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 . . . 14
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
3.5 Tidy data . . . 28
4 Pokročilá vizualizace v R . . . 29
4.1 Balíčky pro vizualizaci dat . . . 29
4.1.1 `grid` . . . 30
4.1.2 `lattice` . . . 30
4.1.3 `ggplot2` . . . 30
4.2 Balíčky pro prostorovou vizualizaci . . . 32
4.3 Balíčky pro interaktivní vizualizaci dat . . . 34
4.3.1 `Plotly` . . . 34
4.3.2 `dygraphs` . . . 34
4.3.3 `Leaflet` . . . 35
4.4 Balíčky pro webové aplikace . . . 35
4.4.1 `flexdashboard` . . . 35
4.4.2 `Shiny` . . . 35

Praktická část **39**
5 Metodika 39
5.1 Technické řešení 39
5.2 Data 40
5.3 Postprocessing 43
6 Základní rozvržení 44
6.1 Základní mapa 45
6.2 Indikátory sucha 47
6.3 Užívání 48
6.4 Validace 48
6.5 O projektu 50

Výsledky **51**

Diskuse **52**

Závěr **53**

Literatura **54**

Přílohy **59**
R kód 59

Seznam obrázků

1 Kombinace různých vizuálních techník, (Playfair 1801) 3
2 Dobytek odeslaný z celé Francie ke spotřebě v Paříži, Minard (1858), převzato z (Palsky 1996) 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 Tří základní pravidla pro *tidy data*. 28
23 Časová řada srážek na zvoleném povodí v roce 2010, `base` 31
24 Časová řada srážek na zvoleném povodí v roce 2010, `lattice` 31
25 Časová řada srážek na zvoleném povodí v roce 2010, `ggplot2` 32
26 Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `ggplot2` 33
27 Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `raster` 33
28 Ukázka `Shiny` aplikace s použitím funkce `reactive()`. 38
29 Útvary povrchových vod (UPOV) (vyznačené černě) a kraje (vyznačené červeně) ČR 40
30 Struktura složek aplikace 41
31 Základní rozlozeni okna aplikace 44
32 Okno „základní mapa“ 45
33 Okno „indikatory“ 47

34 Okno „uzivani“ 48
35 Okno „validace“ 49

Seznam tabulek

1 Základní atributy parametru ‘type‘ 13
2 Funkce pro práci s rozděleními 14
3 Soubor ve „vícesloupcovém“ formátu. 28
4 Soubor v „dlouhém“ formátu. 28
5 Základní `render*()` funkce 37
6 Výstupy kalibrace denního modelu Bilan 42
7 Parametry denního modelu Bilan 43
8 Přehled funkcí z balíčku `CatCa` pro přípravu dat 44
9 Rozdělení indikátorů sucha do kategorií. 47

Úvod

Cíle práce

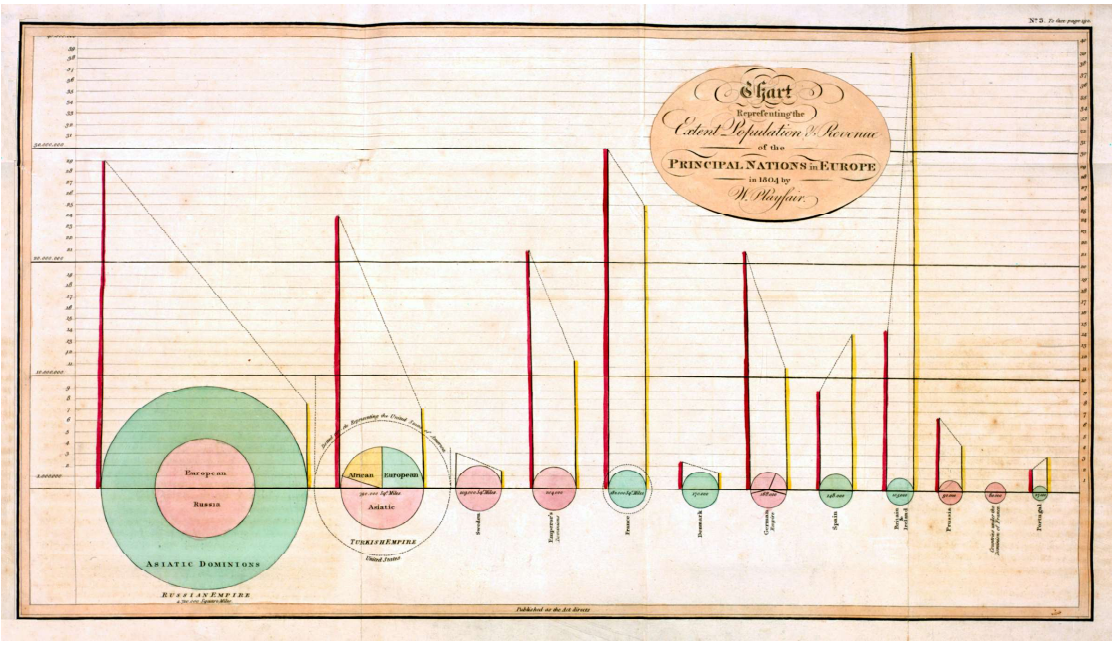
Teoretická část

1 Vizualizace dat

1.1 Historie vizualizace dat

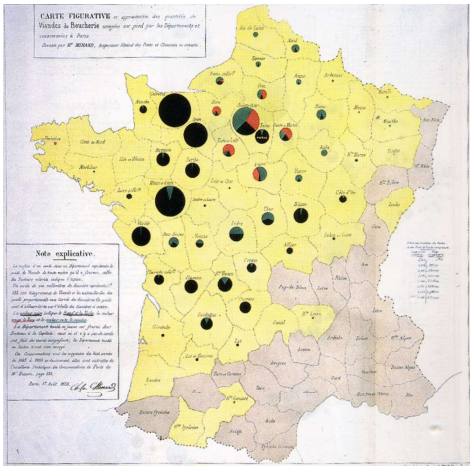
Do 17. století jediné, co by se dalo nazvat jako vizualizace dat byly mapy pro navigaci a průzkum, ale také diagramy, geometrická schémata a tabulky pozic hvězd a jiných nebeských těles. Postupný vývoj statistické teorie a růst zájmu o data na konci 18. století vedly k inovacím a expanzi nových grafických 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“* (Playfair 1786). Později popsal i koláčový graf ve svém breviáři *„Statistical Breviary“* z roku 1801 (Playfair 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).

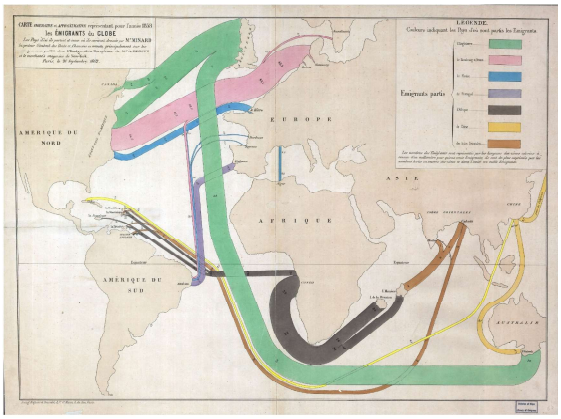


Obrázek 1: Kombinace různých vizuálních techník, (Playfair 1801)

V polovině 19. století byly vytvořeny všechny podmínky pro rychlý vývoj 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 3. Jednou z jeho nejslavnějších prací je zobrazení postupných ztrát francouzské armády během Napoleonského tažení na Moskvu v letech 1812-1813 (obrázek 4), která 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.

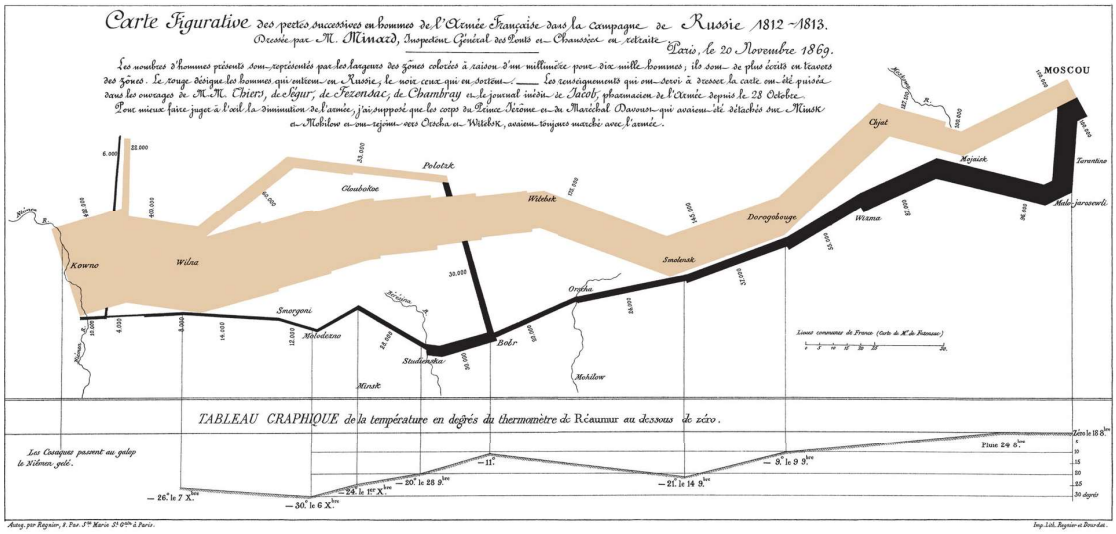


Obrázek 2: Dobytek odeslaný z celé Francie ke spotřebě v Paříži, Minard (1858), převzato z (Palsky 1996)



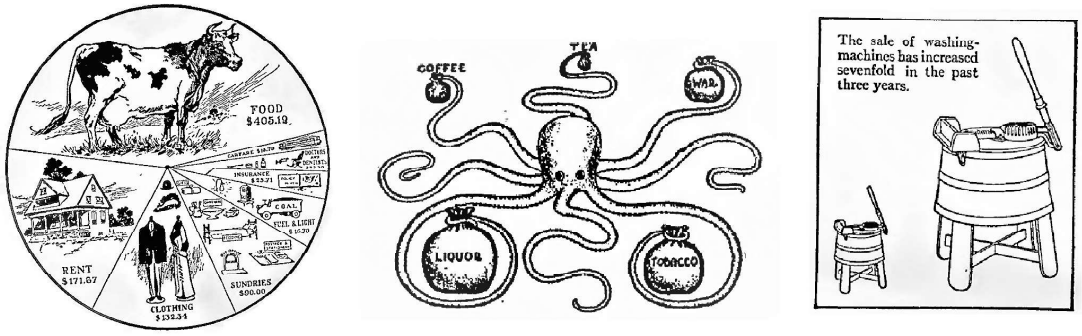
Obrázek 3: Mapa světové migrace, (Minard 1858)

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. (Friendly 2006)



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 5. Tento graf je příkladem nepovedené vizualizace: oko preferenčně soudí dle velikostí obrázků a ne dle uhlů výsečí. Obrázek uprostřed znázorňuje výdaje za různé komodity: 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“ (Brinton 1919).



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 (Tukey 1962). 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, více o tématu viz kapitola 3) (Tukey 1977). Mezi těmito novými grafy je například číslicový histogram (popsaný v kapitole 2.4.3), boxplot neboli krabicový graf (popsaný v kapitole 2.3.2) a další. Mnoho z nich je aktivně používáno ve statistické praxi a implementováno do většiny softwarů (Friendly 2006).

1.2 Zásady vizualizace dat

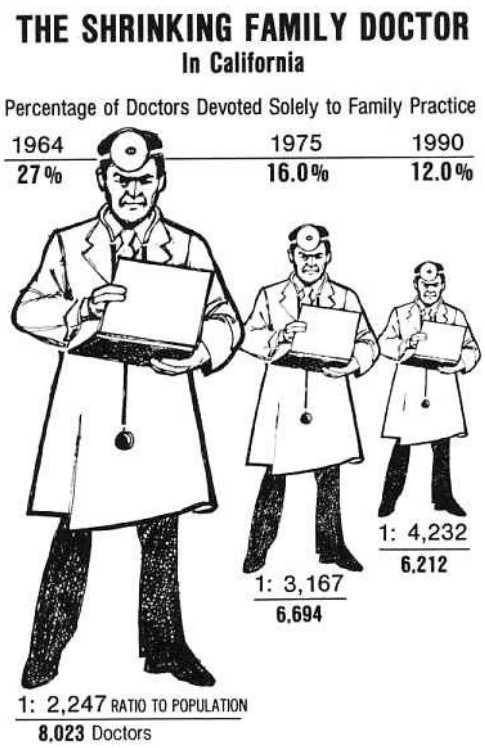
Od roku 1975 se vyvíjí statistické výpočetní systémy a s nimi i nové metody analýzy a vizualizace dat. V tomto období vizualizace začala být vnímána jako vlastní odvětví a to především díky Williamu S. Clevelandovi a Edwardu Tuftemu, kteří položili vědecké základy tohoto odvětví. Tufte vyvinul a popularizoval terminologii a základní principy grafické integrity. Cleveland se zabýval studiemi grafického vnímání, kognitivních procesů, které lidé používají k pochopení grafů, a rozvíjel teorii o správném provedení vizualizace. Výsledek jejich práce se promítá i do současné doby v podobě kvalitních, interaktivních a dynamických vizualizací (Friendly 2006).

1.2.1 Edward Tufte

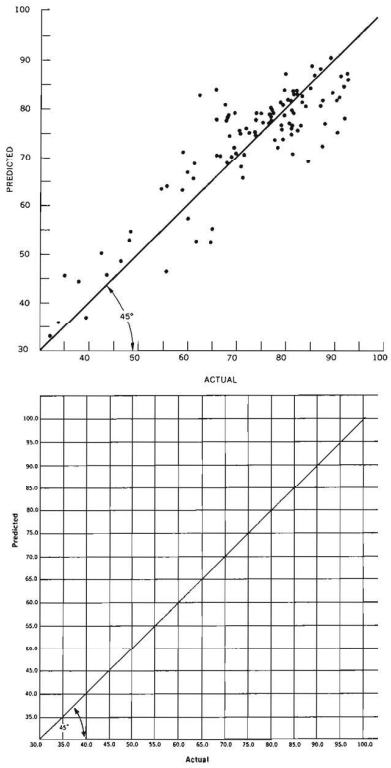
Za revoluční průlom se považuje kniha Edwarda Tufteho *The Visual Display of Quantitative Information* z roku 1983, v kombinaci se dvěma posléze publikovanými knihami *Envisioning Information* a *Visual Explanations* z roku 1990, resp. 1997, patří mezi nejznámější publikace na téma vizualizace dat. Právě v těchto publikacích Tufte originálním způsobem definuje „standard“ vizualizace (Rahlf 2017). 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 4). Tufte říká, že grafická elegance se často nachází v jednoduchosti návrhu a komplexnosti dat (Tufte 1990). Tufte 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
  - prezentovat 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 zavádějící 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. Tufte 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

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

  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.* Tufte 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 (Tufte 1983).

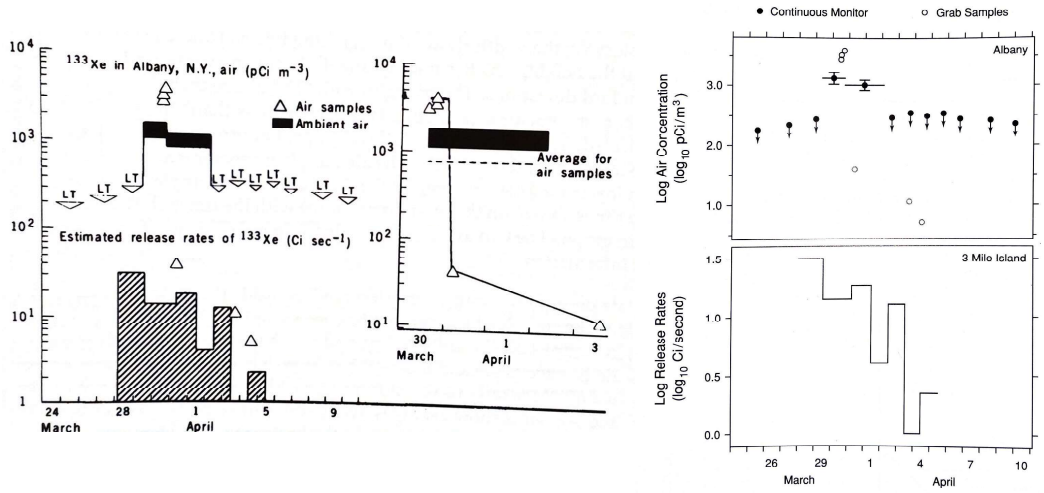
1.2.2 Wiliam S. Cleveland

Kromě práce Edwarda Tufteho měly velký vliv i publikace Wiliama S. Clevelanda. Cleveland se svým kolegou Robertem McGillem publikovali v roce 1984 článek o grafickém vnímání (Cleveland a McGill 1984). 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. 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 přesahoval práci Tufteho. Zásady a principy dle Clevelanda by se daly shrnout do čtyř hlavních kategorií: jasná vize, srozumitelnost, měřítka a obecné postupy (Cleveland 1994).

- **Jasná vize**
  - Data by měla být středem vizualizace, 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.).

- Počet značek na osách by měl být přiměřený.
- 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 se měly vyskytovat mimo oblast grafu (případně v legendě), totéž se týká poznámek a nadpisů, které mohou být také umístěny do textu.
- Překrývající se datasety 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.



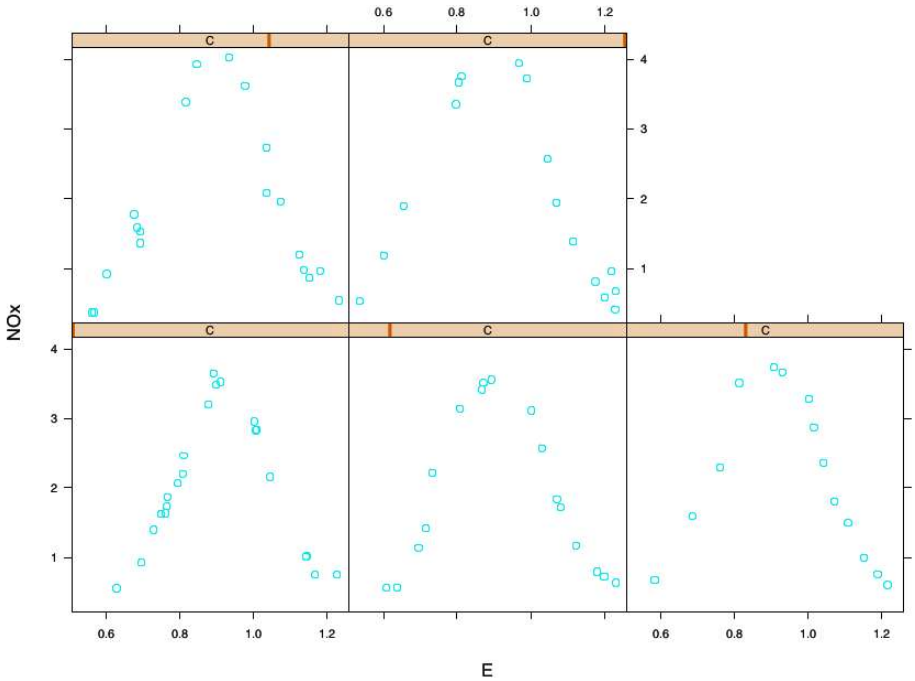
(a) (b)

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 měly 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 (Becker et al. 1996).



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 (Wilkinson 2005) detailně popisuje prvky, které tvoří základ všech statistických grafů a odpovídá na základní otázku: co je statistická grafika? (Wickham 2010b) 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* (Wickham 2010a), který se zaměřuje primárně na vrstvy vizualizační grafiky a jejich zapojení do jazyka R. Následně také pro něj posloužila jako inspirace pro tvorbu balíčku `ggplot` (viz kapitola 4.1.3).

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 použit 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**ingu popisuje, jaké proměnné by měli být použity k rozdělení dat na podmnožiny a jak by tyto podmnožiny měly být uspořádány. Jedná se o mocný nástroj pro zkoumání toho, zda jsou statistické 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 (Cleveland 1993) nebo Tukey (Tukey 1977), 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 (Wickham 2010b). 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 (Murray 2013) nebo *Interactive Visualization* od Billa Ferstera (Ferster 2012).

2 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` (R-Documentation [vid. 22.4.2017]), který obsahuje mnoho užitečných funkcí pro tvorbu grafických prvků. Tato kapitola se soustředí na tento balíček, zatímco v kapitole 4 jsou popsány funkce dalších široce používaných balíčků (například `lattice` 4.1.2 či `ggplot2` 4.1.3), které nabízí podobné funkce, avšak s různým rozsahem nastavení (Teetor 2011).

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 příklady 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 (Teetor 2011):

```
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 zobrazena 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 jako výchozí pro numerické hodnoty. Viz obrázek 10 (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 (Teetor 2011) (viz obrázek 10 (a), (b)). Pro vykreslení liniového

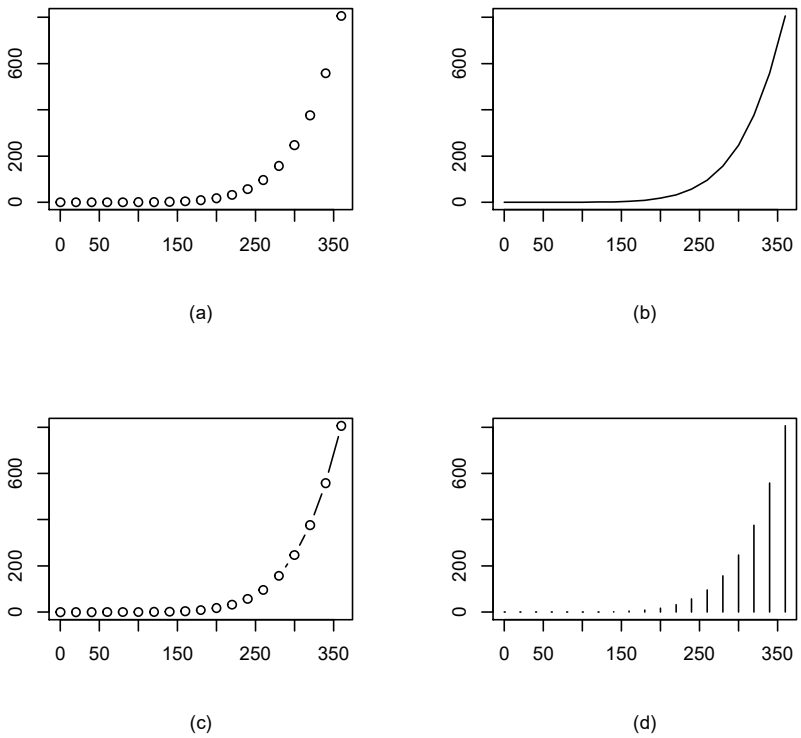
grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

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

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity (R-Documentation [vid. 11.5.2017]):

	Anglický popis	Český popis
p	points	bodový
l	lines	liniový
b	both	složený
h	histogram	tyčkový
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

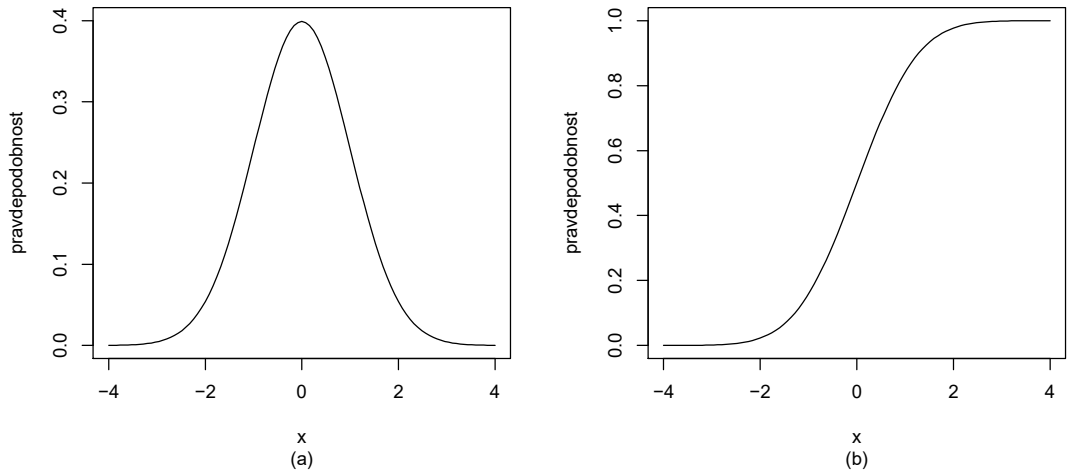
Teorie pravděpodobnosti je základem statistiky a R má hodně nástrojů pro práci s pravděpodobností, rozdělením pravděpodobnosti a náhodnými proměnnými. R má zkrácený název pro každé rozdělení pravděpodobnosti (Teetor 2011). Tyto názvy slouží k identifikaci funkcí spojených s rozděleními. Například zkrácený 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 11):

```
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 12) 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 2.4.1), 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 (nebo jiného vybraného rozdělení) a na druhé ose kvantily naměřené (pozorované). Pokud data mají přesně požadované rozdělení, všechny body grafu leží na přímce pod úhlem 45°. Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 21. (Teetor 2011) (Cleveland 1994)

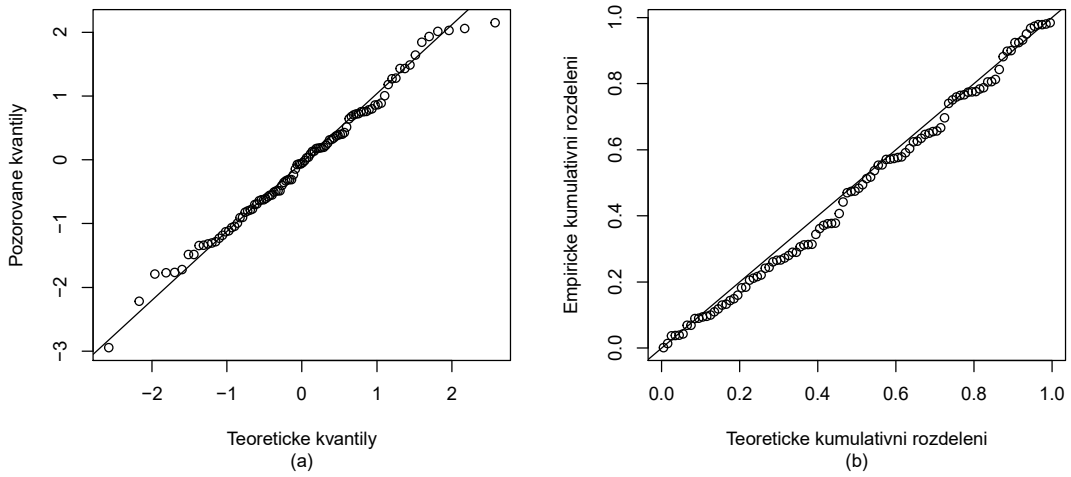
Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkce proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se pravděpodobně o požadované rozdělení. P-P graf se často používá k vyhodnocení koeficientu šikmosti rozdělení.(Wikipedia [vid. 11.8.2017])

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)
```

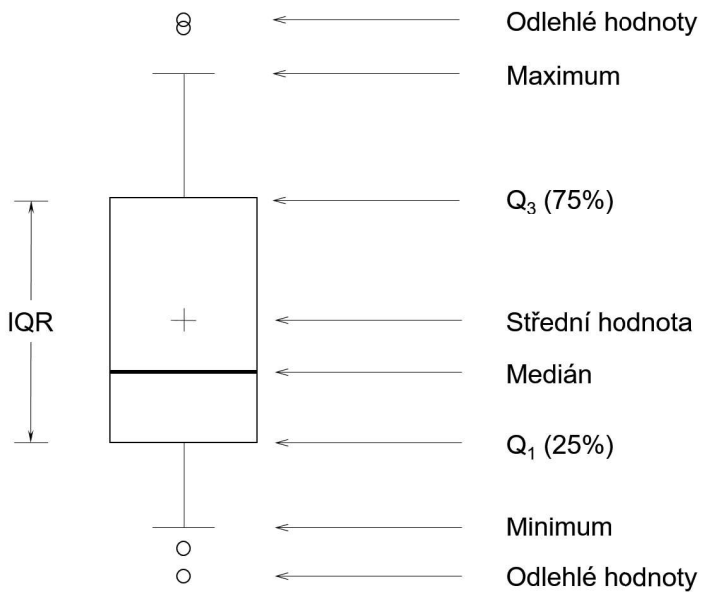


Obrázek 12: Q-Q Graf (a) a P-P Graf (b)

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 13 znázorňuje typický krabicový graf, kde silná čára je medián, krabice kolem ní určuje polohu prvního a třetího kvartilu (dolní Q_1 kvantil 25% a horní Q_3 kvantil 75%). "Vousy" (*whiskers*) nad a pod krabicí 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

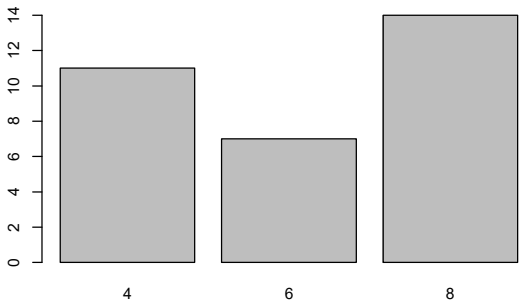
Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.(Chang 2012)

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 14) 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 také známý jako histogram (Chang 2012). 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 (Novovičová 2006). 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 (Maciejewski 2011)

```
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 (Maciejewski 2011)

```
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 (Wikipedia [vid. 6.8.2017])

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

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

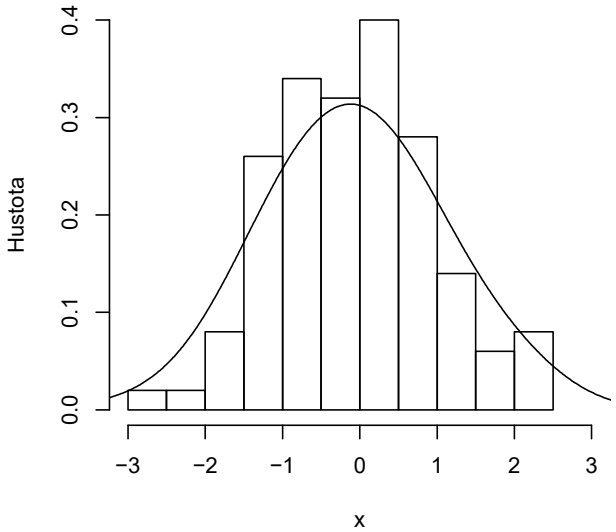
$$h = 2\frac{IQR(x)}{n^{\frac{1}{3}}}$$

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.

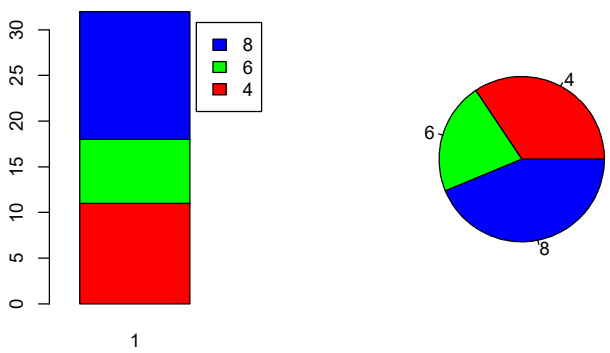
Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 15).



Obrázek 15: Histogram s odhadem hustoty pravděpodobnosti

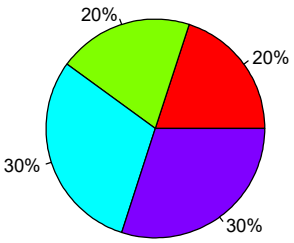
2.4.2 Koláčový graf

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) (Wilkinson 2005).



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

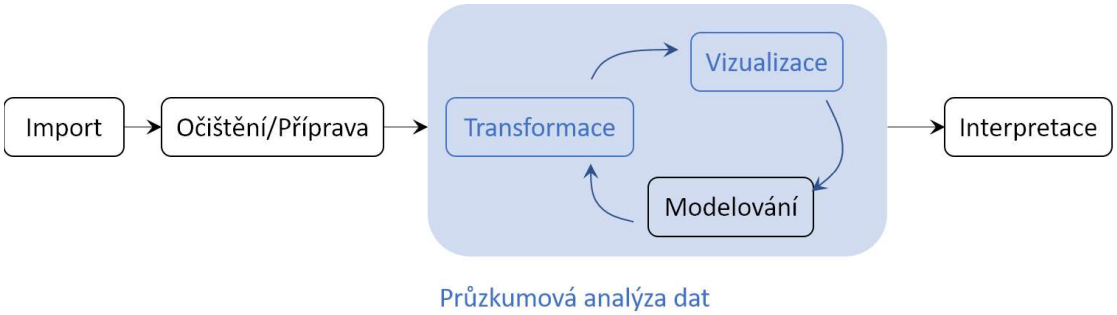
2.4.3 Číslicový histogram (*stem-and-leaf*)

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 datasetů. 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. (Tukey 1977) 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**. Připravou dat je myšleno ukládání dat v konzistentní a systematické formě, odpovídající sémantice původního datasetu. Očištěná data jsou taková data, ve kterých sloupce odpovídají proměnným, 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.

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 rámci 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 (Wickham a Grolemund 2017).

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. Tukeyem v roce 1958 v „*The Annals of Mathematical Statistic*“ (Tukey 1958) a jedná se o speciální případ metody *bootstrap* (více o metodě B. Efron a R. Tibshirani v „*An Introduction to the Bootstrap*“ (Efron a Tibshirani 1994)).

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. Jedna 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. (McIntosh 2016)

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

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

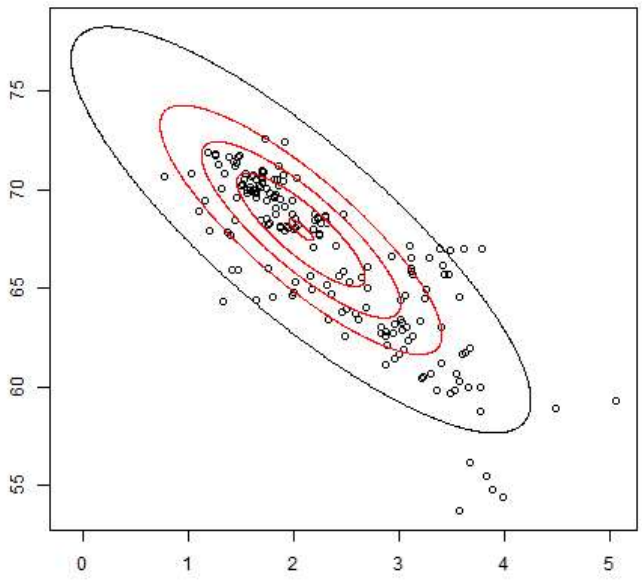
zatímco mahalanobisova vzdálenost se spočítá jako

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

kde $\boldsymbol{C}_x$ je kovarianční matice. (De Maesschalck et al. 2000)

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“* (Mahalanobis 1936). 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. (Cardinali 2014)

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 odhadnuty 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. (Pecáková 2014)

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. (Hebák et al. 2007). 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:

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

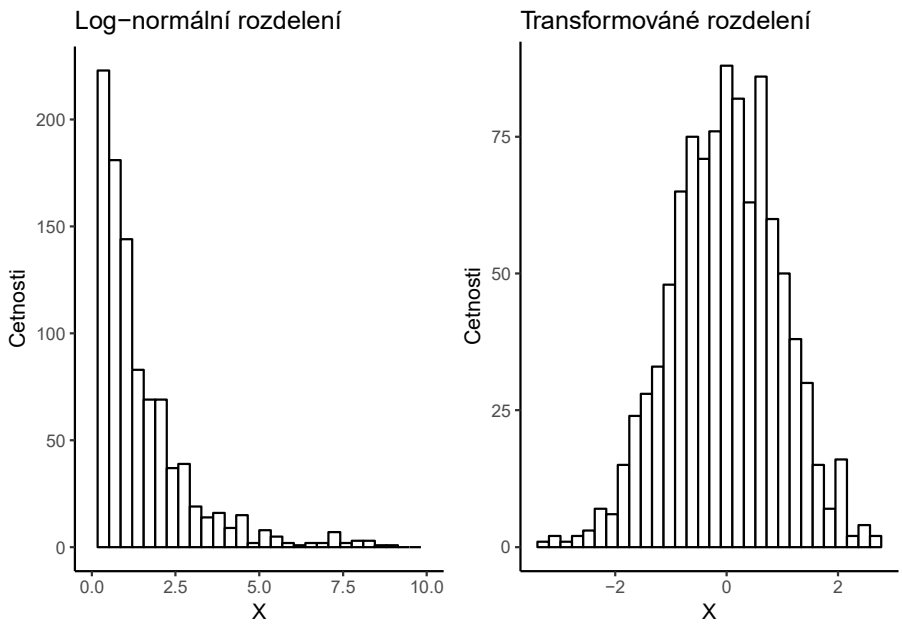
Vektor průměrů $\bar{\boldsymbol{v}}$ je nulový, kovariance a korelace proměnných zůstává nezměněna. (Hebák et al. 2007) Další často využívanou metodou je metoda normalizace dat. Tato metoda transformuje měřítka vzorků pro možnost jejich porovnání (eliminuje jednotky měření), po úpravě střední hodnota vzorku tedy odpovídá nule a směrodatná odchylka jedničce.

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

$\sigma(x_j)$ je směrodatná odchylka sloupce proměnné, vektor průměrů $\bar{\boldsymbol{z}}$ je nulový a kovariance vektoru nových proměnných se shoduje s korelací původního vektoru. (Abdi a Williams 2010)

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. (Zumel a Mount 2014) (Kutner et al. 2004)

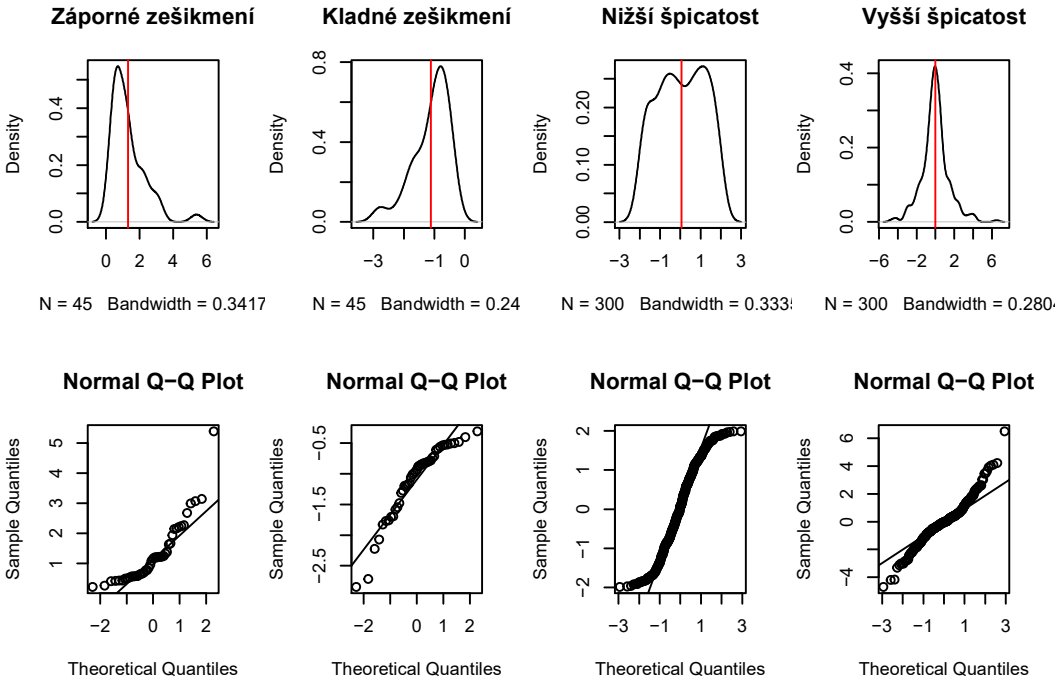


Obrázek 20: Log-normální rozdělení transformováné na normální rozdělení

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 popsány 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 (SHAPIRO a WILK 1965). 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. Kritické hodnoty pro tento test jsou tabelovány.

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

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

kde n je velikost vzorku, $\sqrt{b_1}$ je koeficient šikmosti vzorku a b_2 je koeficient špičatosti. T statistika má přibližně chi-kvadrát rozdělení s dvěma stupni volnosti. Nulová a alternativní hypotéza se schoduje s SW testem. Používá se pro větší datasety nad 2000 pozorování. (Öztuna et al. 2006)

3.5 Tidy data

Průzkumovou analýzou dat se například zabýval Hadley Wickham ve svým článku *Tidy data* z roku 2014 (Wickham 2014). Poskytuje standardy a koncepty pro čištění dat, spolu se základními metody a nástroje pro jejich připravu. Hlavním cílem *tidy dat*, aby každá proměnná odpovídala sloupci, každý pozorování odpovídalo řadku a každá hodnota měla vlastní buňku (viz obrázek 22).

Obrázek 22: Tří základní pravidla pro *tidy data*.

Mezi těmito nástroje jsou takové balíčky R jako `tidyr`, `plyr` a `dplyr`. `tidyr` (tidyr [vid. 17.4.2018]) je balíček vytvořený přesně pro tvorbu *tidy dat* a to pomocí dvou hlavních funkcí `gather()` a `spread()`. Funkce `gather()` převádí vícesloupcové (víceproměnné) data do formátu „klíč-hodnota“ resp. do „dlouhého“ formátu, zatímco funkce `spread()` činí přesný opak. Viz tabulky 3 a 4. Balíček `plyr` (plyr [vid. 17.4.2018]) je soubor nástrojů sloužící k rozdělení velkých souborů dat do stejnorodých menších podsouborů, aplikací funkcí na jednotlivé podsoubory a kombinací vysledků zpatky do jednoho souboru. `dplyr` (dplyr [vid. 17.4.2018]) je další iteraci `plyr` představenou Hadley Wickhamem v roce 2014. Balíček poskytuje sadu nástrojů pro efektivní manipulaci s datasety v R, je rychlejší a jednodušší na použití.

DTM	P	E	R	T
1948-01-01	5.86	1.14	1.54	15.48

Tabulka 3: Soubor ve „vícesloupcovém“ formátu.

DTM	variable	value
1948-01-01	P	5.86
1948-01-01	E	1.14
1948-01-01	R	1.54
1948-01-01	T	15.48

Tabulka 4: Soubor v „dlouhém“ formátu.

4 Pokročilá vizualizace v R

Před samotným popisem nástrojů pro pokročilou vizualizaci je vhodné vysvětlit pár důležitých pojmů, které jsou v této kapitole použity. Tyto pojmy se tykají vizualizace pouze okrajově, avšak v praxi jsou často používány právě v kombinaci s vizualizačními prostředky.

`R Studio` je volné přístupné open source vývojové prostředí (*Integrated Development Environment* nebo *IDE*) pro programovací jazyk R. `R Studio` bylo založeno v roce 2008 J.J. Allairem a hlavním vývojářem je Hadley Wickham. Jedná se o vývojové prostředí obsahující veškeré potřebné nástroje pro práci s R (RStudio [vid. 16.4.2018]). `R Studio` kromě vývoje samotného *IDE* vyvíjí jedny z nepopulárnějších balíčků (např. `ggplot2`, `tidyverse`, `rmakrdown`, ...) a servery pro `Shiny`. `R Studio` je dostupné ve verzi zdarma, která poskytuje všechny klíčové funkce a v komerční verzi, pro kterou je navíc dostupná oficiální podpora (RStudio [vid. 16.4.2018]).

`R Markdown` je založen na značkovacím jazyku `Markdown` a slouží pro úpravu prostého textu a jeho následný převod do HTML, PDF, MS Word a dalších formátů a to rovnou z prostředí R (rmarkdown [vid. 16.4.2018]). Balíček využívá univerzální nástroj pro převod souborů `pandoc`, díky čemuž lze v rámci souboru používat LaTeX příkazy pro pokročilou úpravu PDF výstupů. Dále díky balíčku `knitr` je umožněna integrace R kódu do výstupů. Více o `R Markdown` píše autor a spoluautor těchto balíčků Yihui Xie například (Xie 2018) a (Xie 2015).

4.1 Balíčky pro vizualizaci dat

Vestavěný 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 (není např. možné uložit graf jako proměnnou). `base` obsahuje nástroje pro kreslení jak základních, tak i kompletních grafik. 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 krátký popis vybraných, široce využívaných, 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, uvádějící všechny grafické funkce dostupné v ostatních balíčcích vyvinutých komunitou lze nalézt na stránce http://cran.r-project.org/web/views/Graphics.html. (Wilkinson 2005)

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 (Murrell 1998). Grafické objekty v `grid` mohou být reprezentovány nezávislé na grafu a později upraveny. Systém *viewportů* (každý obsahuje vlastní souřadnicový systém) usnadňuje rozvržení komplexní grafiky. Balíček umožňuje vykreslení jednotlivých výkresů, avšak nemá explicitní nástroje pro tvorbu statistických grafů (Murrell 2003).

4.1.2 `lattice`

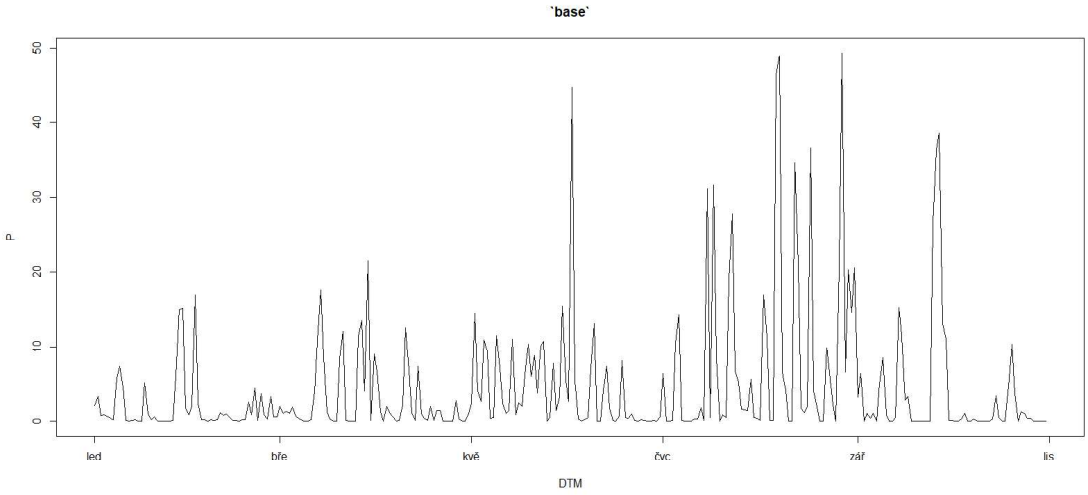
Deepayanem Sarkarem vyvinutý balíček `lattice` používá grafiku `grid` k implementaci Clevelandova *trellis* grafického systému (viz kapitola 1.2.2), což vedlo k značnému vylepšení grafiky oproti `base`. Pomocí `lattice` lze jednoduše vykreslit *trellis* graf a některé detaily výkresu, jako například 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 požadavků.

4.1.3 `ggplot2`

Balíček `ggplot2` byl vyvinut v roce 2005 Hadley Wickhamem na základě *The Grammar of Graphics* Lelanda Wilkinsona (Wilkinson 2005). `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 vzhledu na mnohem nižší úrovni.

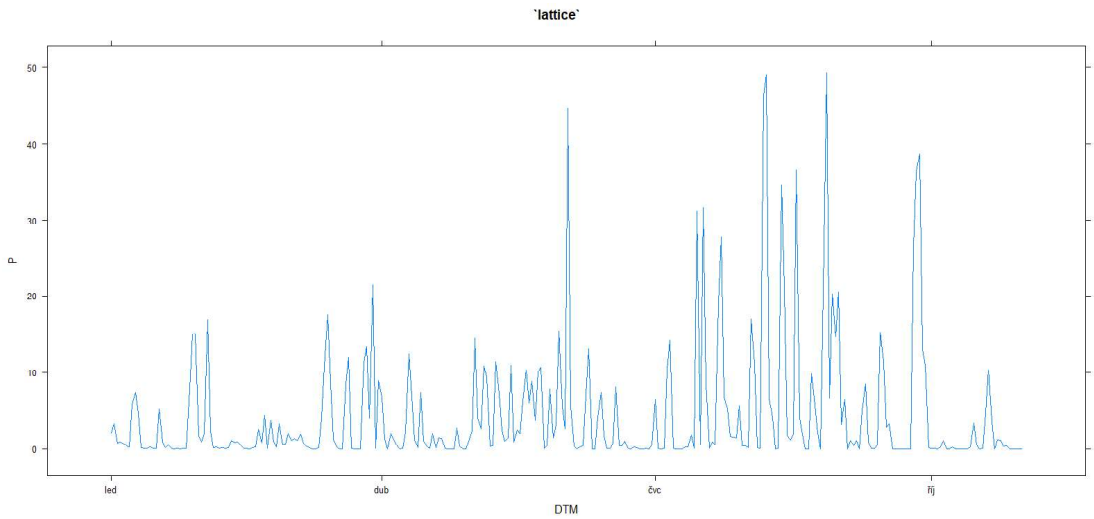
Jako ukázku lze předvést graf vykreslený pomocí různých balíčků (obrázky 23, 24, 25). V představeném kódu byly použity denní časové řady srážek z roku 2010 v oblasti povodí Labe od pramene po Svatopetrský potok včetně. ID tohoto útvaru povrchových vod je `HSL_0010`. `DTM` odpovídá datu v denním kroku a `P` odpovídá úhrnu srážek v milimetrech. Každý balíček má přednastavený počáteční vzhled, který lze upravovat dle požadavků, avšak pro tuto ukázku bylo ponecháno základní nastavení.

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



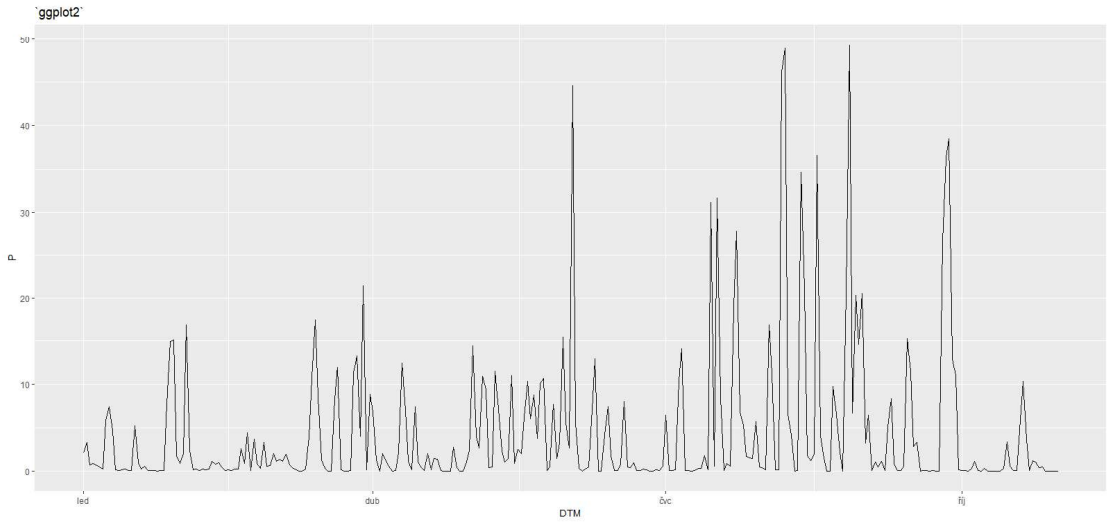
Obrázek 23: Časová řada srážek na zvoleným povodí v roce 2010, `base`

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



Obrázek 24: Časová řada srážek na zvoleným povodí v roce 2010, `lattice`

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

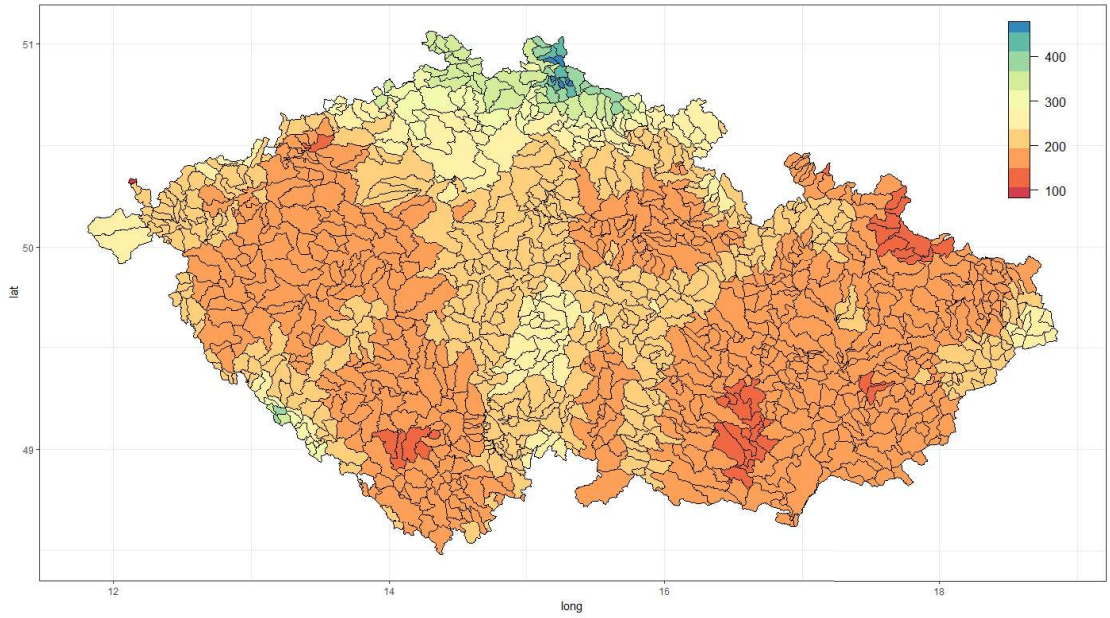


Obrázek 25: Časová řada srážek na zvoleném povodí v roce 2010, `ggplot2`

4.2 Balíčky pro prostorovou vizualizaci

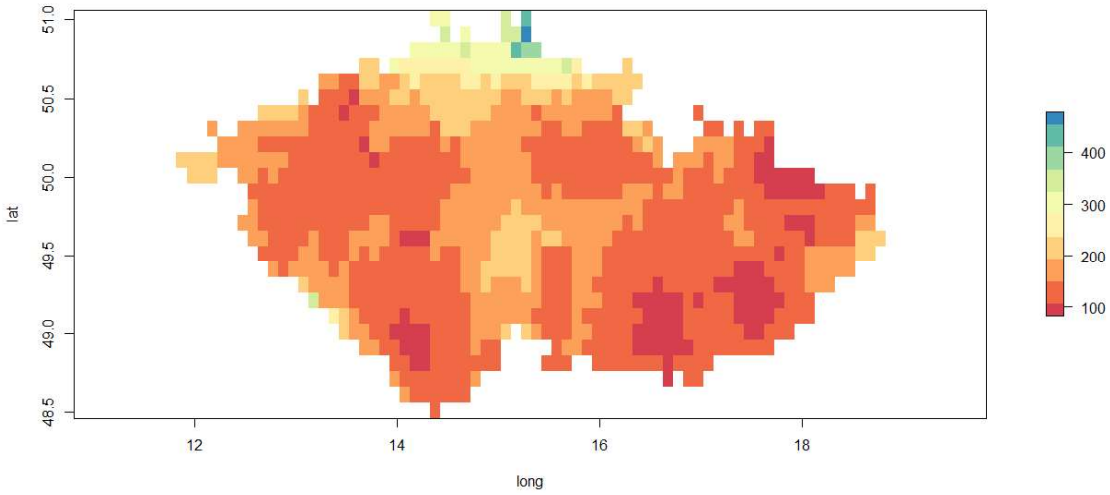
V R existuje celá řada možností pro analýzu prostorových dat a nástrojů k jejich vykreslení. Důležitý rozdíl mezi R a tradičními desktopovými GIS (*geographical information system*) softwary je ten, že GIS je primárně vytvořen k prostorové vizualizaci a zpracovává data jedním přednastaveným způsobem, zatímco v R si uživatel sám musí zvolit vhodné nástroje a odpovídající nastavení. Dále R neobsahuje grafické uživatelské rozhraní (GUI), které by usnadňovalo práci s prostorovými daty. Z tohoto důvodu může být R jako nástroj GIS pro nové uživatele náročné. Hlavní výhodou R je, že přináší do tvorby prostorové vizualizace silné výpočetní prostředky pro úpravu a statistickou analýzu dat a možnost uložení skriptů a jejich další modifikace (Cheshire a Lovelace 2015).

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 (atributovou tabulkou). Dále se také používají prostorová pole, která jsou obvykle reprezentována pomocí *rastrů*. Rastrová data se obvykle používají pro prostorově kontinuální proměnné. Rastry dělí území mřížkou na buňky o stejné velikosti neboli na *pixely*, kterým je přiřazená hodnota dle zájmové proměnné. Většinou se jedná o průměr či většinovou hodnotu pro oblast spadající do konkretního pixelu. Na rozdíl od vektorových dat, v rastrových datech není informace o geometrii objektu uložená jako souřadnice, ale v rámci buněk. Velikost rastrových buněk neboli prostorové rozlišení je dáno počtem řádků a sloupců, na které je území rozdělené. (Hijmans [vid. 13.4.2018])



Obrázek 26: Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `ggplot2`

Vektorová data mohou být vykreslena v R pomocí celé řady balíčků a to jak pomocí již zmíněných `base` (kapitola 4.1) a `ggplot2` (kapitola 4.1.3), tak i pomocí `Leaflet` (kapitola 4.3.3). Příkladem vizualizace vektorových dat pomocí balíčku `ggplot2` je vizualizace na obrázku 26.



Obrázek 27: Úhrn srážek (v mm) na povodích ČR za srpen 2010, vykresleno pomocí balíčků `raster`

Balíček `raster` obsahuje funkce pro vytváření, čtení, manipulaci, modelování, popis a analýzu rastrových dat. Podporuje práci s velkými soubory a vytváření

vlastních specifických funkcí (Hijmans 2017). 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 (Lamigueiro 2018). Vykreslení dat pomocí balíčku `raster` lze nalézt na obrázku 27.

4.3 Balíčky pro interaktivní vizualizaci dat

Možnost přiblížit, filtrovat či zobrazit podrobnosti grafu na požádání je výhodou dynamických a interaktivních grafů. Interaktivní grafy dokážou 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[1] možností je balíček `htmlwidgets`, jehož knihovna obsahuje užitečné nástroje pro tvorbu téměř jakéhokoliv typu grafiky.

Balíček `htmlwidgets` poskytuje rozhraní pro snadné propojení jazyka R a knihoven programovacího jazyka JavaScript, používaného zejména pro webové aplikace. Tímto je umožněna bezproblémová 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`. Tyto nástroje jsou dostupné nejenom pro R, ale i pro další programovací jazyky (například Python) (Vaidyanathan et al. 2018).

4.3.1 `Plotly`

Balíček `Plotly` je vysokoúrovňové rozhraní 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 dvěma 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 (Sievert 2018).

4.3.2 `dygraphs`

Balíček `dygraphs` je R rozhraním 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 (RStudio [vid. 11.4.2018]):

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

- Automatické vykreslení časových řad typu `xts`[2]
- Vysoce konfigurovatelná nastavení os a vykreslení řad (včetně volitelné další osy y)
- Zvýraznění, přiblížení řad či bodů
- Zobrazení predikčních intervalů kolem řad
- Umožňuje vykreslování překrývajících 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 ovládaní `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 stávající hranice mapového výřezu či reakce na kliknutí myši uživatelem. Umožňuje zobrazení map v nesférickým Mercatorově zobrazení a má mnoho dalších funkcí a doplňků (RStudio [vid. 11.4.2018]).

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 zobrazení 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 řádků 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

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

stránek. Pomocí `Shiny` lze vytvořit jak samostatnou aplikaci, běžící na webových stránkách, tak i lokální aplikaci běžící v prostředí R. Vzhled `Shiny` aplikace lze modifikovat pomocí kaskádových stylů[3]. (RStudio [vid. 14.4.2018])

Strukturou se `Shiny` aplikace liší od běžných skriptů. Celá aplikace by měla být umístěna v jednom adresáři a celý skript by měl být obsažen v souboru nazvaném `app.R`.[4] 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)
```

V rámci `ui` lze používat širokou řadu ovládacích 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.
- Postranní panel, který nejčastěji obsahuje ovládací prvky.
- Hlavní panel, který obsahuje výstupy generované funkcí `server`.

Příkladem ovládacího prvku vytvořeného v `ui` může být číselný vstup, který umožňuje uživateli vložit libovolné číslo. Tento vstup je označen pomocí `InputId` a v rámci `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 rámci funkce `render*()`, případně v rámci 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:

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

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

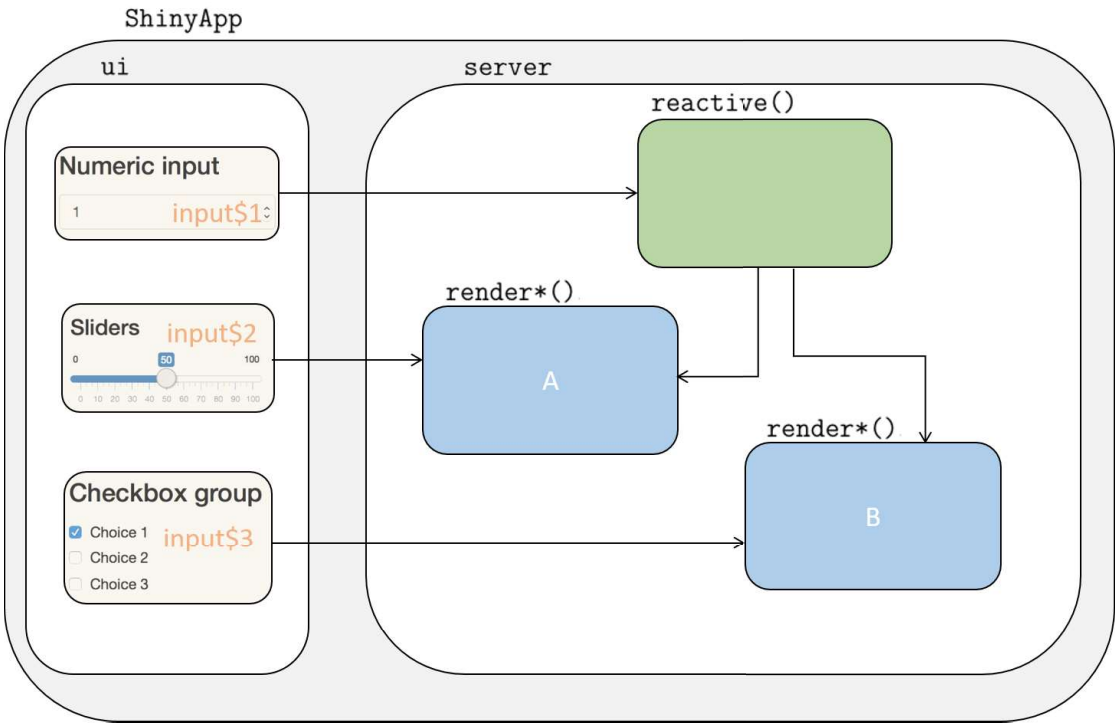
```
output$histPlot <- renderPlot({
  hist(rnorm(input$num, 1, 0))
  })
```

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 vraceli požadovaný typ výstupu. Základní `render*()` funkce jsou vypsány v tabulce 5. 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 5: 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 28 je znázorněn příklad struktury jednoduché `Shiny` aplikace s použitím `reactive()` funkce. V tomto příkladu při změně numerického vstupu (`input$1`) se nejprve přepočítá `reactive()` a následně `render*() A` a `B`, zatímco při změně vstupu `input$2` či `input$3` se přepočítají pouze příslušné `render*()` výstupy `A` či `B` se zachováním informací z `input$1`.



Obrázek 28: Ukázka `Shiny` aplikace s použitím funkce `reactive()`.

Praktická část

Součástí práce je vytvoření webové aplikace pro vizualizaci a analýzu hydrologické bilance a předpověď sucha v útvarech povrchových vod ČR. Aplikace je popsána v části „Metodika“, jejíž součásti je popis použitých dat a balíčků, technického řešení a postprocessingových funkci. Dále následuje popis jednotlivých položek aplikace se zaměřením na účel a funkčnost každé položky.

5 Metodika

Aplikace je vytvořena prostřednictvím programovacího jazyka R. Jedná se o vizualizaci výsledků modelování systému pro předpověď hydrologické situace ***HAMR***. Systém je založen na propojení modelu vláhové bilance půdy `SoilClim`, modelu hydrologické bilance `BILAN` a modelu vodohospodářské soustavy `WATERES` jednotlivých povodí. Jádro aplikace je postaveno na balíčcích `Shiny` a `flexdashboard`. Jak již bylo popsané v kapitole 4.4, `Shiny` umožňuje jednoduché vytváření webových aplikací, interaktivních vizualizací v prostředí R a balíček `flexdashboard` umožňuje pokročilé formátovaní vzhledu. Společně slouží k publikaci dat a jejích přehledné vizualizaci v rámci webového prohlížeče.

5.1 Technické řešení

Aplikace bude přístupná na serveru fakulty[5], lze ji také stáhnout ze stránek GitHubu[6], kde je pro aplikaci založen repozitář. Tento repozitář 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:

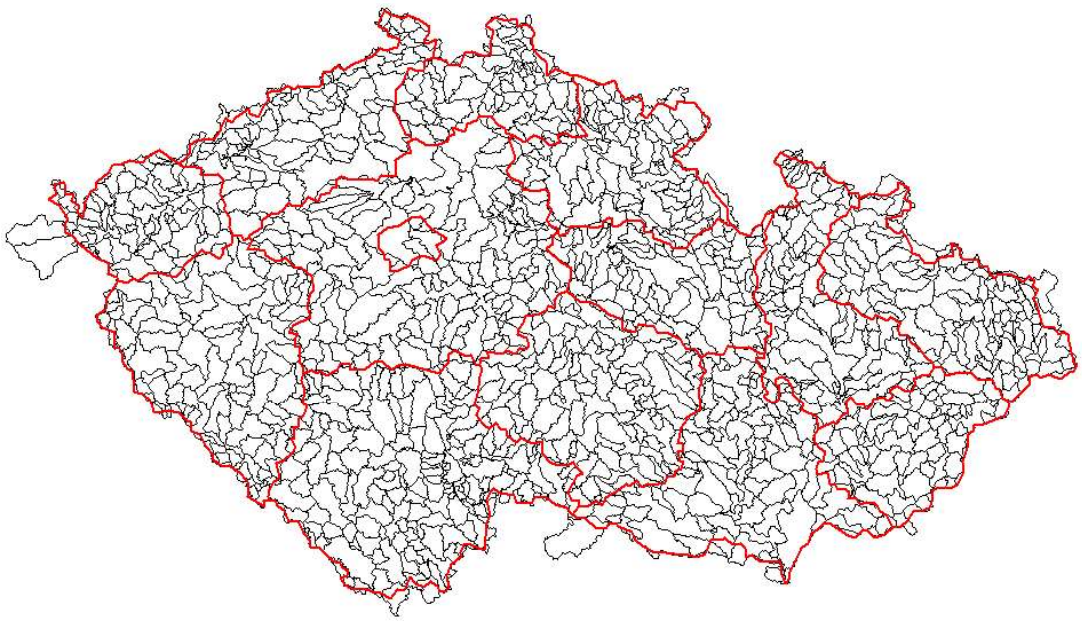
- `leaflet`, umožňující vizualizaci prostorových dat v interaktivních mapách (kapitola 4.3.3)
- `ggplot2`, `dygraphs` a `Plotly` k vykreslení časových řad a čar překročení (kapitoly 4.1.3, 4.3.2 a 4.3.1)
- `data.table`, `dplyr` sloužící k transformaci dat
- `DT` je jedním z *widgets* balíčku `htmlwidgets` (popsáno v kapitole 4.3), sloužící k vytváření interaktivních tabulek.

Veškerý R kód použitý k vytvoření aplikace je uveden v příloze na stráně 59.

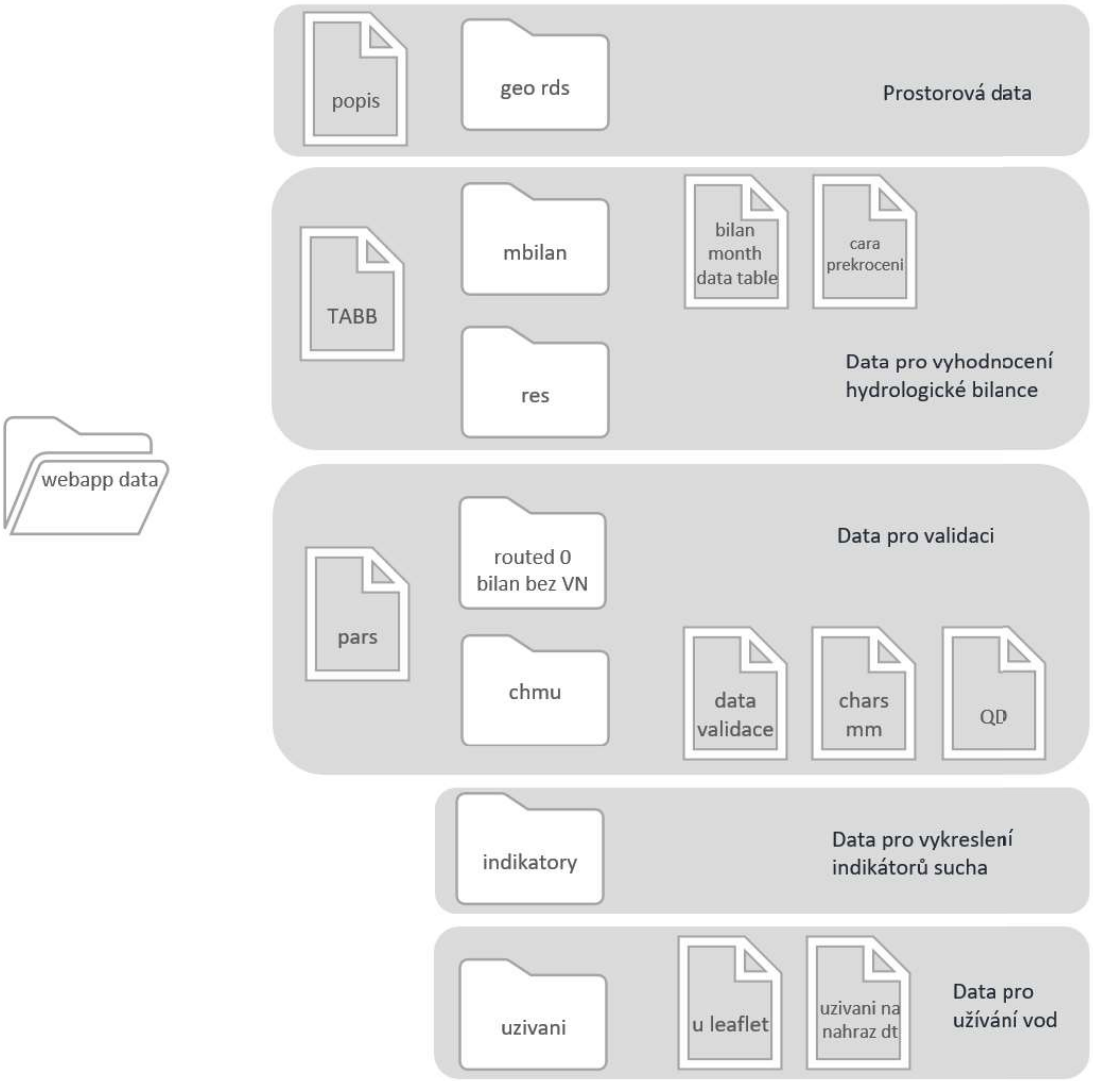
[5]Aplikace na serveru fakulty: https://shiny.fzp.czu.cz/KVHEM/HAMR/
[6]Repozitář na GitHub: https://github.com/KVHEM/Sucho

5.2 Data

Data použitá k vykreslení lze rozdělit do několika skupin: prostorová data, data potřebná pro hydrologickou bilanci, data pro indikátory sucha, data pro užívání vod a data pro validaci. Skupina prostorových dat obsahuje soubory s informacemi o útvarech povrchových vod (UPOV) ČR (`povodi.rds`, `reky.rds`, `jezera.rds`, `nadrze.rds`) a soubory administrativního členění ČR (`kraje.rds`, `okresy.rds` a povodí 3. řadu `povodi_III.rds`) viz obrázek 29. Přesná struktura složek aplikace je vyobrazená na diagramu 30. Tyto soubory byli v rámci přípravy dat převedeny z formátu .shp do formátu .rds. Dále pro rychlejší načítání a jednodušší manipulaci byly tyto soubory transformovány do souřadnicového systému WGS84 (pomocí funkce `spTransform()` z balíčku `sp` a atributu `CRS` - *Coordinate Reference System* - nastaveného na identifikátor pro WGS84 - EPSG:4326) a zjednodušeny (pomocí funkce `ms_simplify()` z balíčku `rmapshaper`) pro snížení náročnosti při vykreslování. Soubor `popis.rds` obsahuje informace o jednotlivých útvarech jako jsou název útvaru, název povodí, název oblasti, kategorie útvaru a typ útvaru. Následně soubor `popis.rds` je spojen pomocí funkce `merge()` z balíčku `sp` se souborem `povodi.rds` přes identy jednotlivých útvaru (`UPOV_ID`). K vykreslení horního povodí se používá soubor `TABB.rds` obsahující informace o tom, která povodí přitékají do jednotlivých UPOVů.



Obrázek 29: Útvary povrchových vod (UPOV) (vyznačené černě) a kraje (vyznačené červeně) ČR



Obrázek 30: Struktura složek aplikace

Další skupinou jsou data potřebná pro vyhodnocení hydrologické bilance. Pomocí modelu Bilan (Vizina et al. 2015) bylo v rámci projektu kalibrováno 1112 UPOVů a výstupem je 18 proměnných pro každý UPOV v denním kroku pro období 1981-2010. Tyto proměnné jsou uvedeny v tabulce 6 (VÚV TGM 2015). Z důvodů šetření vnitřní paměti aplikace se soubory s daty pro hydrologickou bilanci nacházejí na uložišti ve složce „res“ a jsou uloženy ve formátu .rds s názvy odpovídající identům jednotlivých povrchových útvarů (`UPOV_ID`). Aplikace tyto soubory načítá pouze při požadavku uživatele. Měsíční bilance je agregací denních dat a načítá se ze souboru `bilan_month_data_table.rds` (ukládá se do proměnné `BM`). Do teto skupiny lze zařadit i soubor `cara_prekroceni_dt.rds` (v aplikaci proměnná `cp`). Soubor obsahuje předpočítané roční, měsíční a sezónní pravděpodobnosti spočítané přes jednotlivé proměnné dle vzorce $p = (m - 0.3)/(n + 0.4)$, kde po seřazení souboru dle

velikosti v klesajícím pořadí je n počet prvků a m je pořadové číslo.

zkratka	význam	zkratka	význam
P	srážky na povodí	SW	půdní vlhkost (zásoba vody v nenasycené zóně)
R	odtok (pozorovaný)	SS	zásoba vody ve sněhu
RM	celkový odtok (simulovaný)	GS	zásoba podzemní vody
BF	základní odtok (simulovaný)	INF	infiltrace do půdy
B	základní odtok (odvozený)	PERC	perkolace z půdní vrstvy
DS	zásoba pro přímý odtok	RC	dotace zásoby podzemní vody
DR	přímý odtok	T	teplota vzduchu
PET	potenciální evapotranspirace	H	vlhkost vzduchu
ET	územní výpar	WEI	váhy pro kalibraci odtoku

Tabulka 6: Výstupy kalibrace denního modelu Bilan

Data pro vykreslení indikátorů sucha se nacházejí ve samostatné složce „indikatory“. Při zvolení uživatelem indikátoru sucha (v nabídce momentálně jsou pouze indikátory SPI a SPEI spočítány klouzavě s krokem 1, 3, 6, 9 a 12 měsíců) se načte .rds soubor dle odpovídajícího indikátoru a kroku.

Data užívání vod za období 2006-2016 (`uzivani_na_nahraz_dt.rds` ve složce „uzivani“) pocházejí z evidence užívání vod, kterou spravuje VÚV T. G. M. a v rámci aplikace se ukládají do proměnné `u`. Data obsahují informace o poloze odběru ve formě souřadnic (`X` a `Y`), identifikačním čísle odběru (`ICOC`), názvu místa odběru (`NAZICO`) a také o jevu (odběry z podzemních vod `POD`, odběry z povrchových vod `POV` či vypouštění `VYP`). V rámci přípravy dat byly `ICOC`ům, které obsahovali pozorované údaje, ale měli chybějící souřadnice, přiřazeny průměry souřadnic mezi odběry se stejným `ICOC` a jiným jevem. Tyto data slouží k vykreslení časových řad a tabulek. Pro vykreslení bodů do mapy bylo nutné vytvořit soubor (`u_leaflet.rds` ve stejné složce), který obsahuje pouze jeden záznam pro každý `ICOC`. Souřadnice těchto `ICOC`ů byly transformovány do souřadnicového systému WGS84 a následně byly uloženy ve formátu `SpatialPointsDataFrame`.

Kalibrace modelu Bilan proběhla s nastavením modelu na denní časový krok při použití šesti volných parametrů (*Spa*, *Alf*, *Dgm*, *Soc*, *Mec*, *Grd*). Parametry jsou popsány v tabulce 7 (VÚV TGM 2015). Soubor s parametry `pars.rds` se nahrává v aplikaci do stejnojmenné proměnné a obsahuje počáteční hodnoty parametrů (`initial`), jejích dolní a horní meze (`lower`, `upper`) a stávající hodnotu (`current`). Tyto proměnné jsou dány pro každý UPOV. K stanovení hodnot parametrů byl použit globální optimalizační algoritmus diferenciální evoluce (VÚV TGM 2015).

Model byl kalibrován na hydrologické charakteristiky (m-denní průtoky a dlouhodobý průměrný průtok) povodí UPOV poskytnuté ČHMÚ. M-denní průtoky vypočítané na základě pozorovaných hodnot lze načíst ze souboru `chars_mm.rds` a m-denní průtoky pro simulována data ze souboru `chars_mm_sim.rds`.

K validaci denních průtoků (soubor `QD.rds` ze složky „chmu“) bylo využito 156 měrných stanic. Po propojení databankového čísla `DBCN` s povodím `UPOV_ID` zbylo 153 stanic. Mimo `DBCN` a `UPOV_ID` soubor obsahuje hodnoty pozorovaných denních průtoků (`value`) za období 1980-2010 (`DTM`). Simulované průtoky se nacházejí ve složce „routed-0_bilan_bez_VN“ a obdobně jako u denních dat hydrologické bilance obdrží se při zvolení konkrétní stanice (dle patřičného `UPOV_ID` aplikace načte odpovídající .rds soubor). Pro prostorový vykreslení stanic se používá soubor `QD_stanice` (složka „geo_rds“, soubor `stanice.rds`). Původní shapefile byl převeden do souřadnicového systému WGS84 a uložen ve formátu .rds. Soubor obsahuje nejen prostorové informace, ale i informace o názvech toku, ploše povodí atd.

Validace měsíčních průtoků využívá záznamy 542 měrných stanic z období 1982-2010. Prostorová data jsou uložená do proměnné `stanice` (soubor `E04_Vodomerne_stanice.rds` ze složky „geo_rds“). Data s naměřenými (`QMER`) a simulovanými (`QNEX`, `QNEY`) průtoky lze načíst ze souboru `data_validace.rds` ze složky „chmu“. Tento soubor byl vytvořen agregací denních dat ze složky „routed-0_bilan_bez_VN“.

název	`Alf`	`Dgm`	`Grd`
význam	parametr určující odtok ze zásoby pro přímý odtok	koeficient mezi teplotou a táním sněhu	parametr určující odtok ze zásoby podzemní vody
název	`Mec`	`Soc`	`Spa`
význam	parametr rozdělující perkolaci na přímý odtok a na dotaci podzemní vody pro podmínky tání sněhu	koeficient mezi teplotou a táním parametr rozdělující perkolaci na přímý odtok a na dotaci podzemní vody pro letní podmínky	kapacita zásoby půdní vlhkosti

Tabulka 7: Parametry denního modelu Bilan

5.3 Postprocessing

Pro podporu projektu Voda-Sucho byl vytvořen balíček `CatCa`[7], který obsahuje některé důležité funkce pro práci s útvary povrchových vod, vypočet bilanci, m-denních vod atd. Nainstalovat balíček lze pomocí příkazu `devtools::install_github(„KVHEM/CatCa")`. V rámci tvorby aplikace byly do tohoto balíčků přidány také funkce pro přípravu dat. Tyto funkce upravují vstupní data do potřebného formátu a vybírají pouze potřebné proměnné pro snížení využívané paměti a výpočetní náročností. Přehled funkcí k přípravě dat je v tabulce 8. Dále balíček obsahuje funkci „getpath“ pro nastavení pracovních cest k uložišti dat.

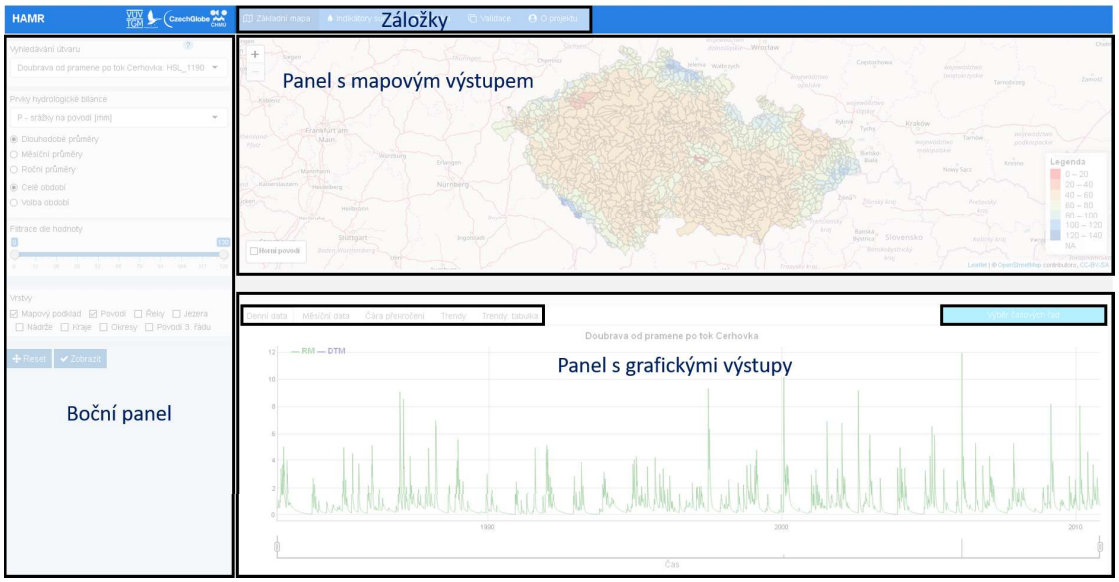
[7]Repozitář na GitHub: https://github.com/KVHEM/CatCa.

Pokud cesta není nastavena pomocí funkce „getpath“ je nutné jí uložit manuálně do proměnné `.datadir`.

`prep_spatial_data()`	příprava prostorových dat pro aplikaci
`prep_bilan_month()`	příprava měsíční bilance pro aplikaci
`prep_QD()`	příprava denních průtoku (validace) pro aplikaci
`prep_uzivani()`	příprava užívaní pro aplikaci
`prep_uzivani_upovid()`	příprava připojeni `UPOV_ID` k užívaní pro aplikaci

Tabulka 8: Přehled funkcí z balíčku `CatCa` pro přípravu dat

6 Základní rozvržení

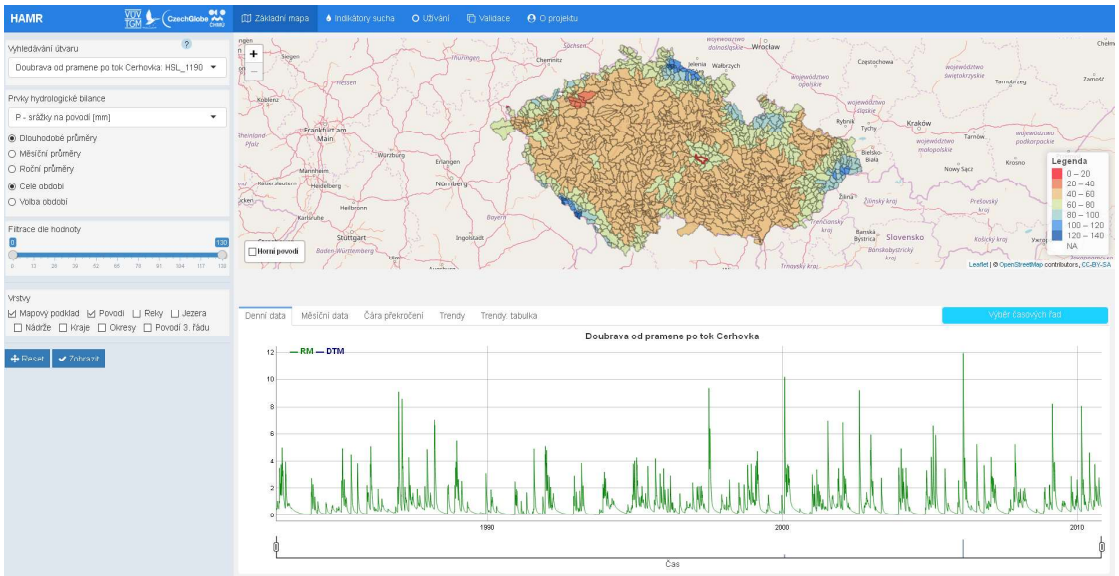


Obrázek 31: Základní rozlozeni okna aplikace

Nejdříve by bylo vhodné uvést některé pojmy, které budou používány pro popis základního rozvržení aplikace. Okno aplikace je většinou rozděleno do několika hlavních části: boční panel, panel s mapovým výstupem a panel s grafickými výstupy. V bočním panelu se nastavují vstupy pro vykreslení mapy, požadované vrstvy a lze také použit pole pro vyhledávaní útvaru. Pole „Výhledávání útvaru“ nabízí seznam všech vykreslených útvarů, ale lze ho také použít k ručnímu vyhledávání; stačí vymazat momentálně zobrazený název a napsat název toku či jeho `UPOV_ID`. Pole „Vrstvy“ v bočním panelu je k dispozici pro každý panel s mapovým výstupem aplikace a umožňuje výběr následujících vrstev: povodí, řeky, jezera, nádrží, mapový podklad, povodí 3. řádu a administrativní členění České republiky (kraje a okresy). Boční panel některých záložek také obsahuje tlačítka „Reset“ a „Zobrazit“. Tlačítko „Reset“ vrací mapový panel na počátečně nastavené souřadnice. Tlačítko „Zobrazit“ slouží k

vykreslení po novem nastavení vstupů. Horní lišta aplikace obsahuje přepínač mezi jednotlivými záložkami aplikace: „Základní mapa“, „Indikátory sucha“, „Užívání“, „Validace“ a „O projektu“. Grafy zpravidla mají vlastní výběr proměnných, který se na obrázku 31 nachází v pravém horním rohu grafického panelu. Tento výběr má tvar **srolovácího ololo tlačítka**. Dále grafický panel může obsahovat vlastní lištu se záložky pro přepínání mezi jednotlivými typy grafů a tabulek.

6.1 Základní mapa



Obrázek 32: Okno „základní mapa“

„Základní mapa“ (obrázek 32) je první záložkou a zobrazí se ihned po spuštění aplikace. Obsahuje informace o hydrologické bilanci povodí České republiky. Záložka základní mapa je rozložena na boční panel, panel s mapovým výstupem a panel s grafickým výstupem.

V bočním panelu se nacházejí pole „Vyhledávání útvaru“, „Prvky hydrologické bilance“, „Filtrace dle hodnoty“, „Vrstvy“ a tlačítka „Reset“ a „Zobrazit“. Uživatel volí proměnnou hydrologické bilance, dle níž budou zbarveny jednotlivá povodí zobrazená na mapě. Hodnoty proměnné jsou agregovány do měsíčních a ročních kroků, lze je také vykreslit jako dlouhodobé průměry, tzn. průměry za celé období nebo za konkrétní periody po 29 letech: 1961-1990, 1971-2000 a 1981-2010. „Filtrace dle hodnoty“ v počátečním stavu obsahuje všechny hodnoty zvolené proměnné a dále umožnuje nastavení rozsahu hodnot, který omezí vykreslená povodí. „Vyhledávání útvaru“ je jedinou částí bočního panelu, která je propojená nejenom s mapou, ale i s grafickými výstupy, a to pomocí funkce `renderUI()` z balíčků `Shiny` (kapitola 4.4.2), která umožňuje, aby element uživatelského rozhrání `ui` obsahoval vstup `input$map_shape_click$id`. Tento vstup je získán po kliknutí na mapový objekt,

vytvořený pomocí `Leaflet` (kapitola 4.3.3) a vrací `UPOV_ID` příslušný tomuto mapovému objektu. Objekt `search.choices` obsahuje seznam všech `UPOV_ID` s přiřazenými názvy toků.

```
renderUI({selectizeInput(inputId = "search.id",
                         label = "Vyhledávání útvaru",
                         selected = input$map_shape_click$id,
                         choices = search.choices)})
```

Mapové objekty jsou vykresleny pomocí `Leaflet`. V rámci objektu `leaflet()` lze použit tyto funkce k nastavení: `setView()` nastaví počáteční souřadnice a přiblížení, `addTiles()` vykreslí mapový podklad. Aplikace používá přednastavený podklad OpenStreetMap (RStudio [vid. 11.4.2018]). Objekty jsou vykresleny pomocí funkce `add*()`, která v názvu obsahuje typ objektu, například `addPolylines()`, `addPolygons()` atd. Sktruktura objektu `leaflet()` pak může vypadat například následovně

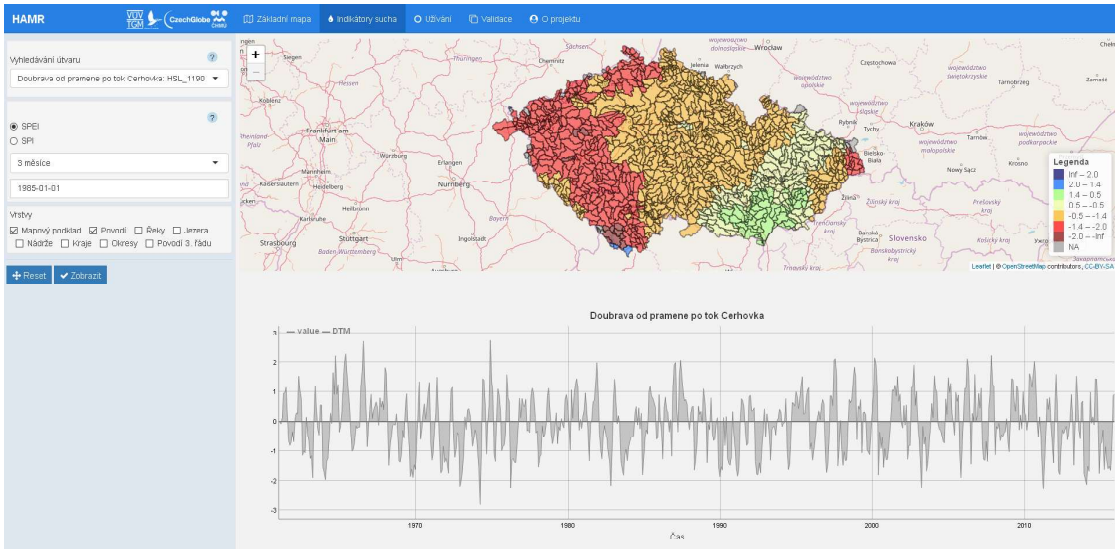
```
map <- leaflet() %>%
          setView(...) %>%
          addTiles(...) %>%
          addPolygons(...)
```

Při změně vstupu v bočním panelů je nutné znovu překreslit mapový objekt. Tento přístup může být pro jiné interaktivní funkce příliš výpočetně náročný. Například vrstva „Horní povodí“, která zobrazuje povodí přitékající do momentálně vybraného povodí, je z toho důvodu vykreslována pomocí `leafletProxy()`. Tento příkaz umožňuje okamžité změny, na již vykresleném mapovém objektu. Konkrétně kód pro aktualizaci horního povodí je uveden níže.

```
observe({
  leafletProxy("map") %>%
    clearGroup("Horní povodí") %>%
    addPolylines(data=horni.povodi(),
                 color = "#00264d", weight = 2.5,
                 opacity = 0.7, stroke = TRUE,
                 group = "Horní povodí")
})
```

Pro zvolené povodí se vypočítají časové řady z měsíčních a denních dat (pomocí balíčku `dygraphs`), čára překročení pro celé období, roční období a měsíce (pomocí balíčku `Plotly`) a trendy: grafické znázornění a tabulka s vyhodnocením statistické významnosti (`ggplot2`). Zvolené povodí se zvýrazní v mapě červeným okrajem pomocí `leafletProxy()`. Kliknutím na jiné povodí se přepočítají grafické výstupy a název nově zvoleného povodí s jeho `UPOV_ID` se promítne do pole „Vyhledávání útvaru“.

6.2 Indikátory sucha



Obrázek 33: Okno „indikatory“

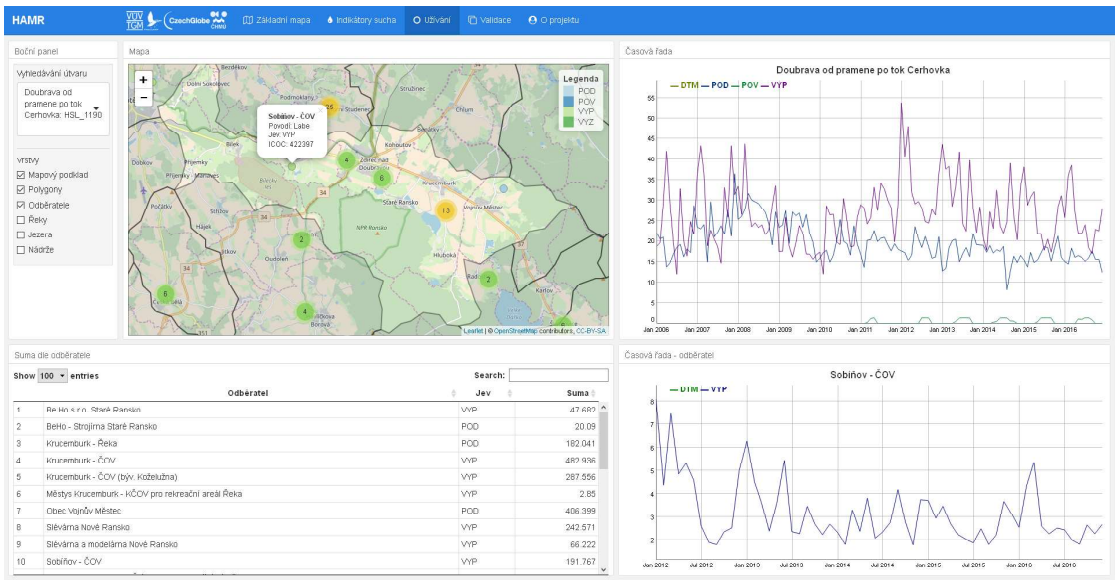
Záložka „Indikátory sucha“ (obrázek 33) se skládá z bočního panelu, mapového panelu a grafického panelu. V bočním panelu se obdobně jako v záložce „Základní mapa“ nachází pole „Vyhledávání útvaru“ a „Vrstvy“. Dále v bočním panelu lze si zvolit indikátor, krok, do kterého budou data agregována a datum. Protože data mají měsíční časový měřítko, volba dne v kalendáře nehraje pro vykreslení žádnou roli, pouze měsíc a rok. Momentálně v mapě jsou zobrazeny indikátory SPI (*Standardized Precipitation Index*) s SPEI (*Standardized Precipitation Evapotranspiration Index*) počítány klouzavě s krokem 1, 3, 6, 9 a 12 měsíců. Časem by do aplikace mohli být přidány indikátory PDSI (*Palmer Drought Severity Index*), SGI (*Standardized Groundwater Index*), SRI (*Standardized Runoff Index*) a nedostatkové objemy ve stejném kroku (Vlnas et al. 2014). Povodí se rozdělují do 7 kategorií, aby se více projevila variabilita:

∞ až $2,0$	?
$2,0$ až $1,4$	?
$1,4$ až $0,5$	?
$0,5$ až $-0,5$	bez výskytu sucha
$-0,5$ až $-1,4$	slabé sucho
$-1,4$ až $-2,0$	silné sucho
$-2,0$ až $-\infty$	mimořádné sucho

Tabulka 9: Rozdělení indikátorů sucha do kategorií.

Mapové objekty jsou vykresleny pomocí `Leaflet`. V grafický panelu se vykresluje časová řada indikátoru pro zvolené povodí pomocí balíčku `dygraphs`.

6.3 Užívání



Obrázek 34: Okno „uzivani“

Záložka „Užívání“ (obrázek 34) obsahuje informace o užívání vody v ČR a je rozdělena do pěti části: boční panel, panel s mapovým výstupem, dva panely s grafickými výstupy a jeden panel s tabulkovým výstupem. Boční panel obsahuje pole „Vyhledávání útvaru“ a „Vrstvy“. Pole „Vrstvy“ je rozšířeno o vrstvu „Odběratele“, avšak postrádá administrativní členění České republiky. Mapový panel spojuje místa odběru do shluků pomoci atributu `clusterOptions = markerClusterOptions()` v ramci funkce `addCircleMarkers()`. Po přiblížení lze na bod kliknout. Po kliknutí se zobrazí popisek s informací o odběrateli a časová řada odběrů. Kliknutím na povodí se obdrží informace o všech odběratelích v tabulkovém panelu a časová řada pro jednotlivé jevy v grafickém panelu (odběry z podzemních vod `POD`, odběry z povrchových vod `POV` či vypouštění `VYP`). Grafické panely jsou vytvořený pomocí `dygraphs`. Tabulka je vytvořená pomocí balíčku `DT` a je interaktivní.

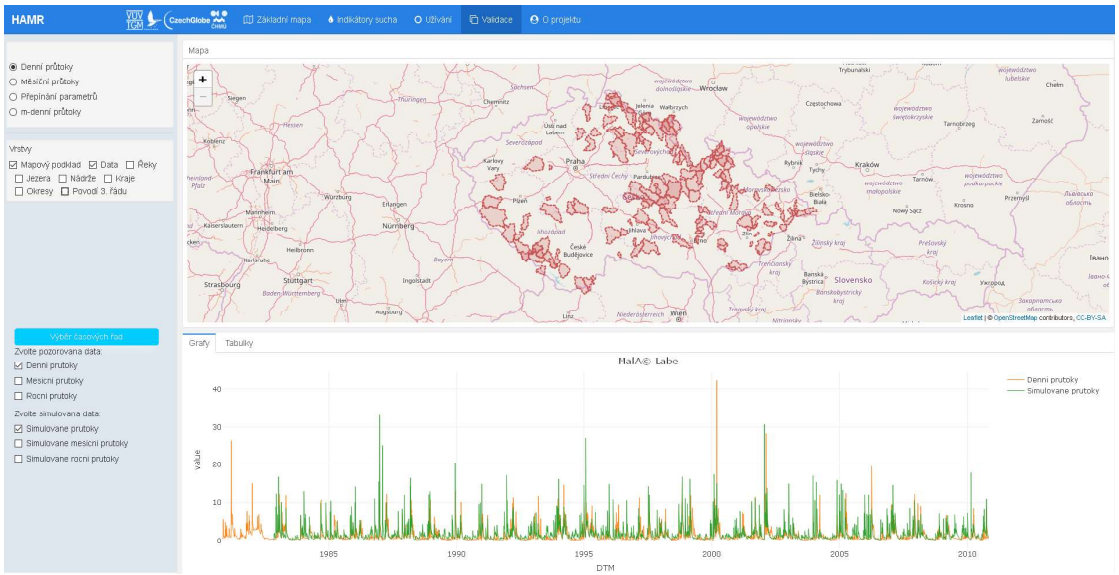
6.4 Validace

Záložka „Validace“ (obrázek 35) se dělí na tří částí: boční panel, panel s mapovým výstupem a panel s grafickými výstupem. Boční panel obsahuje standardní pole „Vrstvy“ a přepínač mezi následujícími možnostmi: denní průtoky, měsíční průtoky, přepínání parametrů a m-denní průtoky. Mapový výstup denních průtoků obsahuje polohu 153 měrných stanic (viz kapitola 5.2) a grafický panel po zvolení konkrétní měrný stanice vytvoří časovou řadou pozorovaných a simulovaných průtoků, které lze vykreslit v denním, měsíčním a ročním kroku pomocí pole „Výběr časových řad“ (nachází se v oblastí bočního panelu). Záložku grafického panelu lze přepnout z

grafů na tabulky, které nejsou interaktivní. Tabulka denních průtoků obsahuje pouze základní přehled o datech (počet pozorování, střední hodnotu atd.).

Mapový výstup měsíčních průtoků obsahuje pozice 542 měrných stanic. Body měrných stanic jsou propojeny s informacemi o UPOVu, do kterého spadají. Kliknutím na bod se objeví popisek stanice a vykreslí se horní povodí, obdobně jako v záložce „Základní mapa“. Graf obsahuje časové řady pozorovaných a simulovaných průtoků a tabulka obsahuje číselné vyhodnocení přesností simulovaných dat vůči pozorovaným (pomocí funkce `gof()` z balíčku `hydroGOR`[8]).

Při změně na „Přepínání parametrů“ v bočním panelu se objeví pole s nabídkou parametrů (*Spa*, *Alf*, *Dgm*, *Soc*, *Mec*, *Grd*). Momentálně „Přepínání parametrů“ obsahuje data pouze v panelu s mapovým výstupem pro vizualizaci plošného rozložení parametrů. UPOVy jsou zbarveny dle stávající hodnoty parametru (`current`). Obdobně jako u „Přepínání parametrů“ objeví se v bočním panelu při změně na „m-denní průtoky“ nabídka m-denních vod ($Q30d$, $Q60d$, $Q90d$, $Q120d$, $Q150d$, $Q180d$, $Q210d$, $Q240d$, $Q270d$, $Q300d$, $Q330d$, $Q355d$, $Q364d$). Také se objeví pole „Vyhledávání útvaru“, které propojí mapový a grafický panel. Útvary mapového výstupu se zbarvují dle hodnoty proměnné, kterou zvolí uživatel. Grafickým výstupem je `Plotly` objekt, který obsahuje seřazený za sebou hodnoty pozorovaných a simulovaných m-denních průtoků pro zvolený povodí. Tabulka obsahuje též hodnoty.



Obrázek 35: Okno „validace“

[8]Dokumentace balíčku je dostupná na adrese https://cran.r-project.org/web/packages/hydroGOF/hydroGOF.pdf

6.5 O projektu

Založka „O projektu“ obsahuje krátký popis projektu a veškeré kontakty. V budoucnu tato záložka bude rozšířena o metodiky k jednotlivým součástem systému. Systém je založen na propojení modelu půdního (SoilClim), modelu hydrologické (Bilan) a vodohospodářské bilance. Tyto modely spolu se vstupními klimatologickými daty reprezentují sucho meteorologické, zemědělské a hydrologické. Na vzniku systému se podílí:

Výzkumný ústav vodohospodářský T. G. Masaryka, v.v.i.

Podbabská 30, 160 00 Praha 6

Ing. Adam Vizina, Ph.D.

Česká zemědělská univerzita v Praze, Fakulta životního prostředí

Kamýcká 129, 165 21 Praha 6

doc. Ing. Martin Hanel, Ph.D.

Ústav výzkumu globální změny AV ČR, v.v.i.

Bělidla 986/4a, 603 00 Brno

prof. Ing. Mgr. Miroslav Trnka, Ph.D.

Český hydrometeorologický ústav

Na Šabatce 2050/17, 143 06 Praha 412

RNDr. Jan Daňhelka, Ph.D.

Systém pro předpověď sucha vznikl v rámci projektu financovaném Ministerstvem životního prostředí v roce 2017. Projekt spadá do činnosti k podpoře výkonu státní správy v problematice SUCHO v roce 2017 (úkol 3702).

Výsledky

Diskuse

Závěr

Literatura

ABDI, Herve a LJ WILLIAMS, 2010. Normalizing data. Encyclopedia of research design. *Thousand Oaks, CA: Sage.*

BECKER, Richard A, William S CLEVELAND, Ming-Jen SHYU, Stephen P KALUZNY a OTHERS, 1996. A tour of Trellis graphics. *Murray Hill, NJ: AT & T Bell Laboratories.*

BRINTON, Willard Cope, 1919. *Graphic Methods for Presenting Facts.* B.m.: The Engineering Magazine Company, New York. ISBN 978-1155058870.

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

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

CHESHIRE, James a Robin LOVELACE, 2015. Spatial data visualisation with R. *Geocomputation: A Practical Primer.* B.m.: Sage.

CLEVELAND, William S., 1993. *Visualizing Data.* B.m.: Hobart Press. ISBN 978-0963488404.

CLEVELAND, William S., 1994. *The Elements of Graphing Data.* B.m.: Hobart Press. ISBN 0963488414.

CLEVELAND, William S. a Robert MCGILL, 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association.* B.m.: Taylor & Francis.

DE MAESSCHALCK, Roy, Delphine JOUAN-RIMBAUD a Désiré L MASSART, 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems.* B.m.: Elsevier.

DPLYR, [vid. 17.4.2018]. *A grammar of data manipulation.* [online]. Dostupné z: https://github.com/tidyverse/dplyr

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

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

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

HEBÁK, P, J HUSTOPECKÝ, E JAROŠOVÁ a I PECÁKOVÁ, 2007. *Vícerozměrné*

statistické metody (1). B.m.: Informatorium, Praha. ISBN 80-7333-025-3.

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

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

KUTNER, Michael H, Christopher J. NACHTSHEIM, John NETER a William LI, 2004. *Applied Linear Statistical Models*. B.m.: McGraw-Hill/Irwin. ISBN 0-07-238688-6.

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

MACIEJEWSKI, Ross, 2011. *Data Representations, Transformations, and Statistics for Visual Reasoning*. B.m.: Morgan & Claypool Publishers. ISBN 978-1-608-45625-3.

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

MCINTOSH, Avery, 2016. The Jackknife Estimation Method. *arXiv*.

MINARD, Charles Joseph, 1858. *Carte figurative et approximative représentant pour l'année 1858 les émigrants du globe, les pays dóu ils partent et ceux oú ils arrivent* [online]. Dostupné z: https://www.loc.gov/resource/g3201e.ct000242/

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

MURRELL, Paul, 2003. *grid: The grid graphics package*. červenec 2003.

MURRELL, Paul R., 1998. *Investigations in Graphical Statistics*. B.m. PhD thesis. The University of Auckland.

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

ÖZTUNA, Derya, Atilla Halil ELHAN a Ersöz TÜCCAR, 2006. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*. B.m.: The Scientific; Technological Research Council of Turkey.

PALSKY, Gilles, 1996. *Des chiffres et des cartes: naissance et développement de la cartographie quantitative française au XIXe siecle*. B.m.: Comité des travaux historiques et scientifiques-CTHS.

PECÁKOVÁ, Iva, 2014. Problém chybějících dat v dotazníkových šetřeních. *Politická ekonomie*. ISSN 2336-8225.

PLAYFAIR, William, 1786. *The Commercial and Political Atlas: Representing, by*

Means of Stained Copper-plate Charts, the Exports, Imports, and General Trade of England; the National Debt, And Other Public Accounts; With Observations And Remarks: by William Playfair,(author Of Regulations For The Interest Of Money.) To which are Added, Charts of the Revenue and Debts of Ireland, Done In The Same Manner, by James Corry, Esq. The Commercial Part is Taken from the Custom-House Books, and the Public Accounts from the Journals of the House of Commons, and Other Papers Belonging to that House Not Yet Published. B.m.: J. Debrett, Piccadilly; GG; J. Robinson, Pater-Noster Row; J. Sewell, Cornhill; the engraver, SJ Neele, NO. 352, Strand; W. Creech; C. Elliot, Edinburgh;; L. White, Dublin.

PLAYFAIR, William, 1801. *The Statistical Breviary: Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe; Illustrated with Stained Copper-plate Charts the Physical Powers of Each Distinct Nation with Ease and Perspicuity: to which is Added, a Similar Exhibition of the Ruling Powers of Hindoostan.* B.m.: T. Bensley, Bolt Court, Fleet Street.

PLYR, [vid. 17.4.2018]. *A R package for splitting, applying and combining large problems into simpler problems.* [online]. Dostupné z: https://github.com/hadley/plyr

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

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

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

RMARKDOWN, [vid. 16.4.2018]. *Dynamic Documents for R* [online]. Dostupné z: https://github.com/rstudio/rmarkdown

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

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

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

RSTUDIO, [vid. 16.4.2018]. *RStudio - Open source and enterprise-ready professional software for R* [online]. Dostupné z: https://www.rstudio.com/

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

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

TEETOR, P., 2011. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and*

Graphics. B.m.: O'Reilly Media. ISBN 9781449307264.

TIDYR, [vid. 17.4.2018]. *Easily tidy data with spread and gather functions*. [online]. Dostupné z: https://github.com/tidyverse/tidyr

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

TUFTE, Edward R., 1990. *Envisioning Information*. B.m.: Graphics Pr. ISBN 978-0961392116.

TUKEY, John W., 1958. Bias and Confidence in Not Quite Large Samples. *Annals of Mathematical Statistics*.

TUKEY, John W., 1962. The Future of Data Analysis. *Annals of Mathematical Statistics*. B.m.: The Institute of Mathematical Statistics.

TUKEY, John W., 1977. *Exploratory Data Analysis*. B.m.: Addison-Wesley. ISBN 0201076160.

VAIDYANATHAN, Ramnath, Yihui XIE, JJ ALLAIRE, Joe CHENG a Kenton RUSSELL, 2018. htmlwidgets: HTML Widgets for R. *R package version 1.0* [online]. Dostupné z: https://cran.r-project.org/web/packages/htmlwidgets/htmlwidgets.pdf

VIZINA, Adam, Stanislav HORÁČEK a Martin HANEL, 2015. Nové možnosti modelu Bilan. *Vodohospodářské technicko-ekonomické informace*. B.m.: Výzkumný ústav vodohospodářský TG Masaryka, veřejná výzkumná instituce.

VLNAS, Radek, Adam BERAN, Martin HANEL, Anna HRABÁNKOVÁ, Tomáš HRDINKA, Ladislav KAŠPÁREK, Marta MARTÍNKOVÁ, Martina PELÁKOVÁ, Pavel TREML, Adam VIZINA, Petr BAŠTA, Lukáš JAČKA, Petr MÁCA, Jiří PAVLÁSEK a Pavel PECH, 2014. *Metodika pro stanovení mezních hodnot indikátorů hydrologického sucha*. Praha: Výzkumný ústav vodohospodářský TG Masaryka, veřejná výzkumná instituce.

VÚV TGM, 2015. *Model hydrologické bilance Bilan*. B.m.: Výzkumný ústav vodohospodářský TG Masaryka, veřejná výzkumná instituce.

WICKHAM, Hadley, 2010a. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*.

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

WICKHAM, Hadley, 2014. Tidy data. *The Journal of Statistical Software* [online]. Dostupné z: http://www.jstatsoft.org/v59/i10/

WICKHAM, Hadley a Garrett GROLEMUND, 2017. *R for Data Science*. B.m.: O'Reilly Media. ISBN 1491910399.

WIKIPEDIA, [vid. 11.8.2017]. *P–P plot - Wikipedia* [online]. Dostupné z: https:

//en.wikipedia.org/wiki/P%E2%80%93P_plot

WIKIPEDIA, [vid. 6.8.2017]. *Histogram* [online]. Dostupné z: https://en.wikipedia.org/wiki/Histogram

WILKINSON, Leland, 2005. *The Grammar of Graphics*. B.m.: Springer. ISBN 9780387286952.

XIE, Yihui, 2015. *Dynamic Documents with R and Knitr* [online]. B.m.: Boca Raton, Florida: Chapman; Hall/CRC. Dostupné z: http://yihui.name/knitr/

XIE, Yihui, 2018. *Authoring Books and Technical Documents with R Markdown*. B.m.: bookdown.

ZUMEL, Nina a John MOUNT, 2014. *Practical Data Science with R*. B.m.: Manning. ISBN 9781617291562.

Přílohy

R kód