Grafické vytěžování dat a jeho praktické uplatnění

Lubor Homolka

Bakalářská práce
2008

Univerzita Tomáše Bati ve Zlíně
Fakulta managementu a ekonomiky

Univerzita Tomáše Bati ve Zlíně
Fakulta managementu a ekonomiky
Ústav informatiky a statistiky
akademický rok: 2007/2008

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Lubor HOMOLKA**
Studijní program:	**B 6208 Ekonomika a management**
Studijní obor:	**Management a ekonomika**
Téma práce:	**Grafické vytěžování dat a jeho praktické uplatnění**

Zásady pro vypracování:

Úvod

I. Teoretická část

- **Zpracujte literární rešerši ke grafickému vytěžování dat.**

II. Praktická část

- **Charakterizujte a analyzujte stávající způsoby prezentování a vyhodnocování dat návštěvnosti webových stránek v nakladatelství Martin Stříž, Bučovice.**
- **Navrhněte postupy a zapište nezbytné návody analýz dat užitím vizualizačních a/nebo statistických programů.**
- **Doporučte nakladatelství Martin Stříž, Bučovice, nové postupy a implementaci navržených postupů.**

Závěr

Rozsah práce: **40 stran**

Rozsah příloh:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

[1] GENTLE, James. Handbook of Computational Statistics : Concepts and methods. Berlin : Springer 2004. 1070 s. ISBN 978-3-540-40464-4.
[2] HRONOVÁ, Stanislava, SEGER, Jan. Statistika pro ekonomy. 4. dopl. vyd. Praha : Professional Publishing, 2003. 415 s. ISBN 80-86419-52-5.
[3] KLÍMEK, Petr, RYTÍŘ, Vladimír. Statistické metody pro ekonomy. 1. vyd. Zlín : Univerzita Tomáše Bati ve Zlíně, 2001. 244 s. ISBN 80-7318-013-8.
[4] SAMUELS, Myra L., WITMER, Jeffrey A. Statistics for the Life Sciences . 3rd edition. Prentice : Prentice Hall, 2003. 680 s. ISBN 9780130413161.
[5] SOUKUP, Tom, DAVIDSON, Ian. Visual Data Mining: Techniques and Tools for Data Visualization and Mining. Indianapolis : Wiley Computer, 2002. 416 s. ISBN 0471149993.

Vedoucí bakalářské práce: **Ing. Pavel Stříž, Ph.D.**
Ústav informatiky a statistiky

Datum zadání bakalářské práce: **17. března 2008**

Termín odevzdání bakalářské práce: **23. května 2008**

Ve Zlíně dne 17. března 2008

L.S.

doc. Dr. Ing. Drahomíra Pavelková
děkan

doc. Ing. Rudolf Pomazal, CSc.
ředitel ústavu

ABSTRAKT

Cílem této práce je popsání základních vizualizačních technik užívaných při statickém vyhodnocování dat. Důraz je kladen na využití nekomerčních programů. V praktické části jsou popsány funkce internetové aplikace Google Analytics. Další část se zaobírá regresní a korelační analýzou. V závěru je představena internetová aplikace Rpad, která umožňuje analyzovat soubor dat na principu klient – server.

Klíčová slova:

Vizuální dolování dat, statistické grafy, analýza návštěvnosti internetových stránek, nekomerční statistický software

ABSTRACT

The aim of this work is to describe the basic visualization techniques used in the statistical evaluation of data. The main emphasis is put on the use of non-commercial programs. The practical section describes the functions of the Internet application Google Analytics. Another part deals with correlation and regression analysis. In the conclusion is presented the internet application Rpad that allows us to analyze the data set on a client - server principle.

Keywords:

Visual Data Mining, Statistics graphs, analysis of web page visit rate, noncommercial statistics software

Rád bych touto cestou poděkoval Ing. Pavlu Střížovi, PhD za odborné vedení, nové myšlenky a nápady přesahující obsah této bakalářské práce a za nezměrnou ochotu, která mne provázela po celou dobu naší spolupráce.

OBSAH

ÚVOD

V současné době jsme svědky fenoménu, který se nazývá informační přesycení. Množství dat, které nás obklopuje, mnohdy nečiní náš život jednodušším. Ačkoliv odpadly mnohé bariéry jejich získání, například prostřednictvím Internetu, dostat se k potřebné informaci je pro mnoho lidí nepřekonatelný problém. To může být dáno neschopností nalezení správných zdrojů, nebo neschopností správného porozumění či analyzování dat.

Schopnost převést soubor dat na smysluplné, bezchybné a konzistentní informace je dnes považována za jeden z klíčových faktorů přežití v tržním prostředí. Vlastnictví těch správných informací představuje silnou konkurenční výhodu.

Ve své práci jsem se zaměřil na techniky vizualizace dat. Techniky, které usnadňují orientaci v nepřehledných datových souborech zejména pomocí statistických grafů. S ohledem na požadavek ekonomického principu MINIMAX, tedy získání co nejvíce informací za nejnižší cenu, jsem se snažil využít volně dostupné programy.

I. TEORETICKÁ ČÁST

1 ANALÝZA DAT

Samotný proces získání informací je mnohdy velmi jednoduchý. Počítačem řízené výrobní linky poskytují zprávy o průběhu činnosti, oznámení o pohybu akcií a jejich cen je k dispozici téměř okamžitě po provedené operaci. Mnoho dalších činností, které jsou prováděny pomocí výpočetní techniky, poskytují standardizovaná data. Většina takto získaných dat ovšem nemá pro koncového uživatele skutečnou hodnotu. Z těchto dat lze získat pouze rámcovou představu. Z dat se stanou hodnotné informace až ve chvíli, kdy se dají využít ke zlepšení stávající situace. Proces získávání informací z těchto dat se dá shrnout pod název analýza dat.

1.1 Data mining

Data mining, někdy též mylně označován jako analýza dat, je metodologie typicky užívaná ke sledování chování objemných datových souborů. K odhalování základních principů nebo odchylek, které by měly být dále podrobněji analyzovány. Tyto podrobnější analýzy by měly být řešeny specifickými metodami, odpovídajícími konkrétnímu problému. Tyto problémy, díky své rozmanitosti, není možné převést na automatický algoritmus. A proto se jimi musí zabývat odborníci. Pro všeobecnou představu o charakteristikách datového souboru lze ovšem tyto mechanismy zavést – v podobě mechanického učení nebo umělé inteligence. Dá se tedy říci, že data mining je metodologie, která užívá automatizovaných technik za účelem nalezení relevantních částí datových souborů.

[1]

1.2 Průzkum dat (Data exploration)

Oproti data miningu je průzkum dat metodologie, která užívá manuální techniky k porozumění specifických problémů. Automatizované metody data miningu jsou limitovány standardizovanou podobou, šablonou dat. S průzkumem dat je spojena mnohem větší variabilita datového souboru. Tyto datové soubory jsou oproti těm, kterými se zaobírá data mining, mnohem menší.

[2]

2 ZÁKLADNÍ PŘÍSTUPY ANALÝZ DAT

Na analýzu dat lze pohlížet ze tří základních pohledů.

1. Reduktivní analýza
2. Matematická analýza
3. Visuální analýza

2.1 Reduktivní analýza

Tato analýza je založena na metodologii, ve které individuální fakt, nebo skupina faktů, je považována za základ pro analýzu. Tato analýza v sobě zahrnuje základní statistické metody a souhrny. S tímto typem analýzy se setkáváme v běžném životě zřejmě nejčastěji. Příkladem je konstatování výše průměrné čisté mzdy, aniž by bylo zmíněno značně vychýlené mzdové složení celé populace. Jedná se o nejjednodušší možnou analýzu dat.

2.2 Matematická analýza

Tato analýza, někdy označována za klasickou, je založena na aplikaci matematických modelů jako základu pro analýzu. Současným trendem je aplikovat model a poté testovat jeho přesnost a správnost. Součástí těchto analýz jsou komplexní statistické a Bayesovské metody. Matematické modelování je důležitou technikou studia dat, protože umožňují redukovat množství nekontrolovatelných dat, která brání odhalení takových atributů v celkové populaci, jakými jsou například předpoklad normality nebo linearity.

Užití této analýzy klade vysoké nároky na uživatele. Mnoho začínajících uživatelů tyto metody užívá (a třeba i správně), ovšem problémem je jejich správná interpretace. Porozumění intervalovým odhadům, testování statistických hypotéz, regresní a korelační analýze je naprostou nezbytností pro smysluplné užití tohoto typu analýzy dat.

2.3 Visuální analýza

Tato metodologie považuje celý soubor dat za základ analyzování. Ne všechna data lze popsat matematickými modely, a někdy je mnohem přínosnější spolehnout se na grafické řešení. Typickým příkladem, kdy je účelnější visuální podoba dat, je technická analýza. Tu lze použít například k předpovědi chování trhu s akciemi.

Visuální analýza je obzvláště silná díky:

- Naší vrozené schopnosti interpretovat data holisticky.

- Odhaluje atributy dat (struktura, trend), která se jen obtížně vyčtou z modelů.
- Pomocí grafického výstupu jsme schopni dedukce, úsudku o modelu, zejména o nematematickém modelu.

[1]

3 DATOVÉ SOUBORY

Aby bylo možné s daty operovat, je nutné nalézt vhodný prostředek pro jejich přenos. Vzhledem k velkému množství analytických programů a celosvětovému šíření informací bylo nutné tyto datové soubory standardizovat. V následujícím výčtu je popis základních a nejvíce užívaných datových souborů.

3.1 Formát TSV (The tab separated values)

Tento formát je textovým formátem, který umožňuje převod mezi aplikacemi, které využívají různé interní formátování. Tento formát je standardizován a oficiálně registrován jako Internet media Type (MIME type) pod jménem text/tab-separated-values. Důležitou výhodou tohoto formátu je skutečnost, že data uložená v tabulce lze zobrazit v klasickém textovém editoru.

[3]

3.2 Formát CSV (Comma separated values)

Tento formát je užíván k přenosu dat, zejména mezi databázemi. Každý řádek obsahuje několik záznamů, které jsou nejčastěji odděleny jednoduchými uvozovkami, čárkou či středníkem. Každý řádek musí být označen dvojitými uvozovkami na konci a na začátku, pokud jsou data oddělena jednoduchými uvozovkami.

[4]

3.3 XML (Extensive markup language)

Tento datový formát byl původně navržen k publikování rozsáhlých datových zdrojů. Velmi důležitou vlastností je jeho aplikace v SQL databázích.

[5]

3.4 PDF (Portable document format)

PDF je souborový formát, který slouží k zobrazení dokumentů nezávisle na použitém softwarovém a hardwarovém zařízení nebo operačním systému. PDF dokument se skládá ze souboru objektů, které ve vzájemné součinnosti popisují výstup dat na jedné či více stránkách, který může být doplněn interaktivními, či zvýrazňujícími prvky. Soubor PDF obsahuje objekty, které vytváří informační strukturu, která je reprezentována sekvencí baj-

tů. Navíc, k popisu statického zobrazení, PDF soubor může obsahovat interaktivní elementy, které jsou možné pouze v elektronické podobě. PDF podporuje potřebný základ k mnoha takovým objektům, například k hypertextovému odkazu, zvukovým přílohám nebo k přehrávání videosekvencí.

[6]

4 GRAF

Velmi důležitou formou zobrazování statistických dat jsou grafy. Oproti statistickým tabulkám poskytují rychlou a přehlednou představu o charakteristických rysech a trendech analyzovaných dat. Graf je vzájemný vztah dvou nebo více proměnných veličin pomocí přehledných symbolů. Pod pojmem grafický symbol si lze představit schematické obrázky, číslice, matematické značky nebo barvy.

Primitivním grafem rozumíme graf, který není možné dále rozložit. Složený graf lze zobrazit pomocí více primitivních grafů. Každý primitivní graf má svou vlastní kapacitu vzhledem k počtu měření a k počtu dimenzí.

Univariate - Grafy, které zobrazují vlastnosti jedné náhodné veličiny.

Bivariate - Grafy, které zobrazují vztah dvou náhodných veličin.

Trivariate - Grafy, které zobrazují vztah tří náhodných veličin.

Multivariate - Souhrnný název pro grafy, které zobrazují více než jednu náhodnou veličinu.

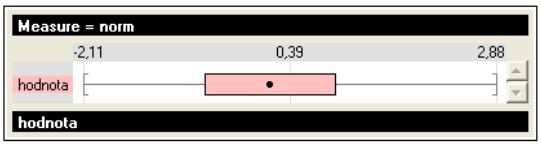
[1]

Mezi tyto primitivní grafy řadíme:

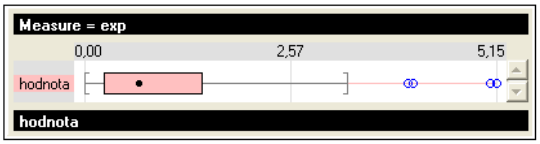
- Krabicový graf (Boxlpot)
- Histogram
- Stem and leaf graf
- Výsečový graf (Pie chart)
- Rozptylový graf (Scatter plot)
- Kvantilový graf (Quantile plot)
- QQ graf (Quantile – Quantile plot)
- Grafy rozpětí (Rangle plot)
- Graf časové řady (Time series plot)

4.1 Krabicový graf (Boxplot)

Jedná se o univariate graf, který statisticky popisuje rozdělení souboru hodnot pomocí variačního rozpětí, kvantilového rozpětí a střední hodnoty-mediánu. Použitím tohoto grafu získáme velmi rychle představu o rozdělení souboru. Proto je vhodný pro přímé porovnávání dvou a více souborů.



Obrázek 1 Krabicový graf (VisiCube)

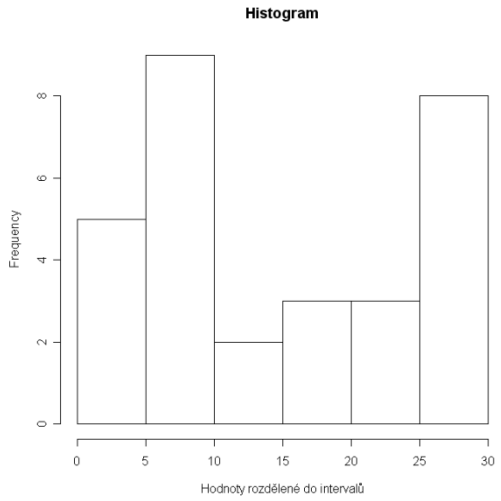


Obrázek 2 Krabicový graf s odlehlými hodnotami (VisiCube)

4.2 Histogram

4.2.1 Základní funkce histogramu

Pro grafické vyjádření intervalového rozdělení četností se používá histogram četností. Tento graf získal své jméno roku 1895, kdy ho tak pojmenoval slavný statistik Pearson. Je to graf, který je tvořen čtyřúhelníky, jejichž základna představuje interval hodnot a jejichž výška představuje velikost třídních četností.



Obrázek 3 Histogram (R)

Při tvorbě histogramu je důležité správné určení hranic, velikosti intervalů. Nestejné intervaly volíme u takových znaků, kde se jejich četnosti vyvíjí nesymetricky. Meze intervalů jsou tedy určitými vzájemnými násobky. Například, pokud logaritmujeme stupnici určitého levostranně vychýleného znaku, toto rozdělení se stane více symetrické. Pro stanovení počtu stejně velkých intervalů používáme některá pravidla.

a, $s \doteq 1 + 3{,}3 \log n$ (Stugarsovo pravidlo)

b, $s \doteq 5 \dot{\log} n$

c, $s \doteq \sqrt{n}$

s = počet intervalů

n = celkový počet údajů

[7]

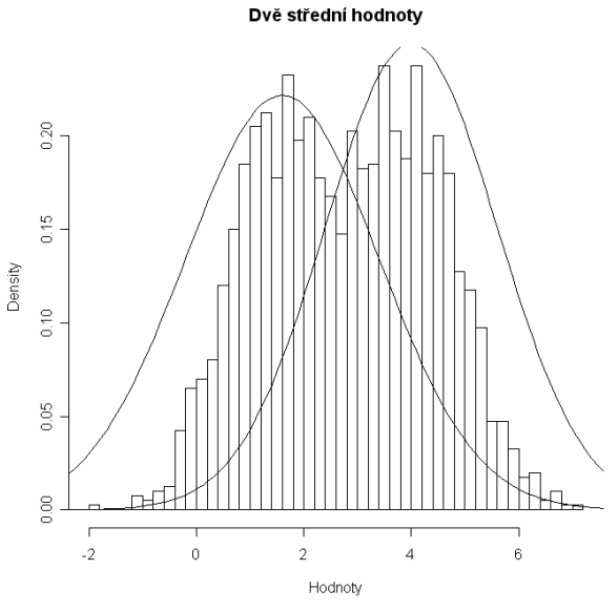
4.2.2 Pokročilé funkce histogramu

Histogram nám ovšem svou grafickou podobou nesděluje pouze četnosti výskytu, ale mnohdy též mnohem důležitější informace. Lze z něj vyčíst:

1. Přibližnou střední hodnotu
2. Další vysoký vrchol značí potenciální další střední hodnotu
3. Rozptyl
4. Symetričnost

5. Špičatost
6. Lze srovnat s předpokládaným rozdělením

Je důležité si uvědomit, že při prokládání histogramu předpokládanou křivkou hustoty pravděpodobnosti musíme změnit osu Y. Ta již nebude zobrazovat četnosti absolutně, ale pravděpodobnosti výskytu.
I přes vysokou vypovídací hodnotu histogramu nesmíme utvářet závěry o multimodalitě pouze z grafu. Blíže v příloze III.



Obrázek 4 Histogram bimodálních dat proložený
Gaussovými křivkami (R)

4.3 Stem and leaf

Jedná se variaci histogramu. První číslice před svislou linkou znamenají základní řád čísla (desítky, stovky), za linkou jsou jednotky tohoto čísla.
Následující řada čísel se pomocí stem and leaf grafu interpretuje následujícím způsobem:

A={5,6,8,11,15,18,19,22,26,31}

0 | 568

1 | 1589

2 | 26

3 | 1

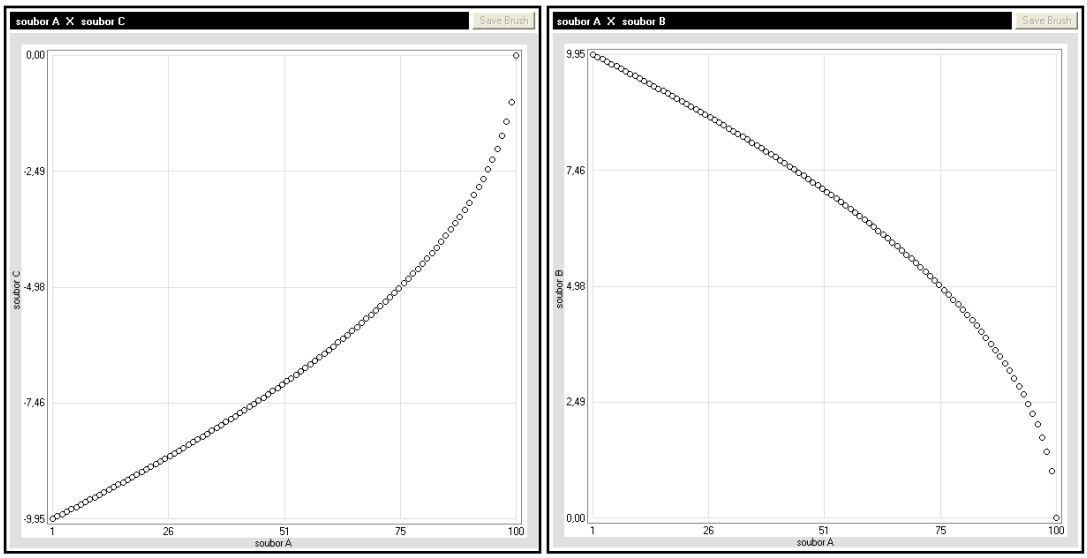
[8]

4.4 Výsečový graf (Pie chart)

Výsečový graf je způsob prezentace kategoriálních dat. Zobrazuje podíl jednotlivých kategorií na celku. Mezi statistiky je značně nepopulární zejména z důvodu nepřehlednosti popisu a velikosti jednotlivých výsečí. Každý výsečový graf lze znázornit takzvaným Dot chart – bodovým grafem. Tento graf je podobný horizontálnímu histogramu.

4.5 Rozptylový graf (Scatter plot)

Rozptylový graf je multivariate graf, protože zobrazuje vztah dvou veličin. Každá vztahová veličina je zobrazena na své vlastní ose. Tento graf je velmi užitečným zejména při zkoumání korelací veličin. Graf silné korelace je charakteristická stálým růstem (přímá) či klesáním (nepřímá) trendu. Na následujících grafech vidíme silnou korelaci mezi souborem A a C.



Obrázek 5 Zobrazení přímé a nepřímé korelace (VisiCube)

Velikost této korelace lze vyjádřit pomocí Pearsonova korelačního koeficientu:

$$r_{x,y} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * (y_i - \bar{y})^2}} \quad (1)$$

První graf vystihuje přímou korelaci, $r_{A,C} = 0{,}976$

Druhý graf vystihuje nepřímou korelaci $r_{A,B} = -0{,}976$

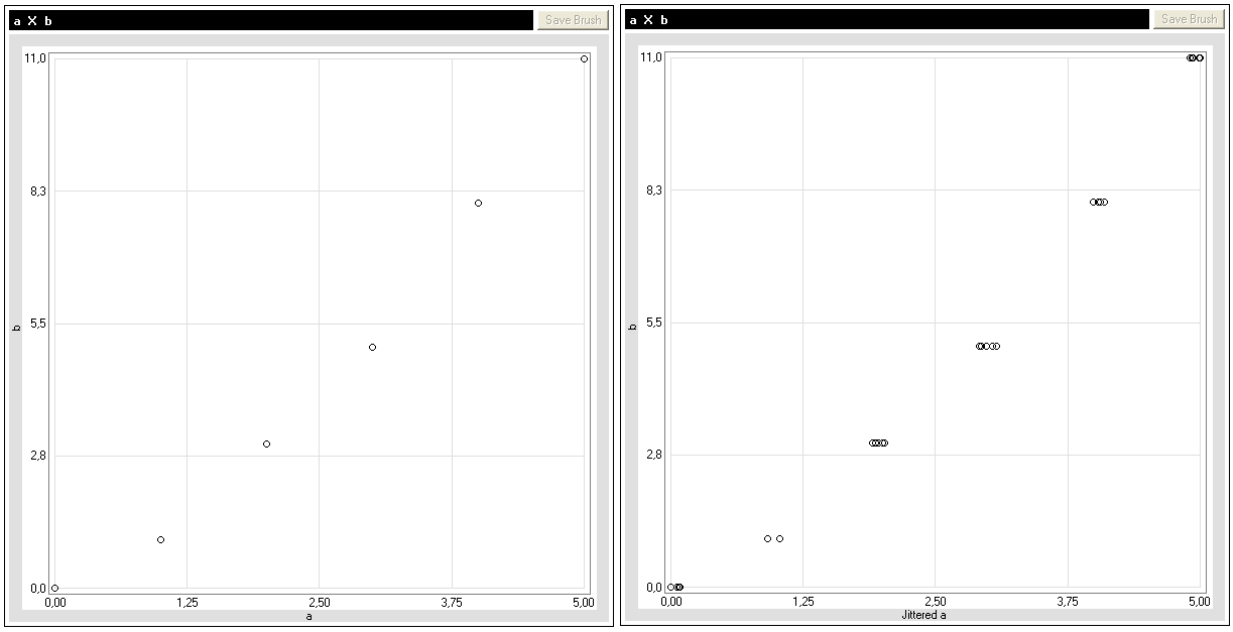
Dalším důležitým aspektem zkoumání vzájemného vlivu dvou proměnných je zakřivení. Neexistuje žádný automatický test, kterým bychom vzájemný vztah zakřivení odhalili. Pearsonovým koeficientem dokážeme změřit lineární vztah, některými neparametrickými korelačními testy, jakým je například Spearmanův korelační koeficient lze měřit nelineární závislost, ovšem pouze u monotónních vztahů. Prozkoumáním rozptylového grafu nám umožní identifikovat tvar vztahu, což nám pomůže při výběru vhodného modelu.

4.5.1 Jittering

Zejména u rozptylových grafů vyvstává problém se zobrazováním hodnot. Problémem je výskyt několika stejných hodnot. Tyto hodnoty jsou poté graficky interpretovány jako jeden bod. Na první pohled tedy není jasné, kde je největší hustota hodnot.

K odstranění tohoto problému slouží metoda zvaný Jittering. Jedná se o vizualizační techniku, jejíž podstata spočívá v přidání malého množství stejného, ale náhodného šumu před samotným vykreslením grafu. To má za následek, že body jsou v mírně odlišných pozicích, než jak je tomu ve skutečnosti. Ačkoliv se záměrně dopouštíme určitého zkreslení tyto body nebyly vůbec viděny, takže dochází k zlepšení vypovídací hodnoty grafu.

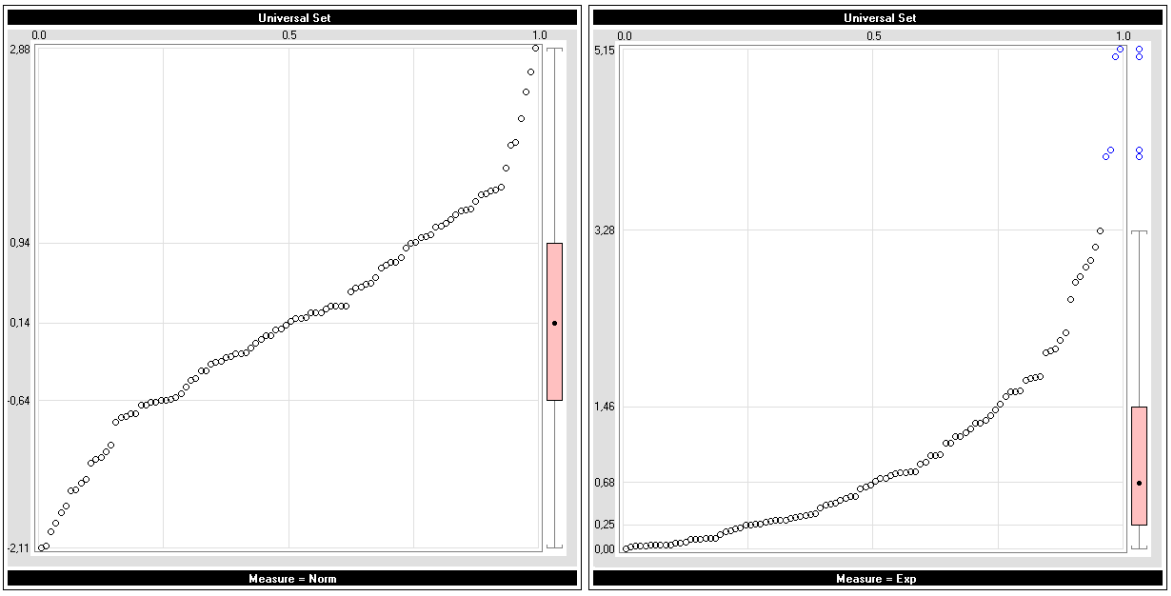
[1]



Obrázek 6 Příklad aplikace Jittering(u) (VisiCube)

4.6 Kvantilový graf (Quantile plot)

V tomto typu grafu je zobrazeno statistické rozdělení náhodné veličiny pomocí distribuční funkce. Tento graf zobrazuje minimální hodnoty, maximální hodnoty, hodnotu dolního kvartilu, hodnotu horního kvartilu a mediánu.

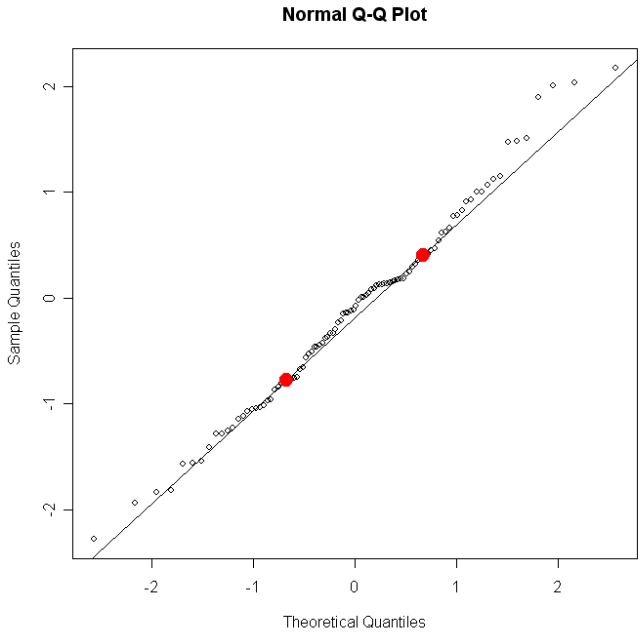


Obrázek 7 Zobrazení distribuční funkce normálního rozdělení (vlevo) a exponenciálního(vpravo) pomocí kvantilového grafu (VisiCube)

Grafy znázorňující studentovo a χ^2 rozdělení jsou uvedeny v příloze II.

4.7 QQ graf (Quantile- Quantile plot)

Jedná se o graf, jenž se využívá při hledání nejvhodnější distribuční funkce, která by nejpřesněji popsala soubor dat. Červenými body jsou označeny hranice dolního a horního kvartilu. Na stupnicích os lze odečíst střední hodnotu a počet směrodatných odchylek od střední hodnoty.

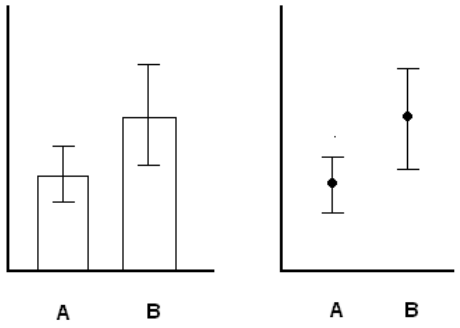


Obrázek 8 Kvantilový graf normálního rozdělení s vyznačenými kvartily (R)

4.8 Grafy rozpětí (Range plot)

Tento typ grafů je svým zobrazením velmi podobný krabicovému grafu. Plní ovšem jinou funkci. Zobrazuje rozsah hodnot, nebo chybové úsečky, související s konkrétní naměřenou hodnotou, a to ve formě krabicového grafu, nebo takzvaných whiskers (vousů). Oproti krabicovému grafu ovšem velikost chybových úseček není vypočítána z celkových dat, ale je definována uživatelem. (například +-2%)

V praxi je běžné dvojí zobrazení.



Obrázek 9 Dvojí zobrazení Range plot (vlastní)

Horizontální grafy rozpětí

Tento typ grafů rozpětí využíváme zejména v situacích, kdy hodláme zobrazit hodnoty nezávislých faktorů při jejich vzájemném porovnávání.

Vertikální grafy rozpětí

Vertikálně položený graf rozpětí obvykle slouží k zaznamenávání pohybu cen na trhu, vývoj předpokládaných tržeb apod.

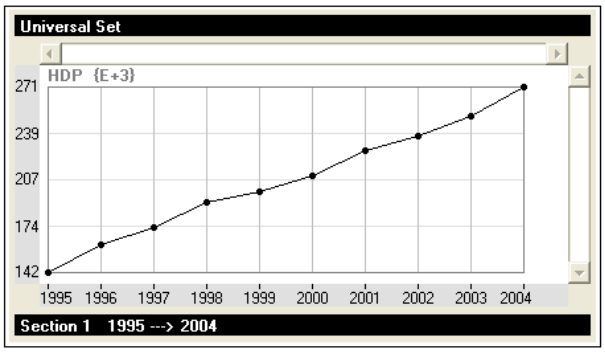
Speciálním typem rozpěťových grafů je graf burzovní, který byl vytvořen pro potřeby akciového trhu. Do tohoto grafu je možné zahrnou minimální, maximální, konečnou cenu a množství obchodovaných akcií.

Graf rozpětí je detailně popsán v nápovědě programu Statistica 7. V této nápovědě jsou uvedeny i ostatní grafy, zmiňované v této kapitole.

[9]

4.9 Graf časové řady (Time series plot)

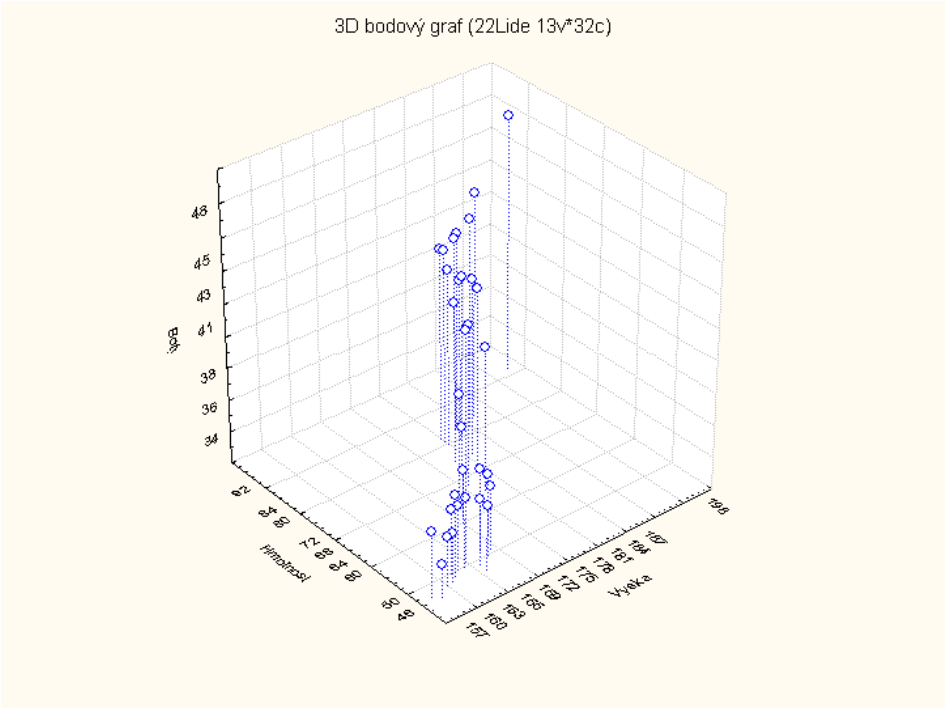
Jedná se o univariate graf, ve kterém jsou zaznamenány naměřené hodnoty v čase. Tento graf je využíván pro základní odhadnutí vývoje, trendu časových řad.



Obrázek 10 Graf časové řady (VisiCube)

4.10 Rozptylový 3D graf

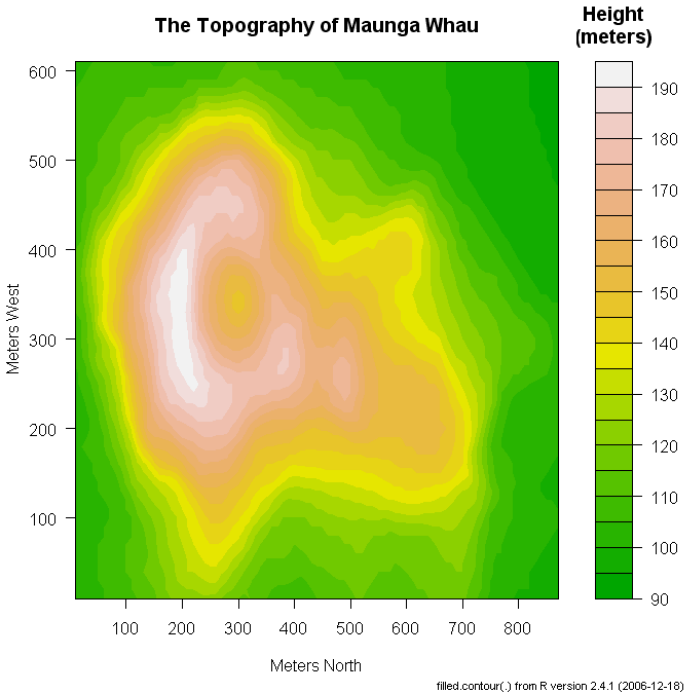
Trojrozměrný rozptylový graf lze využít k prozkoumání závislostí, princip je stejný jako v případě dvourozměrného rozptylového grafu.



Obrázek 11 3D rozptylový graf (Statistica); data použita z Militký

4.11 Povrchové grafy

Trojrozměrný soubor dat je zobrazen pomocí dvou os, třetí osa je definována škálou barev či odstínů. Typickým příkladem využití tohoto typu grafu v praxi jsou mapy.

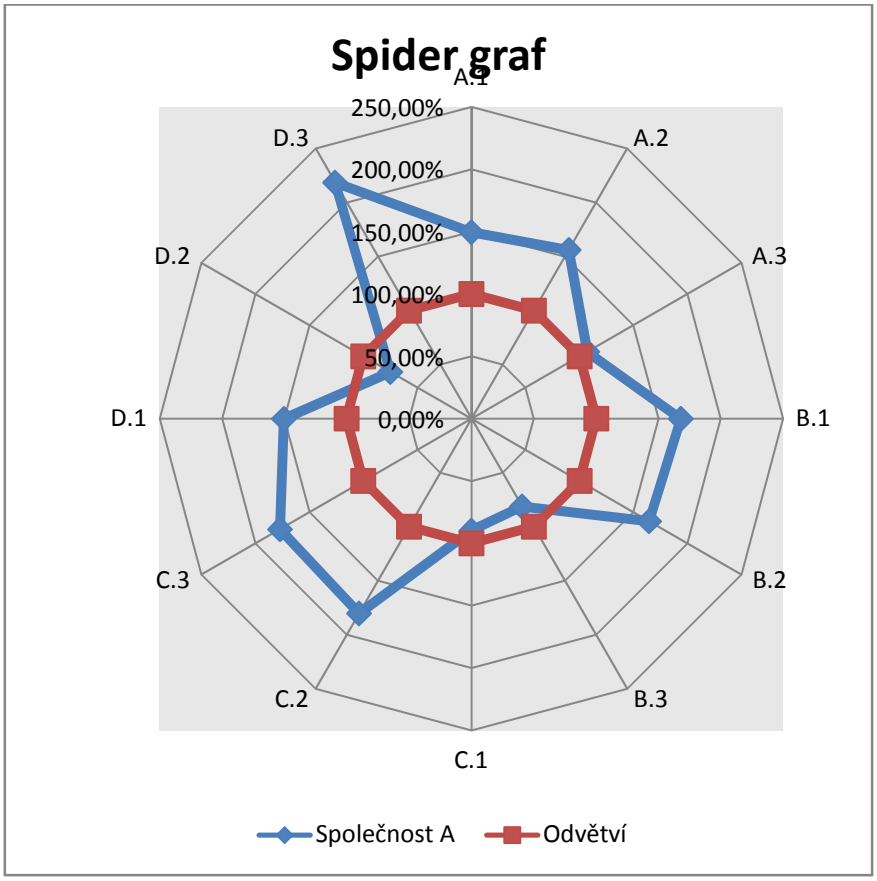


Obrázek 12 Povrchový graf v programu R

4.12 Pavučinový graf (Spider plot)

Tento typ grafu slouží k rychlému zhodnocení, porovnání většího množství ukazatelů. Tento typ grafu se používá zejména v oblasti finančních a makroekonomických analýz (Magický n-úhelník)

[10]



Obrázek 13 Pavučinový graf ekonomických ukazatelů A.1 - D.3 (MS Excel)

4.13 Symbolové grafy

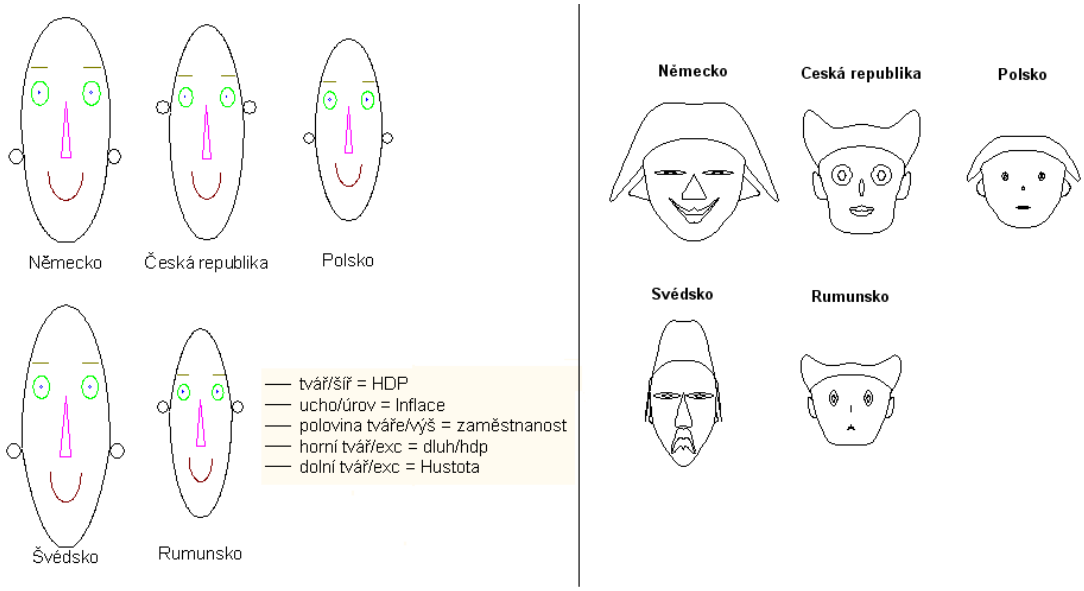
V situaci, kdy srovnáváme větší množství znaků, je vhodné namísto pavučinového grafu využít takzvaných symbolových grafů. Jednotlivé znaky jsou převedeny do určitých geometrických tvarů nebo symbolů. Vyhodnocení dat je poté možné pouhým srovnáním, hledáním podobných obrazců. Nejčastěji užívané symboly podle Melouna a Militkého:

1. Profilové sloupce
2. Profilové křivky

3. Chernoffovy obličeje
4. Profily znaků
5. Sluníčka – polygony
6. Hvězdičky – polygony

[11]

Všechny uvedené typy grafů lze vytvořit v komerčním programu Statistica. V programu R lze vytvořit Chernoffovy obličeje. Na následujících grafech jsou vyhodnoceny stejné makroekonomické údaje.



Obrázek 14 Cheroffovy obličeje v programech Statistica(vlevo) a R

5 NEKOMERČNÍ STATISTICKÝ SOFTWARE

Nekomerční software vzniká především pro účely vědecké nebo výukové. Vzhledem ke snadné přístupnosti a faktu, že jsou zcela zdarma nebo za drobný poplatek, je ve světě tento typ programů velice populární. Díky své vysoké popularitě dochází k aktualizaci nejen ze strany vývojářů, ale i z řad běžných uživatelů. Některá nekomerční programy jsou ovšem již na tak vysoké úrovni, že svými vlastnosti předčí i svou komerční konkurenci. U převážné většiny programů neexistuje oficiální technická podpora, která je k dispozici při řešení každého problému spojeného s užíváním či instalací programu. Na straně druhé, technická pomoc přichází od uživatelů, kteří se pohybují zejména na určených diskusních fórech. Oproti komerčním programům většinou nikdo z tvůrců nenese za (nejen špatné) užívání a za výsledky zodpovědnost.

5.1 R

R je programovací jazyk a prostředí pro statistické vyhodnocování a tvorbu grafických výstupů.

R poskytuje širokou paletu statistických (lineární, nelineární modelování, klasické statistické testy, analýzu časových řad, klasifikace, analýzu klastrů atd.) a grafických technik. Jednou z největších výhod prostředí R je jednoduchost, se kterou se vytváří velmi sofistikované a kvalitní grafické výstupy, které, pokud je třeba, mohou zahrnovat matematické symboly a formule. Jeho jednoduchá rozšiřitelnost z něj činí program použitelný k řešení téměř všech disciplín statistiky. Možný je též export výsledků do grafických programů (například GNUplot) nebo do TeXu k dalšímu zpracování.

Jedná se o GNU projekt. Je podobný jazyku a prostředí S, které bylo vyvinuto Bellovými laboratořemi. R je ovšem považováno za odlišný způsob implementace jazyka S. Ačkoliv je mezi nimi mnoho rozdílů, většina kódů psaných v S fungují i v R. Prostředí S je chápáno jako hnací motor k utváření statistické metodologie výzkumníky. R je poskytováno jako Open Source, čímž poskytuje běžným uživatelům možnost spolupodílet se na vývoji.

Vývojáři tohoto produktu o R nemluví pouze jako o statistickém programu. Spíše jako o prostředí, do kterého je možné implementovat statistické techniky. Tyto techniky se implementují pomocí přídavných balíčků, které vytváří samotní uživatelé a vývojáři. V základní instalaci nalezneme omezené množství, které je ovšem dostačující pro základní sta-

tistickou analýzu. Ostatní specifické balíčky je možné stáhnout na serverech (CRAN) rozmístěných po celém světě.

Někteří uživatelé mohou považovat za zásadní nedostatek grafické prostředí. Jak jsem se již zmínil, statistické výpočty a vykreslování grafů je prováděno přes příkazy z příkazového řádku. Pracovní prostředí je tedy omezeno na klasickou nabídku základních provozních funkcí, nabídku instalace balíčků a nápovědy.

[12]

5.2 Rpad

Rpad je interaktivní analytický program, který slouží jako webové rozhraní pro program R. Stránky Rpadu jsou typem pracovního listu založeného na zdrojovém kódu R. Rpad je tedy analytický balíček a nástroj pro design webových stránek v jednom. Díky Rpadu je možné jednoduchým způsobem sdílet složité statistické analýzy vytvořené v R, a to nejčastěji pomocí intranetu. Koncový uživatel nemusí mít nainstalováno, kromě webového prohlížeče, vůbec nic. Z pohledu tohoto uživatele k Rpadu nepotřebuje žádnou dokumentaci, protože takto vytvořená webová stránka má již předem nadefinovány postupy analýzy a nezbytné skripty k jejich spuštění.

Existují dvě verze Rpadu:

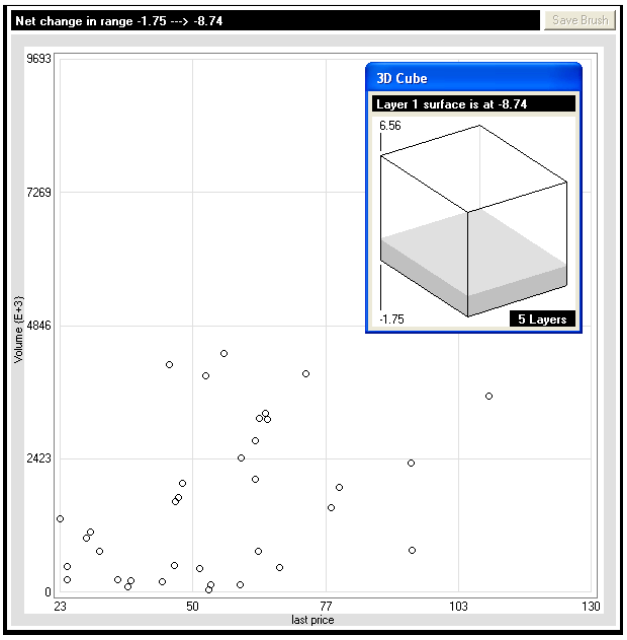
1. Local version, která využívá místní instalace
2. Intranet/Internet version , která pracuje na principu klient/server

[13]

5.3 VisiCube

Jedná se o nekomerční program společnosti Datamology. Je určen primárně pro vizuální analýzu. Neobsahuje žádné mechanismy matematického modelování. Tento program je založen na předem definovaných grafech. Jedinou úlohou uživatele je tedy správně definovat zdroj dat a určit datový typ a následně vyhodnotit grafický výstup.

V závislosti na datovém typu sloupců se vytvoří sada grafů, které lze velmi jednoduchým interaktivním způsobem obsluhovat. K dispozici jsou univariate, ale i multivariate grafy. Třírozměrné grafy jsou řešeny pomocí vrstevnic. Zobrazuje tedy rozptylové grafy (osy x, y) podle vertikální osy (osa z), která je rozdělena do uživatelem zvoleného počtu intervalů. Pro snazší pochopení je k dispozici animace, která tyto vrstvy zobrazuje.



Obrázek 15 Vrstvený rozptylový graf (Visicube)

Práce v tomto programu je pro mnoho uživatelů jednoduchá. Program pracuje ve třech režimech:

1. Sources - definuje zdroj dat
2. Projects - určení datových typů měření (number, date, string) a dimenzí
3. Explore - samotný průzkum dat pomocí předem definovaných grafů

Nevýhodou je absence matematického výstupu.

[1]

5.4 Tanagra

Jedná se o program, který slouží k základním statistickým analýzám. Byl navržen jako výukový program. Umožňuje jak parametrickou tak neparametrickou statistiku (znaménkový test, Kruskall -Wallis test, Mann-Whitney test a další), analýzu shluků nebo rozhodovací stromy. Za velmi užitečné považuji navazování jednotlivých kroků analýzy metodou drag and drop, kdy se automaticky vytváří schéma postupu analýzy. Je tedy velmi jednoduché zkontrolovat dosavadní postup a případně navázat na již dříve provedené kroky alternativní analýzou. Celkový postup analýzy je poté velmi přehledný. Za největší nevýhodu považuji absenci ucelené nápovědy. Na uživatele jsou kladeny znalosti statistických metod. Oproti R je množství dostupných analýz minimální, ovšem pro běžného uživatele dostačující.

[14]

5.5 Google analytics

Google Analytics je volně poskytovaná služba společnosti Google. Nabízí detailní základní statistiky o návštěvnících webových stránek. Vyvinula se z komerčního produktu Urchin společnosti Urchin software, kterou v roce 2005 společnost Google koupila. Google Analytics pracuje na platformě JavaScript. Pomocí tohoto skriptu se uživatel připojuje na centrální server Google. Google Analytics se tedy jako takový neinstaluje. Uživatel potřebuje mít k dispozici pouze webový prohlížeč, který podporuje JavaScript.

Největší výhodou oproti ostatním nástrojům vytvořených k analyzování internetových přístupů je dokonale propracovaný systém AdWords advertisement. Ten umožňuje vytváření marketingových kampaní a cílených reklam v návaznosti na výsledky analýz. Shromažďuje informace o původu návštěvníka, jeho čas strávený na stránkách, způsob, jak se na stránky dostal nebo o jeho geografické poloze.

Administrátor systému má možnost definovat si cíle. Tyto cíle mohou obsahovat zvýšení obratu, tržní pozici, sledovanost konkrétní stránky nebo popularitu stahovaného souboru. Pomocí tohoto nástroje je poté možné určit, která reklama nebo odkaz má žádoucí efekt.

Současnou tendencí v rozvoji Google Analytics je sdílení dat. Toto sdílení slouží k vzájemnému poměřování úspěšnosti internetové komunikace konkurence se zákazníky. Poměřují se s nejlepšími v oboru, jedná se o takzvaný benchmarking.

[15]

6 KOMERČNÍ STATISTICKÝ SOFTWARE

Komerční produkty jsou charakteristické zejména širokým okruhem statistických analýz. Velmi často bývá implementována průmyslová statistika, neuronové sítě, klastrová analýza, kanonická analýza, pokročilé nelineární modely a další.

Velký důraz je kladen na uživatelské prostředí, které by měl dokázat intuitivně obsluhovat i méně zkušený uživatel. Ve většině případů je k dispozici oficiální statistický rádce a případná technická pomoc.

6.1 Minitab

Aplikace Minitab vznikla především pro začínající a příležitostné uživatele statistických analýz. Její ovládání je intuitivní. Obsahuje ovšem mnoho dalších nástrojů, které z Minitabu činí silný nástroj, zejména co se týče kontroly kvality, statistického řízení procesů nebo analýzy spolehlivost/přežití.

6.2 Statistica

Produkt společnosti StatSoft nabízí širokou paletu statistických nástrojů, určených především pro analýzu dat, ale i pro data mining. Umožňuje též správu databází nebo tvorbu nových uživatelských aplikací využívající jádro programu Statistica. Aplikaci Statistica je možné využívat i k analýzám přes webové rozhraní. Nespornou výhodou pro uživatele je pracovní prostředí v českém jazyce. Velmi podrobný statistický rádce a návody k užívání programu jsou ovšem v angličtině.

6.3 MS Excel

Produkt společnosti Microsoft není pravým zástupcem statistického softwaru. Jedná se o takzvaný Spreadsheet program neboli tabulkový kalkulátor. Ačkoliv obsahuje základní statistické funkce, vyniká především ve fázi přípravy dat, nebo při rychlém orientačním vyhodnocování. Jeho další výhodou je schopnost pracovat s velkým množstvím datových souborů. Jeho masivní rozšíření a fakt, že se jeho užívání vyučuje téměř na všech stupních škol, z něj činí nejpoužívanější program pro analýzu dat.

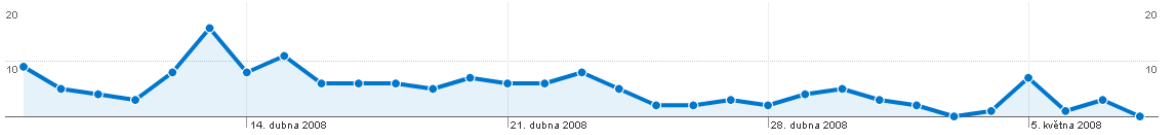
II. PRAKTICKÁ ČÁST

7 SOUČASNÝ ZPŮSOB VYHODNOCOVÁNÍ NÁVŠTĚVNOSTI

V současné době nakladatelství Stříž využívá k analýze návštěvnosti webových stránek internetovou aplikaci Google Analytics. Tato služba se aktivuje vložením zdrojového kódu, do kterého je vloženo předem automaticky generované identifikační číslo sledovaných webových stránek. Pomocí GA je možné sledovat následující oblasti:

Počet návštěvníků:

Jedná se o základní a nejdůležitější faktor, který je zároveň měřítkem úspěšnosti internetové prezentace. Na základě identifikace jedinečného identifikátoru internetového protokolu (IP) je možné identifikovat nové, ale i vracející se návštěvníky.



Obrázek 16 Časová řada

Zobrazení stránek:

Důležitým faktorem pro správné směřování internetové reklamy je četnost otevření konkrétní stránky internetové prezentace. Pro administrátora webové prezentace je to také důležitý údaj. Sledováním těchto vytíženějších stránek lze předcházet nadměrnému zatížení serveru a webové prezentace.

Průměrné zobrazení stránek:

Měřítkem interakce mezi návštěvníkem a internetovými stránkami je průměrný počet zobrazených stránek. Nelze říci, že velký počet průměrně zobrazených stránek je lepší než počet malý. Stránky mohou být výborně strukturovány a návštěvník okamžitě nalezne, co potřebuje. Na straně druhé pokud stránky nabízejí kvalitní obsah, návštěvník nemusí zůstat pouze u toho, co původně hledal.

Čas na stránkách:

Samotný údaj o počtu návštěvníků je nedostatečný. Vysoká hodnota může znamenat spokojenost s poskytovanou prezentací. Nelze ovšem opomenout skutečnost, že mnoho internetových uživatelů užívá takzvané internetové záložky (okna). V těchto záložkách jsou otevřeny internetové prezentace, i když je uživatel aktivně nevyužívá. Na straně druhé uživatelé, kteří jsou na stránky odkázáni pomocí internetových vyhledávačů, mohou stránku

opustit během několika vteřin. (Například kvůli nepřehlednosti webové prezentace, nekompatibility, zdlouhavému načítání obrázků a podobně).

Míra odchodů:

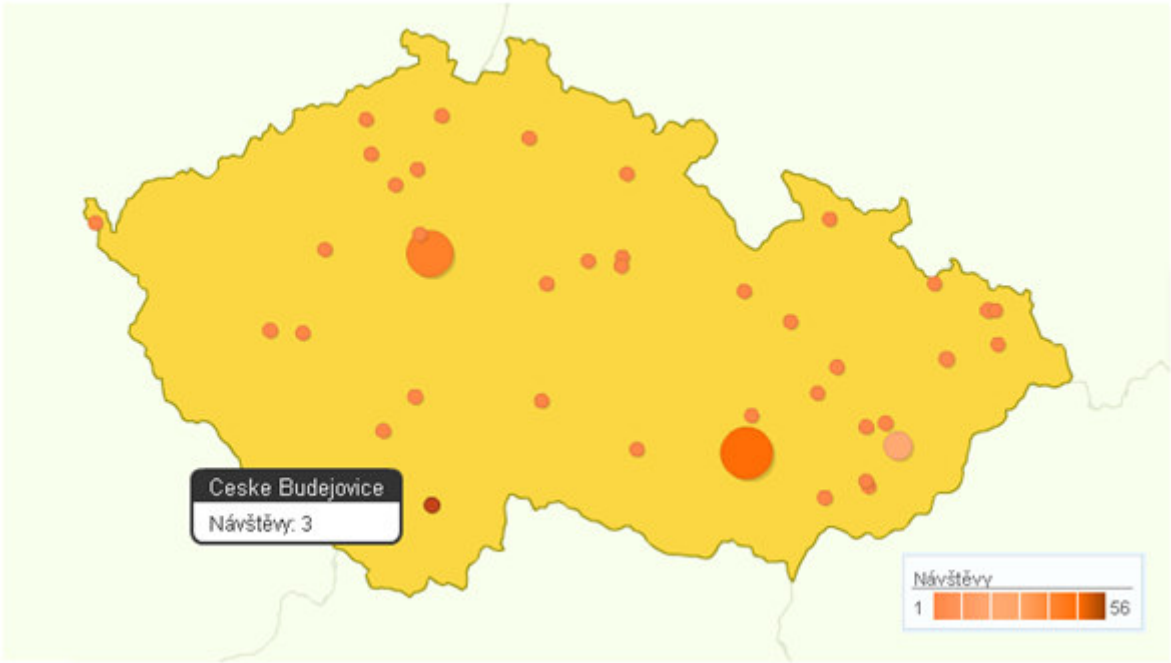
Jedná se o statistiku, která nás informuje o podílu návštěvníků, kteří internetovou prezentaci ukončí na úvodní stránce. Pokud je toto číslo vysoké, většinou to znamená, že úvodní stránka neobsahuje relevantní informace, které návštěvník požaduje. Doporučuje se tedy změnit strukturu s důrazem na kategorizaci nabízených dat a jejich přehlednost.

Návštěvníci - nový vs. staří:

Opět na základě jedinečné IP adresy lze určit, zdali se uživatel na stránky vrací nebo je zde poprvé.

Geografická segmentace klientů:

Google analytics je schopna na základě analýzy IP adresy zjistit polohu uživatele. Poté, pomocí grafického interaktivního výstupu lze tato data analyzovat. Výsledkem této analýzy může být například zvýšený podíl reklamy v určitých regionech, otevření pobočky firmy apod.



Obrázek 17 Geografická segmentace návštěvníků stránek

Jazykové nastavení:

Internetový prohlížeč uživatelů je nakonfigurován pro určitou jazykovou sadu znaků. Říká se tomu jazykové kódování. Google Analytics tyto informace sbírá a třídí. Geografická segmentace nám sice může napovědět, že naše stránky pravidelně sleduje skupina cizinců. To ovšem nemusí být důvod k tomu, abychom pro ně vytvářeli cizojazyčnou mutaci

původních stránek, pokud například používají domácí kódování.

Aktuálnost:

Úroveň zájmu o společnosti, výrobku, nebo prezentované značce je měřitelná též pomocí frekvence opakovaných návštěv.

Prohlížeče:

Důležitý údaj zejména pro administrátory internetové prezentace. Vzhledem k rozmanitosti internetových prohlížečů nelze spoléhat na to, že všichni využívají nejčastěji používaný Internet Explorer, popřípadě jeho mutace založené na jeho jádru. Stránky je proto nutné optimalizovat, aby byly čitelné pro všechny.

Operační systém:

Další segmentací návštěvníků je možná podle užívání operačního systému. V České republice jsou jako nejčastěji užívanými operačními systémy systémy společnosti Microsoft. Menší, nikoliv však zanedbávající, podíl zaujímá Linux a Unix.

Název hostitele:

Internetový hostitel je zařízení, které umožňuje odesílat a přijímat pakety. Názvy internetových hostitelů mohou poskytovat informaci o názvu organizace, ze které návštěvník přichází.

[16]

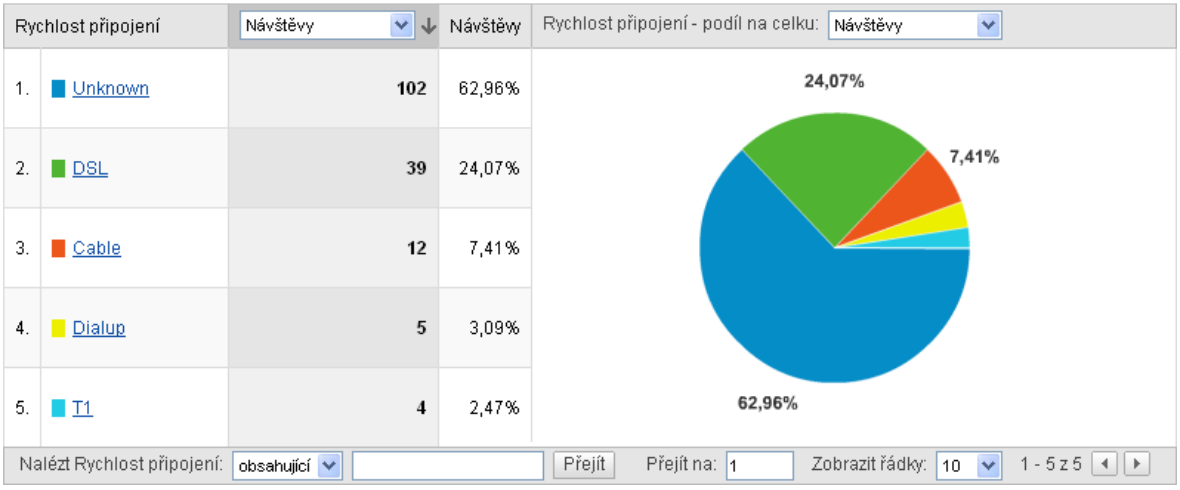
Umístění sítě:

Přináší informaci o poskytovatelích internetového připojení klientovi. Je tedy možné sledovat domény poskytovatelů internetového připojení a také IP adresu.

Další informace:

Z nastavení uživatelova počítače lze vyčíst mnohé. Například množství barev nebo rozlišení obrazovky, podporu programu Flash nebo Java. Způsob a rychlost připojení a další in-

formace, které návštěvník sám poskytne (například ve formulářích). Tyto informace opět slouží spíše k technické optimalizaci stránek.



Obrázek 18 Nejčastější způsob připojení návštěvníků internetových stránek

8 DOPORUČENÉ ZPŮSOBY VYHODNOCOVÁNÍ

Pro další vyhodnocování návštěvnosti lze využít data z následujících zdrojů:

1. Google analytics
2. Externí datový soubor získaný z PHP čítače

Výhody zdrojů Google Analytics:

Pomocí přehledného grafického prostředí lze jednoduchým způsobem zvolit, která data chceme uložit pro vlastní potřebu. GA nabízí širokou paletu datových souborů, do kterých mohou být získaná data uložena, export je tedy snadný. Nevýhodou je ovšem skutečnost, že pro operaci s daty je nutné přihlášení do aplikace. Není tedy možné poskytnout aktuální informace neautorizovaným osobám. Nelze totiž zpřístupnit pouze část aplikace.

Výhody zdroje PHP:

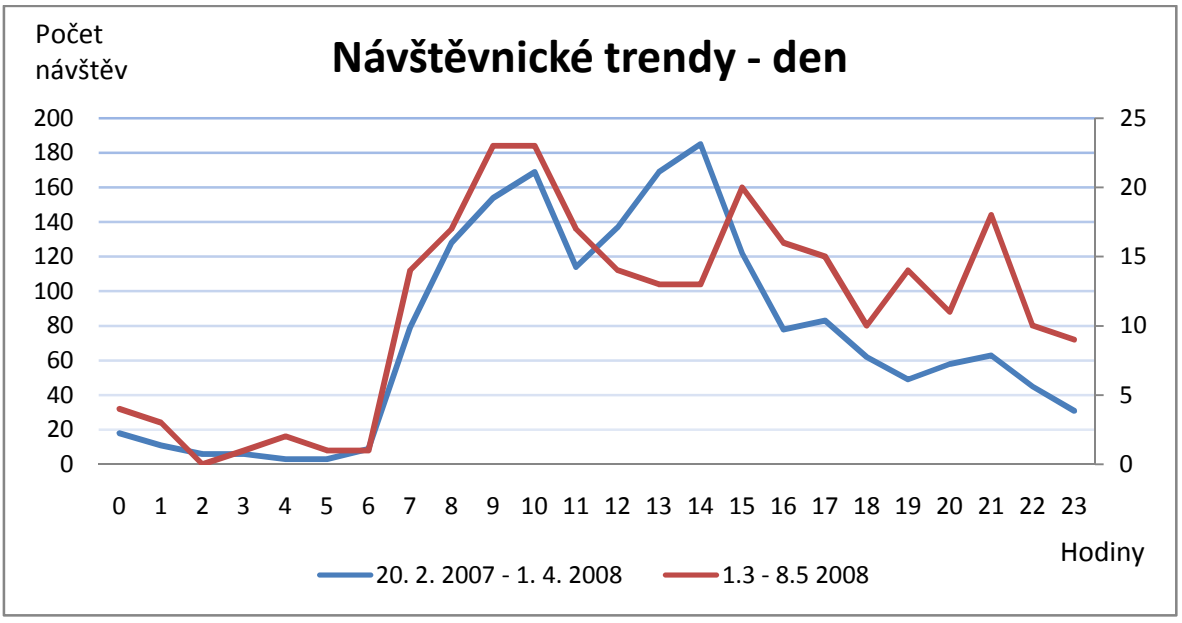
Pomocí krátkého PHP souboru, uloženého na serveru, kde je umístěna webová prezentace, se automaticky ukládají informace, definované ve zdrojovém kódu. Omezujícím faktorem pro uživatele je tedy znalost programovacího jazyka. Výhodou pro koncového uživatele je okamžitý přístup přes internetový prohlížeč. Data mají pevnou strukturu a mohou být exportována například do textového editoru. Ve své práci tento PHP čítač využiji k vytvoření interaktivní webové stránky pomocí Rpadu.

8.1 Analýza návštěvnosti – nejčastější návštěvní hodiny

Vzhledem k předcházení přetížení serveru je důležité identifikovat časový interval, kdy nejčastěji dochází k návštěvám internetové prezentace. Z nejčastější doby lze též usuzovat o složení návštěvníků. Například, zdali navštěvují stránky především v pracovní době, nebo soukromě.

Analyzuje data, která nasbíral GA v časovém období 20. února 2007 - 1. dubna 2008. Celkový počet návštěv za časové období je 1782. Dále se graficky snažíme zjistit, zdali se za časové období 1.3 2008 do 8. 5. 2008 změnili preference doby návštěvy návštěvníků.

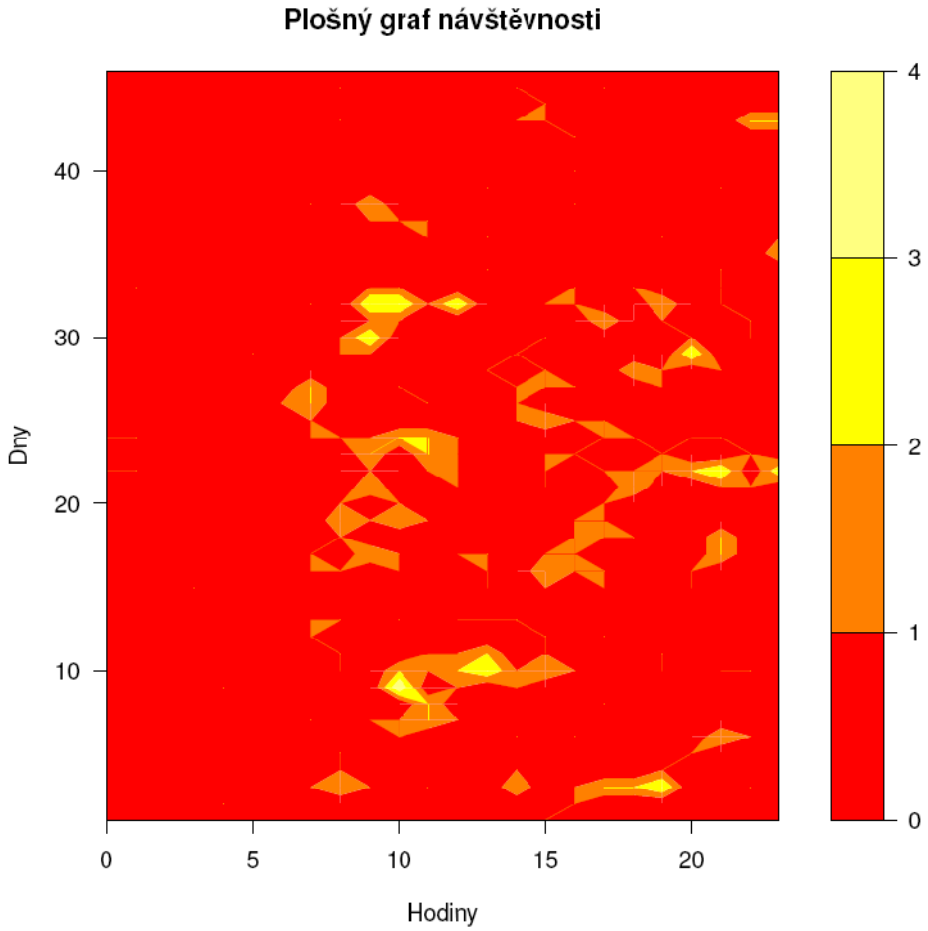
Základním grafem, který tuto situaci vystihuje, je graf časové řady jednoho dne. Vzhledem k tomu, že máme k dispozici údaje za delší časové období a hodláme srovnat návštěvnické trendy a ne absolutní hodnoty, vytvoříme jeden graf s hlavní osou (delší časové období) a vedlejší osou (kratší časové období).



Obrázek 19 Celková návštěvnost (MS Excel)

Viditelný posun preferencí oproti delšímu období zaznamenáme zejména ve zvýšené oblíbenosti návštěv v odpoledních a večerních hodinách.

Pomocí plošného grafu zaznamenáme vývoj návštěvnosti po jednotlivých dnech. Stupnice počtu návštěv je vytvořena podle odstínů barev.



Obrázek 20 Plošný graf návštěvnosti (R)

8.2 Analýza doby strávené na internetové prezentaci

Samotný údaj o počtu návštěvníků může být zavádějící. Důležité je sledování vzájemné souvztažnosti počtu zobrazených stránek a doby strávené na stránkách.

Pokud bychom provedli výpočet síly korelace se všemi hodnotami, tedy i s odlehlými, došli bychom podle (1) k závěru, že r = 0,77.

Z grafu je ovšem čitelné, že tento výpočet je nespolehlivý. Vyznačené Adjescent values (odlehlé hodnoty), se určí jako 1,5 násobek velikosti kvartilového rozpětí od střední hodnoty oběma směry.

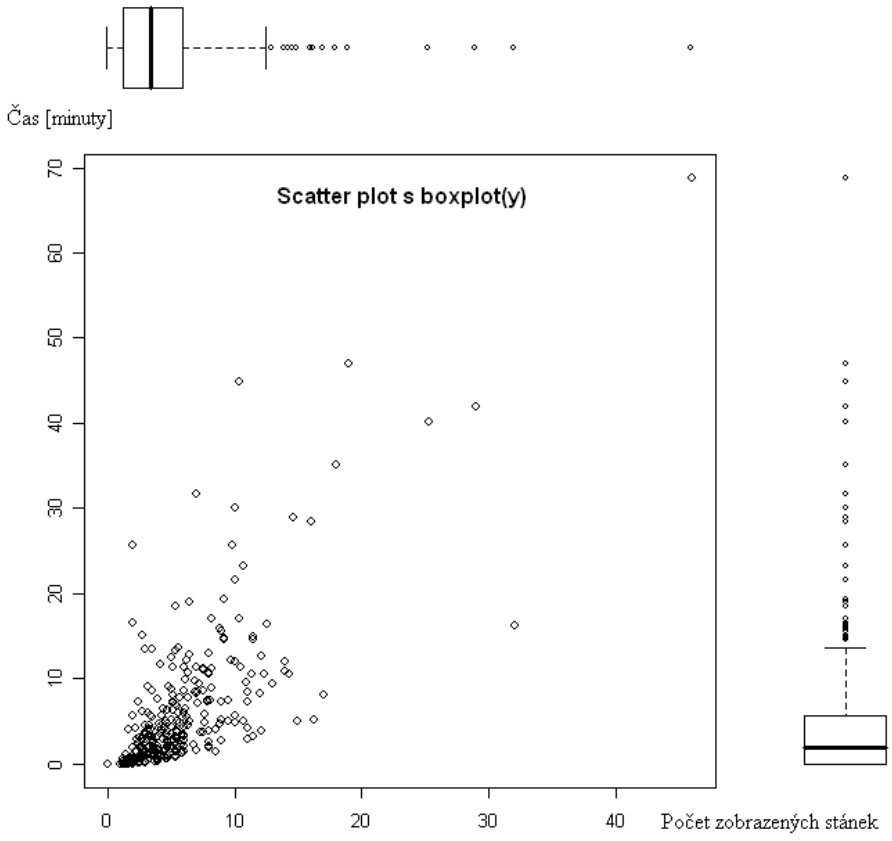
K výpočtu užijeme program R:

	Min	1st Qu	Median	Mean	3rd Qu	Max
Počet zobrazení	0	1,365	3,47	4,462	6	46
Čas [minuty]	0	0,041	1,867	4,526	5,625	69,733

Po drobných výpočtech zjistíme korelaci, opět pomocí (1), bez vlivů odlehlých hodnot.

r = 0,72.

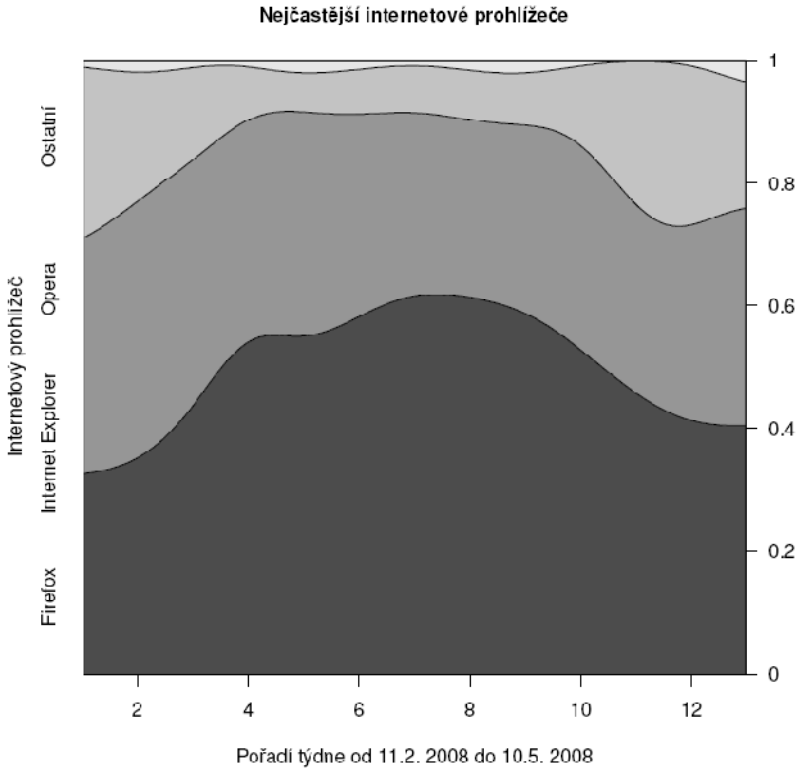
Na následujícím grafu je tato souvztažnost znázorněna.



Obrázek 21 Souvztažnost počtu zobrazených stránek a času stráveného na stránkách (R)

8.3 Analýza přístupů podle internetového prohlížeče

Důležitost tohoto ukazatele spočívá zejména v optimalizaci nastavení stránek. Některé prohlížeče mohou mít problémy s interpretací znaků, tabulek a podobně.



Obrázek 22 CD plot (R)

Na obrázku 23 je znázorněn vývoj podílu jednotlivých prohlížečů uživatelů stránek. V současné době je nejčastějším prohlížečem Firefox, jehož podíl se pohybuje nad úrovní 40%.

8.4 Regresní analýza

Pro odhad vývoje návštěvnosti jsem se rozhodl provést vícenásobnou regresní analýzu. Vstupní data byla sbírána v období od 1. 2. 2007 do 1.4 2008.

Obrázek 23 Vývoj počtu návštěv

Data jsem roztřídil podle dnů v týdnu. Následující výpočty byly provedeny v programu R.

pcd= pořadové číslo dne

Tabulka 1 Celková regresní analýza

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.685309	0.600182	-1.142	0.254
po	4.227641	0.708296	5.969	5.28e-09
ut	3.086066	0.705291	4.376	1.55e-05
st	4.003229	0.708324	5.652	3.03e-08
ct	3.420526	0.708311	4.829	1.96e-06
pa	1.682649	0.708302	2.376	0.018
so	-0.606951	0.708296	-0.857	0.392
ne	NA	NA	NA	NA
pcd	0.013738	0.001609	8.537	2.93e-16

Tabulka 2 Poměr determinace a F test neupravené regresní analýzy

Multiple R-Squared: 0.2893	
F-statistic: 23.2 on 7 and 399 DF	p-value: < 2.2e-16

Z tabulky 1 je zřejmé, že vliv Intercept (hranice) a návštěvnosti ve dnech sobota a neděle je statisticky nevýznamná ve vztahu na celkovou návštěvnost. Proto je z modelu vyřadíme.

Navíc tento model popisuje pouze 28,9% celkového rozptylu.

Tabulka 3 Regresní analýza bez statisticky nevýznamných faktorů

	Estimate	Std. Error	t value	Pr(>|t|)
po	4.004760	0.558768	7.167	3.71e-12
ut	2.885787	0.553630	5.162	3.86e-07
st	3.769151	0.556212	6.776	4.41e-11
ct	3.188687	0.556719	5.728	2.00e-08
pa	1.453050	0.557228	2.608	0.00946
pcd	0.011499	0.001182	9.730	< 2e-16

Tabulka 4 Poměr determinace a F test upravené regresní analýzy

Multiple R-Squared: 0.632	
F-statistic: 114.8 on 6 and 401 DF	p-value: < 2.2e-16

Nezkreslený odhad determinace je Multiple R-Squared: 0.632, to znamená, že daným regresním modelem lze vystihnout 63% celkového rozptylu.

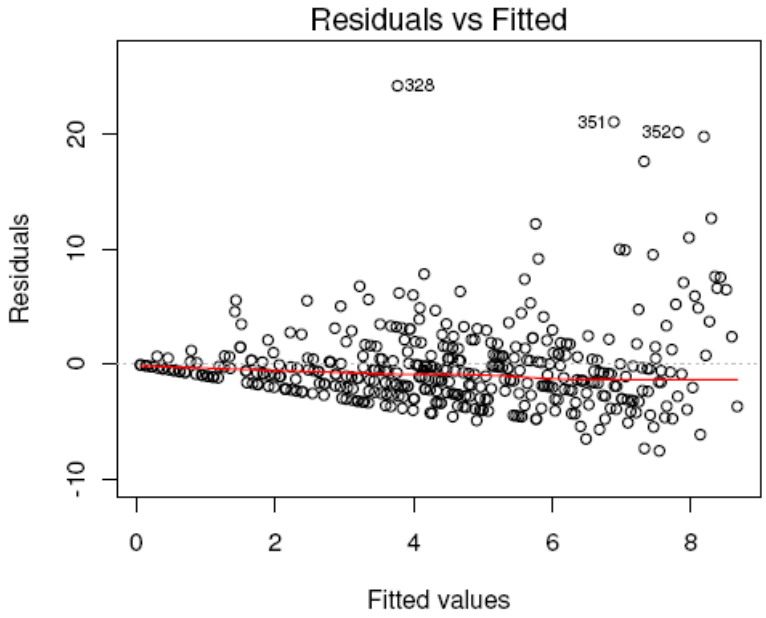
O vhodnosti zvoleného modelu lze usuzovat z celkového F testu. Testová hodnota je dostatečně vysoká. Kritický obor je v našem případě vymezen intervalem F>2,12.

[17]

Vizuálně lze o vhodnosti zvoleného grafu rozhodnout z následujících grafů. Nicméně základem pro rozhodování je statistická analýza - testy důležitosti regresorů a celkový F test.

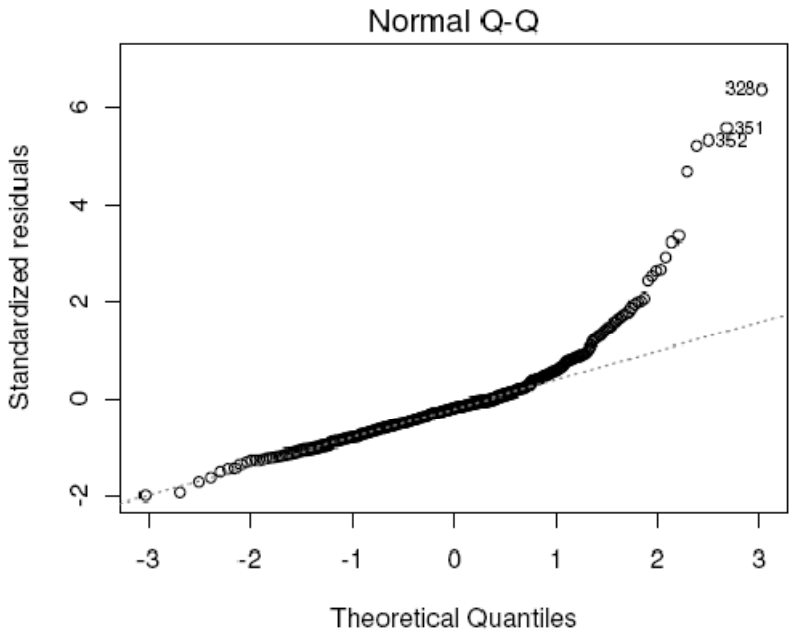
Na následujících grafech jsou analyzována rezidua. Reziduum představuje určitou ztrátu informace vlivem aproximace původního souboru dat. Velká rezidua ukazují, že vytvořený model nedostatečně popisuje data.

[11]



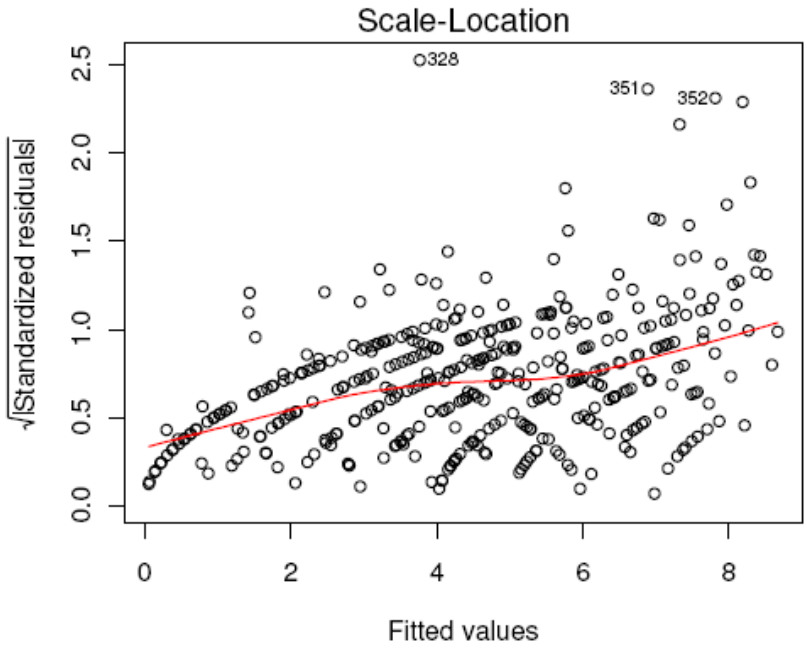
Obrázek 24 Analýza rozptylu vzniklého modelem časové řady (R)

Na Obrázku 24 je zachycena výše reziduí. Naší snahou bylo minimalizovat reziduální složku, neboli dosáhnout co největší hustoty okolo červené přímky. Skutečná hodnota v této rovině je stejná jako hodnota vypočtená z modelu, a proto je reziduum nulové.



Obrázek 25 Kvantilový - reziduální graf (R)

Z kvantilového grafu je zřejmé, že rozdělení rezidua je velmi blízký normálnímu rozdělení.



Obrázek 26Směrodatné odchylky bodů rezidua (R)

Tento graf zobrazuje směrodatnou odchylku hodnot reziduí.

8.5 Korelační analýza

Úkolem korelační analýzy je odhalení intenzity vzájemných vztahů. Nejčastěji se jedná o lineární závislosti.

Numericky lze vyjádřit sílu vztahu pomocí Pearsonova koeficientu (1). Protože srovnáváme 4 veličiny:

- Počet připojení s určitým prohlížečem
- počet návštěvníků, jejichž IP adresa již byla právě jednou zaznamenána = jednou
- počet návštěvníků, jejichž IP adresa již byla zaznamenána 15krát až 25krát =stálí
- průměrný čas na stránkách = minut_prum

výsledné hodnoty uvádím v korelační matici:

Tabulka 5 Korelační matice – Firefox (R)

	Firefox	jednou	stali	minut_prum
Firefox	1,00	0,72	0,47	0,29
jednou	0,72	1,00	0,07	0,23
stali	0,47	0,07	1,00	-0,23
minut_prum	0,29	0,23	-0,23	1,00

Tabulka 6 Korelační matice – Internet Explorer (R)

	Internet.Explorer	Jednou	stali	minut_prum
Internet,Explorer	1,00	0,53	0,57	0,06
jednou	0,53	1,00	0,07	0,23
stali	0,57	0,07	1,00	-0,23
minut_prum	0,06	0,23	-0,23	1,00

Tabulka 7 Korelační matice – Opera (R)

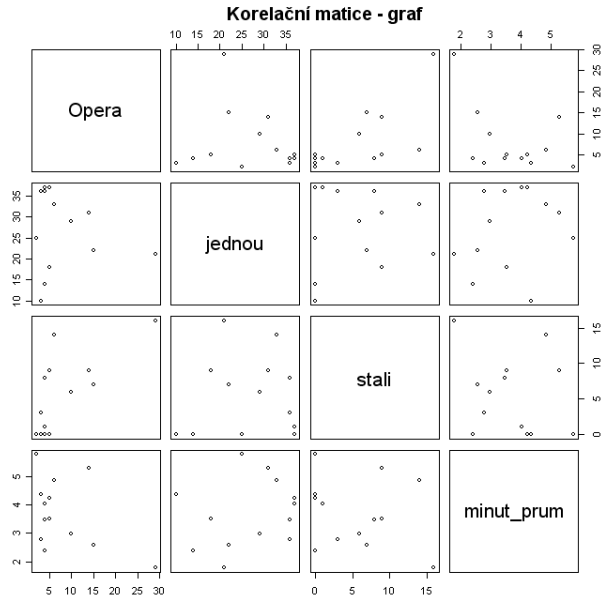
	Opera	jednou	stali	minut_prum
Opera	1,00	-0,14	0,69	-0,46
jednou	-0,14	1,00	0,07	0,23
stali	0,69	0,07	1,00	-0,23
minut_prum	-0,46	0,23	-0,23	1,00

Výsledky korelační analýzy jsou zajímavé. Zatímco zvyšující se počet návštěv uživatelů užívajících Firefox má kladný vliv na velikost průměrné doby strávené na stránkách, u uživatelů Opery zaznamenáváme středně silnou negativní korelaci. U uživatelů Internet Explorer je tento vliv zanedbatelný.

Na straně druhé, s růstem počtu uživatelů Opery nejvíce roste počet stálých návštěvníků.

Zajímavé je též zjištění, že růst počtu stálých návštěvníků snižuje průměrnou dobu prohlížení stránek.

[18]



Obrázek 27 Korelační graf – Pairs (R)

8.6 Analýza stránek pomocí programu Rpad

Pro potřeby aktuálního vývoje a pro poskytování informací ostatním osobám (ti, co nemají přístupová práva na server nebo ke službě Google Analytics) jsem vytvořil stránku, pracující na principu popsaném v kapitole věnované Rpadu. Vzhledem ke skutečnosti, že nakladatelství Pavel Stříž nevlastní administrativní práva k instalování na doméně striz.cz, analýzu stránek je nutné provést pomocí virtuálního serveru. Tento virtuální server se vytvoří automaticky při spuštění aplikace Rpad v programu R, podle zdrojového kódu uvedeného v příloze P I.

Přednastaveny jsou následující analýzy:

1. Základní popisná statistika datového souboru s krabicovými grafy
2. Regresní analýza popsaná blíže v kapitole 9.4
3. Korelační analýza 9.5

ZÁVĚR

Ve své práci jsem chtěl poukázat na možnosti, které poskytuje vizuální analýza dat. Ačkoli matematická analýza musí být základem, nelze význam statistických grafů opomíjet. V úvodní fázi analýzy dat, pomocí rychle vykreslených grafů, je možné přibližně určit, kterou metodu analýzy zvolit. Na závěr matematické analýzy je možné tyto grafy použít k interpretaci výsledků. Zaměřil jsem se na využití zejména nekomerčních programů.

Zjistil jsem, že pro kvalitní analýzu internetové návštěvnosti si běžný uživatel vystačí se službou Google Analytics, která poskytuje sběr dat a jejich následné zobrazení v podobě primitivních grafů nebo schémat.

Pro pokročilou analýzu jsem vytvořil internetové prostředí, které funguje na principu komunikace klient – server. Vytvořil jsem internetovou stránku, do které jsem vložil zdrojový kód programu Rpad. Tato stránka nyní na výzvu návštěvníka automaticky zpracuje datový soubor, který je uložen na serveru nakladatelství Stříž. Výstupem je základní popisná statistka datového souboru, regresní a korelační analýza, která je návštěvníkovi zobrazena přímo v internetovém prohlížeči.

SEZNAM POUŽITÉ LITERATURY

[1] *VisiCube : The Data Microscope*. 2004. 192 s. Dostupný z WWW: <http://www.datamology.com/doc/VisiCube%201.4%20Manual.pdf>.

[2] GENTLE, James. *Handbook of Computational Statistics : Concepts and methods*. Berlin : Springer 2004. 1070 s. ISBN 978-3-540-40464-4.

[3] KORPELLA, Jukka. *Tab Separated Values (TSV): a format for tabular data exchange* [online]. 2000-09-01 , 2005-02-12 [cit. 2008-03-19]. Dostupný z WWW: <http://www.cs.tut.fi/~jkorpela/TSV.html>.

[4] *FILExt - The File Extension Source* [online]. 2000 , 2007-01-20 [cit. 2008-05-19]. Dostupný z WWW: <http://filext.com/file-extension/TSV>.

[5] *SQL Server Developer Center* [online]. 2005 [cit. 2008-03-18]. Dostupný z WWW: <http://msdn2.microsoft.com/en-us/library/ms189887.aspx>.

[6] *PDF Reference : Adobe Portable Document Format*. 6th edition. 2006. 1310 s. Dostupný z WWW: <http://www.adobe.com/devnet/acrobat/pdfs/pdf_reference.pdf>.

[7] KOŽÍŠEK, Jan. *Ekonomická statistika a ekonometrie*. 2. přeprac. vyd. Praha : Vydavatelství ČVUT - výroba, 2005. 175 s. ISBN 80-01-03229-9.

[8] SAMUELS, Myra L., WITMER, Jeffrey A. *Statistics for the Life Sciences* . 3rd edition. Prentice : Prentice Hall, 2003. 680 s. ISBN 9780130413161.

[9] *Statsoft CR : Statistica 8* [online]. c2004- [cit. 2008-05-21]. Dostupný z WWW: http://www.statsoft.cz/page/index.php

[10] SOUKUP, Tom, DAVIDSON, Ian. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. Indianapolis : Wiley Computer, 2002. 416 s. ISBN 0471149993.

[11] MELOUN, Milan, MILITKÝ, Jiří, HILL, Martin. Počítačová analýza vícerozměrných dat v příkladech. Redaktor Aleš Baďura. 1. vyd. Praha : Academia, 2005. 449 s. , 1 CD-ROM. ISBN 80-200-1335-0.

[12] LEISCH, Friedrich. The R project for Statistical Computing [online]. 2003 [cit. 2007-12-09]. Dostupný z WWW: <http://www.r-project.org/>.

[13] *Rpad Documentation* [online]. c2005 [cit. 2008-05-19]. Dostupný z WWW: <http://www.rpad.org/Rpad/BasicDocumentation.html>.

[14] RAKOTOMALALA , Ricco . *A free data mining software for research and education* [online]. 2004 , April 23, 2008 [cit. 2008-05-19]. Dostupný z WWW: <http://eric.univ-lyon2.fr/~ricco/tanagra/index.html>.

[15] *Centrum nápovědy služby Google Analytics* [online]. c2008 [cit. 2008-05-19]. Dostupný z WWW: <http://www.google.com/support/googleanalytics/?hl=cs_CZ>.

[16] HABRMAN, Robert. *Server nejen o internetu, webu a ekomerci* [online]. 25.08.2007 [cit. 2008-05-13]. Dostupný z WWW: <http://www.owebu.cz/pc-site/vypis.php?clanek=1263>.

[17] HRONOVÁ, Stanislava, SEGER, Jan. *Statistika pro ekonomy*. 4. dopl. vyd. Praha : Professional Publishing, 2003. 415 s. ISBN 80-86419-52-5.

[18]KLÍMEK, Petr, RYTÍŘ, Vladimír. *Statistické metody pro ekonomy*. 1. vyd. Zlín : Univerzita Tomáše Bati ve Zlíně, 2001. 244 s. ISBN 80-7318-013-8.

[19] Posuzování bimodality na základě histogramu. DOŠLÁ, Šárka. *Informační Bulletin České statistické společnosti*. 2008. s. 24-33. Dostupný z WWW: <http://www.statspol.cz/bulletiny/ib-08-1.pdf>. ISSN 1210 – 8022.

Použité programy

[A] *Statsoft CR : Statistica 8* [online]. c2004- [cit. 2008-05-21]. Dostupný z WWW: <http://www.statsoft.cz/page/index.php>

[B] The datamology company : The Data Microscope [online]. 2002 [cit. 2007-12-09]. Dostupný z WWW: <http://www.datamology.com/>

[C] Statistical Analysis : Data Analysis and Statistics Software and Training [online]. 2007 [cit. 2007-12-09]. Dostupný z WWW: <http://www.minitab.com/>

[D] LEISCH, Friedrich. The R project for Statistical Computing [online]. 2003 [cit. 2007-12-09]. Dostupný z WWW: <http://www.r-project.org/>.

[E] *Microsoft Office Online* [online]. 2007 , 2008 [cit. 2008-05-19]. Dostupný z WWW: <http://office.microsoft.com/cs-cz/suites/FX101674091029.aspx>.

Manuály programů

[A] VisiCube : The data microscope version 1.4. Reedwood Valley, 2004. 176 s. Dostupný z WWW: <http://www.datamology.com/doc/VisiCube%201.4%20Manual.pdf>.

[B] Meet minitab 15 : for windows. USA, 2007. 140 s. Dostupný z WWW: <http://www.minitab.com/support/docs/rel15/MeetMinitab.pdf>. ISBN 978-0925636-51-5.

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM PŘÍLOH

PŘÍLOHA P I: ZDROJOVÉ KÓDY (R)

1. Zdrojový kód k obrázku číslo 4

```
a<-read.table("hist.txt",header=T)
hist(a$norm1,probab=T,breaks=50,main="Dvě střední hodnoty",xlab="Hodnoty")
curve(dnorm(x,mean=1.6,sd=1.8),add=T)
curve(dnorm(x,mean=4,sd=1.6),add=T)
```

2. Zdrojový kód k obrázku číslo 7

```
qqnorm(precip, pch=1)
qqline(precip)
points( qnorm(c(.25,.75)),
quantile(precip, c(.25, .75)) ,
 pch=16, col=2, cex=2)
```

3. Zdrojový kód k obrázku číslo 11

```
filled.contour(volcano, color = terrain.colors, asp = 1)# simple
x <- 10*1:nrow(volcano)
y <- 10*1:ncol(volcano)
filled.contour(x, y, volcano, color = terrain.colors,
plot.title = title(main = "The Topography of Maunga Whau",
xlab = "Meters North", ylab = "Meters West"),
plot.axes = { axis(1, seq(100, 800, by = 100)) axis(2, seq(100, 600, by = 100)) },key.title = title(main="Height\n(meters)"),key.axes = axis(4, seq(90, 190, by = 10)))# maybe also asp=1mtext(paste("filled.contour(.) from", R.version.string), side = 1, line = 4, adj = 1, cex = .66)
```

[3]

4. Zdrojový kód k obrázku číslo 13

```
staty<-data.frame(zeme=a[1],HDP=a[2],hustota=a[3],zam=a[4],dluh=a[5])
a<-read.table("dohromady.txt",header=T)
státy<-data.frame(zeme=a[1],HDP=a[2],hustota=a[3],zam=a[4],dluh=a[5])
staty2<-data.matrix(staty)
```

```
jmena2<-c("Německo","Česká republika","Polsko","Švédsko","Rumunsko")
faces2(staty2,labels=jmena2)
```

5. Zdrojový kód k obrázku číslo 21

```
a<-read.csv("matrix4.csv",sep = ";",header=F)
a<-as.matrix(a)
a1<-matrix(a,nrow=24,ncol=46)
t<-0:23
r<-1:46
 filled.contour(t,r,a1,nlevels=5,plot.title = title(main = "Plošný graf návštěvnosti", xlab =
"Hodiny", ylab ="Dny"),color=heat.colors)
```

6. Zdrojový kód k obrázku číslo 22

```
a<-read.csv("sc_box.csv", header=T,sep=";")
xbox <- boxplot(a$zobrazeno, plot=FALSE)
ybox <- boxplot(a$cas_minuty, plot=FALSE)
nf <- layout(matrix(c(2,0,1,3),2,2,byrow=TRUE), c(3,1), c(1,3), TRUE)
par(mar=c(3,3,1,1))
plot(a$zobrazeno,a$cas_minuty ,xlab="Počet zobrazených stánek", ylab="Čas [minuty]",
main="Scatter s boxploty" )
par(mar=c(0,3,1,1))
boxplot(a$zobrazeno, axes=FALSE, space=0,horizontal=T)
par(mar=c(3,0,1,1))
boxplot(a$cas_minuty, axes=FALSE, space=0, horiz=T)
```

7. Zdrojový kód k obrázku číslo 23

```
a<-read.csv("prohli2.csv",sep=";", header=T)
cd_plot(a$Prohlizec~a$Tyden,xlab="Pořadí týdne od 11.2. 2008 do 10.5.
2008",ylab="Internetový prohlížeč",main="Nejčastější internetové prohlížeče")
```

```
spineplot(a$Prohlizec~a$Tyden)
```

8. Zdrojový kód k obrázku číslo 24-26

Jedná se o sadu grafů, které se automatiky vytvoří po definování tvaru regresní funkce příkazem:

```
d<- lm(pocet_zobraz ~ doba_na_strankach + prumerne_zobrazeno + poradove_cislo, data = b)
plot(d)
```

9. Zdrojový kód k obrázku číslo 28

```
a<-read.csv("Korelace1b.csv",header=T,sep=";")
a2<-a[c(2,5,6,7)] #Mozilla
pairs(a3)
cor(a3)
```

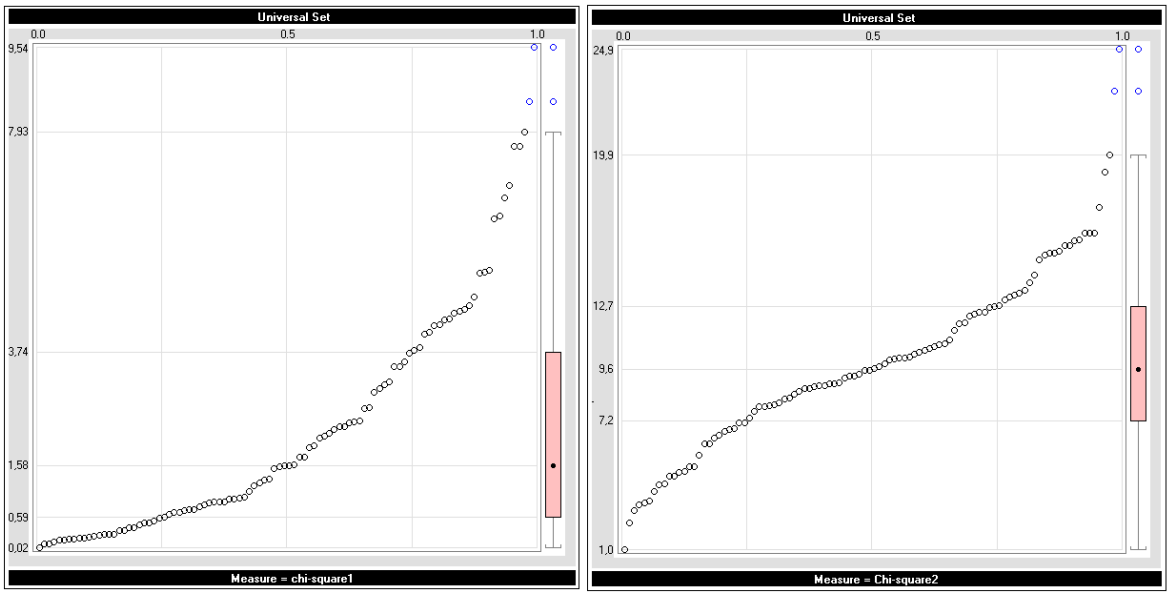
10. Zdrojový kód a návod k vytvoření virtuálního serveru pro Rpad

```
library(Rpad)
Rpad()
```

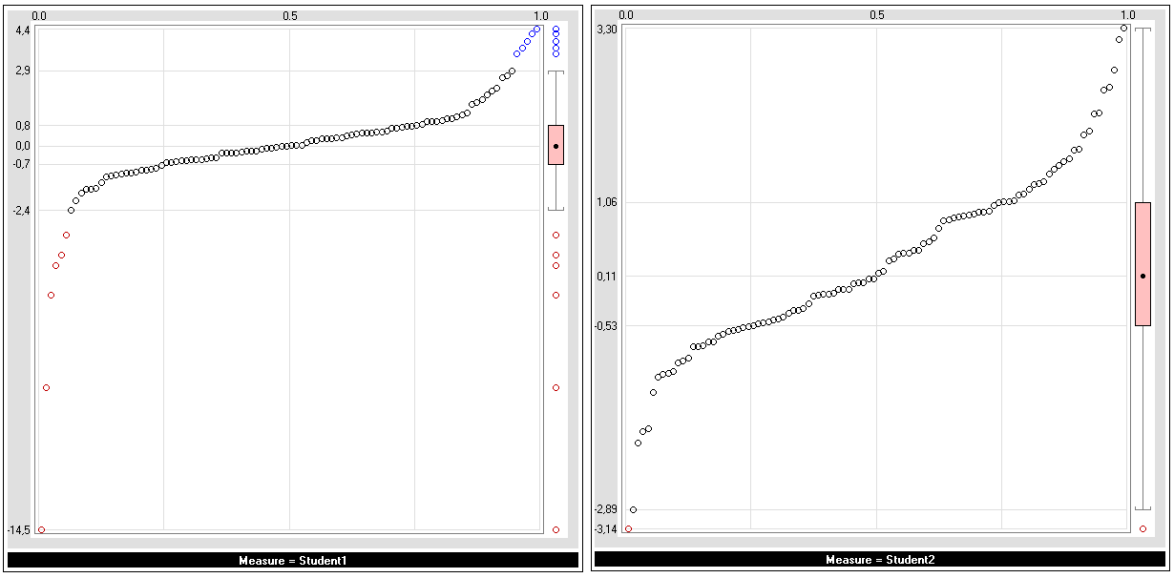
Automaticky se spustí internetový prohlížeč, ve kterém již existuje odkaz na stránku poskytující analýzu návštěvnosti.

PŘÍLOHA P II: OSTATNÍ GRAFY

1. Grafy vygenerovaných dat distribučních funkcí ze základních statistických rozdělení:



Obrázek 28 χ^2 Počet stupňů volnosti = 2 (vlevo) ; 10(vpravo) (VisiCube)



Obrázek 29 Studentovo rozdělení. Počet stupňů volnosti =2(vlevo); 10(vpravo). (Visicube)

PŘÍLOHA P III: STATISTICKÉ POZADÍ – HISTOGRAM A MULTIMODALITA

Pokud jsme nabyli podezření, že zkoumaný soubor dat nepochází pouze z jednoho rozdělení, ale z takzvané směsi dvou rozdělení s různými hustotami a různými středními hodnotami, provedeme Hartiganův test unimodality. Tento test jsme schopni provést pomocí programu R. V balíčku diptest, který není obsažen v základní instalaci, nalezneme funkci dip().

Postup analýzy:

Nulová hypotéza = Rozdělení je unimodální.

Dip = Vypočte testové kritérium.

data(qDiptab) = Tabulky, které značí hranici(kvantil) kritického oboru.

Dip>qDiptab = zamítáme unimodalitu

Musím ovšem upozornit, že tento test je nejvíce účinný při větším rozsahu výběru. V následující tabulce je uveden podíl odhalených bimodalit z bimodálních rozdělení. Pro μ =2 se jedná o unimodální data.

Tabulka 8 Úspěšnost odhalení bimodality

	Rozsah výběru n		
μ	100	1000	5000
2	0,00458	0,00061	0,00008
2,5	0,0209	0,04888	0,3021
2,8	0,05634	0,4279	0,00584
3	0,06856	0,82634	1
3,5	0,38187	0,999998	1

Z následující tabulky například vyplývá, že při velikosti výběru 1000 a rozdílu středních hodnot dvou statistických rozdělení = 0,8 při konstantním rozptylu =1, bylo odhaleno pouze 42,8% případů bimodality. Generování náhodných čísel s již zmíněnými parametry se provádělo 100 000.

[19]

PŘÍLOHA P IV: INTERNETOVÉ STRÁNKY NAKLADATELSTVÍ

Martin Stříž – Nakladatelství

MENU

DOMŮ
VYDALI JSME
UKÁZKY TVORBY
NAŠE SLUŽBY
O NÁS

Kontakt
Martin Stříž
U Škol 940
685 01 Bučovice

Tel.: +420 515 537 515
Mob.: +420 608 885 772

martin zavináč **striz.cz**

Domů

Autor, text → typografie, sazba, korektura → tisk a vazba → kniha

Chybí Vám monografie nebo jste napsali nějakou knihu a nikdo Vám ji nechce vydat? Velká vydavatelství nevydají vše, neboť se bojí, že se kniha nebude prodávat a zakázka bude ztrátová. My jsme však schopni vydat knihu i v jednotkách kusů. Neleží tak žádné knihy ve skladu, neváží peníze a není tak problém vydat knihu, která se prodává velmi pomalu.

Zabýváme se vydáváním knih v malém a středním nákladu, přidělením ISBN, tiskem, vazbou a vším ostatním, co s vydáváním knih souvisí.

Jaké knihy jsme již vydali, si můžete přečíst v sekci **Vydali jsme**. Pokud nemáte zájem přímo o knihu, ale o jiné tiskařské práce, podívejte se do sekce **Služby**, co vše vám můžeme nabídnout.

Chcete-li vidět, jak naše výtvory vypadají, podívejte se do **Ukázek**, kde najdete mnoho fotografií různých výrobků.

Pokud vás nabídka zaujala, můžete nás kontaktovat, viz stránka **s kontakty**.

Rád bych poděkoval Luboru Homolkovi za pomoc s webovou prezentací a statistickým vyhodnocováním její návštěvnosti.

Poslední aktualizace stránky: 4. 2. 2008

Aktuality:

DPH

Od 1. 4. 2008 jsme se stali plátci DPH! Uvedené ceny jsou bez DPH.

Petr Klímek

02/2008: Odborná kniha (monografie) s názvem *„Ekonomické aplikace statistiky a data miningu"*.
[+] **Více**

Laminace

2008: Rozšířili jsme **naše služby** o možnost laminace dokumentů.

Rozšíření webu

Přidali jsme ukázky **kalendářů** a odkazy na stažení **disertační práce**.

Matlab

02/2008: Vydána kniha zabývající se matematickým programem MATLAB.
[+] **Více**

Studie vlivu eura

01/2008: Vydána Studie zavedení eura na ekonomiku ČR.
[+] **Více**

PŘÍLOHA P V: RPAD STRÁNKY PRO ANALÝZU DAT

Podoba neaktivní stránky, která se aktivuje stiskem tlačítka Calculate [F9]

Calculate [F9]

Základní popisná statistika

```
a<-read.csv("1.csv",sep=";",header=T)
a1<-a[3:5]
summary(a1)
a2<-a[3]
b<-a[4]
b1<-b/60
boxplot(a2,horizontal=T,main="Pocet zobrazeni")
boxplot(b1,horizontal=T,main="Prumerna doba v minutach")


HTMLon()
showgraph()
```

Regresní analýza

```
pn<-a$pocet_zobraz
po<-a$pondeli
ut<-a$utery
st<-a$streda
ct<-a$ctvrtek
pa<-a$patek
so<-a$sobota
ne<-a$nedele
pcd<-a$poradove_cislo
d<-summary(lm(pn~po+ut+st+ct+pa+so+ne+pcd))
d
#par(mfrow=c(4,1))
d1<-(lm(pn~po+ut+st+ct+pa+so+ne+pcd))
plot(d1)

HTMLon()
showgraph()
```

Pro správné vyhodnocení regresní analýzy je nutné odstranit z modelu ty faktory, které jsou **statisticky nevýznamné**. Statistická významnost je značena následovně:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
přičemž nejvýznamnější mají ***. Vzhledem k tomu, že se jedná o automatizovaný proces, neodstranil jsem implicitně některé (nejčastěji nevýznamné so+ne) faktory. Tato činnost náleží spíše do oblasti analýzy dat než do klasického Data Miningu.

Korelační analýza
Vytvoření korelační matice a rozptylového grafu. Zvolil jsem prohlížeč Firefox, protože je nečastěji užívaným prohlížečem návštevníky stránek.

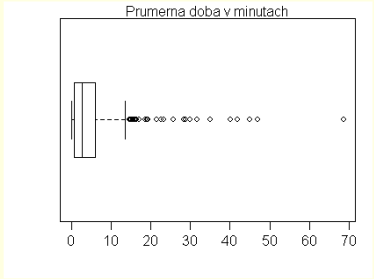
```
n<-read.csv("Kor.csv",header=T,sep=";")
nmo<-n[c(3,5,6,7)] #Firefox
cor(nmo)
pairs(nmo)


HTMLon()
showgraph()
```

Na následujícím obrázku je zachycen grafický výstup analýz definované na této stránce. Pro přehlednost uvádím výřezy. Neobsahuje tedy zdrojové kódy a popis analýz, které jsou ovšem ve skutečnosti viditelné.

Calculate (F9)

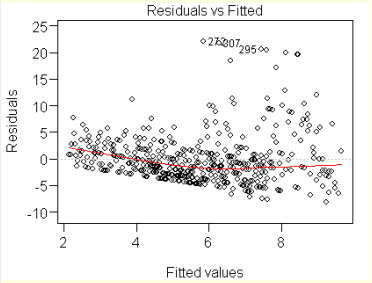
Základní popisná statistika

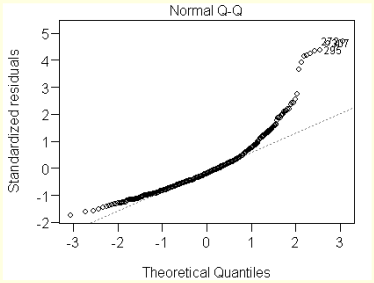
```
  pocet_zobraz     doba_na_strankach prumerne_zobrazeno
 Min.   : 0.000   Min.   :   0.0    1      : 38
 1st Qu.: 2.000   1st Qu.:  38.0    2      : 27
 Median : 5.000   Median : 165.0    4      : 17
 Mean   : 5.759   Mean   : 293.8    3      : 15
 3rd Qu.: 7.000   3rd Qu.: 362.8    6      : 12
 Max.   :28.000   Max.   :4124.0    2,5    :  9
                                    (Other):334
```

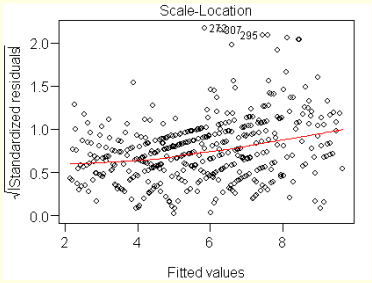


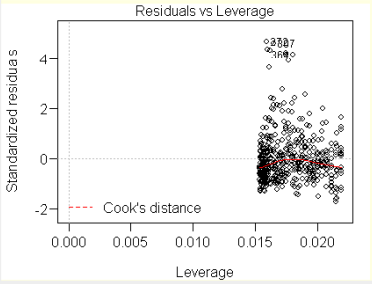
Regresní analýza

```
Call:
lm(formula = pn ~ po + ut + st + ct + pa + so + ne + pcd)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2169 -2.9784 -0.9203  1.6143 22.1512

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.54938    0.70815   3.600 0.000354 ***
po           1.47225    0.83845   1.756 0.079793 .
ut           1.30594    0.83522   1.564 0.118629
st           1.63227    0.83522   1.954 0.051293 .
ct          -0.31833    0.83522  -0.381 0.703291
pa          -0.45353    0.83522  -0.543 0.587397
so          -0.40975    0.83845  -0.489 0.625301
ne                NA         NA      NA       NA
pcd          0.01213    0.00171   7.094 5.16e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.743 on 444 degrees of freedom
Multiple R-Squared: 0.1295, Adjusted R-squared: 0.1158
F-statistic:  9.44 on 7 and 444 DF,  p-value: 6.168e-11
```









Korelační analýza

```
             Firefox     jednou       stali minut_prum
Firefox    1.0000000 0.71672875  0.46714734  0.2942261
jednou     0.7167287 1.00000000  0.06605397  0.2333490
stali      0.4671473 0.06605397  1.00000000 -0.2261669
minut_prum 0.2942261 0.23334905 -0.22616687  1.0000000
```