Univerzita Komenského, Bratislava
Fakulta Matematiky, Fyziky a Informatiky

Vizualizácia verifikácie predpovedných modelov počasia

Diplomová práca

Bratislava, 2015
Bc. Marek Kružliak

Univerzita Komenského, Bratislava

Fakulta Matematiky, Fyziky a Informatiky

Vizualizácia verifikácie predpovedných modelov počasia

Diplomová práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	2511 Aplikovaná informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	RNDr. Andrej Lúčny, PhD.

Bratislava, 2015

Bc. Marek Kružliak

Tu bude zadanie

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím citovaných zdrojov.

. . . .

TODO - tu bude podakovanie

Abstrakt

TODO

Kľúčové slová: vizualizácia informácií, verifikácia predpovedí počasia

Abstract

TODO.

Keywords: information visualization, verification of weather forecasts

Obsah

Zoznam obrázkov

Zoznam tabuliek

Zoznam skratiek

WRF	Weather Research and Forecasting
NCEP	National Centers for Environmental Prediction
NCAR	National Center for Atmospheric Research

Kapitola 1

Úvod

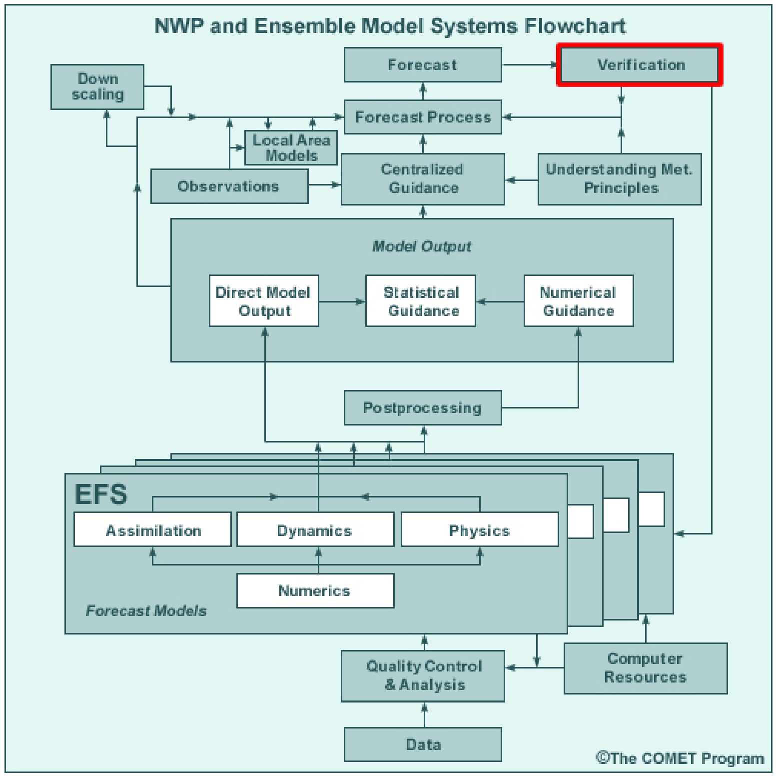
Kapitola 2

Verifikácia predpovedných modelov počasia

Verifikácia je proces, ktorý má overiť správnosť fungovania predpovedného modelu počasia. Z tohto dôvodu je nepostrádateľnou súčasťou meteorologického výskumu a taktiež celkového procesu predpovedania počasia. [CWS$^+$08] Ciele verifikácie môžeme rozdeliť do troch skupín: *administratívne*, *vedecké* a *ekonomické*. Medzi *administratívne ciele* patrí monitorovanie úspešnosti predpovedania modelu a nasmerovanie užívateľov na jeho správnu konfiguráciu alebo voľbu iného modelu. *Vedeckými cieľmi* sú identifikovanie a oprava slabín modelu a taktiež vylepšovanie predpovedí. *Ekonomickými cieľmi* sú rozhodovanie, kam majú smerovať investície do výskumu a iné závažné ekonomické rozhodnutia. [FJB12]

2.1 Predpovedný model počasia

Už v 19. storočí vývoj termodynamiky na základe Newtonovskej fyziky vyvrcholil v ucelení množiny fundamentálnych princípov, ktoré riadia prúdenie plynov v atmosfére. Začiatkom 20. storočia sa o matematický prístup k predpovedaniu počasia najviac zaslúžili osobnosti ako Vilhelm Bjerknes alebo Lewis F.Richardson. Avšak na ďalší úspech, v tejto oblasti, sa muselo čakať až na vynájdenie prvých počítačov počas 2. svetovej vojny ako bol IAS alebo ENIAC. [Lyn07] Prvá úspešná predpoveď bola vykonaná v 50. rokoch minulého storočia a to hlavne vďaka práci Jula Charneyho. Následný vývoj vo výpočtovej sile počítačov,



Obr. 2.1: Flowchart systému predpovedného model počasia od edukačného programu The COMET [LE11]. Na obrázku je zvýraznená časť, ktorej sa venujeme v tejto práci.

používanie satelitných pozorovaní a vývoj samotnej meteorológie ako vedy zapríčinil, že je numerická predpoveď počasia (NWP) dnes najúspešnejším prístupom ako predpovedať počasie. [Gol]

Odvtedy vzniklo veľké množstvo modelov, ako sú napríklad GFS, NAM, RUC, WRF, SREF, GEFS, ECMWF, ALADIN a mnoho ďalších. Naša práca sa zameriava konkrétne na verifikáciu modelu *WRF*. Taktiež pokračuje neustály vývoj aj vďaka novým modelovacím technikám, novým parametrizáciám, a zvyšovaniu výkonu výpočtových zdrojov.

Ako môžeme vidieť na obrázku 2.1 proces predpovedania počasia má okrem numerického modelu, ktorý je jej jadrom, aj iné časti. Ako príklad môžme spomenúť získavanie vstupných dát, ich predspracovanie, postprocesing, spracovanie výstupu a následne poskladanie samotnej predpovede. Cieľom nášho záujmu, celého procesu predpovedania, je *verifikácia*. Ako môžme vidieť z obrázka, verifikácia vplýva na vyladenie parametrov modelu, avšak tento proces sa nedeje automaticky, ale vyžaduje prácu meteorológov a ich chápanie základných meteorologických princípov.

2.1.1 WRF model

Ako sme už spomenuli *The Weather Research and Forecasting* (WRF) model je *numerická predpoveď počasia* (NWP) a systém atmosferickej simulácie.

WRF je podporovaný, ako bežný nástroj pre univerzity, výskum a operačné komunity, pričom sa usiluje o splnenie požiadaviek ich všetkých súčasne. Vývoj WRF modelu bol snahou mnohých spoločností ako napríklad *The National Center for Atmospheric Research's* (NCAR), *Mesoscale and Microscale Meteorology* (MMM), *The National Oceanic and Atmospheric Administration's* (NOAA) *National Centers for Environmental Prediction* (NCEP) a *Earth System Research Laboratory* (ESRL), oddelenie ministerstva obrany *Air Force Weather Agency* (AFWA) a *Naval Research Laboratory* (NRL), *The Center for Analysis and Prediction of Storms* (CAPS) [WCSDG+08].

WRF model je vhodný pre širokú škálu aplikácií od *metódy vzdušných vírov* (Large Eddy Simulation - LES) až po globálne simulácie počasia. Takéto aplikácie vyžadujú numerické predpovede v reálnom čase, vývoj a štúdium asimilácie dát, výskum parametrizovanej fyziky, výskum parametrizovanej fyziky ,modelovanie kvality ovzdušia, idealizované simulácie, čo všetko WRF model spĺňa.

V roku 2008 evidovala WRF viac ako 6000 užívateľov, no dnes (2014) eviduje viac ako 25000 užívateľov vo viac ako 130 krajinách sveta. Tieto fakty poukazujú na to, že WRF model má nie len veľkú základňu užívateľov, ale aj vývojárov a má v budúcnosti istotne svoje miesto a preto si myslíme, že sa oplatí investovať čas a úsilie do verifikácie tohto modelu.

2.2 Dáta

Na správne zhodnotenie úspešnosti modelu potrebujeme dva druhy dát. V prvom rade sa jedná o dáta, ktoré sú výstupom z daného predpovedného modelu počasia, teda **predpovedané dáta**. Tieto umelo získané dáta chceme konfrontovať s realitou, aby sme si mohli vytvoriť obraz o správnom fungovaní celého modelu. Realitu v našom prípade predstavujú dáta namerané špecializovanými meteorologickými senzormi, ktoré označujeme ako

pozorované dáta alebo skrátene *pozorovania*.

2.2.1 Predpovedané dáta

Predpovedané dáta z modelu WRF sa ukladajú vo formáte **GRIB**, čo je skratka pre *GRIdded Binary* [WMO94] alebo na iných miestach uvádzané ako *General Regularly-distributed Information in Binary form* [WMO03]. Tento formát je štandardom Svetovej meteorologickej organizácie teda *World Meteorological Organization* (WMO). Jedná sa o pomerne rozšírený formát, používaný pri veľkom množstve meteorologických aplikácií a je taktiež používaný ako výstupný formát pre iné predpovedné modely ako WRF, či už ECMWF, GFS, NAM, SREF alebo mnohé iné [NCE14].

Doteraz boli vyvinuté 3 verzie tohoto formátu od 0 po 2. Verzia 0 bola určená pre malé projekty typu TOGA a to iba s limitovaným použitím a dnes sa táto verzia už vôbec nepoužíva. Verzia grib 1 [WMO94], grib 2 [WMO03] sú dnes bežne používané väčšinou meteorologických centier.

Medzi verziami 1 a 2 nie sú žiadne rozdiely v obsahovej filozofii, preto popis obsahu gribovského formátu, ktorý tu uvádzame je spoločný pre obe tieto verzie. *Gribovský súbor* (ďalej iba *Grib*) pozostáva z viacerých *Gribovských záznamov*, pričom jeden záznam môže existovať ako samostatný Grib. Vďaka tomu je možné ľahko spájať Griby, a to tiež v ľubovoľnom poradí, bez toho, aby sme ich nejako poškodili. Samozrejme musí byť zachovaná homogenita, čo sa týka verzií Gribov, teda verziu 1 nemožno miešať s verziou 2 a naopak. Už samotný názov *Gridded Binary* nám napovedá, že dáta sú usporiadané v pravidelnej mriežke. Každý Gribovský záznam obsahuje dvojrozmernú mriežku (zemepisná šírka x zemepisná dĺžka) hodnôt v určitom čase a vertikálnej hladine. Taktiež v hlavičke záznamu sa nachádzajú metainformácie, ktoré nám hovoria o aké dáta ide, teda o akú premennú sa jedná, čas predpovede, výškovú hladinu a podobne. Grib je zvyčajne z tohto dôvodu 2 až 5 rozmerná dátová štruktúra s veľkým množstvom veličín ako je napríklad teplota, tlak, relatívna vlhkosť, rosný bod, u a v súradnice vetra a ďalšie, ktoré sú definované v rôznych hladinách. Taktiež je dôležité povedať, že Grib zriedkakedy zachytáva povrch celej planéty, ale iba vymedzenú skúmanú oblasť - *doménu*.

2.2.2 Pozorované dáta

Pozorovania sa získavajú meraním priamo v teréne pomocou špecializovaných meracích zariadení, ktoré sú súčasťou meteo staníc. Každá stanica môže obsahovať iné vybavenie, ku príkladu teplomer, zrážkomer, barometer, vetromer a im podobné [Vas98], ktorými môžme zachytávať informácie o rôznych skúmaných veličinách.

Majoritná časť meraní sa deje pri povrchu zeme priamo na meteo staniciach a nazývajú sa *surface* merania. Tieto merania najlepšie popisujú dianie v oblasti najväčšieho záujmu (biosfére), avšak neobsahujú informáciu o dianí v iných výškových hladinách. Pozorovania týchto hladín sa dejú pomocou *radiosondy*, ktorá je pripojená k meteo balónu alebo vypustená z lietadla smerom k zemi. Takéto pozorovania sa nazývajú *upper air* merania.

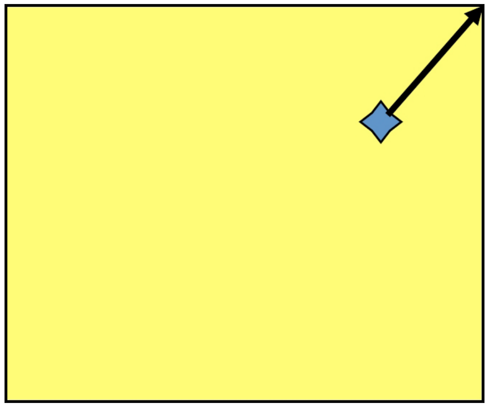
Narozdiel od predpovedaných dát, pozorované dáta nemajú štandardizovaný formát a zvyčajne sa ukladajú do databázy. Aby sme zhrnuli charakteristiku týchto dát, jedná sa o niekoľko meraných veličín, nameraných v konštantných časových krokoch - napríklad každú minútu alebo každú hodinu - v jednom konkrétnom geografickom bode a zvyčajne pri povrchu zeme, teda ak sa nejedná o upper air merania, ktoré sa uskutočňujú v štandardných výškových hladinách, ktoré sa merajú v hPa.

2.2.3 Párovanie dát

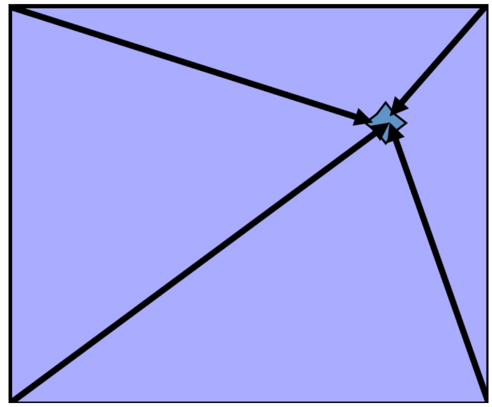
Z predpovedného modelu a rovnako aj z merania získame veľké množstvo hodnôt. Aby sme mohli korektne porovnať predpovede s pozorovaniami, je nevyhnutné nájsť správne párovanie týchto hodnôt, teda zistiť, ktorú hodnotu porovnať s ktorou, aby sme získali zmysluplný výsledok.

Vždy sa snažíme nájsť správnu predpoveď pre pozrovanie a nie naopak. Dôvodom je, že chceme skúmať vzťah predpovede s realitou a preto v párovaní musí byť zahrnutých čo najviac **meraných** hodnôt, ak nie všetky.

Každá pozorovaná hodnota, ktorú chceme spárovať má štyri kľúče podľa ktorých hľadáme pár: *meraná veličina* (napríklad teplota), *čas merania*, *výšková hladina* a *geografická poloha*. Nájsť všetky hodnoty podľa kľúča meranej veličiny v Gribe je ľahké, keďže sa jedná o kategorickú premennú, teda môže nadobúdať iba určitý konečný počet hodnôt.



(a) Point-to-Grid



(b) Grid-to-Point

Obr. 2.2: Vizuálne znázornenie dvoch bežne používaných metód na získavanie hodnôt z mriežky

Toto sa však nedá povedať o čase, hladine a polohe, ktoré sú spojitými premennými.

Pre čas pozorovania, čas predpovede a výškovú hladinu existujú štandardy, ktoré určujú v akých časoch resp hladinách sa robia merania a predpovede, čo nám uľahčuje prácu. Ak sa napriek tomu čas alebo hladina v Gribe nevyskytuje, tak pár vyhadzujeme z párovania.

V prípade polohy zo samozrejmých dôvodov neexistuje žiaden štandard a hustota mriežky v Gribe nemôže byť nikdy tak veľká, aby poloha našej stanice vždy dopadla na presný bod mriežky. Z tohoto dôvodu získavame hodnoty z mriežky z okolitých bodov a to dvoma metódami *Point-to-Grid* a *Grid-to-Point* [FJB12], ktoré sú znázornené na obrázku 2.2. Jedná sa vlastne o dve interpolačné metódy. Point-to-Grid predstavuje metódu *najbližší sused* (*Nearest Neighbour*) a Grid-to-Point *bilineárnu interpolačnú metódu*.

Výber správnej metódy môže značne ovplyvniť výsledok. Dôvodom je, že môžu byť veľké rozdiely hodnôt v okolitých mrežových bodoch a tak, ak pomocou Point-To-Grid metódy získame nízku hodnotu, tak pomocou Grid-To-Point môžeme získať hodnotu omnoho väčšiu, vplyvom zvyšných troch bodov, ktoré vstúpili do interpolácie. Nemožno však jednoznačne povedať, ktorá z metód je lepšia, keďže obe môžu v istých prípadoch dávať lepšie výsledky.

2.3 Meranie chyby predpovede

Výsledkom procesu párovania je n párov (predpoveď, pozorovanie), ktoré je možné porovnať. Z porovnania týchto dvojíc získame numerickú hodnotu, ktorá nám hovorí o veľkosti chyby predpovede daného modelu pre vybrané predpovedané časy.
Chybu predpovede e_i pre i-tu dvojicu $(y_i, \hat{y}_i)$ definujeme takto:

$$e_i = (y_i - \hat{y}_i)$$

Kde y_i je predpoveď a $\hat{y}_i$ je pozorovanie. Takýmto spôsobom z n párov získame n chýb, ktoré agregujeme pomocou rôznych štatistických metód, ktoré sú bežne používané pri verifikácií predpovedí, ako sa spomína v [Nur03], [FJB12] a [Cas09]. Výsledkom agregácie je numerická hodnota, ktorá sa nazýva *skóre* predpovede.

2.3.1 Stredná chyba predpovede

Budeme ju označovať ako *MFE* z anglického *Mean Forecast Error*, ale v literatúre je možné ju nájsť ako *ME* [Nur03], teda *stredná chyba* alebo ako *Linear Bias* [Cas09], [FJB12]. Vzorec pre výpočet MFE vyzerá nasledovne:

$$MFE = \frac{1}{n} \sum_{i=0}^{n} e_i$$

MFE je možné vypočítať aj ako rozdiel priemerov predpovedí a pozorovaní.

$$MFE = \bar{y} - \bar{\hat{y}}$$

MFE vyjadruje priemerný smer chyby. To znamená, že pozitívny výsledok indikuje *over-forecast*, teda nadhodnotenú predpoveď a negatívny výsledok *under-forecast*, teda podhodnotenú predpoveď. Avšak MFE **nevyjadruje veľkosť** chyby v tomto smere, keďže kladné a záporné chyby sa navzájom môžu zrušiť. Napríklad máme množinu chýb $E = \{2, -5\}$, tak MFE pre E je -1.5, ale priemerná veľkosť chyby je 3.5.

2.3.2 Stredná absolútna chyba

Budeme ju označovať ako *MAE* z anglického *Mean Absolute Error*. Vzorec pre výpočet MAE vyzerá nasledovne:

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|e_i|$$

Narozdiel od MFE **neurčuje smer chyby**, ale vyjadruje veľkosť chyby. Z týchto dôvodov je v praxi odporúčané zobrazovať MFE a MAE súčasne [Nur03].

2.3.3 Stredná kvadratická chyba

Budeme ju označovať ako *RMSE* z anglického *Root Mean Square Error*. Vzorec pre výpočet RMSE vyzerá nasledovne:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}e_i^2}$$

Z povahy vzorca pre RMSE je jasné, že rovnako ako MAE, ani RMSE neurčuje smer chyby, pretože nadobúda vždy iba kladné hodnoty. Ďalšou vlastnosťou RMSE je, že nadobúda hodnoty vždy väčšie alebo rovné ako MAE, pričom výsledok RMSE je citlivý na veľké hodnoty chýb.

V praxy sa zvykne používať aj *MSE* (*Mean Square Error*):

$$MSE = \frac{1}{n}\sum_{i=0}^{n}e_i^2$$

Má podobné vlastnosti ako RMSE s jediným rozdielom, že RMSE meria veľkosť chyby zachovávajúc jednotky danej veličiny (napr. ℃), zatiaľ čo MSE jednotky nezachováva [Nur03]. Preto sme si pre náš účel zvolili RMSE, ktoré je jednoduchšie zobraziť spolu s MFE a MAE v jednom grafe, keďže sa zachováva konzistentnosť jednotiek veličín.

2.3.4 Všeobecná kumulovaná chyba

V našom systéme sme navrhli všeobecný vzorec na výpočet kumulovaného skóre, ktorým možno vyjadriť ľubovoľnú zo spomenutých štatistických metód. Takéto vyjadrenie umožňuje

nie len všeobecnosť, ale aj jednoduché rozšírenie systému o ďalšie metódy a to nie len programátorom, ale aj samotným užívateľom systému.
Všeobecný vzorec na výpočet *skóre* pre danú predpoveď vyzerá takto:

$$Score = \Phi(\sum_{i=0}^{n} \varepsilon(e_i))$$

Kde Φ je ľubovoľná funkcia z $\mathbb{R}$ do $\mathbb{R}$, teda $\Phi : \mathbb{R} \to \mathbb{R}$ a podobne funkcia $\varepsilon : \mathbb{R} \to \mathbb{R}$. Spomenuté metódy môžme teda skonštruovať zadefinovaním správneho Φ a ε.
Napríklad pre *MFE*:

$$\Phi(x) = \frac{x}{n}$$

$$\varepsilon(e) = e$$

Pre *MAE*:

$$\Phi(x) = \frac{x}{n}$$

$$\varepsilon(e) = |e|$$

Pre *RMSE*:

$$\Phi(x) = \sqrt{\frac{x}{n}}$$

$$\varepsilon(e) = e^2$$

Pre *MSE*:

$$\Phi(x) = \frac{x}{n}$$

$$\varepsilon(e) = e^2$$

Ako sme spomenuli, je možné rozšírenie o ďalšie metódy a to napríklad o Brownov a Triggov *signál chybných predikcií*, ktorý budeme označovať ako *TS* z anglického *Tracking Signal*. Tieto metódy sme vyššie nespomenuli, keďže sa v meteorologickej praxi nepoužívajú. Uvádzame ich však ako možné rozšírenie, keďže sú tieto metódy bežne používané pri verifikácii iných predpovedných modeloch, ako sú tie meteorologické.

2.3.5 Medián absolútnych chýb

Budeme ju označovať ako *MAD* z anglického *Median Absolute Deviation*. Vzorec pre výpočet MAD vyzerá nasledovne:

$$MAD = median(|e|) = \widetilde{|e|}$$

Nech je daná usporiadaná postupnosť $Y_1, \ldots, Y_N$, tak potom *median* náhodnej premennej x je definovaný rovnako ako v [Wei14]:

$$median(x) = \tilde{x} \equiv \begin{cases} Y_{(N+1)/2} & \text{ak } N \mod 2 = 0 \\ \frac{1}{2}(Y_{(N+1)/2} + Y_{(N+1)/2+1}) & \text{ak } N \mod 2 = 1 \end{cases}$$

Z daného vzorca môžeme vidieť podobné vlastnosti ako má MAE, avšak MAD je robustnejší a extrémne chyby nemajú na skóre žiaden efekt.

Kapitola 3

Predchádzajúce riešenia 1

Verifikačný softvér

Verifikácia predpovedných modelov počasia je úloha dokonale stvorená pre automatizáciu. Z tohto dôvodu meteorológovia začali využívať dostupný štatistický softvér a neskôr boli taktiež vyvíjané špecializované nástroje určené pre verifikáciu. Môžeme teda rozdeliť verifikačný softvér do dvoch základných kategórií a to *štatistický* a *špecializovaný*, ktorý je zväčša podporovaný rôznymi národnými a medzinárodnými organizáciami.

3.1 Štatistický softvér

Spoločnými črtami štatistických programov, ktoré v tejto časti spomenieme sú:

- Obmedzená verifikačná funkčnosť
- Slabé / Žiadne GUI
- Potrebná znalosť špecifického programovacieho jazyka

Taktiež ide zväčša o platený softvér s pomerne vysokými cenami licencií. No aj napriek týmto slabým stránkam sú často používané v meteorologickej komunite, či už kvôli dobrej podpore a veľkému množstvu tutoriálov, alebo kvôli tomu, že poskytujú rôzne štatistické funkcie a umožňujú rýchlo doimplementovať nové verifikačné metódy.

3.1.1 Tabuľkový softvér

Napriek tomu, že je tabuľkový softvér na výpočet štatistík zamietnutý komunitou vedcov a štatistikov ako nevhodný a neprofesionálny, tak je využívaný, a to pomerne často, aj vo vedeckých kruhoch. Výhodou je, že novému užívateľovi umožňuje okamžite vidieť všetky kroky v základných procedúrach verifikácie a teda je výborný pre výučbové účely. [Poc11] Najznámejší kus softvéru z pomedzi komerčných produktov je *Microsoft Excel* [Mic15] a z voľne dostupných je jeho opensoruce náprotivok *Open Office Calculate* [Ope15]. Oba programy zahrňujú základné štatistické funkcie ako napríklad stredná kvadratická chyba (*MSE*) pre spojité predpovede (pozri odsek 2.3.3) a taktiež umožňujú generovanie jednoduchých grafov na základe tabuľkových dát. Tabuľkový softvér neposkytuje priamo funkcionalitu na výpočet ďalších sofistikovanejších verifikačných štatistík, avšak umožňuje ich implementáciu pomocou makro programovania v špecifickom jazyku. Pre Microsoft Excel je to *Microsoft Visual Basic for Applications*(VBA) [Mic13] a pre Open Office Calculate zasa *OpenOffice.org Basic* [Ope13]. Oba jazyky patria do rodiny *Basic* jazykov, takže majú mnoho podobných prvkov.

3.1.2 MATLAB

MATLAB je interaktívne prostredie s vlastným programovacím jazykom, ktorý je využívaný miliónmi inžinierov a vedcov po celom svete [TM15] a tým nevynímajúc meteorológov a ďalších odborníkov pracujúcich v atmosférickom výskume. Zvyčajne sa MATLAB využíva na výskum a protoypovanie nových metód a procedúr [Poc11], pretože umožňuje rýchlu a jednoduchú implementáciu, keďže jeho súčasťou je mnoho matematických knižníc a je prispôsobený na prácu s maticami dát. Výhodou MATLABU je, že umožňuje tvorbu GUI a taktiež poskytuje kreslenie rôznorodých grafov a diagramov. Mali by sme však podotknúť, že podobne ako väčšina štatistického softvéru, aj *MATLAB* je komerčný produkt. Jeho cena za jednu licenciu je $2,650 (k roku 2015), čo je pomerne vysoká suma, ak vezmeme do úvahy za akým účelom chceme tento softvér využívať a ako dobre je naň prispôsobený.

3.1.3 R

Často používaným a pomerne mocným nástrojom je *open source* skriptovací jazyk *R* [Fou15]. V posledných desaťročiach sa stal dominantným jazykom v oblasti štatistického výskumu. Napriek tomu, že ide o voľne stiahnuteľný softvér, tak jeho základný balík obsahuje všetky funkcie, ktoré obsahujú aj platené produkty. R-ko však nezostáva len pri tom, pretože v dobe písania tejto práce (marec 2015) bolo dostupných vyše 6400 užívateľských balíkov s rôznorodou funkcionalitou. Pre nás je dôležití, že medzi týmito balíkmi sa objavil aj balík určený na verifikáciu s názvom ***verification*** [Lab14]. Tento balík obsahuje základné funkcie verifikácie na výpočet štatistík pre spojité, kategorické ale i pravdepodobnostné predpovede.

Jazyk R neslúži iba na rôznorodé štatistické výpočty, ale poskytuje aj veľmi dobre parametrizovateľnú vizualizáciu. V balíkoch jazyka sa nachádzajú funkcie pre čiarové diagramy, krabicové diagramy, bodové grafy a mnohé iné komplexnejšie vizualizácie, ale taktiež funkcie na zobrazenie základných vizuálnych prvkov, ktorými možno vytvoriť úplne novú osobitnú vizualizáciu.

3.1.4 Statistical Analysis Software (SAS)

Statistical Analysis Software [Ins15], skrátene SAS, je opäť štatistický programovací jazyk aj so svojim vývojovým prostredím. V oblasti bioštatistiky a farmakológie je veľmi uznávaným a často používaným jazykom. Keďže je veľká podobnosť v používaných metódach medzi verifikáciou predpovedí a spomínanými odvetviami [Poc11], SAS poskytuje funkcionalitu použiteľnú aj pre verifikáciu. Okrem iného SAS ponúka základné, ale aj niektoré pokročilejšie nástroje na vizualizáciu dát. Opäť však musíme podotknúť, že ide o komerčný produkt, ktorého cena licencie je pomerne vysoká.

3.1.5 Interactive Data Language (IDL)

IDL, teda *Interactive Data Language* [Sol15] je opäť jeden z matematických programovacích jazykov, ktoré patria medzi menej používané v komunite atmosferického výskumu [Poc11]. Napriek tomu niektorí výskumníci medzi, ktorými je aj *Beth Ebert* z *Centre for*

Australian Weather and Climate Research (CAWCR) uverejnili na svojich webstránkach kód obsahujúci metódy verifikácie napísané v IDL:

- Metódy pre verifikáciu pravdepodobnosti zrážok (`http://www.cawcr.gov.au/projects/verification/POP3/POP3.html`)
- Priestorové metódy (`http://www.cawcr.gov.au/staff/eee/#Interests`)

IDL na používanie požaduje taktiež získanie platenej licencie, čo obmedzuje počet užívateľov a rovnako aj zdieľanie kódu medzi, ktorý by si mohol ktokoľvek spustiť.

3.2 Špecializovaný softvér

Tento typ softvéru je často podporovaný veľkými meteorologickými inštitúciami, ktoré majú veľa skúseností v tejto oblasti. Preto obsahuje zvyčajne špecializované metódy na prácu s meteorologickými dátami alebo aj priamo metódy slúžiace pri verifikácii predpovedných modelov počasia. Jedná sa výlučne o open source produkty, ktorých vývoj je dobre financovaný a neustále napreduje. Ďalšou spoločnou črtou je, že všetky spomenuté programy poskytujú viacmenej rovnaké vizualizačné nástroje, v ktorých vidíme priestor na výrazné vylepšenia.

3.2.1 NCAR Command Language (NCL)

Ako väčšina softvérových riešení v predchádzajúcej sekcii, tak aj *NCL*, teda *NCAR Command Language* [UCA15] je skriptovací jazyk s vlastnou syntaxou a interpreterom. Ide o voľne šíriteľný produkt od *National Center for Atmospheric Research* (NCAR), a jeho zameranie je pochopiteľne na atmosferický výskum. Jeho primárnym cieľom je spravovanie a manipulácia klimatologických dát z predpovedných modelov, čo je jedna zo základných častí verifikácie. NCL obsahuje balík funkcií, ktorými je ľahké implementovať štatistiky verifikácie spojitých premenných, ktoré sme spomínali v časti 2.3.

Súčasťou jazyka je aj možnosť vizualizácie predpovedných dát, ale taktiež aj niektoré jednoduché štatistické vizualizačné nástroje použiteľné vo verifikácii. Vzhľadom na to, že ide o meteorologický softvér, tak množstvo funkcií slúžiacich na verifikáciu je nedostačujúci, čoho príčinou môže byť, že verifikácia predpovedných modelov je ešte stále vo svojich začiatkoch [Poc11]. Sme si istý, že ako metódy aj používané praktiky pokročia, tak budú vytvorené komunitou NCL vytvorené na tento účel funkcie.

3.2.2 Model Evaluation Tools (MET)

Jeden z najprepracovanejších softvérov z oblasti verifikácie predpovedných modelov je bezpochyby *MET*, teda *Model Evaluation Tools* [DTC15]. Jeho autorstvo je pripísané opäť americkej organizácii *NCAR* v spolupráci s *Developmental Testbed Center* (DTC) a celý projekt financovala agentúra *AFWA* spolu s *NOAA*.

Rovnako ako naša aplikácia, aj MET sa sústredí na verifikáciu modelu WRF a taktiež môže byť rozšírený na použitie pre iné predpovedné modely počasia alebo iné typy predpovedí. O tomto svedčia aj niektoré z hlavných filozofických cieľov pri návrhu aplikácie a tými u modularita a prispôsobivosť softvéru. Samotný nástroj MET nie je teda samostatne existujúcou aplikáciou, ale skladá sa z viacerých modulov, ktoré môžu fungovať aj ako samostatné nástroje a taktiež môže byť MET týmto spôsobom ľahko rozšíriteľný [Cen14].

Čo sa týka verifikácie, tak MET sa sústreďuje na verifikáciu spojitých premenných a to na bodovo ale aj objektovo založenými verifikačnými metódami. *Bodovo založené* sú orientované na verifikáciu v konkrétnom geografickom bode, zatiaľ čo *objektovo založené* pristupujú k dátam geometricky a identifikujú objekty, ktorými môžu byť oblasti s istou tlakovou hladinou, búrky, oblačnosť a podobne.

Na vizualizáciu slúžia *Grid-Stat tool*, ktorý slúži na výpočet štatistík a ich vizualizáciu pre bodovo založené metódy, zatiaľ čo *MODE tool* poskytuje túto funkcionalitu pre objektovo založené metódy [Cen14]. Oba nástroje neposkytujú nijak zvlášť veľkú vizualizačnú silu, avšak v roku 2011 vznikol produkt *METViewer* [OGJ11], ktorý sa snaží tento problém riešiť. METViewer je webová aplikácia, špeciálne navrhnutá pre výstupy z MET, ktoré spracúva a vizualizuje pomocou R skriptov. Samotná aplikácia je vo verzii 1.0 ako aj 1.1 dostupná na webe (`http://www.dtcenter.org/met/metviewer/metviewer.jsp`, `http://www.dtcenter.org/met/metviewer/db/mv_hmt_2010`) a podporuje všetky typy diagramov, ako väčšina spomenutého softvéru, teda krabicové, bodové a čiarové diagramy a taktiež histogramy.

Silnými stránkami MET teda zostáva silná podpora inštitúcií, ktoré podporujú a financujú vývoj. Vďaka tomu sa vývoj posúva stále dopredu, organizujú sa rôzne workshopy a je dostupné množstvo návodov. Nevýhodou je silná závislosť na softvérových riešeniach tretej strany a taktiež silná modularita , čo zapríčiňuje komplikovanú inštaláciu softvéru.

3.2.3 Ensemble Verification System (EVS)

EVS, teda *Ensemble Verification System* [NWS15] je program vyvíjaný pod záštitou skupiny s názvom *Hydrological Ensemble Prediction* (HEP), ktorá patrí pod oddelenie *Office of Hydrologic Development (OHD)*, patriace do *National Weather Service* [Bro15].

Narozdiel od MET, EVS nepodporuje získavanie dát z rôznych modelových dátových formátov ako je GRIB1, GRIB2, BUFR a podobne (pozri sekciu 2.2), ale je potrebné, aby boli pozorovania aj predpovede vo formáte určenom pre EVS. Jeden súbor predstavuje tabuľku hodnôt oddelených medzerami, kde prvá hodnota v riadku je časová známka. EVS podporuje rôzne štatistiky pre predpovede spojitých premenných, ale taktiež aj pravdepodobnostné predpovede, ktorých výpočet je podrobne popísaný priamo v programe pri konfigurácii verifikácie.

Výsledné štatistiky, môžme v EVS vizualizovať pomocou už mnohokrát spomínaných diagramov, ktorých podrobný popis nájdeme v nasledujúcej kapitole. Výstupy z EVS je potom možné uložiť do formátu *png*, alebo ako mnoho-stránkový *pdf* súbor.

V závere môžme povedať, že EVS je kvalitný softvér ktorý sa sústreďuje na verifikáciu hydrologických a hydrometeorologických premenných. Na jeho ďalšom vývoji sa neustále pracuje a v dobe písania tejto práce (20.1. 2015) bola zverejnená verzia 5.4 [NWS15].

3.3 Zhrnutie

V tejto časti sme sa snažili prehľadným spôsobom v tabuľke 3.1 zhrnúť porovnanie spomenutého softvéru využívaného pri verifikácii.

Tabuľka 3.1: Porovnanie verifikačného softvéru

Názov	**Štatistický / Špecializovaný**	**Programovací jazyk**	**Open Source**	**Zameranie**	**Verifikácia**	**Vizualizácia**
Tabuľkový softvér	Štatistický	NIE	Rôzne	Rôzne	Žiadne explicitné verifikačné funkcie. Možnosť doimplementovať.	Základné diagramy
MATLAB	Štatistický	ÁNO	NIE	Štatistický výskum, Akademické účely	Žiadne explicitné verifikačné funkcie. Možnosť doimplementovať.	Vizualizačné nástroje v rámci jazyka
R	Štatistický	ÁNO	ÁNO	Štatistický výskum	Mnoho štatistických funkcií. Verifikačný balík ’verification’.	Veľmi dobre parametrizovateľná vizualizácia. Diagramy ľubovoľného typu.
SAS	Štatistický	ÁNO	NIE	Bioštatistika, zdravotníctvo	Funkcionalita príbuzná verifikačnej	Vizualizačné nástroje v rámci jazyka
IDL	Štatistický	ÁNO	NIE	Atmosferické vedy	Niektoré užívateľmi vytvorené knižnice.	Vizualizačné nástroje v rámci jazyka
NCL	Špecializovaný	ÁNO	ÁNO	Predpovedanie počasia a výskum v tejto oblasti	Niekoľko verifikačných funkcií.	Vizualizačné nástroje v rámci jazyka
MET	Špecializovaný	NIE	ÁNO	Predpovedanie počasia a výskum v tejto oblasti	Navrhnuté pre verifikáciu.	Nástroj METViewer. Základné diagramy.
EVS	Špecializovaný	NIE	ÁNO	Predpovedanie počasia a výskum v tejto oblasti	Navrhnuté pre verifikáciu.	Základné diagramy

Kapitola 4

Predchádzajúce riešenia 2

Techniky vizualizácie vo verifikácii

4.1 Bodový graf

Najjednoduchším spôsobom ako analyzovať vzťah dvoch náhodných premenných je *bodový graf*, ktorý je inak nazývaný aj *korelačný diagram* a známy je tiež pod svojim anglickým pomenovaním *scatter plot*. Bodový graf je vhodný na štúdium kolerácie dvoch premenných a taktiež je výborný pri odhaľovaní takzvaných *outlier*-ov, teda hodnôt, ktoré sa nejakým spôsobom výrazne odlišujú od tých ostatných. Taktiež nám nepriamo podáva správu o distribúcii hodnôt, čo však pri ich veľkom počte môže byť skreslené, keďže sa body začnú postupne prekrývať, a tak nemožno určiť v akej oblasti je viacej, či menej bodov. Tento jav sa nazýva *overplotting*.

4.1.1 Konštrukcia bodového grafu

Skonštruovanie bodového grafu je veľmi jednoduchá a aj preto je často používaným prostriedkom na vizualizáciu. Na svoju konštrukciu využíva karteziánsku sústavu súradníc. Dve náhodné premenné, ktoré chceme porovnať, vizualizujeme tak, že spravíme zobrazenie jednej premennej na x-ovú súradnicovú os a zobrazenie druhej premennej na y-ovú os. Následne v danom bode nakreslíme určený symbol, čo zvyčajne býva čierna bodka alebo krúžok. Na obrázku XYZ vidíme príklad výsledného bodového grafu, ktorý vznikne takýmto postupom.

Bodový graf ponúka mnoho spôsobov, ako pridať ďalší rozmer informácie do vizualizácie. Môžeme zvoliť odlišnú súradnicovú sústavu, čím veľmi jednoducho vytvoríme napríklad 3D bodový graf pre trojrozmerné dáta. Ďalšou možnosťou rozšírenia je zmena vykresľovaného symbolu, ktorému môžme nastavovať rôzne parametre, do ktorých možno zakódovať nové informácie. Zvyčajne sú týmito parametrami farba, alfa-transparencia [Few08], veľkosť [VWvH+07], tvar [CCM13, JHM+13] a podobne. Tieto modifikácie si v komunite vizuálnej analýzy vyslúžili aj vlastné názvy a teoretické zázemie, avšak v našej práci sa týmto druhom diagramov nebudeme venovať.

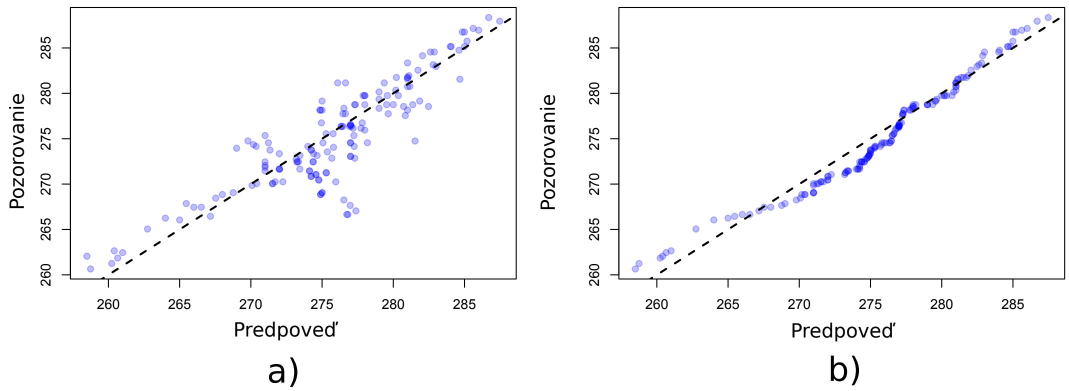
4.1.2 Kantil-kvantil graf

Dôležitou a často využívanou variáciou pre bodový graf je *kvantil-kvantil graf*, skrátene *Q-Q graf* (Q z anglického *quantil*). Jediným rozdielom medzi bodovým grafom a Q-Q grafom je, že zatiaľ, čo bodový graf vizualizuje surové dáta, Q-Q graf zobrazuje iba kvantily z oboch dátových množín.

Kvantilmi sú hodnoty, ktoré rozdeľujú usporiadanú dátovú množinu na niekoľko rovnako veľkých častí. Pre jednu dátovú množinu je $(q-1)$ q-kvantilov. Najznámejším príkladom je 2-kvantil, ktorý sa nazýva medián. Ďalšími bežnejšie používanými sú 4-kvantily, 10-kvantily a 100-kvantily, čo sú vlastne kvartily, decily a percentily.

To, čo sa zvyčajne robí pri Q-Q grafe je, že sa nevyberá konkrétny typ kvantilu, ale hodnoty sa jednoducho usporiadajú podľa veľkosti a zobrazia. Ak máme množinu A a jej veľkosť je $n = |A|$, tak vizualizjeme n q-kvantilov, kde $q = n + 1$. Tým, že nezobrazujeme na grafe surové dáta, ale kvantily, k-ty bod v grafe nie je zobrazením k-teho páru hodnôt, ale zobrazením k-teho q-kvantilu.

Na obrázku 4.1 môžme vidieť porovnanie bodového grafu a Q-Q grafu pre rovnakú dátovú množinu.



Obr. 4.1: Porovnanie bodového grafu a Q-Q grafu pre rovnaké dáta. Oba grafy boli vygenerované v programe EVS [NWS15]. a) Bodový graf b) Q-Q graf

4.1.3 Úloha bodového grafu vo verifikácii

Úloha bodového grafu a Q-Q grafu vo verifikácii je rovnaká, avšak oba nám dávajú trocha iný pohľad na dáta. Vo všeobecnosti nám tieto dva grafy dávajú informáciu o vzťahu medzi dvoma veličinami, ich korelácii a taktiež aj o ich distribúcii.

Dôležitým faktorom teda zostáva aké premenné umiestnime na x-ovú a aké na y-ovú os. Vo verifikácii predpovedných modelov počasia sú to zvyčajne predpovede na jednej osi a pozorovania na druhej. Taktiež sa zvyknú robiť dvojice (predpoveď, chyba predpovede), (pozorovanie, chyba predpovede) alebo (čas predpovede, chyba predpovede) a im podobné. V našej aplikácii sme testovali poslednú zo spomenutých dvojíc.

4.2 Krabicový diagram

Krabicový diagram je v anglickej literatúre zvyčajne nazývaný *box plot* alebo na niektorých miestach označovaný tiež ako *box and whisker*[1] *plot*. Odkedy bol prvýkrát publikovaný v roku 1977 [Tuk77], uplynulo už takmer 40 rokov a dnes ho považujeme už za štandardnú techniku ako vizualizovať distribúciu hodnôt kompaktným spôsobom. Na svoju reprezentáciu využíva súbor 5 čísel (tzv. *5-number summary*) [Pot06], ktoré charakterizujú distribúciu dát robustným spôsobom. Tým, že zredukujeme zvyčajne veľkú dátovú množinu na týchto pár hodnôt ušetríme nielen vzácny vizuálny priestor [WS12], ale taktiež námahu

[1] Slovo *whisker* znamená po slovensky fúz, čo naznačuje, že čiary, ktoré spájajú horný a dolný kvartál s hraničnými hodnotami pripomínajú fúzy.

analytika, ktorý sa snaží preskúmať iba niektoré vybrané charakteristiky.

4.2.1 Konštrukcia krabicového diagramu

Na zostavenie krabicového diagramu potrebujeme týchto 5 hodnôt: medián, horný a dolný kvartil, maximum, minimum. (pozri obrázok 4.2) Prvé tri hodnoty sú takzvané kvartily (Q1, Q2, Q3), ktoré rozdeľujú súbor dát na 4 rovnako veľké časti a ďalšie dve sú extrémne hodnoty, ktoré ohraničujú celú dátovú množinu (pozri podsekciu 4.1.2 pre vysvetlenie pojmu kvantil).

Kvartil Q2 je medián hodnôt a je definovaný rovnako ako v časti 2.3.5. Ďalej horný (Q1) a dolný (Q3) kvartil získame ako medián hodnôt pod a nad hodnotou Q2, pričom hodnotu Q2 nezahŕňame do výpočtov.

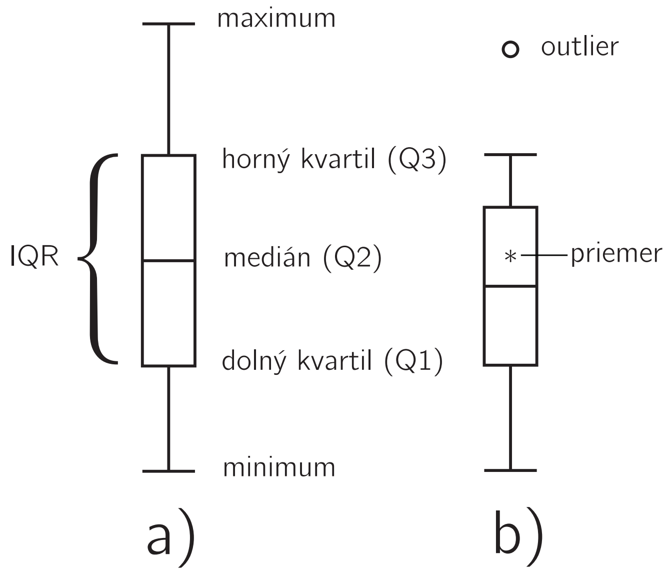
Na obrázku 4.2 vidíme, že krabica v grafe určuje pozície horného a dolného kvartilu, zatiaľ čo vnútro krabice znázorňuje takzvané *IQR*. Táto skratka označuje *interquartile range*, čo sa dá preložiť ako *medzikvartilový rozsah*. IQR definujeme ako rozdiel kvartilov Q3 a Q1:

$$IQR = Q3 - Q1$$

IQR nám hovorí o vzdialenosti týchto dvoch kvartilov, preto nám môže byť tento vzorec na pohľad podozrivý, keďže sa javí, že IQR by mohlo nadobúdať aj záporné hodnoty. My však vieme z definície Q3 a Q1, že $Q3 > Q1$ a ich rozdiel je teda vždy nezáporný (Hovoríme o rozdiele Q3 od Q1, tak ako je definované IQR).

Malú obmenu pôvodného návrhu krabicového diagramu od Tukeyho, vidíme na obrázku 4.2 b), kde malé bodky znázorňujú hodnoty nazývané *outlier*, teda hodnoty ležiace ďaleko od hlavného dátového tela, a hviezdička v strede diagramu určuje priemer hodnôt. Môžme si všimnúť, že konce čiar vychádzajúcich z boxu nemôžu byť extrémy celej množiny dát, ale sú iba extrémami vypočítaných z dát bez *outlier*-ov.

Otázkou zostáva ako určiť, ktorá hodnota je *outlier* a ktorá nie je. Na zodpovedanie tejto otázky sa využíva už spomínaný rozsah IQR. Pomocou neho sa definujú hranice *inner fences* (f_1, f_2) a *outer fences* (F_1, F_2), za ktorými hovoríme už o *outlier*-och alebo o ďalekých *outlier*-och [SOA04]. Definované sú nasledovne:



Obr. 4.2: Pôvodný návrh krabicového diagramu, ako bol prezentovaný v práci *Exploratory Data Analysis*(1977) [Tuk77]

$$f_1 = Q1 - c \times IQR$$

$$f_2 = Q3 + c \times IQR$$

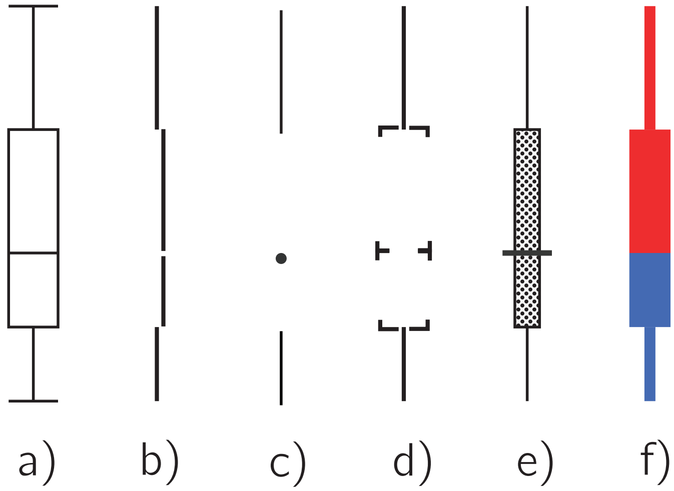
$$F_1 = Q1 - C \times IQR$$

$$F_2 = Q3 + C \times IQR$$

Konštanty c a C sú v niektorých zdrojoch definované rôzne. Najčastejšie sa však vyskytujú hodnoty $c = 1.5$ a $C = 3$, tak ako ich určil pôvodný autor krabicového diagramu [Tuk77].

4.2.2 Ďalšie variácie krabicového diagramu

Popularita krabicového diagramu nevyhnutne viedla k jeho vývoji a modifikáciám. Môžeme hovoriť o dvoch druhoch modifikácií. Jednak *syntaktickej* (vizuálnej), kedy sa zachovávajú všetky vlastnosti a informácie ako v pôvodnom diagrame, len sa menia vizuálne prvky grafu. A modifikácii *sémantickej* pridaním ďalšej popisnej informácie do grafu, čo má na záver vplyv aj na jeho vizuálnu stránku.



Obr. 4.3: a) Klasický krabicový diagram b-f) Vizuálne variácie krabicového diagramu b-c) 2 variácie pre kvartilový graf [Tuf83] c) Skrátený krabicový diagram [PKR07] e) Range-bar chart [Spe52] f) Farebná variácia [Car94]

N a obrázkoch 4.3 b-f) môžme vidieť niektoré vizuálne variácie krabicového diagramu. Vznik prvých troch motivovala snaha maximalizovať takzvaný *data-ink* [Tuf83], teda množstvo atramentu alebo počet pixlov, ktoré zodpovedajú nejakým dátam. Autori sa teda snažili čo najviac znížiť počet vizuálnych prvkov a ponechať len tie, ktoré skutočne nesú nejakú informáciu.

Na obrázku 4.3 b) a c) vidíme dve z viacero riešení navrhnutých v knihe *The Visual Display of Quantitative* [Tuf83], ktoré autor nazýva *kvartilový graf*. Preceptuálne štúdie [SB91] však ukázali, že tieto variácie sú výrazne menej presné ako originálny návrh.

Návrh d) ukazuje skrátený krabicový diagram [PKR07], ktorý sa taktiež snažil zredukovať množstvo okupovaného vizuálneho priestoru. Rozdielom je však to, že jeho účelom nie je existovať samostatne, ale ako súčasť *summary plot*-u, ktorý zahŕňa histogram a ďalšie glify znázorňujúce informácie ako priemer, štandardná odchylka alebo koeficient asymetrie.

Obrázok 4.3 e) znázorňuje predchodcu krabicového diagramu *range* graf alebo tiež nazývaný *range-bar* graf, ktorého autorkou je Mary Eleanor Spear [Spe52].

Ako posledný príklad vizuálnej modifikácie uvádzame pridanie farieb do krabicového diagramu. Táto farebná variácia uchováva tvar diagramu, avšak časť nad mediánom je zafarbená inou farbou ako hodnoty pod. Autori článku odporúčajú červenú a modrú farbu, tak ako na obrázku 4.3 f). Cieľom tohto prístupu bolo chápať krabicový diagram ako jednu percepčnú jednotku a nahradiť 5 symbolov jedným, čím sa mala znížiť námaha pozorovateľa pri analýze a taktiež uľahčiť porovnávanie viacerých diagramov navzájom.

-

K rabicový diagram umožňuje svojim vzhľadom zakódovanie ďalšej informácie do grafu. Na obrázku **??** vidíme aspoň niektoré najčastejšie úpravy...

Len rok po oficiálnom publikovaní krabicového diagramu vznikol článok [MTL78], ktorý zhŕňa jeho tri najčastejšie používané modifikácie, z ktorých prvé dve môžme vidieť na obrázkoch **??**a) a **??**b) a tretí je ich kombináciou. V prvom prípade sa využíva šírka boxu na zakódovanie veľkosti množiny, ktorú skrýva za sebou diagram. Takýto graf sa nazýva *Krabicový diagram s variabilnou šírkou*. V druhom prípade ide o takzvaný *Vrúbkovaný krabicový diagram*. V tomto grafe sú pridané *vrúbky*, ktoré zhruba naznačujú ako výrazné sú rozdiely v miere spoľahlivosti rôznych dátových množín.

V niektorých prípadoch krabicový diagram zakrýva skutočný tvar dát, teda jeho šikmosť alebo modalitu, keďže jeho vzhľad nabáda k tomu, aby si užívateľ myslel, že sú dát zacentrované na stred a unimodálne. V práci s názvom *Can the Box Plot be Improved?* [CM05] autor uvádza príklad kedy skutočne rôznorodé dáta generujú rovnaký krabicový diagram. Tento problém rieši elegantným a čistým spôsobom pridaním hrubej čiary na základe koeficientu asymetrickosti γ. Na obrázku **??**c) môžme vidieť zľava asymetrické dáta, na stred zarovnané dáta a bimodálne dáta.

Krabicový diagram umožňuje mnoho ďalších rozšírení napríklad pridaním informácie o hustote dát (histogramový krabicový diagram, vázový diagram [Yoa88] , huslový diagram [HN98]), rozšírením pre viacrozmerné dáta (vrecový graf [RRT99], 2D krabicový diagram [Ton05]) alebo zobrazením hodnôt v inej súradnicovej sústave (napríklad polárnej - vejárový graf [Fis10]). Opis týchto techník je však nad rámec tejto práce, preto ho tu ani nebudeme uvádzať.

4.2.3 Úloha krabicového diagramu vo verifikácii

Ako sme spomenuli v úvode tejto sekcie, vo všeobecnosti je úlohou krabicového diagramu zobraziť distribúciu dát v kompaktnom tvare a teda slúži na rýchle porovnanie distribúcií viacerých skupín dát. Pri verifikácii spojitej predpovede sa používa na viacero účelov a my tu spomenieme len niekoľko z nich.

V prvom rade ide o porovnanie distribúcie predpovedí s distribúciou pozorovaní za istý časový interval. V takomto prípade máme vedľa seba iba dva krabicové diagramy, ktoré navzájom porovnávame. Použitie krabicového diagramu v takomto prípade, kedy jeden graf pozostáva iba z niekoľkých (2 až 4) krabicových diagramov považujeme za zbytočné. Pri takomto počte nie je potreba na redukciu vizuálnych prvkov a existujú lepšie techniky na vizualizáciu distribúcie, ktoré sprostredkúvajú viacej informácie a teda môže byť analýza efektívnejšia.

Ďalším použitím je porovnanie distribúcie chýb predpovedí, či už pre rôzne merania, konkrétne predpovedané časy, rôzne predpovedné modely a podobne. Tu považujeme použitie krabicového diagramu za opodstatnené, keďže ide zväčša o porovnávanie väčšieho množstva distribúcií, a tak je jeho jednoduchosť, čitateľnosť, kompaktnosť a iné jeho vlastnosti potrebné.

4.3 Histogram

Ďalším veľmi bežným spôsobom, ako vizualizovať distribúciu dát je pomocou *histogramu*, ktorého existencia sa datuje do roku 1895, kedy ho uviedol vo svojej práci Karl Pearson [Pea95]. Histogram využíva veľmi jednoduchú myšlienku vizualizovať frekvencie hodnôt ako stĺpce rôznej výšky, ktorá sa v priebehu histórie analýzy dát ukázala ako veľmi užitočná.

4.3.1 Konštrukcia histogramu

Konštrukcia histogramu je pomerne jednoduchá. V prvom rade potrebujeme rozdeliť interval hodnôt na disjunktné podintervaly rovnakej dĺžky. Na vygenerovanie týchto intervalov využíva jeho autor počiatok O, od ktorého začneme generovať intervaly, a veľkosť intervalu w. Prvý interval je $\langle O, O+w)$ a všetky nasledujúce vzniknú posúvaním intervalu o dĺžku w. Ďalej pre každý takto vzniknutý interval vypočítame koľko hodnôt doň spadá. Potom už môžme vykresliť jednotlivé stĺpce pre každý interval ako obdĺžniky so šírkou w a výškou, ktorá sa určí na základe početnosti hodnôt pre daný interval. Na obrázku 4.4 môžeme vidieť takýmto spôsobom vygenerovaný histogram.

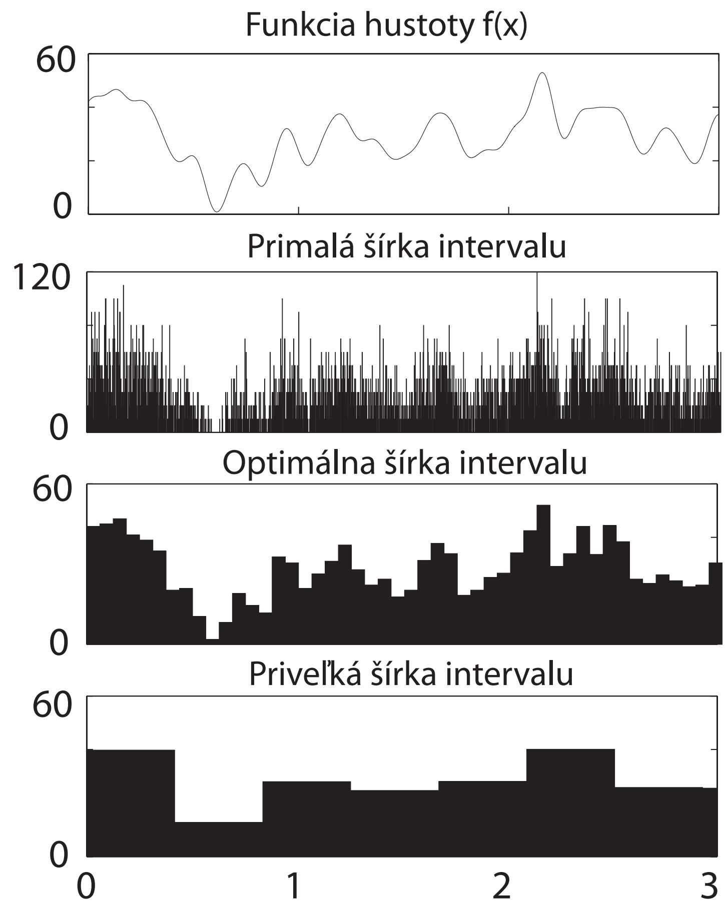
Z vyššie popísaného postupu a rovnako aj z obrázka 4.4 môžme vidieť, že konštrukcia histogramu závisí od zvoleného počiatku O, no v prvom rade na zvolenej veľkosti intervalu w. Z tohto dôvodu vzniklo mnoho prístupov, ktoré sa snažia optimalizovať w, tak aby sa získal optimálny histogram. Optimálny histogram $h(x)$ je taký, ktorý minimalizuje integrovanú strednú kvadratickú chybu (*MISE*) vzhľadom na pôvodnú funkciu hustoty $f(x)$, ktorú vypočítame ako $MISE = \int (h(x) - f(x))^2 dx$. Jedným z príkladov, ktorý toto využíva je algoritmus na generovanie optimálnej hodnoty w, ktorý navrhol *Shimazaki* a *Shinomoto* [SS07]. Autori algoritmu požívajú dekompozíciu MISE na vytvorenie cenovej funkcie $C_n(w)$ pre dĺžku intervalu w:

$$C_n(w) = \frac{2\bar{k} - v}{(nw)^2}$$

kde N je počet intervalov šírky w, n je počet všetkých hodnôt množiny, $\bar{k}$ je priemerný počet hodnôt v jednom intervale a v je variancia počtu hodnôt v intervaloch. Priebeh algoritmu je potom už len taký, že generuje postupne rôzne w, kým nenájde také s najmenšou cenou $C_n(w)$.

4.3.2 Úloha histogramu vo verifikácii

Ako sme už spomenuli, primárna úloha histogramu je detailná vizualizácia distribúcie dát. Vzhľadom k vysokej úrovni detailu, nie je jednoduché porovnávanie mnoho histogramov súčasne, tak ako to bolo u krabicového diagramu. Z tohto dôvodu sa zvyčajne v praxi porovnáva navzájom len niekoľko rôznych histogramov. Histogramy vo verifikácii slúžia na porovnanie distribúcie predpovedí a pozorovaní. Porovnávanie stĺpcov sa robí buď v dvoch rôznych diagramoch umiestnených vedľa seba alebo v jednom diagrame, kde sa



Obr. 4.4: Porovnanie rôznych dĺžok intervalov. Obrázok je upravený z pôvodného článku [SS07]

prislúchajúce stĺpce umiestnia pri sebe a farebne označia, aby ich bolo možné ľahšie rozlíšiť.

4.4 Čiarový diagram

Čiarový diagram je najjednoduchší spôsob ako vizualizovať jednorozmerné dáta. Využíva sa najmä na vizualizáciu spojitej premennej, keďže čiara narozdiel od bodov podporuje vizuálny dojem spojitosti. Pôvod čiarového diagramu sa datuje do prapočiatkov vizualizácie informácií. Jeho autorom je zakladateľ grafických metód v štatistike William Playfair, ktorý objavil 4 typy diagramov medzi, ktorými bol aj čiarový diagram v roku 1786 [Fri09].

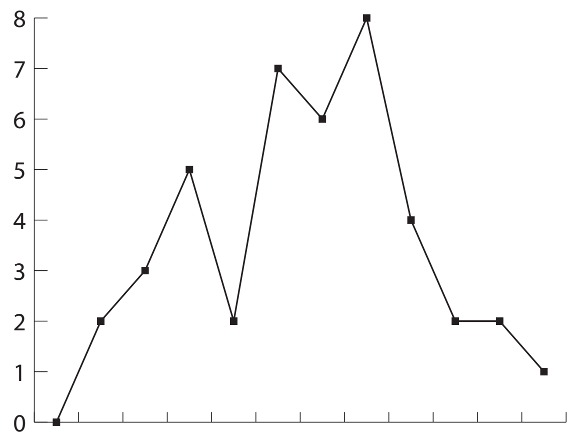
4.4.1 Konštrukcia čiarového diagramu

Pri konštrukcii čiarového diagramu sa hodnoty z množiny zobrazia v karteziánskej sústave ako body a jednotlivé dvojice susedných bodov sa následne pospájajú čiarou. Na obrázku 4.5 môžme vidieť príklad čiarového grafu vytvoreného týmto postupom.

Čiarový graf poskytuje rôznorodé úpravy, ktoré pridávajú ďalšiu informáciu do grafu alebo zlepšujú niektoré vizuálne vlastnosti. Napríklad je možné upravovať šírku, tvar alebo farbu čiary, pridávať ďalšie čiary do grafu, zvýrazniť jednotlivé body pre dané hodnoty, použiť iný typ interpolácie medzi bodmi (zvyčajne sa používa lineárna interpolácia) a podobne.

4.4.2 Úloha čiarového diagramu vo verifikácii

Vo verifikácii je účelom čiarového diagramu vizualizácia časových radov. Na x-ovej osi zvykne byť časový interval s konkrétnymi dátumami predpovedí alebo hodiny predpovede. Rozdiel medzi nimi je ten, že konkrétny dátum je v histórii iba raz, zatiaľ čo model predpovedá pravidelne, takže predpovedné hodiny môžu byť rovnaké pre rôzne dátumy. Na y-ovej osi teda bývajú buď jednotlivé hodnoty chýb alebo pre viacero hodnôt vypočítané štatistiky ako boli spomenuté v časti 2.3.



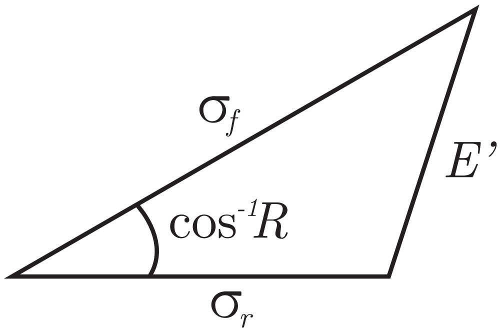
Obr. 4.5: Príklad čiarového grafu vygenerovaného v programe Adobe Illustrator

4.5 Taylorov diagram

Taylorov diagram bol špeciálne navrhnutý pre verifikáciu modelov počasia, ktorého autorom je Karl. E. Taylor [Tay01]. Podľa slov autora úlohou diagramu je štatistické zhrnutie, ako dobre si dátové vzory zodpovedajú v zmysle troch metrík: koeficient korelácie, strednej kvadratickej chyby a smerodajnej odchýlky. Použitie taylorovho diagramu nie je tak časté ako pri predchádzajúcich vizualizačných technikách, možno aj z dôvodu horšej čitateľnosti grafu a taktiež nemá takú veľkú tradíciu v štatistike, ako predošlé techniky. Podobnú snahu zobraziť tri štatistiky súčasne v jednom grafe mal aj Boer a Lambert vo svojom BLT diagrame [BL01]. Tento typ grafu je veľmi podobný taylorovmu grafu, a keďže jeho použitie je ešte zriedkavejšie, tak sa mu v tejto práci nebudeme venovať.

4.5.1 Konštrukcia taylorovho diagramu

Taylorov graf porovnáva jednu alebo viacero dátových množín (f) s jednou referenčnou množinou r. Označme si hodnoty v porovnávanej množine f_n a s nimi spárované referenčné hodnoty r_n, ktoré sú definované na N diskrétnych bodoch (napríklad v čase alebo priestore). Ako sme už spomenuli, taylorov graf vizualizuje 3 štatistiky súčasne, na čo využíva ich vzájomný geometrický vzťah. Týmito štatistikami sú: koeficient korelácie R, centrovaná stredná kvadratická chyba E' a smerodajné odchýlky σ_f a σ_r pre f, r. Vzorce



Obr. 4.6: Geometrický vzťah pre popisné štatistiky $R, E', \sigma_r, \sigma_f$

na výpočet sú nasledovné:

$$R = \frac{\frac{1}{N}\sum_{n=1}^{N}(f_n - \bar{f})(r_n - \bar{r})}{\sigma_f \sigma_r}$$

, kde $\bar{f}$ a $\bar{r}$ sú priemerné hodnoty daných množín.

$$E' = \sqrt{\frac{1}{N}\sum_{n=1}^{N}[(f_n - \bar{f}) - (r_n - \bar{r})]^2}$$

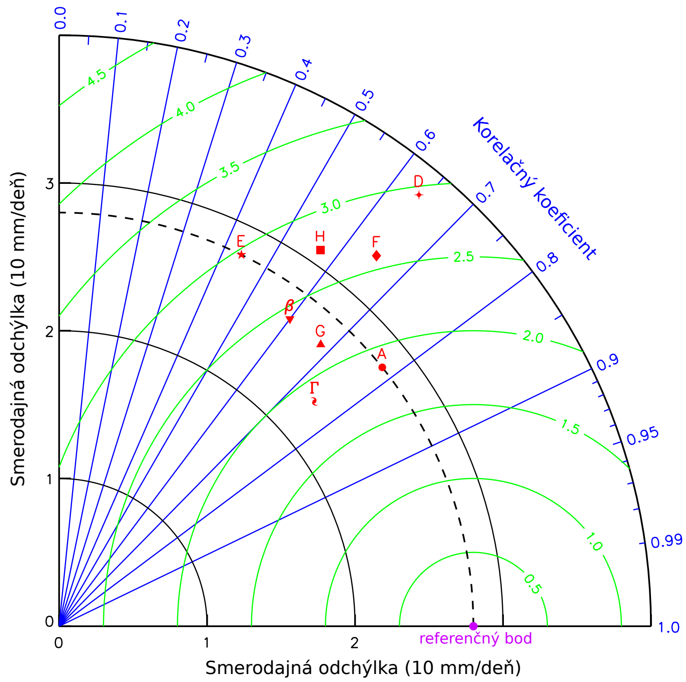
Pre pripomenutie chceme upozorniť, že nejde o strednú kvadratickú chybu, ktorá bola spomenutá v časti 2.3, ale o *centrovanú* strednú kvadratickú chybu, ktorá je relatívna k stredu, teda priemeru skúmaných množín.

Rovnako pre σ_f aj σ_r je výpočet nasledovný:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{n=1}^{N}[(x_n - \bar{x})]^2}$$

Kľúčovým bodom v konštrukcii diagramu je, že s využitím týchto troch respektíve štyroch štatistík môžme skonštruovať nasledovný vzťah:

$$E'^2 = \sigma_f^2 + \sigma_r^2 - 2\sigma_f \sigma_r R$$



Obr. 4.7: Taylorov diagram [Tay01]

, ktorý sa nápadne podobá na kosínusovú vetu:

$$a^2 = b^2 + c^2 - 2bc\cos\alpha$$

, kde a, b a c sú dĺžky strán trojuholníka a uhol α je protiľahlý uhol pre stranu a. Geometrický vzťah pre R, E', σ_f a σ_r je zobrazený na obrázku 4.6. Ako vidíme na obrázku 4.7, jedna dátová množina sa pomocou týchto troch spomenutých hodnôt a vyššie uvedeného geometrického vzťahu zobrazí na grafe s polárnymi súradnicami. Azimutálna pozícia bodu nám hovorí o korelačnom koeficiente medzi množinami, preto sa vždy referenčný bod nachádza v najnižšej úrovni grafu, kde je korelácia rovná jednej. Vzdialenosť od počiatočného bodu 0 nám udáva smerodajnú odchýlku σ a vzdialenosť od referenčného bodu udáva centrovanú RMSE E'. Pre jednoduchšie odčítavanie hodnôt sa pridávajú aj označené radiálne čiary jednak od bodu 0, ale aj od referenčného bodu, ktoré sa však v praxi zvyknú vynechávať.

Popísaným spôsobom vytvoríme základný taylorov diagram, ktorý však ponúka niekoľko modifikácií [Tay05], z ktorých niektoré používané spomenieme:

- Diagram môže byť rozšírený o ďalší kvadrant vľavo, na zobrazenie zápornej kolerácie
- Štatistiky porovnávaných hodnôt môžu byť normalizované pomocou štatistík referenčných, čím dosiahneme, že na jednom diagrame možno zobraziť rôzne jednotky veličín.
- Izočiary sa vynechávajú, aby bolo možné lepšie vidieť zobrazené body.
- Pri porovnávaní výsledkov z viacerých verzií predpovedných modelov sa zvykne nakresliť šípka medzi týmito bodmi, aby bolo lepšie vidieť vzťah medzi nimi.
- Miera E' môže byť nahradená inou. Niektoré príklady mier možno nájsť v Taylorovom článku [Tay01]
- Do grafu je možno pridať ďalšiu informáciu modifikovaním vizuálnych vlastností bodu podobne ako v bodovom diagrame. Jedným z príkladov je zobrazenie percentuálnej priemernej chyby pomocou rôznych symbolov.

4.5.2 Úloha taylorovho diagramu vo verifikácii

Tento typ diagramu bol priamo navrhnutý pre účely verifikácie predpovedných modelov počasia. Hlavným účelom diagramu je porovnávanie výkonu viacerých modelov alebo tiež jedného modelu s rôznymi nastaveniami parametrov.

Taylorov diagram skrýva dáta pod komplikovaný matematický model, ktorý nám poskytuje 3 vyššie popísané charakteristiky dát na vizualizáciu. Ak sa k tomu pridáva fakt, že diagram umožňuje iba jednu referenčnú dátovú množinu, tak použitie tohto diagramu je značne obmedzené. Využitím spomínaných troch charakteristík taktiež výrazne strácame pohľad na detail dát, ale na druhej strane nám to umožňuje jednoduché porovnávanie pomerne veľkých a komplikovaných množín na malo priestore.

Ďalšou slabinou je, že charakteristika E' sa počíta ako centrované RMSE, ktoré je relatívne vzhľadom na stred teda priemernú hodnotu dátovej množiny. Tento fakt implikuje to, že centrované RMSE nemá jednu dôležitú vlastnosť, ktorú naopak klasické RMSE má. Touto vlastnosťou je, že čím je RMSE bližšie k nule, tým podobnejšie sú aj dátové množiny. Keďže štatistická charakteristika E' nemá túto vlastnosť, tak nám euklidovská vzdialenosť skúmaného bodu od referenčného nedáva dobrú informáciu o podobnosti dvoch množín a môže pôsobiť mylne.

Kapitola 5

Návrh vizualizácie

5.1 Charakteristika dát

5.2 Špecifikácia požiadaviek na vizualizáciu

5.3 Návrh rozloženia prvkov vizualizácie

5.4 Návrh vizualizácie štatistík verifikácie

5.5 Návrh vizualizácie distribúcie chýb

Pri verifikácii predpovede spojitej premennej sme použili štatistické metódy spomenuté v sekcii 2.3, ktorých výsledok sme následne vizualizovali. Pôvodné dáta však zostali skryté za použitým matematickým modelom, a tak sme stratili informáciu o distribúcii chyby. Pri verifikácii sa štandardne používajú dve metódy na priamu, či nepriamu vizualizáciu a analýzu distribúcie, ktoré sme opísali v sekcii 4. Týmito metódami sú bodový graf (pozri podsekciu 4.1) a krabicový diagram (pozri podsekciu 4.2).

Pri návrhu vizualizácie sme vyskúšali niekoľko vizualizačných techník a zvážili ich silné a slabé stránky.

5.5.1 Graf hustoty

Jedným z viacerých spôsobov, ako pomerne presne určiť distribúciu chýb je pomocou *grafu hustoty*. Ten sa skonštruuje jednoducho z *funkcie hustoty*, ktorú získame *odhadom hustoty* z dát.

Prvý pohľad na dáta by nám vravel, že ide o dvojrozmerné dáta a teda je potrebné použiť odhad hustoty dvoch premenných. Takýto postup by samozrejme bol možný, ale doviedol by nás k chybnej vizualizácii a tak aj k mylnej predstave o dátach. Dôvodom je to, že máme záujem o analýzu distribúcie chýb pre každú hodinu predpovede zvlášť, čo znamená, že chcem zistiť distribúciu iba v jednom smere.

Pre vytvorenie grafu hustoty, v prvom rade je potrebné vybrať správny spôsob odhadu hustoty. Jedným z bežne používaným štandardným spôsobom je odhad hustoty pomocou jadra, po anglicky známy ako *kernel density estimation* (KDE) [ref Rosenblatt 56, parzen 62].

Nech máme n hodnôt $x_i, 1 \leq i \leq n$, z ktorých chceme určiť odhad hustoty, potom estimátor hustoty $\hat{f}_h(x)$, ktorý aproximuje *funkciu hustoty pravdepodobnosti* (PDF) f, sa vypočíta takto:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=0}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

kde h je šírka jadra a $K(x)$ je funkcia jadra (skrátene iba jadro), ktorá by mala spĺňať nasledovné vlastnosti:

$$K(x) \geq 0$$

$$\int K(x)dx = 1$$

tieto vlastnosti hovoria, že $K(x)$ je na celom definičnom obore nezáporná a jej integrál je rovný 1, teda sa jedná o normalizovanú funkciu. Bolo preštudovaných mnoho jadier, ako napríklad uniformné, tri-angulárne, Epanechnikovo [29], kvadratické, Gaussové, kosínusové a veľa ďalších. Najbežnejšie a zrejme aj najpraktickejšie [18] je Gaussovo (normálne) jadro, ktoré sme použili aj my:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Voľba jadra však nemá na výsledok až taký vplyv, ako voľba šírky jadra h. My sme použili výpočet šírky jadra na základe dátovej množiny, ktorý aproximuje optimálnu šírku jadra [Scott 1992; Bowman and Azzalini 1997] [1]. Všeobecne pre d dimenzionálne dáta je vzorec nasledovný:

$$h = \sigma\Big(\frac{4}{(d+2)n}\Big)^{\frac{1}{d+4}}$$

kde σ je smerodajná odchýlka vypočítaná z daných dát a n je veľkosť dátovej množiny. V našom prípade je $d = 1$, a tak sa nám vzorec zjednoduší na

$$h = \sigma\Big(\frac{4}{3n}\Big)^{\frac{1}{5}}$$

Aby sme zjednodušili výpočet, tak sme si konštanty vypočítali predom a zaokrúhlili na 2 desatinné miesta, čo považujeme za dostačujúce. Výsledný vzorec, ktorý sa nakoniec objavil v aplikácii je takýto:

$$h = 1.06 \times \sigma \times n^{-\frac{1}{5}}$$

Takýmto spôsobom sme si pre každú hodinu predpovede určili samostatnú funkciu $\hat{f}_h(x)$, ktorú môžme vizualizovať. Zvyčajne sa funkcie hustoty vizualizujú ako bežné funkcie, teda pomocou čiarového diagramu, tak ako na obrázku **??**. V našom prípade by bol tento prístup nepraktický, keďže máme veľké množstvo funkcií, tak jednak by bola takáto vizualizácia nepraktická pri porovnávaní distribúcií a taktiež na to nemáme potrebný vizuálny priestor.

Opäť sme teda zvolili štandardné riešenie, ako ušetriť vzácny priestor a to tak, že hodnoty, ktoré by boli zobrazené na y-ovú os zobrazíme na zvolenú farebnú škálu. Vďaka tomu by teoreticky mohol mať graf hustoty pre jeden čas predpovede šírku 1 pixel bez straty akejkoľvek informácie.

Vieme, že pre ľudí nie je také jednoduché pozorovať malé rozdiely medzi dvoma farbami. Testovaním sme zistili, že pri takejto vizualizácii, že tento fakt spôsobuje problémy aj pri našej vizualizácii, kedy sa pre pozorovateľa strácajú výkyvy hodnôt

[1] Optimálna šírka jadra je taká, ktorá minimalizuje *strednú integrovanú kvadratickú chybu*.

5.5.2 Pruhový kvantilový diagram

Z časti 4.1 sme už dobre oboznámený s pojmom kvantil. Klasický kvantilový diagram [ref] zobrazuje kvantil hodnôt pre jednu dimenziu. Ak si vezmeme naše dáta, kde pre každý čas je niekoľko chýb predpovedí, tak kvantilový diagram skonštruujeme tak, že pre každý čas vypočítame kvantil, ktorý zobrazíme ako bod alebo ako súčasť lomenej čiary v grafe. Prirodzeným rozšírením je zobrazovať nielen jeden kvantil, ale mnoho kvantilov súčasne. Zvyčajne sú to tieto kvantily $Q_{0.02}, Q_{0.98}, Q_{0.25}, Q_{0.75}, Q_{0.5}$. V našej práci sme využili toto rozšírenie na lepšie zobrazenie distribúcie a navrhli sme takzvaný *pruhový kvantilový diagram*.

Jeden *pruh* v grafe definujeme pomocou dvojíc hodnôt v čase - spodným a jeho protiľahlým kvantilom. Spodným kvantilom je kvantil Q_α a k nemu protiľahlý je kvantil $Q_{1-\alpha}$, kde $0 \leq \alpha \leq \frac{1}{2}$. Vidíme teda, že pruh ohraničuje hodnoty v okolí stredu usporiadanej množiny dát. Špeciálnym prípadom pruhu je pre $\alpha = \frac{1}{2}$, vtedy spodný aj horný kvantil je $Q_{0.5}$, čo je vlastne medián.

Pri návrhu vizualizácie sme sa snažili, aby mohol mať diagram variabilný počet pruhov a taktiež, aby rozostup medzi pruhmi bol pravidelný. Pri riešení tohto problému, sme sa inšpirovali krabicovým diagramom, kde sa hodnoty delia mediánom na dve časti, ktoré sa ďalej taktiež delia ich mediánom. Ide teda o rekurzívne delenie usporiadanej množiny na polovicu do hĺbky 2. Túto myšlienku sme rozšírili na ľubovoľnú hĺbku delenia d. Potom i-ty pruh $\mathcal{P}_d(i)$ pre hĺbku d definujeme takto:

$$p_d(i) = (Q_\alpha, Q_{1-\alpha}), \alpha = i \times (0.5)^d$$

$$\mathcal{P}_d(i) = \{(t, p_d(i)) : t \in I\}$$

a množina všetkých pruhov grafu pre hĺbku delenia d je definovaná takto:

$$\{\mathcal{P}_d(i) : 0 \leq i \leq 2^{d-1}, i \in \mathbb{N}\}$$

Aby sme sa vyhli rekruzii, vypočítali sme si krok medzi susednými kvantilmi pri hĺbke d, ktorý je $(0.5)^d$, a jednotlivé pruhy sme generovali s týmto krokom. Pri rekruzívnom delení sa vygeneruje 2^d kvantilov a z nich je možné vyrobiť 2^{d-1} pruhov, preto sme index i obmedzili na $i \leq 2^{d-1}$. Z tohto vidíme, že počet pruhov grafe s rastúcou hĺbkou rastie exponenciálne, preto odporúčame, aby d bolo maximálne 4, kedy sa nám množina rozdelí na 16 častí 15 hexadecilmi a tak vznikne 8 pruhov.

Na obrázku **??** vidíme, že takto definovaný pruh sa potom vizualizuje, ako plocha medzi krivkami, ktoré tvoria dvojice hodnôt patriace danému pruhu, s výnimkou špeciálneho prípadu $Q_{0.5}$, ktorý vizualizujeme ako krivku.

5.5.3 Funkčný krabicový diagram

Pre pochopenie dát je dôležité, aby sme sa vedeli pozrieť na hodnoty v ich kontexte. Všetky predošlé techniky uvažovali o chybe ako o samostatnej hodnote pre určitý predpovedný čas, avšak chyby sa nenachádzajú len v kontexte predpovedného času, ale aj v kontexte konkrétnej predpovede. Preto môžme uvažovať o predpovediach ako o funkciách $x_i(t)$, kde $i \in \{1..n\}$ je poradie predpovede a $t \in I$ je čas predpovede, kde I je časový interval predpovede z $\mathbb{R}$ (v našom prípade sa jednalo o dvojdňovú, teda 48 hodinovú predpoveď).

Takýmto spôsobom sme sa dostali do novej situácie, kedy nechceme vizualizovať distribúciu jednotlivých chýb, ale celých predpovedí, ktoré chápeme ako funkcie. Na riešenie tohto problému existuje niekoľko spôsobov, z ktorých sme si zvolili *funkčný krabicový diagram* [SG11], keďže myšlienkovo vychádza z klasického krabicového diagramu, ktorý je jednak na túto situáciu vhodný ale je tiež medzi užívateľmi dobre známy a zaužívaný.

Ako sme spomenuli v časti 4.2, klasický krabicový diagram potrebuje na svoju konštrukciu 5 hodnôt: 3 kvartily a 2 extrémy. Aby sme tieto hodnoty našli pre funkcie, musíme ich vedieť porovnať a povedať, ktorá je "väčšia"alebo "menšia". Autori funkčného krabicového diagramu riešia problém s využitím takzvanej pásmovej hĺbky (*band depth*) [LPR09]. Grafom G funkcie x je množina bodov $G = \{(t, x(t)) : t \in I\}$. Pásmo $\mathcal{B}$ (*band*) v $\mathbb{R}^2$ ohraničené krivkami $x_{i_1}, x_{i_2}, .., x_{i_k}$, kde $k \geq 2$ je definované takto:

$$\mathcal{B}(x_{i_1}, x_{i_2}, .., x_{i_k}) = \{(t, y) : t \in I, \min_{r=1..k} x_{i_r}(t) \leq y \leq \min_{r=1..k} x_{i_r}(t)\}$$

Pásmo $\mathcal{B}$ je teda množina všetkých bodov existujúcich medzi extrémami všetkých kriviek, ktoré doň vstupujú ako parameter. Pomocou týchto dvoch funkcií môžme definovať pomocnú funkciu $BD_n^{(j)}(x)$ pre krivku x, ktorá vyzerá takto:

$$BD_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq n} \mathcal{I}\{G(x) \subset \mathcal{B}(x_{i_1}, ..., x_{i_j})\}, j \geq 2$$

kde j je počet kriviek definujúce pásmo $\mathcal{B}$, n je celkový počet kriviek a $\mathcal{I}$ je takáto funkcia:

$$\mathcal{I}(x) = \begin{cases} 1 & \text{ak platí x} \\ 0 & \text{ak neplatí x} \end{cases}$$

Pomocná funkcia BD pre krivku x definuje pomer všetkých pásem zložených z j kriviek, v ktorých sa graf $G(x)$ nachádza, ku všetkým možným j-ticiam kriviek vybraným z n. Samotná funkcia pásmovej hĺbky $\mathcal{BD}$ pre krivku x je definovaná takto:

$$\mathcal{BD}_{n,J}(x) = \sum_{j=2}^{J} BD_n^{(j)}(x), J \geq 2$$

Hĺbka pásma $\mathcal{BD}$ je teda suma všetkých BD pre počet kriviek 2 až J.

Autor článku definujúci pojem pásmová hĺbka navrhol taktiež flexibilnejšiu verziu s použitím pomocnej funkcie MBD (*modified band depth*) [LPR09]. V pravom rade je potrebné zadefinovať si funkciu A, ktorá určí všetky časové body, kedy sa krivka x nachádza v pásme B.

$$A(x, B) = \{t \in I : (t, x(t)) \in G(x) \wedge (t, x(t)) \in B\}$$

V spomínanom článku autori využívajú alternatívnu definíciu funkcie A, do ktorej vstupuje $j+1$ kriviek. Jej význam zostáva rovnaký ako pri našej definícii, avšak náš prístup považujeme za jednoduchší a zrozumiteľnejší. S využitím *Lebesguevoej miery* λ autori ďalej definujú funkciu λ_r, ktorá nám dáva "pomer času, ktorý krivka strávi v pásme":

$$\lambda_r(A) = \frac{\lambda(A)}{\lambda(I)}$$

Nová pomocná funkcia MBD je definovaná nasledovne:

$$MBD_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < ... < i_j \le n} \lambda_r \{A(x, \mathcal{B}(x_{i_1}, ..., x_{i_j}))\}, j \ge 2$$

Ak platí, že $G(x) \subset \mathcal{B}(x_{i_1}, ..., x_{i_j})$, tak funkcia MBD sa degeneruje na BD [SG11].

V našej aplikácii sme sa rozhodli, že budeme pásmo definovať pomocou iba dvoch kriviek, čo nám vzorec výrazne zjednodušilo. Taktiež to implikovalo fakt, že pri výpočte MBD nie je potrebné $\binom{n}{j}^{-1}$, keďže berieme pásma zložené vždy z rovnakého počtu kriviek. Pre naše účely sme si taktiež zjednodušili funkciu λ_r na λ, keďže nepotrebujeme vlastnosť tejto funkcie, ktorá dosahovala to, že MBD pre špeciálny prípad degeneruje na BD. Po týchto úpravách výsledný vzorec pre naše $\mathcal{BD}'$ vyzerá nasledovne:

$$\mathcal{BD}'_n(x) = \sum_{1 \le i_1 < i_2 \le n} \lambda \{A(x, \mathcal{B}(x_{i_1}, x_{i_2}))\}$$

Teraz, keď sme úspešne definovali mieru, podľa ktorej môžme usporiadať funkcie resp. ich krivky, je veľmi ľahké skonštruovať funkčný krabicový diagram.

Nech sú naše funkcie predpovedí $x_1, .., x_n$ usporiadané zostupne podľa $\mathcal{BD}'$. Potom krivka pre funkciu x_1 má najvyššiu pásmovú hĺbku a predstavuje strednú hodnotu pre množinu funkcií, teda niečo podobné ako medián hodnôt pri krabicovom diagrame. Pri konštrukcii funkcionálneho diagramu nám taktiež pomôže koncept centrálneho regiónu [LPS99]. Centrálny región C pre 50% kriviek je pásmo vytvorené z 50% najhlbších kriviek, teda:

$$C_{0.5} = B(x_1, x_2, .., x_{\lceil n/2 \rceil})$$

Vidíme, že Centrálny región $C_{0.5}$ zaobaľuje 50% najhlbších kriviek, a teda sa jedná o akúsi analógiu pre medzikvartálový rozsah (IQR) v klasickom krabicovom diagrame (pozri sekciu 4.2), ktorý ohraničoval 50% centrálnych dát. Zobrazením hraničných bodov tohto regiónu získame obálku myšlienkovo totožnú boxu v krabicovom diagrame (pozri obrázok 5.1b)). Táto myšlienka sa dá použiť ďalej a môžme, tak ako na obrázku 5.1c) , zobraziť taktiež 25%-ný a 75%-ný centrálny región $C_{0.25}$, $C_{0.75}$, avšak kvôli nižšej čitateľnosti grafu sme túto alternatívu nepoužili.

V prípade, že nepotrebujeme identifikovať *outlier*-ov, tak extremálne hodnoty je už veľmi jednoduché získať, pretože ich tvorí pásmo zložené zo všetkých kriviek $B(x_1, .., x_n)$. V opačnom prípade musíme najprv identifikovať *outlier*-ov, ktorých potom vylúčime z výpočtov. Opäť sa využíva myšlienka z klasického krabicového diagramu, kedy sa *outlier* určil pomocou hodnoty $c \times IQR$, kde c bolo zvyčajne 1.5. Hranice sa teda získajú naškálovaním centrálneho regiónu so škálovacím faktorom 1.5 a všetky krivky, ktoré sa v tomto regióne nenachádzajú, budú považované za *outlier*-ov. Test na *outlier*-a vyzerá teda takto:

$$y_i(t) = 1.5 \times x_i(t)$$

$$isOutlier(x) = [G(x) \nsubseteq B(y_1, .., y_{\lceil n/2 \rceil})]$$

Na obrázku 5.1b) môžme vidieť červené prerušované čiary, ktorými sú *outlier*-i znázornené.

V našej práci sme do diagramu pridali ešte jednu krivku pre ľubovoľnú štatistiku vypočítanú z chýb ako napríklad MFE, MAE alebo RMSE (pozri sekciu 2.3). Takto môžme spraviť porovnanie distribúcie chýb s vypočítanou štatistikou.

5.5.4 Porovnanie metód

Tu bude pekný obrázok a obkeci.

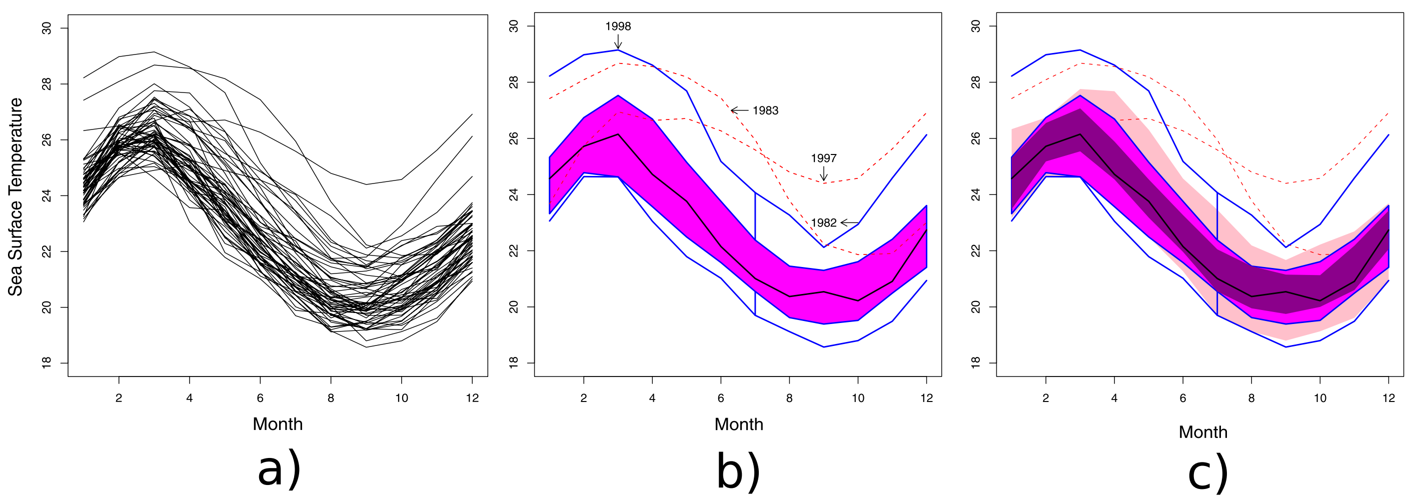
Graf hustoty je presnejsi, ale menej citatelny.

Pruhovy kvantilovy diagram nahradzuje krabicovy diagram.

Funkcionalny krabicovy riesi distribuciu funkcii.

5.6 Návrh farebnej palety

- rainbow a preco zle
- monochrome cez 3 farby
- monochrome 2 farby ostro rozdelene
- ekvalizacia histogramom / boxplotom
- navrh farby pre density, purhovy, functional boxplot



Obr. 5.1: Obrázky z článku *Functional Boxplots* [SG11] a) Funkcie meraní teploty hladiny mora b) Funkčný krabicový diagram c) Rozšírený Funkčný krabicový diagram o centrálne regióny $C_{0.25}$ a $C_{0.75}$

Kapitola 6

Návrh systému a Implementácia

6.1 Návrh systému

6.2 Použité technológie

6.2.1 Java

6.2.2 JavaScript

6.2.3 d3.js

6.3 Extrakcia a spracovanie dát

6.3.1 CSV

6.3.2 Webové zdroje

6.3.3 GRIB

6.3.4 ...

6.4 Konfigurácia systému

6.5 Obrazovka vizualizácie

Kapitola 7

Výsledky

Literatúra

[BL01] G. J. Boer and S. J. Lambert. Second-order space-time climate difference statistics. *Climate Dynamics*, 17(2):213–218, January 2001.

[Bro15] Dr. James D. Brown. *Ensemble Verification Service (EVS) version 5.4 User's Manual*. National Weather Service's Office of Hydrologic Development (OHD), Marec 2015.

[Car94] Daniel B. Carr. A Colorful Variation On Box Plots. *Statistical Computing & Statistical Graphics Newsletter*, 5(3):19–23, December 1994.

[Cas09] Barbara Casati. *Verification of continuous predictands*. Joint Working Group on Forecast Verification Research (JWGFVR), Jún 2009.

[CCM13] Yu-Hsuan Chan, Carlos D. Correa, and Kwan-Liu Ma. The generalized sensitivity scatterplot. *IEEE Trans. Vis. Comput. Graph.*, 19(10):1768–1781, 2013.

[Cen14] Developmental Testbed Center. *Model Evaluation Tools Version 5.0, User's Guide 5.0*. DTC, September 2014.

[CM05] Chamnein Choonpradub and Don McNeil. Can the box plot be improved? *Songklanakarin Journal of Science and Technology*, 27(3):649–657, 2005.

[CWS+08] B. Casati, L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason. Forecast verification: current status and future directions. *Meteorological Applications*, 15(1):3–18, 2008.

[DTC15] DTC Developmental Testbed Center. Model Evaluation Tools. http://www.dtcenter.org/met/users/, Marec 2015. [Prístupné online: 25.3. 2015].

[Few08] Stephen Few. Solutions to the Problem of Over-Plotting in Graphs. *Perceptual Edge*, 2008.

[Fis10] Wolfram Fischer. *Neue Grafiken zur Datenvisualisierung*. Z I M – Zentrum für Informatik und wirtschaftliche Medizin., 2010.

[FJB12] Tressa L. Fowler, Tara L .Jensen, and Barbara G. Brown. *Introduction to Forecast Verification*. 2012.

[Fou15] The R Foundation. The R Project for Statistical Computing. http://www.r-project.org/, Marec 2015. [Prístupné online: 19.3. 2015].

[Fri09] Michael Friendly. Milestones in the history of thematic cartography, statistical graphics, and data visualization. 2009. Na webe: http://datavis.ca/milestones/.

[Gol] Professor Brian Golding. Weather forecasting part 1. http://www.rmets.org/weather-and-climate/weather/weather-forecasting. [Prístupné online: 6.12.2014].

[HN98] Jerry L. Hintze and Ray D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, Máj 1998.

[Ins15] SAS Institute. Statistical Analysis Software. http://www.sas.com, Marec 2015. [Prístupné online: 19.3. 2015].

[JHM+13] Halldor Janetzko, Ming C. Hao, Sebastian Mittelstädt, Umeshwar Dayal, and Daniel A. Keim. Enhancing scatter plots using ellipsoid pixel placement and shading. In *46th Hawaii International Conference on System Sciences, HICSS 2013, Wailea, HI, USA, January 7-10, 2013*, pages 1522–1531, 2013.

[Lab14] NCAR Research Applications Laboratory. *Weather Forecast Verification Utilities*. NCAR, Júl 2014.

[LE11] Dr. Arlene Laing and Dr. Jenni-Louise Evans. *Introduction to Tropical Meteorology 2nd Edition*. UCAR, Október 2011.

[LPR09] Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

[LPS99] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *Ann. Statist.*, 27(3):783–858, 06 1999.

[Lyn07] Peter Lynch. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics 227 (2008) 3431–3444*, Február 2007.

[Mic13] Microsoft. ??? visual basic for applications. `https://`, Január 2013. [Prístupné online: 19.1. 2013].

[Mic15] Microsoft. Microsoft Excel. `https://products.office.com/en-us/excel`, 2015. [Prístupné online: 18.3.2015].

[MTL78] Robert McGill, John W. Tukey, and Wayne A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, Február 1978.

[NCE14] NCEP. Inventory of Data Products on the NOAA Servers. `http://www.nco.ncep.noaa.gov/pmb/products/`, November 2014. [Prístupné online: 10.11.2014].

[Nur03] Pertti Nurmi. *Recommendations on the verification of local weather forecasts*. European Centre for Medium Range Weather Forecasts, Decmeber 2003.

[NWS15] NWS National Weather Service. The Ensemble Verification System (EVS). `http://amazon.nws.noaa.gov/ohd/evs/evs.html`, Marec 2015. [Prístupné online: 23.3. 2015].

[OGJ11] P. Oldenburg, J.Halley Gotway, and T. Jensen. Model Evaluation Tools (MET) verification statistics visualization, 2011. *METViewer*.

[Ope13] Apache OpenOffice. Openoffice.org basic programming guide. `https://wiki.openoffice.org/wiki/Documentation/BASIC_Guide`, Január 2013. [Prístupné online: 19.1. 2013].

[Ope15] Apache OpenOffice. OpenOffice.org Calculate. `https://www.openoffice.org/product/calc.html`, 2015. [Prístupné online: 18.3.2015].

[Pea95] K. Pearson. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Royal Society of London Philosophical Transactions Series A*, 186:343–414, 1895.

[PKR07] Kristin Potter, Joe Kniss, and Richard Riesenfeld. Visual summary statistics. Technical Report UUCS-07-004, University of Utah, 2007.

[Poc11] Matthew Pocernich. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, chapter Appendix - Verification Software, pages 232–240. John Wiley & Sons, Ltd., 2nd edition, December 2011.

[Pot06] Kristin Potter. Methods for presenting statistical information: The box plot. In Hans Hagen, Andreas Kerren, and Peter Dannenmann, editors, *Visualization of Large and Unstructured Data Sets*, volume S-4 of *GI-Edition Lecture Notes in Informatics (LNI)*, pages 97–106. 2006.

[RRT99] Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–287, November 1999.

[SB91] William Stock and John Behrens. Box, Line, and Midgap Plots: Effects of Display Characteristics on the Accuracy and Bias of Estimates of Whisker Length. *Journal of Educational Statistics*, 16(1):1–20, 1991.

[SG11] Ying Sun and Marc G. Genton. Functional Boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, Jún 2011.

[SOA04] Neil C. Schwertman, Margaret Ann Owens, and Robiah Adnan. A simple more general boxplot method for identifying outliers. *Computational Statistics & Data Analysis*, 47(1):165–174, 2004.

[Sol15] Exelis Visual Information Solutions. Interactive Data Language. `http://www.exelisvis.com/ProductsServices/IDL.aspx`, Marec 2015. [Prístupné online: 19.3. 2015].

[Spe52] Mary Eleanor Spear. *Charting Statistics*. McGraw-Hill Book Company, Inc., 1952.

[SS07] Hideaki Shimazaki and Shigeru Shinomoto. A method for selecting the bin size of a time histogram. *Neural Comput.*, 19(6):1503–1527, June 2007.

[Tay01] Karl E. Taylor. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106(D7):7183–7192, April 2001.

[Tay05] Karl E. Taylor. Taylor Diagram Primer. January 2005.

[TM15] Inc. The MathWorks. Matlab the language of technical computing. `http://www.mathworks.com/products/matlab/`, Marec 2015. [Prístupné online: 18.3. 2015].

[Ton05] Phattrawan Tongkumchum. Two-dimensional box plot. *Songklanakarin Journal of Science and Technology*, 27(4):859–866, 2005.

[Tuf83] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 1983.

[Tuk77] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[UCA15] UCAR. NCAR Command Language. `http://www.ncl.ucar.edu/`, Marec 2015. [Prístupné online: 19.3. 2015].

[Vas98] Tim Vasquez. *Observer Handbook*. International Weather Watchers, 1995, 1998.

[VWvH+07] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. Manyeyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, November 2007.

[WCSDG+08] Joseph B. Klemp William C. Skamarock, Jimy Dudhia, David O. Gill, Dale M. Barker, Michael G. Duda, Xiang-Yu Huang, Wei Wang, and Jordan G. Powers. *A Description of the Advanced Research WRF Version 3*. National Center for Atmospheric Research, Jún 2008.

[Wei14] Eric W. Weisstein. Statistical Median. `http://mathworld.wolfram.com/StatisticalMedian.html`, 2014.

[WMO94] WMO. *A GUIDE TO THE CODE FORM FM 92-IX Ext. GRIB Edition 1*. WMO, Máj 1994.

[WMO03] WMO. *Introduction to GRIB Edition1 and GRIB Edition 2*. WMO, Jún 2003.

[WS12] Hadley Wickham and Lisa Stryjewski. 40 years of boxplots. Technical report, had.co.nz, 2012.

[Yoa88] YoavBenjamini. Opening the box of a boxplot. *The American Statistician*, 42(4):257–262, November 1988.