SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

4/2016
ročník/volume 26

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 2

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 21 – 34

Dátum vydania/Publication date: 15. október 2016/October 15, 2016

Roman PAVELKA
Statistický úřad SR

PŘÍKLADY VYUŽITÍ PROGRAMOVACÍHO JAZYKA R PRO STÁTNÍ STATISTIKU

EXAMPLES OF USING THE R PROGRAMMING LANGUAGE IN OFFICIAL STATISTICS

ABSTRAKT
Popularita programovacího jazyka R se v národních statistických úřadech neustále zvyšuje, a to nejen pouze pro simulační a podpůrné úlohy. V současnosti se programovací jazyk R také využívá v procesu produkce statistických výstupů. V rámci programovacího jazyka R bylo v průběhu posledních let vyvinuto množství nových funkcionalit pro rozmanité úlohy na poli státní statistiky, přičemž tyto funkcionality jsou volně dostupné ve formě doplňkových programových balíčků. Příspěvek nejprve krátce představí samotný programovací jazyk R. V další části budou sumarizovány funkcionality programovacího jazyka R pro oblast státní statistiky, které byly vyvinuty a které jsou používány na národních statistických úřadech.

ABSTRACT
The popularity of the R programming language is steadily increasing in national statistical offices and not only for simulation or supporting tasks. Nowadays the R programming language is also used in the production of statistical outputs. In the context of the R programming language, a lot of new features for various tasks in official statistics have been developed over the last few years while these features are freely available in the form of add-on packages. This contribution initially briefly presents the R programming language itself. Other features of this programming language for the field of official statistics developed and used at national statistics institutes will be summarized in the next part of the contribution.

KLÍČOVÁ SLOVA
jazyk R, statistické analýzy, počítačová věda, státní statistika

KEY WORDS
R language, statistical analysis, computer science, official statistics

1. ÚVOD

R je programovací jazyk určený především pro zpracování dat a je specializovaný zejména na statistické výpočty a grafiku. Programovací jazyk R vychází z jazyka S, který byl vyvinut v Bell Laboratories (předtím AT&T, nyní Lucent Technologies) jako komerční softwarový nástroj pro analýzy a vyhodnocování dat. Na rozdíl od tohoto komerčního nástroje představuje jazyk R volně dostupný programovací jazyk, který je šířen a používán v rámci tzv. open-source[1] projektů nadace *Free Software*

[1] *Otevřený software (anglicky open-source software nebo open software, zkratka OSS) je počítačový software s otevřeným zdrojovým kódem. Otevřenost zde znamená jak technickou dostupnost kódu, tak legální dostupnost – licenci software, která umožňuje, při dodržení jistých podmínek, uživatelům zdrojový kód využívat, například prohlížet a upravovat.*

Foundation[2]. Uvedený způsob šíření a používání programovacího jazyka R umožňuje jeho neustálý rozvoj, a to zejména v podobě mnoha doplňkových programových balíčků (*add-on packages*) s knihovnami a funkcemi na různé typy specializovaných analýz. Z těchto uvedených důvodů programovací jazyk R získává stále větší význam nejen v komerční a akademické sféře, resp. vědeckovýzkumných pracovištích, ale nabývá na důležitosti i při produkci výstupů státní statistiky v národních statistických úřadech. Projekt programovacího jazyka R je řízen nekomerční institucí nazývající se *R Foundation for Statistical Computing* [20] se sídlem ve Vídni.

Z pohledu ekonomické náročnosti je programovací jazyk R volně dostupný, a tedy zadarmo. Proto zavedení a využívání programovacího jazyka R by mohlo znamenat značný ekonomický efekt pro národní statistický úřad (samozřejmě s nutnou technickou a organizační podporou).

Nejdůležitějším cílem tohoto příspěvku je poukázat na to, že v současnosti už existuje široké a úspěšné využívání programovacího jazyka R na různých statistických úřadech Evropského statistického systému, zejména ve statistickém úřadu v Rakousku. Příspěvek chce podnítit diskusi o postupném přechodu statistického produkčního procesu založeném na využívání programovacího jazyka *R v národním statistickém úřadu*.

2. NEJDŮLEŽITĚJŠÍ VLASTNOSTI PROGRAMOVACÍHO JAZYKA R

Programovací jazyk R je dostupný jako volně šiřitelný software (*Free Software*) při dodržení podmínek *General Public License* nadace *Free Software Foundation* pro statistické výpočty a analytickou grafiku při zpracování a vyhodnocování dat. Protože je zdarma, programovací jazyk R již předstihl počtem uživatelů komerční jazyk S a stal se faktickým standardem v řadě oblastí statistiky. Programovací jazyk R lze použít ve statistických analýzách při práci s rozdělením pravděpodobnosti, v maticovém počtu, v rámci řešení optimalizačních problémů, s regresními modely různých typů apod. Programovací jazyk R umožňuje různé metody a techniky výběru ve výběrových šetřeních, odhady populačních parametrů, editační a imputační techniky atd. Vedle výborně zvládnuté funkcionality v manipulaci s daty pro výpočty, v práci s vektory, maticemi a šikovnými nástroji pro datovou analýzu a vizualizaci je toto softwarové prostředí neustále rozvíjeno týmem odborníků soustředěných v rámci *The Comprehensive R Archive Network* (dále „CRAN“) [36]. Základní informace o programovacím jazyku R, o manipulaci s daty a základních možnostech jejich vizualizace poskytl článek v časopisu Slovenská štatistika a demografia 1/2016 [17]. Uvedený článek představil praktické využití programovacího jazyka R při analýze portfolia, při regresní a korelační analýze se základními statistickými grafy z pohledu především analýzy rizika.

Podle jiné studie je další hlavní silnou stránkou programovacího jazyka R vysoký stupeň interakce s ostatními již zavedenými – komerčními i nekomerčními – statistickými pakety i jinými programy [23], což prakticky znamená:

- programové rozhraní je kompatibilní vůči dalším programovacím jazykům, takovým jako je C, C++, Java nebo Python;

[2] *Nadace Free Software Foundation, česky Nadace pro svobodný software, byla založena v roce 1985 s cílem podporovat práva uživatelů počítačů používat, studovat, kopírovat, modifikovat a redistribuovat počítačové programy.*

- obsahuje vynikající nástroje pro import/export dat, když výměna dat může být realizována ve formátu CSV, EXCEL, SDMX, XML, Stata, SPSS, SAS (Xport sas7bdat), JSON, ve formátech pevné šířky, v binárních formátech;
- zahrnuje funkce, které umožňují propojení na důležité databáze, např. DB2 (ODBC, JDBC), MySQL, PostgreSQL, Oracle.

Funkcionalitu programovacího jazyka R lze využívat i v komerčních softwarových systémech jako je SAS, SPSS [2] anebo Stata [19], případně v jiných programech. Uživatelé programu SAS mohou pro programovací jazyk R využít SAS/IML studio, kde příkazem *ExportDatasetToR* se posílají data do programovacího jazyka R. Ke spouštění kódu programovacího jazyka R v prostředí SAS jeho uživatele použijí příkaz „*submit /R*;“ a „*endsubmit*;“. SPSS má implementovanou schopnost spouštět programy R od verze 16, přičemž je nutné nainstalovat do programu SPSS doplněk pro programovací jazyk R. Pro uživatele programu Stata je nejjednodušší cesta použití programovacího jazyka R uložení datových souborů a jejich následný import do programovacího jazyka R. Podobným způsobem lze využívat programovací jazyk R i v jiných statistických programech.

I když je programovací jazyk volně dostupným a velmi výkonným softwarovým systémem, vyznačuje se několika omezeními [2]:

- programovací jazyk R je mnoha uživateli považován jako složitější k pochopení než jiný programový systém. Nápověda k funkcionalitám programovacího jazyka R je napsána pro relativně pokročilé uživatele. Příkladem takové nápovědy je "*funkce print tiskne svůj argument a vrací tento argument neviditelně. Jedná se o generickou funkci, což znamená, že nové tiskové metody mohou být jednoduše přidané do nových tříd*.", což je mnohem méně jasné a zřetelné, než v jiných programech;
- významným omezením je skutečnost, že veškerá zpracovávaná data v programovacím jazyku R musí být v hlavní paměti počítače. Znamená to tedy, že programovací jazyk R je schopen vyhodnotit v jednom okamžiku pouze takové množství údajů, které se vejde do hlavní paměti počítače.

Pro uživatele a vývojáře pracující v prostředí programovacího jazyka R byly založeny také i specializované diskusní fóra, a to uživatelské diskusní fórum *r-help* [21] a *r-dev* [22] jako diskusní fórum pro vývojáře. Tato pomáhají v řešení různých otázek a problémů i pro výměnu zkušeností týkajících se programovacího jazyka R a jeho využívání.

Softwarové funkce, třídy a metody v programovacím jazyku R mohou být definovány a vytvářeny uživatelsky, což poskytuje mnohem více volnosti a flexibility než například nějaký standardní programovací jazyk skládající se z maker. Je nutno poznamenat, že uživatelé programovacího jazyka R mají zcela stejný přístup ke stejným softwarovým nástrojům jako programoví vývojáři. A tedy každý uživatel stejně jako softwarový vývojář může vytvářet doplňkové programové funkcionality. To je ten důvod, proč je téměř 6 000 doplňkových programových balíčků (*add-on packages*) připravených ke stažení a použití v rozsáhlém archívu programovacího jazyka R v systému *CRAN*. Nejvýznamnější část rozsáhlého archívu je hostována na serveru *Institute for Statistics and Mathematics* [8] na *Wirtschaftsuniversität* Wien [42].

V současné době je programovací jazyk R vyučován především na matematickofyzikálních fakultách, a to na Karlově univerzitě v Praze i na Univerzitě Komenského v Bratislavě hlavně jako integrální součást výuky matematiky a fyziky i aktuárských věd. Především z tohoto důvodu je literatura ke studiu o programovacím jazyku R, případně základů jazyka R pro statistickou analýzu dostupná na uvedených českých a slovenských univerzitách. Vznikají i internetovské (on-line) učebnice a kurzy programovacího jazyka R, čehož příkladem je učebnice R jako programovací jazyk [23].

3. DOPLŇKOVÉ PROGRAMOVÉ BALÍČKY (ADD-ON PACKAGES) PRO POUŽITÍ VE STÁTNÍ STATISTICE

Všechny doplňkové balíčky pro použití v rámci programovacího jazyka R jsou umístěny v archívu doplňkových balíčků *CRAN* v sekci *CRAN Task Views* [39]. Pro státní (oficiální) statistiku jsou doplňkové funkce a metody obsaženy a stručně popsány v sekci *The CRAN Task Views on Official Statistics and Survey Methodology*. Mezi témata státní (oficiální) statistiky jsou uvažovány následující oblasti statistiky:

- návrh komplexních statistických zjišťování;
- editace a vizuální inspekce mikrodat;
- imputace;
- řízení přístupu ke statistickým údajům (*statistical disclosure control*);
- sezónní očišťování;
- statistical record matching;
- odhady v malých oblastech (*small area estimation*);
- indexy a indikátory.

Vybrané softwarové doplňky byly vytvořeny odborníky z metodického odboru Rakouského statistického úřadu (některé z nich ve spolupráci i s jinými institucemi). Podrobnější přehled vybraných doplňkových balíčků bude náplní následující části tohoto příspěvku.

V následující části budou přehledně popsány opravdu jen ty nejvýznamnější programové balíčky, které se dotýkají možností pro státní statistiku. Prostředí programovacího jazyka R může využívat velký počet doplňkových instalací, přičemž každý takový doplněk může být specializován na konkrétní oblast. Každý programový balíček má v sobě zabudován určité množství programových funkcí (v mnoha případech se jedná o desítky), které lze po jeho instalaci využívat v programovacím jazyku R. I z tohoto důvodu není reálné provést výčet všech možností programového doplňkového balíčku. Proto bude proveden pouze popis nejdůležitějších funkcí u vybraných doplňkových balíčků.

3.1. Doplňkové programové balíčky *survey* a *sampling*

Doplňkový programový balíček *survey* [34] byl naprogramován z pohledu splnění 2 hlavních důvodů [14]:

- spojení nezbytných designových metadat k datům, aby mohly být provedeny spolehlivě a automaticky správné analytické úpravy. Tento požadavek je zajištěn konstrukcí na programovaných funkcí, které jsou využívány při řešení návrhu designu šetření a odhadu příslušných populačních parametrů;
- poskytovat platné odhady rozptylů pro statistiky počítané nad daty. Odhad rozptylu je realizován buď pomocí replikací anebo linearizace využitím Taylorových řad.

Výpočty se provádějí využíváním již existujících funkcí v programovacím jazyku R, například pro modelování zobecněnými lineárními modely s modifikovanými úpravami pro získání správných hodnot variancí.

Doplňkový programový balíček *sampling* [26] obsahuje především funkce pro realizaci výběrů z opory a kalibraci vah pro odhady populačních parametrů. Použití tohoto doplňkového balíčku poskytuje [40]:

- jedno- a dvojstupňovou stratifikaci, výběry s nestejnými pravděpodobnostmi, vyvážené výběry;
- odhady s kalibrovanými vahami, regresní odhady;
- výpočty pravděpodobností zahrnutí, řeší problematiku překrývajících se strat;
- obsahuje 2 databáze švýcarské a belgické municipality.

3.2. Doplňkové programové balíčky *stratification* a *SamplingStrata*

Doplňkový programový balíček *stratification* [33] umožňuje jednorozměrnou stratifikaci populace pomocí zobecněné metody konstrukce strat podle Lavalleea-Hidirogloua. Tato zobecněná metoda bere do úvahy rozdíly mezi proměnnou stratifikační a proměnnou šetřenou. Pro alokaci ve stratech lze určit optimální meze, a to i za podmínky předpokládané míry neodpovědí, i výběry s nestejnými pravděpodobnostmi zahrnutí (například úměrnými velikosti vybíraných jednotek). V balíčku je také implementována alokace na základě pravidla kumulované frekvence od Dalenia a Hodgese a geometrického pravidla Gunninga a Horgana. Doplněk lze použít na výběry z populací jednostupňové i vícestupňové podle různých výběrových plánů. Uživatelům také poskytuje odhady z výběrových šetření včetně kalibrace pravděpodobnostních vah. Při použití useknutých a logitových metod kalibrační váhy by měly ležet ve specifikovaném rozsahu. Také provádí kontrolu validity kalibrace.

Doplňkový programový balíček *SamplingStrata* [27] nabízí v oblasti stratifikovaného designu zjišťování přístup k vymezení nejlepší stratifikace opory, což zajišťuje minimální náklady výběru za podmínek splnění předem stanovené přesnosti na odhady parametrů. Uvedené řešení je založeno na použití genetického algoritmu, kde každé řešení (tj. velikost vzorku v jednotlivém stratu) je považováno za individuální v celé populaci. Funkce v programovém doplňku umožňují:

- analyzovat získané výsledky z optimalizačního kroku;
- označit vytvořená strata v opoře výběru;
- vybírat vzorky jednotek z opory podle vymezené nejlepší alokace.

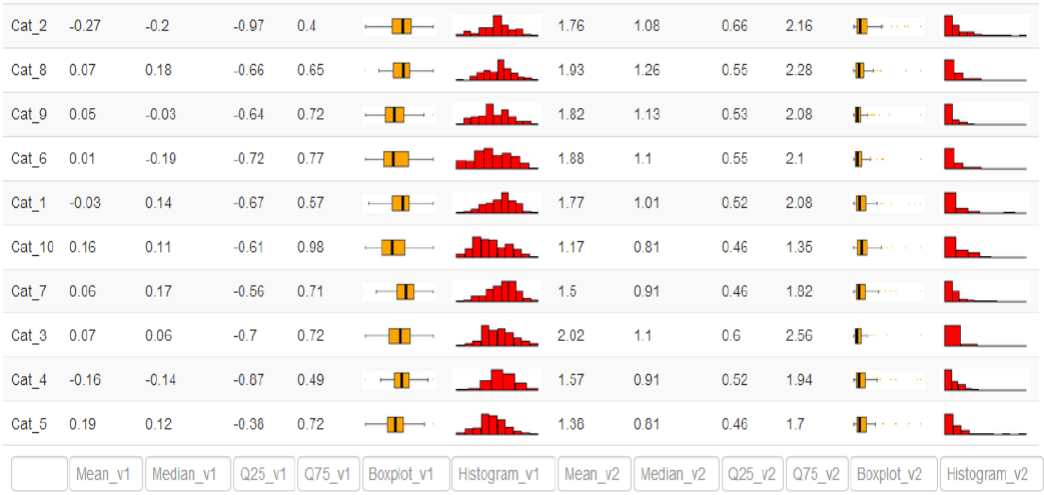
3.3. Doplňkový programový balíček *laeken*

Programový balíček *laeken* určený pro programovací jazyk R je objektově orientovaný nástroj pro odhady indikátorů z komplexních statistických zjišťování pomocí standardních nebo robustních metod [1]. Balíček se zejména používá k odhadům indikátorů sociálního vyloučení a chudoby, což předurčilo název tohoto doplňkového balíku. Umožňuje také využití kalibrované bootstrapové metody odhadu rozptylu indikátorů u většiny výběrových plánů (designu zjišťování). Balíček dále obsahuje synteticky generovaná (blížící se realitě) data pro EU SILC a SES, která jsou použita k demonstraci možností funkcí tohoto softwarového doplňku.

3.4. Doplňkový programový balíček *sparkTable*

Doplňkový programový balíček *sparkTable* podporuje metody pro tvorbu statistických tabulek s různými typy minigrafů, které mohou být použity na webových stránkách, prezentacích i dokumentech [37]. Jedná se o druh grafů, které představují jednoduché, výrazné a ilustrativní grafy dostatečně malé, aby se vešly do jednoho řádku. Balíček zahrnuje minigrafy od čárových až po sloupcové (histogramy), a tak umožňuje tvorbu uživatelských grafů vložených do standardních statistických tabulek. Ukázka vizualizace dat pomocí funkcí tohoto doplňkového balíčku je připojena níže (viz obrázek 1).

Obrázek č. 1: Ukázka statistické tabulky s minigrafy



Zdroj: [15]

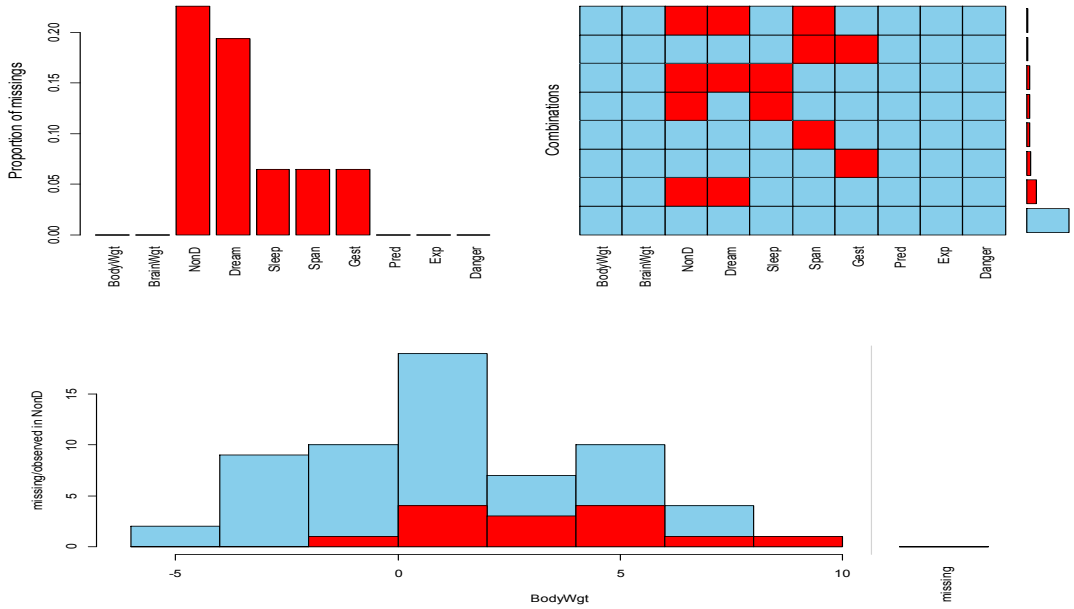
3.5. Doplňkové programové balíčky *sdcMicro*, *sdcMicroGUI* a *sdcTable*

Data z národních statistických úřadů a jiných institucí jsou většinou důvěrné. Doplňkové programové balíčky *sdcMicro* a *sdcMicroGUI* mohou být použity pro generování anonymizovaných mikrodat. Toto znamená, že tyto nástroje slouží ke tvorbě souborů mikrodat určených pro veřejnost nebo pro vědecké účely [28]. Uvedený nástroj je navíc vybaven různými metodami odhadu rizik. Uživatelský komfort při používání tohoto nástroje je zajišťován balíčkem *sdcMicroGUI* [29], který poskytuje grafické rozhraní pro tyto různé metody ochrany dat. Balíček *sdcTable* poskytuje možnost ochrany dat, které jsou rozmístěny ve statistických tabulkách. Uvedený nástroj umožňuje řízení přístupu uživatelů pro tabulkové údaje (více či méně agregovaná) [30].

3.6. Doplňkové programové balíčky *VIM* a *VIMGUI*

Funkce programového doplňkového balíčku jsou vhodné pro průzkum dat a struktury chybějících údajů a jejich vizualizaci (viz obrázek 2). V závislosti na struktuře chybějících hodnot může tento nástroj být užitečný pro identifikaci mechanismu generujícího chybějící hodnoty. Znalost tohoto mechanismu je důležitá pro výběr vhodné imputační metody na spolehlivý odhad chybějících hodnot [35]. Nástroj poskytuje uživatelům implementaci různých algoritmů pro imputaci chybějících údajů.

Obrázek č. 2: Ukázka grafické analýzy chybějících hodnot pomocí balíčku VIM



Zdroj: vlastní výpočty

Grafické rozhraní tohoto vizualizačního nástroje je velmi uživatelsky příjemné a dovoluje jednoduchý výběr jednotlivých zobrazovacích metod. Nástroj nemusí být používán pouze ve státní statistice, ale také nad daty různých oblastí. Zvláštní pozornost je věnována prostorovým datům, u kterých mohou být chybějící hodnoty zobrazovány pomocí map.

3.7. Doplňkové programové balíčky *x12* a *x12GUI*

Sezónní očišťování metodou X-12-ARIMA se používá nad měsíčními a čtvrtletními řadami jako standardní metoda sezónních úprav v mnoha národních statistických úřadech. Procedura sezónních úprav používá aditivní nebo multiplikativní metody očišťování a vytváří výstupní data obsahující adjustované časové řady a bezprostřední výpočty [12]. Doplňkový programový balíček nabízí programové funkce ke zpracování vhodných časových řad metodou X-12-ARIMA. Umožňuje sumarizaci, modifikaci a ukládání výstupu z metody X-12-ARIMA objektově orientované implementace. Grafické uživatelské rozhraní umožňuje uživatelům přístup k metodě X-12-ARIMA bez větších znalostí programovacího jazyka R. Uživatelé mohou interaktivně vybírat odlehlé hodnoty, úrovně posunu a dočasné změny a zaznamenávat bezprostřední účinek těchto změn. Pomocí tohoto nástroje je metoda sezónního očišťování X-12-ARIMA dostupná přímo z prostředí programovacího jazyka R všem uživatelům, což dává všem obrovské možnosti pro zpracování sezónních časových řad.

3.8. Doplňkové programové balíčky *editrules* a *deducorrect*

Funkce programových doplňků *editrules* a *deducorrect* slouží k automatizované editaci dat a jejich opravě pomocí uživatelsky nastavených lineárních i nelineárních omezení. Doplňkový programový balíček *editrules* [11] je navržen s cílem usnadnit automatizovanou editaci dat pomocí funkcionality pro čtení a manipulaci editačních pravidel k jejich použití nad daty a lokalizovanými chybami na základě (zobecněného) Fellegiho-Holtova principu (1976) [10]. Doplněk *deducorrect* využívá metody pro deduktivní opravy znaménka, zaokrouhlování a chyb zápisu. Principem

balíčku *deducorrect* je využití informace v chybném záznamu k vystopování správné hodnoty [9]. Lze také využít automatizovaných pravidel k odvození možných řešení (variant správných hodnot).

4. PŘÍKLADY STATISTICKÝCH ÚŘADŮ VYUŽÍVAJÍCÍ PROGRAMOVACÍ JAZYK R V PRODUKČNÍM PROCESU

4.1. Statistický úřad Slovenské republiky

Použití programovacího jazyka R ve Statistickém úřadu SR se koncentruje v produkčním statistickém procesu především na oblast kalibrace pravděpodobnostních vah. K tomuto účelu byl na úřadu vyvinut sofistikovaný nástroj kalibrace vah v programovacím jazyku R [41] nazývaný CALIF[3]. CALIF nabízí jednoduše použitelné a uživatelsky příjemné grafické rozhraní, které zvyšuje kvalitu odhadů ve statistickém zjišťování. Příkladem použití tohoto nástroje na kalibrování vah je statistika příjmů a výdajů. V ostatních oblastech statistického produkčního procesu není používání programovacího jazyka R příliš rozšířeno a ve statistickém produkčním procesu převažuje programový systém SAS.

4.2. Italský statistický úřad

Programovací jazyk R se speciálně vyvinutými programovými doplňky Národní statistický úřad Itálie využívá především ve fázi návrhu statistického zjišťování a ve fázi zpracování sesbíraných statistických dat [16]. Vybrané softwarové doplňky byly vytvořeny odborníky z metodického odboru italského Národního statistického úřadu (některé z nich ve spolupráci i s jinými institucemi). Přehled vybraných doplňkových balíčků bude v krátkosti proveden v následující části tohoto příspěvku.

Při návrhu opory statistického zjišťování a realizaci výběrů zpravodajských jednotek jsou v italském Národním statistickém úřadu v rámci programovacího jazyka R používány následující programové doplňky:

- doplněk *FS4* [6] (název doplňku je odvozen od anglického *First Stage Stratification and Selection in Sampling*) implementuje metody pro stratifikaci prvního stupně a výběru zpravodajských jednotek do výběrového souboru ve dvou nebo více stupních. Grafické rozhraní tohoto doplňku je velmi uživatelsky příjemné a dovoluje velmi intuitivní výběr jednotlivých statistických metod;
- doplněk *Mauss-R* [15] (anglicky *Multivariate Allocation of Units in Sampling Surveys*, z čehož se odvíjí i název doplňku) – jedná se o programový doplněk pro programovací jazyk R, s jehož pomocí je vymezena výběrová alokace v mnohorozměrném případě a pro více domén odhadu současně v jednostupňových výběrech;
- doplněk *SamplingStrata*, který byl zmíněn již v podkapitole 3.2 tohoto příspěvku.

V rámci zpracování (editace a imputace údajů a případného kódování) dat sesbíraných ze statistického zjišťování italským Národním statistickým úřadem jsou v programovacím jazyku R uplatňovány následující programové doplňky:

- v oblasti integrace dat se jedná o doplňky:
  - doplněk *RELAIS* (anglicky REcord Linkage At IStat), který představuje nástroj pro editaci a imputaci údajů;

[3] *Název CALIF je odvozen od anglického výrazu „Calibration“ a iniciálů autora tohoto kalibračního nástroje (CALI+F).*

  - doplněk *StatMatch* [32], jehož úlohou je řešení zejména imputačních úloh ve statistickém zjišťování;
- v oblasti kódování textových odpovědí se v úřadu využívá programový doplněk *CIRCE* [4] (název byl vytvořen z anglického názvu *Comprehensive Istat R Coding Environment*), který implementuje automatizovaný kódovací systém pro textové odpovědi ve statistických zjišťováních;
- k detekci chyb měření a imputací částečných (položkových) neodpovědí úřad využívá doplňky programovacího jazyka R, a to *CANCEIS* [3] (*CANadian Census Edit and Imputation System*), *CONCORDJava* [5] (podle názvu *CONtrollo e CORrezione dei Dati version with Java interface*). Také používá programové doplňky specializované na editaci a imputaci chybějících údajů *SeleMix* [31] (podle anglického názvu *Selective editing via Mixture models*), které současně slouží k identifikaci zpravodajských jednotek s potenciálně velkým vlivem na odhadované hodnoty, apod.

Pro odhady, vážení, kalibraci vah i vyhodnocování výběrových chyb se v italském statistickém úřadu používá programový doplněk *ReGenesees* [24] (*R Evolved Generalized Software for Sampling Estimates and Errors in Surveys*). Tato procedura patří ke komplexním programovým doplňkům, které zahrnují velmi rozmanité funkcionality z oblasti vyhodnocování statistických šetření.

Zmiňované programové doplňky pro programovací jazyk R byly vytvořeny zaměstnanci italského Národního statistického úřadu s cílem nastavení moderních standardů statistického produkčního procesu pro všechny fáze návrhu, zpracování a vyhodnocování statistických šetření.

4.3. Holandský statistický úřad

Holandský statistický úřad začal v roce 2010 s pilotním projektem ke zjištění nejoptimálnějšího způsobu rozložení nástrojů programovacího jazyka R v rámci své organizace. Úspěchem tohoto pilotního projektu byl výběr sady programových balíčků jazyka R jako standardních pro uživatele a sady programových balíčků jako podpůrných nástrojů. Výsledkem byla distribuce programovacího jazyka R ve 3 variantách, která byla odvislá na použití programovacího jazyka R ve statistické produkci, statistickém výzkumu nebo výzkumu v metodách a výpočtech [13].

K zajištění dostatečné míry kompetencí v používání programovacího jazyka R musel každý nový uživatel projít několikadenním tréninkovým kursem vytvořeným expertním centrem. Tímto kurzem prošla během 18 měsíců více než stovka zaměstnanců. Mimo tohoto kurzu byly často organizovány setkání uživatelů, na kterých uživatelé prezentovali svoje projekty a programové kódy vytvářené na základě programování v jazyku R. Příležitostně byli zváni experti, kteří předávali své zkušenosti.

Spolu se členy pilotního projektu a zaměstnanci IT oddělení byly vyvinuty tzv. kódovací standardy, které vymezují základní zásady architektury softwarového řešení v programovacím jazyku R. Vedle kódovacích standardů byly také vypracovány „software development" standardy, které obsahují doporučení, jak budovat a distribuovat rozvoj produkčního statistického procesu pomocí programovacího jazyka R. Jak kódovací standardy, tak i „software development" standardy jsou součástí procesu řízení kvality v Holandském statistickém úřadu.

4.4. Rakouský statistický úřad

V Rakouském statistickém úřadu je v současnosti programovací jazyk R instalován na více než 60 počítačích s operačním systémem založeným na platformě Windows (především na Windows 7), dále pak založených na virtuálních serverech typu PowerPC a SUSE Linux Enterprise Server. Uvedené serverové řešení je použito zejména z důvodu pokrytí velké paměťové náročnosti a paralelního zpracování [36].

Vedení R-teamu (který zastřešuje správu a využívání programovacího jazyka R) pozůstává v Rakouském statistickém úřadu z 3 expertů z metodického odboru. Navíc každý odbor má 1 pracovníka vybraného jako první kontaktní osobu pro otázky a problémy týkající se programovacího jazyka R, které mohou být jednoduše zodpovězeny. Dále byla v úřadu vytvořena následující organizační struktura zajišťující příslušné využívání programovacího jazyka R:

- experti R-teamu v metodickém odboru (administrátoři) se starají o užívání správné verze programovacího jazyka R, rozhodují o vzhledu a funkcích programovacího jazyka R a potřebných doplňkových programových balíčcích pro defaultní (implicitní) instalaci. Všechny nutné informace a soubory (samotný programovací jazyk R, Rstudio [25], doplňkové balíčky, dokumentace a příklady) jsou uloženy na vyhrazeném serveru;
- distribuci potřebných souborů a instalací programovacího jazyka R zajišťuje uživatelům IT odbor (včetně vzhledu a funkcí prostředí jazyka R, tj. Rstudia). Toto zajišťuje výlučně standardizované instalace na všechny počítače;
- obecná podpora programovacího jazyka R je centralizována prostřednictvím uživatelského emailového seznamu *R-podpora* (vyjma otázek, které mohou být zodpovězeny pracovníky prvního kontaktu);
- vytvoření vnitřní znalostní báze, která je použita ke shromažďování informací a zkušeností o programování v programovacím jazyku R;
- přístupová práva na servery, k souborům, IT hotline, apod. jednotlivým konkrétním uživatelům jsou definována administrátory metodického odboru. Momentálně jsou založeny 2 základní skupiny uživatelů, a to:
  - administrátoři programovacího jazyka R s plným přístupem a odpovědností za složku obsahující úplné softwarové vybavení, dokumentaci i interní znalostní databázi, atd.;
  - team R s omezeným přístupem (pouze čtení) k dokumentaci o programování v programovacím jazyku R a plným přístupem ke znalostní databázi a členskými právy v uživatelském emailovém seznamu *R-podpora*.

Rakouský statistický úřad nabízí svým zaměstnancům k prohlubování znalostí o programování v programovacím jazyku R kurzy ve 2 znalostních úrovních, a to kurz pro základní uživatele a kurz pro pokročilé uživatele [36]. Do úřadu také nastupuje mnoho absolventů, kteří začínají pracovat na úseku státní statiky a už ovládají programování v programovacím jazyku R [7].

5. ZÁVĚRY A DOPORUČENÍ

Výše provedený přehled statistických úřadů zahrnuje pouze vybrané statistické úřady, ve kterých se významněji využívá programovací jazyk R ve statistickém produkčním procesu. Toto se týká zejména statistického úřadu v Rakousku a statistického úřadu v Holandsku a částečně i statistického úřadu v Itálii. Všechny tyto úřady systematicky přecházejí na produkční statistický proces založený na softwarovém řešení v prostředí programovacího jazyka R. Ve jmenovaných

institucích byl dosažen cíl implementace metod statistického produkčního procesu pomocí využití programovacího jazyka R a snižování závislosti produkčního statistického procesu na programovém systému SAS. Vytvořením softwarové infrastruktury na bázi programovacího jazyka R včetně tréninku a podpory je použití programovacího jazyka R v statistickém produkčním procesu realizovatelné a mnohdy upřednostňované před jinými softwarovými řešeními.

Využití specializovaných programových balíčků v rámci programovacího jazyka R na metody státní statistiky dovoluje zvládat problémy, které nejsou jednoduše řešitelné ostatními statistickými programy. Toto zahrnuje především výběrová šetření, kalibraci, editaci a imputaci dat, řízení přístupu k datům stejně jako odhadování, vizualizaci a další odborné otázky státní statistiky. Neustálý vývoj nových programových balíčků, na jejichž vzniku se zpravidla podílejí významní odborníci i zkušení praktici v této oblasti, zajišťuje možnost přístupu k moderním metodám výběrového šetření (i jiných oblastí statistiky a pravděpodobnosti) celé komunitě uživatelů programovacího jazyka R. Navíc většina nových zaměstnanců s akademickým vzděláním ve statistice již ovládá programování v programovacím jazyku R a jsou velmi motivovaní pokračovat v jeho využívání.

Cílem článku bylo především upozornit na obrovský potenciál a velmi vhodné možnosti používání programovacího jazyka R ve státní statistice. Článek dále poskytl – aspoň pro některé programovatelné balíčky, které jsou zajímavé z pohledu metod státní statistiky – stručný přehled a předložil pár jednoduchých a názorných příkladů praxe v používání programovacího jazyka R.

LITERATURA

[1] ALFONS, A. – TEMPL, M.: Estimation of social exclusion indicators from complex surveys: The R package laeken. Journal of Statistical Software, Volume 54, Issue 15, August 2013, pp. 1-25. ISSN 1548-7660. WWW: <http://www.jstatsoft.org/v54/i15/paper>.

[2] Calling R from Other Software [online]. 26. 04. 2016. [cit. 18. 06. 2016]. WWW: <http://r4stats.com/articles/calling-r/>.

[3] CANCEIS [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/canceis>.

[4] CIRCE [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/circe>.

[5] CONCORDJava [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW:

[6] FS4 (First Stage Stratification and Selection in Sampling) [online]. 16. 05. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/design-tools/fs4>.

[7] GENTLEMEN, R.: Data analysts captivated by R' power, 2009. [online]. 07. 01. 2009. [cit. 23. 04. 2016]. WWW:<http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all&_r=0.A>.

[8] Institute for Statistics and Mathematics [online]. 13. 05. 2011. [cit. 23. 04. 2016]. WWW: <http://statmath.wu.ac.at/>.<http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/concordjava>.

[9] JONGE de, E. – van der LOO, M.: Data editing with editrules and deducorrect. amst-R-dam user meeting 02. 04. 2012. Amsterdam, 2012. WWW: <www.markvanderloo.eu/files/statistics/amstRdam.pdf>.

[10] JONGE de, E. – van der LOO, M.: Error localization as a mixed-integer programming problem with editrules. Discussion paper 2014/07, Statistics Netherlands, The Hague. 2014. WWW: <https://www.cbs.nl/-/media/imported/documents/2014/15/2014-07-x10-pub.pdf>.
[11] JONGE de, E. – van der LOO, M.: Manipulation of linear edits and error localization with the editrules package. Technical Report 201120, Statistics Netherlands, The Hague. 2011. WWW: <https://www.cbs.nl/-/media/imported/documents/2011/36/2011-x10-20.pdf>.
[12] KOWARIK, A. – MERANER, A. – TEMPL, M. – SCHOPHAUSER, D.: Seasonal Adjustment with the R Packages x12 and x12GUI. Journal of Statistical Software, Volume 62, Issue 2, November 2014, pp. 1-21. WWW: <https://www.jstatsoft.org/article/view/v062i02/v62i02.pdf>.
[13] LOO, van der, M.: The introduction and use of R software at Statistics Netherlands. In Proceedings of the Third International Conference of Establishment Surveys (CD-ROM). Canada, Montreal, 2012. American Statistical Association [online]. 07. 01. 2009. [cit. 23. 04. 2016]. WWW: <http://www.amstat.org/meetings/ices/2012/papers/302187.pdf>.
[14] LUMLEY, T.: Analysis of complex survey samples. Journal of Statistical Software. Volume 9, Issue 8, April 2004, pp. 1-19. ISSN 1548-7660. WWW: <https://www.jstatsoft.org/index.php/jss/article/view/v009i08/paper-5.pdf>.
[15] MAUSS-R (Multivariate Allocation of Units in Sampling Surveys) [online]. 26. 04. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/design-tools/mauss-r>.
[16] Methods and IT Tools for statistical production [online]. 16. 06. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools>.
[17] PÁLEŠ, M.: Grafická podpora jazyka R pri štatistických analýzach. In: Slovenská štatistika a demografia, 2016, č. 1. ISSN 1339-6854 (online), 1210-1095 (tlačené vydanie).
[18] R Developer Page [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <https://developer.r-project.org/>.
[19] R for Stata Users [online]. 26. 04. 2016. [cit. 19. 06. 2016]. WWW: <http://r4stats.com/books/r4stata/>.
[20] R Foundation for Statistical Computing [online]. 01. 02. 2006. [cit. 23. 04. 2016]. WWW: <https://www.r-project.org/foundation/>.
[21] R-help Info Page [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <https://stat.ethz.ch/mailman/listinfo/r-help>.
[22] R Developer Page [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <https://developer.r-project.org/>.
[23] R jako programovací jazyk [online]. 26. 05. 2016. [cit. 20. 06. 2016]. WWW: <http://portal.matematickabiologie.cz/index.php?pg=zaklady-informatiky-pro-biology--analyza-dat-v-r--rozsirene-zaklady-r--r-jako-programovaci-jazyk>.
[24] ReGenesees [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/regenesees>.
[25] RStudio - Open source [online]. 01. 02. 2006. [cit. 24. 04. 2016]. WWW: <http://www.rstudio.org>.
[26] Sampling: Survey Sampling [online]. 01. 07. 2015. [cit. 26. 04. 2016]. WWW: <https://cran.r-project.org/web/packages/sampling/sampling.pdf>.
[27] SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys [online]. 12. 01. 2016. [cit. 27. 04. 2016]. WWW: <https://cran.r-project.org/web/packages/SamplingStrata/SamplingStrata.pdf>.

[28] sdcMicro: Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation [online]. 01. 10. 2015. [cit. 26. 04. 2016]. WWW: <https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf>.
[29] sdcMicroGUI: Graphical User Interface for Package 'sdcMicro' [online]. 06. 05. 2015. [cit. 26. 04. 2016]. WWW: <https://cran.r-project.org/web/packages/sdcMicroGUI/sdcMicroGUI.pdf>.
[30] sdcTable: Methods for Statistical Disclosure Control in Tabular Data [online]. 22. 04. 2016. [cit. 26. 04. 2016]. WWW: <https://cran.r-project.org/web/packages/sdcTable/sdcTable.pdf>.
[31] SeleMix [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/selemix>.
[32] StatMatch [online]. 26. 05. 2016. [cit. 18. 06. 2016]. WWW: <http://www.istat.it/en/tools/methods-and-it-tools/processing-tools/statmatch>.
[33] Stratification: Univariate Stratification of Survey Populations [online]. 01. 07. 2015. [cit. 26. 04. 2016]. WWW: <https://cran.r-project.org/web/ packages/ sampling/sampling.pdf>.
[34] Survey: analysis of complex survey samples [online]. 15. 08. 2014. [cit. 26. 04. 2016]. WWW: <https://cran.r-project.org/web/packages/survey/survey.pdf>.
[35] TEMPL, M. – FILZMOSER, P.: Visualization of missing values using the R-package VIM. Research report cs-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology, 2008. URL <http://www.statistik.tuwien.ac.at/forschung/CS/ CS-2008-1complete.pdf>.
[36] TEMPL, M. – KOWARIK, A. – MEINDL, B.: Development and Current Practice in Using R at Statistics Austria. In: Romanian Statistical Review, No. 2, Vol. 2014. ISIN (printed) 1018-046X, ISSN (on-line) 1844-7694. pp. 172-184. WWW: <http://www.revistadestatistica.ro/wp-content/uploads/2014/07/RRS_2_2014_a14.pdf>.
[37] TEMPL, M. – KOWARIK, A. – MEINDL, B.: sparkTable: Generating Graphical Tables for Websites and Documents with R. In: The R Journal, Volume 7, Issue 1, pp. 24-37. June 2015. ISSN 2073-4859. WWW: <https://journal.r-project.org/archive/2015-1/templ-kowarik-meindl.pdf>.
[38] The Comprehensive R Archive Network [online]. 01. 02. 2016. [cit. 23. 04. 2016]. WWW: <http://cran.at.r-project.org>.
[39] The Comprehensive R Archive Network Task Vies [online]. 01. 02. 2016. [cit. 23. 04. 2016]. WWW: <http://cran.at.r-project.org>.
[40] TILLÉ, Y. – MATEI, A.: The R sampling package. Euskal Estatistika Erakundea, XXII Seminario Internacional de Estadistica, November 2010 [online]. 15.11.2010. [cit. 26. 04. 2016]. WWW: <http://www.eustat.eus/productos Servicios/52.3_R_sampling_package.pdf>.
[41] VLAČUHA, R. – FRANKOVIČ, B.: The Calibration of Weights by Calif Tool in the Practice of the Statistical Office of the Slovak Republic. In: Romanian Statistical Review, No. 2, Vol. 2016, ISIN (printed) 1018-046X, ISSN (online) 1844-7694, pp. 153-164. WWW: <http://www.revistadestatistica.ro/wp-content/uploads/2015/04/RRS2_2015_A15.pdf>.
[42] Wirtschaftsuniversität Wien [online]. 10. 02. 2016. [cit. 23. 04. 2016]. WWW: <https://www.wu.ac.at/>.

RESUME
The aim of this article was to provide a brief overview of the R programming language for the purposes of state statistics. Currently, the national statistical offices using the R programming language are constantly increasing, not only for supporting or simulation tasks, but within the statistical production process as well. Expanding the functionality of this program is made by using additional software packages. The article also presents some selected program packages applicable to state statistics.

PROFESNÍ ŽIVOTOPIS
***Ing. Roman Pavelka, PhD.**, v letech 1995 – 2010 pracoval v poradenské společnosti Trexima, s. r. o. Na pozici statistik-analytik se zabýval analýzami zejména mzdových a personálních dat. Podílel se na tvorbě pravidelných statistických přehledů a reportů. Spolupracoval s akademickými pracovišti, agenturami i soukromými subjekty na realizaci a vyhodnocování ad-hoc statistických výzkumů. Oblast jeho vědeckého zájmu představují výběrová šetření, odhady a statistické modely. V letech 2012 až 2013 se zúčastnil zahraniční stáže ve Velké Británii. Od roku 2013 působil v Národním ústavě certifikovaných měření vzdelávaní (NÚCEM), kde zajišťoval statistické vyhodnocování výsledků testování žáků a studentů. Od roku 2015 pracuje v odboru metod statistických zjišťování Statistického úřadu SR.*

KONTAKT
Roman.Pavelka@statistics.sk