ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

FAKULTA ŽIVOTNÍHO PROSTŘEDÍ

KATEDRA VODNÍHO HOSPODÁŘSTVÍ
A ENVIRONMENTÁLNÍHO MODELOVÁNÍ

Vizualizace enviromentálních dat

BAKALÁŘSKÁ PRÁCE

Vedoucí práce: **doc. Ing. Martin Hanel, Ph.D.**

Bakalant: **Irina Georgievová**

2018

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Fakulta životního prostředí

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Irina Georgievová

Vodní hospodářství

Název práce

Vizualizace environmentálních dat

Název anglicky

Visualization of environmental data

Cíle práce

Představení klíčových poznatků týkajících se vizualizace a průzkumové analýzy dat z teoretického hlediska i z hlediska praktické implementace v R. Zhodnoceny budou jak nástroje obsažené v základní distribuci R, tak nástroje dostupné v balících lattice, grid, ggplot2, raster, rasterVis, případně i nástroje pro tvorbu dynamických vizualizací (htmlwidgets, shiny apod.).

Součástí práce bude i vytvoření webové aplikace pro vizualizaci a analýzu hydrologické bilance a předpověď sucha v útvarech povrchových vod ČR.

Metodika

Teoretická část:

- rešerše základních poznatků o vizualizaci dat
- popis vizualizačních prostředků se zaměřením na využití v hydrologii, porovnání výhod/nevýhod
- popis nejpoužívanějších R balíků, jejich základních funkcí a demonstrace jejich využití

Praktická část:

- tvorba aplikace s využitím Shiny dle průběžné specifikace
- stručný popis aplikace v BP

Doporučený rozsah práce

40-60 stran

Klíčová slova

vizualizace dat, grammar of graphics, průzkumová analýza dat

Doporučené zdroje informací

CLEVELAND, W S. *The elements of graphing data.* Murray Hill: AT&T Bell Laboratories, 1994. ISBN 0-9634884-1-4.

TUFTE, E.R. The Visual Display of Quantitative Information. Graphics Press, 1983. ISBN 978-0-961-39210-9.

TUKEY, J.W. Exploratory Data Analysis. Addison-Wesley, 1977. ISBN 0201076160

WICKHAM, H. *Ggplot2 : elegant graphics for data analysis.* Dordrecht: Springer, 2009. ISBN 978-0-387-98140-6.

WILKINSON, L. The Grammar of Graphics. Springer, 2006. ISBN 978-0-387-28695-2.

Předběžný termín obhajoby

2017/18 LS – FŽP

Vedoucí práce

doc. Ing. Martin Hanel, Ph.D.

Garantující pracoviště

Katedra vodního hospodářství a environmentálního modelování

Elektronicky schváleno dne 8. 3. 2018

doc. Ing. Martin Hanel, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 9. 3. 2018

prof. RNDr. Vladimír Bejček, CSc.

Děkan

V Praze dne 09. 03. 2018

Prohlášení:
Prohlašuji, že jsem bakalářskou práci *Vizualizace enviromentálních dat* zpracovala samostatně. Veškerou literaturu a další podkladové materiály uvádím v seznamu na straně

V Praze dne .....

.....
Irina Georgievová

Poděkování:

Abstrakt

Vložte abstrakt o rozsahu cca 100–200 slov.

Klíčová slova: vizualizace dat, grammar of graphics, průzkumová analýza dat

Abstract

Keywords: Data visualization, grammar of graphics, exploratory data analysis

Obsah

Úvod **1**

Cíle práce **2**

Teoretická část **3**

1 Vizualizace dat . . . 3
1.1 Historie vizualizace dat . . . 3
1.2 Zásady vizualizace dat . . . 6
1.2.1 Edward Tufte . . . 6
1.2.2 Wiliam S. Cleveland . . . 8
1.3 Grammar of graphics . . . 11
2 Základní grafy v R . . . 12
2.1 Bodový graf . . . 12
2.2 Liniový graf . . . 12
2.3 Vykreslení rozdělení v R . . . 13
2.3.1 Q-Q graf a P-P graf . . . 15
2.3.2 Krabicový graf . . . 16
2.4 Sloupcový graf . . . 17
2.4.1 Histogram . . . 17
2.4.2 Koláčový graf . . . 19
2.4.3 Číslicový histogram (*stem-and-leaf*) . . . 20
3 Průzkumová analýza dat . . . 21
3.1 Odlehlá pozorování . . . 22
3.1.1 *Jackknife* . . . 22
3.1.2 Mahalanobisovy vzdálenosti . . . 23
3.1.3 Leverages . . . 24
3.2 Náhrada chybějících pozorování . . . 25
3.3 Transformace dat . . . 25
3.4 Ověřování normality . . . 26
4 Pokročilá vizualizace v R . . . 28
4.1 Balíčky pro vizualizaci dat . . . 28
4.1.1 `grid` . . . 28
4.1.2 `lattice` . . . 28
4.1.3 `ggplot2` . . . 29
4.2 Balíčky pro prostorovou vizualizaci . . . 31
4.3 Balíčky pro interaktivní vizualizaci dat . . . 32
4.3.1 `Plotly` . . . 33
4.3.2 `dygraphs` . . . 33
4.3.3 `Leaflet` . . . 33
4.4 Balíčky pro webové aplikace . . . 34
4.4.1 `flexdashboard` . . . 34
4.4.2 `Shiny` . . . 34

Praktická část **37**
5 Metodika . . . 37
5.1 Technické řešení . . . 37
5.2 Postprocessing . . . 37
6 Základní rozvržení . . . 38
6.1 Základní mapa . . . 38
6.2 Indikátory sucha . . . 40
6.3 Užívání . . . 41
6.4 Validace . . . 42
6.5 O projektu . . . 43

Výsledky **44**

Diskuse **45**

Závěr **46**

Přílohy **47**
Seznam obrázků . . . 47
R kód . . . 48
Literatura . . . 49

Úvod

Cíle práce

Teoretická část

1 Vizualizace dat

1.1 Historie vizualizace dat

Do 17. století jediné, co by se dalo klasifikovat jako vizualizace dat byly mapy pro navigaci a průzkum, ale také diagramy, geometrická schémata a tabulky pozic hvězd a jiných nebeských těles. Postupný vývoj statistické teorie a růst zájmu o data na konci 18. století vedly k inovacím a expanzi nových grafických forem. Kartografové se pokoušeli zaznamenat více, než pouhou geografickou polohu na mapě a objevily se první pokusy o tematické mapování geologických, ekonomických a medicínských dat.

Wiliam Playfair (1759-1823) je obecně znám jako průkopník v oblasti vizualizace dat a je považován za vynálezce několika typů grafů. Například liniový a sloupcový graf či grafy časových řad byly popsány v jeho atlasu z roku 1786 *„Commercial and Political Atlas“*[1]. Později popsal i koláčový graf ve svém breviáři *„Statistical Breviary“* z roku 1801. Obrázek 1 ukazuje příklad jeho kreativní kombinace různých vizualizačních technik (kruhy, koláče, linie), pomocí kterých se snažil porovnat daňovou zátěž mezi Británii a dalšími zeměmi. Na tomto grafu také ukázal možnost použití více měřítek pro různé ukazatele (graficky vyjádřena populace a výše daní).

Obrázek 1: Kombinace různých vuzuálních techník, Playfair 1801

[1] *„Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century“*

V polovině 19. století byly vytvořeny všechny podmínky pro rychlý růst vizualizace. V důsledku rostoucí významnosti číselných informací pro sociální plánovaní, industrializaci, obchod a dopravu byly zřízeny oficiální statistické úřady po celé Evropě. Rozvoj statistické teorie, iniciovaný Johannem Carlem Friedrichem Gaussem a Pierrem Simonem de Laplacem, měl odezvu ve společností a poskytl prostředky ke zpracování velkého množství dat. Pro vizualizaci dat se stalo období 1850-1900 „zlatým věkem“, s velkým množstvím inovací. S těmito inovacemi je hlavně spojeno jméno Charlese Josepha Minarda [1781-1870]. Minardem bylo například zavedeno použití koláčových grafů s výsečemi na mapách (obrázek 2), kde velikost koláčového grafu ukazuje sumu sledované proměnné pro každou oblast na mapě a výseče reprezentují dílčí součty za jednotlivé kategorie. Dále se také zabýval znázorněním geografických pohybů úměrně jejich velikostí jako je přeprava lidí, zboží, import či export. Tento typ vizualizace se nazývá *„flow maps“*, viz obrázek 25. Jednou z jeho nejslavnějších prací je zobrazení postupných ztrát francouzské armády během Napoleonského tažení na Moskvu v letech 1812-1813 (obrázek 26). Je považována za nejlepší informativní vizualizaci vůbec. I přestože v tomto grafu je celkem 6 proměnných (množství, lokace ve dvou rozměrech, postup armády, teplota, datum a skupiny), podařilo vše zobrazit tak, aniž by graf byl přeplněný a matoucí.

Obrázek 2: Dobytek odeslaný z celé Francie ke spotřebě v Paříži, Minard 1858

Obrázek 3: Mapa světové migrace, Minard 1858

Začátek 20. století je občas nazýván „moderním temným věkem“ vizualizace. V letech 1900-1950 bylo jen málo inovací. Nadšení pro vizualizací, které charakterizovalo 19. století bylo nahrazeno formálními (z velké části statistickými) grafy a modely z oblasti sociologie. Hlavní zájem byl o přesná čísla, odhady parametrů a směrodatné odchylky. Vizualizace byla považována za pouhé hezké obrázky bez schopnosti podat přesná data. [1]

Obrázek 4: Postup Napoleonských vojsk v letech 1812-13, Minard 1869

Ve své knize *„Graphic Methods for Presenting Facts“* z roku 1919 Willord C. Brinton [1880-1957] kritizoval a vysvětloval chyby takovýchto grafů. Například koláčový graf rozdělení rodinných příjmů (od 900$ do 1000$) na obrázku 27. Tento graf je příkladem nepovedené vizualizace: oko preferenčně soudí dle velikostí obrázků a ne dle uhlů výsečí. Obrázek uprostřed znázorňuje druhy utracení: je to zábavný způsob vizualizace, avšak nelze přesně určit velikost brašen, ani je porovnat mezi sebou. Další obrázek by měl čtenáři sdělit informaci, že prodej praček za poslední tří roky vzrostl sedmkrát. Z obrázku není patrný poměr sedmi ku jedné ani přesné roky kdy bylo provedeno porovnání údajů. Dále Brinton ve své práci upozorňoval, že neúspěšná prezentace dat může vést k chybným závěrům a také zmiňoval potřebu jakéhosi standardu, souhrnu „gramatických pravidel pro grafický jazyk“. [2]

Obrázek 5: Ukázky vizualizací ze začátku 20. století, Brinton 1919

Ke „znovuzrození“ vizualizace došlo v polovině šedesátých let 20. století po zveřejnění článku Johna W. Tukeyho [1915-2000] *„The Future of Data Analysis“*, ve kterém vyzývá společnost k uznání analýzy dat jako samostatného oboru statistiky odlišného od statistiky matematické. [3] Brzy poté začal Tukey s vývojem široké řady nových a efektivních grafů pod společným názvem „průzkumová analýza dat“ (popsáno v jeho knize *„Explanatory Data Analysis“* z roku 1977, viz o tématu kapitola 3). [4] Mezi těmito novými grafy je například číslicový histogram (popsaný v kapitole 2.4.3), boxplot nebo krabicový graf (popsaný v kapitole 2.3.2) a další. Mnoho z nich je aktivně používáno ve statistické praxi a implementováno do většiny softwarů. [1]

1.2 Zásady vizualizace dat

Od roku 1975 se vyvíjí statistické výpočetní systémy a s nimi i nové metody analýzy a vizualizace dat. V tomto období vizualizace začala být vnímána jako vlastní odvětví a to především díky Williamu S. Clevelandu a Edwardu Tuftemu, kteří položili vědecké základy tohoto odvětví. Tufte vyvinul a popularizoval terminologii a základní principy grafické integrity. Cleveland se zabýval studií grafického vnímání, kognitivních procesů, které lidé používají k pochopení grafů, a rozvíjel teorii o správném provedení vizualizace. [5] Výsledek jejich práce se promítá i do současné doby kvalitní, interaktivní a dynamickou vizualizací. [1]

1.2.1 Edward Tufte

Za revoluční průlom se považuje kniha Edwarda Tufte *The Visual Display of Quantitative Information* z roku 1983, v kombinaci se dvěma posléze publikovanými knihami *Envisioning Information* a *Visual Explanations* z roku 1990, resp. 1997, patří mezi nejznámější publikace na téma vizualizace dat. Právě v těchto publikacích Tufte originálním způsobem definuje „standard“ vizualizace. [6] Ideální vizualizace dle Tufteho je stručná, elegantní a informativní. Příkladem ideálního grafu je pro Tufteho graf postupu Napoleonských vojsk v letech 1812-13, vytvořený Minardem (viz obrázek 26). Tufte říká, že grafická elegance se často nachází v jednoduchosti návrhu a komplexnosti dat. [7] Tafte formuluje základní principy vizualizace jako grafickou dokonalost a grafickou integritu.

- **Grafická dokonalost** - grafika by měla:
  - být o datech a během jejich reprezentace by nemělo dojít ke zkreslení
  - vyvolávat otázky o datech, ne o metodologii a technikách vizualizace
  - ukazovat velké množství dat na malém prostoru
  - podávat velké datasety souvisle a logicky
  - sloužit rozumnému a jasnému cíli (popisu, průzkumu, ...)
  - být jednotná se statistickým nebo slovním popisem datasetu

- Pravidla pro **Grafickou integritu** neboli grafickou celistvost a jednoznačnost:
  - reprezentace čísel zobrazených v grafu by měla být přímo úměrná číselným veličinám datasetu
  - jasná, detailní a svědomitá označení v grafech pro potlačení zkreslení, nejasností a dvojznačností
  - popisky jsou důležité
  - ukazovat variaci dat, nikoliv designu
  - v případě časových řad, představujících peníze, používat obecně známé jednotky
  - počet rozměrů představených v grafu by neměl přesahovat počet proměnných datasetu
  - reprezentace by neměla zahrnovat neúmyslný kontext

Obrázek 6: Snižující se procento rodinných lékařů, Los Angeles Times, 1979

Obrázek 7: Vztah skutečné míry volební registrace k předpovídaným hodnotám, převzato E. Tuftem, 1983

Ve spojení s těmito principy byly zavedeny Edwardem Tuftem následující termíny:

- **Lie factor** je definován jako poměr velikosti efektu zobrazeného v grafu oproti velikosti efektu v datech. Pokud se rovná jedné, považují se reprezentované hodnoty za přesné. Pokud je faktor větší než 1.05 či menší než 0.95, indikuje podstatné zkreslení, přesahující míru drobných nepřesností vyskytujících se při vykreslování grafu. Tafte ve své práci uvádí jako jeden z příkladů graf na

obrázku 6. Tento graf zobrazující snižující se procento lékařů věnujících se výhradně rodinné praxi má *lie factor* odpovídající hodnotě 2.8, tedy skutečný pokles je značně nadhodnocen.

- **Data ink ratio** - poměr, který vyhodnocuje hustotu grafu a obsah informací. Dal by se vyjádřit vzorcem

$$Data\ ink\ ratio = \frac{data\text{-}ink}{celkový\ inkoust\ použitý\ v\ datech},$$

  kde *data-ink* je nezbytné jádro grafu a smazání jakékoliv jeho části znamená ztrátu informací. Tento vztah také odpovídá podílu grafického inkoustu požitého k vykreslení nepodstatných informací. Dalo by se to také vyjádřit jako jedna mínus *podíl grafiky, která může být vymazána bez ztráty informací.* Tafte doporučuje tento faktor maximalizovat v rozumných mezích, nejlépe se vyhnout těžkým mřížkovým liniím na pozadí (dokonce i horizontálním referenčním liniím). V příkladu na obrázku 7 jsou zobrazeny dvě verze stejného grafu. Horní má hodnotu *data ink ratio* kolem 0.7, dolní graf, však protože neobsahuje informaci o datech, pouze nápomocné čáry, má *data ink ratio* roven nule.

- **Chartjunk** - se vztahuje ke všem vizuálním elementům, které neslouží ke komunikaci informací zobrazených v grafu nebo odvádějí pozornost od těchto informací. [8]

1.2.2 Wiliam S. Cleveland

Kromě práce Edwarda Tufte velkýho měly vliv i publikace Wiliama S. Clevelanda. Cleveland se svým kolegou Robertem McGillem publikovali v roce 1984 článek o grafickém vnímání [9]. Prováděli studie zabývající se rozdílem ve vnímání sloupcových grafů (pozice a obecné měřítko), koláčových grafů (úhel), skládaných sloupcových grafů (plocha), barevných a stínovaných map (saturace barev a stínování) a dalších. [5] Ve svých knihách *Visualizing data* z roku 1993 a *The Elements of Graphing Data* z roku 1994 se Cleveland zabýval principy vizualizace, grafickými metodami a technikami či vykreslením tří a více proměnných (rozměrů). Některé z jeho principů se shodují s principy vymezenými Tuftem, avšak výzkum Clevelanda v této oblasti přesahovala práci Tufteho. Zásady a principy dle Clevelanda by se daly shrnout do čtyř hlavních kategorií: jasná vize, srozumitelnost, měřítka a obecné postupy. [10]

- **Jasná vize**
  - Data by měla vyčnívat, bez vykreslení nadbytečných prvků (neboli *chartjunk* dle Tufteho)
  - K zobrazení dat by se měly používat výrazné grafické prvky.
  - Pro každou proměnnou by měla být použita dvojice os, prostor v takto vytvořeném obdélníku je určen k vykreslení grafu, značky na osách by měly směrovat mimo oblast grafu.
  - Prostor grafu by neměl být přeplněný (legenda mimo oblast grafu atd.).
  - S počtem značek na osách by se nemělo přehánět.

- Pokud je to vhodné, referenční linie mohou být použity, avšak nesmějí zasahovat do dat.
- Popisky by neměly zasahovat do kvantitativních dat a nesmějí znepřehledňovat graf.
- Značky a klíče by měly vyskytovat mimo oblast grafu (případně v legendě), totéž se týká poznámek a nadpisů, které mohou být také umístěny do textu.
- Překrývající se data sety či symboly musí být vizuálně snadně rozpoznatelné.
- Jasnost obrazu musí být zachována při reprodukci i při snížení kvality a zmenšení rozlišení.

Cleveland jako příklad špatně zpracované vizualizace vybral graf 8a, na kterém je zobrazeno množství izotopu xenonu ^{133}Xe ve vzduchu ($pCi.m^{-3}$) po havárii elektrárny Three Mile Island v Albany, ve státě New York koncem března a začátkem dubna roku 1979. Vše, včetně popisků os, klíčů a popisků dat bylo umístěno do oblasti grafu, není dodržena žádná ze zásad Clevelanda. Výsledkem je matoucí graf, který je obtížně čitelný. Stejná data byla vizualizována Clevelandem na grafu 8b s dodržením veškerých zásad: odstranění zbytečných objektů a detailů z oblasti grafu, datasety se zobrazují ve vlastních panelech, oprava popisků popisujících měření.

(a) (b)

Obrázek 8: Radioaktivní oblak při havárii elektrárny Three Mile Island: ^{133}Xe ve vzduchu ve vzdálenosti 375 km (a) a stejný graf přepracovaný Clevelandem (b), 1994

- **Jasná srozumitelnost**
  - Hlavní závěry by měly být obsaženy v grafické formě. Legenda a nadpisy by měly být srozumitelné a vyčerpávající.
  - Grafy by měli být zkontrolovány.
  - Mělo by se usilovat o přehlednost (viz „jasná víze“).

- **Měřítka**
  - Volit rozsah os tak, aby obsahoval, případně téměř obsahoval, rozsah dat.
  - Volit takové měřítko, aby data vyplňovala co největší prostor.
  - Občas je užitečné mít pro proměnnou dvě osy pro rozdílná měřítka.
  - Volit vhodné měřítko pokud data jsou porovnávány na více panelech.
  - Osy grafu nemusejí vždy nutně zahrnovat nulu pro ukázku rozsahu.
  - Použít logaritmická měřítka, když je důležité pochopit procentní změny nebo multiplikativní faktory.
  - Použít přerušené měřítko pouze v případě potřeby. Alternativou může být logaritmizace měřítka.
- **Obecné postupy**
  - Velké množství kvantitativní informace může být vměstnáno do relativně malých oblastí.
  - Tvorba grafů by měla být opakující se, iterativní a experimentální činností.
  - Data by měla být vykreslena tolikrát, kolikrát je třeba.
  - Užitečné grafy vyžadují pečlivou a detailní práci.

Cleveland se mimo jiné podílel na tvorbě řady technik pro prohlížení komplexních datových sad s více proměnnými, kterým se říká *Trellis Graphics* nebo *Trellis Plots*. Technika obdržela svůj název *trellis* kvůli obvyklému výsledku řady obdélníkových grafů, připomínajících zahradní mříž. Na obrázku 9 je ukázka *trellis* grafu, zobrazující údaje o emisích motoru. [11]

Obrázek 9: *Trellis* graf, zobrazující údaje o emisích motoru, Becker et al., 1996

1.3 Grammar of graphics

The Grammar of Graphics publikována Lelandem Wilkinsonem v roce 2005 [12] detailně popisuje prvky, které tvoří základ všech statistických grafů a odpovídá na základní otázku: co je statistická grafika? [13] Tato publikace měla extrémně velký vliv na myšlení o grafech. Hadley Wickham na základě Wilkinsonovy gramatiky publikoval v roce 2009 článek *A Layered Grammar of Graphics* [14], který se zaměřuje primárně na vrstvy a jejich zapojení do jazyka R. Následně také pro něj posloužila jako inspirace pro tvorbu balíčku `ggplot` (viz kapitola 4.2.1).

The Grammar of Graphics říká, že statistická grafika je mapováním dat k estetickým atributům (barva, tvar, velikost) geometrických objektů (body, linie, sloupce). Graf také může obsahovat statistickou transformaci dat a být vykreslen ve specifickém souřadnicovém systému. *Faceting* může být použito k vygenerování stejného grafu pro různé podmnožiny datasetu. Kombinace těchto nezávislých komponent tvoří grafiku. Jednotlivé komponenty tvořící graf, dle Wilkinsonovy syntaxe, lze zapsat následovně:

- Vizualizovaná **data** a soubor estetických mapování (**mapping**s) popisujících jak jsou proměnné z dat mapovány na vnímané estetické atributy.
- Geometrické objekty (**geom**s) reprezentují to, co je doopravdy na grafu: body, linie, polygony atd.
- Statistické transformace (**stat**s) sumarizují data mnoha užitečnými způsoby. Jako příklad by se daly použít výpočty intervalů a počty pozorování při tvorbě histogramu (kapitola 2.4.1) nebo tvorba lineárního modelu. Statistické transformace patří k nepovinným, ale velmi užitečným komponentům.
- Měřítka (**scale**s) reprezentují hodnoty v datovém prostoru převedené na hodnoty v estetickém prostoru, ať už se jedná o barvu, velikost či tvar. Na měřítku závisí legenda a osy, tvořené inverzním mapováním umožňující číst z grafu původní hodnoty datasetu.
- Souřadnicový systém (**coord**) popisuje, jak jsou souřadnice dat mapovány do roviny grafiky. Rovněž poskytuje osy a mřížky, umožňující čtení grafů. Běžně se používá kartézský souřadnicový systém, ale je k dispozici i řada dalších systémů včetně polárních souřadnic a mapových projekcí.
- Specifikace **facet**ing popisuje, jaké proměnné by měli být použity k rozdělení dat na podmnožiny a jak tyto podmnožiny by měly být uspořádány. Jedná se o mocný nástroj pro zkoumání toho, zda jsou modely stejné nebo odlišné v různých podmínkách.

Je také důležité zmínit, o čem Wilkinsova gramatika není. Nenaznačuje jaký typ grafů by se měl použít k zodpovězení otázek o datech, jak to dělali Cleveland [15] nebo Tukey [4], zaměřuje se konkrétně na jejich tvorbu. Ironií je, že *The Grammar of Graphics* neurčuje, jak by měla vypadat grafika, nespecifikuje velikost písma ani barvu pozadí. [13] Otázkou vzhledu grafů se zabývali Tufte a Cleveland (kapitoly 1.2.1 a 1.2.2). Dále Wilkinsova gramatika nepopisuje interaktivní ani dynamické vizualizace, obsahuje pouze statické grafy. Při tvorbě dynamických či interaktivních grafů je třeba se obrátit na jiný zdroj, například *Interactive Data Visualization for the Web* od Scottea Murrayho [16] nebo *Interactive Visualization* od Billa Ferstera [17].

2 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` [18], který obsahuje mnoho užitečných funkcí pro tvorbu grafických prvků. První kapitola se soustředí na tyto funkce tohoto balíčku a v dalších kapitolách jsou popsány funkce balíčků dalších (například `lattice`, `ggplot2`,...), které zastávají podobné funkce, avšak s různým rozsahem nastavení [19].

V následujících příkladech nejsou grafy doplněny o barvy, popisky os, legendy ani názvy a to především proto, že záměrem této kapitoly je popsat základní grafy a funkce pro jejich tvorbu v prostředí R. Všechny tyto prvky mohou být přidány do grafu, ale tím by příkazy obsahovali irelevantní parametry vzhledem k zaměření této kapitoly. Základní funkce `plot(x)` jejímž voláním se obdrží pole s grafickou reprezentaci proměnné „x“, by při doplnění kódu o veškeré parametry vypadala následovaně [19]:

```
plot(x, main = "Název grafu", xlab = "popis osy x",
+    ylab = "popis osy y", col = c("red", "black", "green"))
```

Záměrem je tedy používání příkazů s pouze relevantními parametry.

2.1 Bodový graf

Bodový graf je rychlým způsobem, jak znázornit vztahy a souvislosti mezi proměnnými datasetu, případně k zjištění jejich neexistence. Data jsou zobrazeny v kartézském souřadném systému a mají pro každou hodnotu proměnné dané místo na vodorovné a svislé ose. V případě existence závislostí mezi proměnnými lze tuto závislost interpolovat přímkou, křivkou či dalším vhodným vyobrazením této závislosti.

Pro vytvoření bodového grafu v základním prostředí R (pomocí `graphics`) použijeme funkci `plot()`, která má tento typ grafu předdefinovaný pro numerické hodnoty. Viz obrázek 28 (a). Nečíselná data vytvoří jiný typ grafu.

```
plot(cars)
```

2.2 Liniový graf

Jediný rozdíl mezi bodovým a liniovým grafem je, že jeden zobrazuje body a druhý je spojuje.[19] (viz. obrázek 28 (a), (b)). Pro vykreslení liniového grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

```
plot(x, type="l")
```

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity [20]:

	Anglický popis	Český popis
p	points	bodový
l	lines	liniový
b	both	složený
h	histogram	histogram
n	no plotting	bez vykreslení

Tabulka 1: Základní atributy parametru ‘type‘

Obrázek 10: Porovnání základních typů grafů: (a) - bodový, (b) - liniový, (c) - složený, (d) - histogram

Popis a všechny atributy dalších parametrů funkce `plot()` lze nalézt v nápovědě zadáním příkazu `?plot()`.

2.3 Vykreslení rozdělení v R

Teorie pravděpodobnosti je základem statistiky a R má hodně nástrojů pro práci s pravděpodobnosti, rozdělením pravděpodobnosti a náhodnými proměnnými. R má zkrácený název pro každé rozdělení pravděpodobnosti. [19] Tyto názvy slouží

k identifikaci funkcí spojených s rozděleními. Například zkrácený název „norm“ pro normální rozdělení, „exp“ pro exponenciální rozdělení a další. Funkce pak mají formu:

Funkce	Účel
dxxxx	Hustota pravděpodobnosti
pxxxx	Distribuční funkce
qxxxx	Kvantilová funkce
rxxxx	Generátor náhodných čísel z daného rozdělení

Tabulka 2: Funkce pro práci s rozděleními

Funkce v R lze vykreslovat pomocí funkce `curve()` z balíčku `graphics`. Lze vykreslit jak standardní funkce, tak i funkce definované uživatelem. Například hustotu pravděpodobnosti normálního rozdělení a její distribuční funkci můžeme vykreslit tímto způsobem (Obrázek 29):

```
curve(dnorm(x))
curve(pnorm(x))
```

Obrázek 11: Hustota pravděpodobnosti normálního rozdělení (a) a její distribuční funkce (b)

2.3.1 Q-Q graf a P-P graf

Q-Q (*quantile-quantile*) graf a P-P (*probability–probability* nebo *percent–percent*) graf (Obrázek 30) se používají hlavně k testování normality při průzkumové analýze dat 3.4. Další způsob, jak zjistit zda-li data mají normální rozdělení je sestrojení histogramu (viz. sekce 1.4.1), avšak použití Q-Q grafu je přesnější.

Princip Q-Q grafu spočívá v porovnání dvou rozdělení pravděpodobnosti pomocí vykreslení jejich kvantilů proti sobě. Na jedné ose se nacházejí teoretické kvantily normálního rozdělení a na druhé ose kvantily naměřené (pozorované). Pokud data mají přesně normální rozdělení, všechny body grafu leží na přímce 45°. Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 21. [19] [10]

Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkcí proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se o normální rozdělení. Z velké části se P-P graf používá k vyhodnocení koeficientu šikmosti rozdělení.[21]

V R se Q-Q graf vykreslí takto:

```
qqnorm(x)
qqline(x)
```

P-P graf v R lze vykreslit například následovně:

```
plot(ppoints(length(x)), sort(pnorm(x)))
abline(0,1)
```

Obrázek 12: Q-Q Graf (a) a P-P Graf (b)

2.3.2 Krabicový graf

Krabicový graf poskytuje rychlé a jednoduché vizuální shrnutí datasetu. V základním prostředí R se vykreslí pomocí funkce `boxplot()` z balíčku `graphics`. Obrázek 31 znázorňuje typický krabicový graf, kde silná čára je medián, krabice kolem ní určuje polohu prvního a třetího kvartilů (dolní Q_1 kvantil 25% a horní Q_3 kvantil 75%). "Vousy" (*whiskers*) nad a pod krabici znázorňují rozpětí dat bez odlehlých hodnot. Odlehlé hodnoty jsou definovaný jako hodnoty ležící ve větší vzdálenosti od krabice než 1,5 × IQR, kde IQR je mezikvartilové rozpětí (*interquartile range*) neboli $Q_3 - Q_1$.

```
boxplot(x)
```

Obrázek 13: Boxplot

2.4 Sloupcový graf

Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.[22]

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 32) je použit data set `mtcars`, konkretně atribut `cyl` - počet válců v motoru.

```
table(mtcars$cyl)
```

```
##
##  4  6  8
## 11  7 14
```

```
barplot(table(mtcars$cyl))
```

Obrázek 14: Ukázka jednoduchého sloupcového grafu

2.4.1 Histogram

Sloupcový graf s četnostmi na souvislé ose je taky známý jako histogram.[22] Četnosti mohou být absolutní či relativní. Absolutní četnost zobrazuje počet statistických jednotek s hodnotou znaku, který patří do určitého intervalu. Podíl příslušné četnosti a rozsahu datového souboru se nazývá relativní četnost.[23] Šířka sloupce reprezentuje jednotlivé intervaly, které mají stejnou délku. Pro výpočet optimální délky intervalu existují různé metody. Základní histogram se vytváří pomocí funkci `hist()` a její atribut `breaks` udává buď hranice intervalů, jejich preferovaný počet nebo metodu výpočtu intervalu. V R jsou vestavěny 3 metody výpočtu:

1. Sturges [24]

```
hist(x, breaks = "Sturges")
```

$$k = [log_2(n)] + 1$$

Kde k je počet intervalů a n je počet prvků neboli počet pozorování výběru x. Tato metoda je výchozí pro funkci `hist()`.

2. Scott [24]

```
hist(x, breaks = "Scott")
```

Scotovo pravidlo je následující:

$$h = \frac{3.5\sigma}{n^{\frac{1}{3}}}$$

kde σ je směrodatná odchylka a h je předpokládaná šířka intervalu.

Počet intervalů může být vypočítán pomocí vztahu:

$$k = \left[\frac{max(x) - min(x)}{h}\right]$$

Případně oba vztahy lze shrnout do jednoho:

$$k = \left[n^{\frac{1}{3}} \frac{max(x) - min(x)}{3.5\sigma}\right]$$

3. Freedman–Diaconis [25]

```
hist(x, breaks = "FD")
```

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

$$h = 2\frac{IQR(x)}{n^{\frac{1}{3}}}$$

Po dosazení:

$$k = \left[n^{\frac{1}{3}} \frac{max(x) - min(x)}{2IQR(x)}\right]$$

kde IQR je mezikvartilové rozpětí, které definujeme jako rozdíl třetího a prvního kvartilů.

Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 15).

Obrázek 15: Histogram s odhadem hustoty pravděpodobnosti

2.4.2 Koláčový graf

Koláčový graf představuje plný kruh (360°), který je rozdělen na jednotlivé výseče pro znázornění číselných proporci mezi proměnnými. Koláčový graf je tvořen transformaci skládaného sloupcového grafu do polárního souřadnicového systému (Obrázek 16). [12]

Obrázek 16: Skládaný sloupcový graf transformovaný do polárního souřadnicového systému

Jednoduché koláčové grafy se vykreslují pomoci funkci `pie()` (Obrázek 17).

Obrázek 17: Ukázka jednoduchého koláčového grafu

2.4.3 Číslicový histogram (*stem-and-leaf*)

Číslicový histogram, jinak známy jako *stem-and-leaf plot*, podobně jako histogram pomáhá vizualizovat tvar rozdělení. Jedná se spíše o historický typ grafu, který byl populární v osmdesátých letech, kvůli obtížnějšímu vykreslování velkých datasetu. Vstupní údaje jsou rozdělené vertikální linií na dva sloupce. Pravý sloupec obsahuje listy (*leaf*) - poslední číslice po desetinné čárce a levý sloupec obsahuje stonek (*stem*) - číslice před desetinnou čárkou. Každý stonek je uveden pouze jednou i pokud neobsahuje žádné listy. Listy se uvádějí od nejmenšího po největší. [4] Proto v příkladu uvedeném níže je v prvním řádku stonkem číslice -2 a listy jsou číslice 9 a 2. Víme tak, že v datasetu se vyskytli čísla -2.9 a -2.2. Tento typ grafu v prostředí R se vykresluje pomoci funkce `stem()`:

```r
stem(x)
```

```
##
##   The decimal point is at the |
##
##   -2 | 92
##   -1 | 888755333332211100
##   -0 | 99888877666666655554433332111100
##    0 | 00111222222333344444567777888888999
##    1 | 0233445689
##    2 | 0012
```

3 Průzkumová analýza dat

Obrázek 18: Posloupnost datové analýzy

Úkolem průzkumové analýzy dat (*Explanatory Data Analysis*, zkráceně EDA) je vizualizace a transformace dat systematickým způsobem za účelem maximálního pochopení dat, určení vztahu mezi nimi a posouzení jejích kvality. EDA je důležitou části datové analýzy a měla by být jedním z jejích prvních kroků.

Zařazení průzkumové analýzy dat do procesu datové analýzy je zobrazeno v diagramu 18. Prvním krokem datové analýzy je **import** dat. Obecně to v tomto případě znamená nahrání obdržených dat ze souboru či databáze do prostředí R. Bez tohoto kroku datová analýza nemůže být vykonána. V momentě když data jsou importována do R je vhodné je **očistit** neboli **připravit**. Tím je myšleno ukládání dat v konzistentní a systematické formě, odpovídající sémantice původního datasetu. Zkrátka očištěná data jsou taková data, ve kterých sloupce odpovídají proměnným a řádky odpovídají pozorováním. Takováto příprava dat usnadňuje další práci s nimi.

Jakmile jsou data očištěna, je obvyklým krokem jejich **transformace**. Transformací se rozumí omezení pozorování (například dle zájmového území či povodí), vytváření nových proměnných na základě již existujících, agregace (např. z denního do měsíčního kroku), výpočet souhrnných statistik (středních hodnot, kvantilů atd.), odstranění odlehlých pozorování a normalizace. Poté, co jsou data očištěná a obsahují veškeré potřebné proměnné, je možné na ně aplikovat dva nejdůležitější nástroje k zjištění informací: vizualizaci a modelování. Jakákoliv analýza tyto nástroje opakovaně využívá, ačkoliv samozřejmě mají své výhody a nevýhody.

Vizualizace je schopná odhalit neočekávané chování dat a napovědět další směr analýzy. Vizualizaci lze odhalit nevhodně zvolená či špatně připravená data a nekorektní dotazování. I přesto, že vizualizace je dobrým nástrojem datové analýzy, její aplikace na větší datasety je značně náročná a interpretace výsledků je subjektivní, tudíž závisí na analytikovi.

Modelování je v ramci průzkumové analýzy dat doplňkem vizualizace. Jedná se o zásadně matematický a výpočetní nástroj, který se obecně hodí i na větší datasety. Téměř každý model musí splňovat své předpoklady, které by měli být ověřeny před jejich aplikací, na rozdíl od vizualizace, která žádné předpoklady nevyžaduje.[26]

Důležitou součástí analýzy je **interpretace** výsledků a formulace závěrů. Vyhodnocuje, jak dobře zvolený model či vizualizace slouží k pochopení dat a jejich popisu. Je také důležité si uvědomit komu jsou výsledky interpretovány, kdo je cílová skupina. Dobře provedené grafické výstupy podložené jejich správnou interpretaci jsou jedním z nejlepších způsobů prezentace dat.

Průzkumová analýza dat není specifikována jako konkrétní soubor pravidel a postupů, ale jako přístup k analýze dat. Obvykle zahrnuje následující kroky:

- Vyhledávání vybočujících (odlehlých) pozorování
- Náhrada chybějících hodnot
- Transformace dat
- Změny typu proměnných
- Ověřování normality

3.1 Odlehlá pozorování

Odlehlá pozorování (*outliers*) jsou významně odlišná vůči ostatním hodnotám datasetu. Definice toho, jak moc odlišná taková pozorování mají být je dáno analytikem na základě konkretního datasetu a kontextu problematiky. Tato pozorování mohou být indikátorem chybných dat nebo vzácných událostí. Důvody proč se tato pozorování vyskytují by měli být pečlivě zkoumány. Dále je důležité posoudit, jak je jimi výsledek analýzy ovlivněn, případně zdali je předpoklady metody připouštějí.

Hledání odlehlých, vybočujících, pozorování a jiných anomálií pro jednotlivé veličiny lze provést graficky například pomoci boxplotu (viz. sekce 2.3.2), bodových grafů (2.1) nebo číslicových histogramů (2.4.3). Dají se také vypočítat pomocí různých statistik, například metodou *jackknife*, která je popsána v následující kapitole (3.1.1). V momentech, kdy je vizualizace obtížná (velké datasety, větší množství navzájem se ovlivňujících proměnných, atd.), využívají se nástroje vícerozměrné, například Mahalanobisovy vzdálenosti (3.1.2), *leverages* (3.1.3) a další.

3.1.1 *Jackknife*

Metoda byla původně představená Johnem W. Tukey v roce 1958 v „*The Annals of Mathematical Statistic*“ [27] a jedná se o speciální případ metody *bootstrap* (více o metodě B. Efron a R. Tibshirani v „*An Introduction to the Bootstrap*“ [28]).

Postup metody *jackknife* je založen na celkem jednoduché myšlence. Zjišťují se souhrnné statistiky podsouborů (*Jackknife Samples*), které se vytvářejí postupným vypouštěním jednotlivých pozorování z původního datasetu. Jinými slovy existuje n unikátních Jackknife podsouborů a i-tý Jackknife podsoubor je definován jako vektor.

Pomocí porovnání souhrnných statistik původního datasetu a vytvořených Jackknife podsouborů se odhadne vliv jednotlivých pozorování na původní dataset. Jedná ze souhrnných statistik, kterou lze použít je střední hodnota $\bar{x}$. Pro původní dataset

obsahující n pozorování lze střední hodnotu odhadnout dle vzorce $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_i$. Střední hodnota Jackknife podsouborů se vyhodnotí následovně:

$$\bar{x}_i = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} x_j, \quad \text{kde } i = 1, \ldots, n.$$

Porovnání lze provést dle vzorce $Var(\bar{x}) = \frac{n-1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2$, kde $Var(\bar{x})$ je odhad rozptylu, který indikuje, jak moc jednotlivá pozorování ovlivňují dataset, tj. přítomnost odlehlých pozorování. Metoda může být také použita k odhadu skutečné, neovlivněné střední hodnoty datasetu. [29]

3.1.2 Mahalanobisovy vzdálenosti

K měření vzdálenosti mezi objekty se často používá euklidovská vzdálenost. Euklidovská vzdálenost je jednoduchá na výpočet a interpretaci, ale není schopná brát v úvahu vztahy mezi daty. Proto je v řádě případů vhodné použít mahalanobisovou vzdálenost. Je definovaná matice $\boldsymbol{X}(n \times p)$, obsahující n objektů $\boldsymbol{x}_i$ a p proměnných. Euklidovská vzdálenost mezi vektorem i-tého řádku $\boldsymbol{x}_i(1 \times p)$ této matice a vektoru středních hodnot $\bar{\boldsymbol{x}}(1 \times p)$ se spočítá jako

$$ED_i = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T}, \quad \text{pro } i = 1, \ldots, n$$

zatímco mahalanobisova vzdálenost se spočítá jako

$$MD_i = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{C}_x^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T}, \quad \text{pro } i = 1, \ldots, n$$

kde $\boldsymbol{C}_x$ je kovarianční matice. [30]

Na obrázku 19 jsou znázorněny elipsy mahalanobisových vzdáleností, kde každá elipsa představuje vzdálenost od průměru. Z tohoto je zřejmé, že vzdálenost roste pomaleji ve směru korelace. Pozorování, které je výrazně vzdáleno od středu, ale leží ve směru závislosti, má nižší mahalanobisovou vzdálenost než pozorování, které je stejně vzdáleno od středu, ale neleží ve směru závislosti. Tato vlastnost mahalanobisových vzdálenosti umožňuje identifikaci odlehlých pozorování.

Metoda byla představena P.C. Mahalanobisem v roce 1936 ve článku *„On the Generalized Distance in Statistics“* [31]. Mahalanobisové vzdálenosti se používají nejenom k nalezení odlehlých pozorování, ale i ke zkoumání reprezentativity mezi dvěma data sety, aplikuje se v algoritmu k-nejbližších sousedů, v diskriminační analýze a má mnoho dalších uplatnění.

Obrázek 19: Mahalanobisovy vzdálenosti

3.1.3 Leverages

Leverage (případně též efekt, vliv nebo projekční h prvek) se používá v regresní analýze k měření velikosti vlivu pozorování na regresní odhad. Princip metody spočívá v kontrole diagonálních prvků projekční matice $\boldsymbol{H}$, která je produktem metody nejmenších čtverců a je definována

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

Model lineární regrese může být zapsán následovně:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde vektor vysvětlované proměnné je $\boldsymbol{y}$, matice vysvětlujících proměnných je $\boldsymbol{X}$, vektor regresních koeficientů, který je odhadován, je $\boldsymbol{\beta}$ a vektor náhodné složky je $\boldsymbol{\varepsilon}$. Metoda nejmenších čtverců poskytuje řešení regresních rovnic:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Lze dosadit:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Výsledný vektor má tvar $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, kde $\boldsymbol{H}$ je projekční matice. [32]

3.2 Náhrada chybějících pozorování

Problém chybějících pozorování spočívá v neschopnosti jejich zpracovávání některými metodami. Takové hodnoty lze vynechat nebo doplnit (nahradit) jednou z řady metod. Vynechání hodnot vede k nežádoucímu zmenšení datasetu, proto je výhodnější chybějící údaje doplnit. Nejednodušším nástrojem pro náhradu chybějících hodnot je aritmetický průměr příslušné proměnné. Tento způsob může vést ke zkresleným odhadům (neplatí-li předpoklad, že chybějící údaje jsou zcela náhodné) a podhodnocuje variabilitu a kovarianci datasetu, a proto se nedoporučuje v případě vyššího podílu chybějících údajů. Další možnou metodou je náhrada náhodným číslem generovaným z příslušného rozdělení (parametry jsou odhaduty z výběru). V tomto případě se respektuje variabilita datasetu, ale nerespektuje se jeho kovariance. Chybějící údaje lze také odvodit pomocí známých hodnot na základě pomocné jednoduché lineární regresní funkce. Tato metoda respektuje nejenom variabilitu vzorku, ale i jeho korelační strukturu. [33]

3.3 Transformace dat

Jedním z cílů transformace dat je dosažení srovnatelnosti proměnných: sjednocení měřítka, variace a typu proměnných. Hlavním využitím je splnění podmínek vyžadovaných metodami, například podmínky normality, kde je snaha převést data na normální rozdělení, snížení vlivu rušivých proměnných (odlehlých hodnot) atd. [34] Rozdělujeme transformaci lineární (centrování, normování) a nelineární (plynoucí z typu a charakteru dat).

Lineární transformace zachovává lineární vztahy mezi proměnnými. Jedním z příkladů takovéto úpravy dat je metoda centrování, která se používá u vícerozměrných analýz. Podstata metody spočívá v zachování měřítka vzorku při změně hodnot: od původních hodnot se odečítá průměr proměnné (od prvků sloupce se odečte jejich sloupcový průměr), průměry získaných nových proměnných se tudíž rovnají nule. Toto lze zapsat následovně:

$$v_{ij} = x_{ij} - \bar{x}_j$$

Vektor průměrů $\bar{\boldsymbol{v}}$ je nulový, kovariance a korelace proměnných zůstává nezměněna. [35] Další často využívanou metodou je metoda normalizace dat. Tato metoda transformuje měřítka vzorků pro možnost jejich porovnání (eliminuje jednotky měření), po úpravě střední hodnota vzorku tedy odpovídá nule a odchylka jedničce (normální rozdělení).

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma(x_j)}$$

$\sigma(x_j)$ je směrodatná odchylka sloupce proměnné, vektor průměrů $\bar{\boldsymbol{z}}$ je nulový a kovariance vektoru nových proměnných se shoduje s korelací původního vektoru. [36]

Nelineární transformace vyplývá z typu dat a mění (snižuje či zvyšuje) lineární vztahy mezi proměnnými a to znamená, že nezachovává korelaci mezi nimi. Pokud data

mají charakter absolutní četnosti, používá se odmocninová transformace $X' = \sqrt{X}$, pokud odpovídají log-normálnímu rozdělení, používá se logaritmická transformace $X' = \log_{10} X$ atd. Logaritmus náhodné veličiny s log-normálním rozdělením má normální rozdělení (viz obrázek 20). Logaritmická transformace může být použita pouze u nezáporných rozdělení. [37] [38]

Obrázek 20: Log-normální rozdělení transformováné na normální rozdělení

3.4 Ověřování normality

Důležitým aspektem popisu proměnné je tvar jejího rozdělení, který udává četnosti hodnot z různých rozsahů proměnné. Většina statistických testů a metod se zakládá na předpokladu, že proměnná má normální rozdělení. Z tohoto důvodu je vhodné ověřovat normalitu rozdělení analyzovaného vzorku.

Zjistit zda-li vzorek pochází z normálního rozdělení lze grafickým posouzením nebo pomocí testů normality. Mezi nástroje grafického posouzení normality se řadí histogram rozdělení četnosti (kapitola 2.4.1), graf výběrové distribuční funkce (2.3), Q-Q graf a P-P graf (2.3.1). Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 21. Dále existuje řada testů normality, zde jsou popsany testy Shapiro-Wilk (SW) a Jarqua-Bera (JB).

Obrázek 21: Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality

Shapiro-Wilk test byl poprvé představen v roce 1965 S. S. Shapirem a M. Wilkem [39]. Metoda dokáže pracovat se vzorky velikosti 12 až 5000 pozorování. Nulová hypotéza tohoto testu předpokládá, že vzorek má normální rozdělení. Pokud p-hodnota je menší, než zvolená hladina významnosti, zamítá se nulová hypotéza, jinými slovy vzorek nemá normální rozdělení. Statistika testu vypadá následovně:

$$W = \frac{\left(\sum\limits_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2},$$

kde $x_{(i)}$ je i-tý nejmenší prvek (statistika i-tého řádu), $\bar{x}$ je průměr vzorku, n je počet pozorování.

Jarqua-Bera test závisí na koeficientech šikmosti a špičatosti. Statistika JB testu může být zapsána:

$$T = n\left(\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2 - 3)^2}{24}\right),$$

kde n je velikost vzorku, $\sqrt{b_1}$ je koeficient šikmosti vzorku a b_2 je koeficient špičatosti. Nulová a alternativní hypotéza se schoduje s SW testem. Používá se pro větší datasety nad 2000 pozorování. [40]

4 Pokročilá vizualizace v R

4.1 Balíčky pro vizualizaci dat

Vestavený balíček `base` byl vyvinut Rossem Ihaka na základě zkušeností s implementací grafických ovladačů do S (předchůdce R). Grafy v `base` mají charakter grafů na papíře: stávající obsah nelze modifikovat ani odstranit. Do grafu lze přidávat potřebné prvky, které jsou vykreslovány na povrch grafu a po vykreslení nemohou být dále měněny. V `base` neexistuje jiná uživateli přístupná reprezentace, než ta, co se objeví na obrazovce. `base` obsahuje nástroje pro kreslení jak primitivních, tak i kompletních výkresů. Funkce tohoto balíčku jsou obecně rychlé, ale mají omezené možnosti. Vykreslení základních grafů v R je popsáno v kapitole 2.

Mimo `base` má uživatel možnost využít rozsáhlou nabídku dalších balíčků. Následující kapitoly obsahují kratký popis vybraných balíčku (`grid`, `lattice`, `ggplot2`). V R existuje mnoho dalších balíčků, například `vcd`, `plotrix` a `gplot`, které implementují speciální grafiku, avšak žádný z nich neposkytuje rámec pro tvorbu statistických grafů. Pro instalaci balíčků se používá příkaz `install.packages()`. Komplexní zdroj, úvádějící všechny grafické funkce dostupné v ostatních balíčcích vyvinutých komunitou lze nalézt na stránce http://cran.r-project.org/web/views/Graphics.html. [12]

4.1.1 `grid`

`grid` je alternativou jednoduché grafiky z `base` s rozsáhlejšími možnosti pro úpravu grafu. Vývoj balíčku začal v roce 2000. Autorem je Paul Murrell a balíček vznikl na základě jeho doktorské práce *Investigations in Graphical Statistics* z roku 1998 [41]. Grafické objekty v `grid` mohou být reprezentovány nezávislé na grafu a později upraveny. Systém *viewport*ů (každý obsahuje vlastní souřadnicový systém) usnadňuje rozvržení komlexní grafiky. Balíček umožňuje vykreslení jednotlivých výkresů, avšak nemá nástroje pro tvorbu statistických grafů. [42]

4.1.2 `lattice`

Deepayanem Sarkarem vyvinutý balíček `lattice` používá grafiku `grid` k implementaci Clevelandového *trellis* grafického systému (viz kapitola 1.2.2), což vedlo k značnému vylepšení grafiky oproti `base`. Pomocí `lattice` lze jednoduše vykreslit *trellis* graf a některé detaily výkresu, jako například legenda, se vytváří automaticky. Avšak `lattice` postrádá formální model, což může ztížit jeho rozšíření. Tento balíček je dostačující pro typické grafické potřeby a je dostatečně flexibilní i pro zvládnutí většiny nestandardních požádavků.

4.1.3 ggplot2

Balíček `ggplot2` byl vyvinut v roce 2005 Hadley Wickhamem na základě *The Grammar of Graphics* Lelanda Wilkinsona. `ggplot2` přebírá přednosti balíčků `base` a `lattice` a vylepšuje je silným základním modelem, který podporuje tvorbu libovolného statistického grafu založeného na principech popsaných v kapitole 1.3. Silný základní model `ggplot2` umožňuje popsat širokou škálu grafiky pomocí kompaktní syntaxe a nezávislé komponenty zjednodušují rozšíření. Obdobně jako `lattice`, využívá `ggplot2` mřížky k vykreslení grafiky, což znamená, že umožňuje úpravu vyhledu na mnohem nižší úrovni.

Jako ukázku lze předvést graf vykreslený pomocí různých balíčků (obrázky 22, 23, 24). V představeném kódu byly použity denní časové řady srážek z roku 2010 v oblasti povodí Labe od pramene po Svatopetrský potok včetně. ID tohoto útvaru povrchových vod je `HSL_0010`. `DTM` odpovídá datumu v denním kroce a `P` odpovídá úhrnu srážek v milimetrech. Každý balíček má přednastavený počáteční vzhled, který lze úpravovat dle požádavků, avšak pro tuto ukázku je nutno počáteční rozdíly naopak ponechat.

```
plot(HSL_0010$DTM, HSL_0010$P, type = "l",
                               xlab = "DTM", ylab = "P",
                               main = "`base`")
```

Obrázek 22: Časová řada srážek na zvoleném povodí v roce 2010, `base`

```r
xyplot(P~DTM, data = HSL_0010, type = "l",
                               xlab = "DTM", ylab = "P",
                               main = "`lattice`" ,)
```

Obrázek 23: Časová řada srážek na zvoleném povodí v roce 2010, `lattice`

```r
ggplot(data=HSL_0010, aes(DTM,P)) +
  geom_line() +
  labs(title = "`ggplot2`")
```

Obrázek 24: Časová řada srážek na zvoleném povodí v roce 2010, `ggplot2`

4.2 Balíčky pro prostorovou vizualizaci

V R existuje celá řada možností pro analýzu prostorových dat a nástrojů k jejich vykreslení. Důležitý rozdíl mezi R a tradičními desktopovými GIS (*geographical information system*) softwary je ten, že GIS je primárně vytvořen k prostorové vizualizaci a zpracovává data jedním přednastaveným způsobem, zatímco v R uživatel si sám musí zvolit vhodné nástroje a odpovídající nastavení. Dále R neobsahuje GUI, které by usnadňoválo práci s prostorovými daty. Z tohoto důvodu může být R pro nové uživatele naročné. Hlavní výhodou R je, že přináší do tvorby prostorové vizualizace silné výpočetní prostředky pro úpravu dat a možnost uložení skriptů a jejich další modifikace. [43]

Obrázek 25: Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `ggplot2`

Prostorové objekty jsou často reprezentovány *vektorovými* daty. Takováto data obsahují popis „geometrie“ nebo „tvaru“ území (body, linie a polygony) a většinou také obsahují proměnné s dodatečnými informacemi o území. Dále se také používají prostorová pole, která jsou obvýkle reprezentována pomocí *rastrů*. Rastrová data se obvykle používají pro kontinuálná proměnné. Rastry dělí území mřížkou na buňky o stejné velikosti neboli na *pixely*, kterým je přiřazená hodnota dle zajmové proměnné. Většinou se jedná o průměr či většinovou hodnotu pro oblast spadájící do konkretního pixelu. Na rozdíl od vektorových dat, v rastrových datech není informace o geometrii objektu uložená jako souřadnice, ale v rámci buňek. Velikost rastrových buněk neboli prostorové rozlišení je dáno počtem řadků a sloupců, na které je území rozdělené. [44]

Vektorová data mohou být vykreslena v R pomocí celé řady balíčků a to jak pomocí již zmíněných `base` (kapitola 4.1) a `ggplot2` (kapitola 4.1.2), tak i pomocí

`Leaflet` (kapitola 4.3.3). Příkladem vizualizace vektorových dat pomocí balíčku `ggplot2` je zobrazeno na obrázku 25.

Obrázek 26: Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `raster`

Balíček `raster` obsahuje funkce pro vytvaření, čtení, manipulaci, modelování, popis a analýzu rastrových dat. Podporuje prácí s velkými soubory a vytvaření vlastních specifických funkcí. [45] Balíček `rasterVis` představuje metody pro vylepšenou vizualizaci a interakci s rastrovými daty. Implementuje vizualizační metody pro kvantitativní a kvalitativní data a to jak pro jednorozměrné, tak i vícerozměrné rastry. Dále poskytuje metody k zobrazení rastrů, měnících se v čase, a vektorových polí. [46] Vykreslení dat pomocí balíčku `raster` lze nalézt na obrázku 26.

4.3 Balíčky pro interaktivní vizualizaci dat

Možnost přiblížit, filtrovat či zobrazit podrobnosti grafu na požádání je výhodou dynamických a interakticních grafů. Interakticní grafy dokažou zobrazovat více informací a tím umožňují i hlubší pochopení dat. Pro R existuje široká nabídka balíčků a nástrojů k vytváření interaktivních grafů. Jednou z nejpopulárnějších[2] možností je balíček `htmlwidgets`, jehož knihovna obsahuje užitečné nástroje pro tvorbu téměř jakéhokoliv typu grafiky.

Balíček `htmlwidgets` poskýtuje rozhrání pro snadné propojení jazyku R a knihoven programovácího jazyku JavaScript, používáného zejména pro webové aplikace. Tímto je umožněna bezproblemová a konzistentní práce interaktivních map, grafů a tabulek a to jak v dokumentech `R Markdown` a v aplikacích `Shiny` (kapitola 4.4.2), tak i v rámci samotného R Studio. `htmlwidgets` umožňuje vytváření vlastních *widgets*, ale především nabízí řadů již vytvořených, například `Plotly`, `Leaflet` a `dygraphs`.

[2]Dle žebříčku rdocumentation.org je `htmlwidgets` 58. nejstahovanější balíček.

Tyto nástroje jsou dostupné nejenom pro R, ale i pro další programovácí jazyky (například Python). [47]

4.3.1 `Plotly`

Balíček `Plotly` je vysokoúrovňové rozhrání pro open source JavaScript vizualizační knihovnu `plotly.js`. Pro tvorbu interaktivních vizualizací využívá jako základ balíček `htmlwidgets`. Dále umožňuje jednoduchý převod `ggplot2` grafů na interaktivní verzi. `Plotly` objekt lze vytvořit dvěmi způsoby, a to pomocí funkce `plot_ly()`, která převádí data na `Plotly` objekt, případně pomocí funkce `ggplotly()`, která převádí `ggplot2` objekt na `Plotly` objekt. Bez ohledu na to, jak je `Plotly` objekt vytvořen, výsledkem je interaktivní vizualizace, která ve své přednastavené verzi obsahuje nástrojovou lištu a umožňuje přiblížení a pohyb po vizualizaci. [48]

4.3.2 `dygraphs`

Balíček `dygraphs` je R rozhráním pro práci s JavaScriptovou vizualizační knihovnou `dygraphs`. Poskytuje bohaté možnosti k vykreslení časových řad v R, včetně podpory interaktivních funkcí. Mezi tyto funkce patří například [49]:

- Automatické vykreslení časových řad typu `xts`[3]
- Vysoce konfigurovatelná nastavení os a vykreslení řad (včetně volitelné další osy y)
- Zvyraznění, přiblížení řad či bodů
- Zobrázení predikčních interválů kolem řad
- Umožnuje vykreslování překrývájících se grafů, vyznačení oblastí zastíněním a označení určitých událostí svislými liniemi a popisky.

4.3.3 `Leaflet`

`Leaflet` je další z populárních open source JavaScript vizualizačních knihoven pro interaktivní mapy. Využívají ji takové webové stránky jako jsou *The New York Times* a *The Washington Post*, ale i *GitHub* a *Flickr*. `Leaflet` je také využíván GIS platformami jako jsou *OpenStreetMap*, *Mapbox* a *CartoDB*.

Tento R balíček umožňuje integrování a ovladání `Leaflet` map. Mezi jeho funkce patří pohyb/přiblížení v rámci mapy, možnost vytvářet mapy z libovolných kombinací (polygony, linie, markery, atd.). Dále snadno vykresluje prostorové objekty z balíčků `sp` nebo `sf` a datové soubory se sloupci zeměpisné šířky a délky. `Leaflet` také poskytuje možnost ovládání interakcí v `Shiny` aplikacích přes stavající hranice mapového výřezu či reakce na kliknutí myši uživatelem. Umožňuje zobrazení map v nesférickým Mercatorově zobrazení a má mnoho dalších funkcí a doplňků. [50]

[3] Objekt `xts` pochází z balíčku `xts`, který je rozšířením populárního balíčku pro práci s časovými řadami `zoo`.

4.4 Balíčky pro webové aplikace

4.4.1 flexdashboard

flexdashboard slouží k publikaci dat a jejích přehledné vizualizaci v rámci webového prohlížeče. Využívá R Markdown k publikaci souvisejících vizualizací do jednotného zobrázení neboli *dashboard*u. Balíček podporuje široký výběr komponentů, včetně htmlwidgets, base, lattice a grid grafiky, tabulek, textových poznámek a dalších. Vyznačuje se mimo jiné i jednoduchým a flexibilním nastavením samotného rozvržení dashboardu a to definováním řadků a sloupců. Komponenty takového dashboardu se pak inteligentně přizpůsobí oknu prohlížeče, případně obrazovce mobilního zařízení. Dále balíček umožňuje nastavení a přepínání mezi jednotlivými záložkami dashboardu. Pro dynamické vizualizace lze kombinovat flexdashboard a Shiny aplikace.

4.4.2 Shiny

Shiny je balíček, umožňující jednoduché vytváření interaktivních webových aplikací kombinací výpočetních možností R s interaktivitou moderních webových stránek. Pomocí Shiny lze vytvořit jak samostatnou aplikaci, běžící na webových strankách, tak i lokální aplikaci bežící v prostředí R. Vzhled Shiny aplikace lze modifikovat pomocí kaskádových stylů[4]. [51]

Strukturou se Shiny aplikace líší od běžných skriptů. Celá aplikace by měla být umístěna v jednom adresáři a celý skript by měl být obsažen v souboru nazvaném app.R.[5] Aplikaci pak lze spustit pomocí příkazu runApp(„cesta"), kde cesta je cesta k adresáři s aplikací. Skript app.R se skládá ze tří komponentů:

- Uživatelské rozhraní ui, které obsahuje rozložení ovládacích prvků a nastavení vzhledu.
- server funkce, který obsahuje funkce generující samotné interaktivní výstupy.
- ShinyApp funkce, která spojuje ui a server.

```
library(shiny)

ui <- fluidPage()

server <- function(input, output){}

shinyApp(ui = ui, server = server)
```

[4]Kaskádové styly (CSS) jazyk určený pro popis vzhledu elementů napsaných ve značkovácích jazycích (například HTML).

[5]V dřívějších verzích Shiny bylo nutné skript rozdělit do dvou částí ui a server.

V rámci `ui` lze používat širokou řadu ovldácích prvků jako jsou například tlačítka, zatrhávací políčka, přepínače, textové vstupy, atd. Běžně se `ui` dělí do tří částí:

- Titulní panel, který obsahuje metadata, název aplikace a další relevantní informace.
- Postraní panel, který nejčastěji obsahuje ovládací prvky.
- Hlavní panel, který obsahuje výstupy generované funkcí `server`.

Příkladem ovladacího prvku vytvořeného v `ui` může být číselný vstup, který umožňuje uživateli vložit libovolné číslo. Tento vstup je označen pomocí `InputId` a v ramcí `server` funkcí lze na něj dle toho id odkázat. Konkretně tento vstup lze získat z proměnné `input$num`. Obecně by to tedy mělo tvar `input$inputId`. Vstup z `ui` lze použit pouze v ramci funkce `render*()`, případně v ramcí funkce `reactive()` (viz dále).

```
ui <- fluidPage(numericInput(inputId = "num",
                             label = "Číselný vstup",
                             value = 10))
```

Funkce `server` přebírá vstupy z ovládacích prvků `ui` a mapuje jednotlivé funkce k výstupům. Příkladem takové funkce může být:

```
output$histPlot <- renderPlot({
  hist(rnorm(input$num, 1, 0))
  })
```

Tato funkce generuje grafický výstup v podobě histogramu normálního rozdělení s interaktivním počtem pozorování. Obecně pro generování výstupu se používá funkce `render*()`. Tyto funkce mohou obsahovat libovolný kód (například příprava dat, doprovodné výpočty, atd.), ale je nutné, aby vráceli požadovaný typ výstupu. Základní `render*()` funkce jsou vypsáne v tabulce 3. Při změně vstupu z `ui` v `render*()` dochází k přepočítání.

Funkce	Výstup
`renderDataTable()`	Interaktivní tabulka
`renderImage()`	Obrázek
`renderPlot()`	Graf
`renderPrint()`	Vytištěný blok kódu
`renderTable()`	Tabulka
`renderText()`	Text
`renderUI()`	`Shiny ui` element

Tabulka 3: Základní `render*()` funkce

Pro komplexní `Shiny` aplikaci s větším počtem interaktivních prvků je vhodné použit funkce `reactive()` (tzv reaktivní výrazy). Tato funkce slouží zejména jako prostředník mezi `ui` a `render*()`. Na obrázku 27 je znazorněn příklad struktury jednoduché `Shiny` aplikace s použitím `reactive()` funkce. V tomto příkladu při změně numerického vstupu (`input$1`) se nejprvé přepočítá `reactive()` a následně `render*()` `A` a `B`, zatímco při změně vstupu `input$2` či `input$3` se přepočítají pouze přislušné `render*()` výstupy `A` či `B` se zachováním informací z `input$1`.

Obrázek 27: Úkazka `Shiny` aplikace s použitím funkce `reactive()`.

Praktická část

5 Metodika

Aplikace je vytvořena prostřednictvím programovacího jazyka R. Jedná se o vizualizaci výsledků modelování projektu ***HAMR*** (viz záložka „O projektu“). Jádro aplikace je postaveno na balíčcích `Shiny` a `flexdashboard`.

`Shiny` umožňuje jednoduché vytváření webových aplikací, interaktivních vizualizací v prostředí R. Balíček `flexdashboard` je rozšířením `Shiny` a umožňuje pokročilé formátovaní vzhledu. Společně slouží k publikaci dat a jejích přehledné vizualizaci v rámci webového prohlížeče.

5.1 Technické řešení

Aplikace je přístupná na serveru[6] nebo ji lze také stáhnout[7]. Repositář na github[8] obsahuje následující soubory:

- soubor s aplikací „flex_app.Rmd“
- skript připravující vstupní data pro aplikací „prep.R“
- skript pro automatickou instalaci potřebných balíčku „install.packages.R“
- soubor s kaskádovými styly pro nastavení vzhledu aplikace „styles.css“

Mimo již zmíněné balíčky `Shiny` a `flexdashboard` byly použité následující balíčky:

- `leaflet`, umožňující vizualizaci prostorových dat v interaktivních mapách
- `ggplot2`, `dygraphs` k vykreslení časových řad a čar překročení
- `sp`, `rgeos` pro práci s prostorovými daty
- `data.table`, `dplyr` sloužící k transformaci dat

5.2 Postprocessing

Na základě kalibrovaných parametrů modelu Bilan jsou vygenerovány denní časové řady odtoku a ostatních bilančních veličin pro období 1961-2015. Pro aplikaci jsou dále vygenerovány další datové produkty:

- měsíční data - pro teplotu jsou průměry z denních hodnot, pro ostatní ukazatele jsou součty denních hodnot
- data pro čáru překročení - pochází z měsíčních dat, které jsou řazeny dle celého období, měsíců a sezon

Na základě evidence užívání vod, kterou spravuje VÚV T. G. M., jsou generovány měsíční časové řady pro období 2006-2016 a tabulky:

[6] https://shiny.fzp.czu.cz/KVHEM/HAMR/
[7] https://github.com/KVHEM/Sucho
[8] https://github.com/

- data pro časovou řadu za vybrané povodí jsou sumami jednotlivých odběratelů seskupenými podle jevů
- tabulka „Suma dle odběratele“ obsahuje sumy za celé období za každého odběratele dle jevu

Aplikace pracuje s velkým množstvím dat, které se v rámci aplikace filtrují, agregují a průměrují. Pro snížení výpočetního času a náročnosti by bylo vhodné některé z těchto úkonů předpočítat a uložit jako datový soubor Rds. Téměř všechny výpočty mohou být předpočítany, ale implementace tohoto postupu proběhne v budoucnu. Je nutné nejdříve změřit a kvantifikovat rozdíl ve výkonu před a po implementaci.

6 Základní rozvržení

6.1 Základní mapa

Obrázek 28: Okno „základní mapa“

Okno prohlížeče je rozděleno do tří části: boční panel, panel s mapovým výstupem a panel s grafickými výstupy.

V bočním panelu se nastavují vstupy pro vykreslení mapy. Volí se zde proměnná hydrologické bilance, podle níž budou zbarveny jednotlivá povodí zobrazené na mapě. Hodnoty jsou agregovány do měsíčních a ročních kroků, lze také vykreslit dlouhodobé průměry, tzn. průměry za celé období a za periody 1961-1990, 1971-2000 a 1981-2010.

Filtrace dle hodnoty v počátečním stavu obsahuje všechny hodnoty zvolené proměnné. Po nastavení rozsahu hodnot se vykreslí pouze povodí odpovídající požadavku.

Vyhledávání útvaru je jedinou částí bočního panelu, která je propojená nejenom s mapou, ale i s grafickými výstupy. Pro zvolené povodí se vypočítají časové řady z měsíčních a denních dat a čára překročení pro celé období, roční období a měsíce (dlouhodobé charakteristiky). Zvolené povodí se zvýrazní v mapě červeným okrajem. Kliknutím na jiné povodí se přepočítají grafické výstupy a název nově zvoleného povodí s jeho UPOV_ID se promítne do pole „Vyhledávání útvaru“. Toto pole lze také použít k ručnímu vyhledávání; stačí vymazat momentálně zobrazený název a začít psát název toku či jeho UPOV_ID.

Grafy mají vlastní výběr proměnných z prvků hydrologické bilance, který se nachází v pravém horním rohu grafického panelu.

Po nastavení údajů k vykreslení lze použít tlačítko „Zobrazit“.

Tlačítko „Reset“ vrací mapový panel na počátečně nastavené souřadnice

Mapový panel obsahuje základní vrstvy (jsou obsaženy v každém mapovém výstupu aplikace), jsou to vrstvy řek, jezer a nádrží a vrstva mapového podkladu. Vrstva povodí se zbarví dle nastavení v bočním panelu. Dále jsou k dispozici vrstvy administrativního členění ČR, vrstva povodí 3. řádu a vrstva horního povodí. Vrstva horního povodí zobrazuje pouze povodí přitékající do vybraného povodí.

V budoucnu bude umožněno přiblížení zvoleného polygonu a jemu odpovídacího okresu, kraje či oblasti povodí 3. řádu. Tento systém přibližování by pak měl být použít i u ostatních mapových výstupů v aplikaci.

Momentálně jsou v grafickém panelu u časových řad zobrazeny trendy pro celé pozorované období. V budoucnu budou grafické výstupy rozšířeny o záložku „Trendy“, kde bude umožněna vizualizace trendů a vyhodnocení statistické významnosti pro vybrané období.

6.2 Indikátory sucha

Obrázek 29: Okno „indikatory“

Záložka indikátory sucha se skládá z bočního panelu, mapového panelu a grafického panelu. V bočním panelu si lze zvolit indikátor, krok, do kterého budou data agregována a datum. Momentálně v mapě je zobrazen indikátor SPI počítány klouzavě s krokem 1, 3, 6 a 12 měsíců. Časem budou přidány indikátory SPEI, PDSI, SGI, SRI a nedostatkové objemy ve stejném kroku. Povodí se rozdělují do 4 kategorií dle hodnot indexu:

- 0 - bez výskytu sucha (-0,7 a vyšší)
- 1 - slabé sucho (-0,8 až -1,3)
- 2 - silné sucho (-1,4 až -1,9)
- 3 - mimořádné sucho (- 2,0 a nižší)

V grafický panelu se vykresluje časová řada indikátoru v konkretním povodí.

6.3 Užívání

Obrázek 30: Okno „uzivani“

Záložka užívání je rozdělena do čtyř části: panel s mapovým výstupem, dva panely s grafickými výstupy a jeden panel s tabulkovým výstupem.

Mapový panel obsahuje informace o užívání vody v ČR. Místa užívání se spojují do shluků. Po přiblížení lze na bod kliknout. Po kliknutí se zobrazí informace o odběrateli a jeho časová řada. Kliknutím na povodí je možno obdržet informace o všech odběratelích v tabulkovém panelu a časovou řadu pro jednotlivé jevy v grafickém panelu.

Podobně jako v záložce „Základní mapa“ zde existuje panel s názvem povodí a jeho UPOV_ID, ve kterém lze nalézt požadovaná povodí. Tento panel lze přetáhnout na libovolné místo v rámci okna prohlížeče.

6.4 Validace

Obrázek 31: Okno „validace“

V současné verzi záložka „Validace“ se dělí na dvě částí: mapový výstup denních průtoků/parametrů (lze přepnout v příslušném panelu) a grafický panel s časovou řadou zvoleného povodí (lze vykreslit v denním, měsíčním a ročním kroku pomocí panelu „Výběr časových řad“).

V dalších verzích bude umožněno porovnání výstupu modelu s pozorováním:

- ve vodoměrných stanicích
- v profilech nádrží
- charakteristiky UPOVu (m-denni vody)
- plošné rozložení parametrů

6.5 O projektu

Obrázek 32: Okno „o projektu“

Krátký popis projektu a veškeré kontakty.V budoucnu bude rozšířeno o metodiky k jednotlivým součástem systému.

Výsledky

Diskuse

Závěr

Přílohy

Seznam obrázků

1 Kombinace různých vuzuálních techník, Playfair 1801 3
2 Dobytek odeslaný z celé Francie ke spotřebě v Paříži, Minard 1858 . . 4
3 Mapa světové migrace, Minard 1858 4
4 Postup Napoleonských vojsk v letech 1812-13, Minard 1869 5
5 Ukázky vizualizací ze začátku 20. století, Brinton 1919 5
6 Snižující se procento rodinných lékařů, Los Angeles Times, 1979 . . . 7
7 Vztah skutečné míry volební registrace k předpovídaným hodnotám, převzato E. Tuftem, 1983 7
8 Radioaktivní oblak při havárii elektrárny Three Mile Island: ^{133}Xe ve vzduchu ve vzdálenosti 375 km (a) a stejný graf přepracovaný Clevelandem (b), 1994 9
9 *Trellis* graf, zobrazující údaje o emisích motoru, Becker et al., 1996 . 10
10 Porovnání základních typů grafů: (a) - bodový, (b) - liniový, (c) - složený, (d) - histogram 13
11 Hustota pravděpodobnosti normálního rozdělení (a) a její distribuční funkce (b) 14
12 Q-Q Graf (a) a P-P Graf (b) 15
13 Boxplot 16
14 Ukázka jednoduchého sloupcového grafu 17
15 Histogram s odhadem hustoty pravděpodobnosti 19
16 Skládaný sloupcový graf transformovaný do polárního souřadnicového systému 19
17 Ukázka jednoduchého koláčového grafu 20
18 Posloupnost datové analýzy 21
19 Mahalanobisovy vzdálenosti 24
20 Log-normální rozdělení transformováné na normální rozdělení 26
21 Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality . . 27
22 Časová řada srážek na zvoleném povodí v roce 2010, `base` 29
23 Časová řada srážek na zvoleném povodí v roce 2010, `lattice` 30
24 Časová řada srážek na zvoleném povodí v roce 2010, `ggplot2` 30
25 Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `ggplot2` 31
26 Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `raster` 32
27 Úkazka `Shiny` aplikace s použitím funkce `reactive()`. 36
28 Okno „základní mapa“ 38
29 Okno „indikatory“ 40
30 Okno „uzivani“ 41
31 Okno „validace“ 42
32 Okno „o projektu“ 43

R kód

Literatura

[1] FRIENDLY, M. A Brief History of Data Visualization. In: *Handbook of Computational Statistics: Data Visualization*. B.m.: Springer-Verlag, 2006. ISBN 978-3-540-32825-4.

[2] BRINTON, W.C. *Graphic Methods for Presenting Facts*. B.m.: The Engineering Magazine Company, New York, 1919. ISBN 978-1155058870.

[3] TUKEY, J.W. The Future of Data Analysis. *Annals of Mathematical Statistics*. 1962.

[4] TUKEY, J.W. *Exploratory Data Analysis*. B.m.: Addison-Wesley, 1977. ISBN 0201076160.

[5] *How William Cleveland Turned Data Visualization Into a Science [vid. 2.11.2017]* [online]. Dostupné z: https://priceonomics.com/

[6] RAHLF, T. *Data Visualisation with R: 100 Examples*. B.m.: Springer, 2017. ISBN 978-3-319-49751-8.

[7] TUFTE, E.R. *Envisioning Information*. B.m.: Graphics Pr, 1990. ISBN 978-0961392116.

[8] TUFTE, E.R. *The Visual Display of Quantitative Information*. B.m.: Graphics Press, 1983. ISBN 978-0961392109.

[9] CLEVELAND, W.S. a R. MCGILL. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the*

American Statistical Association. 1984.

[10] CLEVELAND, W.S. *The Elements of Graphing Data*. B.m.: Hobart Press, 1994. ISBN 0963488414.

[11] BECKER, R.A., W.S. CLEVELAND, M.-J. SHYU, aj. A tour of Trellis graphics. *Murray Hill, NJ: AT & T Bell Laboratories*. 1996.

[12] WILKINSON, L. *The Grammar of Graphics*. B.m.: Springer, 2005. ISBN 9780387286952.

[13] WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis (Use R!)*. B.m.: Springer, 2010. ISBN 978-0387981406.

[14] WICKHAM, H. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*. 2010.

[15] CLEVELAND, W.S. *Visualizing Data*. B.m.: Hobart Press, 1993. ISBN 978-0963488404.

[16] MURRAY, S. *Interactive Data Visualization for the Web*. B.m.: O'Reilly Media, 2013. ISBN 978-1449339739.

[17] FERSTER, B. *Interactive Visualization: Insight through Inquiry (MIT Press)*. B.m.: The MIT Press, 2012. ISBN 978-0262018159.

[18] *R: The R Graphics Package [vid. 22.4.2017]* [online]. Dostupné z: https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html

[19] TEETOR, P. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. B.m.: O'Reilly Media, 2011. ISBN 9781449307264.

[20] *R: Generic X-Y Plotting [vid. 11.5.2017]* [online]. Dostupné z: https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html

[21] *P–P plot - Wikipedia [vid. 11.8.2017]* [online]. Dostupné z: https://en.wikipedia.org/wiki/P%E2%80%93P_plot

[22] CHANG, W. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. B.m.: O'Reilly Media, 2012. ISBN 9781449363116.

[23] NOVOVIČOVÁ, J. *Pravděpodobnost a matematická statistika*. B.m.: Praha: Vydavatelství ČVUT, 2006. ISBN 80-01-01980-2.

[24] MACIEJEWSKI, R. *Data Representations, Transformations, and Statistics for Visual Reasoning*. B.m.: Morgan & Claypool Publishers, 2011. ISBN 978-1-608-45625-

3.

[25] *Histogram - Wikipedia [vid. 6.8.2017]* [online]. Dostupné z: https://en.wikipedia.org/wiki/Histogram

[26] WICKHAM, H. a G. GROLEMUND. *R for Data Science*. B.m.: O'Reilly Media, 2017. ISBN 1491910399.

[27] TUKEY, J.W. Bias and Confidence in Not Quite Large Samples. *Annals of Mathematical Statistics*. 1958, roč. 29.

[28] EFRON, B. a R. TIBSHIRANI. *An Introduction to the Bootstrap*. B.m.: Taylor & Francis Ltd, 1994. ISBN 0-412-04231-2.

[29] MCINTOSH, A. The Jackknife Estimation Method. *arXiv*. 2016.

[30] DE MAESSCHALCK, R., D. JOUAN-RIMBAUD a D.L. MASSART. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*. 2000, roč. 50, č. 1.

[31] MAHALANOBIS, P.C. On the generalised distance in statistics. *Proceedings National Institute of Science*. 1936.

[32] CARDINALI, C. Observation influence diagnostic of a data assimilation system. *Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics: Special Issue*. 2014.

[33] PECÁKOVÁ, I. Problém chybějících dat v dotazníkových šetřeních. *Politická ekonomie*. 2014. ISSN 2336-8225.

[34] *Transformation of Data [vid. 3.2.2018]* [online]. Dostupné z: http://statisticalconcepts.blogspot.cz/2010/02/transformation-of-data-validity-of.html

[35] HEBÁK, P., J. HUSTOPECKÝ, E. JAROŠOVÁ, aj. *Vícerozměrné statistické metody (1)*. B.m.: Informatorium, Praha, 2007. ISBN 80-7333-025-3.

[36] ABDI, H. a L. WILLIAMS. Normalizing data. Encyclopedia of research design. *Thousand Oaks, CA: Sage*. 2010.

[37] ZUMEL, N. a J. MOUNT. *Practical Data Science with R*. B.m.: Manning, 2014. ISBN 9781617291562.

[38] KUTNER, M.H., C.J. NACHTSHEIM, J. NETER, aj. *Applied Linear Statistical Models*. B.m.: McGraw-Hill/Irwin, 2004. ISBN 0-07-238688-6.

[39] SHAPIRO, S.S. a M.B. WILK. An analysis of variance test for normality (complete samples)†. *Biometrika*. 1965.

[40] ÖZTUNA, D., A.H. ELHAN a E. TÜCCAR. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions.

Turkish Journal of Medical Sciences. 2006.

[41] MURRELL, P.R. *Investigations in Graphical Statistics*. B.m., 1998. PhD thesis. The University of Auckland.

[42] MURRELL, P. *grid: The grid graphics package*. červenec 2003

[43] CHESHIRE, J. a R. LOVELACE. Spatial data visualisation with R. *Geocomputation: A Practical Primer*. 2015.

[44] HIJMANS, R. *Spatial Data Analysis and Modeling with R [vid. 13.4.2018]* [online]. 2016. Dostupné z: http://rspatial.org/

[45] HIJMANS, R.J. raster: Geographic Data Analysis and Modeling. *R package version 2.6-7* [online]. 2017. Dostupné z: https://cran.r-project.org/web/packages/raster/raster.pdf

[46] LAMIGUEIRO, O.P. rasterVis: Visualization Methods for Raster Data. *R package version 0.44* [online]. 2018. Dostupné z: https://cran.r-project.org/web/packages/rasterVis/rasterVis.pdf

[47] VAIDYANATHAN, R., Y. XIE, J. ALLAIRE, aj. htmlwidgets: HTML Widgets for R. *R package version 1.0* [online]. 2018. Dostupné z: https://cran.r-project.org/web/packages/htmlwidgets/htmlwidgets.pdf

[48] SIEVERT, C. *plotly for R* [online]. B.m.: bookdown, 2018. Dostupné z: https://plotly-book.cpsievert.me/

[49] *dygraphs for R [vid. 11.4.2018]* [online]. Dostupné z: http://rstudio.github.io/dygraphs/

[50] *Leaflet for R [vid. 11.4.2018]* [online]. Dostupné z: http://rstudio.github.io/leaflet/

[51] *Shiny from R Studio [vid. 14.4.2018]* [online]. Dostupné z: https://shiny.rstudio.com/