TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Tvorba specializované webové aplikace pro statistickou analýzu dat

Bakalářská práce

Studijní program: B2612 – Elektrotechnika a informatika
Studijní obor: 2612R011 – Elektronické informační a řídicí systémy

Autor práce: **Ladislav Hochman**
Vedoucí práce: Ing. Vratislav Žabka, Ph.D.

Liberec 2019

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Creating a specialized web application for statistical analysis of data

Bachelor thesis

Study programme: B2612 – Electrical Engineering and Informatics
Study branch: 2612R011 – Electronic Information and Control Systems

Author: **Ladislav Hochman**
Supervisor: Ing. Vratislav Žabka, Ph.D.

Liberec 2019

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Zadání bakalářské práce

Tvorba specializované webové aplikace pro statistickou analýzu dat

Jméno a příjmení: **Ladislav Hochman**
Osobní číslo: M15000093
Studijní program: B2612 Elektrotechnika a informatika
Studijní obor: Elektronické informační a řídicí systémy
Zadávající katedra: Ústav mechatroniky a technické informatiky
Akademický rok: **2018/2019**

Zásady pro vypracování:

1. Seznamte se s jazykem R a s prací v Rstudiu. Naučte se načítání časových řad, jejich základní statistickou analýzu a zobrazování výsledků pomocí různých typů grafů.
2. Zpracujte data z měření elektrických veličin. Na těchto datech otestujte pokročilejší statistické postupy pro analýzu průběhu jednotlivých veličin. Připravte a otestujte algoritmy, které automaticky prověří vlastnosti úseků datové řady o zadané délce. Informace statisticky zpracujte.
3. Vytvořte webovou aplikaci v prostředí Shiny, která umožní online zpracování a grafickou vizualizaci vybraných dat z oboru elektrotechniky.

Rozsah grafických prací:	dle potřeby dokumentace
Rozsah pracovní zprávy:	30–40 stran
Forma zpracování práce:	tištěná/elektronická

Seznam odborné literatury:

[1] Šmilauer Petr, Moderní regresní metody. Biologická fakulta JU, České Budějovice, 2007.

[2] Lawrence Leemis. Learning Base R. Lightning Source, 2016. ISBN 978-0-9829174-8-0.

[3] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.4. https://CRAN.R-project.org/package=shiny.

Vedoucí práce:	Ing. Vratislav Žabka, Ph.D. Ústav mechatroniky a technické informatiky
Datum zadání práce:	10. října 2018
Předpokládaný termín odevzdání:	30. dubna 2019

L. S.

prof. Ing. Zdeněk Plíva, Ph.D.
děkan

doc. Ing. Milan Kolář, CSc.
vedoucí ústavu

V Liberci 10. října 2018

Prohlášení

Byl jsem seznámen s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu TUL.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Bakalářskou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Současně čestně prohlašuji, že texty tištěné verze práce a elektronické verze práce vložené do IS STAG se shodují.

30. 4. 2019 Ladislav Hochman

Poděkování

Na tomto místě bych rád poděkoval Ing. Vratislavu Žabkovy, Ph.D. za cenné připomínky a odborné rady, kterými přispěl k vypracování této bakalářské práce.

Abstrakt

Bakalářská práce se zabývá analýzou dat elektrických veličin. Podstatou práce je nalezení vhodné délky intervalu pro ukládání časové řady. Pro práci s daty slouží webová aplikace vytvořená v prostředí Shiny jazyka R. Aplikace umožnuje analyzovat vstupní data. Dále automaticky prověřovat časový úsek a vytvořit příslušný graf. Během práce jsou data rozdělena do intervalů. Změna délky intervalu má vliv na procentuální výskyt rozdělení. Mezi zkoumaná rozdělení patří normální, logistické a weibullovo. Intervaly, které obsahují rozdělení, lze efektivně uložit. Jednou z možností je ukládání dat pomocí směrodatné odchylky a průměru. Podmínkou pro popis intervalu pomocí rozdělení je p hodnota. Počet intervalů s rozdělením je zaznamenán. Nalezením optimální délky intervalu je možné docílit menšího počtu uložených dat a tím zefektivnit práci s daty.

Klíčová slova: analýza dat, aplikace, délka intervalu, jazyk R, p hodnota, rozdělení, Shiny

Abstract

The bachelor thesis deals with the analysis of electrical quantities data. The subject of this thesis is to find a suitable length of time interval for saving time series. A web application created in the Shiny R environment is used to work with data. The application allows analyzing input data. Next, automatically scan the time slot and create the appropriate graph. During the work, the data is divided into intervals. Changing the interval length affects the percentage distribution. The distributions examined include normal, logistic, and Weibull. Intervals that contain splits can be saved effectively. One option is to store data using standard deviation and diameter. The prerequisite for the description of the interval is p value. The number of distribution intervals is saved. By finding the optimum interval time, it is possible to achieve a smaller number of stored data and thus to make the work with data more efficient.

Keywords: application, data analysis, distribution, interval length, R language, p value, Shiny

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Bakalářské práce se zabývá efektivním ukládáním měřených veličin. Nalezením vhodné délky intervalu pro ukládání časové řady. Během procesu ukládání je zapotřebí značné množství místa. Právě z tohoto důvodu je snahou proces ukládání zefektivnit. Jednou z možností je interpretace intervalu pomocí maxima, minima a průměru. Při ukládání však může dojít ke ztrátě informace vlivem skokových změn měřených dat. Z tohoto důvodu je vhodné měřená data ukládat pomocí směrodatné odchylky a průměru.

Cílem práce je seznámení se se statistickým jazykem R. Využití pokročilých statistických metod pro automatické prověření vlastností úseků datové řady o zadané délce. Následné statistické zpracování analýzy průběhu jednotlivých veličin. Posledním cílem je vytvoření webové aplikace v prostředí Shiny pro vizualizaci vybraných dat.

Teoretická část bakalářské práce se věnuje v první kapitole možnostem efektivního ukládání. Druhá kapitola je věnována popisu jazyka R, programu RStudio a využitých knihoven. Kapitole 3 pojednává o funkci boxplot. Boxplot umožňuje snadnou interpretaci chování celé řady nebo intervalu. Udává průměr a poukazuje na odlehlé hodnoty. V kapitole 4 jsou zmíněny typy porovnávaných rozdělení. Mezi použitá patří normální, logistické a weibullovo rozdělení. Kapitola 5 se zabývá nejvíce využívanými grafy v bakalářské práci. Jedná se o CDF graf a histogram.

Na začátku praktické části v kapitole 6 jsou popsány vstupní data. Jedná se o fyzikální veličiny napětí, proudu, frekvence a výkonu. Dále je zde vykreslen jeden konkrétní interval. Interval reprezentuje časový úsek fyzikální veličiny. Na intervalu je popsán význam p hodnoty. P hodnota slouží pro vyhodnocení jednotlivých intervalů na odhadovaný výskyt rozdělení. Tedy zdali sledovaný interval lze interpretovat pomocí některého ze tří zmíněných rozdělení. Pokud ano, je takový interval zaznamenán. Právě tomuto účelu je věnována kapitola 8. Jedná se o stěžejní část práce.

Kapitola 7 slouží pro popis webové aplikace. Podstatou vzniku aplikace je usnadnění práce s daty. Aplikace uživateli umožní automaticky prověřovat jednotlivé intervaly. Dále posuzovat vzájemné chování intervalů. Program umožní tento proces ještě více zjednodušit, posuzováním intervalů o různé časové délce. Aplikace posoudí, zdali intervaly odpovídají zkoumaným rozdělením. Nakonec vypíše procentuální úspěšnost pro jednotlivá rozdělení v porovnání s počtem intervalů.

Hlavní část bakalářské práce kapitola 9 analyzuje měřená data a rozděluje je na série intervalů. Měřené intervaly jsou následně podrobovány testu na jednotlivá rozdělení. Počet kladných výskytů je uložen a procentuálně porovnáván s celkovým množstvím intervalů. Měřená data jsou v této fázi rozdělována na různě dlouhé časové řady. Výsledkem je procentuální porovnání různě dlouhých intervalů na výskyt jednotlivých rozdělení.

Nakonec se bakalářská práce zabývá nalezením optimální délky intervalů. Nahrazením intervalu jen několika body je možné zefektivnit práci s daty. Z tohoto důvodu je zmíněna možnost ukládání dat pomocí směrodatné odchylky a průměru.

1 Možnosti efektivního ukládání

Existuje mnoho variant, jak zefektivnit ukládání hodnot. V současnosti je využito maximum, minimum a průměr. V budoucnu je možné využít směrodatnou odchylku společně s průměrem. To by výrazně zmenšilo množství nutných ukládaných dat a zjednodušilo analýzu těch zbývajících. Dobrým příkladem je identifikaci odlehlých hodnot. Ty v praxi mohou být předzvěstí zkratu elektrické veličiny a tím umožní předvídat možné selhání zařízení.

Jednou ze slabin průměru je ztráta informace o odlišnosti hodnot uvnitř zkoumaného intervalu. V případě, že jedna hodnota nabyde vysoké hodnoty a druhá naopak nízké, průměrem bude vyvážená hodnota. Takováto mylná informace by celkově zkreslila představu o chování řady. Průměr tedy nic neříká o variabilitě uvnitř zkoumaného intervalu. Ve statistice však existují i jiné ukazatele variability. [1]

Jedním z nich je rozptyl. Ten je chápán jako průměrná čtvercová odchylka od průměru. Naznačuje, jak moc jsou hodnoty ve sledovaném intervalu odchýlené nebo vzdálené od průměru. Z důvodu možného vniku záporných odchylek jsou hodnoty umocněny na druhou. Problém rozptylu spočívá v jeho horší vizualizaci což znesnadňuje jeho následnou interpretaci. [1]

Dalším způsobem, jak popsat chování řady je směrodatná odchylka. Jednodušeji ji lze stanovit jako průměrnou odchylku od průměru. Obecně se jedná o druhou odmocninu z rozptylu. Ta je mnohem přesnější než samotný průměr a na rozdíl o směrodatné odchylky také lépe představitelná. V případě že bychom chtěli vědět kolik procent průměru představuje směrodatná odchylka, použijeme variační koeficient. [1]

Variační koeficient vypovídá o relativním významu průměrné odchylky od průměru. Jedná se o bezrozměrnou veličinu většinou vyjádřenou v procentech. Variační koeficient se užívá při porovnávání různě velikých intervalů. [1]

Běžnou variantou je ukládání dat pomocí maxima, minima a průměru. Cílem této práce je nalezení vhodné délky intervalu. Délka časového intervalu mý vliv na procentuální výskyt rozdělení. V případě, že interval obsahuje některé rozdělení, je možné časový úsek nahradit. U intervalů s normálním rozdělením lze jednotlivé úseky ukládat pomocí směrodatné odchylky a průměru. U ostatních rozdělení v bakalářské práci by mohlo dojít ke ztrátě důležité informace.

2 Popis jazyka R a využitých knihoven

2.1 R

R je programovací jazyk vytvořený pro statistickou analýzu dat. Zároveň je také schopný jejich vizuálního zobrazení. Jazyk R vychází z komerčního jazyku S. Na rozdíl od něho byl uveden na trh pod svobodnou licencí. Díky své bezplatnosti je hojně využíván v mnoha oblastech statistiky. Tomu napomáhá i fakt neustálého rozšiřování funkcí pomocí knihoven, které jsou označovány jako balíčky. Ty umožňují využití specializovaných statistických nástrojů, import nebo export dat případně grafické zobrazení. Ačkoliv je R volně šiřitelný, na jeho vývoji pracuje celá řada komerčních společnosti nabízející technickou podporu. [2]

Jazyk R disponuje celou řadou statistických případně grafických technik jako lineární nebo nelineární modelování, analýza časových řad a mnoho dalších [3]. Jejich dalšímu rozšíření vypomáhá i aktivní R komunita, která se dále stará o aktualizaci stávajících balíčků. Tyto balíčky pak umožňují například vytvářet dynamické i interaktivní grafy, které jsou využity v této práci. Na rozdíl od ostatních statistických programů je jazyk R silně objektově orientovaný jazyk. [2]

Jazyk R komunikuje pomocí příkazového řádku, existuje však mnoho frontendů s grafickým rozhraním, které práci s ním zjednodušují. Jedním z nich je RStudio. [2]

2.2 RStudio

Jedná se o volně šiřitelné otevřené vývojové prostředí pro R. Obsahuje konzoly, nástroje pro grafy nebo editor pro zvýraznění syntaxe, který podporuje přímé spuštění kódu [4]. RStudio je k dispozici ve volně šiřitelných nebo komerčních vydáních. Ty fungují na všech současných operačních systémech. [5] Práce s nimi může probíhat buď na lokálním zařízení nebo na vzdáleném serveru. V prvním případě je program spuštěn jako běžná desktopová aplikace. U druhé možnosti se zajistí přístup ke službě pomocí webového prohlížeče [6]. Samotné základy RStudia byly vytvořeny v programovacím jazyce C++. Na jeho tvorbu však byly využity i jiné jazyky. [7]

Mezi zajímavé funkce tohoto vývojového prostředí patří zvýraznění a doplnění kódu. Dále snadné vyhledávání v kódu, integrovaná nápověda nebo procházení historie [6]. V RStudiu se nalézá množství důležitých balíčků, které byly využity v této práci. Jedním z nich je interaktivní webová technologie Shiny.

2.3 Shiny

Jedním z cílů této práce bylo vytvoření webové aplikace, čehož bylo dosaženo pomocí již zmíněného RStudia a také prostřednictvím jeho balíčku Shiny. Ten umožňuje snadné vytváření interaktivních webových aplikací přímo z prostředí jazyka R [8]. Z hlediska programátora tedy není nutné žádných dalších programovacích jazyků jako jsou HTML nebo JavaSctript. Úplná webová aplikace tak může být budována čistě v jazyce R. Tato skutečnost značně snižuje nároky na znalosti programátorů a usnadňuje tvorbu webových aplikací. [9]

Typická webová aplikace se skládá z několika prvků uživatelského rozhraní. Mezi takové členy patří tlačítka nebo zaškrtávací políčka. Každý z těchto prvků může být ovlivněn

například stisknutím tlačítka. Po stisknutí tlačítka se spustí kód programu. Na základě toho se může změnit stav webové aplikace jako je načtení informací ze serveru nebo vykreslení obrázku v aplikaci. [9]

Shiny aplikace obvykle obsahují krom již zmíněného uživatelského rozhraní také serverovou logiku. Uživatelské rozhraní slouží pro zobrazení tlačítek nebo tabulátorů se kterými uživatelé pracují [8]. Na druhé straně server obsahuje pracovní kód vývojáře aplikace. Většinou se jedná o dvě samostatné části. Nicméně pokud je program příliš malý, může být výhodné sloučit oba soubory do jednoho a pracovat pouze s jedním souborem. [9]

Mezi zajímavé funkce Shiny aplikací patří okamžitá reakce změn nastavení vstupů vyvolané uživatelem. Výstupy jsou automaticky aktivní a upraví se včetně tabulek. Tyto změny se provádí, aniž by uživatel musel znovu načíst prohlížeč. Další důležitou funkcí je reaktivní programovací program. Ten do určité míry eliminuje chybový kód vzniklý během zpracování událostí. To vývojáři umožňuje snáze rozpoznat místo chyby a problém vyřešit. V neposlední řadě je důležité zmínit, že samotnou webovou aplikaci lze vytvořit pouze s několika řádky kódu. Tento fakt výrazně šetří čas vývojářů aplikací, což ostatně platí pro veškerou práci v jazyce R [10].

2.4 Fitdistrplus

Ve statistice je určování rozdělení dat běžným jevem. Obecně spočívá v nalezení odhadů parametrů určitého rozdělení u náhodné proměnné. K tomuto účelu slouží funkce fitdistr implementovaná v balíčku fitdistrplus. Funkce fitdist využívá pro zjištění parametrů rozdělení metodu maximálního odhadu vhodnosti. [11]

Tato shoda modelu s pozorováním, anglicky goodness of fit vyjadřuje, s jakou určitostí statistický model odpovídá zkoumaným vzorkům. [12]

3 Boxplot

Box-plot česky krabicový graf je jednoduchý nástroj pro vizualizaci a posouzení skupiny dat využívaný ve statistice. Pro svou jednoduchost a srozumitelnost je využíván u automatických zkušebních a měřících zařízení, které uchovávají a vyhodnocují nasbírané údaje. Tato funkce umožňuje popis chování dat pomocí kvantilů přesněji kvartilů. Kvantily vyjadřují jednotlivé hodnoty. Rozdělením uspořádaného souboru hodnot na předem daný počet stejně obsažených částí vzniká kvantil. Kvantil lze obecně vyjádřit pomocí výrazu „x“ a jeho dolního indexu „p“ označujícího procentní kvantil. [13]

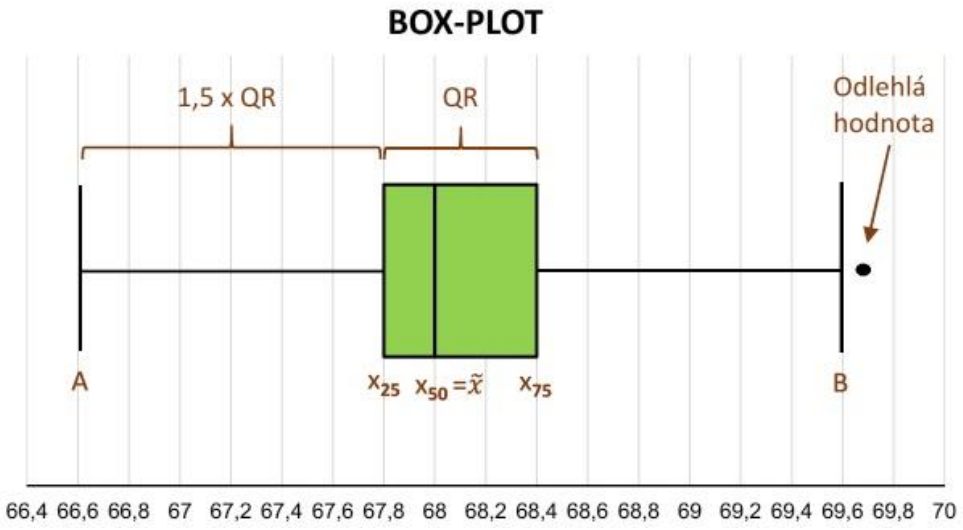
Mezi nejpoužívanější kvantily patří kvartily. Kvartily rozdělí zkoumaný vzorek na čtyři části. Každý úsek obsahuje 25 % z celkového souboru. Jednotlivé kvartily jsou často označovány jako Q1, Q2 a Q3. První kvartil x_{25} značí dolní kvartil. Druhý kvartil x_{50} vyjadřující též medián, může být také označován jako x̃. Třetí kvartil x_{75} označuje horní kvartil. Ve statistických programech, například v jazyku R, mohou být popsány jako Q_{25}, Q_{50} a Q_{75}. [13]

Pro výpočet se zkoumaný vzorek seřadí podle velikosti. Obvykle od nejnižší po nejvyšší. Medián se bude v tomto souboru hodnot nalézat uprostřed. V případě, že má skupina dat sudý počet hodnot, je medián vypočten jako průměr dvou prostředních hodnot. [13]

3.1 Konstrukce krabicového grafu

Vytvoření boxplotu ze zadaných hodnot je možné u většiny běžných statistických programů. Jedním z takových příkladů je RStudio pro jazyk R. Na samotnou tvorbu krabicového grafu však žádný speciální statistický program být nemusí. Na jeho sestrojení stačí tabulkové programy, které bývají běžně dostupné coby součástí kancelářských balíků. [13]

Pro sestrojení krabicového grafu je nutné určit kvartilové rozpětí označováno QR. Kvartilové rozpětí lze vyjádřit jako rozdíl mezi horním a dolním kvartilem. Kvartilové rozpětí je následně využito pro určení koncových bodů paprsků grafu. Kde první bod A určíme jako 1,5násobek hodnoty QR, který je odečten od velikosti prvního kvartilu. Určení druhého bodu B je stanoveno přičtením 1,5násobku QR ke třetímu kvartilu. To jsou linie vycházející ze střední části diagramu, jsou někdy označovány jako tzv. vousy. Vzdálenost mezi jednotlivými body indikuje stupeň rozptylu, odborně disperzi, dále ukazuje odlehlé hodnoty. Vše je znázorněno na Obr. 3-1. [13]



Obr. 3-1 Zobrazení rozptylu boxplotu [14]

3.2 Využití Box-Plotu

Krabicový graf má mnoho funkcí. Mezi jedny z nejdůležitějších patří identifikace odlehlých hodnot, cizím slovem outliers. Dále umožňuje rozpoznat symetrie u konců rozdělení. Jednou z mnoha výhod je také schopnost porovnat rozptyl u dvou a více uskupení vzorků. Popřípadě zhodnotit způsobilost procesů. [13]

3.2.1 Identifikace odlehlých hodnot

Odlehlé hodnoty mohou při klasickém zpracování dat zkreslit výsledné charakteristiky. Mezi výsledné charakteristiky patří průměr, rozptyl nebo index způsobilosti aj. Pro zamezení ovlivnění špatnou interpretací lze krabicový graf využít jako nástroj pro identifikaci odlehlých hodnot. Právě schopnost boxplotu identifikovat odlehlé hodnoty je často využívána v této práci. Při snaze nahradit celý interval pouze několika málo body, je rozpoznání a uložení odlehlých hodnot klíčové. [13]

Odlehlé hodnoty mohou být někdy nežádoucí, kvůli zhoršení popisu chování systému. Vznik takových odlehlých hodnot však nemusí být vždy způsoben nestabilitou procesu. Ani nemusí být příčinou nízká způsobilost procesu jako takového. Původce vzniku odlehlých hodnot můžou být technické nebo technologické důvody. Dobrým příkladem může být elektrický zkrat na tiskárně, jejichž hodnotami se tato práce zabývá. Ten může sloužit pro predikci budoucích problémů. Jeho vznik se projeví na souboru dat právě jako odlehlá hodnota. Může tak být reálným odrazem skutečného stavu tiskárny. Z toho je patrné, že odlehlé hodnoty v sobě nesou důležitou informaci. Jejich přínos tak může být užitečný. [13]

Příkladem nežádoucích odlehlých hodnot je případ, kdy je součástí takového systému člověk. V takovém případě může za podobnými chybami stát i selhání lidského faktoru při zápisu hodnot. Chybný zápis tak může při zpracování dat výrazně pozměnit nebo ovlivnit celý výsledek. [13]

Z tohoto důvodu je boxplot velice užitečným nástrojem pro identifikaci odlehlých hodnot, které bývají v jazyce R obvykle zobrazovány jako puntíky nebo křížky na vnější straně paprsků.

3.2.2 Posouzení asymetrie

V případě, že data odpovídají Normálnímu rozdělení, nalézá se mediánová čára uprostřed krabicového grafu. Pokud tato čára atakuje hranici dolního nebo naopak horního kvartilu, lze nastolit předpoklad, že zkoumaná data odpovídají i jinému rozdělení.

V případě technologické výroby, může být asymetrie procesu způsobena větší odchylkou od Mediánu u hůře zpracovatelných materiálů. Pokud bude měřenou veličinou hmotnost, tak u materiálů, které část své hmotnosti ztratí během zpracování dojde k odklonu od očekávané hmotnosti. [13]

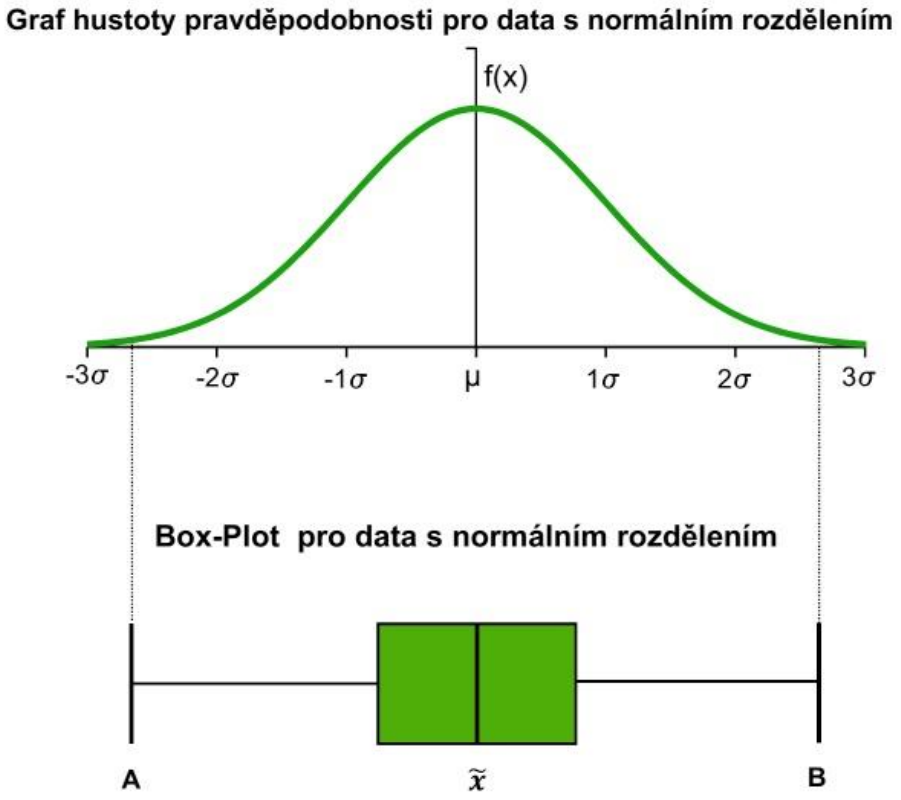
V případě elektrické sítě může být asymetrie pozorována u hodnot napětí. Kde v třífázové distribuční soustavě je všeobecně způsobena nerovnoměrným zatížením jednofázovými zátěžemi. Buď ve dvou nebo ve všech třech fázích. [15]

3.2.3 Porovnávání rozptylů dvou a více souborů

Pro budoucí výpočty je nejprve nutné zjistit směrodatnou odchylku. Pro její zjištění je nejprve nutné vypočítat rozptyl ze zkoumaného úseku dat.

Pro objektivní posuzování zkoumané řady je třeba vypočíst více rozptylů z rozdílných souborů dat. Samotná hodnota rozptylu o chování trendu nevypovídá. Aby byl výsledek relevantní musí být u všech sledován stejný parametr u srovnatelného procesu. V případě této práce změna délky intervalu u shodné veličiny.

Krabicové grafy se také hodí pro vizuální interpretaci chování celé řady. Tuto vlastnost podtrhuje samotná konstrukce boxplotu. [13]



Obr. 3-2 Graf hustoty pravděpodobnosti [16]

Pokud se Mediánová čára uvnitř boxplotu nachází uprostřed, znamená to, že je hodnota mediánu blízká zkoumané hodnotě. Z toho vyplívá snazší popis chování řady, jak naznačuje Obr. 3-2. Pro snazší ilustraci Obr. 3-2 neobsahuje odlehlé hodnoty. Vzdálenost rozptylu je dána koncovými body. S menší vzdáleností koncových bodů klesá rozptyl.

3.3 Boxplot v jazyce R

Pro zobrazení boxplotu je vhodné využít funkci ggplot2. Vnitřní nástroj funkce ggplot2 se jmenuje geom. Jednou z možností příkazu geom je ovládání barvy boxplotu. Umožňuje však i jiná přizpůsobení. Mezi prováděné úkony patří práce s vícečlennou skupinou, nastavování šířky nebo zobrazování průměrné hodnoty. [17]

4 Typy porovnávaných rozdělení

Pro posouzení vlastností intervalů byly vybrány tři rozdělení popsané v následující kapitole. Jedná se o Normální, Logistické a Weibullovo rozdělení. Existují i jiná rozdělení jako je Gama nebo Logaritmicko-normální. Těmito distribucemi se však bakalářská práce nezabývá.

4.1 Normální rozdělení

Normální neboli Gaussovo rozdělení pravděpodobnosti je jedno z nejpoužívanějších spojitých rozložení četností výskytu určitého jevu. Pro toto rozdělení je typický zvoncovitý tvar. Normální rozdělení má jen jeden vrchol, je jednomodální. Jedná se o symetrické rozdělení s hodnotami soustředícími se okolo jeho průměru. Jedním ze zajímavých údajů pro statistiku je, že aritmetický průměr je zároveň mediánem i modem. Jeho výskyt je častý pro mnoho sociálních, psychických, případně biologických jevů. Pojem „Normální“ v tomto případě vychází z historicích pramenů a znamená "řídící se předpisem, zákonem případně modelem". Gaussovo rozdělení je ostatně jako většina jiných statistických rozdělení především matematickým ideálem a myšlenkovým modelem. Ve statistice je to však model velice významný. Podobnost s normálním rozdělením je zkoumána u většiny měřených hodnot v této práci. [18] [19]

4.1.1 Gaussova křivka

Gaussova křivka je grafem hustoty pravděpodobnosti normálního rozdělení. Gaussova křivka bývá funkce o dvou proměnných. Prvním parametrem je střední hodnota, druhým rozptyl. U Gaussovy křivky leží střední hodnota pod jejím vrcholem. Jedná se o symetrickou funkci. V místě střední hodnoty se nalézá extrém. Ten udává, že v případě opakování náhodného pokusu, se budou výsledné hodnoty nejvíce vyskytovat právě v okolí střední hodnoty. Pakliže se opakovaný pokus řídí normálním rozdělením. Přibližně stejně častý výskyt u výsledků vychýlených pod nebo nad střední hodnotu stanovuje symetrie křivky. Parametr rozptylu určuje, do jaké míry se křivka přimyká ke střední hodnotě. Vizuálně si to lze představit tak, že čím nižší je tato proměnná, tím se graf jeví „ostřejší“. [19]

4.1.2 Význam Gaussovy křivky

Dobrým příkladem pro představu významu této křivky sloužící pro popis hustoty pravděpodobnosti je konstrukce histogramu. Sloupcového diagramu tvořeného obdélníky. Ty mají pevně stanovou šíři základny na horizontální ose odpovídající vybranému intervalu. Četnost výskytu v dané třídě odpovídá pro změnu výšce obdélníka. Stanovení hustoty pravděpodobnosti normálního rozdělení se provede tak, že u zaznamenaných výsledků pokusu s normálním rozdělením, bude šířka jedné kategorie snižována limitně k nule. V případě, kdy sledovaný znak odpovídá normálnímu rozdělení, je rozumné stanovit odhad střední hodnoty, popřípadě rozptylu. [19]

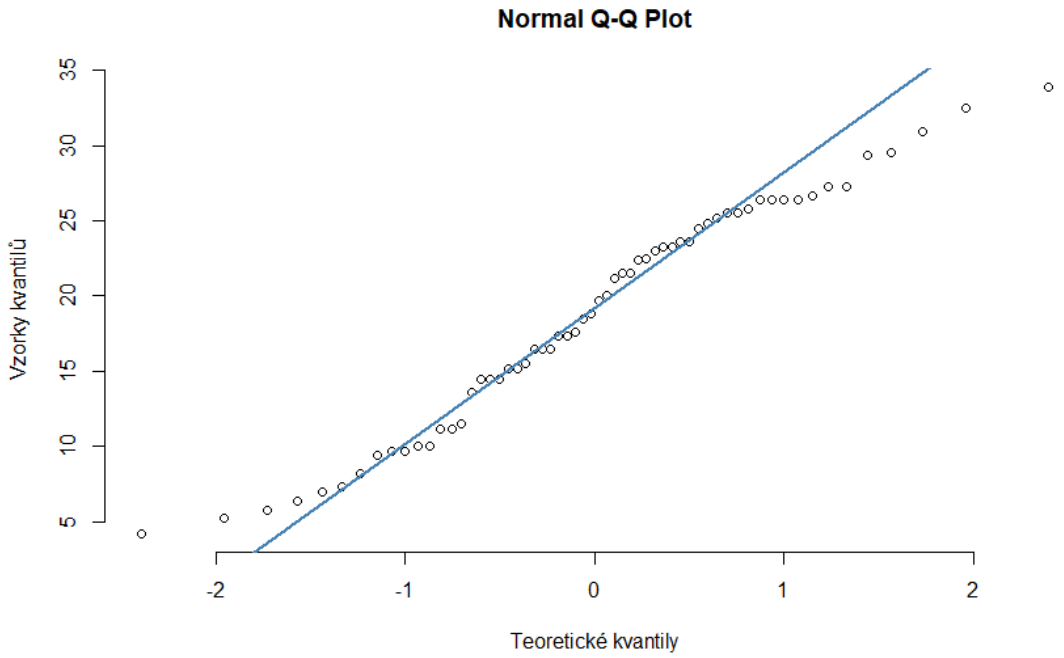
Dalším zajímavým parametrem je směrodatná odchylka, kterou lze získat druhou odmocninou z již získaného rozptylu. Z odhadu střední hodnoty pro konečný počet pokusů lze získat aritmetický průměr. Rovněž lze určit medián i modus. U normálního rozdělení bývají číselně shodné. Často se proto vypočítá jen jedna z veličin. Většinou aritmetický průměr, jehož výpočet je relativně snadný. Podle aritmetického průměru se následně stanovuje odhad střední hodnoty. Přesněji v jakém pásmu kolem aritmetického průměru se

nachází skutečná střední hodnota s předem stanovenou pravděpodobností. O takovémto vzniklém intervalu se někdy hovoří jako o konfidenčním intervalu. Mnohem častěji je však užíván název interval spolehlivosti. [19]

Visuálně by úzký interval spolehlivosti odpovídal vysokému štíhlému grafu. Zvětší-li se rozsah výběru, je křivka hustoty rozložení pravděpodobnosti užší a zároveň vyšší. S přibývajícím počtem měření, se zmenšuje interval spolehlivosti a spolu s tím zvyšuje jistota odhadu skutečné střední hodnoty. V případě obráceného postupu, se k předem dané šířce intervalu spolehlivosti dopočítá, kolik měření je nutné provést. [19]

4.1.3 Testy normality

Některé statistické metody mohou být aplikovány jen u specifických souborů vycházející z normálního rozdělení. Pro zjištění, zdali dané rozdělení dat splňuje kritéria normálního rozdělení slouží test normality. Takovéto testy jsou ve větší či menší míře implementovány u většiny stávajících statistických softwarů. Za dobrý příklad může sloužit jazyk R využitý v této práci. K visuálnímu ohodnocení normality rozdělení dat může sloužit histogram nebo již zmíněný boxplot. Pro testování jiných rozdělení je vhodné použití Q-Q grafu Obr. 4-1 nebo P-P grafu. Ty jsou mnohem přesnější. Oba tyto grafy jsou zastoupeny v jazyce R. Hrubého odhadu míry normality zkoumaných vzorků lze dosáhnout prostým srovnáním mediánu a aritmetického průměru. [19] [20]



Obr. 4-1 Q-Q graf

4.2 Logistické rozdělení

Logistické rozdělení je spojitou distribucí pravděpodobnosti [21]. Jedná se o symetrické rozdělení s jedním vrcholem. Logistická distribuce bývá často využita pro růstové modely například růst populace [22]. Jeho kumulativní distribuční funkce je známá jako logistická

funkce, která je někdy označována jako logistická křivka [23]. Ta se objevuje v určitém typu regrese, konkrétně logistické regresi a dopředných neuronových sítí. Tvary logistického a normálního rozdělení jsou si velmi podobné, akorát s tím rozdílem, že při visuálním porovnání se logistický graf jeví špičatější. Tedy oproti normálnímu rozdělení má vyšší koeficient špičatosti. Jednou z mnoho aplikací je modelování životních údajů pomocí logistické regrese. [21] [22]

4.2.1 Logistická regrese

Jednou z nejběžnějších aplikací logistického rozdělení je logistická regrese. Logistická regrese neboli logistický regresivní model je název pro regresní model s binární závisle proměnnou. Ta se využívá pro modelování kategoricky závislých proměnných. Pro takovéto modelování je zde využita logistická funkce ve své základní formě. Další možným využitím je modelování spojitých proměnných například přijmu. Příkladem závislých proměnných jsou binárně závislé proměnné jako je volba ano/ne. Model binární logistické regrese může být rozšířen na více než dvě závislé proměnné. Takový model nabývá více možných stavů. Jedná se o ordinální logistickou regresi. Existují modely se třemi nebo čtyřmi proměnnými. V takovém případě se modeluje pravděpodobnost výstupu z hlediska vstupu a neprovádí se statistická klasifikace. Model tedy není klasifikátorem. [24] [25]

Logistická regrese je využita v různých oblastech, včetně strojního učení, většiny lékařských oborů nebo společenských věd. Závislá proměnná představuje například nepřítomnost nebo naopak přítomnost choroby. Logistická regrese pak umožňuje modelovat například vznik srdeční choroby jako funkci několika parametrů mezi něž patří pohlaví nebo věk. Logistická regrese určuje, které veličiny se významně podílely na úspěšnosti daného výrobku. Určení podílu jednotlivých veličin se využívá v průmyslu při posuzování úspěšnosti nebo neúspěšnosti produktu. Logistická regrese tak může predikovat pravděpodobnost vlivu nějakého jevu na výrobek. [26]

4.3 Weibullovo rozdělení

Relativně hojně aplikovaným teoretickým rozdělením pravděpodobnosti využívaným při řešení problémů v oblasti spolehlivosti je Weibullovo rozdělení. Jedná se o zobecněné exponenciální rozdělení, navržené pro popis životnosti materiálů. [27] Jeho význam spočívá v modelování dat, u nichž nebere v potaz, zdali je intenzita poruch klesající, rostoucí nebo setrvává na současné hodnotě. [28] Weibullovo rozdělení je v porovnání s jinými distribucemi přizpůsobivější hodnotám s velikým rozptylem. Dále nepředpokládá stále stejné riziko výskytu sledované události v čase. Naproti tomu uvažuje jednotvárnou rizikovou funkci. Takováto funkce bude buď monotónně klesající nebo rostoucí spolu s časem. Z toho vyplývá jeho širší využití v praxi. [27]

Weibullovo rozdělení je definováno pomocí dvou veličin, tvaru hustoty pravděpodobnosti a škály hodnot. Samotný tvar rizikové funkce náhodné veličiny zásadním způsobem závisí na tvaru hustoty pravděpodobnosti. Pokud je jeho hodnota menší než jedna, je riziková funkce náhodné veličiny monotónně klesající. Jeli tvar hustoty pravděpodobnosti roven jedné je tato funkce konstantní. V případě, kdy je jeho hodnota větší než jedna, je riziková funkce náhodné

veličiny monotónně rostoucí. Z toho vyplívá, že rizikovou funkci tedy není možné specifikovat jako zároveň rostoucí i klesající. [27]

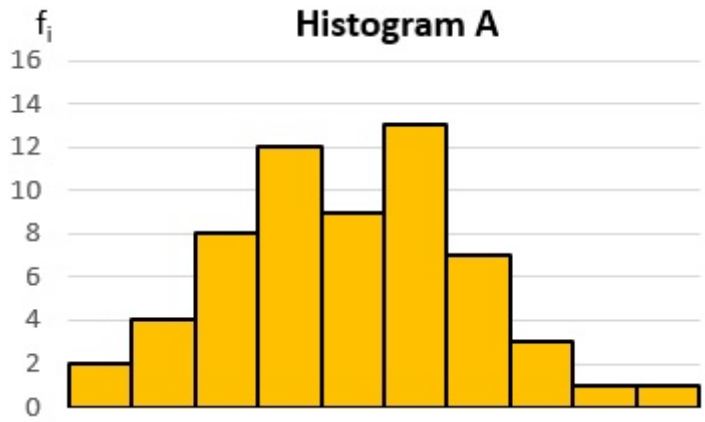
Jeho uplatnění lze nalézt u medicínského výzkumu monitorující přežití pacientů. Za zmínku stojí aplikace při modelování systémů pouze v období stabilního života. [27] V neposlední řadě je schopný modelovat dobu do výskytu událostí u systémů, které jsou v období častých poruch. Tedy tam kde se projevuje únava materiálu případně mechanické poškození. [29]

5 Další využité grafy

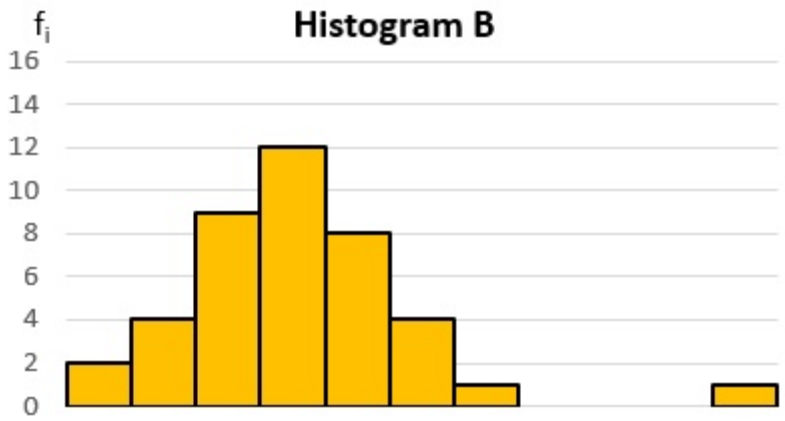
5.1 Histogram

Histogram je jedním z nejpoužívanějších grafických nástrojů. Slouží pro vizualizaci intervalových a poměrových hodnot. Histogram je relativně jednoduchý nástroj, který umožňuje základní analýzu dat. Svojí podobou připomíná sloupcový graf. Na rozdíl od něho však každý sloupec histogramu znázorňuje relativní nebo absolutní četnost. Tuto četnost ve vztahu k jednotce sledované veličiny pak znázorňuje na vodorovné ose. V porovnání se sloupcovým grafem, který zobrazuje kvalitativní data a dále již nepracuje s jednotkami na vodorovné ose, je histogram mnohem flexibilnější. [30]

Při vytvoření histogramu jsou hodnoty sledované veličiny nejdříve seřazeny dle velikosti a následně přiřazeny do vzájemně disjunktivních intervalů, tedy intervalů bez společných prvků. Tyto intervaly jsou následně vytvořeny na horizontální ose. Je důležité, aby tyto intervaly měly stejnou šířku, a to z důvodu srovnatelnosti. Rozdílná šířka by mohla vést k zavádějící interpretaci. Dalším důležitým aspektem je samotný počet takovýchto intervalů. Malý počet intervalů může způsobit zamaskování charakteru dat. Velký počet může mít za následek velkou variabilitu v četnostech hodnot uvnitř jednotlivých intervalů. Počet intervalů tak zásadně ovlivňuje výsledný obraz histogramu. Proto je vhodné najít způsob, jak volit adekvátní počet intervalů. Jednou takovou metodou je volba takového počtu intervalů, aby byl výsledný počet intervalů roven druhé odmocnině ze zkoumaných vzorků. Pro přehlednost je vložena modelová ilustrace souboru dat na Obr. 5-1. Z histogramu je patrné, že sloupce odpovídají normálnímu rozdělení. Naproti tomu Obr. 5-2 s největší pravděpodobností obsahuje odlehlé hodnoty. [30]



Obr. 5-1 Histogram s normálním rozdělením [31]

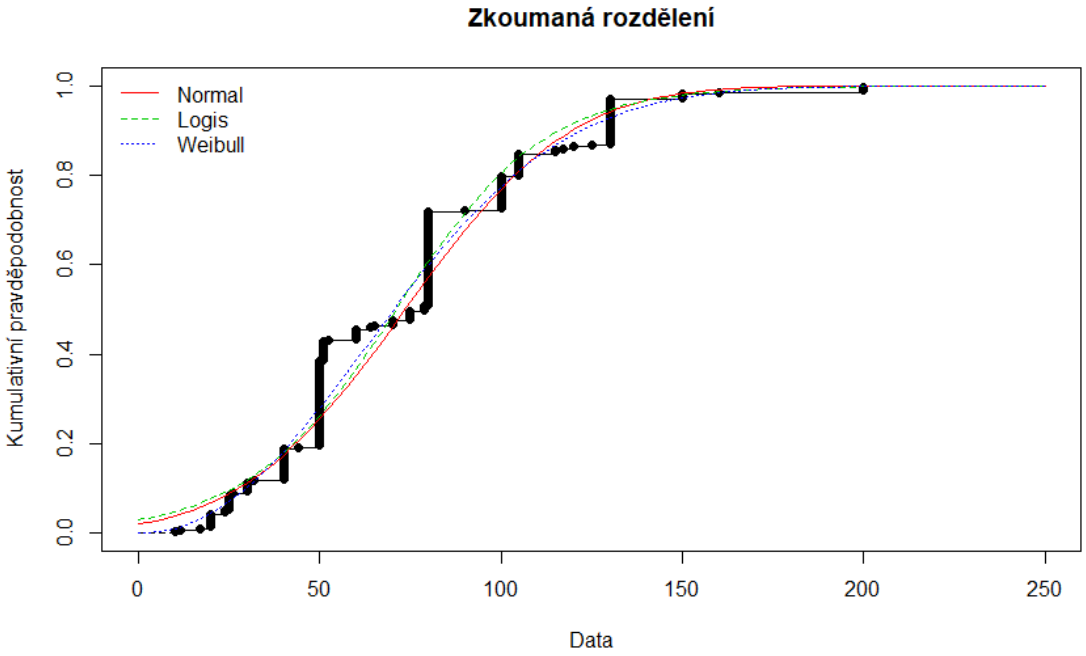


Obr. 5-2 Histogram s odlehlou hodnotou [32]

5.2 Empirický CDF graf

Funkce empirické distribuce je odhadem kumulativní distribuční funkce. Kumulativní distribuční funkce, zkratka CDF z anglického Cumulative Distribution Function, je taková funkce, která určuje pravděpodobnost mezi náhodnou proměnnou a zadanou hodnotou. [33] Přesněji s jakou pravděpodobností je hodnota náhodné proměnné menší než zadaná hodnota. Vzájemný vztah by se dal popsat jako neostrá nerovnost. Z toho vyplývá, že distribuční funkce jednoznačně určuje rozdělení pravděpodobnosti. [34]

V této práci je však pro vykreslení grafů jednotlivých rozdělení využita empirická distribuční funkce. Ta slouží jako odhad skutečné distribuční funkce náhodné veličiny. [35] Pro jednodušší představu je nejprve spočteno kolik náhodných veličin v našem náhodném výběru nabylo hodnoty menší nebo rovno zadané hodnotě. Tento výsledek je následně podělen velikostí náhodného výběru. S přibývající velikostí náhodného výběru se empirická distribuční funkce přibližuje k teoretické distribuční funkci. Jedná se o schodovitou funkci. Pro lepší ilustraci na Obr. 5-3 přiložen CDF graf vytvořený v jazyku R.



Obr. 5-3 CDF graf

6 Vstupní data

Jako vstupní data bakalářské práce slouží veličiny naměřené na veřejné tiskárně. Data se zálohují pomocí speciálního zařízení. Konkrétně se jedná o analyzátor SMC 144. Jeho úkolem je záznam spotřeby elektrické energie a zároveň umožňuje její dálkové sledování. Analyzátor je navržen jako zdroj dat pro moderní systémy, kde do vnitřní paměti ukládá informace o měřených veličinách [36]. Mezi sledované veličiny patří proud, napětí, frekvence a výkon. U těchto vzorků je dále zapisována jejich minimální, maximální a průměrná hodnota. Pro každou veličinu je k dispozici více verzí zapisovaných hodnot. V práci se nejčastěji pracuje s hodnotami prvního harmonického napětí označované jako avg.U1. Jedná se tedy o první sloupec z balíku hodnot harmonického napětí. V tabulce měřených veličin se dále nachází postupně hodnoty avg, min a max pro frekvenci f. Veličina harmonického napětí je zastoupena sedmi sloupci s pořadovým číslem 1-4, 12, 23 a 31. Hodnota minimálního a maximálního napětí je reprezentována sloupci 1-4. Pro hodnoty sloupců avg, minima a maxima proudu, je označení stejné jako v předchozím případě. U výkonu taktéž avšak s tím rozdílem, že počet sloupců pro každou variantu je omezený na tři.

Měřená data se rozdělily do několika intervalů. Následně se intervaly podrobovaly analytickému testování. Po načtení celého souboru se hodnoty uložily do lokální proměnné. Z proměnné se následně vybírala konkrétní veličina například již zmíněný sloupec U1. V něm bylo k dispozici celkem 6284 hodnot. Jednalo se o zkrácený testovací soubor o časové délce necelých 105 hodin. Časová řada se následně rozdělila na několik stejně dlouhých intervalů. Každý interval je tak reprezentován závislostí vstupní veličiny na čase. Počet intervalů je nepřímo úměrný na jejich časové délce. Zkoumání událostí při změně délky jednotlivých intervalů bylo jedním z cílů bakalářské práce.

Pro snadnější porovnání jednotlivých intervalů byly zkoumané úseky vykresleny do grafů. Nejčastěji se jednalo o histogram a CDF graf. Oba grafy přehledně zobrazují velikost pozorovaných hodnot a jejich četnost výskytu.

Do grafů se následně vložily křivky normálního, logistického a weibullova rozdělení. Aby mohlo být posouzeno, zdali zkoumaný interval odpovídá některému z rozdělení, vypočetla se i p hodnota, která byla pro větší přehlednost vypsána do titulku grafu.

Před samotným zjišťováním p hodnoty jsou nejprve stanoveny dvě hypotézy. Testovanou, ta bývá označována jako nulová, a alternativní. U nulové hypotézy je cílem prokázat pravdivost výroku. V našem případě, zdali testovaná data odpovídají zkoumanému rozdělení. Pokud se podaří zamítnout nulovou platí alternativní hypotéza. V našem případě, pokud zamítneme nulovou hypotézu, lze z určitostí tvrdit, že zkoumaný interval neodpovídá posuzovanému rozdělení. Na druhé straně, pokud se nepodaří vyvrátit nulovou hypotézu nelze s určitostí zamítnout alternativní hypotézu, pouze nezamítáme nulovou hypotézu. Mluvíme tak pouze o odhadu, kdy zkoumaný interval pravděpodobně odpovídá posuzovanému rozdělení. [37]

Kritéria pro vyhodnocení těchto hypotéz stanovuje p hodnota. Pro testování p hodnoty je formuluvána testovací a alternativní hypotéza. Testovaná hypotéza je ohraničena hodnotou 0,05 na které se ještě připouští platnost testované hypotézy. V případě, kdy p hodnota klesne

pod tuto hranici, je stanovena alternativní hypotézu. Při platnosti alternativní hypotézy je zamítnut výskyt rozdělení u zkoumaného intervalu. Posouzení velikosti p hodnoty hraje důležitou roli při posuzování, zdali je výsledek statistického testu pro práci významný nebo nikoliv [38]. Tato práce zaznamenává kladné hodnoty a ukládá je pro další využití. Hranice 0,05 interpretuje 5% chybovost testu, tedy případ kdy je zamítnuta testovací hypotéza i když platí. Pokud p hodnota pro testovaný vzorek dat překročí hranici 0,05, tak pozorovaný interval s určitou pravděpodobností odpovídá posuzovanému rozdělení. [38]

7 Popis webové aplikace

V této části je popsána naprogramovaná webová aplikace. Jsou zde popsány důvody vzniku webové aplikace a popisem samotného programu. Další část se věnuje grafické podobě programu.

7.1 Důvody vzniku Shiny aplikace

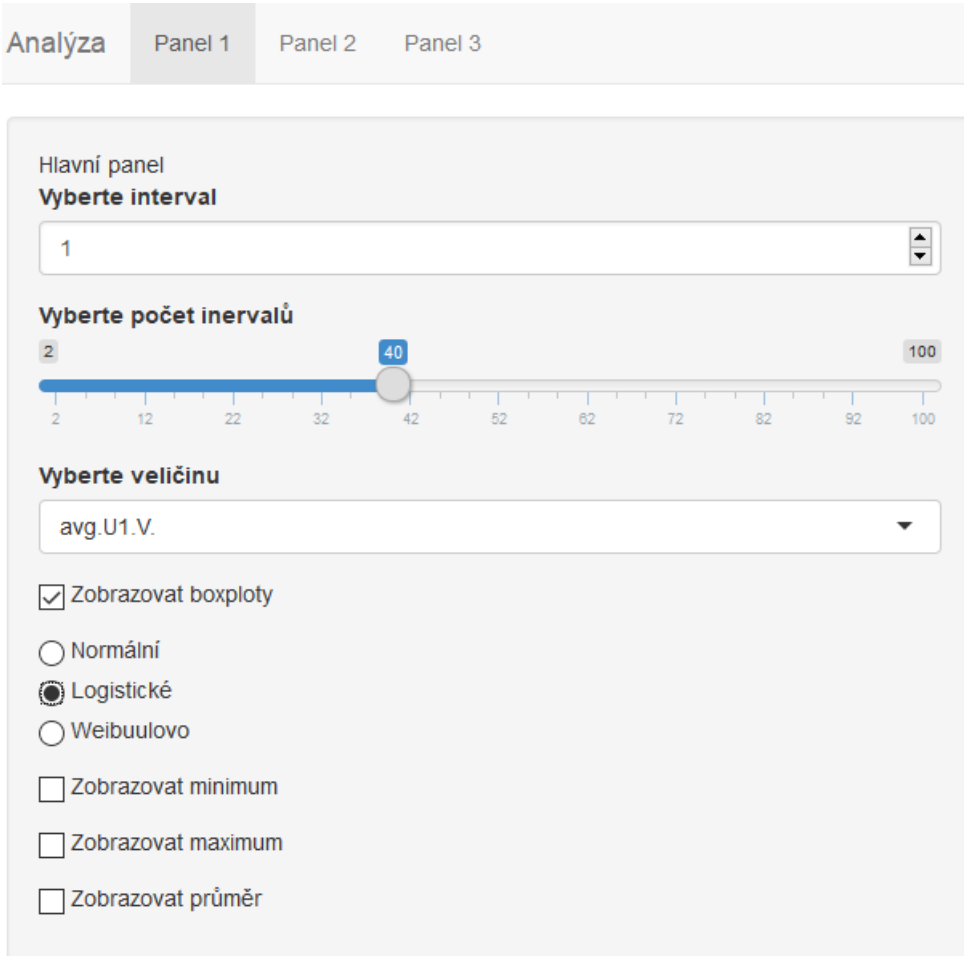
Tato část se zabývá důvodem vzniku, samotným programem a popisem webové aplikace vytvořené v prostřední Shiny. Webová aplikace má hned několik využití. V prvé řadě slouží pro snadnější práci a porozumění měřeným hodnotám. Díky vizualizaci výsledků výpočetního procesu jsou na první pohled zřejmé rozdíly mezi jednotlivými veličinami případně konkrétními intervaly. Je tak mnohem zřetelnější, proč daný interval neodpovídá porovnávaným rozdělením, než by tomu bylo s pouhým programovým výpočtem. Aplikace tak ulehčuje interpretaci chování měřených veličin případně dílčích intervalů, ať už se jedná o pravděpodobnostní rozdělení nebo zobrazení vzniklých odlehlých hodnot. Díky p hodnotám vloženým do popisků grafů lze naprosto intuitivně vytušit, které rozdělení interval splňuje a které nikoliv. Druhým důvodem je odstranění mechanického výpočtu p hodnoty. Bylo by velmi komplikované a zdlouhavé vypočítávat p hodnotu pro všechna rozdělení u každého intervalu. Intervaly by se musely následně sečíst a s každou změnou délky intervalu provést výpočet znovu. Shiny aplikace tak automatizuje výpočet a zpřístupní najednou všechny potřebné výsledky. Tento proces je navíc zvětší části automatizovaný.

7.2 Práce programu

Program na začátku načte vstupní data tiskárny. Všechny měřené fyzikální veličiny se uloží. Aktuálně zkoumaná veličina je rozdělena na sérii intervalů. Každý interval je prověřen testováním na odhad výskytů pravděpodobnostních rozdělení. Porovná tedy každý interval, zdali odpovídá normálnímu, logistickému nebo weibullovu rozdělení. Pro každý interval vypočte p hodnoty pro všechna tři zkoumaná rozdělení. Výsledné hodnoty spolu s křivkami jednotlivých rozdělení poté zanese do grafu. Výsledky u jednotlivých intervalů se mění spolu se zvyšujícím se nebo zmenšujícím se počtem intervalů. V programu se vypočte procentuální hodnota výskytů jednotlivých rozdělení při stanoveném množství intervalů. Pro zvýšení efektivity procesu je možné vypočíst kladné výskyty rozdělení pro více intervalů naráz.

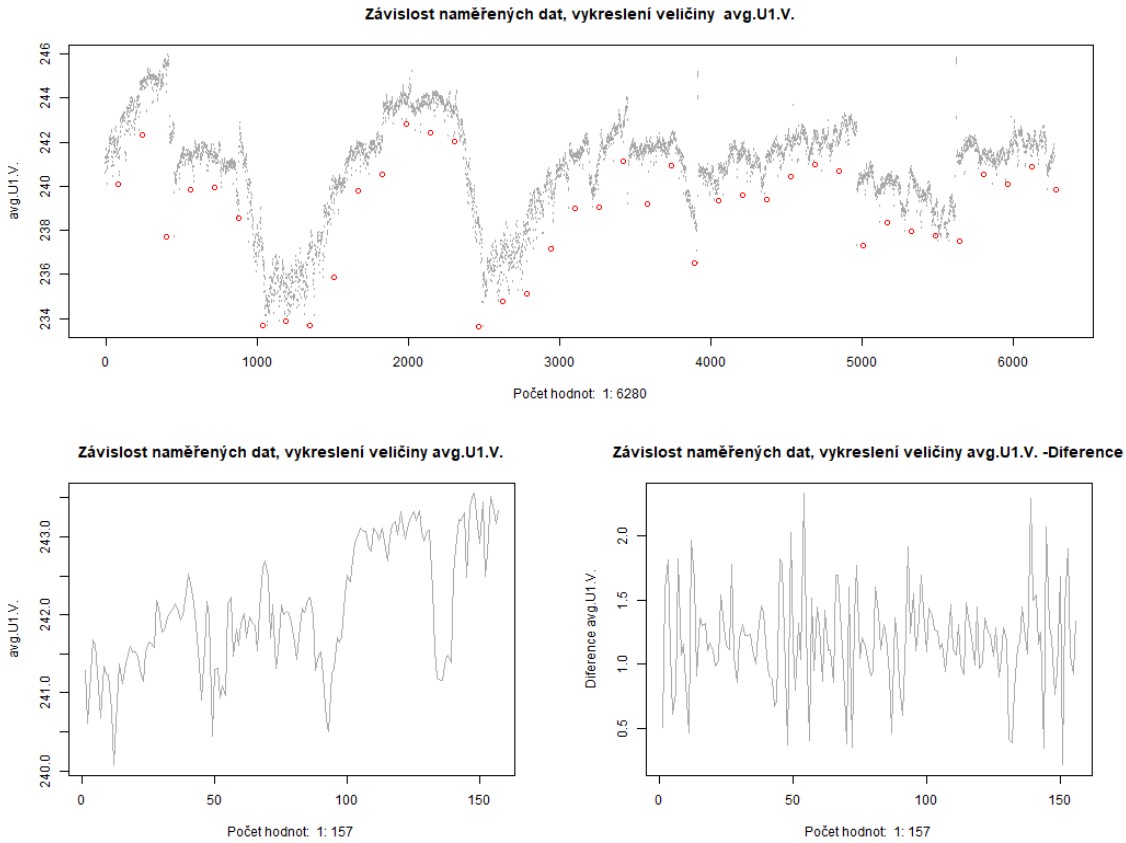
7.3 Vizuální podoba webové aplikace

V první záložce *Panel 1* je umístěn uživatelský panel společně se sérií grafů. Panel se nachází v levém horním rohu. Úkolem panelu je načíst všechny veličiny z nichž si uživatel vybere jednu konkrétní. Po stanovení veličiny se posuvníkem určí na kolik intervalů se sledovaná veličina rozdělí. Následně je k dispozici seznam jednotlivých intervalů. V kartě panelu je možné zaškrtnout zobrazování boxplotů jednotlivých intervalů. Barva boxplotů je stanovena podle p hodnoty jednotlivých rozdělení. Červená barva boxplotů značí interval bez rozdělení. Žlutá, zelená a modrá barva znázorňují výskyt rozdělení. Uživatel tak má možnost vizuálně posoudit výskyt kladných intervalů, společně s p hodnotou vypočtenou v některém z grafů. Další možností je zobrazení minima, maxima a průměru. Zobrazení boxplotů, stejně jako maxima, minima průměru je k dispozici u prvního grafu.

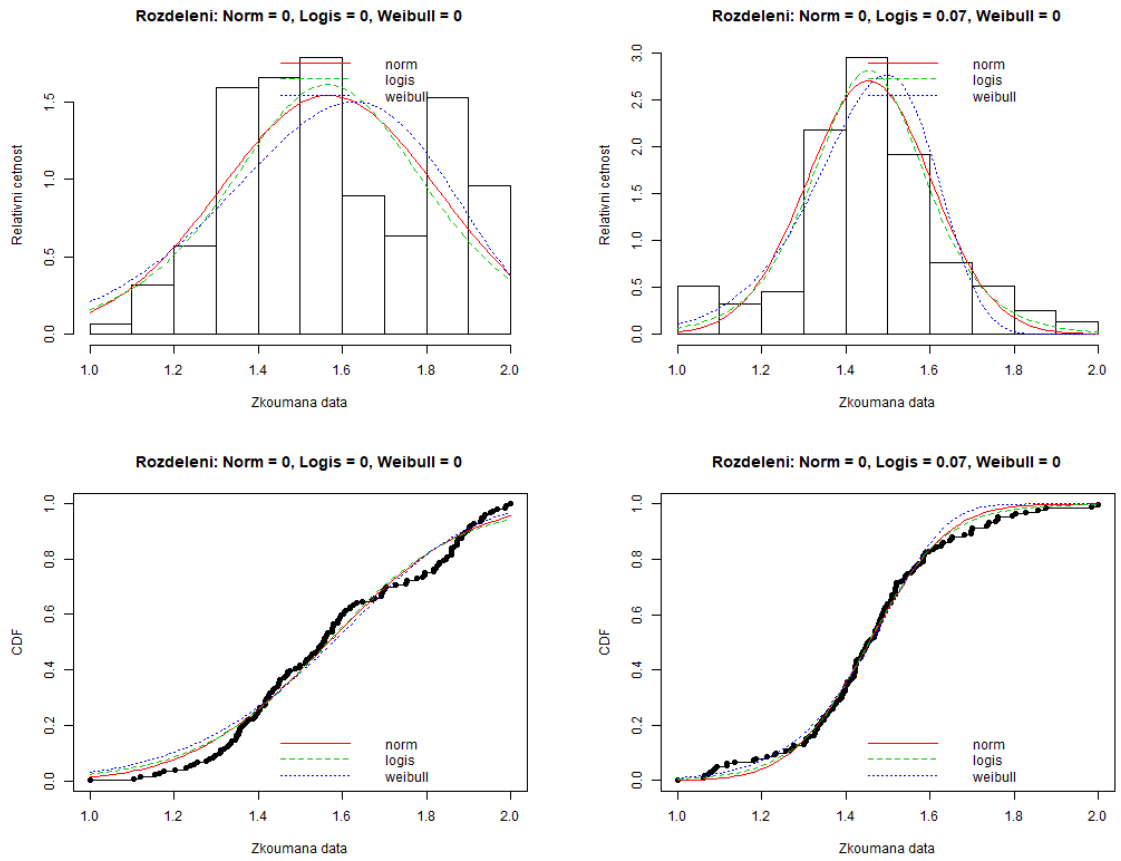


Obr. 7-1 Panel 1

Originální hodnoty veličiny se zobrazí v horním a zároveň největším grafu. Pouze velký graf je možné reprezentovat boxploty, maxima, minima a průměry jednotlivých intervalů. Hodnoty konkrétního intervalu jsou vyobrazeny vždy v prvním z trojice grafů. Jedná se o neupravená data. Chování intervalu je popsáno pomocí histogramu a CDF grafu. V obou grafech jsou umístěny křivky zkoumaných rozdělení včetně p hodnoty. Pro lepší názornost pozorovaných změn s délkou intervalů, jsou zde duplikovány tytéž grafy, avšak s upravenými hodnotami pomocí diference. Velká hodnota diference se může projevit nečekaným skokem v datech. Přidání diference je snahou podobné problémy vyřešit.

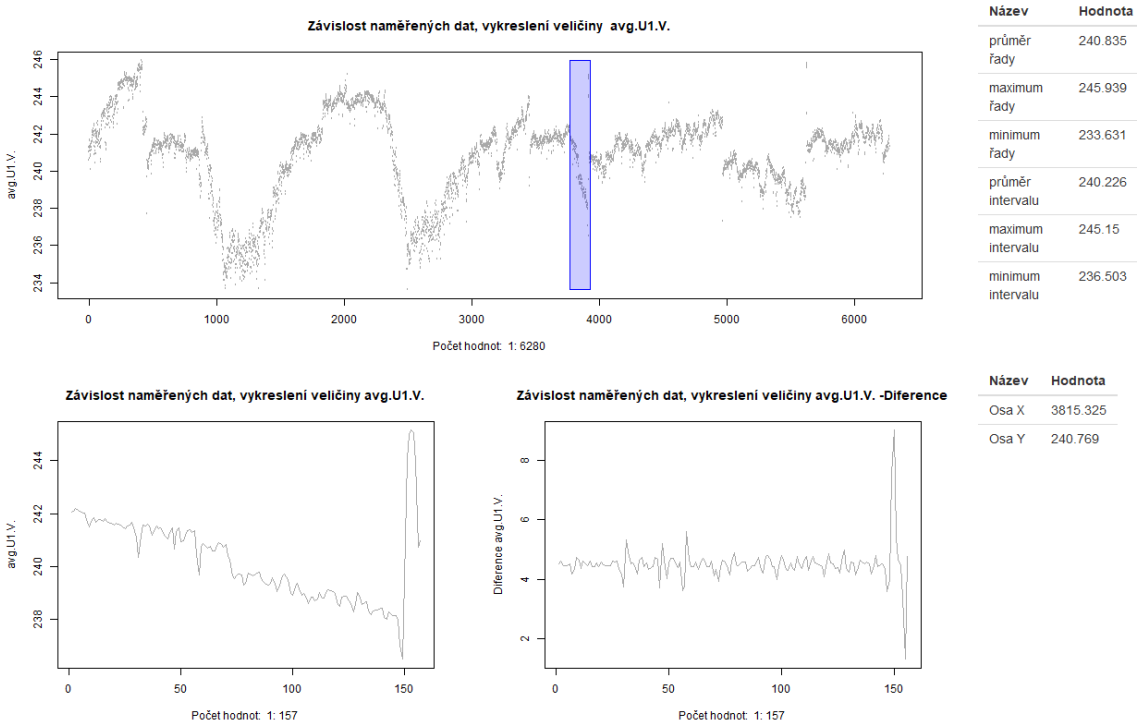


Obr. 7-2 Shiny aplikace zobrazující minima



Obr. 7-3 Data vykreslená histogramem a CDF grafem

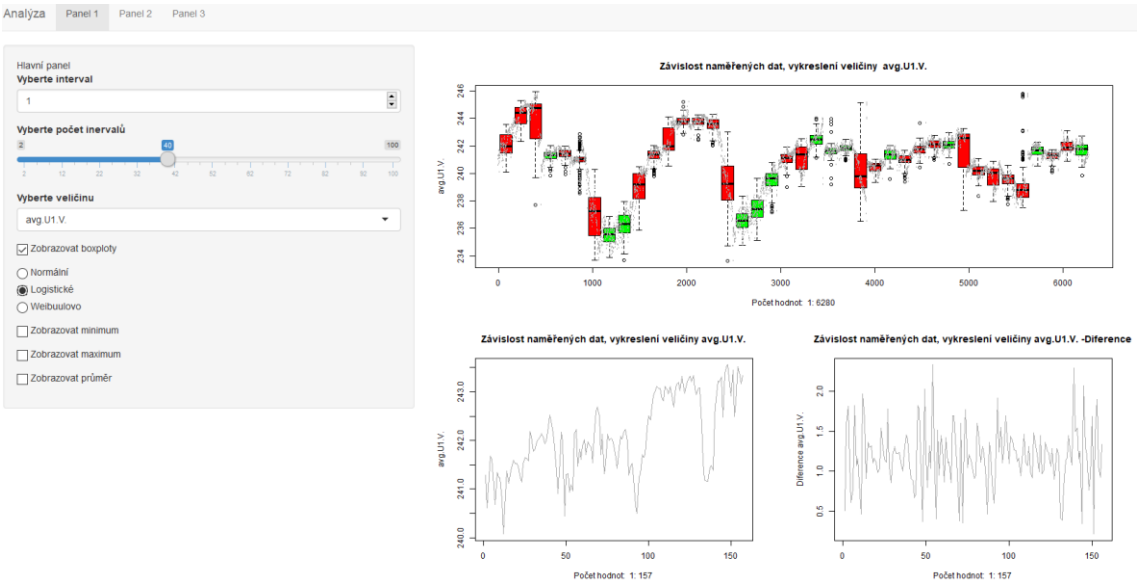
Kliknutím do velkého grafu se vybere jeden interval. Díky tomu je možné si na základě zobrazených měřených hodnot vybrat konkrétní interval. Důvodem mohou být zajímavé hodnoty uvnitř intervalu například v podobě nějakého extrému, tedy odlehlé hodnoty. Zároveň se v pravé dolní části zobrazí tabulka určující průměrnou hodnotu daného intervalu pro obě osy. Spolu s vybráním intervalu se interaktivně změní všechny grafy pracující s intervalem a zároveň se v levém panelu zobrazí konkrétní interval. Ten už je možné dále rozdělovat. Postupně je tak možné studovat chování jednotlivých intervalů dle vlastního uvážení.



Název	Hodnota
průměr řady	240.835
maximum řady	245.939
minimum řady	233.631
průměr intervalu	240.226
maximum intervalu	245.15
minimum intervalu	236.503

Název	Hodnota
Osa X	3815.325
Osa Y	240.769

Obr. 7-4 Shiny s vybraným intervalem

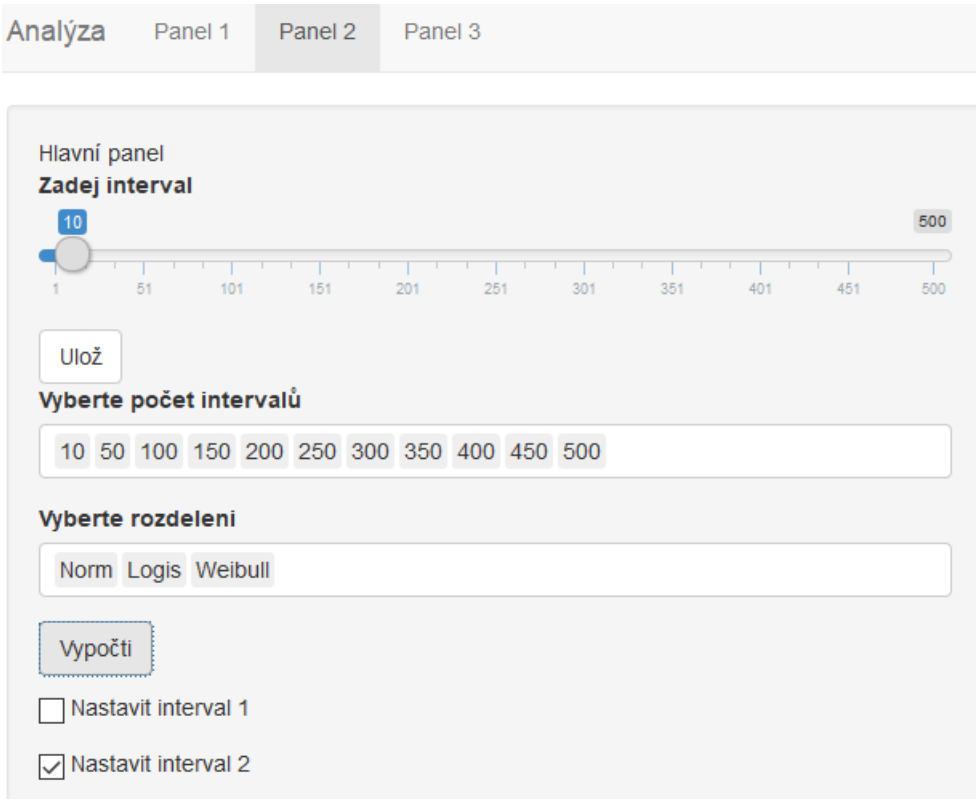


Obr. 7-5 Shiny aplikace zobrazující boxploty intervalů s logistickým rozdělením

Ve druhé záložce *Panel 2* se celý proces v mnohém zjednodušuje. Panel dva počítá globální statistiky všech intervalů, kdežto panel jedna slouží hlavně pro procházení zajímavých úseků a studování chování veličin v daném úseku. Uživatel si zde vybírá pouze ta rozdělení, která ho zajímají. Po jejich výběru se zobrazí tabulka s výsledky. Pro výpočet více intervalů zároveň je zde vytvořen graf, který přehledně ilustruje chování celé řady.

Postranní panel obsahuje posuvník pro zadání počtu intervalů. Pomocí tlačítka pro ukládání se hodnota objeví v seznamu zadaných hodnot. Hodnota v seznamu znázorňuje počet

intervalů vzniklých rozdělením veličiny. Pro usnadnění zadávání jsou hodnoty počtů intervalů předdefinovány ve dvou zaškrtávacích polích, z nichž si uživatel vybírá. Výhodou oproti *Panelu 1* je možná zadávání různorodého počtu intervalů najednou. Celý proces vybírání a zadávání intervalů se tak oproti prvnímu panelu značně zjednodušil na úkor komplikovanějšího výpočtu. V dalším uživatelském vstupu se vybere ze seznamu požadovaných rozdělení. Nakonec se tlačítkem provede výpočet.



Obr. 7-6 Nastavení panelu 2

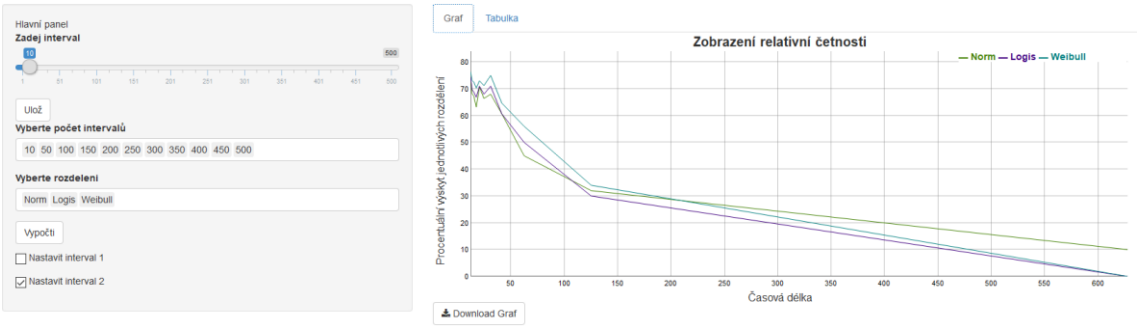
Obr. 7-7 pochází z panelu dva. Jsou vybrány všechny tři rozdělení. K označeným rozdělením se provede pravděpodobnostní odhad. Pro zpřehlednění výpočtu se vytvoří tabulka vypočtených údajů. V tabulce je obsaženo pět sloupců, kde tři z nich odpovídají sledovaným rozdělením. Zbylé dva sloupce reprezentují počet intervalů a délku intervalu.

Graf | Tabulka

Int	Int2	Norm	Logis	Weibull
10.00	628.00	10.00	0.00	0.00
50.00	125.60	32.00	30.00	34.00
100.00	62.80	45.00	50.00	56.00
150.00	41.87	60.67	60.67	64.67
200.00	31.40	68.00	71.00	75.00
250.00	25.12	66.40	68.00	71.20
300.00	20.93	70.67	71.00	73.00
350.00	17.94	63.14	66.86	70.29
400.00	15.70	67.25	68.75	72.50
450.00	13.96	68.22	69.78	73.11
500.00	12.56	73.20	74.40	76.40

Obr. 7-7 Rozdělení veličiny na intervaly za běhu webové aplikace

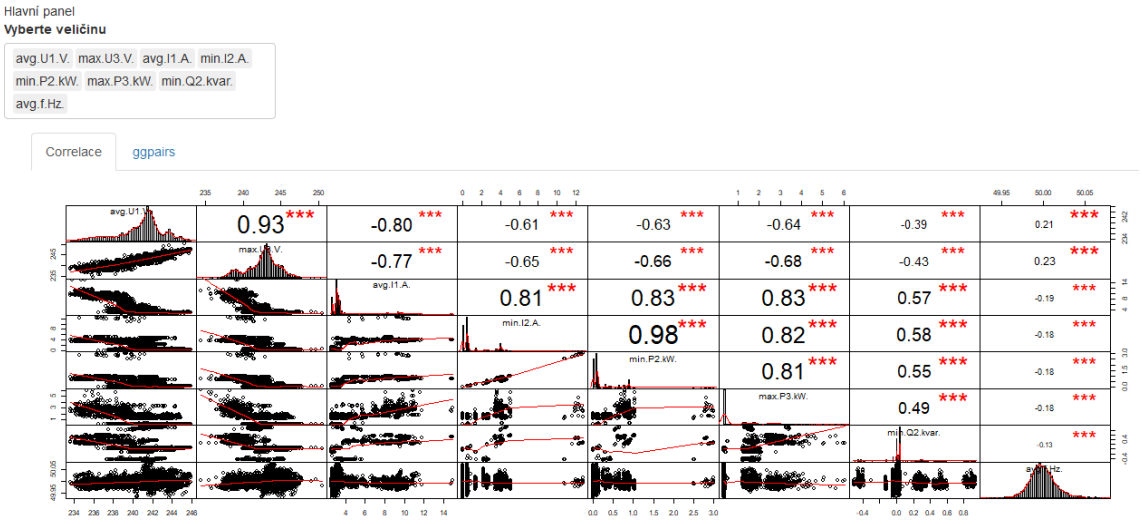
V přiložené tabulce první sloupec zobrazuje právě zadané počty intervalů. Pro pochopení je dále pro každý sloupec uveden příklad. Hodnota dvacet prvního sloupce udává, že se veličina rozdělila na dvacet intervalů. Ve druhém sloupci je vypočtena délka jednoho intervalu. Délka každého sloupce ze dvaceti intervalů je 314 minut. Třetí až pátý sloupec udává výsledky pro jednotlivá rozdělení. Ve veličině rozdělené na dvacet intervalů se normální rozdělení vyskytlo ve dvaceti procentech případů. Logistické a weibullovo rozdělení se vyskytlo ve 25 a 5 procentech případů. Hodnoty počtu výskytů pro zkoumaná rozdělení vůči časové délce intervalů jsou pro snazší představu nastíněny v grafu. V něm každá křivka představuje právě jedno rozdělení.



Obr. 7-8 Zobrazení výpočetních výsledků v grafu

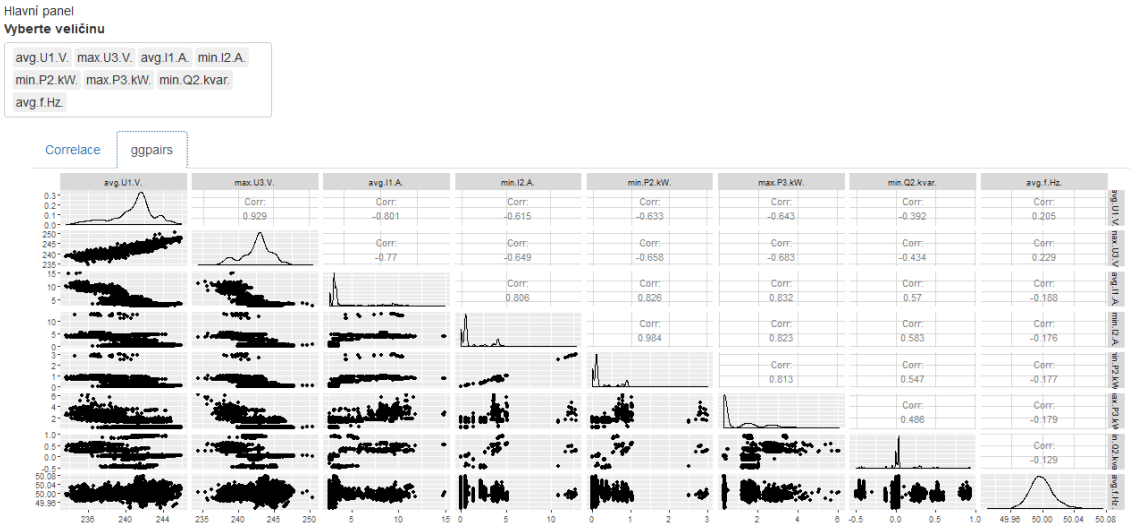
Ve třetí záložce *Panel 3* si uživatel vybírá pouze konkrétní veličiny. *Panel 3* přímo nesouvisí s předchozí prací, ale může být vhodný při studiu vzájemných souvislostí mezi řadami. Umožňuje srovnávat vzájemné korelace mezi libovolným počtem vybraných veličin. Jejich vzájemná závislost je poté přehledně zobrazena v grafu.

Obr. 7-9 představuje vizualizovanou korelační matici. Na vrcholu se nachází absolutní hodnota korelace doplněná řešením korelačního testu ve formě hvězdiček. Ve spodní části se nachází korelační diagram neboli bodový graf vyjadřující závislost proměnných. Tento graf je protnut křivkou, která co nejlépe vystihuje vzájemnou závislost. V tomto případě graf zobrazuje veličiny napětí, proudu, výkonu a frekvence. [39]



Obr. 7-9 Korelace

V tomhle případě je na diagonále vždy histogram a distribuční funkce dané veličiny. V ostatních oknech se nachází různé varianty vzájemného srovnání dvojic.



Obr. 7-10 Graf funkce ggpairs

Pro lepší názornost je přiložen ggpairs graf. Jedná se o tutéž funkci akorát s využitím jiného prostředku.

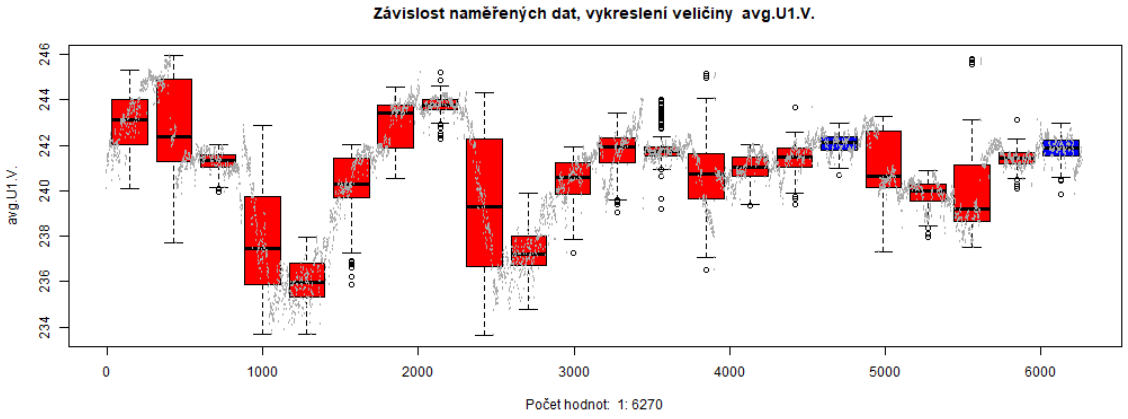
8 Automatické prověřování časové řady

Na začátku kapitoly je proveden rozbor časové řady nejprve pomocí boxplotů. Další část se věnuje analýze časové řady pomocí sledovaných rozdělení. Časová řada je rozdělena mezi intervaly reprezentující konkrétní úsek popsaný p hodnotou. Posouzením p hodnoty je stanoven odhad rozdělení. Ke každému intervalu dat přísluší jeden interval diferencí.

Jako příklady různých výsledků jsou zobrazeny intervaly s normálním, logistickým a weibullovým rozdělením. Pro úplnost jsou uvedeny intervaly se všemi rozděleními, doplněné o intervaly, které odpovídaly jenom některým případně neobsahovaly žádné rozdělení.

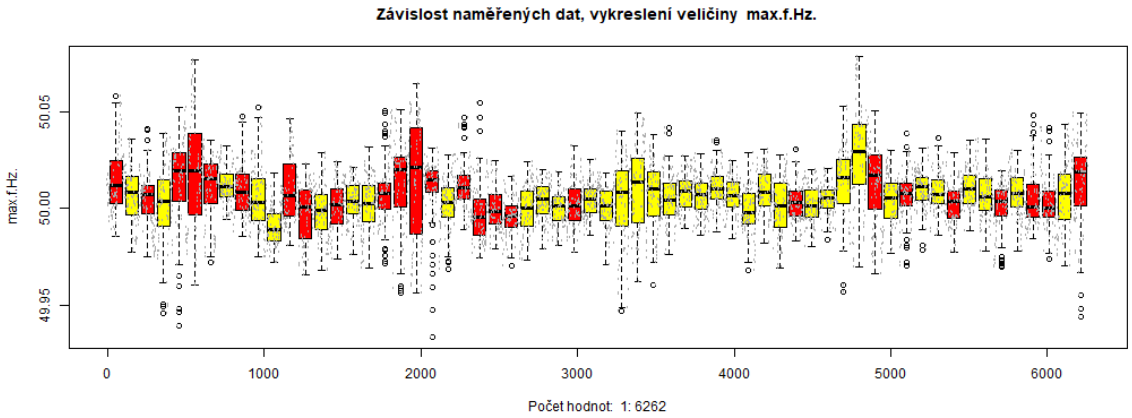
Tato část popisuje chování jednotlivých intervalů a popisuje praktické využití p hodnoty. Na začátku se měřené veličiny načetli. Poté došlo na sledování a vyhodnocování chování celých řad. Pro ukázku je zde vykreslen graf při rozdělení měřených úseků na 22 intervalů Obr. 8-1. Pro názornost je zde uveden příklad při rozdělení veličiny na 62 intervalů Obr. 8-2.

Na Obr. 8-1 jeden boxplot reprezentuje celý interval, kde dva intervaly pravděpodobně splňují weibullovo rozdělení.



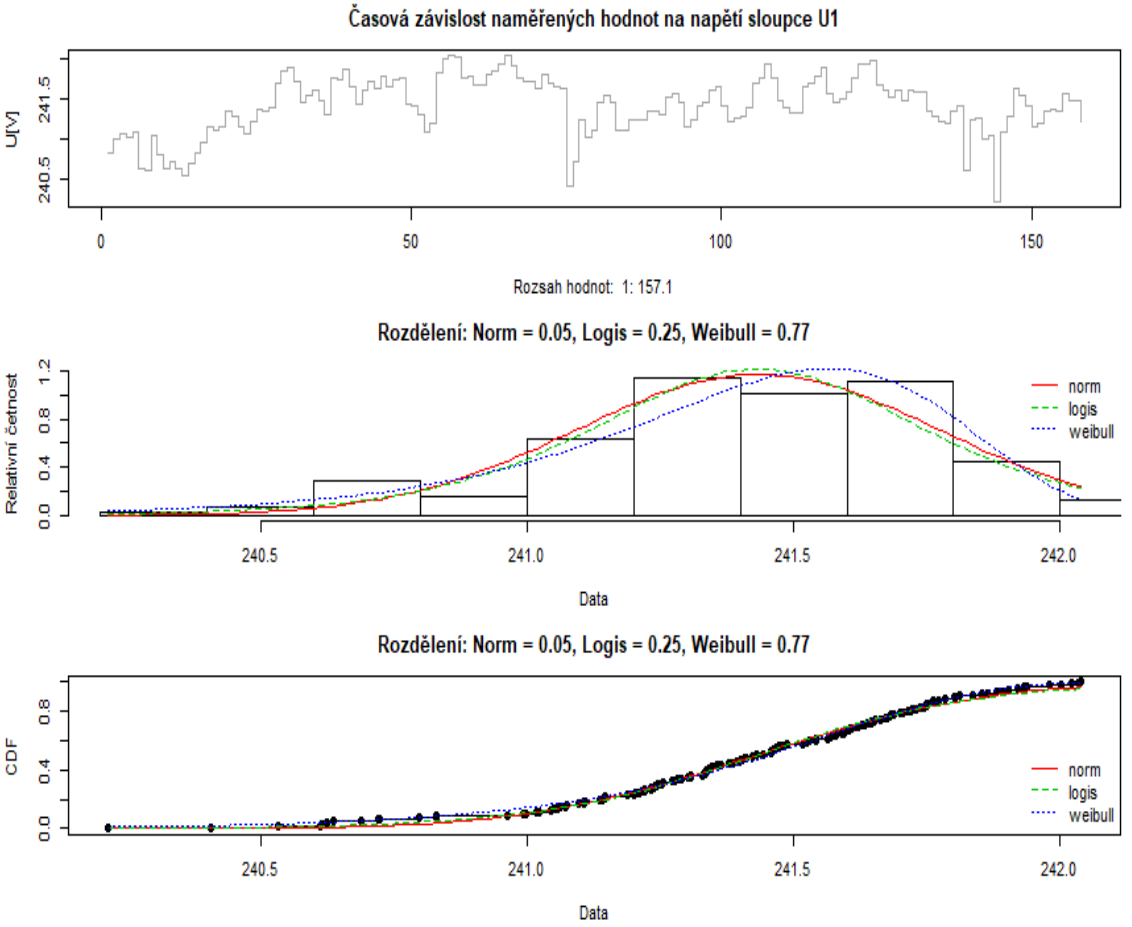
Obr. 8-1 Boxploty intervalů veličiny avg U1

Zde je na první pohled vidět mnohem větší počet pravděpodobného výskytu tentokráte pro normální rozdělení.



Obr. 8-2 Boxploty intervalů pro max f

Dále došlo na sledování samostatných intervalů. Tento proces znamená, že bylo nutné projít všechny intervaly v dané veličině a spočítat pro ně p hodnotu. P hodnota, jak již bylo popsáno víše naznačuje, zdali sledovaný úsek s jistou určitostí odpovídá hledaným rozdělením. Pro vizualizace celého procesu jsou k dispozici série několika intervalů s rozdílnou p hodnotou.

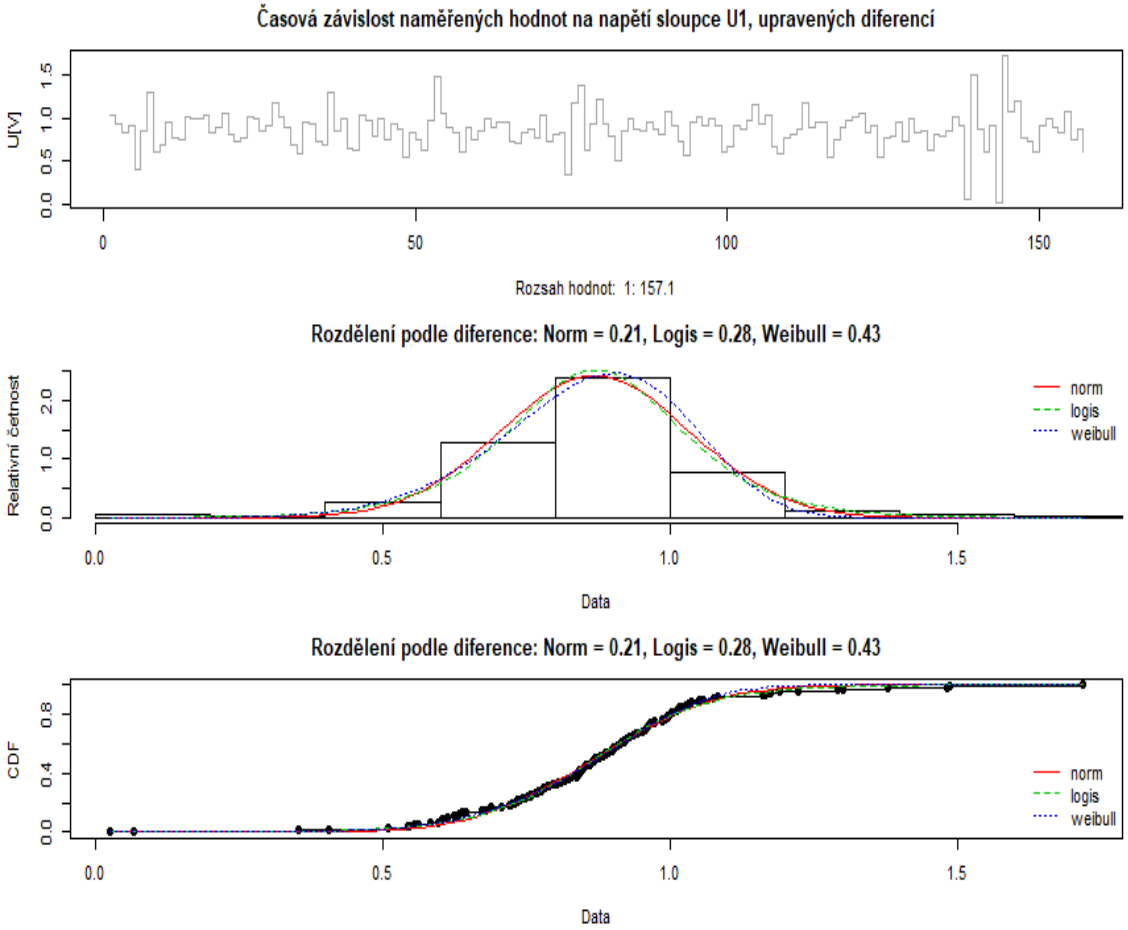


Obr. 8-3 Veličina U1 obsahující všechna zkoumaná rozdělení

Aby bylo dosaženo co nejlepší možné názornosti je každý úsek dat vyobrazen. Pro vizuální kontrolu s vypočtenou p hodnotou jsou zde k dispozici dva grafy. Jedná se o histogram a CDF

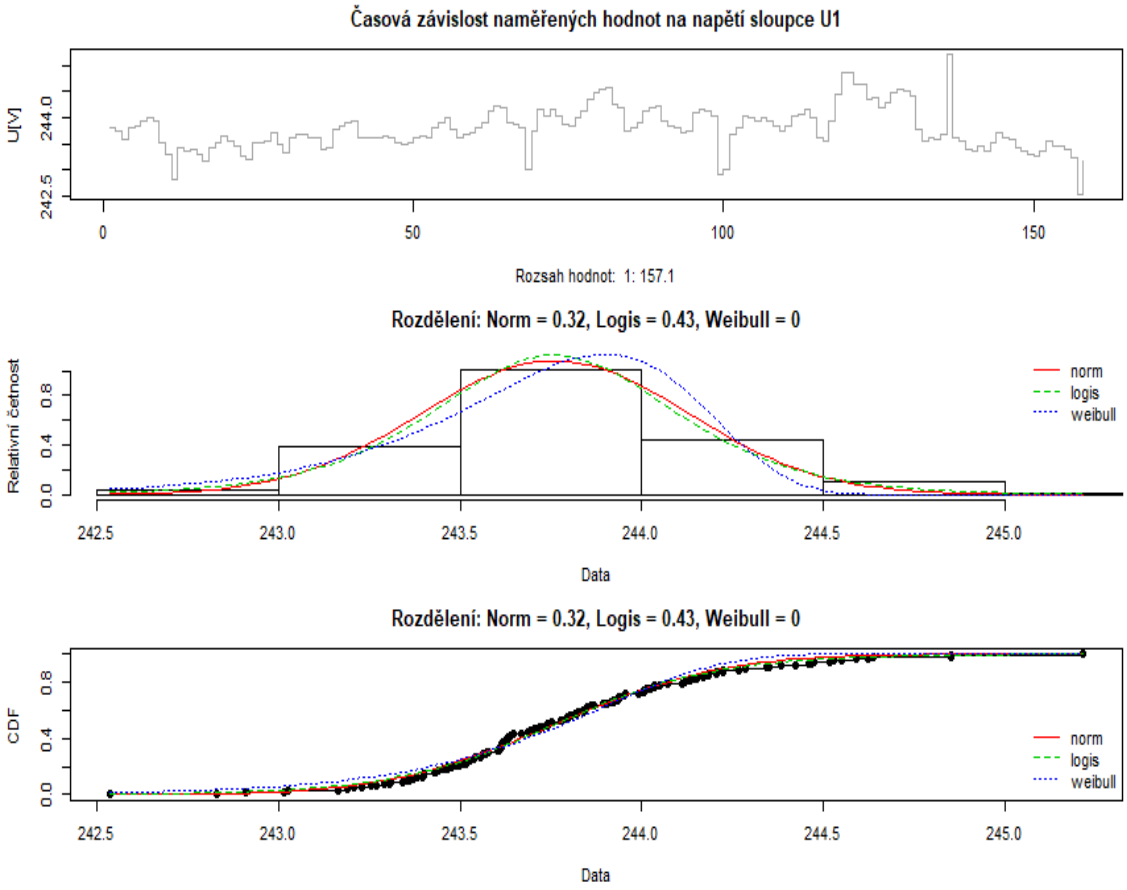
graf. Oba mají podobný význam. Vypočtené hodnoty v grafu jsou proloženy úsečkami jednotlivých pravděpodobnostních rozdělení.

V tomto případě podle vypočtené p hodnoty lze stanovit odhad, kde pozorovaný interval odpovídá všem třem rozdělení. Odhad lze stanovit díky kritériu a sice minimální velikosti p hodnoty. Podmínkou rozdělení je vyšší p hodnota než 0,05 což opravdu všechny vypočtené p hodnoty splňují. Ostatně p hodnotu udává i samotný graf, kde je možné vidět, jak moc se úsečky rozdělení k naměřeným datům přibližují.



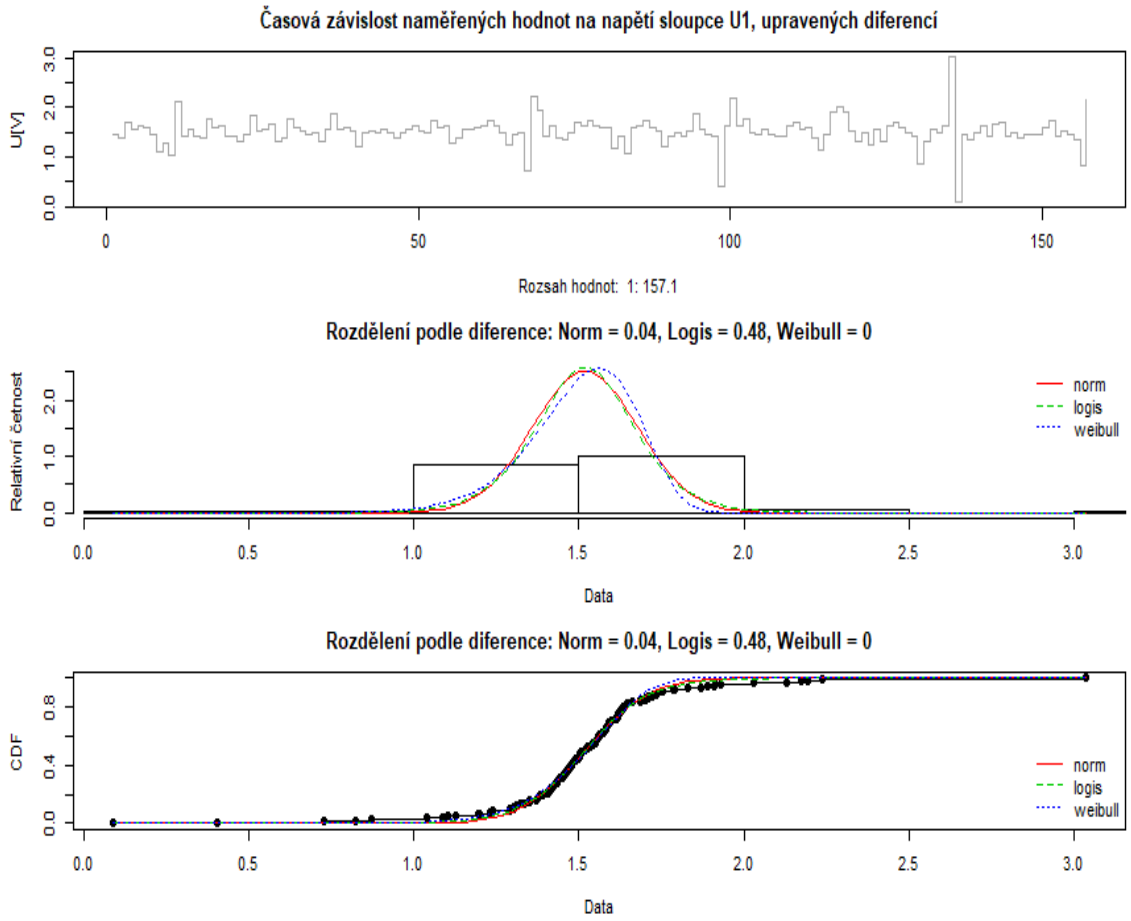
Obr. 8-4 Diference pro veličinu U1 odpovídající všem hledaným rozdělením

Jedná se o tentýž interval. V tomto případě je však vytvořený graf upraven pomocí diference. Diference představuje rozdíl dvou sousedních absolutních přírůstků. Diference umocňuje výrazné jednorázové změny. Při porovnání různě kolísavých časových řad by mohlo dojít ke zkreslenému výsledku. Diference rozdíly normalizuje a tím zabraňuje chybné interpretaci.



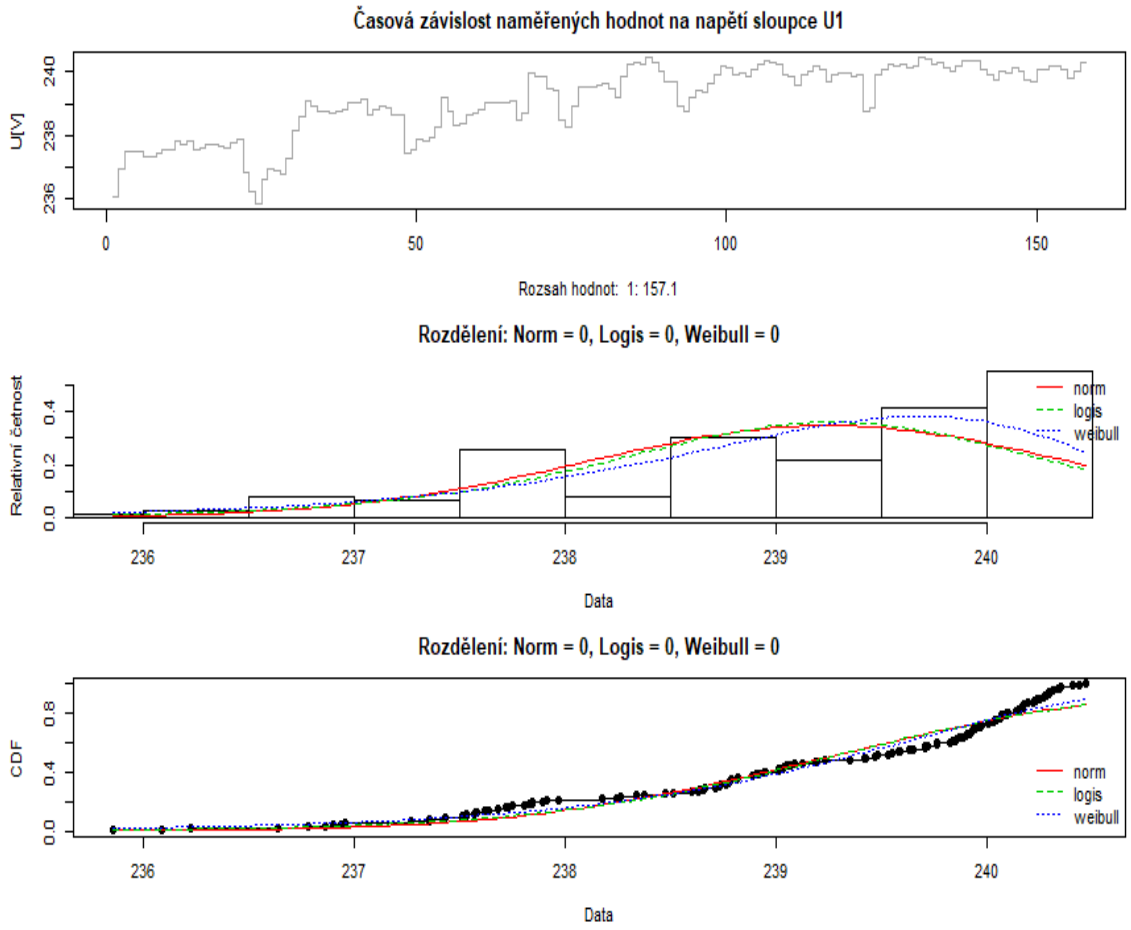
Obr. 8-5 Veličina U1 s normálním a logistickým rozdělením

Oproti předchozím případům obrázků zde došlo k viditelné změně. Zatímco pro první dvě rozdělení se lze domnívat, že intervaly odpovídají zkoumaným rozdělením. Ve třetím případě pro weibullovo rozdělení je možné s jistotou tvrdit, že danému intervalu neodpovídá. Zdali daný interval lze interpretovat pomocí jednoho či druhého rozdělení souvisí již se zmíněnou p hodnotou. Jedině po zamítnutí nulové hypotézy lze s jistotou tvrdit, že je získána důvěryhodná zpráva o chování celé řady.



Obr. 8-6 Diference U1 s logistickým rozdělením

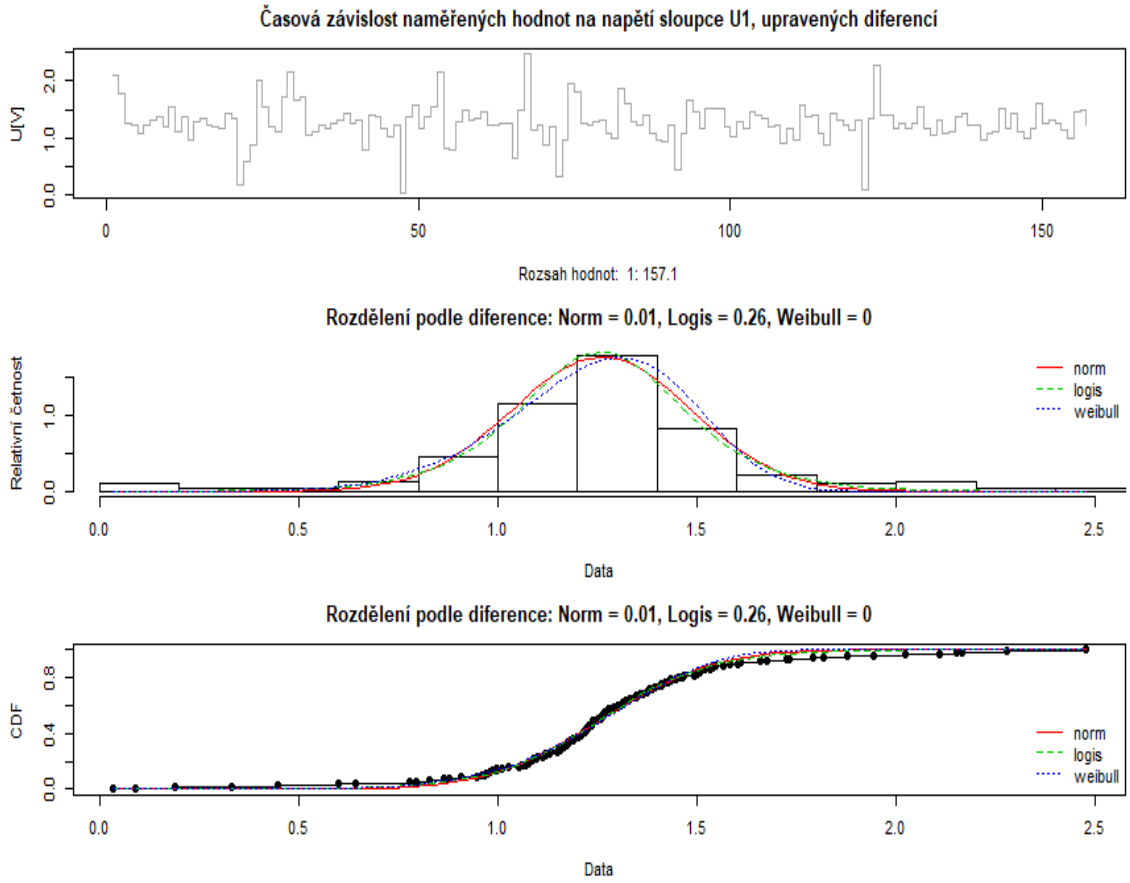
Znovu stejný interval. Taktéž jako v předchozím intervalu upravený pomocí diference. V tomto případě je vidět, že vlastnosti p hodnoty splňuje pouze logistické rozdělení. Pro zbylá dvě rozdělení lze s určitostí tvrdit, že danému intervalu neodpovídají.



Obr. 8-7 Veličina U1 neodpovídající žádnému z rozdělení

Tento interval neodpovídá žádnému ze zkoumaných rozdělení.

Pro diferenci platí, že pravděpodobně odpovídá pouze logistickému rozdělení.



Obr. 8-8 Diference U1 s logistickým rozdělením

9 Měření a porovnání četností jednotlivých intervalů

Hlavní část této práce je věnována nalezení co možná nejdelšího úseku časového intervalu, který odpovídá některému ze zkoumaných rozdělení. Mezi taková rozdělení patří normální, logistické a weibullovo, která již byla zmíněna v průběhu práce. Motivací pro tuto kapitolu je nalezení vhodné metody pro ukládání dat. Je vhodné ukládat co nejmenší objem dat, proto se časový úsek rozdělí do několika intervalů. Hlavní podmínkou je nepřijít o důležité informace. Jednou z méně vhodných metod používaných v současnosti je ukládání minima, maxima a průměru. Při tomto procesu může dojít ke zkreslení informace, dokonce i její ztrátě viz kapitola 1. Práce se věnuje jiné metodě. Každá veličina v bakalářské práci představuje časovou řadu. Během rozkladu časové řady na intervaly je každý sloupec reprezentující interval zkoumán na výskyt jednotlivých rozdělení. Kapitola se zabývá souvislostí mezí délkou intervalu a procentuálním počtem rozdělení v intervalu.

Pro některé časové úseky platí, že nedochází k žádnému výskytu odlehlých hodnot. Odlehlé hodnoty mohou být způsobeny kolísáním proudu, nebo jiné fyzikální veličiny. I v takovém případě jsou data uložena. Bylo by však výhodné uložit více dat najednou. Z tohoto důvodu se vyplatí hledat co možná nejdelší časové intervaly, které by mohly být nahrazeny směrodatnou odchylkou a průměrem. V intervalu, kde hodnoty hodně kolísají, je nutné uložit všechna data. Analyzováním vzorků jednotlivých intervalů lze nalézt optimální hranici délky intervalu.

Pro lepší srovnání jsou přidána data diference. Diference umocňuje výrazné jednorázové změny. Jestli jsou data rostoucí nebo klesající nemá na diferenci vliv. Kapitola 9 o testování normality je snahou o zlepšení metody ukládání. Nahrazení maxima, minima a průměru pomocí směrodatné odchylky a průměru.

Výskyt jednotlivých rozdělení vypovídá o minimálních skokových změnách hodnot uvnitř časového úseku. Tedy, že v celém analyzovaném intervalu nejsou žádné odlehlé hodnoty. Největší zastoupení normálního rozdělení bylo dosaženo v délce intervalu 10 minut. Z toho vyplývá, že nejlepší metodou je ukládání dat v 10-ti minutových intervalech. Z přiložených grafů křivek a přidaných tabulek je dále patrné, že při délce intervalu 30 minut je procentuální výskyt normálního rozdělení také velmi vysoký. Ukládáním 30-ti minutových intervalů lze dosáhnout výhodnějšího poměru mezi vysokým procentuálním výskytem rozdělení a ušetřeným místem. Z tohoto důvodu je ukládání po 30 minutách mnohem výhodnější metodou.

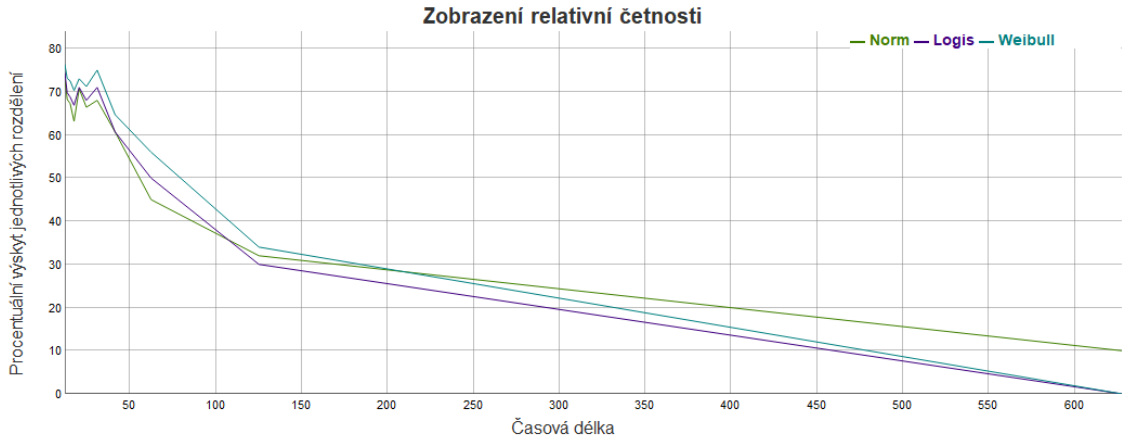
V této kapitole je několik tabulek popisující rozdílnost naměřených hodnot jednotlivých rozdělení pro různě dlouhé časové řady. Každá popisuje, kolik hodnot z našich vzorků odpovídá porovnávaným rozdělením. V našem případě tedy normálnímu, logistickému a weibullovu rozdělení. Po analýze byly výsledné hodnoty zaznamenány a porovnány mezi sebou. Hodnoty obsažené v intervalu odpovídají časové délce naměřených dat vyjádřených v minutách. Pro názornost těchto výpočtů je zde přiložen graf.

Tabulka 1 Počet odhadovaných výskytů pro veličinu avg U1

Počet odhadovaných výskytů u jednotlivých rozdělení			
Délka intervalů	Normální [%]	Logistické [%]	Weibullovo [%]
628	10	0	0
126	32	30	34
63	45	50	56
42	61	61	65
31	68	71	75
25	66	68	71
21	71	71	73
18	63	67	70
16	67	69	73
14	68	70	73
13	73	74	76

Jedná se o tabulku, kde jednotlivé intervaly nabývaly hodnot v rozmezí 13 min až 628 minut. Pro porovnání vzájemné závislosti délky intervalů a vlivu na distribuce jsou zde zaznamenány odhady normálního, logistického, weibullova rozdělení. Každý sloupec obsahuje jeho procentuální výskyt vzhledem k délce jednotlivých intervalů. Vypočtená procenta uvnitř tabulky tedy popisují, jak častý byl výskyt u jednotlivých rozdělení.

Pro představu je zde ukázkový příklad. Intervaly, jež reprezentují časový úsek v délce 14 minut z celkových 6284 minut, jsou po 14 minutách analyzovány. Prováděná analýza porovnává, zdali daný interval splňuje podmínky pro výskyt rozdělení. Tedy zdali zkoumaný časový úsek lze aproximovat odhadem buď normálního, logistického, nebo weibullova rozdělení. Pokud je výsledek pro interval kladný, je uložen. Tímto způsobem jsou porovnány všechny intervaly. Takto se postupuje, dokud nejsou analyzovány všechny časové intervaly, v tomto případě veličiny avg U1. Výsledné odhady pravděpodobnosti u jednotlivých rozdělení se sečtou a zapíší do tabulky. Podle výše uvedeného platí, že intervalům v délce 14 minut přísluší 68 % pravděpodobnost výskytu normálního rozdělení. Procento úseků logistického rozdělení je 70 % a u weibullova rozdělení 73 % při stejné délce intervalu.



Obr. 9-1 Graf pro jednotlivá rozdělení

Analyzování dat ukázalo, že se zkracující délkou intervalu se zvětšuje procento úseků normálního, logistického a weibullova rozdělení. Z analýzy vyplývá, že zvýšení pravděpodobnosti procentuálního výskytu rozdělení lze dosáhnout sestavením co možná nejkratších intervalů.

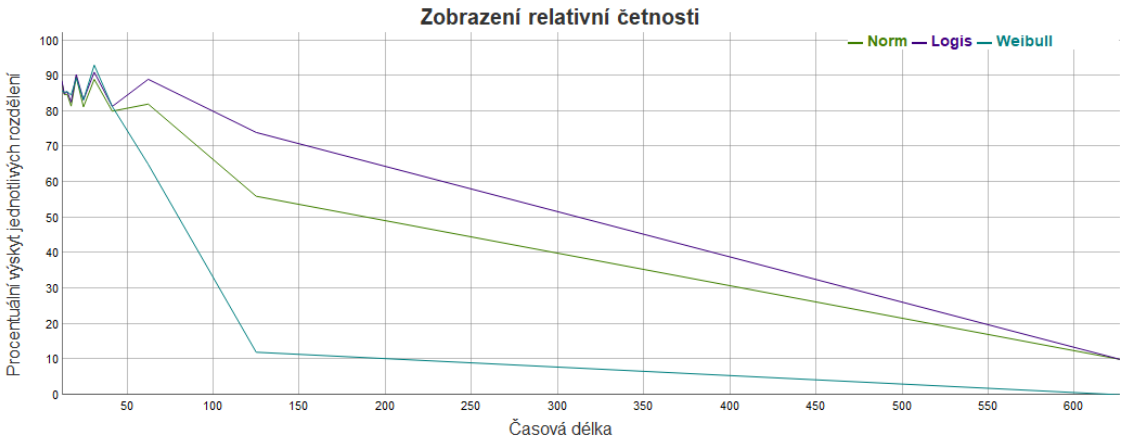
Dále je zde vyobrazena tabulka diferencí pro U1.

Tabulka 2 Počet odhadovaných výskytů diferencí avg U1

Počet odhadovaných výskytů u jednotlivých rozdělení			
Délka intervalů	Normální [%]	Logistické [%]	Weibullovo [%]
628	10	0	0
126	56	74	12
63	82	89	65
42	80	81	81
31	89	91	93
25	81	83	83
21	90	90	89
18	81	83	85
16	85	86	86
14	85	85	85
13	88	89	86

Dále je zde vyobrazena tabulka diferencí pro U1.

Tabulka 2 popisuje stejná vstupní data, avšak upravená pomocí diference. Diference zvýrazňuje skokové rozdíly sousedních hodnot, což má kladný vliv na procentuální výskyt u všech tří rozdělení.



Obr. 9-2 Četnost jednotlivých rozdělení pro diferenci U1

Z grafu je patrné, že se výsledné hodnoty u všech rozdělení zvýšily o jednotky procent. Taková normalizace dat zvýšila počet pravděpodobnostních výskytů u všech sledovaných rozdělení a umožnila tak sestrojit přesnější odhad, který by lépe interpretoval časový úsek.

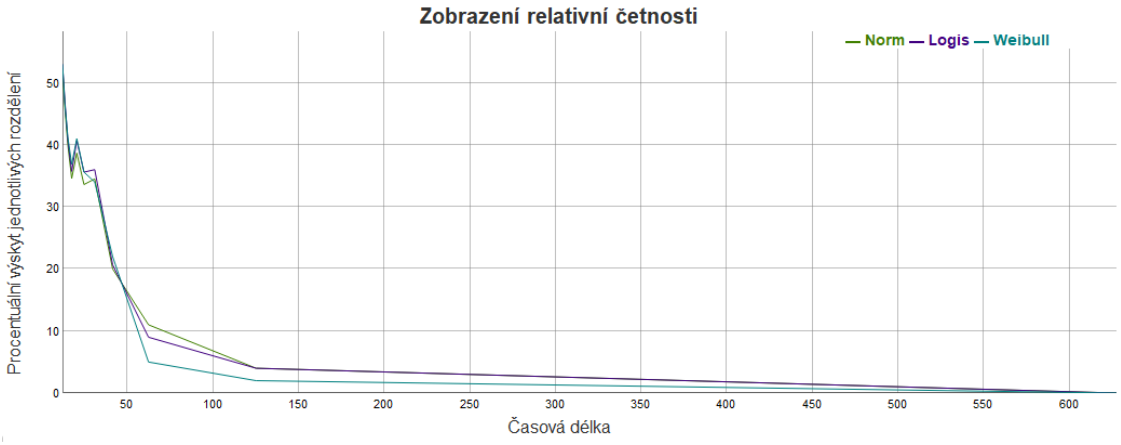
V případě jiných hodnot napětí konkrétně min U3 nedošlo k výrazným odchylkám od proměnné U1. Data se příliš nelišila ani v hodnotách diference. Z tohoto důvodu byla pro lepší srovnání zavedena jiná veličina.

Aby bylo možné určit, zdali se zachová vzrůstající tendence u kratších intervalů i pro jiné veličiny, je zde přiložena Tabulka 3 proměnné max I4 Obr. 9-3. Analýza této proměnné by měla určit, zdali je možné popisovat intervaly pomocí distribucí a spolu s tím takové úseky dat vypouštět a nahrazovat je pouze průměrem a směrodatnou odchylkou.

Tabulka 3 Počet odhadovaných výskytů pro veličinu max I4

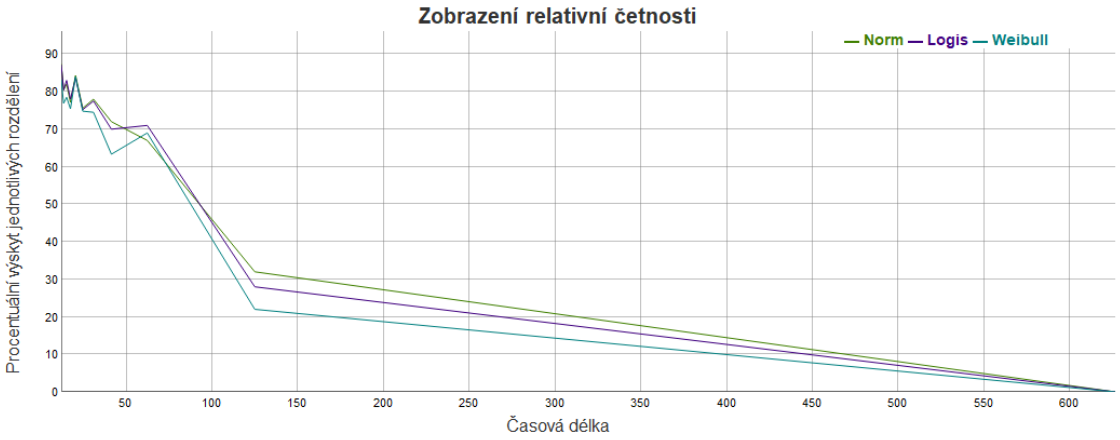
Počet odhadovaných výskytů u jednotlivých rozdělení			
Délka intervalů	Normální [%]	Logistické [%]	Weibullovo [%]
628	0	0	0
126	4	4	2
63	11	9	5
42	20	21	22
31	34.5	36	34
25	34	36	36
21	39	41	41
18	35	36	37
16	40	41	41
14	46	47	47
13	52	53	53

Hodnoty v této tabulce se liší od předchozích. Jsou zde patrné zákmity pro všechna sledovaná rozdělení. Navíc procentuální výsledky jsou nižší pro všechna rozdělení, a to v řádu několika desítek procent.



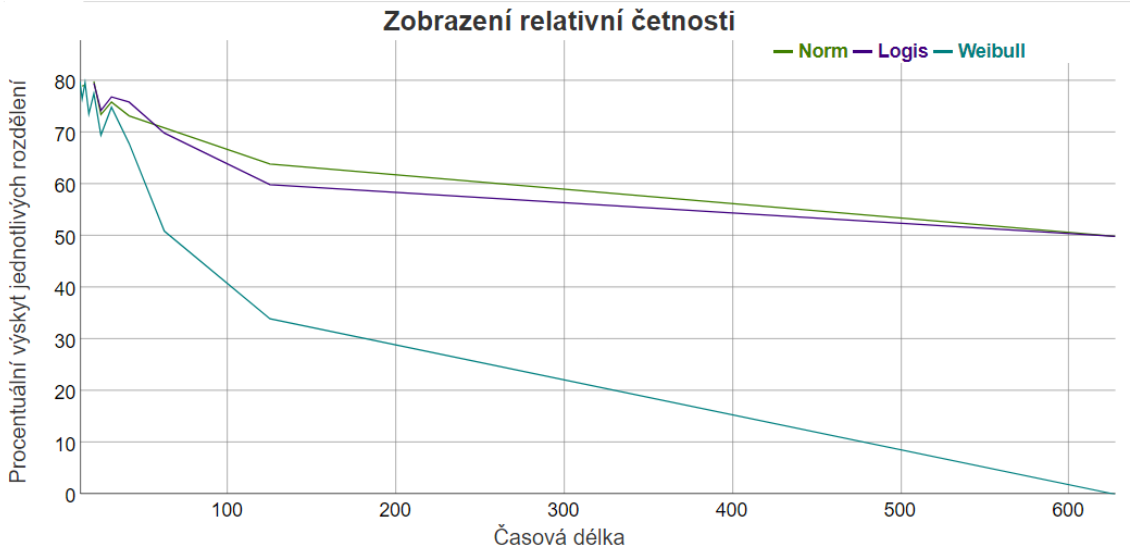
Obr. 9-3 Četnost výskytů jednotlivých rozdělení pro veličinu I4

Sledování normality diference napomáhá při výběru nejvhodnější metody pro ukládání dat. Poskytuje mnoho zajímavých informací o chování systému. Výpočet diference veličiny max I4 na Obr. 9-4 napomáhá chování systému objasnit, případně využít. O tom ostatně vypovídá i následující graf.



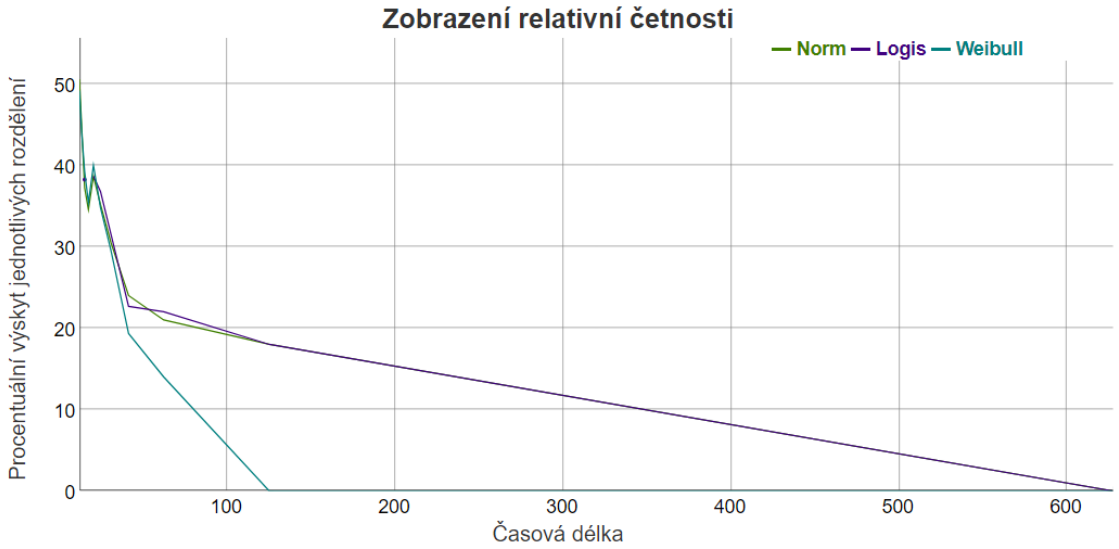
Obr. 9-4 Četnost výskytů jednotlivých rozdělení pro diferenci I4

U diference došlo k výraznému zlepšení pro všechna rozdělení. Při porovnání měření jednotlivých veličin bylo zjištěno několik rozdílností. Data napětí je možné nahradit pouze směrodatnou odchylkou a průměrem. V případě proudu se hodnoty mezi sebou značně lišily. U takového chování by mohlo dojít ke zkreslení výsledku a ztrátě informace.



Obr. 9-5 Četnost výskytů jednotlivých rozdělení pro veličinu avg f

Pro úplnost jsou zde přiloženy ještě grafy frekvence a výkonu. Jedná se o avg f a max P1.



Obr. 9-6 Četnost výskytů jednotlivých rozdělení pro veličinu P1

Z analýzy je patrné, že rostoucí pravděpodobnost výskytu u jednotlivých rozdělení platí pouze pro některé veličiny. V případě této práce pro napětí. U proudu se situace změnila až s úpravou diference. Cílem této kapitoly bylo zanalyzovat chování dat při různě dlouhé časové délce intervalů. Pokud se intervaly dostatečně zkrátí, začnou ve větší míře odpovídat porovnávaným rozdělením. Díky tomu bude možné většinu takových intervalů nahradit pouze směrodatnou odchylkou a průměrem. To by mělo výrazně snížit celkové nároky na ukládání dat. Zároveň by mělo být možné snáze identifikovat odlehlé hodnoty nebo jinak zajímavé vzorky, popřípadě celé intervaly.

Z výše uvedených průběhů je možné sestavit návrh délky intervalu. Nejvhodnější délka ukládání časového úseku je v 30-ti minutových intervalech. Podle kritéria nejvyšších dosažených hodnot je nejvhodnějším rozdělením logistické rozdělení. Logistické rozdělení mělo u všech veličin stabilně nejvyšší zastoupení. U logistického rozdělení zároveň nedocházelo k výrazným výkyvům procentuálního zastoupení. Rozptyl mezi veličinami byl celkem značný. Napětí mělo většinou poměr logistických intervalů kolem 80 % , kdežto proud zhruba 40%. Při nastavení delších intervalů než 60 minut dochází k výraznému poklesu procentuálního zastoupení u obou.

Závěr

V průběhu práce byla pomocí funkcí v jazyku R zpracována data z měření elektrických veličin. Cílem bakalářské práce bylo automatické prověření vlastností úseků o zadané délce. Nalézt vhodnou délku intervalu, kde nedochází ke ztrátám informací. Zároveň určit vhodné rozdělení pro efektivní ukládání. Mezi sledovaná rozdělení patřily normální, logistické a weibullovo.

Pro výpočet byla vytvořena webová aplikace v prostředí Shiny, což byl ostatně jeden z cílů bakalářské práce. Webová aplikace umožnila vizualizaci dat měřených veličin. Dále snadnější práci s daty, jakožto vytvoření tabulek a grafů využitých v praktické části.

V praktické části byla nejprve pospána vstupní data. Práce se zabývala veličinami napětí, proudu, frekvence a výkonu. Měřená data byla v kapitole 8 rozdělena na několik intervalů. Kritériem pro odhadovaný výskyt rozdělení v intervalu bylo překonání hranice 0,05 u p hodnoty. V případě, že daný interval splnil tuto podmínku, určilo se, že interval pravděpodobně obsahuje dané rozdělení.

V hlavní části práce byly v kapitole 9 sledované veličiny rozděleny na sérii intervalů. Postupně se měnila časová délka a s ní počet intervalů. Byl pozorován vliv časové délky na odhadovaný výskyt rozdělení. Pro lepší názornost byla data upravena pomocí diference. Výsledkem bylo zjištění, že rostoucí pravděpodobnost výskytu u rozdělení je dána kratší časovou délkou. Příkladem je změna procentuálního počtu výskytů u napětí. Neplatí to však obecně. U ostatních veličin docházelo ke značnému kolísaní.

Z výše uvedeného chování je možné sestavit optimální délku intervalu. Podle použitých statistických metod je nejvhodnější délka ukládání dat po 30-ti minutových intervalech. Podle provedených výpočtů je nejlepší logistické rozdělení, u kterého bylo dosaženo nejvyššího procentuálního zastoupení.

V bakalářské práci se podařilo splnit všechny vytyčené cíle. Do budoucna by bylo zajímavé provést výpočet pro více rozdělení. Popřípadě se zaměřit na nalezení další metody efektivního ukládání.

Seznam použité literatury

[1] KOŠŤÁKOVÁ, Tereza. Rozptyl, směrodatná odchylka a variační koeficient. *Statistika&My* [online]. Praha: Český statistický úřad, 2017 [cit. 2018-11-30]. Dostupné z: http://www.statistikaamy.cz/2017/01/rozptyl-smerodatna-odchylka-a-variacni-koeficient/

[2] *Hornik, Kurt (2017-10-04). "R FAQ". The Comprehensive R Archive Network. 2.1 What is R?* [cit. 2018-12-2].

[3] LEWIN-KOH, Nicholas. CRAN Task View: Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization. *cran.r-project.org*. 2015-01-07. Dostupné z: https://cran.r-project.org/web/views/Graphics.html [cit. 2018-11-29].

[4] Verzani, John. *Getting Started with RStudio*. O'Reilly Media, Inc. p. 4. ISBN 9781449309039.

[5] RStudio Team. *RStudio, new open-source IDE for R* [online]. RStudio Blog. RStudio, Inc., 2011-02-28 [cit. 2018-11-29]. Dostupné z: https://blog.rstudio.com/2011/02/28/rstudio-new-open-source-ide-for-r

[6] *RStudio IDE features* [online]. RStudio, 2018-07-19 [cit. 2018-11-29]. Dostupné z: https://www.rstudio.com/products/rstudio/features/

[7] GitHub, Inc. [cit. 2018-11-29]. Dostupné z: https://github.com/rstudio/rstudio

[8] Build your first web app dashboard using Shiny and R. *Free Code Camp* [online]. San Francisco, 2018 [cit. 2018-11-30]. Dostupné z: https://medium.freecodecamp.org/build-your-first-web-app-dashboard-using-shiny-and-r-ec433c9f3f6c

[9] Tutorial: creating webapps with R using Shiny. *Science and Technology Corporation* [online]. Olof Palmestraat: Paul Hiemstra [cit. 2018-11-29]. Dostupné z: http://stcorp.nl/R_course/tutorial_shiny.html

[10] GitHub, Inc. [cit. 2018-11-29]. Dostupné z: https://github.com/rstudio/shiny

[11] Marie Laure Delignette-Muller, Christophe Dutang (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. Dostupné z: http://www.jstatsoft.org/v64/i04/

[12] Goodness-of-Fit Test. *Penn State* [online]. State College, c2018 [cit. 2018 11-30]. Dostupné z: https://onlinecourses.science.psu.edu/stat504/node/60/

[13] DUDEK, Martin. Box-Plot neboli Krabicový graf. *Kvalita jednoduše* [online]. Martin Dudek, 2017 [cit. 2018-11-28]. Dostupné z: http://kvalita-jednoduse.cz/box-plot/

[14] DUDEK, Martin. *Kvalitajednoduse.cz* [online]. [cit. 6.12.2018]. Dostupné z: http://kvalita-jednoduse.cz/wp-content/uploads/2017/02/Box1.jpg

[15] Nesymetrie napětí v distribuční soustavě. *ElektroPrůmysl* [online]. Brno: Jaroslav Bubeníček, 2016 [cit. 2018-11-28]. Dostupné z: http://www.elektroprumysl.cz/energetika/nesymetrie-napeti-v-distribucni-soustave

[16] DUDEK, Martin. *Kvalitajednoduse.cz* [online]. [cit. 6.12.2018]. Dostupné z: http://kvalita-jednoduse.cz/wp-content/uploads/2017/02/box2.jpg

[17] The R Graph Gallery. *THE R GRAPH GALLERY* [online]. Yan Holtz, 2017 [cit. 2018-11-28]. Dostupné z: https://www.r-graph-gallery.com/boxplot/#content

[18] Mareš, P., Rabušic, L., Soukup, P. (2015). Analýza sociálněvědních dat (nejen) v SPSS. 1. vyd. Brno: Masarykova univerzita. ISBN 978-80-210-6362-4

[19] ZVÁROVÁ, Jana. *Základy statistiky pro biomedicínské obory.* 2. vydání. Praha : Karolinum, 2011. 219 s. Biomedicínská statistika; sv. I. ISBN 978-80-246-1931-6.

[20] GEIZEROVÁ, Helena, et al. *Epidemiologie : vybrané kapitoly pro seminární a praktická cvičení.* 1. vydání. Praha : Karolinum, 1995. 83 s. ISBN 80-7184-179-X.

[21] John S. deCani & Robert A. Stine (1986). "A note on deriving the information matrix for a logistic distribution". The American Statistician. American Statistical Association. 40: 220–222. doi:10.2307/2684541.

[22] *N., Balakrishnan (1992). Handbook of the Logistic Distribution. Marcel Dekker, New York. ISBN 0-8247-8587-8.*

[23] Shulman, Bonnie (1998). "Math-alive! using original sources to teach mathematics in social context". *PRIMUS* (March): 1–14. doi:10.1080/10511979808965879.

[24] Palei, S. K.; Das, S. K. (2009). "Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach". Safety Science. 47: 88–96. doi:10.1016/j.ssci.2008.01.002.

[25] M. Strano; B.M. Colosimo (2006). "Logistic regression analysis for experimental determination of forming limit diagrams". International Journal of Machine Tools and Manufacture. 46 (6): 673–682. doi:10.1016/j.ijmachtools.2005.07.005.

[26] ZLÁMAL, Filip. *Logisticka regrese v R* [online]. Brno, 2013 [cit. 2018-11-28]. Dostupné z: https://is.muni.cz/th/jkx3p/bakalarska_prace_Filip_Zlamal.pdf?so=nx. Bakalářská práce. Masarykova univerzita, Přírodovědecká fakulta, Ústav matematiky a statistiky.

[27] Weibull W. A statistical distribution function of wide applicability. J. Appl. Mech.-Trans. ASME, 1951; 18 (3): 293–297.

[28] Weibullovo rozdělení náhodných veličin. *Česká společnost pro jakost* [online]. 2016, , 1-41 [cit. 2018-11-28]. Dostupné z: https://www.csq.cz/fileadmin/user_upload/Spolkova_cinnost/Odborne_skupiny/Spolehlivost/Sborniky/Sbornik192_64.pdf

[22] Spojité rozdělení pravděpodobnosti. *Statistika* [online]. Ostrava: Janurová Kateřina, 2011 [cit. 2018-11-28]. Dostupné z: https://homel.vsb.cz/~jan939/STA/cviceni/06_Spojite_rozdeleni_pravdepodobnosti.pdf

[30] PAVLÍK, Tomáš a Ladislav DUŠEK. *Biostatistika* [online]. Brno, 2012 [cit. 2018-11-29]. Dostupné z: https://is.muni.cz/www/98951/41610771/43823411/43823458/44159634/44707073/Pavlik_-_Biostatistika_-_kapitola_3.pdf. Masarykova univerzita, Přírodovědecká fakulta.

[31] DUDEK, Martin. *Kvalitajednoduse.cz* [online]. [cit. 6.12.2018]. Dostupné z: http://kvalita-jednoduse.cz/wp-content/uploads/2016/04/histogram4.jpg

[32] DUDEK, Martin. *Kvalitajednoduse.cz* [online]. [cit. 6.12.2018]. Dostupné z: http://kvalita-jednoduse.cz/wp-content/uploads/2016/04/histogram5.jpg

[33] Zwillinger, Daniel; Kokoska, Stephen (2010). *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press. p. 49. ISBN 978-1-58488-059-2.

[34] *Monti, K.L. (1995). "Folded Empirical Distribution Function Curves (Mountain Plots)". The American Statistician.* ***49****: 342–345.* doi:10.2307/2684570. JSTOR 2684570.

[35] Empirická distribuční funkce. In: *Pravděpodobnost a statistika II* [online]. Brno: Forbelská, 2013 [cit. 2018-11-29]. ISBN 978-80-210-6711-0. ISSN 1802-128X. Dostupné z: https://mathstat.econ.muni.cz/media/12558/emp_dist.pdf

[36] Analyzátor SMC 144. *KMB* [online]. Liberec: KMB systems, c2011 [cit. 2018-11-30]. Dostupné z: http://www.kmb.cz/index.php/cs/digitalni-merici-pristroj-s-pameti/smc-144-pro-smart-metering

[37] KLASCHKA, Jan. *Testování statistických hypotéz* [přednáška k předmětu Zdravotnická statistika 1,2, obor Všeobecné lékařství, 1. lékařská fakulta Univerzita Karlova]. Praha. 26. 4. 2011.

[38] StatSoft Nebojte se p-hodnot!. *StatSoft* [online]. Praha: StatSoft CR, c2004-2018 [cit. 2018-11-30]. Dostupné z: http://www.statsoft.cz/file1/PDF/newsletter/2014_06_26_StatSoft_Nebojte_se_p-hodnot.pdf

[39] Brian G. Peterson and Peter Carl (2018). PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. R package version 1.5.2. https://CRAN.R-project.org/package=PerformanceAnalytics