MAPY AKO NÁSTROJ DÁTOVEJ VIZUALIZÁCIE

Michal Páleš
Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky
Ekonomickej univerzity v Bratislave
Dolnozemská cesta 1, 852 35 Bratislava, SK; pales.euba@gmail.com

Abstract: *One of the options that best presents the content of the data and its character is to visualize it. This is a representation of required data in a relevant and fetching form for further decision making. Data science is one of the fastest growing and most prosperous areas of the 21st century. Part of data science is also the presentation of the results, especially in graphical form. We will know several possibilities of data visualization. In the paper, we focus on visualization through presentation maps (static and dynamic) using the functionality of the open-source programming language R.*

Key words: R language, data science, data visualization, maps

ÚVOD

R je programovací jazyk špecializovaný predovšetkým na štatistické výpočty a grafiku. Ide o projekt GNU podobný jazyku S, ktorý vyvinul v Bell Laboratories (predtým AT&T, teraz Lucent Technologies) John Chambers so svojím kolektívom. V roku 2009 New York Times zverejnil článok, v ktorom jazyk R získava veľké uznanie medzi dátovými analytikmi, a naznačil, že predstavuje vážnu konkurenciu pre komerčný softvér. V posledných rokoch v rôznych prieskumoch sa dostáva do top 10 najpopulárnejších programovacích jazykov sveta. Jazyk R je voľne dostupný a využíva ho akademická, vedecká i komerčná sféra. Rôzne analýzy možno realizovať už v štandardnej verzii, prípadne po inštalácii konkrétnych podporných balíčkov (knižníc, *packages*), kde je implementované veľké množstvo pokročilých funkcií. Samozrejmosťou je aj možnosť vytvárať vlastné funkcie a skripty. Medzi hlavné výhody systému R patrí:

- *dostupnosť* – open-source programovací jazyk zaradený v rámci projektu GNU nadácie Free Software Foundantion,
- *kompatibilita* – kompatibilný s operačnými systémami Windows, Linux, Mac,
- *množstvo analytických nástrojov* – pre štatistiku, **dátovú vedu**, demografiu, biometriu, taxonómiu, genetiku, geografiu, finančné analýzy a pod. s využitím mnohých vyvinutých doplňujúcich balíčkov s knižnicami funkcií na rôzne typy analýz,
- *aktuálnosť* – rýchle reakcie na vývoj nových metód v štatistike, obsahuje často metódy, ktoré ešte nie sú implementované do klasického komerčného softvéru,
- *grafické výstupy* – štandardné aj nové moderné grafické výstupy s interaktívnym zásahom užívateľa, vrátane možnosti doplniť matematické vzorce a symboly,
- *zdieľanie* – vytvorené kódy možno pohodlne zdieľať buď s inými užívateľmi, resp. v publikáciách, učebniciach, monografiách a pod., čo zaručuje názornejšie pochopenie problematiky (v porovnaní so zdieľaním napr. zošita MS Excel).

Pre viac informácií o jazyku R pozri napr. [1], [3] – [5] a [7].

K rastu popularity jazyka R prispievajú nie len jeho kľúčové uvedené vyššie vlastnosti, ale aj rýchly vývoj v oblasti tzv. *Big Data* a *Data Science*, kde je uplatnenie R veľmi vysoké, nakoľko sa javí ako jeden z najvhodnejších a (vďaka *open source*) najdostupnejších prostriedkov na dátové a grafické analýzy, či rôzne modelovania a predikciu. V súčasnosti aj viaceré významné spoločnosti – ako napríklad SAP, si uvedomujú význam R a vytvárajú produkty, ktoré dokážu priamo s jazykom R natívne pracovať, bez nutnosti zložitého

nastavenia ovládačov a použitia emulátorov. Napríklad jeden z najdôležitejších produktov od SAP-u v súčasnosti je in-memory databáza SAP HANA, umožňuje spúšťanie skriptov napísaných v R priamo v natívnom prostredí databázy.[1]

1. DÁTOVÁ VEDA A VIZUALIZÁCIA ÚDAJOV

Big Data môžeme označiť enormné množstvo údajov (dát), ktoré sa v čase rýchlo a dynamicky menia, pričom tempo rastu týchto dát je čoraz rýchlejšie. Patria sem napríklad každodenné údaje o počasí, letoch, vlakoch, prepravených pasažieroch, ale ja údaje zo sociálnych sietí, rôzne statusy, obrázky či iné príspevky. Ak zoberieme do úvahy napríklad dáta o teplote na celej planéte – každú hodinu na každej meteo stanici na zemi, tak ide o obrovský súbor dát, ktorý dosahuje enormné rozmery, často tempo rastu takýchto súborov dosahuje niekoľko gigabajtov za deň. Na firemnej úrovni sú to napríklad dáta o zamestnancoch, materiáloch, majetku či služobných cestách. Príkladom môžu byť databázy veľkých softvérových spoločností, ktoré vo svojich databázach uchovávajú desiatky miliónov záznamov o zamestnancoch, partneroch či projektoch – všetko sú to rýchlo sa vyvíjajúce a meniace sa dáta. Tieto dáta sa stále hromadia a je ich viac a viac – pritom nie je žiadnym tajomstvom, že takéto záznamy a databázy majú značnú hodnotu. Preto je namieste otázka ako je možné tieto dáta s úžitkom využiť.

Odpoveď na túto otázku sa snaží dať *Data Science* (v doslovnom preklade dátová veda, pozri tiež [6]). Ide o multidisciplinárny vedný odbor, ktorý kombinuje znalosti programovania, štatistiky, matematiky, teórie pravdepodobnosti a analytického myslenia. Podstatou Data Science je analýza spomínaných *Big Data* – ide o snahu nájsť v nich pridanú hodnotu. Pomocou rôznych analýz a reportov sa firmy snažia optimalizovať výrobné procesy, náklady a maximalizovať zisk. V súčasnosti je práca *Data Scientist*-u (teda osoby, ktorý sa takouto analýzou zaoberá), podľa viacerých prieskumov, jedným z najžiadanejších a najlepšie platených povolaní na svete. Veľké korporácie ako napríklad mobilní operátori majú celé vlastné oddelenia takýchto pracovníkov.

Neoddeliteľnou súčasťou *Data Science* je práve *reporting*, tzn. prezentovanie dosiahnutých výsledkov čo najužitočnejšou a najatraktívnejšou formou, napríklad pre manažment podniku. Pre lepšiu prezentáciu výsledkov sa namiesto "obyčajných čísiel" používajú rôzne vizualizácie. Tie sa stávajú čoraz populárnejšou formou zobrazovania informácií – hovoríme o takzvanej *Data Visualization* (dátovej vizualizácií). Jej podstata, ako sme už naznačili, nie je zložitá – pútavým spôsobom sa snažíme zobraziť dáta, ktoré máme k dispozícií. Tieto zobrazenia často bývajú interaktívne, meniace sa v čase, čím sa dosiahne detailný prehľad opisovanej problematiky. Nakoľko rôzne vizualizácie – či už statické alebo dynamické, dotvárajú a zlepšujú výstupy *Data Science*, sú veľmi žiadanou a potrebnou súčasťou dátových analýz a tak *Data Visualization* predstavuje jednu z najdôležitejších súčastí samotnej *Data Science*.

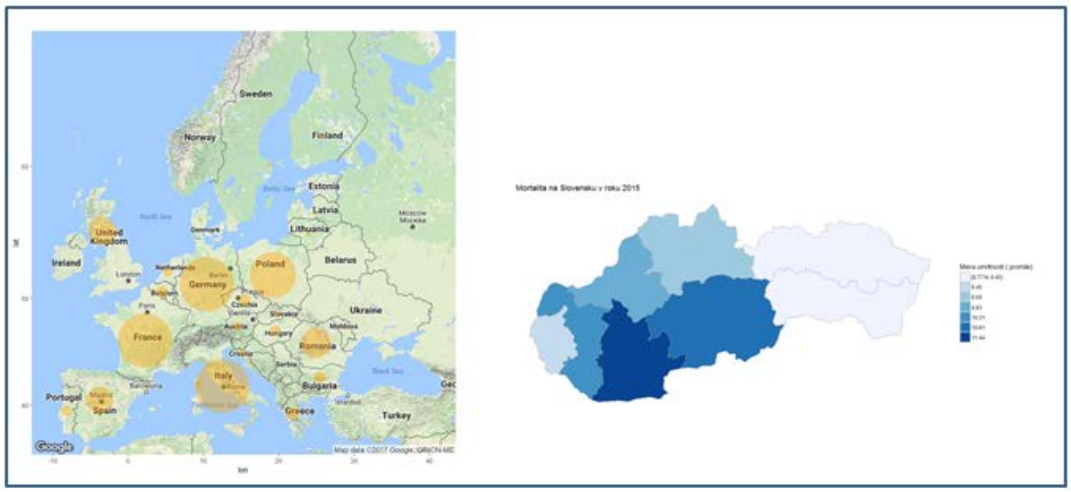
Grafy, diagramy, mapy, obrázky, tabuľky, interaktívne reporty, tzv. *heatmaps* alebo rôzne iné pokročilé vizualizácie historických udalostí môžeme zaradiť do tejto vednej disciplíny. Väčšina významných hráčov na trhu v oblasti informačných technológií ponúka vlastné produkty, ktoré ponúkajú rôzne možnosti a formy vizualizácií – napr. *Power Bi* od spoločnosti Microsoft, *SAP Lumira* od SAP-u alebo *Visual Analytics* od Oracle. Ich výhodou je pomerne jednoduchá implementácia do infraštruktúry a integrácia s rôznymi zdrojmi údajov – relačné databázy, Hadoop, SAP, MS Excel a pod.

Jednou z oblastí, kde má dátová vizualizácia sľubné využitie, je pochopiteľne aj aktuárstvo (pozri [3], [5]) – pomocou vhodne zvolených metód a postupov môžu aktuári zjednodušiť a zefektívniť svoju prácu, napríklad pomocou máp povodní či iných prírodných katastrof

rozdeliť rôzne regióny či krajiny do rizikových oblastí, vizualizovať demografické údaje v súvislosti so životným poistením a pod.[1]

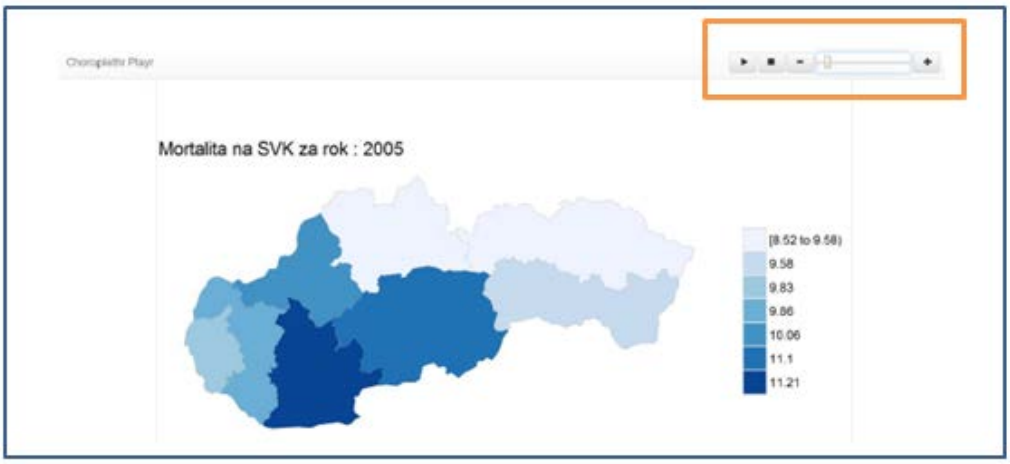
2. VIZUALIZÁCIA PROSTREDNÍCTVOM MÁP V JAZYKU R

Jednoduché je členiť prezentačné mapy na dva hlavné typy – a to na statické a interaktívne (dynamické) mapy. Statickou mapou rozumieme takú mapu, s ktorou po jej vytvorení už nevieme ďalej manipulovať ani interagovať – nevieme ju priblížiť, posúvať, po kliknutí na mapu sa mapa nijakým spôsobom nemení.



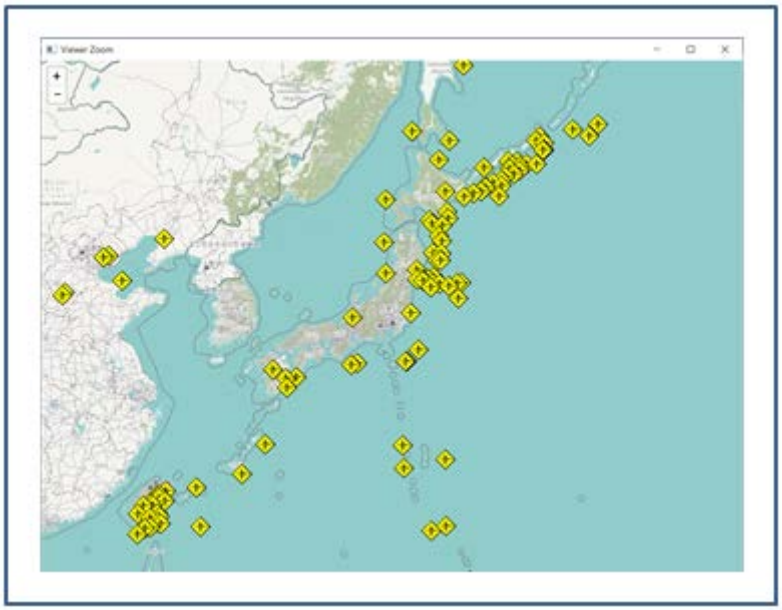
Obr. 1 Mapa smrteľných nehôd v krajinách EÚ za rok 2015 (`ggmaps`)
a mapa mortality na Slovensku za rok 2015 podľa krajov (`choroplethr`)
Zdroj: [1]

Naopak, pri práci s interaktívnymi mapami zobrazujeme tok dát alebo mapa s nami priamo komunikuje po kliknutí alebo priblížení. Pre statické mapy sa najčastejšie v R využívajú knižnice `ggmap` a `choroplethr`. Pre vytváranie interaktívnych máp existuje v R množstvo knižníc, ktoré túto funkcionalitu umožňujú – frekventovaná je knižnica `choroplethr Admin1` a knižnica `leaflet`.



Obr. 2 Dynamická mapa úmrtnosti v SR v rokoch 2005 – 2015;
prostredie *Choroplethr Playr*
Zdroj: [1]

Obrázky 1 až 3 popisujú výstupy z vyššie uvedených knižníc, pričom všetky zdrojové kódy a postupy tvorby týchto máp popisuje práca [1] (autor príspevku bol vedúci tejto záverečnej práce). Tieto pre rozsiahlosť a podmienené príkazy neuvádzame.



Obr. 3 Mapa významných zemetrasení v Japonsku od roku 1950
s použitím preddefinovanej ikony (`leaflet`, funkcia `makeIcon`)
Zdroj: [1]

3. KOMPARÁCIA KNIŽNÍC A INÉ MOŽNOSTI VIZUALIZÁCIE

Každá z knižníc, ktorú sme v tejto práci použili, má svoje výhody aj nevýhody. Pre používateľa je však dôležité poznať možnosti, ktorá knižnica čo ponúka – či dokáže s danou knižnicou generovať interaktívne mapy, či dokáže aktualizovať údaje z internetu alebo či je možné s mapou ďalej pracovať, napr. pridať na ňu nápisy či vysvetlivky. Základné porovnania kľúčových funkcionalít sú dostupné v tabuľke 1.

Podmienka	`ggmap`	`Choropleth/ ChoroplethAdmin1`	`leaflet`
Vlastná databáza údajov		+	
Internetová databáza údajov	+		+
Interaktivita		+	+
Variabilita máp	+		+
Integrácia s Google Maps	+	+	
Dodatočná práca s mapou		+	+

Tab. 1 Porovnanie základných funkcionalít prezentovaných knižníc
v R pre tvorbu máp
Zdroj: [1], upravené

Prostredie jazyka R nie je jedinou formou, ako jednoducho a rýchlo vizualizovať údaje. V článku [2] je uvedený postup obdobnej vizualizácie v MS Excel s využitím doplnkov *Power View* a *Power Map*. Ide o doplnky, ktoré sú dostupné len v programových verziách MS Office Professional Plus (najčastejšie od verzie 2013, resp. 2017) a Office 365 Professional

Plus, takže nejde o základné doplnky, ktoré je možné pridať štandardne do MS Excel a pre ich používanie je potrebné mať licencované spomínané produkty. Oba doplnky poskytujú pomerne jednoduché, kvalitné a interaktívne zobrazovanie údajov. MS Excel samotný ponúka užívateľsky intuitívnu navigáciu, ktorá používateľa sprevádza celým procesom vizualizácie. Ďalšou výhodou je to, že dáta vizualizujeme priamo v prostredí, v ktorom sa nachádzajú – nepotrebujeme žiadne exporty a importy dát, môžeme s nimi pracovať priamo v programe MS Excel. Pre používateľov, ktorí potrebujú len základné dátové vizualizácie a nechcú resp. nepotrebujú sa učiť programovací jazyk, je tak použitie týchto doplnkov vhodnou voľbou.

Obr. 4 Interaktívne zvýraznenie demografických údajov s *pie charts* na mape s využitím doplnku *Power View*
Zdroj: [2]

Na obrázku 4 môžeme vidieť na mape aj vizualizáciu s využitím „koláčového“ grafu pri zobrazení predmetného poľa. V R možno využiť rôzne knižnice (napr. `ggtree`, `ggplot2` a i.), potom ukážku výstupu uvádza obrázok 5.

Obr. 5 Interaktívne zvýraznenie údajov s *pie charts*
Zdroj: https://stackoverflow.com

ZÁVER

Využitie knižnice `ggmap` má nespornú výhodu, že ide o implementáciu Google Maps do prostredia R, a teda knižnica je schopná načítavať údaje priamo zo serverov samotného Google. Potom máme k dispozícii mapy z celého sveta, pričom môžeme využiť rôzne základy máp. Používateľ si môže vybrať a pomocou jedného argumentu zvoliť, či chce používať mapu terénu, satelitné snímky alebo mapu ciest. Tým pádom sa majú podklady máp z `ggmap` univerzálne využitie. Ďalšou veľkou výhodou je aj funkcia `geocode`, ktorá je schopná vrátiť

súradnice zadanej lokality. S jej využitím sme schopný získať a vizualizovať aj dáta, ktorých geografickú polohu nepoznáme. Toto všetko je dostupné jednoducho, pomocou niekoľkých príkazov. Naopak, medzi nevýhody `ggmap` patrí až prílišná statickosť vytvorených máp – ako náhle mapu vytvoríme, nie sme schopní s ňou už ďalej manipulovať, realizovať jej zoom alebo pridanie popisu na mapu nie je možné, a používateľ sa musí uspokojiť so základnou vrstvou mapy.

Najväčšou výhodou knižníc `choroplethr` aj `choroplethrAdmin1` je databáza máp a administratívnych celkov pre všetky svetové krajiny. To pomerne jednoducho umožňuje vizualizovať jednotlivé dáta podľa krajín, bez nutnosti nahrávať dodatočné mapy administratívnych území krajín komplikovaným spôsobom – či už použitím HTML templates alebo máp vo formáte SHP či JSON. To, že tieto knižnice pracujú len s touto spomínanou databázou, je zároveň aj ich nevýhodou – dáta je potrebné mať v potrebnou formáte už pre vstupom do programu.

Knižnica `leaflet` umožňuje tvorbu iba vysoko interaktívnych máp. Takže mapy vytvorené prostredníctvom tejto knižnice dokážeme ľubovoľne oddialiť, priblížiť či posunúť. Vysoká intuitívnosť sa prejavuje aj v možnostiach, ktoré môžeme na mapu pridať a využiť – dokážeme pridať rôzne body, kruhy, ikony, geometrické tvary a pod. Mapa s nami interaguje – dokážeme pridať na mapu rôzne pop-up okná, či odkazy, ktoré sú k dispozícií po kliknutí na mapu a poskytujú používateľovi dodatočné informácie (`leaflet` je prepojená s *OpenStreetMap*).[1]

Ako komparáciu funkcionality iného prostredia pre vizualizáciu sme prezentovali doplnky MS Excel (*Power View* a *Power Map*).

Nakoľko jazyk R sa veľmi rýchlo vyvíja, nemožno vylúčiť, že v blízkej dobe budú k dispozícií ďalšie nové knižnice, ktoré umožnia ešte kvalitnejšie a náročnejšie vizualizácie, prostredníctvom máp.

Je dôležité tiež spomenúť, že v roku 2015 na konferencii používateľov ESRI v San Diegu došlo k oznámeniu iniciatívy na preklenutie R s profesionálnym prostredím geografického informačného systému **ArcGIS**. Dá sa povedať, že v podstate ESRI vytvoril R „knižnicu“ (R-bridge), ktorá je schopná komunikovať a vymieňať dáta medzi ArcGIS a R, takže je možné vytvoriť ArcGIS toolboxes pomocou R skriptov. O túto aplikáciu je medzi používateľmi mimoriadny záujem, pretože R sa v posledných rokoch stal silným nástrojom aj pri analýze priestorových údajov. V rámci geografickej komunity sa R stále považuje za trochu „outsidera“. Je to preto, lebo hlavná aplikácia GIS, t.j. ArcGIS, je založená na jazyku Python a preto napríklad kurzy v oddeleniach geografie a geomatiky majú tendenciu sústrediť sa na výučbu Pythonu namiesto R. Pre viac informácií pozri napr. [8], [9].

LITERATÚRA

[1] MASÁR, Ján. Využitie jazyka R pri tvorbe prezentačných máp. bakalárska práca. Bratislava: Ekonomická univerzita v Bratislave, 2017.

[2] MUCHA, Vladimír. Vizualizácia údajov pomocou doplnkov Power View a Power Map v Microsoft Excel. In *Slovenská štatistika a demografia* : vedecký časopis. - Bratislava : Štatistický úrad Slovenskej republiky, 2017. ISSN 1210-1095, 2017, roč. 27, č. 1.

[3] PÁLEŠ, Michal. *Aktuárstvo v režime Solventnosť II. S riešenými príkladmi v jazyku R*. Bratislava: Vydavateľstvo EKONÓM, 2016. ISBN 978-80-225-4288-3.

[4] PÁLEŠ, Michal. Grafická podpora jazyka R pri štatistických analýzach. In *Slovenská štatistika a demografia* : vedecký časopis. - Bratislava : Štatistický úrad Slovenskej republiky, 2016. ISSN 1210-1095, 2016, roč. 26, č. 1.

[5] PÁLEŠ, Michal. *Jazyk R v aktuárskych analýzach*. Bratislava: Vydavateľstvo EKONÓM, 2017. ISBN 978-80-225-4331-6.
[6] PROVOST, Foster, FAWCETT, Tom. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. 1st. edition. Sebastopol: O'Reilly Media, 2013. ISBN 978-14-493-6132-7.
[7] R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. <www.r-project.org>
[8] https://www.r-bloggers.com/combine-arcgis-and-r-clustering-toolbox/
[9] https://r-arcgis.github.io/

Príspevok bol vytvorený v rámci projektu VEGA č. 1/0120/18 – Moderné nástroje riadenia rizika v interných modeloch poisťovní v kontexte direktívy Solvency II.