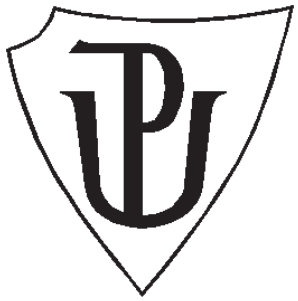
UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA GEOINFORMATIKY - DEPARTMENT OF GEOINFORMATICS

BAKALÁŘSKÁ PRÁCE

Využití Chí kvadrát testů na příkladech experimentálních dat s využitím Geostatistical Analyst v softwaru ArcMap



Vypracovala: **Markéta Papaková**

Vedoucí: **Mgr. Pavel Tuček, Ph.D.**

Olomouc 2010

Prohlášení

Prohlašuji, že na základě zadání bakalářské práce s cílem vypracovat pojednání o využití Chí kvadrát testů na příkladech experimentálních dat s využitím Geostatistical Analyst v softwaru ArcMap, jsem vytvořila tuto bakalářskou práci samostatně pod vedením Mgr. Pavla Tučka, Ph.D. Dále prohlašuji, že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování diplomové práce.

V Olomouci dne 25. května 2010

.....
Markéta Papaková

Poděkování

Děkuji svému vedoucímu bakalářské práce Mgr. Pavlu Tučkovi, Ph.D za veškerou pomoc a čas, který mi při zpracování této bakalářské práci poskytl. Dále bych ráda poděkovala svému konzultantovi Mgr. Vítu Pásztovi, za nespočet konzultací, rad a připomínek. Mé poděkování patří také Mgr. Michaele Tučkové, především za její pomoc při práci s typografickým systémem TeX.

Obsah

Úvod

Chí kvadrát testů se využívá napříč všemi obory a to nejen přírodovědnými, ale i socioekonomickými. Ve většině případů je výsledek testu zobrazen pouze ve formě tabulky. V lepším případě je pak tabulka doplněna slovním popisem. Přitom grafická vizualizace těchto testů by byla běžnému uživateli rozhodně bližší a měla by daleko větší vypovídající hodnotu.

Hlavní využití chí kvadrát testů spočívá v analýzách závislosti. Otázku, zda jsou jevy na sobě závislé, řeší v běžném životě každý v podstatě neustále. Například kdokoliv, kdo dojíždí do zaměstnání autobusem, si jistě všiml, že zaplněnost autobusu je závislá na denní době. Pokud jede do práce ráno, tlačí se v autobuse s ostatními, kteří vstávají na ranní směnu. Ale když jednou zaspí a jede o něco později, může se v autobuse pohodlně posadit. Tato práce je zaměřena na matematické prokázání právě této závislosti.

Práce není zaměřena pouze na teoretické vysvětlení tohoto testu, ale také na jeho praktické použití. Výsledkem pak nejsou jen tabulky či části programového kódu, pomocí něhož je možné závislost vypočítat, ale nedílnou součástí je také grafická vizualizace závislosti.

Data použitá v této práci byla získána dotazníkovým šetřením na území konkrétního města provedeném pod záštitou kraje. V tomto městě obsluhuje městskou hromadnou dopravu více autobusových dopravců. Vzhledem k ochraně soukromých údajů těchto dopravců nebylo možné v průběhu práce tato data blíže popsat, či ukázat. Veškeré mapové i jiné přílohy této práce proto není možné poskytnout třetí osobě k dalšímu zpracování a využití.

1 Cíle práce

Cílem bakalářské práce bylo sepsání rešerše na téma χ^2 testy se zaměřením na popis a aplikace této teorie (testy dobré shody). Dále byla práce zaměřena na sofistikovaný popis použité teorie a rozebrání dostupnosti popisovaných metod v běžně používaných softwarech. Všechny potřebné vizualizace byly provedeny v doporučeném softwaru firmy ESRI a byly vyhodnoceny možnosti takovýchto analýz v jiném běžně používaném softwaru na katedře geoinformatiky Univerzity Palackého v Olomouci. V závěru práce byl vyhodnocen datový soubor a doplněn o signifikantní vizualizace.

2 Metody a postup práce

Nejprve bylo potřeba prostudovat projekty, které se tématem zabývaly již v minulosti a také pročíst literaturu, která obsahuje použitou teorii. V rámci teoretické části jsou popsány nejen chí kvadrát testy, ale také testy a možnosti, které s tímto tématem úzce souvisí. Praktická část je zaměřena na uplatnění popsané teorie na konkrétním datovém souboru. Následně byly výpočty zvizualizovány pomocí tabulek a map, ze kterých jsou výsledky testů snadno čitelné.

Sesbíraná data v terénu byla nejprve převedena z analogové formy do softwaru Microsoft Exel, kde proběhla také jejich základní úprava a zpracování. Dále následoval export vybraných údajů do formátu CSV (Comma Separated Value). Veškeré výpočty byly provedeny v softwaru R-project ve verzi 3.10.1, který pracuje právě se zmíněným CSV formátem. Některé vizualizace byly provedeny exportem ze softwaru R-project, avšak většina grafických výstupů byla provedena v softwaru ArcMap ve verzi 9.3. Celá práce byla sepsána v typovém editoru TeXnic Center. Webové stránky byly vytvořeny pomocí editoru PS Pad a jejich grafika upravená pomocí softwaru TopStyleLite 3.0.

2.1 Současné řešení problematiky

Aplikace kontingenčních tabulek a χ^2 testů do GISů se využívá v mnoha výzkumech a pro různé obory. Snad nejvíce se χ^2 testů využívá v genetice, nikde však nebylo nalezeno další zpracování výsledků v GISech.

2.1.1 Rešerše literatury

Chí kvadrát (χ^2) testu dobré schody se věnují také Jaromír Antoch a Dana Vorlíčková v díle nazvaném *Vybrané metody statistické analýzy dat* [4]. Na začátku kapitoly upozorňují autoři, že data pro počítání testu musí být setříděná, ať už si třídy musíme vytvořit sami, či je jejich setřídění přirozené. V dalším odstavci se zabývají náhodnou veličinou při asymptotickém rozdělení $2k - 1$ při multinomickém rozdělení náhodného vektoru s parametry. Počet hypotetických četností v

této statistice musí být větší, roven pěti, aby bylo možné využít asymptotického rozdělení χ^2. Jestliže hodnota testu přesáhne jeho kvantil, zamítáme hypotézu, že pravděpodobnosti jednotlivých tříd jsou na hladině významnosti α. Autoři se také dále zabývají případem, kdy distribuční funkce F závisí na neznámých parametrech, kde došli k závěru, že jedinou změnou oproti předchozí statistice bude změna stupňů volnosti. V závěru kapitoly je upozorněno na nutnost dostatečného množství pozorování v rámci jednotlivých tříd a zároveň dostatečný počet tříd vzhledem k počtu neznámých parametrů.

Mezi nejvýznamnější zahraniční statistiky patří Rudolf Dutter, který se zabývá Chí kvadrát (χ^2) testem v díle *Statistik und Wahrscheinlich-keitsrechnung für InformatikerInnen* [7]. Zdůrazňuje zde, že je tento test nejrozšířenějším testem hypotézy o normálním rozdělení. Dutter zasazuje do setříděných intervalů histogram. Odchylky dat od histogramu jsou tak výrazné, že hypotézu o normálním rozdělení lze vždy přijmout nebo zamítnout. Dále Dutter uvádí vzorec pro výpočet četností záznamů a pravděpodobnost spádu hodnoty do dané třídy. Autor se pak věnuje samotnému testu, kdy při výrazných odchylkách hypotézu zamítá a definuje kritický obor na hladině významnosti α. Také Dutter upozorňuje, že test lze aplikovat pouze na rozsáhlé množství datových záznamů a že jedna třída musí obsahovat minimálně pět záznamů. Poslední odstavec je věnován možnosti, že rozdělení obsahuje neznámé parametry. Celý proces pak bude stejný, přičemž se změní pouze počet stupňů volnosti, který se sníží o odhadovaný parametr.

V díle *Statistika pro ekonomy* se kolektiv autorů zaměřil na χ^2 testy v kapitole věnované zpracování dat z výběrových šetření [3]. Zabývají se zde využitím χ^2 testu pro prokázání nezávislosti sledovaných znaků, či pro ověření předpokladu o určitém typu rozdělení. U χ^2 testu dobré shody rozlišují autoři úplně specifikovaný a neúplně specifikovaný model, které se liší pouze v tom, zda známe, či neznáme parametry rozdělení dat. Dále je uveden matematický vztah pro výpočet samotného χ^2 testu dobré shody, který je pak aplikován na několika příkladech. I zde autoři upozorňují, že musí být dostatečné obsazení dat ve všech skupinách a že pokud test provádíme při nedostatečném rozsahu výběru, mohou být závěry zpochybnitelné. V případě malého rozsahu výběru pak autoři doporučují použit

Kolmogorovův-Smirnovův test pro jeden výběr. Dále se autoři věnují χ^2 testu nezávislosti v kombinační (kontingenční) tabulce. Také vyvození tohoto vzorce ukazují na obecných i konkrétních případech kontingenčních tabulek.

Jiří Anděl patří mezi nejvýznamnějším české statistiky. Věnuje χ^2 testu dobré schody celou kapitolu ve svém díle nazvaném *Statistické metody* [1]. V první podkapitole se autor zabývá multinomickým rozdělením, jehož předpokladem je opakování pokusu několikrát nezávislé na sobě a jistota, že jeden jev musí vždy nastat. Zvláštním případem multinomického rozdělení je rozdělení binomické. V závěru této podkapitoly pak odvozuje autor matematický vztah pro výpočet samotného χ^2 testu. Ve druhé podkapitole se Anděl zaměřuje na χ^2 test při známých parametrech. Upozorňuje zde na to, že je tento test asymptotický a tedy použitelný pro dostatečné množství dat. Anděl se však přikloňuje k Yarnoldovu kritériu, které se vztahuje k počtu dat v rámci jedné třídy. V tomto tvrzení se liší od ostatních autorů, kteří používají tvrzení, že v rámci jedné třídy musí být minimálně pět pozorování. Při χ^2 testu při neznámých parametrech zavádí do vzorce i tento parametr. Dochází k soustavě rovnic, které dokazují, že čím je více pozorování (neboli větší množství dat), tím je vliv parametru menší. Při tomto testu již ale nelze použít Yornoldovo kritérium, jelikož se vztahuje pouze na testy o známých parametrech. Počet pozorování v rámci jedné třídy musí být tedy větší či roven pěti. V další kapitole se Anděl věnuje testu normality a testu Poissonova rozdělení. V testu normality jde o testování hypotézy, že mají data normální rozdělení $N(\mu, \sigma^2)$, kde jsou parametry neznámé. Autor nejprve rozdělí data minimálně do čtyř tříd. Potom odvodí vzorec pro výpočet parametru a nakonec provede samotný χ^2 test. Normalitu výběru lze také testovat pomocí šikmosti a špičatosti. Metodu χ^2 lze také použít na test Poissonova rozdělení, kde se testuje hypotéza, že jde o výběr dat právě tohoto rozdělení, přičemž parametr λ je neznámý. Na tuto kapitolu navazuje Anděl kapitolou věnovanou kontingenčním tabulkám, které s χ^2 úzce souvisí [10]. Autor se zde zaměřuje na vysvětlení kontingenčních tabulek obecně, větší část je však věnována jejich jednodušší variací, a to sice čtyřpolní tabulce. S pomocí této čtyřpolní tabulky pak uvádí vzorce pro výpočet χ^2 testu. V závěru kapitoly se pak autor věnuje ještě Fisherovu faktoriálovému testu a McNemanovu testu.

2.1.2 Rešerše provedených studií

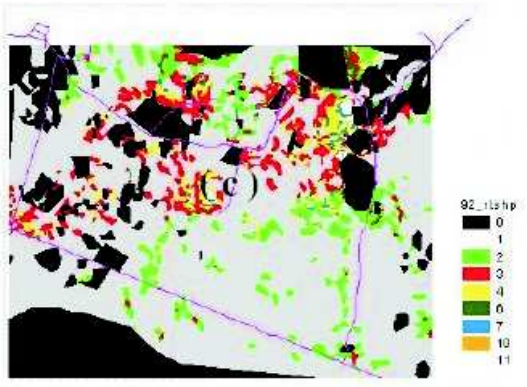
Chí kvadrát (χ^2) testů bylo využito například na projektu *Překryvné analýzy rastrových dat typu využití a pokryvu území* [10], kde byl použit pro hodnocení překryvu dvou vrstev rastrových dat. Pomocí kontingenčních tabulek byly zjištěny velikostí ploch pro kombinace jednotlivých tříd vzniklých při překryvu vrstev. Největší shoda byla zjištěna pro klasifikaci jehličnatého lesa. Celkový rozdíl mezi klasifikacemi je však 45%, což znamená, že leze shodu odhadnout na 55%. Po vytvoření kontingenční tabulky byl proveden χ^2 test, který byl použit pro výpočet míry asociace mezi těmito vazbami. Tento test byl proveden v programu SPSS, který však není vhodný pro rozsáhlejší datový soubor, proto do výpočtu nebyly vkládány dvojce odpovídající jednotlivým pixelům, ale odpovídající každé stovce pixelů. Chí kvadrát (χ^2) test potvrdil nenáhodnost těchto vazeb. Tato nenáhodnost byla způsobena vysokým počtem pozorování a tento výsledek byl tedy očekáván.

		CLC				Total
		jehličnatý	listnatý	ostatní	smíšený	
ETM	jehličnatý	4911	286	625	1220	7042
	listnatý	13	484	324	123	944
	ostatní	449	834	1100	414	2797
	smíšený	431	1057	527	1117	3132
Total		5804	2661	2576	2874	13915

Obrázek 1: Kontingenční tabulka pro výpočet χ^2 testu [10]

Dalším příkladem může být studie *GIS modelling of land degradation in Northern-Jordan using landsat imagery* [8] provedená v severním Jordánsku. Tuto studii provedla S. Essa z univerzity ve Spojených Arabských Emirátech. Šlo zde o vytvoření modelu v GIS, který by znázorňoval degradační dopady na změnu krajinného pokryvu v prostorových souvislostech. Tento model je založen na odhadech ročních ztrát vody. Chí kvadrát (χ^2) test byl použit na srovnání oblastí s rony a oblastí s průměrnou odhadovanou ztrátou půdy. Výsledek testu byl však

velmi nejednoznačný, a to sice, že ztráta půdy možná souvisí s výskytem ronů.



Obrázek 2: Předpokládané ztráty půdy pro rok 1992. Černou barvou jsou maskovány oblastí čedičů, zastavěné a obdělávané plochy. [8]

Využitím χ^2 testů v GIS se zabývá také projekt *Analysis of traffic flow in Varanasi city by using GIS* [11] proveden v městě Varanasi v Indii. Šlo zde o vytvoření GIS modelu pro analýzu míst ve městě, kde se nacházejí přetížené dopravní úseky s častými dopravními zácpami a navíc jsou charakteristické vysokou návštěvností turistů. V době největší dopravní špičky byla posbírána data, pomocí kterých byl vytvořen model v GIS. Tento model byl testován na dalším sběru dat, pomocí χ^2 testu, který předpověděl charakter dopravního proudu.

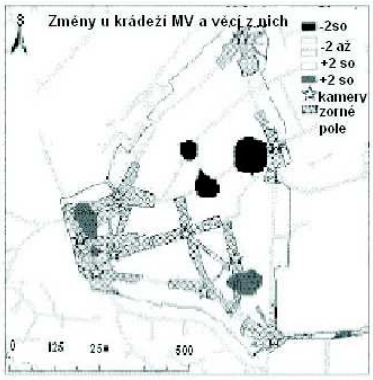
Nemělo by být zapomenuto ani na studii *GIS technologies for aquatic macrophyte studies: Modeling applications* [13] provedenou v Jižní Karolíně na jezeře Marion. Šlo o zjištění existence vztahu mezi vodními mykrofyty a parametry růstu rostlin. Pomocí χ^2 testů a následných překryvných analýz v GIS byl testován význam hloubky vody, sedimentů, dusíku, fosforu, rozpuštěného kyslíku a osvětlení při růstu mykrofyt. Výsledkem bylo potvrzení prostorové závislosti výskytu mykrofyt a kartografický model popisující stav optimálního růstu vodních mykrofyt. Tento výsledek pomohl k identifikaci oblastí jezer s nadměrnou náchylností k růstu mykrofyt, která tedy vyžadují další pozornost.

Parameter	N[1]	d.f.[2]	Calc.[3] X^2	Crit X^2	Pearson's[4] C	Cramer's[5] V
Bathymetry	2146	14	1363.4	36.1	0.62	0.56
Nitrogen	2188	30	1241.3	59.7	0.60	0.53
TDO	2383	23	619.9	49.7	0.45	0.36
Phosphorus	2541	17	251.7	40.8	0.30	0.22
Absolute Lt.	2570	11	233.4	31.3	0.29	0.21
Percent Lt.	2525	14	202.9	36.1	0.27	0.20
BDO	2314	30	194.9	59.7	0.28	0.21
Sediment	2591	11	170.8	31.3	0.25	0.18

[1] N = total number of observations
[2] degrees of freedom (d.f.) = (r * c) − 1
where r = number of rows
c = number of columns
[3] Chi Square $(X^2) = \Sigma i = \frac{r}{1} \Sigma j = \frac{c}{1} (f_{ij} - F_{ij})^2 / F_{ij}$
where f_{ij} = observed frequency in $cell_{ij}$
F_{ij} = expected frequency in $cell_{ij}$
[4] Pearson's $C = \sqrt{X^2 / (X^2 + N)}$
where N = total number of observations
[5] Cramer's $V = \sqrt{X^2 / (N * \min (r-1, c-1)}$
where N = total number of observations
r = number of rows
c = number of columns

Obrázek 3: Hodnoty χ^2 a nominální měření pro vodní vegetaci a environmentální parametry. [13]

Posledním příkladem použití χ^2 rozdělení v GIS je studie o *Vyhodnocení účinku kamerových systémů* [9] provedena v Anglii. Byl zjišťován účinek kamerových systému ve třinácti obvodech v různých podmínkách. Chí kvadrát (χ^2) test zde zjistil přímý vztah mezi viktimitou (náchylnost člověka k tomu, aby se stal snadnější obětí trestného činu) a obavami z možnosti stát se obětí trestného činu v daném místě. Následně byly v GIS zpracovány geografické trendy kriminality, což pomohlo při interpretaci míry ovlivnění kriminality kamerovými systémy.



Obrázek 4: Oblasti umístění kamer, jejich zorný úhel a oblasti se zvýšenou kriminalitou na dvojnásobek [9]

3 Testy normality

Celá řada statistických hypotéz je založena na skutečnosti, že je předem známe pravděpodobnostní rozložení dat. Testy normality sledují, zda má daný datový soubor rozdělení pravděpodobností v náhodném výběru roven normálnímu rozdělení $N(\mu, \sigma^2)$.

Testy normality můžou být rozděleny na testy o známých parametrech a testy o neznámých parametrech. Testy o známých parametrech počítají s parametry z původního datového souboru, kde se většinou předpokládá, že má normální rozdělení. Díky centrální limitní větě je možné je použít bez ohledu na to, jestli má základní soubor normální rozdělení, za předpokladu, že máme dostatečně rozsáhlý výběr ($n > 50$). Bohužel však výsledky testů můžou výrazně ovlivnit extrémní hodnoty v datech. Testy o neznámých parametrech se neřídí parametry z původního datového souboru. Někdy jsou též označovány jako testy s volným rozdělením. Tyto testy pak nepracují přímo s naměřenými daty, ale s jejich četnostmi či pořadovými čísly, které byly původním datům přiděleny. Při malém rozsahu výběru je však síla testu velmi malá. Když tedy výsledek testu zamítá hypotézu o normálním rozdělení, je jisté, že jsou data jiného rozdělení. Pokud však výsledek testu normalitu nezamítá, není závěr jasný. Výsledek může znamenat, že se data blíží normálnímu rozdělení, ale také, že nemáme dostatečné množství dat k prokázání normality [14].

Při testování, že náhodný výběr pochází z normálního rozdělení $N(\mu, \sigma^2)$ je nutné si nejprve vytvořit třídy. Je nutné dodržet podmínku, že počet tříd $k >= 4$. Dále musí být vypočteny pravděpodobnosti, že daná veličina padne právě do třídy Ji.

$$p_i = p_i(\mu, \sigma) = \int J_i f(x) dx$$

$$f(x) \frac{1}{\sqrt{2\pi\sigma}} exp - \left[\frac{(x-\mu)^2}{2\sigma^2} \right]$$

Z těchto matematických vztahů lze vypočítat neznámé parametry

$$\mu = \frac{1}{n}\sum_{i=1}^{k}\int J_i x f(x)dx$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}\frac{X_i}{p_i}\int J_i (x-\mu)^2 f(x)dx$$

Vzhledem k tomu, že vypočítané parametry závisejí jak na p_i, tak na $f(x)$, které jsou na pravé straně daných rovnic, musíme je označit jako odhady parametru. Dále lze pak pokračovat statistikou

$$\chi^2 = \sum_{i=1}^{k}\frac{[X_i - np_i(\hat{\mu},\hat{\sigma})^2]}{np_i(\hat{\mu},\hat{\sigma})}$$

Pokud hodnota testové statistiky překročí kritický obor o $k-3$ stupních volnosti, zamítáme hypotézu H_0, že data pocházejí z normálního rozdělení na hladině významnosti α [2].

Normalitu lze testovat také pomocí testů šikmosti a špičatosti [2]. Nejjednodušími a nejnázornějšími testy normality jsou grafické vizualizace. Nejznámějšími početními testy normality jsou pak χ^2 testy, kterým jsou věnovány ostatní kapitoly této práce.

3.1 Grafické vizualizace testů normality

Grafické ověřování normality je velmi jednoduché, a proto také velmi často používané. Nevrací jako výsledek žádná přesná čísla, ale z grafiky vyčte výsledek i člověk s menší odborností v dané problematice.

3.1.1 Histogram

Histogram je jednou z možností ověření normality pomocí grafu. Z histogramu lze také zjistit, kde jsou data umístěna, jak jsou rozložená, zda jsou symetrická či nikoliv a jaké mají odchylky. Je možné ho vytvořit vynesením intervalů na vodorovnou osu x a vynesením četností na svislou osu y. Pokud jsou data v normálním rozdělení, je histogram "zvonovitého tvaru".

Pro sestrojení histogramu je nutné najít maximum a minimum, k čemuž pomůže seřazení dat. Po nalezení těchto hodnot je možné určit variační rozpětí histogramu R, kde

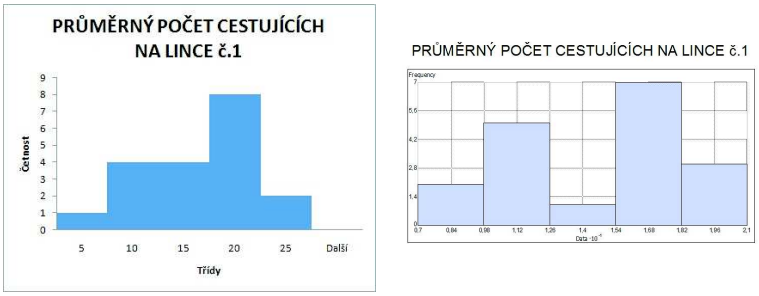
$$R = x_{max} - x_{min}$$

Pokud je známo toto rozpětí, stačí už jen vypočítat počet intervalů k a šířku intervalů h. Počet intervalů lze vypočítat různými způsob v závislosti na rozsahu souboru. Doporučený počet tříd je $7 - 20$. Pokud je rozsah souboru vyšší než 100, pak lze počet tříd definovat jako $k = 10 \log n$. Při rozsahu souboru $40 < n <= 100$, pak je počet intervalů roven $k = \sqrt{n}$. Poslední možností je, že je rozsah datového souboru $n < 40$. V tomto případě se vypočítá počet tříd pomocí matematického vztahu $k = 1 + 1,4426 \ln n$. Nyní už k sestrojení stačit jednoduchým matematickým vztahem $h = R/k$ určit šířku intervalů [15].

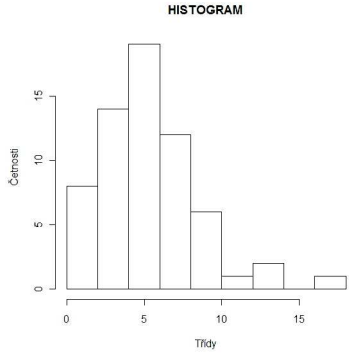
Před tvorbou samotného histogramu je vhodné si vytvořit tabulku, do které se zaznamenají intervaly, středy intervalů a četnosti výskytu v jednotlivých intervalech.

3.1.2 Box Plot (Krabicový diagram)

Box plot slouží k jednoduché vizualizaci datového souboru. Na první pohled vidíme odlehlé hodnoty (tzv. outliery), horní kvantil, dolní kvantil, medián, maximum a minimum.



Obrázek 5: Histogram vytvořený v softwaru MS Excel (vlevo) a v softwaru ArcMap (vpravo)



Obrázek 6: Histogram vytvořený v softwaru R-project

Uvnitř obdélníku se nachází 50 % všech dat. Dolní hranice obdélníku značí 25. percentil. To znamená, že 25 % všech dat je nižších než hodnota tohoto dolního kvartilu. Horní hrana obdélníku značí 75. percentil, který znamená, že 75 % všech hodnot je nižších než hodnota tohoto horního kvartilu. Čára uvnitř rámečku značí medián, tedy prostření hodnotu datového souboru. Dále box plot obsahuje tzv. vousy, na jejichž koncích se nachází minimum a maximum datového souboru. Pokud data obsahují extrémní hodnoty, jsou v box-plotu znázorněny pomocí hvězdiček či koleček zobrazených za vousy [14].

Pro sestrojení box-plotu je třeba nejprve vypočítat kvantily pomocí matematického vztahu

$$z_p = \frac{np}{100} + 0,5,$$

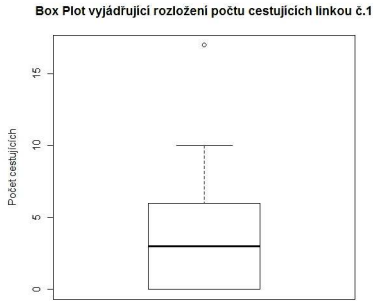
kde p znamená pořadí kvantilu xp a n je rozsah souboru. Pak už zbývá určit jen délku obdélníku R, kde

$$R = x_{75} - x_{25}$$

a určit konce vousů pomocí matematických vztahů [15]

$$A = x_{25} - 1,5R \qquad B = x_{75} - 1,5R$$

Data mají normální rozdělení, v případě že je box-plot symetrický.



Obrázek 7: Box plot vytvořený v softwaru R-project

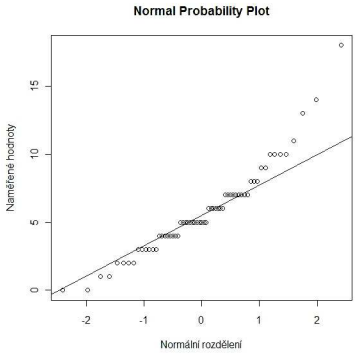
3.1.3 N-P Plot (Normal Probability Plot)

Normal probability plot je další z možností, jak graficky ověřit normalitu. Pokud mají data normální rozdělení, tvoří body zanesené do grafu přibližně přímku. Jakékoliv odchylky od této přímky znamenají odchylky od normálního rozdělení.

N-P plot lze vytvořit nanesením uspořádaných hodnot datového souboru na vodorovnou osu a na vodorovnou osu vynášíme kvantily odpovídající normálnímu rozložení dat [5].

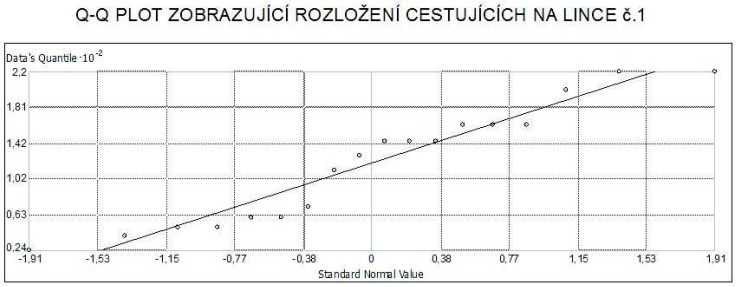
3.1.4 Q-Q Plot (Quantile-Quantile Plot)

Quantile - quantile plot umožňuje zjistit, zda mají data opravdu předem známé rozložení. Proto lze s jeho pomocí taktéž porovnávat normalitu dat. Tato grafická



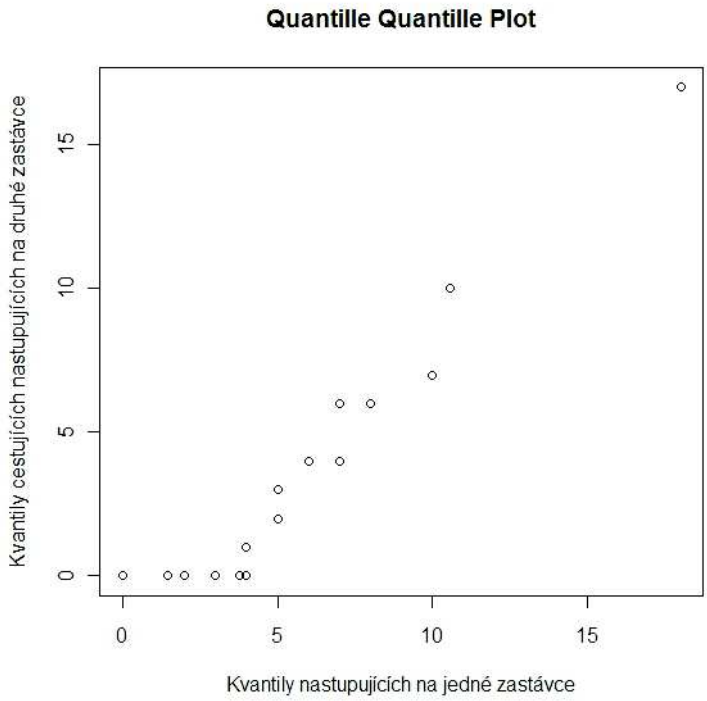
Obrázek 8: Normal Probability plot vytvořený v softwaru R-project

metoda spočívá v nanesení kvantilů dvou porovnávaných pravděpodobností rozložení dat proti sobě. Vynesenými body se pomocí metody nejmenších čtverců proloží přímka. Čím více leží body pávě na této přímce, tím více se rozložení dat blíží námi porovnávanému rozdělení [5].



Obrázek 9: Quantil Quantil plot vytvořený v softwaru ArcMap

Na obrázku 10 je možno sledovat oddalování a následně opět přibližování k přímce. To znamená, že data nemají normální rozložení, jehož kvantily jsou vyneseny na ose x.



Obrázek 10: Quantil Quantil plot vytvořený v softwaru R-project

3.1.5 P-P Plot (Probability-Probability Plot)

Poslední z uvedených možností grafického testování normality je P-P plot. Jeho konstrukce spočívá v nanesení teoretického rozložení distribuční kumulativní funkce na jednu osu a empirické kumulativní funkce na druhou osu [5]. Stejně jako u Q-Q plotu i v tomto případě proložíme vynesenými body pomocí metody nejmenších čtverců přímku. Pokud námi vynesené body leží právě na této přímce, pak jde o shodu teoretického a empirického rozložení. Odchýlení bodů od této přímky pak znamená, že rozložení dat není takové, jak se předpokládalo.

Všechny grafické metody testů normality zle vytvořit v softwaru R-project. Tyto grafické výstupy jsou statisticky správné a při vhodné úpravě je i grafická vizualizace přijatelná. Software ArcMap umí z výše uvedených vizualizací vytvořit histogram a Q-Q plot. U těchto výstupu se však bohužel nedá moc měnit nastavení ani popisky dat. Histogram umí vytvořit i software INDRISI ve verzi

Taiga a tabulkový software MS Excel.

3.2 Chí kvadrát (χ^2) test dobré shody

Chí kvadrát (χ^2) test dobré shody umožňuje zjistit, zda má sledovaná veličina opravdu předem známý typ rozdělení. Tato metoda vychází z hledání rozdělení, které by odpovídalo provedenému náhodnému výběru a sloužilo tak jako případný model. Jde tedy o porovnání skutečného rozdělení četností s námi teoreticky zvoleným rozdělení, které je určeno na základě úvah, předchozích zkušeností či pomocí grafické vizualizace rozdělení četností [3]. Shodu teoretického a skutečného modelu je nutno ověřit právě pomocí χ^2 testu dobré shody.

Jde o převedení jakéhokoliv multinomického rozdělení pravděpodobností sledované veličiny na veličinu s rozdělením pravděpodobností asymptoticky rovnu χ^2 rozdělení.

3.2.1 Multinomické rozdělení

Nechť $A_1, \ldots, A_k$ jsou disjunktní jevy (nemohou nastat dva nebo více současně) a zároveň právě jeden z nich v průběhu náhodného pokusu musí nastat. Jejich pravděpodobnosti označme $pi = P(Ai)$, $i = 1, \ldots, k$, kdy $p_1 + \ldots + p_k = 1$. Předpokládejme, že náhodný pokus opakujeme n-krát a počet výskytu jevu Ai označíme Xi. Pak pro nezáporná celočíselná $x_1, \ldots, X_k$, jejichž součet je roven n platí:

$$P(X_1 = x_1, \ldots, x_k = x_k) = \frac{n!}{x_{1!} \ldots x_{k!}} p_1^{x_1} \ldots p_k^{x_k}$$

Takto dané rozdělení pravděpodobností se nazývá multinomické rozdělení. Může nastat případ, kdy $k = 2$. Pak se toto rozdělení nazývá binomické [1].

3.2.2 Chí kvadrát (χ^2) test při známých parametrech

V případě, že nulová hypotéza H_0 udává nejen typ rozdělení, ale také jeho parametry, mluvíme o χ^2 testu při známých parametrech [1], nebo také o úplně

specifikovaném modelu [3].

Tento typ testu se používá pro náhodné veličiny nominální, ordinální či diskrétní. Zde je počet možných nabývaných hodnot nebo také kategorií označen k. V případě spojitých funkcí je setříděn obor hodnot do k vzájemně se nepřekrývajících intervalů.

Za testové kritérium se volí statistika,

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i},$$

kde X_i jsou empirické četnosti, nebo-li pozorované četnosti náhodné veličiny a np_i jsou teoretické nebo také očekávané četnosti v i-té skupině, $i = 1, 2, \ldots, k$. Pokud není třeba znát hodnotu sčítanců na pravé straně, ale stačí znát pouze výslednou hodnotu testového kritéria, je vhodnější upravit vzorec do podoby [1]:

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{k} \frac{X_i^2}{p_i} - n.$$

Hranice mezi důvěrou a nedůvěrou ve správnost hypotézy H_0 se nazývá kritický obor. Ten je omezen hodnotou $\chi^2_{1-\alpha}$ při $k-1$ stupních volnosti na hladině významnosti α . Jestliže je výsledná hodnota χ^2 vyšší než hodnota kvantilu kritického oboru, pak zamítáme nulovou hypotézu H_0, že pravděpodobnosti multinomického rozdělení jednotlivých tříd jsou právě rovny číslům $p_1, \ldots, p_k$ na hladině významnosti α . Pokud hodnota χ^2 testu spadá do kritického oboru, pak nulovou hypotézu H_0 nelze zamítnout na hladině významnosti α [3].

Jelikož je tento test asymptotický, je nutné mít dostatečný rozsah výběru n, jinak by nebylo přesné rozdělení testového kritéria při platnosti nulové hypotézy H_0 dobře aproximováno χ^2 rozdělením a výsledky testu pak mohou být zpochybnitelné. Ve všech třídách, do kterých je počáteční soubor pozorování roztřízen, je nutné mít dostatečné obsazení. Doporučuje se rozdělení [4]

$$np_i > 5 \text{ pro každé } \; i = 1, 2, ..., k.$$

Mnohem přesněji však toto obsazení definuje Yarnoldovo kritérium, které říká, že dobrá shoda s limitním rozdělením $2k - 1$ je zaručena, platí-li

$$np_i \geq 5q \text{ pro všechna } \ i = 1, 2, ..., k \text{ při } \ k \geq 3,$$

kde q je podíl tříd, pro než platí $np_i < 5$[1].

V případě, že tyto podmínky nemohou být splněny, je nutné sloučit třídy, které jsou si příbuzné nebo které jsou nedostatečně obsazené. Většinou jde o sloučení okrajových skupin. Pokud však k tomuto sloučení tříd dojde, je nutné změnit i počet stupňů volnosti, který se snižuje[3].

3.2.3 Chí kvadrát (χ^2) test při neznámých parametrech

Mnohem častěji se v praxi setkáváme s možností, že nulová hypotéza H_0 specifikuje jen typ pravděpodobnostního rozdělení, avšak hodnoty parametrů rozdělení známy nejsou a je třeba je předem z daného výběru odhadnout. V takovém případě mluvíme o χ^2 testu při neznámých parametrech [1], případně o neúplně specifikovaném modelu [3].

Výpočet celé statistiky je obdobný výpočtu χ^2 testu při známých parametrech. V tomto případě však pravděpodobnosti $p_1, \ldots, p_k$ závisí na určitém množství neznámých parametrů. Pravděpodobnosti jednotlivých intervalů, které nejsou známy, je možno určit jako rozdíl mezi hodnotami distribuční funkce, který odpovídá horní a dolní hranici intervalů. Odhad parametru μ může být nahrazen průměrem zjištěných hodnot $\tilde{x}$. Parametr σ^2 je možno nahradit jeho výběrovým protějškem s_x^2 [1]. Tímto jsou vypočteny odhady neznámých parametrů a může být tedy použito stejného testovacího kritéria, jako v případě χ^2 testu při známých parametrech.

Dalším možným způsobem jak odhadnout neznámé parametry z výběru je pomocí modifikované metody minimálního χ^2. Hodnota parametru a minimalizuje statistiku χ^2. Tuto hodnotu dostaneme pomocí řešení soustavy rovnic

$$\frac{\delta\chi^2(\mathbf{a})}{\delta a_j} = -\frac{1}{n}\sum_{i=1}^{k}\frac{X_i^2}{p_i^2(\mathbf{a})}\frac{\delta p_i(\mathbf{a})}{\delta a_j} = 0, \qquad j = 1, \ldots, m.$$

Pomocí různých derivací a jiných metod zjednodušení rovnic do formy vhodné pro další výpočty, dojdeme nakonec k soustavě

$$\sum_{i=1}^{k}\frac{X_i}{p_i(\mathbf{a})}\frac{\delta p_i(\mathbf{a})}{\delta a_j} = 0, \qquad j = 1, \ldots, m.$$

Z této soustavy již vypočítáme neznámé parametry a můžeme vypočítat hodnotu χ^2 testu [1].

Na každý odhad neznámého parametru se počet stupňů volnosti snižuje o 1. Pokud tedy odhadujeme dva parametry, je počet stupňů volnosti $df = k-1-2$ [4].

Příkladem neparametrického testu dobré shody může být například Shapirův-Wilkův test. Výsledná hodnota testu říká, na kolik data korelují s křivkou normálního rozdělení. Hodnota p-value udává chybu s jakou jsou data odchýleny od křivky normálního rozdělení. Dalším příkladem je Wilkoxonův párový test. V tomto testu se sledují rozdíly mezi pozorováními. Je třeba seřadit pozorování podle velikosti absolutních čísel od největšího po nejmenší. Poté se spočítá součet pořadí kladných a záporných rozdílů a porovná se s kritickou hodnotou.

4 Chí kvadrát (χ^2) test v kontingenčních tabulkách

4.1 Kontingenční tabulky

Kontingenční tabulky slouží k zjednodušení práce s obsáhlými datovými soubory, které zobrazují závislost dvou sledovaných znaků. Pomocí kontingenčních tabulek jsme schopni si data setřídit do srozumitelnější formy a snadněji z nich získat dalšími výpočty informace, které nejsou na první pohled patrné.

A/B	B_1	B_2	...	B_j	$\sum_j$
A_1	n_{11}	n_{12}	...	n_{1j}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	$n_{2.}$
...	...	...	...	...	...
A_i	n_{i1}	n_{i2}	...	n_{ij}	$n_{i.}$
$\sum_i$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	n

Tabulka 1: Obecné vzjádření kontingenční tabulky

Písmeny A a B značíme dva znaky, které sledujeme. V jednotlivých buňkách označených písmeny n pak zapisujeme počet (četnosti) případů, ve kterých jednotlivé možnosti znaků nastaly. Pod označením n_i jsou i-té varianty znaku A. V případě označení n_j jde pak o j-tou variantu znaku B. Součty hodnot ve sloupcích ($n_{.j}$) pak označují počet výskytu znaku B bez ohledu na znak A, naopak součty hodnot v řádcích ($n_{i.}$) označují počet výskytu znaku A bez ohledu na znak B [3].

Pro lepší názornost je možné si místo písmene A představit názvy měsíců v roce a místo písmene B názvy krajů České republiky. Pomocí kontingenční tabulky je sledována návštěvnost zahraničních turistů v jednotlivých krajích. V jednotlivých buňkách tabulky označených písmenem n s indexy je pak zaznamenán počet zahraničních turistů v konkrétním kraji za konkrétní měsíc. Pokud je v buňce tečka na prvním místě v indexování, jde o součet zahraničních turistů v jednotlivém kraji, bez ohledu na to, v kterém měsíci tento kraj navštívili. Po-

kud je však tečka v indexování na druhém místě, jde o součet všech zahraničních turistů v jednotlivých kalendářních měsících, bez ohledu na to, který kraj navštívili. Buňka označená písmenem n bez indexu pak obsahuje celkový součet všech zahraničních turistů v ČR.

Kontingenční tabulky se určují podle počtu řádku r a sloupců s na $r \times s$. Zvláštním případem je kontingenční tabulka typu 2×2. V tomto případě mluvíme o takzvané čtyřpolní tabulce. Tato varianta kontingenční tabulky nastane v případě, že porovnáváme dva dichotomické znaky. Dichotomický znak je každý takový znak, který může nabývat pouze dvou hodnot, které se vzájemně vylučují.

Využití kontingenčních tabulek je možné pouze v případě, že znaky které zkoumáme, nabývají pouze konečných hodnot, případně konečný počet kategorií. Pokud jsou data setříděna do kontingenčních tabulek, je jejich zkoumání a prokazování různých statistických hypotéz jednodušší. Například je možné zkoumat nezávislost znaků, homogenitu, či symetrii vztahu [16].

Kontingenční tabulky umí samozřejmě vytvořit statistický software R-project a tabulkový software MS Excel. Z GIS softwaru má tuto možnost software IDRISI ve verzi Taiga. Software ArcMap umí stejně jako IDRISI vypočítat relativní četnosti pravděpodobností a vytvořit z nich graf.

4.2 Chí kvadrát (χ^2) test nezávislosti

Chí kvadrát (χ^2) test nezávislosti je nejčastější a nejvíce používanou statistikou při rozboru kontingenčních tabulek. Je založen na chí kvadrát (χ^2) testu o neznámých parametrech a jeho úkolem je zjistit, zda jsou na sobě dva pozorované jevy závislé či nikoliv. V tomto testu jde o porovnání empirických neboli pozorovaných četností a četností očekávaných.

Aby bylo možné dál hovořit o tesu nezávislosti, je třeba si nejprve vysvětlit, co to vlastně nezávislost je. Nezávislost znamená, že ani jeden znak neovlivňuje,

jakých konkrétních hodnot nabývá znak druhý. Definice nezávislosti vyplývá z klasické teorie pravděpodobnosti. Pro zjednodušení je možné říct, že veličiny Y a Z jsou nezávislé tehdy a jen tehdy, platí-li $p_{ij} = p_{i.} \times p_{.j}$ pro všechny dvojice (i, j) [2].

Číslům $p_{i.}$ a $p_{.j}$ se říká marginální pravděpodobnosti. Jsou to pravděpodobnosti, s jakými nabývají buňky konkrétních hodnot v jednotlivých řádcích či sloupcích.

V rámci tohoto testu musí být počítáno s neznámými parametry. Do těch však už není možno počítat jako neznámé parametry $p_{r.}$ a $p_{.s}$, jelikož ty lze z ostatních marginálních pravděpodobností vypočítat. Počet neznámých parametrů je tudíž stanoven na

$$m = r - 1 + s - 1 = r + s - 2$$

Při výpočtu testové statistiky je třeba se omezit pouze na ty marginální pravděpodobnosti, které nabývají kladných hodnot. V opačném případě by se mohlo stát, že by se některé řádky či sloupce odečetly a výsledek testu by pak byl nepřesný.

Testové kritérium chí kvadrát (χ^2) testu nezávislosti je

$$\chi^2 = \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Matematický vytah $\frac{n_{i.}n_{.j}}{n}$ znázorňuje výpočet očekávaných četností. Označení n_{ij} pak znamená četnosti empirické, neboli pozorované.

Tuto testovou statistiku lze pro snadnější výpočty zjednodušit na tvar

$$\chi^2 = n \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{n_{ij}^2}{n_{i.}n_{.j}} - n$$

Tato testová statistika má přibližné rozdělení χ^2 o $(r-1)(s-1)$ stupních volnosti. Pokud je výsledek testové statistiky vyšší než testové kritérium, znamená

to, že je nulová hypotéza H_0 o nezávislosti zamítnuta na hladině významnosti α. Pokud testová statistika spadá do testového kritéria, pak nulovou hypotézu H_0 nezle zamítnout. Sledované jevy jsou tedy závislé.

Vzhledem k tomu, že má testová statistika χ^2 rozdělení, je nutné mít v rámci každé třídy co nejvíce pozorování. Čím je četnost jednotlivých jevů menší, tím je horší kvalita testu. Je doporučeno, aby minimálně 80 % všech četností bylo vyšších než pět a všechny musí být větší než 1. Pokud toto doporučení není splněno, je nutné některé třídy (řádky či sloupce) sloučit.

Chí kvadrát (χ^2) test nezávislosti lze uplatnit i ve čtyřpolní kontingenční tabulce. V tomto případě je test o něco konkrétnější vhledem k malému počtu buněk v rámci tabulky. Testová statistika je pak rovna [2]

$$\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

Pokud vyjde hodnota této statistiky vyšší než je hodnota kritického oboru, pak zamítáme nulovou hypotézu H_0 o nezávislosti. Ve čtyřpolních tabulkách může případ, kdy je počet očekávaných četností menší než pět, nastat častěji. Proto je možné použít korigované χ^2 testy. Jedním z příkladů může být *Yeatesova korekce* [16]

$$\chi^2 = n\frac{\left(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2}\right)^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{21}+n_{22})(n_{12}+n_{22})}$$

Při použití této korekce dojde ke snížení hodnoty testového kritéria, což znamená, že je obtížnější zamítnou nulovou hypotézu H_0 o nezávislosti. Díky tomu se sníží riziko chyby prvního duhu, že zamítneme nulovou hypotézu H_0, která je správná. Bohužel však zároveň dochází ke zvýšení rizika chyby druhého druhu, že nezamítneme nulovou hypotézu H_0, která je špatná. Z tohoto důvodu se použití korekcí moc nedoporučuje. Vzhledem k jednoduchosti χ^2 testu ve čtyřpolní tabulce je velmi často využíván i v případech, kdy jsou jiné testy daleko přesnější

a výhodnější. Často také dochází ke kategorizování spojité veličiny, což vede ke ztrátě důležitých informací, které původní datový soubor obsahoval [16].

4.2.1 McNemarův test

McNemarův test lze použít pro výpočty pouze ve čtyřpolní kontingenční tabulce. Test slouží ke sledování přítomnosti či naopak nepřítomnosti nějakého znaku. V celém datovém souboru je proveden jeden zákrok, po kterém se opět sleduje přítomnost či nepřítomnost daného znaku. Cílem tohoto testu je zjistit, zda zákrok změnil pravděpodobnost výskytu sledovaného znaku či nikoliv. Testová statistika je vyjádřena matematickým vztahem

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

a má asymptotické rozdělení χ^2. Nulovou hypotézu H_0 zamítáme v případě, že je hodnota testové statistiky větší než testové kritérium na hladině významnosti α . Potom můžeme říci, že zákrok ovlivnil pravděpodobnosti výskytu znaku. Pokud však výsledná hodnota testové statistiky spadá do kritického oboru, nelze nulovou hypotézu H_0 zamítnout, což znamená, že provedený zákrok neovlivnil pravděpodobnost výskytu znaku.[2]

Chí kvadrát testy bohužel žádný GIS software vypočítat neumí. Lze je počítat v softwarech R project a MS Excel.

4.2.2 Kruskal-Wallisův test

Kruskal-Wallisův test je jedním z neparametrických testů závislosti. Test slouží k ověření hypotézy o nezávislosti dvou jevů pocházejících z jednoho datového souboru.

Každé hodnotě souboru je přiřazeno vzestupně pořadové číslo, stejným hodnotám pak pořadí průměrné. Následně jsou sečteny pořadové čísla jednotlivých pozorování pro každý původní soubor, čímž jsou získány součty pořadových čísel.

Testová statistika má tvar:

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{2} k \frac{T_i^2}{n_i} - 3(n+1),$$

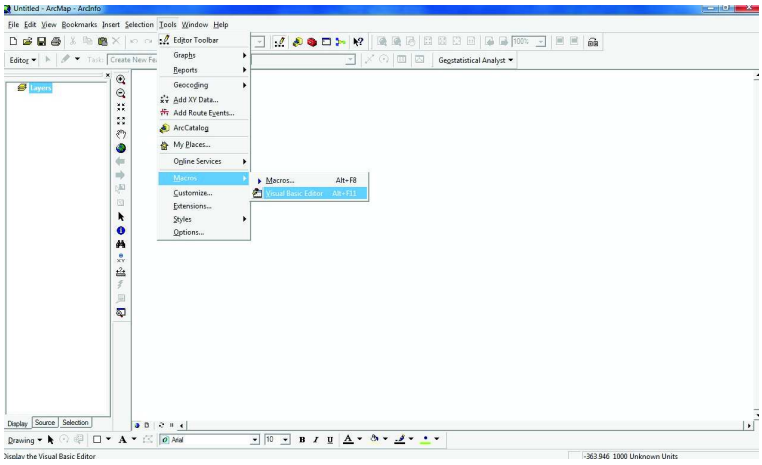
kde $n = n_1 + n_2 + \ldots + n_k$. Tato statistika má asymptotické χ^2 rozdělení o $k-1$ stupních volností.

4.2.3 Lillieforsův test

Lillieforsův test je dalším z používaných neparametrických testů nezávislosti. Test je založen na odhadu průměru a rozptylu datového souboru. Následně je nalezen maximální rozdíl mezi naměřenými a očekávanými četnostmi z normálního rozdělení. Nakonec je testováno, zda je tento rozdíl dostatečně významný k zamítnutí nulové hypotézy.

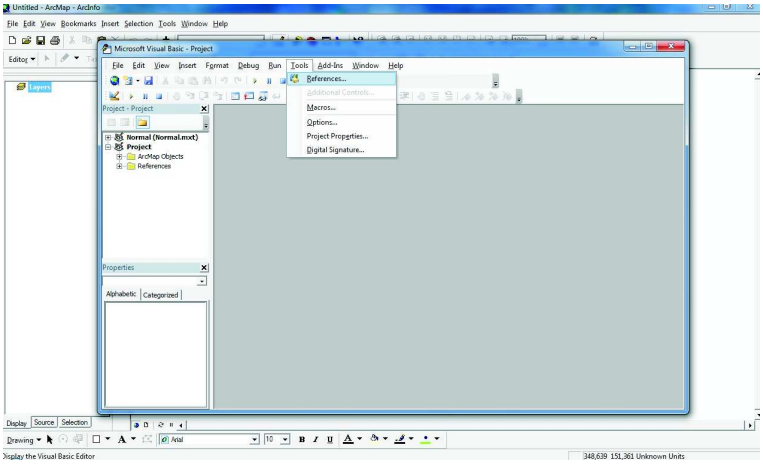
4.2.4 Dostupnost testu nezávislosti v různých softwarech

Software R project lze propojit se softwarem ArcMap pomocí skriptovacího jazyka Visual Basic. Po vytvoření nového projektu je třeba jej uložit a pojmenovat. Pomocí nástroje Tools lze spustit Visual Editor[6].



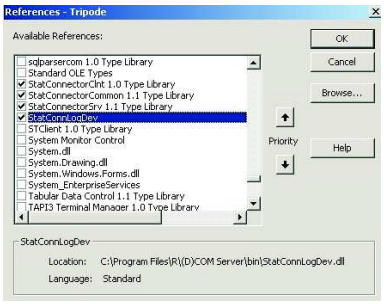
Obrázek 11: Ukázka zapnutí editoru Visual Basic v softwaru ArcMap

Dalším krokem je výběr potřebných referencí z nástroje nástroje Tools.



Obrázek 12: Ukázka zapnutí referencí v editoru Visual Basic v softwaru ArcMap

Nyní už stačí vybrat správné odkazy.



Obrázek 13: Ukázka referencí, které je třeba vybrat

Požívání statistických výpočtů pomocí tohoto propojení je však velice složité. Je jednodušší provádět výpočty v softwaru R-project a jejich výsledky následně správně interpretovat a vizualizovat, aby byly složité výpočetní úlohy snadno interpretovatelné i neodborníkem v dané problematice.

Další možností, jak provádět výpočty přímo v softwaru ArcMap, je vytvoření skriptu a jeho následný import do programu ArcMap. Skript je možno napsat v

softwaru Python, například ve verzi 2.6.2, nebo v softwaru Visual Basic. Oba tyto softwary obsahují standardní nástroje pro matematické operace a lze je snadno importovat přímo do toolboxu v ArcMapu.

5 Vyhodnocení datového souboru

5.1 Popis a sběr dat

Sběr dat proběhl pod záštitou společnosti . Tato společnost dostala zakázku od kraje. Šlo o průzkum skladby jízdních dokladů v okrese Přerov, konkrétně v zóně . Tato studie byla provedena na popud poskytovatelů veřejné dopravy v , kterými jsou . Ti dostávají dotace od kraje na základě toho, kolik cestujících převážejí. Na území je však možné cestovat s dokladem jednoho autobusového dopravce v zóně v autobuse jiného dopravce. Proto byl průzkum zaměřen hlavně na tuto křížovou přepravu. Na základě tohoto průzkumu se pak uvidí, zda kraj přerozděluje finance správně, či je třeba výši těchto dotací jednotlivým poskytovatelům přepravy změnit.

Průzkum proběhl v hlavním sčítacím dni, a to v úterý 13. dubna 2010. Některá dodatečná sčítání byla provedena následující dny, tj. 14. a 15. dubna 2010 a 20. dubna 2010. Záměrně bylo sčítání provedeno v úterý, protože se jedná o běžný pracovní den. V pondělí a v pátek by mohly být výsledky zkresleny, jelikož začíná, případně končí, pracovní týden. To znamená, že je možnost že se studenti či pracující, kteří přechodně přebývají v průběhu týdne právě v , přesunují z/do míst trvalého bydliště.

Sčítání probíhalo jak ve spojích, tak i na autobusových zastávkách od prvního ranního spoje až po poslední večerní spoj. Celkem bylo v terénu 45 sčítacích komisařů z řad studentů (dále jen sčítač), kteří kontrolovali všechny linky. V každém autobusu městské hromadné dopravy (MHD) seděl vždy jeden sčítač. Jeho úkolem bylo zejména hlídat, zda do autobusu MHD, jehož provozovatelem je , nastoupil cestující s předplatnou jízdenkou jiného dopravce. Aby se s daty daly provádět různé početní i grafické analýzy, byly do formuláře (viz příloha 1) přidány také kolonky pro záznam počtu nastupujících cestujících na jednotlivých zastávkách a počet cestujících ve voze mezi jednotlivými zastávkami.

Veřejná linková doprava (VLD) byla sčítána dvěma způsoby. Sčítači, jež kon-

trolovali VLD, se pohybovali nejen v linkách veřejné linkové dopravy, ale také na některých zastávkách. Většinou tak tento sčítač nastoupil do autobusu VLD, projel s ním jeho trasu a dále pokračoval sčítáním cestujících na zastávce. I tito sčítači zaznamenávali do svých formulářů (viz příloha 2) nejen cestující, kteří cestovali na křížový jízdní doklad, ale také počty nastupujících a vystupujících cestujících.

Poslední variantou provedeného sčítání bylo sčítání pouze v zastávce. Na některých zastávkách tak v průběhu celého dne stal sčítač, který zaznamenával jak cestující s městskou hromadnou dopravou, tak cestující s veřejnou linkovou dopravou. I tento sčítač zaznamenával do svého formuláře cestující s křížovými jízdenkami a počty nastupujících a vystupujících cestujících v jednotlivých spojích, které na zastávce zastavovaly.

V průběhu sčítání skladby jízdních dokladů se v autobusech pohybovala kontrolorka z krajského úřadu, která sledovala, zda sčítači opravdu dělají, co mají. Další možností kontroly je porovnat údaj ve formuláři konkrétního spoje, který zastavoval na zastávce, která byla obsazena sčítačem. Záznam o nástupu cestujících se tak musí shodovat jak ve formuláři sčítače v autobuse, tak ve formuláři sčítače v zastávce.

Firma kladla důraz zejména na počty cestujících s křížovým jízdním dokladem a záznam o jaký konkrétní jízdní doklad šlo. Velmi často se proto stávalo, že sčítači zaznamenávali pouze počet nastupujících cestujících a počty a druhy právě těchto křížových dokladů. Bohužel už nezaznamenávali počty cestujících ve voze, což velmi snižuje množství informací, které lze z těchto formulářů získat. Kompletní data (počet nastupujících cestujících, počet cestujících ve voze mezi jednotlivými zastávkami a počty a druhy křížových jízdních dokladů) jsou k dispozici pouze pro linku č. 4 a linku č. 5 městské hromadné dopravy. U ostatních linek jsou tato kompletní data k dispozici pouze u některých spojů.

5.2 Testování normality

Jak bylo uvedeno výše, kompletní datové sady jsou pouze pro autobusy linky č. 4 a linky č. 5. Proto bude většina testování provedena právě na těchto dvou linkách.

Aby bylo možné s daty dále pracovat, je nejprve nutné ověřit, jejich normalitu. Na základě výsledků testů se pak rozhodne, jaký bude další postup při testování. Pro testování normality byl po konzultaci s vedoucím práce použit v softwaru R-project Shapirův test. Normalitu bylo třeba otestovat jak pro počty nastupujících a vystupujících cestujících v konkrétní zastávce v jednotlivých časech (spojích), tak pro počty nastupujících a vystupujících ve všech zastávkách pro každý konkrétní spoj (konkrétní čas) jízdy.

U výsledků testů je třeba sledovat hodnotu p-value. Ta nám udává výsledek testu. Pokud je hodnota p-value menší než $0,05$, pak zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A. Znamená to tedy, že pokud testujeme normalitu dat a p-value vyjde menší než $0,05$, pak řekneme, že tato data nepocházejí z normálního rozdělení.

Do softwaru R-project je třeba ověřovat normalitu pro každou položku zvlášť. Do proměnné x se tedy postupně přiřazují hodnoty, pro které chceme určit normalitu. Následně pak provedeme pro proměnnou x Shapirův test.

```
> x <- c (0,5,4,2,1,8,9,3,2,12,9,8,5,9,12,8,15,14,16,12,7,7,14,5,9,
11,8,6,5,5,10,11,15,6,14,10,30,27,13,6,18,5,16,8,13,2,3,10,6,2,3,5,
3,11,5,1,3,4,2,0,5,0,8)
> shapiro.test (x)

        Shapiro-Wilk normality test

data:  x
W = 0.9005, p-value = 9.41e-05
```

V následující tabulce 2 jsou uvedeny výsledky Shapirova testu, pomocí něhož byla testována normalita rozložení dat v jednotlivých stanicích v průběhu celého dne. V prvním sloupci jsou vypsány všechny autobusové stanice na trase linky č. 5. Jelikož jsou všechny výsledné p-value menší než $0,05$, znamená to, že rozložení dat není v ani jednom případě normální, tedy že tato data nepocházejí z normálního rozdělení.

Autobusové zastávky	**Nástup**	**Výstup**
	$8.188 * 10^{-5}$	
	$6.79 * 10^{-7}$	$4.748 * 10^{-15}$
	$7.389 * 10^{-6}$	$5.969 * 10^{-9}$
	$3.294 * 10^{-6}$	$4.826 * 10^{-5}$
	0.00161	$7.47 * 10^{-9}$
	0.0008005	$1.312 * 10^{-7}$
	$7.878 * 10^{-10}$	$7.22 * 10^{-9}$
	$1.427 * 10^{-7}$	$8.495 * 10^{-7}$
	$3.102 * 10^{-10}$	0.0032
	$3.232 * 10^{-6}$	$7.822 * 10^{-5}$
	$9.41 * 10^{-5}$	$1.373 * 10^{-7}$
	$1.900 * 10^{-6}$	$2.044 * 10^{-9}$
	$2.516 * 10^{-7}$	$1.292 * 10^{-6}$
	$6.679 * 10^{-9}$	$9.744 * 10^{-10}$
	$1.410 * 10^{-7}$	$1.999 * 10^{-6}$
	$2.145 * 10^{-8}$	$5.811 * 10^{-6}$
	$7.873 * 10^{-10}$	$1.229 * 10^{-8}$
	$9.102 * 10^{-16}$	$6.258 * 10^{-6}$

Tabulka 2: Výsledné hodnoty p-value pro nástup a výstup v jednotlivých stanicích linky č. 5 městské hromadné dopravy naměřené 13. dubna 2010

```
> x <- c (0,0,3,0,1,0,2,2,1,1,3,4,1,3,0,2,5,3,2,3,2,0,1,2,2,3,1,8,0,
3,3,3,3,3,2,11,7,6,3,6,8,3,5,5,10,3,4,5,5,3,2,7,5,4,4,4,1,2,2,1,2,0,
```

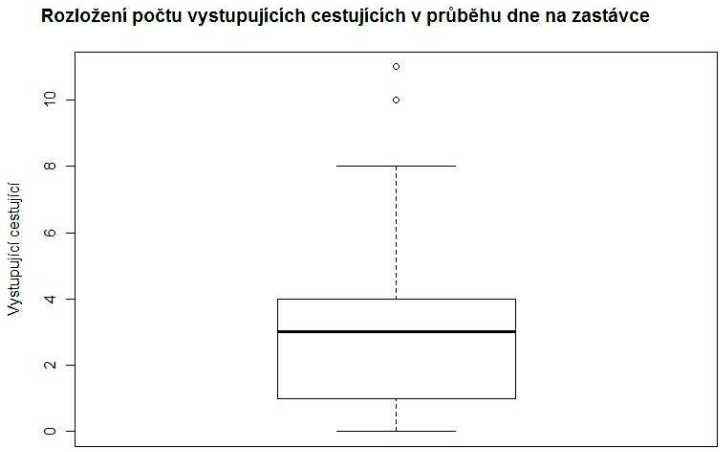
```
0)
> boxplot (x, main="Rozložení počtu vystupujících cestujících v prů-
během dne na zastávce           ", ylab="Vystupující cestující")
```

Autobusové zastávky	Nástup	Výstup
	$4.703 * 10^{-6}$	
	$2.944 * 10^{-5}$	
	$1.621 * 10^{-5}$	$7.086 * 10^{-6}$
	$1.015 * 10^{-7}$	$2.288 * 10^{-6}$
	0.02745	0.0007708
	$7.33 * 10^{-5}$	$4.711 * 10^{-6}$
	$2.355 * 10^{-6}$	$3.963 * 10^{-8}$
	0.001607	0.0004924
	0.00411	0.01162
	0.0005699	0.003469
	$5.623 * 10^{-7}$	$5.685 * 10^{-8}$
	$5.056 * 10^{-5}$	0.004902
	$1.631 * 10^{-5}$	0.0002954
	0.03931	$4.838 * 10^{-5}$
	$8.548 * 10^{-11}$	$4.154 * 10^{-10}$
	0.3352	$7.909 * 10^{-5}$
	0.003455	$6.528 * 10^{-5}$
	$1.548 * 10^{-5}$	0.0001291
	0.0008185	$4.873 * 10^{-5}$
	$7.338 * 10^{-6}$	**0.2292**
	0.0003081	**0.3063**
	$2.106 * 10^{-7}$	0.0004261
		0.08299

Tabulka 3: Výsledné hodnoty p-value pro nástup a výstup v jednotlivých stanicích linky č. 4 městské hromadné dopravy naměřené 13. dubna 2010

Tabulka 3 znázorňuje stejné výsledky jako tabulka 2 s tím rozdílem, že se

jedná o linku městské hromadné dopravy č. 4. Tučně jsou vyznačeny hodnoty p-value vyšší než $0,05$, což znamená, že jsou v těchto zastávkách počty nastupujících, případně vystupujících cestujících v normálním rozdělení. Jde však jen o několik záznamů a pro další testování to můžeme zanedbat.

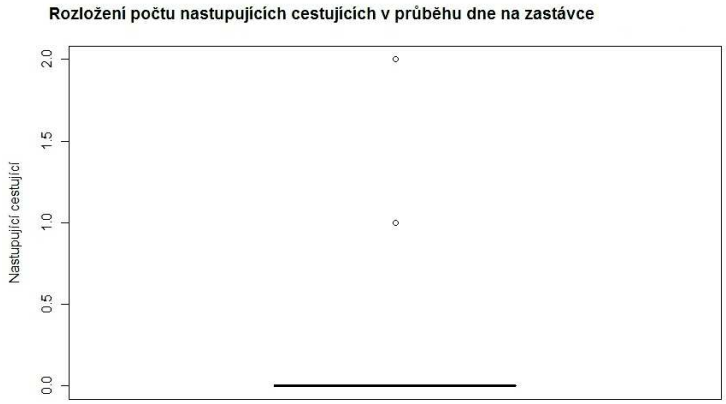


Obrázek 14: Box-plot s rozložením dat blízkým normálnímu rozdělení

Obrázek 14 ukazuje rozložení vystupujících cestujících na zastávce v průběhu celého dne 13. dubna 2010. Z box plotu lze snadno vyčíst, že data sice nemají normální rozdělení, ale velice se mu blíží. Vidíme také, že nejvíce zde vystupovali cestující v počtu 1 až 4. Kolečka za vousem značí extrémní hodnotu. Ta značí, že ve dvou případech vystoupilo na zastávce více než 10 cestujících.

V obrázku 15 byla záměrně využitá stejná metoda (box-plot) jako u předchozího příkladu, aby bylo možné porovnat, jak vypadá box plot dat, které neodpovídají normálnímu rozložení. To je z grafu patrné na první pohled. Je jasné, že na zastávce v průběhu dne prakticky nikdo nenastoupil. Jako extrémní hodnoty se pak jeví pouhé dva případy nástupu cestujících.

Všechny výsledné hodnoty p-value pro test normality jsou zaznamenány v tabulce, která je zařazena v přílohách (příloha 1). Většina hodnot je menších



Obrázek 15: Box plot s rozložením dat, které neodpovídá normálnímu rozdělení

než $0,05$ a proto můžeme zamítnout nulovou hypotézu o normálním rozdělení. Tučně jsou vyznačeny hodnoty vyšší než $0,5$, u kterých tedy nulovou hypotézu H_0 o normálním rozdělení nelze zamítnout. Tato normalita může být způsobena například větším množstvím dat. Vzhledem k tomu, že normalita dat byla prokázána pouze u minimální části vzorku, neovlivní tato normalita další testování.

Tabulka 4 je obdobná jako tabulka 3 s jedinou změnou a to, že jsou hodnoty p-value pro nástup a výstup cestujících ve všech zastávkách každého spoje jsou počítány pro linku č. 4.

Další možností testování je Wilkoxonův test. Byl použit pro zjištění, zda jsou si jednotlivé zastávky v nástupu a výstupu podobné či nikoliv. Toho dosáhneme uspořádáním zastávek podle průměrného počtu nastupujících (vystupujících) cestujících od nejvyššího po nejnižší. Pak porovnáváme vždy dvě po sobě jdoucí zastávky. Stejně jako u předchozího případu i zde udává výsledek testu hodnota p-value. Pokud je p-value menší než $0,05$ pak zamítáme nulovou hypotézu H_0 a řekneme, že si jednotlivé zastávky nejsou podobné v rozložení nastupujících a vystupujících cestujících v průběhu celého dne. Výsledky tohoto testu jsou znázorněny v následujících tabulkách (Tab. 5, Tab. 6)

Čas/spoj	Nástup	Výstup
	$5.358 * 10^{-6}$	$2.383 * 10^{-6}$
	$5.061 * 10^{-6}$	$1.776 * 10^{-7}$
	$8.64 * 10^{-5}$	$8.713 * 10^{-5}$
	$2.032 * 10^{-7}$	$2.935 * 10^{-6}$
	$7.664 * 10^{-6}$	0.0007035
	$4.196 * 10^{-5}$	$9.18 * 10^{-5}$
	$8.715 * 10^{-8}$	0.0006964
	0.0001145	0.02692
	$2.657 * 10^{-5}$	0.0004276
	0.0004441	0.006087
	0.0003935	0.0001741
	0.001991	0.001480
	0.003987	0.0001024
	0.0003081	0.0009455
	0.02066	0.0001685
	$2.936 * 10^{-5}$	0.001494
	0.01819	0.000802
	0.0695	0.0003722
	0.004719	$7.226 * 10^{-5}$
	0.0001331	0.008652
	0.0001571	0.0009119
	$5.747 * 10^{-6}$	$6.304 * 10^{-5}$
	$1.895 * 10^{-6}$	$4.867 * 10^{-6}$
	$2.325 * 10^{-6}$	$1.470 * 10^{-7}$
	$1.967 * 10^{-7}$	$7.569 * 10^{-6}$
	$1.750 * 10^{-7}$	$3.909 * 10^{-8}$

Tabulka 4: Výsledné hodnoty p-value pro nástup a výstup ve všech zastávkách každého spoje linky č. 4 naměřené 13. dubna 2010

Z tabulky 5 je jasně patrné, které zastávky jsou si svým rozložením nastupujících cestujících podobné. Tučně jsou vyznačeny hodnoty p-value vyšší než $0,05$.

Porovnávané stanice	Nástup
	0.0835
	0.03327
	0.0004792
	0.7178
	0.2201
	0.002302
	0.3761
	0.4536
	0.2119
	0.6514
	0.619
	0.3693
	0.0003552
	0.2112
	0.5685
	0.03059
	$\mathbf{2.901 * 10^{-5}}$

Tabulka 5: Výsledné hodnoty p-value pro Wilkoxův test při nástupu cestujících na lince č. 5 naměřeno 13. dubna 2010

To znamená, že zastávky s těmito hodnotami jsou u sledované charakteristiky (nástup cestujících) odlišné. Je možno říci, že zastávky a : jsou si v rozložení cestujících podobné. Další skupinu podobných zastávek tvoří zastávky . Následuje skupina podobných zastávek, tvoří ji . Poslední skupinou podobných zastávek jsou zastávky .

Tabulka 6 je svým systémem velice podobná předchozí tabulce č. 5. Liší se pouze v tom, že jsou zde vypsány hodnoty p-value po výstup a proto se liší i

Porovnávané stanice	Výstup
	0.1315
	0.02879
	0.002303
	0.4031
	0.5724
	0.1042
	0.438
	0.4408
	0.6215
	0.4274
	0.7161
	$\mathbf{5.575 * 10^{-6}}$
	0.2633
	0.7843
	0.02965
	$\mathbf{3.23 * 10^{-14}}$
	0.003687

Tabulka 6: : Výsledné hodnoty p-value pro Wilkoxonův test při výstupu cestujících na lince č. 5 naměřeno 13. dubna 2010

výsledné skupiny zastávek. První skupinu tvoří pouze dvě zastávky, které si jsou rozložením vystupujících cestujících podobné, a to sice
. Následuje velká skupina, do které patří zastávky

. Poslední zastávky, které jsou si rozložením podobné je zastávka

5.3 Test nezávislosti

Pokud víme, že data nepocházejí z normálního rozdělení, můžeme k dalšímu testování požít Kruskalův chí kvadrát test. Tohoto testu bylo využito pro zjištění závislosti v množství nastupujících a vystupujících lidí v jednotlivých zastávkách na konkrétním čase, tedy na spoji, se kterým se pohybují. První testovanou linkou byla linka č. 5. V případě nástupu i výstupu cestujících vyšla výsledná hodnota p-value menší než $2.2 * 10^{-16}$. Jelikož je tato výsledná hodnota p-value menší než $0,05$ pak zamítáme nulovou hypotézu H_0 o nezávislosti, a řekneme tedy, že počet nastupujících a vystupujících cestujících na jednotlivých zastávkách je závislý na čase ve kterém zde autobus zastaví. Další testovanou linkou na závislost nastupujících a vystupujících cestujících v zastávkách na konkrétním čase (spoji) byla linka č. 4. I v tomto případě byla výsledná hodnota p-value menší než $2.2 * 10^{-16}$ a bylo tedy možné zamítnout nulovou hypotézu o nezávislosti. I v lince č. 4 je tedy počet cestujících nastupujících a vystupujících na jednotlivých zastávkách závislý na čase, tedy na spoji, kterým se pohybují.

Nezávislost bylo možné také testovat na příkladu, zda je výstup a nástup do konkrétního spoje závislý na zastávce ve které zastavuje. Opět byl test proveden na linkách číslo 4 a 5. V případě linky číslo 4 vyšla výsledná hodnota p-value pro nástup cestujících $9.476 * 10^{-10}$ a pro výstup $7.47 * 10^{-12}$. Obě hodnoty jsou menší než $0,05$. Můžeme tedy zamítnout nulovou hypotézu o nezávislosti a říct, že nástup a výstup cestujících do konkrétního spoje je závislý na zastávce, ve které spoj zastavuje. Výsledek testu byl stejný i pro linku číslo 5. Jen výsledné hodnoty p value byly pro nástup $5.063 * 10^{-7}$ a pro výstup $2.921 * 10^{-7}$. Obě jsou tedy menší než $0,05$ a prokazují tedy taktéž závislost nástupu a výstupu cestujících do každého spoje na zastávce, ve které zastavují.

Kruskalova chí kvadrát testu nezávislosti bylo také využito pro testování, zda jsou na sobě závislé jednotlivé dny. Jelikož bylo měření některých spojů různých linek provedeno i v jiné dny než hlavní sčítání, lze tuto závislost posoudit. Konkrétně byl tento test proveden na některých spojích linky č. 4, které byly měřeny ve středu 14. dubna 2010 a ve čtvrtek 15. dubna 2010. Byla testována nulová

hypotéza H_0, zda jsou počty nastupujících a vystupujících cestujících na jednotlivých zastávkách nezávislé na dni v týdnu. U všech testovaných spojů vyšla hodnota p-value větší než $0,05$. Testovanou nulovou hypotézu tedy nelze zamítnou na hladině významnosti α. Znamená to tedy, že počty cestujících nastupujících a vystupujících v jednotlivých zastávkách nejsou závislé na dni v týdnu.

Spoj	**Nástup**	**Výstup**
	0.7843	0.4026
	0.531	0.685
	0.4342	0.858
	0.4428	0.6639
	0.1495	0.2810

Tabulka 7: Výsledné hodnoty p-value pro počet nstupujících a vystupujících cestujících ve vybraných spojích linky č. 4 naměřené ve dnech 14. - 15. dubna 2010

```
a <- c (0,6,29,7,4,7,4,5,9,9,6,31,10,17,9,14,7,13,11,7,1,8,
7,8,6,5,1,9,5,7,10,9,6,17,12,12,17,10,16,8,12,19,8,20,13,10,
20,12,11,16,7,27,6,8,3,9,14,6,3,1,4,8,2)
b <- c (0,1,2,0,1,0,1,1,1,0,5,7,0,0,0,5,2,0,2,2,2,1,3,3,2,2,
4,5,2,2,5,0,0,4,0,3,1,5,1,0,2,1,1,4,1,0,1,3,4,0,2,1,0,0,1,0,
2,0,1,0,0,0,0)
c <- c (0,0,5,5,1,1,0,3,1,2,5,4,5,1,0,3,4,3,3,2,2,0,3,0,4,3,
4,0,1,0,0,2,0,2,1,5,5,0,2,1,4,2,4,1,1,4,2,0,2,1,1,1,2,2,0,0,
1,0,0,1,0,1,0)
d <- c (1,19,9,5,1,6,9,10,4,6,7,9,2,8,4,1,11,3,4,1,1,4,0,10,
1,2,3,1,4,0,2,5,1,4,12,6,8,4,7,5,4,2,0,3,5,3,2,2,3,6,4,1,3,1,
1,1,1,6,2,0,1,1,0)
e <- c (0,6,4,13,3,2,5,10,3,7,6,7,4,4,10,3,18,4,10,5,7,6,6,
8,6,7,4,1,5,6,7,2,3,11,9,14,8,2,7,5,6,7,5,7,8,10,5,4,5,5,4,7,
5,3,2,3,9,4,5,5,5,1,0)
f <- c (0,1,3,1,0,1,5,0,3,0,3,6,1,2,4,4,2,2,6,3,10,3,5,4,9,4,
```

```
4,6,3,3,3,6,3,6,10,8,2,3,8,9,8,3,6,4,3,6,5,1,11,6,0,6,0,1,7,0,
1,1,0,0,3,0,0)
g <- c (0,1,8,4,0,0,1,2,1,0,4,1,0,3,0,0,1,1,1,0,1,1,0,3,1,1,2,
3,1,0,1,2,2,2,3,10,2,2,0,1,0,3,3,0,2,0,0,1,5,1,2,1,2,1,0,0,1,0,
0,1,0,0,0)
h <- c (1,1,2,0,1,2,1,1,2,1,4,1,1,1,2,2,3,4,6,0,2,1,4,2,3,0,4,
1,1,0,1,6,0,3,3,0,0,4,9,3,4,1,1,1,4,3,2,0,1,1,0,0,0,0,0,0,0,0,
0,0,0,1,1)
i <- c (1,5,3,1,0,0,4,0,3,4,9,2,0,0,4,3,3,3,20,3,5,3,1,1,0,1,
4,6,2,1,2,1,7,2,6,0,5,1,1,0,1,2,8,1,3,2,5,2,0,0,0,0,1,2,1,1,0,
4,1,0,0,0,0)
j <- c (1,2,7,2,6,2,6,6,3,10,4,9,1,2,4,5,4,1,3,2,3,3,3,2,1,2,
0,3,1,2,1,1,4,3,0,4,0,2,2,1,3,2,4,2,0,1,5,3,0,1,1,1,0,0,0,0,1,
1,0,0,1,0,0)
k <- c (0,5,4,2,1,8,9,3,2,12,9,8,5,9,12,8,15,14,16,12,7,7,14,
5,9,11,8,6,5,5,10,11,15,6,14,10,30,27,13,6,18,5,16,8,13,2,3,10,
6,2,3,5,3,11,5,1,3,4,2,0,5,0,8)
l <- c (3,2,3,1,4,0,2,5,4,1,9,0,3,0,3,1,5,1,3,0,2,1,1,0,4,4,
1,2,1,0,2,1,5,4,4,1,2,1,1,0,7,2,7,0,3,1,8,1,4,1,0,0,3,1,1,0,2,
3,0,0,1,1,0)
m <- c (3,2,0,1,1,8,3,5,1,0,3,1,3,5,4,2,7,3,2,5,0,7,1,3,1,2,
10,5,4,0,2,5,5,4,8,7,3,11,14,0,15,5,6,1,2,1,4,2,3,1,1,1,1,1,2,
0,1,0,0,1,1,0,0)
n <- c (1,1,0,1,0,0,1,0,1,1,0,1,1,0,1,2,2,0,5,1,1,1,0,1,0,1,1,
0,2,0,0,1,5,0,0,2,2,0,4,2,1,0,2,1,3,1,2,0,3,0,1,0,1,1,1,0,1,0,
0,1,0,0,0)
o <- c (0,0,0,0,1,0,1,0,0,0,0,1,0,2,5,3,9,0,1,3,4,2,2,5,0,1,3,
2,7,2,2,5,3,1,9,3,5,2,3,3,7,2,2,2,1,0,4,1,1,0,2,2,1,4,0,1,0,0,
0,0,0,0,0)
p <- c (1,0,0,2,0,3,0,2,4,1,2,2,2,2,0,1,3,0,0,0,2,2,1,1,1,1,6,
0,0,0,5,2,3,4,0,1,5,0,2,0,5,0,1,0,6,1,0,1,0,0,2,1,2,0,0,0,0,0,
0,0,0,0,1)
q <- c (1,2,2,1,0,3,4,0,0,0,1,0,0,0,0,0,2,1,2,1,1,1,1,0,0,0,0,
```

```
0,2,0,0,0,1,0,0,3,1,0,1,0,2,0,1,0,2,0,1,3,2,0,0,0,0,0,0,0,0,0,
0,0,0,0,0)
r <- c (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,1,0,0,0,0,0,0,1,
1,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0)
kruskal.test(list(a, b, c, d, e, f, g, h, i, j, k, l, m, n, o,
p, q, r))
```

Dalším příkladem využití Kruskalova chí kvadrát testu je testování závislostí využívání křížových jízdních dokladů. Konkrétně šlo o testování nulové hypotézy H_0, zda je využívání jízdenek od veřejných linkových dopravců nezávislé. Alternativní hypotéza H_A by tedy znamenala, že využívání jízdenek od dopravce a jsou na sobě závislé. V následující tabulce je vidět, že ve většině případů je využívání jízdenek od autobusového dopravce . nezávislé na využívání jízdních dokladů od dopravce

linka	**p-value**
1	0.03284
2	0.06701
3	0.4195
4	0.8916
5	0.3300
8	0.1290

Tabulka 8: Výsledná hodnota p-value Kruskalova testu nezávislosti křížových dokladů jednotlivých dopravců, naměřeno 13. dubna 2010

5.4 Grafické vizualizace

Pro zjednodušení vyjádření závislosti počtu nastupujících a vystupujících cestujících na denní době byla data rozdělena do tří částí. V první části byla data vztahující se k době mezi 4 : 30 a 9 : 45. Tato skupina pak byla označena jako

dopoledne. Následovala skupina odpoledne, která obsahovala data z doby mezi 10 : 00 a 15 : 45. V rámci poslední skupiny byla data naměřena mezi 16 : 00 a 22 : 30. Tato poslední skupina pak byla označena jako večer. Všechny mapové výstupy jsou dělány pro tyto tři skupiny a je tedy možno porovnat, kolik cestujících se pohybuje v konkrétní denní dobu na konkrétní části trasy. Tyto tři mapy jsou pak doplněny ještě o mapu, která zahrnuje data naměřena v průběhu celého dne.

Pro linku č. 4 a č. 5 byly vytvořeny série tří map vždy pro danou denní dobu. V první mapě byla využita metoda kruhového strukturního kartodiagramu pro znázornění podílu průměrného počtu nastupujících a vystupujících cestujících na jednotlivé zastávce. Díky tomu lze snadno poznat, na které zastávce se více nastupuje či vystupuje.

Pro znázornění trasy linek byla využita stuhová metoda, která vyjadřuje průměrný počet cestujících, kteří se cestují ve voze mezi jednotlivými zastávkami. Pro další dvě mapy byla využita metoda kruhového kartodiagramu. Pomoci měnící se velikosti průměru diagramu je vyjádřen průměrný počet nastupujících případně vystupujících cestujících na jednotlivých stanicích. Pro nástup cestujících je zvolená žluta barva ve dvou odstínech v závislosti na tom, zda se jedná o zastávku ve směru do centra nebo o zastávku ve směru z centra. V případě výstupů je využito odstínů zelené barvy. Trasa linek je opět vyjádřená stuhovou metodou, kdy linie mění šířku podle průměrného množství cestujících, kteří jsou ve vozidle mezi jednotlivými zastávkami.

U linky č. 4 bylo vytvořeno navíc ještě šest mapových výstupů. Jsou tvořeny na stejném principu jako mapy předchozí s rozdílem, že jde pouze o data posbírána v průběhu celého dne. Na těchto mapách je znázorněno několik spojů, které byly naměřeny ve více dnech, aby bylo možno porovnat, zda se počty nastupujících a vystupujících cestujících a počty cestujících ve voze v jednotlivých dnech liší.

6 Shrnutí výsledků

Výsledkem práce je sepsaná rešerše na téma Chí kvadrát (χ^2) testy a ucelená teorie o chí kvadrát testech. U každého konkrétního testu je doplněna jeho dostupnost ve vybraných softwarech, které jsou běžně používané na katedře geoinformatiky Univerzity Palackého v Olomouci.

Podstatnou část tvoří testování analýzy závislosti, ke kterým slouží popisované chí kvadrát (χ^2) testy. Aby bylo možno tyto testy použít, je třeba vědět, z jakého rozdělení jsou data. Proto bylo nutné nejprve pro všechny datové sady provést testy normality.

Testy normality byly použity na testování nástupu a výstupu cestujících konkrétních zastávkách do jednotlivých spojů. Tyto testy byly použity na lince číslo 4 a lince číslo 5, jelikož pro tyto dvě linky byly k dispozici kompletní data o nástupu i výstupu. V případě linky číslo 5 bylo výsledkem testu každé zastávky, že data nepocházejí z normálního rozdělení. U linky číslo 4 vyšlo rozdělení dat u některých zastávek jako normální. Mohlo to být způsobeno větším množstvím a rozsahem dat v těchto konkrétních autobusových zastávkách. Jelikož se však jednalo pouze o malou část dat, bylo možno tento výsledek zanedbat a pokračovat v dalším testování jakoby všechna data nepocházela z normálního rozdělení.

Před provedením testů závislosti bylo ještě zkoumáno, zda jsou si zastávky v rozložení nastupujících a vystupujících cestujících v průběhu dne podobné. Po provedení Wilkoxova testu byly vytvořeny skupiny zastávek linek číslo 4 a 5, které jsou si v nástupu a výstupu cestujících podobné.

U linky číslo 5 bylo tímto způsobem vytvořeno u nástupu 7 skupin podobných zastávek. První skupinu tvoří zastávky
. Do druhé skupiny patří jediná zastávka a to sice . Do třetí skupiny spadají zastávky
zastávka . Čtrvrtou skupinu tvoří největší počet zastávek a to zastávka

. Do páté skupiny spadají tři autobusové zastávky, podobné si svým rozložením nastupujících cestujících v průběhu dne a to sice zastávka (ve směru zpět na). Poslední dvě skupiny tvoří vždy jedna zastávka. Těmito zastávkami jsou (ve směru zpět do) a (také ve směru do).

V případě rozložení vystupujících cestujících v průběhu dne v jednotlivých zastávkách linky číslo 5 vzniklo také sedm skupin zastávek, které si jsou rozložením cestujících podobné. První skupinu zastávky . Ve druhé skupině je samostatná zastávka . Třetí a zároveň největší skupinu tvoří zastávky (ve směru do i z), (ve směru do), . Poslední skupinu zastávek s podobných rozložením vystupujících cestujících tvoří zastávky (ve směru do), . Pak už následují jen poslední tři zastávky, které tvoří každá samostatnou skupinu. Jde o zastávky (ve směru z), (ve směru z).

Nejdůležitější částí práce bylo testování nezávislosti. Postupně bylo testováno několik veličin, u kterých bylo třeba zjistit jejich závislost. Testy nezávislosti byla prokázána závislost mezi nastupujícími cestujícími na konkrétní zastávce a denní době, stejně jako byla prokázána závislost vystupujících cestujících na konkrétní zastávce na denní době. Dalším úkolem bylo prokázat, zda je závislý počet nastupujících a vystupujících cestujících do konkrétního spoje na zastávce ve které do spoje nastupují, respektive vystupují. I v tomto případě byla závislost prokázána testováním.

Dalším případem kdy bylo využito testu nezávislosti, bylo v případě křížových jízdních dokladů. Bylo zde testováno, zda je využívání jízdních dokladů autobusového dopravce ve vozech městské hromadné dopravy v závislé na využívání jízdních dokladů od autobusového dopravce

. Tato závislost byla testována u všech linek MHD .
U linky číslo jedna prokázal výsledek testu závislost mezi využíváním křížových jízdenek od . U ostatních linek MHD $(2, 3, 4, 5, 8)$ byla prokázána nezávislost mezi využíváním těchto křížových jízdních dokladů v jednotlivých spojích.

Posledním a nejdůležitějším testem bylo vyhodnocení, zda jsou na sobě závislé jednotlivé dny. Měření probíhalo ve více sčítacích dnech a některé spoje, které se nepodařilo sčítat v hlavní měřící den, byly dodatečně sečteny v jiné dny. Bylo třeba prokázat, zda jsou počty nastupujících a vystupujících cestujících na jednotlivých zastávkách, případně v jednotlivých spojích, závislé na dni v týdnu. Tento test mohl být použit pouze pro linku 4, jelikož právě na této lince byly sčítány některé spoje ve více dnech. Pomocí těchto získaných údajů se podařilo prokázat nezávislost mezi jednotlivými sčítacími dny. Díky tomuto výsledku mohou být spoje, které byly dodatečně sčítány v jiný den zařazeny do hlavního sčítacího dne a bylo tak možno dát dohromady data pro všechny spoje jednotlivých linek, jako by byly všechny sčítány v jeden jednotlivý den.

V grafických vizualizacích bylo využito stuhové metody pro znázornění počtu cestujících ve voze. Pro počty nastupujících a vystupujících cestujících na jednotlivých zastávkách byla zvolena metoda kartodiagramu. Díky těmto grafickým vizualizacím byla prokázána závislost počtu nastupujících a vystupujících cestujících na denní době. Pro zjednodušení byl den rozdělen do tří období na dopoledne (od 4 : 30 do 9 : 45), odpoledne (od 10 : 00 do 15 : 45) a večer (16 : 00 do 22 : 00). I při tomto zjednodušení lze vidět závislost na denní době, kdy nejvíce cestujících směrem do centra cestuje dopoledne, zatímco z centra jsou větší počty cestujících v odpoledních hodinách. Také závislost nástupu a výstupu cestujících na jednotlivých zastávkách lze z map snadno vyčíst. Díky velikosti kartodiagramu lze snadno rozpoznat počty nastupujících a vystupujících cestujících na jednotlivých zastávkách. Jako poslední vyjadřovací prostředek byl použit strukturní kartodiagram, který znázorňuje podíl nastupujících a vystupujících cestujících na jednotlivých zastávkách. Lze pak snadno rozpoznat, na které autobusové zastávky jsou převážně výstupní či nástupní.

7 Diskuse

V průběhu práce byla vypracována teoretická část na téma Chí kvadrát (χ^2) testy. K sepsání této části bylo třeba prostudovat množství materiálů.

Další částí bylo praktické zpravování dat. Veškeré výpočty byly provedeny v softwaru R project. Existují i možnosti, jak toto testování provést přímo v softwaru ArcMap. Tyto možnosti byly popsány již dříve v podkapitole 5.2.2, která je věnována dostupnosti výpočtu testu nezávislosti v různých softwarech. Byly zde popsány dvě možnosti, jak provést výpočty v softwaru ArcMap. Jednou z nich je propojení softwaru R project se softwarem ArcMap. Toto propojení je velmi jednoduché a je zvládnutelné ve třech krocích. Zde však veškerá jednoduchost končí. Propojením bylo otevřeno prostředí editoru Visual Basicu. Tento editor je objektově orientovaný programovací jazyk. Avšak provádění složitějších matematických a statistických výpočtů vyžaduje velmi složitý programovací kód a výbornou znalost tohoto prostředí. Další popsanou metodou v podkapitole 5.2.2. je napsání skriptu v prostředí Python. Výsledný skript je pak možné importovat do prostředí ArcMap. Takto vytvořený skript by byl však vhodný pouze na jedno použití a na soubor s konkrétním počtem řádků a sloupců. Zobecnění skriptu by bylo velice náročné.

Výsledkem Chí kvadrát (χ^2) testu je jedno číslo. Toto číslo nám pouze řekne, zda jsou jevy na sobě závislé, či nikoliv. Je proto zbytečné, snažit se cokoliv naprogramovat a přidat to do prostředí ArcMap, když například software R project je volně dostupný a je to jeden z nejlepších statistických softwarů. Vzhledem k tomu, že je výsledkem testu pouze jedno číslo, nelze tento test ani zvizualizovat.

Nástroj Geostatistical Analyst, který je součástí softwaru ArcMap disponuje pouze dvěma možnostmi jak využít Chí kvadrát (χ^2) testů a to konkrétné testů normality. Obsahují totiž možnosti grafických vizualizací těchto testů ve formě histogramu a QQ plotu. Tyto dva grafy ukazují rozložení dat.

Mapové výstupy, které znázorňují závislost a nezávislost záleží pouze na auto-

rovi a na tom, jak konkrétní závislost zobrazí. Měly by být vytvořeny tak, ať je z nich výsledek testu nezávislosti jasně viditelný. Mělo by být tedy na první pohled vidět, zda jsou sledované jevy na sobě závislé či nikoliv. Pokud jde o závislost jevu na čase, je vhodné vytvořit několik map v časovém sledu a jev na nich porovnat a doplnit popisem. V jiných případech lze závislost zobrazit v jednom mapovém poli.

8 Závěr

Hlavní náplní práce bylo popsání teorie Chí kvadrát (χ^2) testů a následná aplikace této teorie na praktickém vypracování datového souboru. Aby bylo možno sepsat teoretickou část, bylo třeba nejprve nastudovat odbornou literaturu a projít studie, které se zabývají aplikací Chí kvadrát (χ^2) testu v prostředí ArcMap.

Teoretická část je rozdělena do několika částí. Vzhledem k tomu, že je Chí kvadrát (χ^2) test jedním z testů normality, je v úvodní kapitole vysvětlena teorie právě těchto testů normality, konkrétně jejich grafických vizualizací. Některé z těchto grafických vizualizací lze provést i pomoci nástroje Geostatistical Analyst. Pro srovnání je v textu uveden příklad vytvoření několika grafických možností testování normality pomocí softwaru ArcMap, R-project a Microsoft Excel. Stěžejní kapitola je věnována samotnému Chí kvadrát (χ^2) testu. V této kapitole je možno nalézt vzorce pro výpočet testu normality i testu nezávislosti a vysvětlení, k čemu a za jakých podmínek lze tyto testy využít. Vzhledem k tomu, že jsou testy nezávislosti počítány pro kontingenční tabulky, je teorie doplněna také o popis ukázku kontingenční tabulky. U všech zmíněných testů a grafů je popsána také možnost jejich použití či vytvoření v různých softwarech. Při možnostech softwaru je srovnáván software ArcMap, R-project, Microsoft Excel a software Idrisi ve verzi Taiga.

Popsána teorie je pak použita na získaném data setu. V průběhu práce jsou tato data blíže popsána, včetně způsobu jejich sběru. Na těchto datech byly testovány normality a různé závislosti. Nejprve bylo na datech provedeno početní testování, následně byly tyto testy graficky vyjádřený pomocí mapových výstupů. Při početních úlohách byl použit Shapiro-Wilkoxonův test pro testování normality, kdy u většiny dat bylo prokázáno nenormální rozdělení. V případě některých datových sad byla prokázána normalita. Jednalo se však pouze o minimální vzorek, a proto mohl být i na tyto sady dále použit test nezávislosti. Nezávislost byla testována pomocí Kruskalova Chí kvadrát testu. Při ověřování závislostí mezi počtem nastupujících a vystupujících cestujících a konkrétními zastávkami případně na konkrétní denní době byla prokázána závislost. Dále bylo nutné otestovat, zda

jsou tyto počty cestujících závislé také na dnech v týdnu. Jelikož bylo měření provedeno v několika dnech, bylo potřeba potvrdit nezávislost těchto dat. Pomocí Kruskalova Chí kvadrát testu se tuto nezávislost podařilo prokázat. Tento test byl jedním ze stěžejních testů, které bylo potřeba na datech provést. Posledním použitým testem byl Wilkoxonův test, který umožnil vytvořit skupiny zastávek, které jsou si podobné v rozložení nastupujících případně vystupujících cestujících v průběhu celého dne.

Grafické výstupy byly vytvořeny pouze pro linky č. 4 a č. 5. Data byla rozdělena vždy na tři části, odpovídající denní době - na dopoledne, odpoledne a večer. Následně byly vytvořeny tři mapové výstupy pro každou denní dobu a tři mapové výstupy, které zahrnují data za celý den. Jedním mapovým výstupem je vyjádřen podíl nastupujících a vystupujících cestujících na jednotlivých zastávkách. Trasa linky je znázorněna stuhovou metodou, která vyjadřuje počet cestujících, kteří se pohybovali ve voze mezi jednotlivými zastávkami. V dalších dvou mapových výstupech je použitá metoda kartodiagramů, které vyjadřují počest nastupujících respektive vystupujících cestujících na jednotlivých autobusových zastávkách. Pomoci různých odstínů diagramů je znázorněno, které zastávky jsou ve směru do centra města a které jsou ve směru z centra města.

9 Seznam použité literatury

Reference

[1] ANDĚL, Jiří. Statistické metody. Praha : Matfyzpress, 1998. Testy dobré schody, s. 143-155. ISBN 80-85863-27-8.

[2] ANDĚL, Jiří. Statistické metody. Praha : Matfyzpress, 1998. Kontingenční tabulky, s. 157-174. ISBN 80-85863-27-8.

[3] HINDLS, Richard, et al. Statistika pro ekonomy. Vyd.7. Praha : Professional Publishing, 2006. Zpracování dat z výběrových zjišťování, s. 151-162. ISBN 80-86946-16-9.

[4] JANTOCH, Jaromír; VORLÍČKOVÁ, Dana. Vybrané metody statistické analýzy dat. Praha : Academia, 1992. Testy dobré schody, s. 114-116. ISBN 80-200-0204-9.

[5] NETOLICKÁ, Veronika. Testy normality [online]. [s.l.], 2008. 51 s. Bakalářská práce. Univerzita Palackého v Olomouci.

[6] BIVAND, R.; SAINT-JEAN, C. R / Arcgis Repository [online]. [cit. 2010-04-10]. R / Arcgis Repository. Dostupné z WWW:

<http://perso.univ-lr.fr/csaintje/Recherche/RArcgis/index.html>

[7] DUTTER, Rudolf. Homepage of Rudi Dutter [online]. 2003-09-09 [cit. 2010-02-24]. Chi-Quadrat-Test. Dostupné z WWW:

<http://www.statistik.tuwien.ac.at/public/dutt/>.

[8] ESSA, S. GIS MODELLING OF LAND DEGRADATION IN NORTHERN-JORDAN USING LANDSAT IMAGERY. Commision IV, papers [online]. 2004 [cit. 2009-10-21]. Dostupný z WWW:

<http://www.isprs.org/congresses/istanbul2004/comm4/papers/401.pdf>

.

[9] GILL, Martin; SPRIGGS, Angela. VYHODNOCENÍ ÚČINKU KAMEROVÝCH SYSTÉMŮ [online]. [s.l.] : [s.n.], 2007 [cit. 2010-02-28]. Dostupné z WWW:

<http://www.ok.cz/iksp/docs/336.pdf>. ISBN 978-80-7338-061-8

[10] HANZLOVÁ, Markéta, et al. Překryvné analýzy rastrových dat typu využití a pokryvu území. In HORÁK, Jiří; DĚRGEL, Pavel; KAPIAS, Adrian. Sympozium GIS Ostrava 2007. [s.l.] : [s.n.], 2007 [cit. 2010-02-23]. ISSN 1213-239X. Dostupné z WWW:

<http://gis.vsb.cz/GIS_Ostrava/GIS_Ova_2007/sbornik/default.htm>

[11] KAUSHIK, Inderjeet; OHRI, Anurag. GIS Development [online]. [cit. 2010-03-23]. Analysis of traffic flow in Varanasi city by using GIS. Dostupné z WWW:

<http://www.gisdevelopment.net/application/utility/transport/mwf_144abs.htm>

[12] KOHOUT, Václav. Katedra matematiky [online]. 6.5.2004 [cit. 2010-04-11]. Katedra matematiky. Dostupné z WWW:

<http://www.kmt.zcu.cz/person/Kohout/info_soubory/letnisem/SS/stat19.pdf>

[13] REMILLARD, Marguerite; ROY, Roy. ForestLandscapeEcologyLab [online]. 1993 [cit. 2010-03-20]. GIS technologies for aquatic macrophyte studies: Modeling applications. Dostupné z WWW:

<http://forestlandscape.wisc.edu/landscapeecology/articles/v08i03p163.pdf>

[14] RIMARČÍK, Marián. Štatistická analýza [online]. 2000 [cit. 2010-04-11]. Testy normality. Dostupné z WWW:

<http://rimarcik.com/navigator/normal.html>

.

[15] TOŠOVSKÁ, Libuše. Obchodní akademie a vyšší odborná škola Příbram : Projekt využití ICT [online]. 2009-2010, 9.3.2010 [cit. 2010-04-28]. Statistika. Dostupné z WWW:

<http://www.oapb.cz/skolst/projekt2009/statistika/index.html>

[16] ZVÁROVÁ, Jana. Základy statistiky pro biomedicínské obory [online]. [s.l.] : [s.n.], 2.3.1998 [cit. 2010-04-11]. Analýza kategoriálních dat, s. . Dostupné z WWW:

<http://new.euromise.org/czech/tajne/ucebnice/html/html/statist.html>

.

Summary

The main concern of the Bachelor's Thesis was to describe the theory of Chi square tests and then apply this theory to the practical a data set. In first part was necessary to study literature and pass the study, which involved applications of Chi square tests in the ArcMap. Then could be made the part of theory.

The theoretical part is divided into several parts. In an introductory chapter is explaining the theory of tests for normality, namely the graphical visualization because of the Chi square test is one of the tests of normality. Some of these graphical visualizations can be performed using the Geostatistical Analyst. For comparison, in the text is an example of more graphic capabilities of testing for normality using ArcMap software, R-project and Microsoft Excel. Pivotal chapter is devoted to the Chi square tests. It is possible to find there a formula for calculating the normality test and the test of independence and an explanation of what and under what conditions is possible to use these tests. The theory is supplement by the description of a sample contingency table because the tests of independence are calculates for the independence of contingency tables. For all these tests and graphs are also described their use or creation of different software. In the thesis are compared software ArcMap, R-project, Microsoft Excel and IDRISI version Taiga.

Described theory is then applied to the acquired data set. During the thesis the data is further described, including form of collection. On these data were tested for normality and various dependencies. It was first performed on data from numerical tests, then test was graphically expressed by maps outputs. In the mathematical task were used Shapiro-Wilkoxon test for testing normality, where the majority of the data showed abnormal distribution. For certain data has been demonstrated normality. But it was only a minimal sample and there for could be for these data use the same test of independence. Independence was tested using Kruskal Chi square test. Where verifying the number of embarking passenger and disembarking passengers on specific bus stops or a specific time of the day was demonstrated dependence. It was also necessary to test whether

these passenger numbers also depend on a weekday. Since the measurement was made in a few days would be needed to confirm the independence of these data. Using Kruskal Chi square was demonstrated this independence. This test was one of the key tests that were needed to make on these data. The last test was used Wilkoxon test, which enabled a group of bus station which are similar in distribution of embarking or disembarking passengers through the day.

Graphic outputs were generated for only a few lines of public transport. The data were divided into three parts, each corresponding to the time of the day - at morning, afternoon and the evening. Subsequently were created three maps outputs for each part of the day and three maps outputs include data of the whole day. One map output is expressed in percentage of embarking and disembarking passengers at various bus stops. Line route is shown using method of ribbon, which represents the number passengers who were in the bus between the bus stops. In the other two maps outputs were used the cartodiagrams, which represents embarking respectively disembarking passenger at various bus stops. Using different shades of cartodiagrams are shown the bus stops in the direction into the city centre and the bus stops in the direction out of the city centre.