní funkce, a to především ve funkcích typu **app()** a **tapp()**, jejichž názvy připomínají základní funkce R nazvané

`apply()` nebo `tapply()`. Tím ale výčet těchto funkcí nekončí, existují zde i funkce příbuzné základním funkcím `lapply()`, `sapply()` apod. (tedy `lapp()` a `sapp()`).

Je třeba poznamenat, že pokud se někdo rozhodne aplikovat funkci, která není implementovaná v C++, jistě narazí na značné zpomalení výpočtů. Pro tyto účely vývojáři balíčku **terra** umožňují rozložit proces na více logických jader pracovní stanice, což má většinou za následek urychlení práce. Ovšem velmi záleží na konkrétním úkolu, struktuře dat, na stylu napsání využívané funkce (např. jestli je vektorizovaná) a samozřejmě na konfiguraci samotné pracovní stanice, jako je počet logických jader nebo kapacita RAM. Proto je před rutinním využívání balíčku **terra** nezbytné počáteční experimentování. Teprve po rozpoznání typů úloh a nejlepšího způsobu jejich zpracování na té které pracovní stanici dochází k rapidnímu navrácení času investovaného do prvotních pokusů.

Práce s RAM zde hraje důležitou roli. Rastrová geodata jsou často tak velká, že se do paměti běžného osobního počítače ani nevejdou. Je proto třeba vymýšlet různé strategie, jak k takto velkým datům přistupovat. Jednou z nich je poctivě ukládání si pracně vytvořených výsledků, nebo dokonce mezivýsledků, do souborů na disku. V tomto **terra** závisí na knihovně GDAL (např. GeoTIFF soubory) anebo i jiných knihovnách, jako je např. **netcdf** pro ukládání v klimatologii velmi oblíbených NetCDF souborů. Pokud do souboru výsledky neukládáme rovnou, tak jako tak se k objektům v Globálním prostředí často přiřazují jen tzv. pointery (odkazy) na dočasné soubory, které **terra** ukládá za nás na místo dané funkcí `terraOptions()`, aniž bychom o tom vůbec museli zpočátku vědět. Skutečnost, že se objekt odkazuje na soubor, poznáme tak, že se vedle jeho názvu v Globálním prostředí objevuje sousloví *Formal class SpatRaster* namísto *SpatRaster*, což je třída pro rastrová geodata balíčku **terra** (vedle třídy *SpatVector* pro vektorová geodata). Proto není doporučováno objekty z Globálního prostředí v případě geodat ukládat do velmi známých RData souborů, po jejichž otevření stejně zjistíme, že v takových souborech se žádná rastrová geodata nenachází.

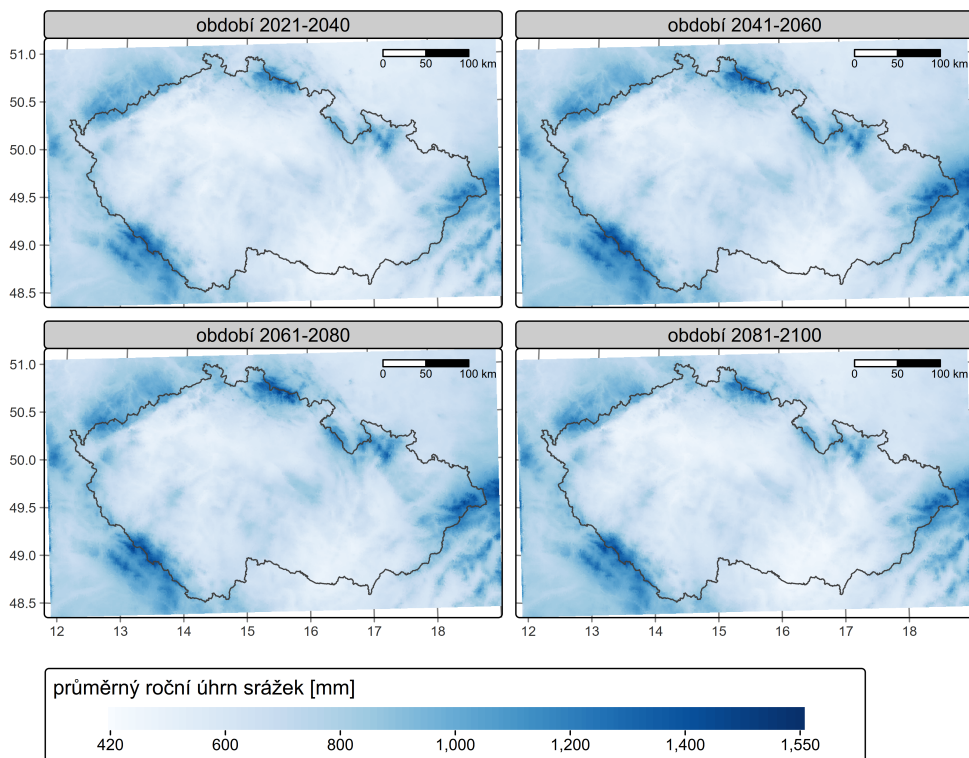
Načítání a ukládání rastrových geodat vyžadují speciální poznámky. Pro načítání všech typů souborů, lhostejno, zda GeoTIFF, nebo NetCDF (dokonce je jedno, zda s jednou, nebo s více vrstvami, pokud je splněn určitý limit na počet vrstev), slouží funkce `rast()`. Této funkci ani nevádí načítat více souborů z adresáře najednou, když budou stejného typu (např. GeoTIFF), případně ani z různých adresářů či podadresářů, což lze zajistit předchozím připravením si vektoru cest k takovým souborům. Funkce `rast()` dokonce umí konvertovat na třídu *SpatRaster* z jiných tříd, jako jsou pole, seznamy nebo datové rámce. Najdeme zde i vlastnost mazání hodnot rastrových buněk a ponechání pouhé geometrické konstrukce rastru, což se hodí před přiřazením nových hodnot. Pro ukládání rastrových geodat naopak existuje více funkcí. V projektu PERUN jsou pro tyto účely nejvíce využívány funkce `writeRaster()` a `writeCDF()`. Tyto funkce existují odděleně patrně z důvodu využívání různých externích knihoven (viz výše) a podpory různého chování ukládaných souborů. Nežřídka se stává, že uživatelé

vyžadují každý bod časové řady gridů mít uložen do separátního souboru. Zřejmě je to proto, že jejich softvér, který používají k dalšímu zpracování, neumí načítat více vrstev najednou, nebo je to v něm velmi nešikovné. V projektu PERUN byly testovány různé strategie pro uložení individuálních vrstev do souborů. Byla k tomu zaměstnána i tzv. paralelizace mapovacích funkcí prostřednictvím tzv. *futures*, což umožňuje R balíček *furrr* [12]. Ovšem při lokálním zpracování bylo ve smyslu rychlosti bezkonkurenční zaměřit se na nepatrnou poznámku v nápovědě k funkci `writeRaster()`, a sice, že, pokud bude jako argument cest nových souborů nastaven vektor o stejném počtu prvků, jako je počet vrstev v původním *SpatRaster* objektu, dojde k sepnutí trhání do souborů implementovanému v C++.

Nedostatek balíčku *terra* lze spatřovat v nízké podpoře ovládání ve smyslu *tidyverse*. Funkce balíčku *terra* připomínají spíše funkce základního R. Je ale třeba podotknout, že z důvodu překlenutí tohoto nedostatku vzniknul další R balíček *tidyterra* [3], díky němuž lze na třídy balíčku *terra* aplikovat celou škálu funkcí známých ze světa *tidyverse*.

Náleží sem rovněž podpora kreslení rastrových geodat ve smyslu funkcí balíčku *ggplot2* (viz např. `tidyterra::geom_spatraster()`), který je součástí jádrového *tidyverse*. Někdy je však zapotřebí dávat si pozor na výkonnost některých takových funkcí. Při zpracování rastrových geodat s velkým počtem vrstev bylo upozorováno, že výběr vrstev pomocí funkce `tidyterra::select()` je velmi pomalý oproti nativnímu *terra* postupu s dvěma hranatými závorkami. Dobrou zprávou je, že objekty vzniklé funkcemi balíčku *terra* lze naskládat do sloupce datového rámce (lépe řečeno *tibble*) jako seznam (např. po aplikaci funkce `terra::as.list()`) a lze tedy na ně velmi dobře aplikovat funkcionální programování ve smyslu balíčků *purrr* či *furrr*. Při paralelizaci se ovšem ukazuje nedostatek implementace funkcí v C++, a sice, že bez načtení dat do paměti funkcí `terra::wrap()` zatím nelze informovat pracovníky clusteru o *SpatRaster* objektech, se kterými mají pracovat. Zdá se ale, že dosavadní vývoj balíčku *terra* má tendenci tuto mezeru vyplnit.

Jde-li o další *terra* funkce využívané v projektu PERUN, patří mezi ně zcela jistě funkce `extract()`, která umožňuje redukovat hodnoty rastru pro místa jednotek danými vektorovými geodaty. Mohou tak rychle vznikat časové řady srážkových úhrnů či teploty vzduchu pro jednotlivá povodí nad vodoměrnými stanicemi pro porovnání s časovými řadami průtoků [5] nebo pro hydrogeologické rajony z důvodu výzkumu reagování vydatnosti pramenů na klimatickou změnu [7, 6]. Užitečnou se ukázala být funkce `tapp()`, jež dokáže rychle dodat agregace pro časové jednotky (např. při přechodu z denního časového kroku na měsíční), což je velmi chtěná úloha. Vizualizaci průměrných ročních srážkových úhrnů na území Česka a jeho okolí podle modelu ALADIN-CLIMATE/CZ (scénáře SSP2-4.5) ve čtyřech budoucích obdobích pak znázorňuje obr. 1.



Obrázek 1. Průměrný roční úhrn srážek na území Česka a jeho okolí ve čtyřech budoucích obdobích podle modelu ALADIN-CLIMATE/CZ (scénář SSP2-4.5; zdroj: <https://www.perun-klima.cz/scenare/index.php>)

4. Závěr

Kombinace GIS funkcí a skriptování nabízí širokou škálu aplikací hlavně tam, kde není kladen důraz na vizuální stránku jako spíše na zpracování velkého množství geodat a vytěžení syntetické informace z více jednotek najednou. Jazyk R kombinuje tyto možnosti i s funkcionálním programováním (např. prostřednictvím přístupu *tidyverse*), které tvoří výborný základ pro aplikování metod strojového učení. Paralelizace procesů nebo implementace některých funkcí v jazyku C++ značně snižují dobu zpracování. Při analýze geodat v R je však třeba pamatovat na několik úskalí, a to především v případě rastrových geodat. Předně je vhodné se odnaučit ukládat objekty celého Globálního prostředí do RData souborů a serializovat jen vybrané objekty, které skutečně měly co dočinění s nahráním dat do paměti pracovní stanice (např. do souborů RDS). Další velmi doporučenou strategií je ukládat si výsledky a mezivýsledky, jejichž zdrojem jsou dočasné soubory, které mizí po končení každého sezení, do stálých souborů

(např. GeoTIFF nebo NetCDF). V neposlední řadě je nutné si ověřit co nejrychlejší způsob zpracování individuálně na té které stanici s daným počtem logických jader a nějakou kapacitou RAM. Rovněž tak se musí volit vhodný postup podle toho, jak zní zadání úlohy, jaká data máme právě po ruce a jestli funkce, které známe, jsou vektorizované či ne. Pokud je funkce vektorizovaná, lze proces snadno paralelizovat (např. s využitím anonymních funkcí), ale u rastrových dat jsme zatím nuceni před paralelizací je nahrát do paměti, což může při lokálních aplikacích činit nemalé potíže. Proto je dobré zauvažovat i o rychlosti postupů s funkcemi implementovaných v C++ a všechny možnosti srovnávat (např. měřením spotřebovaného času).

Jazyk R je obecně dosti pomalý a lze tedy reálně uvažovat o jeho kombinacích s externími knihovnami či dokonce s funkcemi externích GIS, jako je QGIS, SAGA GIS nebo GRASS GIS. To vše je možné přes specializované R balíčky. Při zpracování velkých rastrových geodat se lze poohlédnout i po jiném jazyce, jako je např. Julia. Při tomto jazyce najdeme velmi agilní skupinu programátorů, díky nimž vzniká čím dál více atraktivních balíčků pro zpracování geodat. Nechme se tedy překvapit, co přinese budoucnost.

Poděkování. Tento příspěvek vznikl s podporou Technologické agentury České republiky (TA ČR), projektu SS02030040 „Predikce, hodnocení a výzkum citlivosti vybraných systémů, vlivu sucha a změny klimatu v Česku (PERUN)“.

Reference

- [1] BIVAND, R. – NOWOSAD, J.: *CRAN Task View: Analysis of Spatial Data*, 2025, <https://CRAN.R-project.org/view=Spatial> [cit. 21. 6. 2025].
- [2] FERČÁK, J.: *Variability and Change of Snow Storage in Czechia at Different Elevations*, Bachelor Thesis, 2025, Charles University, Faculty of Science, Prague.
- [3] HERNANGÓMEZ, D.: *Using the Tidyverse with Terra Objects: The Tidyterra Package*, month = nov, Journal of Open Source Software, 2023, Vol. 8, Nr. 91, pp. 5751, ISSN 2475-9066, dostupné z DOI 10.21105/joss.05751 [cit. 20. 4. 2024].
- [4] HIJMANS, R. J.: *Terra: Spatial Data Analysis*, 2025, howpublished = <https://CRAN.R-project.org/package=terra>.
- [5] JENÍČEK, M. – NEDĚLČEV, O. – TYL, R. – LEDVINKA, O. – VAJSKEBR, V. – ŠERCL, P. – BERNSTEINOVÁ, J. – LANGHAMMER, J. – UNUCKA, J.: *CAMELS-CZ: Databáze atributů povodí pro hydrologické a klimatologické studie na velkém vzorku malých povodí*, První konference PERUN, 2023, Praha, Český hydrometeorologický ústav, ISBN 978-80-7653-063-8, s. 99–105, <https://www.chmi.cz/files/portal/docs/reditel/SIS/doi/Perun/2023/978-80-7653-063-8.14.pdf> [cit. 6. 1. 2025].
- [6] LAMAČOVÁ, A. – LEDVINKA, O. – BOHDÁLKOVÁ, L. – OULEHLE, F. – KREISINGER, J. – VLNAS, R.: *Response of Spring Yield Dynamics to Climate Change across Altitude Gradient and Varied Hydrogeological Conditions*, Science of The Total Environment, 2024, Vol. 921, ISSN 0048-9697, dostupné z DOI 10.1016/j.scitotenv.2024.171082 [cit. 11. 4. 2024].
- [7] LEDVINKA, O. – LAMAČOVÁ, A.: *Změny ve vydatnosti pramenů na území Česka v letech 1971–2020*, První konference PERUN, 2023, Praha, Český hydrometeorologický ústav,

- ISBN 978-80-7653-063-8, s. 175–178, <https://www.chmi.cz/files/portal/docs/reditel/SIS/doi/Perun/2023/978-80-7653-063-8.22.pdf>, [cit. 6. 1. 2025].
- [8] PEBESMA, E. – BIVAND, R.: *CRAN Task View: Handling and Analyzing Spatio-Temporal Data*, 2022, <https://CRAN.R-project.org/view=SpatioTemporal> [cit. 21. 6. 2025].
 - [9] PEBESMA, E. – BIVAND, R.: *R-Spatial Evolution: Retirement of Rgdal, Rgeos and Maptools / R-bloggers*, 2022, <https://www.r-bloggers.com/2022/04/r-spatial-evolution-retirement-of-rgdal-rgeos-and-maptools/> [cit. 21. 6. 2025].
 - [10] PEBESMA, E. – BIVAND, R.: *Spatial Data Science: With Applications in R*, The R Series, 2023, CRC Press, Boca Raton, ISBN 978-0-429-45901-6.
 - [11] PEBESMA, E.: *Simple Features for R: Standardized Support for Spatial Vector Data*, The R Journal, 2018, Vol. 10, Nr. 1, pp. 439–446, ISSN 2073-4859, dostupné z DOI 10.32614/RJ-2018-009 [cit. 7. 3. 2021].
 - [12] VAUGHAN, D. – DANCHO, M.: *Furrr: Apply Mapping Functions in Parallel Using Futures*, 2022, <https://CRAN.R-project.org/package=furrr> [cit. 21. 6. 2025].
 - [13] WICKHAM, H. – AVERICK, M. – BRYAN, J. – CHANG, W. – MCGOWAN, L. – FRANÇOIS, R. – GROLEMUND, G. – HAYES, A. – HENRY, L. – HESTER, J. – KUHN, M. – PEDERSEN, T. – MILLER, E. – BACHE, S. – MÜLLER, K. – OOMS, J. – ROBINSON, D. – SEIDEL, D. – SPINU, V. – TAKAHASHI, K. – VAUGHAN, D. – WILKE, C. – WOO, K. – YUTANI, H.: *Welcome to the Tidyverse*, Journal of Open Source Software, 2019, Vol. 4, Nr. 43, pp. 1686, ISSN 2475-9066, dostupné z DOI 10.21105/joss.01686 [cit. 7. 3. 2021].
 - [14] WICKHAM, H. – ÇETINKAYA-RUNDEL, M. – GROLEMUND, G.: *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2023, 2nd, O'Reilly Media, Inc, Sebastopol, CA, ISBN 978-1-4920-9740-2.

Kontaktní adresa

Mgr. Ondřej Ledvinka, Ph.D., Oddělení hydrofondu a bilancí, Úsek hydrologie, Český hydrometeorologický ústav, Na Šabatce 2050/17, 143 06 Praha 4 – Komořany, Česká republika,
E-mailová adresa: ondrej.ledvinka@chmi.cz
 Katedra fyzické geografie a geoekologie, Přírodovědecká fakulta, Univerzita Karlova, Albertov 6, 128 00 Praha 2, Česká republika,
E-mailová adresa: ledvinka@natur.cuni.cz