UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA GEOINFORMATIKY - DEPARTMENTS GEOINFORMATICS

BAKALÁŘSKÁ PRÁCE

METODY ANALÝZY ZÁVISLOSTI S VYUŽITÍM KORELACE A LOGISTICKÉ REGRESE V PROSTŘEDÍ R

Vypracovala: **Zuzana Římská**

Vedoucí: **Mgr. Pavel Tuček, Ph.D.**

Olomouc 2010

Prohlášení

Prohlašuji, že jsem vytvořila tuto bakalářskou práci pod vedením Mgr. Pavla Tučka, Ph.D. a mého konzultanta Mgr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedla veškeré použité zdroje zpracované v mé bakalářské práci.

V Olomouci dne .....

27. května 2010 Zuzana Římská

Poděkování

Na začátek své bakalářské práce bych ráda vyjádřila vděčnost následujícím lidem za spolupráci při tvorbě práce a za konzultaci výsledků. Hlavní poděkování patří vedoucímu práce Mgr. Pavlu Tučkovi, Ph.D. z katedry geoinformatiky a mému konzultantovi Mgr. Karlu Hronovi, Ph. D. z katedry matematické analýzy a aplikací matematiky.

Obsah

Úvod

Matematika a statistika jsou mi velmi blízké již od střední školy, proto jsem chtěla využít své matematické znalosti i při nynějším studiu a odpovědět si na otázku do jaké míry jsou tyto obory slučitelné. Mohlo by se zdát, že logistická regrese a korelační analýza nejsou přímým oborem geoinformatiky, ale spíše matematiky a statistiky. A také používaný software R nepatří k nejběžněji používaným aplikacím pro ztvárnění vizuálního výsledku dat, ale opak je pravdou.

Proč by měl tedy geoinformatik, matematik a statistik křížit své znalosti, když má každý svou oborovou specializaci. Avšak musíme si uvědomit, že veškerá data, která se v oborech využívají, mají prostorová, časová a účelová hlediska, nad kterými se analyzují nebo simulují další výpočty. S tím rozdílem, že matematikovi a statistikovi stačí jako výsledek jeho matematická řeč vzorců a geoinformatik většinou zakončuje vizuálním výstupem, ale vytvořeným právě na základě matematiky a statistiky, což nemusí vůbec vědět, neboť některé softwary zpracují pokyn bez prozrazení, jak byl výpočet proveden.

Pokud spojíme řeč běžně psaných informací, řeč matematickou a řeč programovací, pochopíme daleko lépe problematiku zkoumaného jevu a budeme si je schopni i náležitě ověřit. Nebudeme omezeni svým úzkým jednooborovým pohledem a to nás povede k vlastní logické kontrole výsledku problému.

V mé práci především vysvětlím, jak je logistická regrese a korelační analýza pro práci s daty důležitá. Jednak s pomocí čisté matematické teorie, která je obecná a jejím pochopením získáme povědomí o základech, na kterých pracují všechny softwary.

V praktické části s využitím programování v R softwaru je ukázán postup práce s daty tak, aby byla matematická teorie pochopena i v praxi. Data se týkají poškození lesů polomy.

Vybrané faktory, které nejvíce usnadňují silnému větru vyvrácení stromů, použijeme pro výpočet pravděpodobnosti a metodou *ordinary kriging* zprostřed-

kujeme místa náchylnější k polomům.

Bude vysvětlena i daná teorie v dalších vybraných statistických softwarech, ale také v tradičním geoinformatickém softwaru, který je na naší katedře geoinformatiky University Palackého dostupný, a tím je ArcGIS Desktop 9.3.

Snad se geoinformatik matematicky poučí, aby si nemyslel, že program je jen klikací“ okno, které produkuje správný výsledek, ale že zde probíhají operace, kterým musí rozumět, pokud chce relevantní odpověď. A matematik a statistik by měl pochopit, že běžnému uživateli je prospěšnější svou vypovídající hodnotou vizuální ztvárnění než matematický zápis.

1 Cíle práce

Cílem bakalářské práce je vypracování pojednání o využití logistické regrese a korelační analýzy pro zjištění formy závislosti ve vybrané datové struktuře.

Nejprve je sepsána rešerše na téma analýzy závislosti s využitím logistické regrese a korelační analýzy, která uvede praktické příklady.

Použitá matematická teorie bude podrobně popsána tak, aby byla pochopena i nematematiky, ale udrží si i jistý standard matematického přednesu.

Detailní popsání prostředí R softwaru a dalších vybraných matematických, statistických programů, u kterých byla zjištěna dostupnost metod korelace a logistické regrese. Vyhodnocení možnosti řešení těchto metod také za pomoci softwaru, používaném na katedře geoinformatiky Univerzity Palackého v Olomouci.

Na poskytnutém datovém souboru bude předvedena matematická teorie s konkrétními výpočty a výsledkem. Všechny potřebné výpočty a tvorba grafických vizualizací je provedena v softwaru R.

2 Teoretický modul

Uvedení do matematické a statistické teorie s důrazem na problematiku korelace a logistické regrese s příklady, jak byla teorie využita v praxi.

2.1 Rešerše

Pojem korelace, logistická regrese a jejich analýzy nejsou ve světě ani u nás neznámým pojmem. Setkáváme se s nimi v mnoha odvětvích od využití v ekonomice po vědecké účely. Na zpracování dat v praktických a specifických úlohách lze nahlížet i níže uvedenými způsoby.

Korelace je statistická metoda, která ukáže, zda a jak silně jsou hodnoty ve vzájemném vztahu. Například jak jsou výška a váha vztaženy k faktu, že vyšší lidé jsou zpravidla těžší. Na tomto základě stojí mnohé studie vyhodnocování dat.

Korelační analýzy pomáhají předpovědět trendy, například vztah mezi cenou surové nafty a bilancí amerického dolaru. Pokud lze nakoupit barrel nafty pod 40 dolarů ,nákup je povolen. Pokud je nákup nad 40 dolarů, je otevřen i prodej. Použitím korelace (např. Pearsonovi korelace) lze odfiltrovat některé neprospěšné typy obchodů a vytvořit nová místa pro obchody výhodnější.

Logistická regrese je matematický model pro analýzu binárních dat s velkým rozsahem použitelnosti (fyzika, medicína a behaviorální vědy). Přesné závěry modelů logistické regrese jsou založeny na generování permutací distribuovaných statistikou pro regresní parametry.

Navzdory rychlosti moderních výpočetních systémů může být naše datová sada pro analýzu exaktních metod příliš velká, nebo může obsahovat nevhodně rozložené či nevyvážené pravděpodobnostní rozdělení.

V takovém případě je potřebná k přesnému podmíněnému přístupu alternativa *Monte Carlo* [23], která přemostí trhlinu mezi exaktními a asymptotickými odvozovacími metodami.

Problém je technicky nesnadný, protože tradiční metody vedou k masivnímu zamítaní vzorků dat, které nesplňují podmínku linearity podmíněných distribucích. Odborníci zde proto navrhli vzorkování pro eliminaci zamítaní.

Cyrus R. MEHTA a Nitin R. PATE představují postupy a modely síťového vzorkování a to přímou metodu pro zastoupení sítě v malé datové sadě. Vysvětlí i její konstrukci a zhodnotí příznaky tak, aby analýza dat byla co nejsprávnější a nejhodnotnější. Data v této studii pochází z automobilové společnosti z Massachusetts a pojednávají o nehodovosti automobilů v roce 1989 a o kalibraci "crash" testů pro zlepšení bezpečnosti automobilů.

Pro další nepřímé vzorkovací postupy *Monte Carlo* jsou použity Markovovy řetězce (Markov Chains), které jsou určeny pro práci s maticemi vzorkovacího vektoru. Dostupnost paměti je velice důležitá k správnému síťovému vzorkování, ale nikdy nejsme schopni zachytit její veškerou obsáhlost s dosavadními technickými možnostmi. Proto je nutné vytvořit takové metody, které méně důležité věci vhodně odstraní jako je tato MCMC (Markov Chain Monte Carlo) metoda.

Tento příklad je již více specifický a možná na začátek složitý na pochopení, ale díky této představě, jaké výpočty se dají provádět, lépe nahlédnete do základů korelace a logistické regrese.

Často se stává, že i při veškeré snaze o nejlepší popsání světa chybí v datech proměnné. I zde lze problém řešit pomocí redukování vícečetných konceptů a pomocí logistické regrese. Studie Tapabrata Maiti [22] poukazuje na softwarové balíky pro analýzu dat, které jsou zpravidla vybaveny regresními koeficienty, které jsou získány opakováním, s odhady malého nebo středně velkého datového balíku, ty mohou být neobjektivní (bias, biased). Výběry pak často narazí na to, že poslední parametr diverguje k nekonečnu. Standardní přístupy odhadování neberou v úvahu tento problém. Chybějící proměnné tedy mohou přidat potřebný nový pohled pro nalezení úspěšného řešení. V datové sadě se mohou vyskytnout tři problémy, jako je *bias* (neobjektivnost), špatný výběr a chybějící proměnné. Navržené metody řeší například SAS program.

Logistická regrese je speciální případ regrese. Předběžným průzkumem maximální pravděpodobnosti pomocí logistické regrese oproti exaktní logistické regresi (ELR) se zabívá Elizabeth N.King a Thomas P.Ryan. Maximální pravděpodobnost může generovat velice slabé výsledky za určitých podmínek např. u vzácných událostí. Primárně testujeme maximální pravděpodobnost v bližším rozdělení. Exaktní logistická regrese (ERL) je alternativním řešením, které nabízí potřebné nalezení odhadu.

Když tedy maximální pravděpodobnost nabízí slabé výsledky, nabízí se ERL použití pro řídká data nebo data nevyvážená. Pokud velká většina odpovídajících hodnot není 0 nebo 1.

Procento odpovědi ERL je ustanoveno okolo 50%, pokud se efekt pohyboval kolem této hranice, byl brán v úvahu. Slabé výsledky se tedy produkují, když procento je mnohem menší nebo mnohem větší než hranice. Pravděpodobnostní porovnávání křivky u normální regrese a ERL je důležité z hlediska rozsahu parametrů. Pokud je křivka plochá, tak je i odhad špatně determinován a použije se ERL. Ale přes veškeré výhody ERL můžeme dojít k závěru, že chyba ERL je větší, než bychom požadovali a musíme hledat jinou metodu.

Nejčastěji se běžný uživatel setká se statickými výpočty v lékařství. Lze například spočítat defibrotide pro jaterní VOD u dětí (Bone Marrow Transplantation [15]). Výsledky retrospektivní studie léčby pomocí defibrotide (DF) u dětí s žilně-oklusivní chorobou (VOD). Po nasazení léčby s DF se zpozdil příchod choroby VOD. Pomocí vícemožnostní analýzy logistické regrese se provedlo hledání, zda existuje spojení mezi předpokládanými faktory a s výsledky léčby. Běžná asymptotická logistická regrese je založena na parametrických předpokladech a vícemožnostní analýzu lze použít pro všechny možné proměnné. Modelováním logistickou regresí se vybraly dva parametry s velkou důležitostí (kdy DF pozastavila VOD).

S dostatečným množstvím pacientů a použitím vícemožnostní analýzy, by se statistická chyba snížila, ale to není možné, neboť není tolik testovacích sub-

jektů s nemocí. Ale s metodami analýz pro malé datasety se dá dojít k dobrým výsledkům. Použitý testovací software LogXact.

Je vidět v kolika důležitých směrech se dají korelace a logistická regrese využít. V mé bakalářské práci je uveden teoretický základ k pochopení a praktickou ukázku. Tato rešerše má povzbudit k dalšímu rozšíření znalostí, tedy ukázat kam až se dá v praktickém užití zajít.

2.2 První seznámení

Do statistiky zavedl pojem regrese britský učenec Francis Galton v rámci spojení "regres k průměru". Tím označil fakt, že synové vysokých otců jsou obvykle nižší, než byli jejich otcové, zatímco synové malých otců jsou vyšší než jejich rodiče. A podobně je tomu s jinými vlastnostmi, nejen u lidí.

Galtonův název se z jeho výzkumů přenosu vlastností mezi generacemi rozšířil na jakékoli zkoumání souvislostí mezi náhodnými veličinami (proměnné, jejichž hodnoty jsou jednoznačně určeny). Zde nalezneme vznik statistické regresní analýzy. Ta dnes patří k nejvýznamnějším metodám matematické statistiky a samostatně či ve spojení s jinými metodami se používá prakticky v každé oblasti empirické a aplikované vědy.

Regresní analýza je označení statistických metod, pomocí nichž odhadujeme hodnotu jisté náhodné veličiny na základě znalosti jiných veličin.
Příkladem uvažování v duchu regresní analýzy z běžného života může být, odhadujeme-li ráno, jaké bude přes den počasí na základě znalosti předpovědi počasí a toho, jaké je venku počasí nyní. V praxi je odhadování očekávané pooperační délky života pacientů trpících rakovinou Na základě zkušeností z minulých let, kdy se shromáždily předoperační údaje o zdravotním stavu většího počtu pacientů, například velikost a typ nádorů, věk pacientů apod. jakož i záznamy o délce života po operaci, lze pomocí vhodného typu regresní analýzy stanovit vzorec, s jehož pomocí bude možné u nového pacienta na základě znalosti jeho zdravotního

stavu odhadnout střední hodnotu očekávané doby přežití v případě operace. Je-li navíc k dispozici podobná analýza pro pacienty léčené konzervativně, lze pak tomuto novému pacientovi doporučit, který způsob léčby mu v dané situaci dává naději na delší přežití. Regresní analýzy se dají taktéž namodelovat jako úlohy v prostoru, které mě, jako geoinformatika, zajímají nejvíce. Například určit místa, kde je možné očekávat polom nebo kde je možné očekávat výskyt určité zvěře nebo květeny a apod.

Je důležité vysvětlit také pojem pravděpodobnost [30], se kterou regrese pracuje. Pravděpodobnost náhodného jevu je číslo, které je mírou očekávatelnosti výskytu jevu. Náhodným jevem rozumíme opakovatelnou činnost prováděnou za stejných (nebo přibližně stejných) podmínek, jejíž výsledek je nejistý a závisí na náhodě. Množinu všech možných výsledků pokusu značíme velkou omegou. Podmnožiny množiny všech možných výsledků pokusu se nazývají (náhodné) jevy. Nejvýznamnějším a dodnes inspirativním klasikem teorie pravděpodobnosti je Pierre-Simon Laplace.

Logistická regrese je speciální forma regrese, zabývající se problematikou odhadu pravděpodobnosti nějakého jevu (závisle proměnné) na základě určitých známých skutečností (nezávisle proměnných), které mohou ovlivnit výskyt jevu. Událost, zda zkoumaný jev nastal, se modeluje pomocí náhodné veličiny, která nabývá hodnoty 0, pokud jev nenastal, nebo 1, pokud jev nastal. Náhodná veličina, která nabývá dvou hodnot 0 a 1 má alternativní rozdělení.

Metoda logistické regrese předpokládá, že za podmínek, které určuje vektor X, bude náhodná veličina Y (X) rovna 1 s pravděpodobností, jejíž závislost na X můžeme vyjádřit pomocí logistické funkce.
Součástí logistické regrese je nutná specifikace vstupní proměnné, způsob odhadu pravděpodobnosti úspěchu, propojení vstupní proměnné s vysvětlujícími proměnnými. Odhady koeficientů a intervalů spolehlivosti a posouzení validity modelu. Přínosem logické regrese je možnost modelovat pravděpodobnost jako, funkci spojitých proměnných, predikce jsou mezi 1 a 0. Avšak není jednoduché

interpretovat regresní koeficienty.
Rozdělení veličiny se mění podle hodnot (střední hodnota, rozptyl, tvar a rozdělení). Při sledování vztahu mezi naměřenými hodnotami řešíme zaprvé, jaký je konkrétní matematický tvar závislosti mezi veličinami - regresní analýza a za druhé, zda existuje či neexistuje závislost (vztah, korelace) mezi veličinami. Případně sledujeme míru intenzity (těsnosti) tohoto vztahu - korelační analýza.

Korelace je vzájemný vztah, souvztažnost mezi znaky, veličinami a ději.
V teorii pravděpodobnosti a statistice, korelace, prostřednictvím koeficientu korelace, ukáže sílu a směr lineárního vztahu mezi dvěma náhodnými proměnnými. V širším slova smyslu existuje několik koeficientů, měřicích míru korelace, přizpůsobené k povaze dat.

Často se setkáváme s případem, kdy máme dvě či více proměnných a kdy kterákoliv z nich může vystupovat v roli vysvětlované a druhá či zbývající v roli vysvětlujících proměnných. Zkoumá-li se například souvislost výdajů na zboží A a na zboží B, pak v mnoha případech má význam vysvětlovat změny výdajů na zboží A změnami výdajů na zboží B a rovněž změny výdajů na zboží B vysvětlovat změnami výdajů na zboží A. Jinými slovy, vztah mezi výdaji na oba druhy zboží zkoumáme jako dvoustranný.

V takových případech považujeme n zjištěných dvojic hodnot obou proměnných za hodnoty dvourozměrné náhodné veličiny a hledáme její dvourozměrný pravděpodobnostní model. V případě většího počtu proměnných pak považujeme v případech tohoto druhu n pozorovaných trojic, čtveřic atd. hodnot za hodnoty trojrozměrné, čtyřrozměrné atd. náhodné veličiny a hledáme vhodný trojrozměrný, čtyřrozměrný atd. pravděpodobnostní model.

I když je toto první seznámení s teorií hutné, je třeba si představit celkový rozsah hned na začátku. Uvědoměním si základů se vytvoří větší nadhled na jistě rozsáhlou problematiku, kterou není možné v této práci kompletně shrnout. Tato část musí být dobře pochopena pro stabilní teoretický základ. Přečtením prvního

seznámení je patrné, že nedostatečným prostudováním nějakého z kroků může vést k nepochopení kroku následujícího.

2.3 Základní pojmy

K vysvětlení metod závislosti, korelace a logistické regrese je nutné pochopení matematického jazyka, který statistika využívá a na kterém jsou založeny statistické programy. Avšak některé programy vám nenabídnou celkový pohled, jak bylo testování provedeno a vypočteno. Podají výsledek, který nemusí být vypovídající, natož dobře vyhodnocen. Jak tedy zjistit důvěryhodnost a správnost, kterou vyžadujeme.

Komplexním pochopením obecné matematické a statistické teorie, a na ní založené problematiky výpočtů, budeme schopni pracovat, s jakýmkoli statistickým programem tak, abychom mohli zpětně rozšifrovat, jak bylo naloženo s daty, jaké funkce byly použity a zda vše vedlo k zdárnému výsledku.

Pochopení definic základních pojmů [9] a slovních spojení v textu je nejnutnější základ pro pochopení celkové problematiky statistického odhadu, testování hypotéz a na ně navazující další teorie.

Definice 2.1 Matematická statistika *je analýza číselných údajů, podléhajících náhodným vlivům shromažďovaných pro popis početných souborů. Na vyšetřované údaje pohlíží jako na realizaci náhodných veličin, pro jejichž rozdělení se snaží odvodit závěry. Postupy se opírají o teorii pravděpodobnosti.*

Definice 2.2 Náhodný pokus, *jehož výsledek není jednoznačně určen předepsanými podmínkami pokusu a který je za těchto podmínek mnohokrát opakovatelný.*

Definice 2.3 Náhodný jev *takové tvrzení o výsledku náhodného pokusu, o němž lze po realizaci pokusu jednoznačně rozhodnout, zda je nebo není pravdivé.*

Definice 2.4 Náhodná veličina *je výsledek náhodného pokusu, vyjádřený reálným číslem.*

Definice 2.5 Pravděpodobnost. *Je-li počet všech možných výsledků náhodného pokusu konečný a jednotlivé výsledky lze považovat za stejně pravděpodobné, pak pravděpodobnost libovolného náhodného jevu A lze stanovit tzv. klasickou definicí pravděpodobnosti jako P (A)=K/N. Přitom N je počet všech možných náhodných pokusů a K je počet všech výsledků pokusu, při kterých nastává jev A.*

Definice 2.6 Statistický model *je rozdělení pravděpodobnosti, na jehož základě vznikají číselné údaje pro statistický rozbor, kdy je předem rozhodnuto, z jakého modelu budeme vycházet.*

Definice 2.7 Náhodný výběr *z daného rozdělení pravděpodobností je soubor nezávislých náhodných veličin $X_1,\ldots,X_n$. Také zapisujeme náhodným vektorem $\boldsymbol{X} = (X_1,\ldots,X_n)'$. Číslo n se nazývá rozsah náhodného výběru. Tedy situace kdy n-krát navzájem nezávisle opakujeme za stejných podmínek určitý náhodný pokus.*

2.4 Teorie odhadu

Statistiky a odhady. Hodnoty konstruované na základě náhodného výběru za účelem statistického rozboru se nazývají *statistiky*. Statistiky konstruované k přibližnému určení charakteristik příslušného rozdělení se nazývají *výběrové (empirické) charakteristiky*. Statistiky konstruované k přibližnému určení parametrů se nazývají *odhady*. Vlastnosti odhadů v praktickém použití se preferují jako odhady určitých vlastností.

Kvalita odhadu parametru p se posuzuje podle jeho střední kvadratické chyby, která vyjadřuje stupeň koncentrace odhadu kolem odhadovaného parametru. Střední kvadratická chyba nestranného odhadu je totožná s rozptylem tohoto odhadu. Přirozeným cílem je konstruovat odhady, které mají střední kvadratickou chybu co nejmenší.

2.5 Testování hypotéz

Na základě zjištěných údajů je nutné zvolit právě jedno ze dvou možných, předem daných, navzájem se vylučujících rozhodnutí.

Statistická hypotéza

Nechť rozdělení pravděpodobnosti, z něhož je prováděn náhodný výběr X_1,-...,X_n, závisí na neznámem m-rozměrném parametru p ze známého parametrického prostoru O. Nechť w je daná podmnožina množiny O. Statickou hypotézou se nazývá tvrzení tvaru p náleží w a úkolem je rozhodnout o jeho platnosti na základě pozorovaných hodnot $x_1, \ldots, x_n$ uvažovaného náhodného výběru.

Často se pak píše H_0: p náleží w a H_1 : p náleží O-w, kde H_0 je tzv. *nulová hypotéza* a H_1 je tzv. *alternativní hypotéza*. Říkáme, že testujeme *nulovou hypotézu* H_0 proti *alternativní hypotéze* H_1. Konstrukce takového testu spočívá v nalezení vhodné podmnožiny W výběrového prostoru,jež je nazýván *kritickým oborem*. Platí pro pozorované hodnoty $(x_1, \ldots, x_n)$´náleží W, pak H_0 zamítneme ve prospěch H_1. V opačném případě H_0 nemůžeme zamítnout a dáme přednost H_0 před H_1 [10].

Samozřejmě zde existují rizika chybného rozhodnutí, např. H_0 zamítáme, i když je správné či naopak, zde je nutné zkoumat kritický obor.

3 Korelace a regrese

Vysvětlení pomocí matematických definic, vět a pouček upravených z čisté matematické formy slovním rozšířením.

3.1 Korelace

Objev korelace patří k milníkům rozvoje statistiky. Nejznámější a často skloňovaný je standardně klasický párový korelační koeficient, označující se jménem Pearsonovým podle Karla Pearsona, který rozšířil myšlenky dalšího významného matematika Galtona.

O rozvoj koncepce korelačního koeficientu se také zasloužil R. A. Fisher, který určil rozdělení výběrového korelačního koeficientu. Ten také považoval korelační koeficient za přelomovou míru asociace mezi dvěma proměnnými a zavedl jeho použití ve studijní genetice. I dnes se teorie rozvíjí, ale ne už tak razantně, dnes je spíše problém, kde teorii vypočítat. Řada prací stojí na korelačním koeficientu, neboť jej lze použít k vyjádření souvislosti mezi veličinami [7].

Definice 3.1 *Korelace je vyjádření vztahů mezi složkami* $E_1, \ldots, E_m$ m*-rozměrného náhodného vektoru* $\boldsymbol{E}$ *využitím podmíněné střední hodnoty, a určením intenzity vztahů s využitím korelačního koeficientu.*

Definice 3.2 *Druhý centrální moment náhodné veličiny* X *se nazývá* rozptyl (variance, disperze) náhodné veličiny X. *Obvykle se označuje*

$$\mathrm{var}(X) = \mathsf{E}[X - \mathsf{E}(X)]^2.$$

Druhá odmocnina z rozptylu $\sqrt{\mathrm{var}(X)}$ *se nazývá* směrodatná (standardní, střední kvadratická) odchylka náhodné veličiny X.

Definice 3.3 *Nechť* X, Y *mají konečné druhé momenty a nechť* $\mathrm{var}(X) \neq 0$, $\mathrm{var}(Y) \neq 0$. Korelační koeficient $\rho_{X,Y}$ náhodných veličin X, Y *je číslo definované vztahem*

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X) \cdot \mathrm{var}(Y)}} = \mathrm{cov}\left(\frac{X - \mathsf{E}(X)}{\sqrt{\mathrm{var}(X)}}, \frac{Y - \mathsf{E}(Y)}{\sqrt{\mathrm{var}(Y)}}\right).$$

Definice 3.4 *Mějme náhodný výběr*

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Potom výběrovou kovarianci definujeme následovně:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Definice 3.5 *Jsou-li výběrové rozptyly nenulové, potom jako výběrový korelační koeficient označujeme vzorec:*

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

I když jsou v praxi korelační úlohy méně časté, hlubším pohledem zjistíme, že problémy jako porovnávání analytických metod na odlišných vzorcích nebo hledání souvislostí mezi různými vlastnostmi či charakteristikami látek patří spíše mezi úlohy korelace než klasické regrese.

3.1.1 Korelační modely

Složky náhodných vektorů lze použít k charakteristice chování středních hodnot a rozptylů. Míra intenzity vztahu složek určuje kovarience či korelační koeficient. Korelační modely platí pro dvě i více náhodných veličin.

3.1.2 Korelační koeficienty

Korelační koeficienty slouží pro vyjádření těsnosti vazby mezi složkami náhodného vektoru **E**. K nejpoužívanějším patří: Párový korelační koeficient, Parciální korelační koeficient, Vícenásobný korelační koeficient.

3.2 Regresní analýzy

Hledání závislosti mezi proměnnými je důležité v mnoha oborech (např. výše investic do reklamy k celkové hodnotě tržeb podniku). Závisle proměnná (vysvětlovaná) je určena nezávisle proměnou (vysvětlující). Závisle proměnou je náhodná veličina, která má při dané hodnotě (náhodné) vysvětlující veličiny určité rozdělení pravděpodobnosti.

Regresní funkce. V praktických úlohách se vyšetřuje závislost různých proměnných. Příkladem může být spotřeba benzínu na průměrné rychlosti automobilu. Ve statistice je vhodné pohlížet na závisle proměnnou jako na náhodnou veličinu Y, kdy pak může tvrdit

$$Y = f(x_1 \ldots, x_r) + e.$$

Kde e je náhodná odchylka. Náhodná odchylka by měla kolísat kolem nulové hodnoty, tj. měla by mít nulovou střední hodnotu. Na základě pozorovaných hodnot $y_1, \ldots, y_n$ náhodné veličiny Y, které odpovídají hodnotám $(x_{11}, \ldots, x_{1r})$-,...,$(x_{1_n}, \ldots, x_{n_r})$ proměnných $x_1, \ldots, x_r$, je pak úkolem statistického rozboru specifikovat funkci f. Většinou jde o odhad parametrů, které tuto funkci jednoznačně určí. Odhadnutá funkce f umožní pro libovolné hodnoty proměnných $x_1, \ldots, x_r$ odhadnout odpovídající hodnotu proměnné Y.

V rámci matematické statistiky se f $(x_1, \ldots, x_r)$ nazývá *regresní funkce*, Y se nazývá *vysvětlovaná proměnná* a $x_1, \ldots, x_r$ se nazývají *vysvětlující proměnné (regresory).*

3.3 Lineární regresní modely

Cílem statistické analýzy je objasnit vztah mezi měřenou výstupní závisle proměnnou (vysvětlovanou) veličinou y a nastavovanými, vstupními nezávisle proměnnými (vysvětlujícími) veličinami x.

Typ závislosti, vyjádřený funkcí y = f(x, B), záleží na tom, jaké povahy jsou

veličiny y a x (obě náhodné, y je náhodná a x nenáhodná). Způsob odhadu parametru B uvažovaného modelu závisí na rozdělení náhodné veličiny y.

3.4 Metody nejmenších čtverců

Pro odhad parametrů regresní funkce se nejčastěji používá metoda nejmenších čtverců, která spočívá v minimalizaci výrazu přes všechny možné hodnoty parametrů určujících regresní funkci. Metodu nejmenších čtverců interpretujeme jako metodu umožňující optimálně proložit vhodnou křivku y = f(x_1,...,x_r) danými body [x_{11},...,x_{1r},y_1],...,[x_{n1},...,x_{nr},y_n] [8]. Maticový zápis MNČ:

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{e}$$

V hladině L leží všechny vektory **X B**. Parametry **B** lze chápat jako koeficienty úměrnosti u jednotlivých složek x_j souřadnicového systému, jejich lineární kombinace tvoří regresní model. Bez ohledu na užité kritérium regrese bude u lineárních regresních modelů ležet modelová funkce X b stejně jako teoretický model **X B** v m-rozměrné nadroviněL.

Metoda nejmenších čtverců (MNČ) je v regresní analýze nejužívanější. MNČ je metodou maximální věrohodnosti. Na základě geometrické lineární regrese lze hledat odhady parametrů b tak, aby byla minimalizována vzdálenost mezi vektorem y a nadhladinou L.

Pro dosazení za b a úpravě vyjde obecný a používaný tvar $\mathbf{X^T y} = \mathbf{X^T X b}$.

Odhad b, minimalizující vzdálenost má tvar $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$.

3.5 Lineární regrese na počítači

Určení odhadů b lineárního regresního modelu $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ může být jednoduchou úlohou, když jsou v knihovně programů k dispozici programy pro maticové operace s formálním řešením rovnice $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$. Problémy vznikají u matice $X^T X$. U polynomického modelu, vycházejí často odhady bez fyzikálního

smyslu. Regresní křivka sice prochází v těsné blízkosti experimentálních bodů, ale buď mezi nimi silně osciluju (u polynomů vysokých stupňů), nebo je systematicky posunutá.

Příčiny numerických potíží při počítačovém odhadu parametrů b jsou:

1. Zanedbání omezené přesnosti počítače při sestavování $\mathrm{X}^T\mathrm{X}$ matice.
2. Nevhodné postupy výpočtu inverzní matice $(X^TX)^{-1}$ řešení soustav lineárních rovnic.
3. Multikolinarita vedoucí ke špatné podmíněnosti matice $\mathrm{X}^T\mathrm{X}$.
4. Lineární závislost některých sloupců matice X^TX vedoucí k tomu, že k matici X^TX nelze najít inverzní matici z důvodů singularity.

Kvalitní programy lineární regrese překonávají tyto obtíže a poskytují řešení vždy. Mezi nejefektivnější patří algoritmy, které nesestavují matici X^TX, ale řeší předurčenou soustavu n lineárních rovnic o m neznámých $y = Xb$. Příkladem je algoritmus SVD (*singular value decomposition*), který pracuje i na počítači s malou přesností zobrazení dat.

3.6 Regresní diagnostika

Při vyhodnocení lineárních a nelineárních regresních modelů se tedy používá metody nejmenších čtverců. Tato metoda ještě nezajišťuje nalezení přijatelného modelu, a to jak ze statického, tak i z fyzikálního hlediska. Zdrojem problémů jsou složky tzv. *regresního tripletu* (data, model, model odhadu). Metoda nejmenších čtverců poskytuje optimální výsledky jenom při současném splnění předpokladů o datech a regresním modelu. Pokud tyto předpoklady nejsou splněny, je metoda nejmenších čtverců nevhodná. Regresní diagnostika obsahuje postup k identifikaci:

- kvality dat pro navržený model,

- kvality modelu pro daná data,
- splnění základních předpokladů metody nejmenších čtverců.

Do regresní diagnostiky lze zahrnovat:

1. metody pro průzkumovou analýzu jednotlivých proměnných,
2. metody pro analýzu vlivných bodů,
3. metody pro odhalení porušení předpokladů metody nejmenších čtverců.

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu a jsou přitom odhaleny typy odchylek od ideálního regresního tripletu. Počítač nám slouží jako nástroj analýzy dat, modelu a metody odhadu, ale uživatel ví o datech stejně více. Model je navrhován interaktivně spoluprací uživatele a programem na počítači. Tím by měl být omezen vznik formálních regresních modelů, které nemají smysl.

3.6.1 Využití průzkumové analýzy dat

Metody využíváme:

1. pro určení statistických zvláštností jednotlivých proměnných nebo reziduí,
2. k posouzení párových“ vztahů mezi všemi sledovanými proměnnými,
3. k ověření předpokladů o rozdělení proměnných nebo reziduí.

V řadě případů již pouhé vynesení naměřené veličiny y_i proti indexu i přirozeného pořadí naměřených dat může odhalit skrytou proměnnou, často související s časem nebo pořadím měření. K orientačnímu posouzení vztahů mezi jednotlivými proměnnými se užívá rozptylových grafů.

K ověřování normality dat se často používá Q-Q grafů. Základem průzkumové analýzy je stanovení volby rozsahu a rozmezí dat, jejich variabilitu a přítomnost vybočujících pozorování. K tomu lze využít grafů rozptýlení s kvantily.

Přes svou jednoduchost umožňuje průzkumová analýza identifikovat vlastní regresní analýzu:

1. nevhodnost dat (malé rozmezí nebo přítomnost vybočujících bodů),
2. nesprávnost navrženého modelu (skryté proměnné),
3. multikolienaritu,
4. nenormalitu v případě, kdy jsou vysvětlující proměnné náhodné veličiny.

3.6.2 Kvalita dat a kvalita modelu

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje především výskyt vlivných bodů (VB), které jsou hlavní zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů. Někdy VB zlepšují predikční schopnost modelů. Vlivné body ovlivňují většinu výsledků regrese a lze je rozdělit do tří základních skupin.

- **Hrubé chyby**, které jsou způsobeny měřenou veličinou – vybočující pozorování – nebo nevhodným nastavením vysvětlujících proměnných – extrémy. Jsou obyčejně důsledkem chyb při manipulaci s daty.
- **Body s vysokým vlivem** (tzv. golden points) jsou speciálně vybrané body, které byly přesně změřeny a které obvykle rozšiřují predikční schopnost modelu.
- **Zdánlivě vlivné body** vznikají jako důsledek nesprávně navrženého regresního modelu.

Podle toho, kde se vlivné body vyskytují, lze provést dělení na:

- **vybočující pozorování** (outliers), které se na ose y výrazně liší od ostatních, a
- **extrémy** (high leverage points), které se liší v hodnotách na ose x nebo v jejich kombinaci (v případě multikolinearity) od ostatních bodů.

Kvalitu regresního modelu lze posoudit v případě jedné vysvětlující proměnné x přímo z rozptylového grafu závislosti y na x. V případě více vysvětlujících proměnných je to složitější a k objasnění pomáhají parciální regresní grafy, parciální reziduální grafy.

3.7 Multikolinearita

Porušení metody nejmenších čtverců je často označována jako multikolinearita. Ale multikolinearita neznamená v pravém slova smyslu porušení MNČ, souvisí pouze s předpokladem o pozitivní definitivnosti matice $\mathrm{X}^T\mathrm{X}$, a tím jednoznačnosti rovnice

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\ \mathbf{X}^T\mathbf{y}.$$

Multikolinearita způsobuje špatnou podmíněnost matice X^TX. To má za následek že:

1. determinant matice $\mathrm{X}^T\mathrm{X}$ je číslo blízké nule,
2. některá vlastní čísla matice $\mathrm{X}^T\mathrm{X}$ jsou blízké nule.

Lze tedy očekávat řadu obtíží, když hledáme inverzní matici k matici X^TX, a projevů numerických chyb v závislosti na užitém algoritmu a strojové přesnosti počítače. Kromě numerických problémů způsobuje multikolinearita i řadu statistických obtíží jako nestabilitu odhadů, velké rozptyly či silnou korelovanou mezi prvky.

3.8 Kalibrace

Kalibrace je jedna ze základních úloh praxe, neboť chceme počítat s co nejpřesnějšími daty, abychom dostali i co nejpřesnější výsledek. Kalibrace slouží pro konstrukci snímačů fyzikálních veličin.
Fáze kalibrace:

1. sestavení kalibračního modelu,

2. použití kalibračního modelu.

Kalibrační model se sestavuje stejně jako regresní model. Úloha řeší proměnou odezvu y^* hledá odpovídající hodnota x^* a její statistické charakteristiky.
Druhy kalibrace a kalibrační modely:

- Absolutní kalibrace je nejvhodnější pro technickou praxi. Hledá se vztah mezi měřitelnou veličinou (signálem) a veličinou, která určuje stav nebo vlastnosti systému.

- Komparativní kalibrace znamená, že jeden přístroj kalibruje vůči druhému využitím standardu.

K přesnosti kalibračních metod se obyčejně definují limitní hodnoty a hraniční šum.

3.9 Vyšetřovaní a konstrukce regresního modelu

První fáze: Vyšetřování

Nejprve navrhneme jednoduchý model a předběžně analyzujeme data. Sledují se jednotlivé proměnné a možné párové vztahy. Užívá se proto rozptylových diagramů nebo indexovaných grafů.

Již v této fázi se posuzuje významnost proměnných s ohledem na jejich proměnlivost a přítomnost multikolinearity. Protože lineární vztahy mezi proměnnými v rozptylových grafech indikuje multikolinearitu a vlivné body.

Odhadování parametrů modelu se poté provádí metodou nejmenších čtverců (MNČ) a také se určí základní statistické charakteristiky. Následuje test významnosti jednotlivých parametrů.

Druhá fáze: Diagnostika a konstrukce

Regresní diagnostika využívá rozličných grafů k identifikaci vlivných bodů a ověřuje se samotný předpoklad MNČ. Na základě nalezených vlivných bodů se rozhoduje, zda je nutné tyto body z dat eliminovat, nebo zda je možné tyto body v datech ponechat. Pokud je nutné upravit data, provedeme znovu regresní diagnostiku, zkontrolujeme předpoklad MNČ a posuzujeme opět vliv multikolinearity.

Konstrukce zpřesněného modelu lze provádět s využitím různých metod MNČ (vážených, zobecněných, rozšiřujících atd.) Také je důležité zhodnotit dosavadní kvalitu modelu regresní diagnostikou. U kalibračních modelů se určí dolní hranice. Nakonec můžeme testovat více hypotéz tvorby modelu a porovnat je.

3.10 Logistická regrese

Pojem regrese poprvé použil Francis Galton při vyšetřování závislosti průměrné výšky potomků na výšce rodičů. U matky (dcery) zvětšil výšku o 8%, aby byla srovnatelná s výškou otce (syna). Pearson a Lee pak na základě nových dat došli k poznání, že pokud sledujeme závislost výšky syna na výšce otce, představme si dvě skupiny synů. Charakteristika první je, že otcové měří 170 cm, to je současně průměrnou výškou v populaci otců. Druhá skupina synů je charakterizována výškou jejich otců, která je rovna 180 cm, tedy o 10 cm více, než průměrná výška

všech otců v populaci. Průměrná výška synů z druhé skupiny je jen o 5 cm vyšší, než průměrná výška synů z první skupiny. Odchylka výšky synů sleduje odchylku výšky otců, ale nereprodukuje ji celou, ale krátí na polovinu. Jde o zpětný postup" k průměru.

Vlastní odhady parametrů regrese (lineárního modelu) jsou jen jednou dílčí úlohou, důležitější je odhad vektoru podmíněných středních hodnot závislé proměnné. Na tomto odhadu je založena téměř celá diagnostika.

Cílem bude vysvětlení variability (kolísání) náhodné veličiny Y (závisle proměnné, vysvětlované proměnné, odezvy) v závislosti na její střední hodnotě jedné nebo více náhodných nezávisle proměnných. Pokud by nezávisle proměnné byly náhodnými veličinami, pak se zajímáme o podmíněnou střední hodnotu Y při daných hodnotách $X = x$.

Logistická regrese pracuje s binárními veličinami, které se modelují jako nevýskyt (nula, 0) nebo výskyt (jednička, 1) sledovaného jevu. Stejně jako u lineárního modelu budeme vyjadřovat střední hodnotu vysvětlované závisle proměnné jako funkci nezávisle proměnných. Ovšem střední hodnota bude rovna pravděpodobnosti jedničky, nuly, tedy pravděpodobnosti výskytu, nevýskytu sledovaného jevu [14].

3.10.1 Tvar závislosti

Uvažujme nezávislé náhodné veličiny $Y_1,\ldots Y_n$ s alternativním rozdělením s parametry μ_i. Střední hodnoty μ_i jsou totožné s pravděpodobnostmi jedniček a mohou záviset na nějakých nenáhodných doprovodných veličinách x_i.

Je zřejmé, že platí var Y = μ_i (1 - μ_i), takže rozdíl Y_i závisí na střední hodnotě této veličiny. Toto je podstatný rozdíl v porovnání s normálním lineárním modelem, kde byl rozptyl konstantní.

Pokud bychom předpokládali, jako v lineárním modelu, že bude problém s interpretací, protože nelze zaručit, že pro libovolné x_i. bude μ_i ležet v intervalu

(0,1). Nutné je najít jiný interpretovatelný tvar závislosti, nejlépe v odhadech metodách maximální věrohodnosti.

Pravděpodobnosti dvou možných hodnot $Y_i = 1$ a $Y_i = 0$ lze psát jako $\mathrm{P}(Y_i = \mathrm{j}) = \mu_i^j(1 - \mu_i)1 - \mathrm{j}$, $\mathrm{j} = 0{,}1$.

Logaritmockou věrohodnostní funkci lze tedy zapsat

$$\sum_{i=1}^{n} Y_i \log(\frac{\mu_i}{1 - \mu i}) + \sum_{i=1}^{n} \log(1 - \mu_i)$$

náhodné veličiny se v logaritmické věrohodnostní funkci projevují pouze v součinech s výrazy log $(\mu_{ii})/(\mu_{ii})$. Podíl

$$w(x_{i.}) = \frac{\mu_i}{(1 - \mu_i)} = \frac{Px_{i.}(Y_i = 1)}{Px_{i.}(Y_i = 0)}$$

porovnává pravděpodobnost výskyt sledovaného jevu a nevýskyt sledovaného jevu. Pro tento podíl se v anglickém jazyce užívá označní *odds* (šance). Samotné funkci

$$\eta_(\mu) = \log \frac{\mu}{(1 - \mu)}$$

se říká *logit*. V kontextu zobecněných lineárních modelů (generalized linear model - GML) je *logit* speciálním případem spojovací funkce (link function).

Předpokládejme, že *logit* pravděpodobnosti je lineární funkcí neznámých parametrů

$$\eta_i(ß) = ß'x_{i.}.$$

Někdy se v obecné zápisu explicitně uvádí absolutní člen, protože, ne vždy jej budeme schopni odhadnout. Pak bychom místo regresní matice X psali matici (1, X).

Střední hodnotu $\mathrm{E}Y_i$ vyjádříme jako

$$\mu_i(ß) = exp\frac{(\eta_i(ß))}{1 + exp(\eta_i(ß))}$$

$$= exp\frac{(ß'x_{i.})}{(1 + exp(ß'x_{i.})}$$
$$= \frac{1}{1 + exp(-(ß'x_{i.})),}$$

platí $0 < \mu_i < 1$ a odstraní se jeden ze zmíněných problémů.

3.10.2 Odhad parametrů

Odhad parametrů metodou maximální věrohodnosti. Platí

$$\frac{\partial}{\partial \eta_i} log(1 - \mu_i) = -\frac{\partial}{\partial \eta_i} \log 1 + e^{\eta_i} = -\mu_i$$

a logaritmickou věrohodnostní funkci ve tvaru

$$l(ß) = \sum_{i=1}^{n} Y_i \eta_i(ß) + \sum_{i=1}^{n} \log(1 - \mu_i(ß)).$$

3.10.3 Interpretace parametrů

Prametry $ß_0$ a $ß_1$ v nejjednoduším modelu η_i =$ß_0$+$ß_1$*x_i, podobně jako u lineární regrese. Předpokládejme nezávisle proměnnou danou dvouhodnotovým faktorem. Potom nula-jedničková proměnná x vyjadřuje nepřítomnost či přítomnost jevu.

Pro x = 0 jsou šance rovny

$$w(0) = \frac{P(Y = 1)}{P(Y = 0)} = \frac{\frac{exp(ß_0)}{1+exp(ß_0)}}{\frac{1}{1+exp(ß_0)}} = e^{ß_0}$$

Parametr $ß_0$ je roven *logitu* pravděpodobnosti výskytu sledovaného jevu pro x = 0

$$ß_0 = \log \frac{P(Y = 1)}{P(Y = 0)}$$

.

Pro x = 1 je šance rovna

$$w(1) = \frac{\frac{exp(ß_0+ß_1)}{1+exp(ß_0+ß_1)}}{\frac{1}{1+exp(ß_0+ß_1)}} = e^{ß_0+ß_1}$$

.

Poměr *šancí (odds ratio)* pro dvě hodnoty x je dán takto

$$\frac{w(1)}{w(0)} = \frac{exp(ß_0 + ß_1)}{exp(ß_0)} = e^{ß_1}$$

,

takže parametr $ß_1$ je roven logaritmu poměru *šancí*. Pokud pravděpodobnost sledovaného jevu na hodnotě x nezávisí, je poměr *šancí* roven jedné, tedy $ß_1 = 0$.

Existují i studie, kdy u zvláštních modelů logistické regrese, jde odhadnout vektor **ß**, aniž by byl odhadnut parametr $ß_0$. Jako je například *průřezová studie* (cross-sectional), *prospektivní studie*(cohort) a *retrospektivní studie*(case-control).

4 Praktická část, porovnání softwarů

V praktické části bude na datasetech vysvětlena manipulace s daty v prostředí R softwaru, jeho celkové představení, příkazy k výpočtům logistické regrese a ukázka kódové struktury. Popsání jiných matematických, geoinformatických softwarů se zaměřením na ArcGIS 9.x., SAS Systém a STATA.

4.1 R software

4.1.1 Základní informace

R je open-source matematický software specializovaný na statistiku a grafické výstupy. Jako mnoho dalších statistických programů používá programovací jazyk podobný S. R funguje na široké škále UNIX platforem (včetně FreeBSDa Linux), Windows a MacOS a je možné analyzovat data z veškerých lidských oblastí (zemědělství, astrofyzika, klimatologie, ekonomika, geografie a mnoho dalších) [28].

Ke stažení je zdarma jako Free Software Foundation´s GNU General Public licence na stránkách samotné organizace http://www.r-project.org/, CRAN Mirrors, zde jsou také návody, informace o R, volně stažitelné balíky (Packages), které pomáhají zlepšit práci a ostatní příslušná dokumentace.

R software má i českou lokalizaci http://www.r-project.cz/, ale informace byly čerpány z anglické verze.

R je jazyk a prostředí pro výpočet a grafické znázornění statistických úloh. R může být považováno za jiné provedení programovacího jazyka S, který byl vyvinut Johnem Chambersem a kolegy v laboratořích Bell Laboratories (Lucent Technologies). Mezi R a S jazykem nalezneme rozdíly, ale mnoho kódů napsaných pro S běží nezměněně pod jazykem R. Jazyk S je většinou využíván ve výzkumu statistických metodik, kdežto jazyk R nabízí zapojení se i do samotného vytváření lineárních a nelineárních modelů po klasifikaci, clustering či jen můžeme provádět klasické statistické testování to vše lze zakončit grafickými výstupy. R software

je také schopný dalšího rozšíření.

Nynější R je výsledkem společného úsilí lidí z celého světa. Robert Gentleman a Ross Ihaka započali se psaním softwaru s názvem "R and R" (1993), pod záštitou Statistického oddělení university Auckland (the Statistics Department of the University of Auckland). Ale od roku 1997 se skupina rozrostla o dalších 19 členů a poději o další dva. Například Simon Davies z téže university napsal kód pro funkci glm(), kterou budu využívat.

R by nemohlo být čím je dnes bez neocenitelné pomoci lidí, kteří přispívají darováním jednotlivých kódů a úpravou bagů (chyb).

Dobře navrženými výstupy jsou bezesporu silnou stránkou R softwaru. Dále zahrnuje matematické symboly a rovnice, to vše plně kontrolovatelné uživatelem.

4.1.2 Prostření R softwaru

Prostředí R je charakteristické plně plánovaným a logickým systémem, jeho vybavení zahrnuje manipulaci s daty, výpočty a grafická zobrazení jako

- efektivní zacházení s daty a jejich uskladnění,
- příhodné pro výpočetní operace v sadách, jednotlivých maticích,
- pomocné nástroje pro datové analýzy,
- grafické zobrazení analýz na obrazovku, do kopie,
- rozvinuté, jednoduché a efektivní programování s podmínkami, cykly, uživatel může nadefinovat funkce a použít vstupy či výstupy,
- uživateli je dovoleno přidat další funkce jejich přímým definováním.

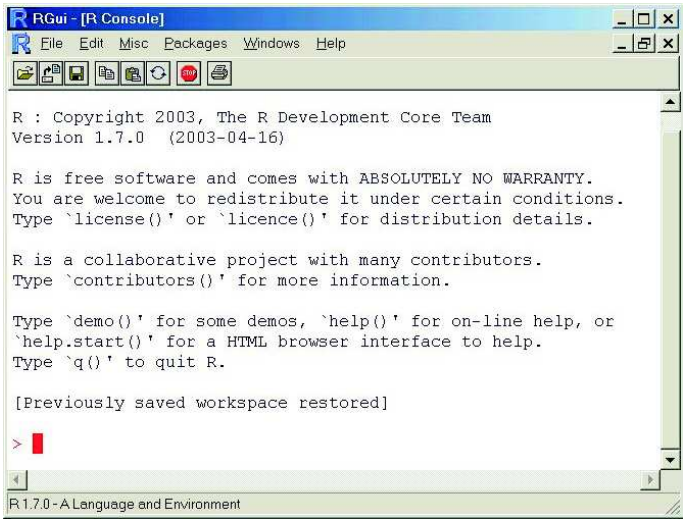
Mnoho ze systému je napsáno v R dialektu S jazyka, což usnadňuje uživatelům následovat algoritmické volby pro výpočetně silné jazyky jako C, C++ či Fortran. Pokročilý uživatel smí napsat C kód pro manipulaci s R objekty přímo.

Mnoho uživatelů smýšlí o R jako o statistickém systému, ale raději o něm uvažujme jako o prostředí, ve kterém jsou statistické techniky implementovány. Pokud chceme rozšířit R, lze stáhnout tzv. packages (balíčky). Základní jsou již spuštěny po instalaci R a mnoho dalších lze najít a získat pomocí CRAN rodiny na stejné adrese stažení softwaru.

LaTeX formát je podporou komplexní dokumentace jako u on-line podoby či tiskové kopie.

Ačkoli je R většinou užíván statistiky a lidmi, kteří se statistikou přímo zabývají, R pracuje jako obecná maticová výpočetní nástrojová sada v měřítku GNU Octave nebo MATLAB. Rozhraní R bylo přidáno do populárního data mining“ softwaru Weka, který spojí schopnosti Weka a statistických analýz v R.

Při zapnutí R softwaru je zobrazeno okno s informacemi o verzi, licenci a pomoci (help). Zde také započneme s psaním našeho kódu, jak napovídá červený kurzor.



Obr. 1. *Rozhraní R softwaru.*

Základna R (R Foundation) softwaru rozvíjí

- poskytování podpory R projektu a dalších inovací v statistických operacích, které se stále zdokonalují,

- společnost pro vývoj R spolupracuje s jednotlivci i institucemi za účelem co nejrychlejšího zkvalitnění systému.

Základna R má stejný cíl jako jiné open source softwary jako Apache Foundation nebo GNOME Foundation a to vyvíjet se ve prospěch uživatelů.

Fyzická základna R sídlí v rakouské Vídni pod správou Technologické University ve Vídni (Vienna University of Technology). Pokud se budete chtít stát běžnými členy R Foundation, musíte být většinově zvoleni jejich radou. Každý se ovšem může stát tzv. suporting (podporujícím) členem po zaplacení poplatku. Nebo vkládat své návrhy, připomínky a chyby přímo na jejich oficiální stránku.

V práci, byl využit R 2.10.1 pro Windows (32 megabytes), ale již 22. dubna vyšla verze novější R 2.11.0.

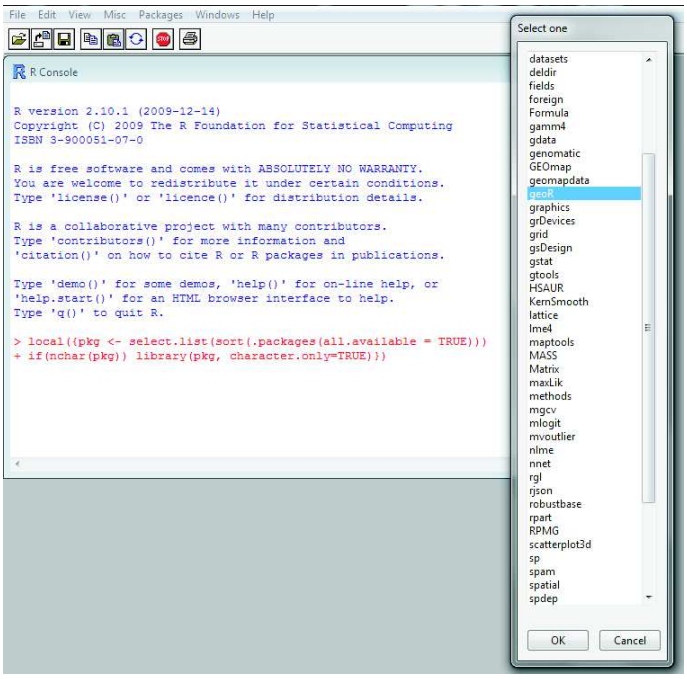
4.1.3 R Packages (rozšiřující balíčky)

Schopnosti R softwaru se mohou rozšířit skrze packages (rozšiřující balíčky), které dovolují specializovat statistické techniky, grafická zařízení či import/export (vstup/výstup) velkého množství datových formátů.

Packages vyvinuli v R, LaTeX, Java, C a Fortan. Základní packages zahrnuje prvotní instalace, jak již bylo řečeno. Nalezneme více než 2 000 dostupných rozšiřujících balíků v kompletní síti R archivu CRAN (Comprehensive R Archive Network). Oficiální shrnutí veškerých R balíků (packages) naleznete http://cran.r-project.org/web/packages/.

Crantastic je komunitní stana pro hodnocení a kontrolování CRAN packages. R-Forge nabízí centrální platformu pro rozvoj R packages. Je zde mnoho nezveřejněných, beta (zkušebních) balíků a vývojových verzí CRAN packages.

Projekt Bioconductor připravuje R balíky pro analýzu genomových datových sad jako Affymetrix a cDNA.



Obr. 2. *Instalace packages.*

4.2 ArcGIS Desktop 9.x

ArGIS je profesionální geografický informační systém (GIS software) spadající do rodiny ESRI produktů. Usnadňuje prohlížení, tvorbu a správu prostorových dat. Operační systém je koncipován pro Windows. Desktop ArcGIS může zahrnovat ArcReader, ArcView (tvorba map a analýzy) a jeho rozšíření v podobě ArcEditor. Ostatní ESRI produkty jsou například MapObjects (mapové objekty), programovací knihovny pro vývojáře a ArcSDE (relační řízení databáze).

ArcGIS 9.x obsahuje prostředí pro geoprocesing (zpracování prostorových dat), vizuální programování (Phyton, VB Script, Perl). Pracuje s formátem shapefile (.shp), geodatabáze ukládá data pomocí objektivně relační databáze i s topologickými informacemi. Třídy prvků jsou body linie a polygony, jedna třída prvků je uložena samostatně ve vrstvě, tedy nemohou být body a polygony v jedné vrstvě. Geodatabáze ukládá i vícenásobné třídy prvků třemi možnými způsoby (.gdb), a to souborová (file) geodatabáze, osobní (personal) geodatabáze a

relační ArcSDE geodatabáze.

ArcGIS Desktop obsahuje integrované aplikace a to ArcMap, ArcCatalog, ArcToolbox a ArcGlobe. Přes aplikaci ArcToolbox je přístupná funkcionalita programu sady nástrojů jsou však dostupné v různých extenzích. K prohlížení dat (atributy, geometrie) a tvorbě výstupů je dostatečné použití aplikace ArcMap. Aplikace ArcCatalog spravuje data, mohu je tvořit geodatabáze, spravovat a editovat včetně jejich metadat [18].

Součástí ArcGIS je i funkční část programu umožňující automatizaci práce a tvorbu nástrojů pro manipulování s daty a to je Geoprocesing.

Pracovní prostředí pro prostorovou analýzu dat:

- Dialog – průvodce, formulář, kde se definují vstupní data a parametry, použití po spuštění geoprocesingu uvnitř aplikace
- Model – interaktivní vizuální model, spojuje procesy, data a parametry dohromady, okno ModelBuilder, bez nutnosti psaní skriptů
- Příkazový řádek – menu příkazů a funkcí
- Skript – napsaný soubor pomocí skriptovacího jazyka (Phyton. JScript, VB Script), automatizace opakujících se úloh, pro pokročilé uživatele ArcGIS Desktop
- ArcObjects – objekty geoprocesingu, vytváření uživatelských nástrojů, Visual Basic 6, C++, tvorba uživatelských dynamických knihoven

ArcGIS Desktop 9.3 je přístupný ve třech licenčních úrovních, lišících se funkcionalitou:

Základní: ArcView

prohlížení dat, editace, analýzy nad daty a tvorbu grafických výstupů (map). Dovoluje také pracovat s daty uloženými v systému relačních databází přístupných přes ArcSDE.

Rozšířené: ArcEditor

zvyšuje funkcionalitu ArcView možnostmi práce s geodatabázemi, včetně tvorby, editace a kompilace.

Nejvyšší: ArcInfo

rozšíření možností pro prostorové operace, tvorbu map a navýšení funkcionality Geoprocessingu.

4.2.1 Geostatistical Analyst

K programům skupiny ArcGIS 9.x byla přidána extenze Geostatistical Analyst, přístupná přes menu Tools. Určená pro práci vyšetřování statických dat, identifikaci nepřesností a jiných zvláštností dat. Díky tomu je možné předpovědět a vyhodnotit přesnost předpovědí a tvořit z nich výsledné povrchy.

Geostatistické modelování tvoří povrchy ze zpracovaných dat. Procedura, která se k modelaci využívá k odhadu neznámé hodnoty ze známých okolních hodnot, nazýváme interpolací či interpolační metodou. Pro nás je nejdůležitější, zda interpolační metody zachovávají původní hodnoty vstupních bodů či je jim přiřazena hodnota nová. Dělíme, podle tohoto tvrzení, interpolační metody na exaktní (zachovávají vstupní hodnoty) a aproximační (dochází k vyhlazení hodnot i v místech se známou hodnotou).

Geostatistical Analyst v ArcInfo však používá odlišné rozdělovací kritérium než exaktní a aproximační podle zohledňování náhodnostních prvků. Deterministické metody jsou označením pro metody, které nepracují s prvkem náhodnosti. U geostatických (stochastických) metod je prvek náhodnosti připuštěn. Metody dostupné v extenzi Geostatistical Analyst stojí na tvrzení, že jevy prostorově bližší jsou si více podobné než jevy prostorově vzdálenější.

Síla Geostatistical Analyst je tedy v samotných nástrojích pro tvorbu spojitého povrchu z naměřených hodnot rozptýlených bodů. Pomůže odhadovat hodnoty povrchu použitím interpolace kriging. Nástroje pro analýzu prostorových dat

umožňují pochopit podstatu dat (rozložení, globální a lokální odchylky, globální trendy, úroveň prostorové autokorelace). Předpovědi obsahují i stupeň nejistoty.

Práce s geostatistical analyst je intuitivní, stačí jen přidat panel Geostatistical Analyst do nástrojové lišty programu poklikem“ pravého tlačítka na volnou plochu lišty a odškrtnou ve výběru. Pokud jsou funkce menu zašedlé, spusťte licenci pro Geostatistical Analyst.

Menu nabízí tři hlavní komponenty extenze – Expore Data (nástroje pro analýzu dat), Geostatistical Wizard (interpolace dat) a Create Subsets (tvorba pracovního a kontrolního datasetu).

4.3 SAS System

SAS System (Statistical Analysis System) je integovaný softwarový systém vyráběný firmou SAS Institute Inc. Slouží běžně ve firmách pro databázové účely, ale je také nástroj pro analýzu a obchodní využití dat. Využívaný pro statistickou analýzu dat také pro vědecké účely.
Vytvořen Anthony J. Barr (1966) na universitě Severní Karolína, SAS je tvořen z mnoha komponentů jako SAS Add-In pro Microsoft Office, Base SAS, Data Integration Studio, SAS Web OLAP Viewer for Java, SAS/GIS, SAS/STAT a mnoho dalších. Jedná se o modulární software, takže si lze pronajmout jen potřebné části, využitelného konkrétní účely. SAS obsahuje vlastní programovací jazyk, označovaný rovněž jako SAS, ale i jazyk C. Nezávislý na platformě [29].

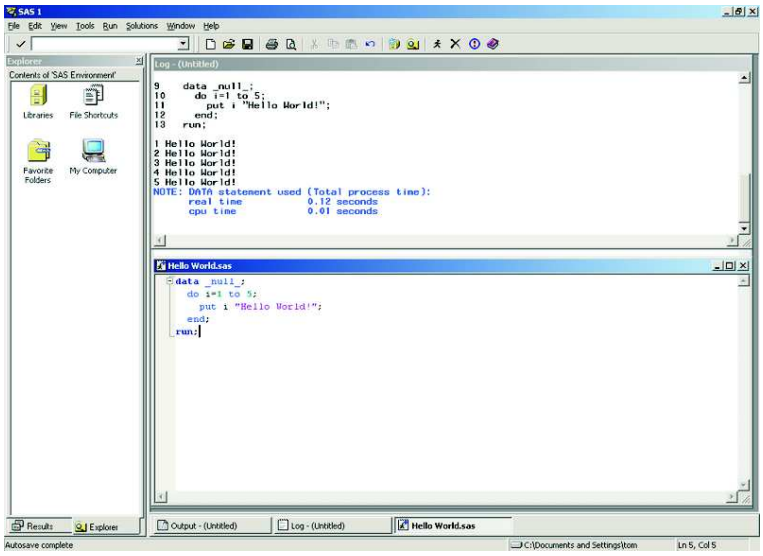
Z nejdůležitějších věcí umožňuje:

- Vstup a správu dat.
- Statistické analýzy.
- Obchodní plánování, předpovídání a podporu.
- Skladování dat.

A zahrnuje:

- Data
- Procedury
- Macro jazyky

Místo tabulek (tables), řídků (rows) a sloupců (columns), SAS užívá termíny datasetu (datasets), pozorování (observation) a proměnné (variable), které jsou numerické nebo zakové (string).



Obr. 3. *Rozhraní programu SAS [16].*

SAS je velmi podobný prostředím R softwaru, samozřejmě, že se liší příkazy a vnitřní výstavbou, ale zde je pár příkazů [34], využitelných pro logistickou regresi.

načti data a proveď

data hsb2;

set hsb2;

hiwrite = write = 52;

run;

proc means data = hsb2 mean std;

run;

použij logistickou logisticou funkci

roc logistic data = hsb2;

smodel hiwrite (event='1') = female ;

ods output ParameterEstimates = model_female;

run;

vypočtené parametry dat,estimate =odhad, Pr min ChiSq pravděpodobnost Chi-testu

Parameter (FEMALE) DF(1) Estimate(0.9928)Error(0.301) ChiSquare(10.8369) Pr min ChiSq (0.0010) ##datový model s exp (exponentem) odhadů

data model_fem;

set model_female;

o = exp(estimate);

run;

proc print data = model_fem;

var variable estimate o;

run;

logaritmust log, a funkce logit pro modelu

proc logistic data = hsb2;

model hiwrite (event='1') = female math;

output out = m2 p = prob xbeta = logit;

run;

vytvoření polotového grafu

symbol1 i = join v=star l=32 c = black;

symbol2 i = join v=circle l = 1 c=black;

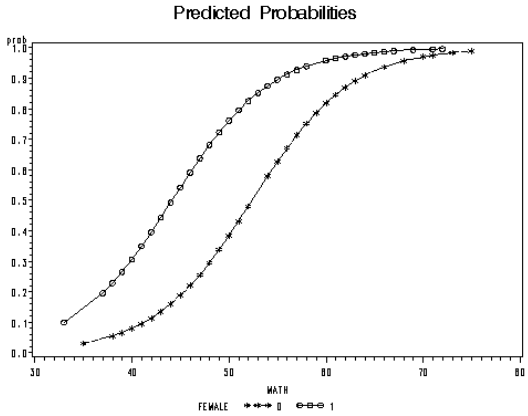
proc gplot data = m2;

plot logit*math = female;

plot prob*math = female;

run;

quit;



Obr. 4. *Pravděpodobnostní předpověď.*

Bohužel SAS neumí tak dobré grafické výstupy jako R software. Pro dopočet úloh je velice vhodný ArcGIS 9.x.

4.4 STATA Software

STATA vznikl složeninou dvou slov statistika (statistic) a slova data, v plném znění: Data Analysis and Satistical Software, tedy lze provádět analýzy nad daty pomocí statistických pravidel. STATA je softwarový balík pro statistické výpočty, autor Bill Gold a zdokonalený firmou StataCorp (1985). Jako SAS Systém je částečně psán v C jazyku, je multiplatformní a s propriety licencí. Najdeme jej v obchodních i akademických institucích zabývajících se jak výzkumem v oblasti ekonomie, ale i v sociologii a jiných vědních oborech [31].

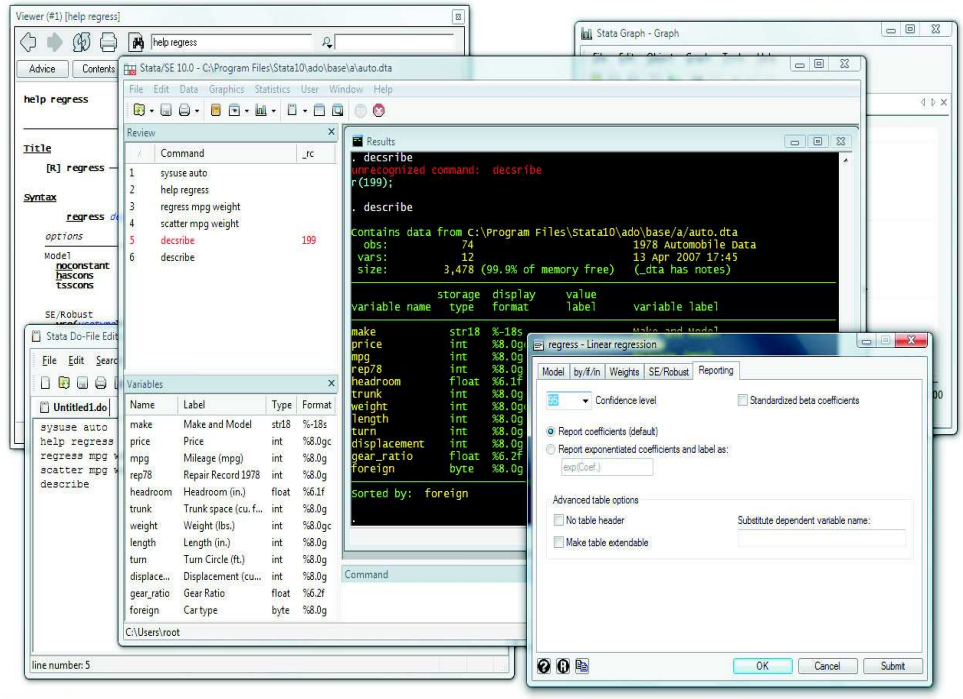
Umožňuje:

- Spravovat data
- Provádět statistické analýzy
- Grafy
- Simulace

- Uživatelské programování

Verze STATA:

- STATA/MP, pro multiprocesní počítače
- STATA/SE, pro velké databáze
- STATA/IC, standardní verze
- Small STATA, studentská verze



Obr. 5. Rozhraní *programu STATA [27].*

STATA se vyznačuje množstvím dialogových boxů pro tvorbu příkazů a flexibilitu skriptovacích příkazů. Vždy pro jeden dataset, všechny datasety potom shromažďuje v paměti. Přičemž, proměnné mají stejná čísla jako pozorování. Rozšíření STATA také funguje pomocí balíků (packages) volně stažitelných na internetu a jednoduchým přidáním do knihoven.

Ukázka psaní kódu pro STATA [33].

načti data, zde pomocí webu

use http://www.ats.ucla.edu/stat/stata/webbooks/logistic-/apilog, clear

použij regresi

regress hiqual avg_ed

předpovídej y a ukaž v grafu **predict yhat**

(option xb assumed; fitted values)

(42 missing values generated)

twoway scatter yhat hiqual avg_ed, connect(l .) symbol(i O)-sort ylabel(01)

5 Data a příkazy

Charakteristika vybraného území, oblast výskytu častých polomů. Hodnocení rizik a volba vhodné interpolační metody. Výpočet pravděpodobnosti výskytu polomů v R softwaru s vysvětlením kódu.

5.1 Význam polomů

Rostoucí povědomí veřejnosti o problematice rizik a problémů lesních polomů v posledních desetiletích vzrostl, což je také jeden z důkazů projevu péče o samotné zkvalitnění života. Setkáváme se s výskytem větrných, sněhových a nárazových polomů. Lesní polomy, čili vyvrácení nebo polámání stromů bouřlivými větry, patří k nejběžnějším a zcela přirozeným událostem, které souvisejí s existencí a vývojem lesních porostů.

Poškození způsobené větrem za spolu účasti dalších faktorů a následná těžba zasaženého dřeva má za následek přímé peněžní ztráty, ale musíme počítat i se zasažením a změnou krajiny a jejího např. hydrického režimu, zvýšené eroze, ztráta původních biotopů a další nepřímé faktory, které druhotně také způsobují peněžní i ekologické ztráty. Lesy jsou tedy důležitou součástí naší ekonomiky. Lidé se přirozeně snaží dopad polomů zmírnit a předpokládat metodami matematické statistiky.

Všestranný geografický pohled na problematiku lesních polomů a analýza jejich rozmístění pomocí geoinformačních technologií (GIS), nástrojů pro podporu prostorového rozhodování (SDSS) a prediktivních modelů jsou možností, jak identifikovat a pochopit řadu zákonitostí dopadů větrných bouří na lesy. Důležité je efektivně hodnotit lesní území v krátkém čase a přijímat vhodná opatření ke snížení negativních dopadů na lesní porosty.

Orkán Kyrill, který přivanul do České republiky v noci z 18. na 19. ledna 2007, postihl velké části Šumavy. Po obou stranách státní hranice úplně nebo částečně poškodil desítky km^2 lesních porostů ve všech zónách národních parků

Šumava a Bayerischer Wald. Pro analýzu působení orkánu Kyrill bylo vybráno testovací území, s rozlohou kolem 70 km^2, neboť je typickým územím Šumavy postihovaným polomy. V průběhu vývoje lesa docházelo k častým vichřicovým kalamitám, ke kterým se přičítalo přemnožení kůrovců. Les byl obnovován výhradně výsadbou smrků s mělkým kořenový systém, náchylnost na vývrat je u nich větší.

Na Šumavě se vyskytují větry s nárazovou rychlostí kolem 110 km/h až pětkrát do roka. Pohroma orkánu Kyrill (až 170 km/h) byla z těch největších, rozsah škody čítá okolo 840 000 m^3 padlého dřeva. Vítr je tak velmi významným fenoménem ovlivňujícím přírodu a ekonomiku Šumavy.

Orkán Kyrill zasáhl různou měrou téměř celou oblast Šumavy, testovací území bylo zvoleno v omezeném rozsahu. Pro území bylo postupně shromažďována nezbytná podkladová materiálová základna a geoinformace z veřejně dostupných zdrojů, z databází Správy národního parku a CHKO Šumava či MŽP ČR.

Území s konkrétními přírodními (lesohospodářskými) parametry jsou více náchylnější k výskytu polomů. Výskyt polomů, množství způsobených škod a náklady na jejich zmírnění tak podléhají jistému zákonitému spolupůsobení přírodních faktorů území a předchozích lesohospodářských zásahů do krajiny. Parametry jako vzduch, voda, energie horniny a zeminy, reliéfu, půdy a vlastní porost se nepochybně podílejí na místu a způsobu poškození.

Řadu parametrů lze zjistit přímým naměřením, další důležité podklady chybí (skutečná rychlost a směr větru v daném momentu, tehdejší vlhkostní poměry atd.). Navíc údaje o některých parametrech jednotlivých přírodních složek nutno získávat odvozováním od známých, avšak zobecněných charakteristik stanovišť. V tomto případě lze použít lesnických typologických map nebo derivovat nové hodnoty z digitálního modelu reliéfu.

Každý hodnocený faktor musel být ještě opatřen dodatečnou vahou podle toho, jak výrazné byly rozdíly v četnostech (velikosti) ploch s konkrétními hodnotami daného faktoru. Došlo k nově vymezeným kategoriím rizikovosti ploch v

území a zjištěna míra poškození.

Geografická analýza potvrdila, že z rozložení polomů vyplývá, že jejich vznik je ovlivněn podmínkami stanovišť (nadmořská výška, zakmenění, vlhkost půdy) a jejich biotickými vlastnostmi (věk, hloubka kořenového systému). Vlastní chování větru také hraje velkou roli v působení škod. Popis a prognóza zůstává dlouhodobým úkolem aplikované vědy.

5.2 Vymezení zájmového území

Z důvodů ochrany citlivých dat nebude tato kapitola veřejně poskytnuta.

5.2.1 Mechanický přístup

Mechanický přístup předpovídá pravděpodobnost poškozeného lesního porostu polomem na základě kritické rychlosti větru, která zapříčiní vývrat či zlomení stromu. Pravděpodobnost výskytu větrů v dané lokalitě [24].

Modely stojí na základě údajů a informací o mechanických vlastnostech stromů, které byly kalibrovány pomocí terénního průzkumu. Charakterizují fyzikální děje, které mají za následek vyvrácení stromů až zlomení kmene stromu.

Výpočet pravděpodobnosti rizika ohrožení porostu polomem je v mechanických modelech rozdělen do dvou fází.

- Fáze první počítá tzv. **kritickou rychlost větru** (critical wind speed), která vyvrací či láme stromy. Síla je určena podle faktorů, jako je místní rychlost větru, nárazový vítr, pozice stromu v lesním pokryvu, vlastnosti koruny (například velikost, aerodynamika, hmota), vlastnosti kmene (tvar, délka, hmota). Odporové (bránící) síly stromu závisí na faktorech, jako jsou vlastnosti kmene (průměr a pevnost dřeva), morfologie kořenů.

- Fáze druhá počítá **pravděpodobnost výskytu větrů**, které překročí kritickou rychlost větru.

5.2.2 Přístup založený na pozorování

V terénu jsou pomocí pozorování zaznamenávány výskyty faktorů, jež mají prokazatelný vliv na zvýšení větrných škod způsobených větrem v terénu. Pozorované charakteristiky asymetrie, chůdové kořeny, tvary korun a nahnilé kořeny. Kombinace faktorů vede k odhadu lokalit, které mohou být v budoucnu postiženy polomem.

Velké množství empirických a mechanických modelů a metod mají za účel sloužit k odhadnutí rizika výskytu lesních polomů. Nejpoužívanější metoda k hodnocení pravděpodobnosti výskytu polomů je tzv. *klasifikace rizika polomů*, kterou zavedl v roce 1985 K. F. Miller.

Pro hodnocení rizika lesních porostů jsou nyní nejvíce rozšířeny hybridní mechanické modely kombinující mechanický a empirický přístup. Jako model ForestGALES [20] a model HWIND [26].

5.3 Přístup založený na pozorování

5.3.1 Metody deterministické

Nejpoužívanější modelace pro náhodné prostorové procesy. Jsou založeny na vstupním měřítku nebo na stupni zarovnání povrchu (Smoothing u ArcGIS Desktop 9.x).

- **Lokální metoda** předpovídání hodnot, nejbližší okolní body k předpovídané lokaci. Mezi lokální metody přístupné v Geostatistical Analyst patří IDW (Inverse Distance Weight, metoda inverzních vzdáleností), lokální polynomická interpolace (pro celý povrch s jednotným trendem) a RBF (Radial Basis Functions, radiální funkce pro menší zakřivení povrchu).
- **Globální metoda** předpovídání hodnot, z celého vstupního datasetu, v Geostatistical Analyst je přístupná globální polynomická interpolace [12].

5.3.2 Metody geostatistické

Předpokládá, že rozdíly sledovaného jevu mohou být namodelovány náhodnými procesy s prostorovou autokorelací. Počítají se statistickým modelem, který obsahuje pravděpodobnost, na rozdíl od metod deterministických. Výsledkem je předpověď jevu a pravděpodobnost s jakou se předpovídá, př. výskyt polomů.

Kriging, základní geostatistická metoda pro předpověď hodnoty sledovaného jevu v určitém místě. Významnou roli hraje předpoklad autokorelace (korelace proměn ných závisí na vzdálenosti a/nebo směru mezi nimi).

Měřeným hodnotám, které jsou nejblíže k počítané hodnotě, je přidělena nejvyšší důležitost. Oproti IDW, která počítá pomocí jednoduchého vzdálenost-

ního algoritmu hodnoty pro kriging jsou počítány pomocí variogramu, semivariogramu, stanoveného na základě znalosti vstupních dat.

Typy krigingu Ordinary Kriging, Simple Kriging, Universal Krigin, Indicator Kriging, Probability Kriging, Disjunctive Kriging a Cokriging.

Nejjednodušší a nejčastěji používané metody krigingu jsou následující: Ordinary Kringing (průměrná hodnota sledovaného jevu v poli) a Simple Kriging (konstantní aritmetický průměr).

5.4 Volba vhodné interpolační metody

Povaha zkoumaného jevu je úzce spjata s výběrem interpolační metody. Přírodní procesy je vhodné modelovat pomocí metod deterministických (například oslunění) a sociální jevy pomocí komplexnějších metod geostatistických, jako u předpovědi výskytu geologických jevů, kdy se počítá prvek náhodnosti.

Exaktní a aproximační metody mají stejný princip s důrazem na to, zda se jev v území mění spojitě (teplota) nebo) skokově (geologické jevy) a zda-li ve výsledku připouštíme hodnoty vyšší či nižší než u vstupních dat.

Porovnáním výsledných hodnot s hodnotami naměřenými v terénu zjistíme přesnost výsledného modelovaného povrchu.

Při rozdělení originálního datasetu na dvě části, lze jednu část využít pro zjištění struktury data tvorbu povrchu a duhou lze využít k ověření přesnosti výsledků. Create Subsets dialog box (ArcGIS Desktop 9.x) umožňuje vytvoření jak zjišťovacího, tak i testovacího datasetů [2].

5.5 Interpolace dat

Důležité je v první řadě pochopit strukturu dat, která zajistí, že vybraná interpolační metoda bude správná. Zkoumání dat na anomálie a test na normálnost rozdělení zjistí struktury dat. Například Geostatistical Analyst (ArcGIS) umí nastavit prohledávanou oblast, pokud se jedná o lokální interpolaci, a prohledat

oblast (Searching Neighborhood) tvaru kruhu, elipsy s parametry tvaru, orientací a rozdělit vše na sektory s uspořádáním. Možné je výběr minimálního/maximálního počtu bodů, který výpočet zahrnuje. A brát v úvahu prostorovou autokorelaci, ke které se vztahuje strukturální funkce a graf semivariogram, poukazující na hodnoty sledované proměnné v závislosti na vzdálenosti.

Hodnota semivariogramu se počítá pro každý pár vstupních lokací. Na základě empirického semivariogramu je pak konstruován model, pomocí něhož je modelována distribuce jevu v prostoru [19].

Zkoumání datové struktury lze pomocí histogramu, Voronoi Map, Normal QQPlot, General QQPlot, Trend Analyst, Semivariogram/Covariance Cloud, Crosscovariance Cloud .

ArcGIS Info a jeho Geostatistical Analyst má v menu Geostatistical Wizard, to nabízí interpolační metody a zvolení jejich parametrů. Ordinary Kriging byl použit i pro naše data. Odstraní zkreslení v trendu dat. Zvolíme parametry modelu a semivariogramu prohledávací oblasti. Křížová validace (Cross Validation) poskytuje informaci o tom, jak dobře model sedí na vstupní hodnoty [4].

5.6 Postupy a metody

Sběr informací a dat, stanovení a konstrukce datové sady závislých a nezávislých proměnných, tvorba výběrového soboru, konstrukce segmentační databáze, metody statistické analýzy, dálkový průzkum Země a GIS.

5.6.1 Zdroje dat

Z důvodů ochrany citlivých dat nebude tato podkapitola veřejně poskytnuta.

5.7 Kód v prostředí R software pro daný dataset

Před programovaním v R softwaru osvěžme teorii, ketou uvedeme do praxe. Vysvětlující a vysvětlovaná proměnná jsou tedy v určitém vztahu, který můžeme modelovat. Závislá proměnná je tedy náhodná veličina s normálním rozdělením a pro odvození modelu se použije metoda nejmenších čtverců. Pokud závislá proměnná není spojitá, ale binární (0 či 1, výskyt polomu ano/ne), nelze použít k odhadu parametrů klasickou regresní analýzu, ale požijeme metodu logistické regrese k hodnocení pravděpodobnosti výskytu polomů. Tedy odhadne závisle proměnnou podle známých (nezávisle proměnných), které mohou ovlivnit výskyt jevu. Cílem je analýza efektů nezávislých proměnných (výška, věk, zakmenění), které mohou být numerického nebo kategoriálního typu.

Rovnice logistické regrese [32]:

$$Y = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{e^{g(x)}}{1 + e^{-g(x)}}$$

$$\mathrm{g(x)} = ß_0 + ß_1\, x_1 + ß_2\, x_2 + \ldots + ß_p\, x_p$$

$$ß_0, \ldots, ß_p = \text{parametery odhadů, a}$$

$$x_1, \ldots, x_p = \text{předpovídané proměnné.}$$

Datová sada nezávislých (vysvětlujících) a závislých (vysvětlovaných) proměnných, kvantitativního (množstevního, číselného) charakteru či kvalitativní (označují kategorie) charakteru. Každá proměnná v modelu je reprezentována sadou znaků. Ke každému znaku je následně přiřazená binární proměnná nabývajících hodnotu 1, prvek daného znaku nabývá (výskyt polomu v určitém zakmenění) a hodnotu 0, prvek daného znaku nenabývá. Celkový počet jednotek 85 142. Konstrukce segmentační databáze z datové sady o jednotkách 25x25metrů vytvořena pomocí nadstavby ArcGIS Spatial Analyst a ArcGIS Anylyst Tools se závislými a nezávislými proměnnými.

K odhadnutí parametrů modelu byla užita metoda maximální věrohodnosti, která spočívá v konstrukci věrohodnostní funkce, která udá pravděpodobnost, s jakou, při daném odhadovaném modelu, nastanou právě všechny pozorované události. Algoritmy pro výpočet parametrů jsou implementovány v běžně dostupných, výše vyjmenovaných, statistických programech.

Před výpočtem parametrů byly pro všechny vysvětlující proměnné spočteny Pearsonovy korelační koeficienty, v případech vysoké korelace mezi nezávislými proměnnými byla jedna z nich z výpočtu parametrů modelu vyloučena z důvodů multikolinearity. Výběr nejvhodnějšího modelu poskytla metoda postupné regrese (nalezne nejvhodnější kombinaci nezávislých proměnných), nejlépe vysvětluje pravděpodobnost výskytu jevu.

Vybrány byly:

- nadm_vyska (nadmořská výška),
- vek (věk),
- zakmeneni (zamenění),
- vlhkost,
- hloubka
- smrk_zast (zastoupení smrku v porostu).

Model logistické regrese byl ověřen testem poměru věrohodností, jimž se testují zejména hypotézy o vlivu konkrétní nezávislé proměnné na pravděpodobnost výskytu zkoumaného jevu. Testování významnosti jednotlivých parametrů provedeno pomocí takzvaného Waldova testu (poměr maximálně věrohodného odhadu a odhadu směrodatné odchylky).

My se však soustřeďme na výpočet a vysvětlení logistické regrese v R softwaru s grafickými výstupy, jak je tomu v zadání bakalářské práce.

5.7.1 Napsání kódu

po zapnutí R změňte adresář do své složky s uloženými daty (File, Change dir)

načteme použitá datové sady ze svého adresáře (příkaz read.table)

´´jméno souboru´´, header =TRUE (rozpozná záhlaví), sep= oddělovač, quote=vrací

argument, dec=desetiný oddělovač

data = read.table("polomy.txt", header = TRUE, sep = ",", quote="¨")

souradnice = read.table("xy_coord.txt", header = TRUE, sep = ";", quote=" ", dec=",")

generalizovaný lineární model (glm) pro všechny faktory ovlivňující pravděpodobnost výskytu polomu

glm.p = glm(data$vyskyt data$krivost+ data$nadm_vyska + data$sklon + data$orientace + data$vek + data$zakmeneni + data$tloustka_s + data$vyska_smrk + data$rychlost_v + data$vlhkost + data$hloubka + data$smrk_zast + data$p_typ, family=binomial(logit))

souhrné výsledky modelu glm

summary(glm.p)

Generalizovaný lineární model

Ve statistice znám pod pojmem GLM (generalized linear model) flexibilní zevšeobecnění normálního regersního modelu. GML je používán k uchycení generalizovaného modelu určeného lineární předpovědí a popisem rozložení chyb.

proběhl výběr nejvýznamějších proměnných

generalizovaný lineární model (glm) pro vybrané faktory ovlivňující pravdě-

podobnost výskytu polomu

glm.p2 = glm(data$vyskyt data$nadm_vyska + data$vek + data-$zakmeneni + data$vlhkost + data$hloubka + data$smrk_zast, family =binomial(logit))

souhrné výsledky modelu glm.p2, vypíše program, zajímá nás Pr (pravděpodobnost, aby byla menší jak 0,05, tedy s maximální chyba 5%)

summary(glm.p2)

Call:glm(formula = data$vyskyt data$nadm_vyska + data$vek + data$zakmeneni + data$vlhkost + data$hloubka + data$smrk_zast, family = binomial(logit))

výsledky modelu glm.p2 vložím do proměnné a odhad (Estimate)

a = summary(glm.p2)

výpočet pravděpodobnosti p s vybranými faktory podle vzorce logistické regrese

p = exp(coef(a)[1] + coef(a)[2]*data$nadm_vyska + coef(a)[3]*data-$vek + coef(a)[4]*data$zakmeneni + coef(a)[5]*data$vlhkost + coef(a)-[6]*data$hloubka + coef(a)[7]*data$smrk_zast)/(1 + exp(coef(a)[1] + coef(a)[2]*data$nadm_vyska + coef(a)[3]*data$vek + coef(a)[4]*data$za-kmeneni + coef(a)[5]*data$vlhkost + coef(a)[6]*data$hloubka + coef(a)-[7]*data$smrk_zast))

I. možnost (85 142 položek)

vytvoříme nový pracovní dataset

c spojení hodnot do vektoru, "název"

jmeno = c("p", "x", "y")

spojení R objektů podle sloupců a řádků

fin = cbind(p, souradnice[,1], souradnice[,2])

vytvoří soubor v adresáři

write.table(fin, file = "fin.csv", sep = ",", dec = ".",

col.names = jmeno, row.names = FALSE)

nutná knihovna, pokud není nainstalovaná zvolte Packages, Set CRAN mirror

(vyber zemmi), Instal packages (vyber gstat) library(gstat)

data_krig = read.csv('fin.csv')

pro gstat není vhodné, pokud data chybí, verzování originálními daty

error = na.omit(data_krig)

převeď jednoduchý datová rámec na speciální datový rámec objektu

coordinates(error) = ~ x+y

testování výsledků pomocí poloměru kruhu "bubliny"

bubble(error, zcol='p', fill=FALSE, do.sqrt=FALSE, maxsize=2)

vytvoření gridu, ve kterém budeme interpolovat

datový rozsah

x.range = as.integer(range(error@coords[,1]))

y.range = as.integer(range(error@coords[,2]))

vytvoř grid se 100 metrovým rozestupem

grd = expand.grid(x=seq(from=x.range[1], to=x.range[2], by=100),

y=seq(from=y.range[1], to=y.range[2], by=100)

převeď na SpatialPixel (prostorově-pixelová třída) class

coordinates(grd) = ~ x+y

gridded(grd) = TRUE

vizuální testování

plot(grd, cex=0.5)

points(error, pch=1, col='red', cex=0.7)

title("Interpolovaný grid a vzorkovací body"

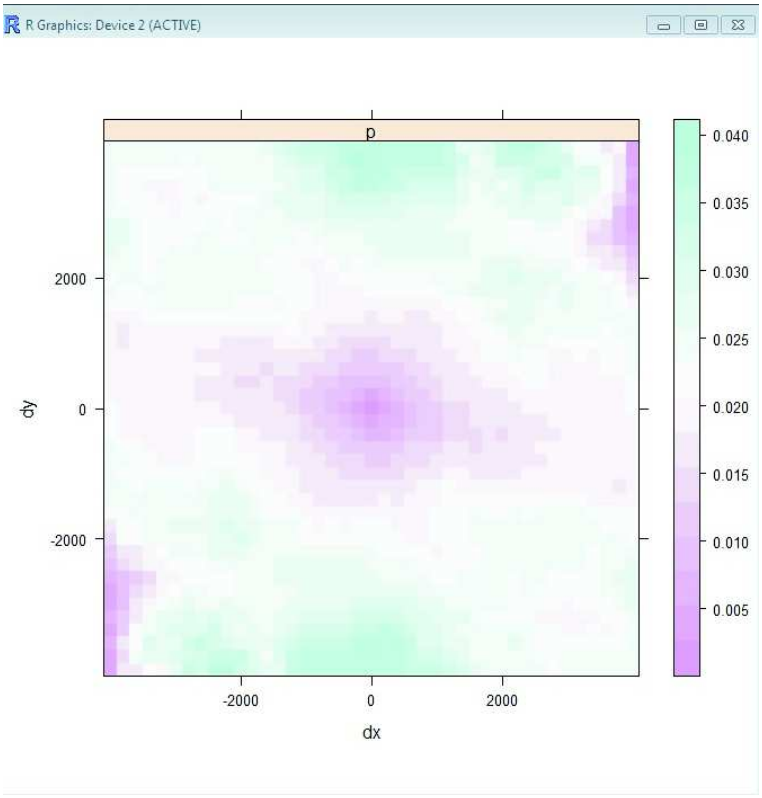
vytvoření gstat objektu

g = gstat(id="p", formula=p 1, data=error

mapa variogramu

plot(variogram(g, map=TRUE, cutoff=4000, width=200),

threshold=10)



Obr. 6. *Variogram.*

nebo můžeme vytvořit směrový variogram v 0°, 45°,90°, 135° od severu

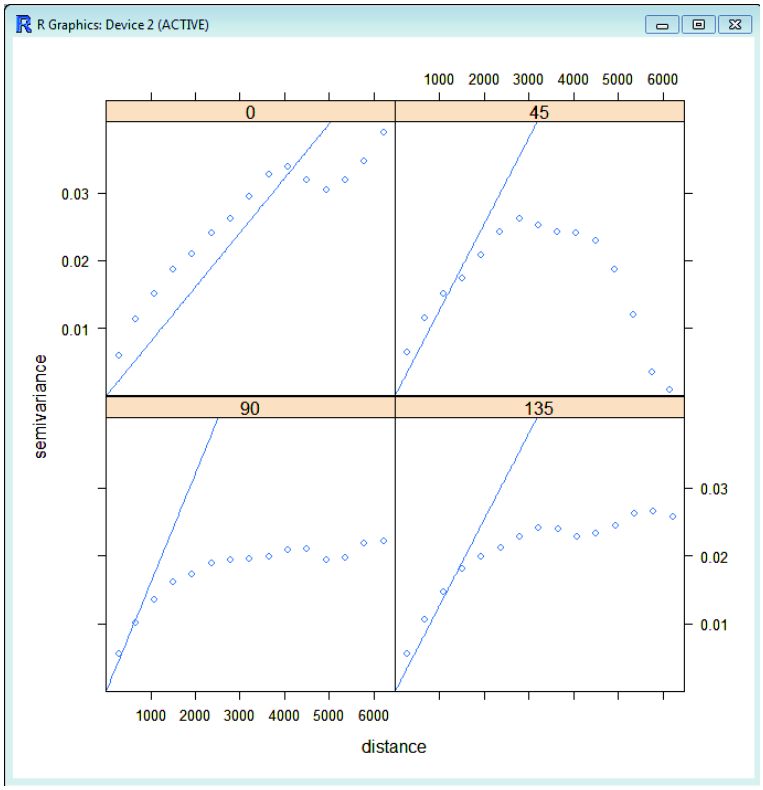
v = variogram(g, alpha=c(0,45,90,135))

0° a 45° vypadají dobře, použijmeje na uchycení linearního modelu variogramu

v.fit = fit.variogram(v, model=vgm(model='Lin', anis=c(0, 0.5)))

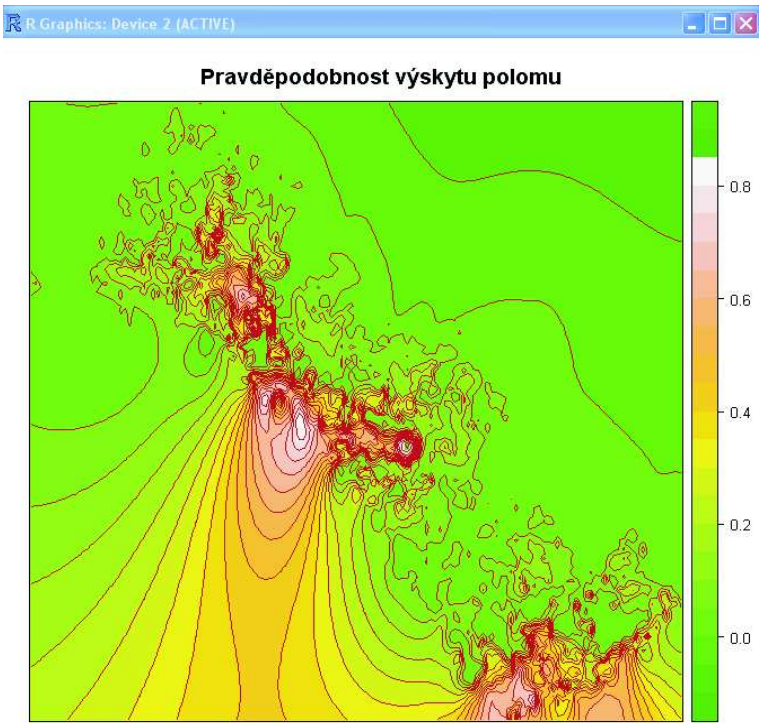
výsledek

plot(v, model=v.fit, as.table=TRUE)



Obr. 7. Vybrání nejvíce přilehajícího modelu

```
    ## aktualizuj gstat objekt
g = gstat(g, id="p", model=v.fit )
    ## proveď predikci ordinary (běžného) kriging
pq = predict(g, model=v.fit, newdata=grd)
    ## vizualizace
## základní grafy
par(mar=c(2,2,2,2))
image(pq, col=terrain.colors(20))
contour(pq, add=TRUE, drawlabels=FALSE, col='brown')
points(error, pch=4, cex=0.5)
title('Pravděpodobnost výskytu polomů')
```



Obr. 8. *Pravděpodobnost výskytu polomů na území (II. možnost).*

II. MOŽNOST

```
## generování náhodných čísel
## x = počet záznamů dat
x = 85142
## generuj náhodná čísla v rozsahu 1 až x
sample(x)
sample(x,replace=TRUE)
write.table(sample(x), file = "sample.csv", sep = ",", dec = ".",
col.names = jmeno, row.names = FALSE)
sample = read.csv("sample.csv", header = TRUE, sep = ";",
quote="¨", dec=",")
jmeno = c("nahodcis", "p", "x", "y")
fin.nah = cbind(sample,p, souradnice[,1], souradnice[,2])
## vytvoř tabulku se jménem fin_nah
write.table(fin_nah, file = "fin_nah.csv", sep = ",", dec = ".",
col.names = jmeno, row.names = FALSE)
library(gstat)
data_krig = read.csv('fin_nah.csv', header = TRUE,sep=";",dec=",")
error = na.omit(data_krig)
```

pokračujeme stejně jako u I. možnosti

Na základě výstupu logicky odhadneme, že většina území spadá do intervalu s nejnižšími hodnotami pravděpodobnosti polomu 0 až 0,2. Plochy se středním rizikem ohrožení, 0,2 až 0,6 zaujímají přibližně jednu třetinu území.

Plochy s nejvyšším pravděpodobnostím rizikem, okolo 0,8, dobře prezentují ohrožení. Intervaly s nejvyššími hodnotami pravděpodobnosti výskytu větrných polomů se smysluplně překrývají a navazují s opravdovými polomovými plochami způsobenými orkánem Kyrill, což potvrzuje správnost použitých postupů a modelu.

6 Diskuse

Nelze kompletně popsat celý svět, proto i míra faktorů, která byla přiřazena datům, nemůže zachytit naprosto veškeré zákonitosti, které k polomům napomáhají. I když je snaha zachytit jich co nejvíce, samozřejmě s kompromisem na finanční a časovou náročnost, nakonec kvůli výpočetní technice vybereme jen ty nejdůležitější faktory, abychom zjistili relevantní výsledky ve smysluplném čase. A proto se částečně ochuzujeme o 100% správný výsledek, pokud nějaký existuje, v poměru aktuální a chtěné využitelnosti.

Samotný výběr parametrů, byl převzat z jiné studie, protože mým nejhlavnějším cílem bylo uvedení do problematiky matematické teorie, představení programů a naučení práce v R prostředí s kompletním vysvětlením kódu, aby jej bylo schopno dále interpretovat a použít. Neboť jsou všechny statistické softwary koncipovány tematicky podobně, v psaní do rozhraní pro výpočet, zvýší se i intuitivní práce i mimo R. Neznamená to, že pokud byste se rozhodli použít jiný z nabízených programů, že si dále neprostudujete jeho detailnější principy.

Použití různých programů vede i k různým výsledkům, to musíme mít vždy na paměti. Například srovnání SAS a R pomocí vzorce logistické funkce pro výpočet pravděpodobnosti s exponenty.

Obecně lze v programech napsat

p = exp(coef(a)[1] + coef(a)[2]*data$nadm_vyska + coef(a)[3]*data$vek + coef(a)[4]*data$zakmeneni + coef(a)[5]*data$vlhkost +coef(a)[6]*data$hloubka + coef(a)[7]*data$smrk_zast)/(1 + exp(coef(a)[1] + coef(a)[2]*data$nadm_vyska + coef(a)[3]*data$vek+ coef(a)[4]*data$zakmeneni + coef(a)[5]*data$vlhkost + coef(a)[6]*data$hloubka + coef(a)[7]*data$smrk_zast))

V SAS 9.1 může dospět k výsledkům

p = exp (-(-16.0743 + ELEV * 0,00794+ VEK * 0,0221+ ZAKM * 0,1198-VLHK * 0,1379+ HLOUB * 0,3226+ ZAST_S * 0,0156)/(1 + exp(-16.0743 + ELEV * 0,00794+ VEK * 0,0221+ ZAKM * 0,1198-VLHK * 0,1379+ HLOUB * 0,3226+ ZAST_S * 0,0156)[6]

A v R 2.10.1 k takovým to

p = exp (-(-14.96 + ELEV * 0.008603 + VEK * 0.01961 + ZAKM * 0.04808 -VLHK * 0.1288 + HLOUB * 0. 3259 + ZAST_S * 0.004872)/(1 + exp(-14.96 + ELEV * 0.008603 + VEK * 0.01961 + ZAKM * 0.04808 -VLHK * 0.1288 + HLOUB * 0. 3259 + ZAST_S * 0.004872)

ELEV = nadm_vyska = nadmořská výška, VEK = vek = věk, ZAKM = zakmeneni = zakmenění, VLHK = vlhkost = vlhkost, HLOUB = hloubka = hloubka, ZAST_S = smrk_zast =zastoupení smrku.

Drobné nuance se mohou stírat, některé podporovat. Jak tento výpočet dále ovlivní průběh dalších kroků, můžeme až s nadhledem zkontrolovat, při vizualizaci výsledků a jejich logického odůvodnění. V mém případě však konečný výstup pravděpodobností výskytu polomů v R softwaru ukazoval velmi podobné, téměř shodné oblastí se stejnou pravděpodobností výskytu polomů jako při výpočtu v SAS systému. Můžeme také polemizovat o správnosti nastavení fit modelu (uchycení lineárního modelu) a jeho následného výběru pro další počty. Zda měl být lineární či sférický. Ale pro tyto konkrétní data, se jevil lépe lineární model.

I přes veškeré silné stránky grafické vizualizace R softwaru, nedosahuje plnohodnotné tvorby výstupů, samozřejmě neumí mapové výstupy, jako u ArcGIS Desktop 9. x, proto i závěrečné výstupy jsou prezentovány způsobem, jakým to R software dovoluje.

Nicméně pokud jej budeme chápat jako výhradně statistický program, je jeden z nejzdařilejších, jaký lze nalézt a který je schopen zodpovědět i GIS otázky. Na druhé straně je vidět, že i samotné GISy se neobejdou bez lepší a průhlednější implementace statistických nadstaveb do jejich prostředí.

Výpočty byly prováděny nad náhodně generovanými a redukovanými daty, ze kterých je i výstup. Je možné, že pokud bychom nechali generovat náhodná čísla v jiném prostředí, vybrala by se jiná data, která poskytnou jiný výsledek. Či lze nechat spočítat veškeré dostupné položky výkonnějším počítačem pomocí sériového napojení více počítačů, kdy každý jeden z nich by počítal část operace. Ale to je pro uživatele začátečníka těžko proveditelné a časové náročné. Pro pokročilejší studie je to však nevyhnutelné a běžně používané.

Proveden výpočet variogramu pro veškerá data a pro vybraná. V jejich porovnání nebyly nalezeny dramatické změny, proto byla uznaná za vhodná. Pro představení logistické regrese, prvotní nahlédnutí a práci s R softwarem bude výsledek dostačující i pro menší objem dat, který si bude moci, více uživatelů zkusit sám a pokud by se setkali s nutností výpočtu veškerých položek, mají také poskytnuté řešení.

Prostudov8no i jiné nezávislé metody výpočtu pravděpodobnosti výskytu polomů v zájmovém území, v rámci STRiM projektu.

7 Závěr

Důkladným pochopením základních matematických pouček budeme schopni porozumět statistickému softwaru a intuitivně jej ovládat. Budeme schopni si také logicky odvodit vyhodnocené závěry a rozhodnou, jak jsme byli úspěšní, což velice zefektivní naši práci a sníží časový interval celého procesu řešení problému.

Korelace a logistická regrese jsou tedy velmi vhodnými metodami pro počítání s binárními daty a s pravděpodobnostním odhadem jevů v prostoru, jak jsme se dozvěděli z matematické teorie a z uvedených příkladů.

Větrné polomy, které byly použity pro praktickou ukázku, ovlivňuje velké množství faktorů a parametrů vyskytujících se ve fyzické i biologické sféře. Výběrem nejdůležitějších faktorů, pomocí korelačních koeficientů, zlepšíme samotné hodnocení pravděpodobnostní mapy. Pochopením jednotlivých vztahů snížíme i náročnost na výpočet.

Logistická regrese vyhodnotila pravděpodobnosti výskytu polomů v zájmové oblasti a odhadla závisle proměnné podle známých proměnných, které mohou ovlivnit výskyt jevu.

Samotná problematika polomů je velmi skloňovaným tématem dnešní doby, neboť je stále více ekonomicky náročnější odstranění následků větrných smrští. Proto je jim třeba předcházet, například výsadbou jiných dřevin či jiným rozmístěním stromů.

Nejprve byla tedy objasněna matematická teorie, byly také představeny jednotlivé statistické a geoinformatické softwary (R, SAS, STATA a ArcGIS 9.x). Detailně popsána problematika polomů a bylo vymezeno zájmové území. Představeny byly také základní přístupy pro hodnocení rizik ohrožení větrem, průzkumové analýzy dat, volba interpolační metody a interpolace a postupy a metody užité na zdroji dat. Samotné řešení úlohy je provedeno pomocí R softwaru s jeho výstupy a s ohodnocením správnosti výsledku. Výsledky mají bezesporu praktické využití v protikrizových opatřeních nebo pro prevenci snížení škod.

Studie hlavně měla seznámit s matematickou teorií, základními satistickými softwary a s prací v R softwaru, který je velmi dobrým nástrojem i pro grafické výstupy.

8 Summary

Mathematical theory can accompany you in every statistical or graphics systems, where you may be confused about the actions, what you have set going. Results and outcomes, what the software offered can be wrong, but still with a nice outcome. Therefore is our logic needed or we should be able to find a way, how can we inspect demanded hypothesis.

Keywords: correlation, logistic regression, windthrow, GIS, statistical software

My work consider just a small, but very effectual, part of statistical area for these days. It is correlation and logistic regression with example in R system upon datasets, for prognose of places with higher posibility of windthrow. Here is also described basic approaches and methods of assessing wind damage disaster, dataset´s developing, spatial prediction methods of windthrow disaster and comparing tools in software (ArcGIS 9.x, SAS, STATA).

In short I introduce you the correlation and logistic regression theory, but more space I leave for a practical example. In my Czech version mathematical theory contain more and detailed information (method of least squares, a computing linear regression, a regression diagnostics, calibration), but for summary is better present outcomes and examples.

Correlation is a statistical technique, what can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people or higher places are more affected by wind erosion. The relationships are usually not perfect. Smaller can be heavier than taller, or high places can be orientated, where wind can not blow, so with minimum wind erosion.

Correlation is able to tell you just how much of the variation in peoples' weights is related to their heights. Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are

correlations, but do not know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data. There are several different correlation techniques. The most common type, called the Pearson or product-moment correlation, which I also used for variable in my datasets. When there was too high correlation between independent variables, one of them was needed to expect, because of multicorrelation. Correlation is only appropriate for certain kinds of data. it works for quantifiable data in which numbers are meaningful. It cannot be used for purely categorical data, such as gender or color. Correlation Coefficients as main result range from -1.0 to +1.0 and say how closely the two variables are related, 0 no related or -1;1 related. Correlation analysis is a powerful tool to enhance the predictive power of analysis [7].

Logistic regression is a mathematical model for the analysis of binary data with widespread applicability in the physical, biomedical, and behavioral sciences. Parameter inference for this model is usually based on maximizing the unconditional likelihood function. For large well-balanced datasets or for datasets with only a few parameters, unconditional maximum likelihood inference is a satisfactory approach.

However unconditional maximum likelihood inference can produce inconsistent point estimates, inaccurate p values, and inaccurate confidence intervals for small or unbalanced datasets and for datasets with a large number of parameters relative to the number of observations, here can help study of Monte Carlo sampling or Markov chains, but it is not one of my main research topic [8].

My datasets are created with windthrow data, what can be caused by wind, snow or ice. Result of these factors is damaging forests, what can have great impact for economy and also the environment, but we can not avoid windthrow at all, because it is a natural process, which can help forests to grown, change old wood and keep balanced biodiversity.

With advanced technologies like GIS (geographical information systems), -SDSS (Spatial decision support systems), predictive models and specialized tools

is possible to find the areas, where is higher likelihood for wind damaging and decide, how set the protection. Although, my data cover just particular territory in the Czech Republic, the principles can be used everywhere and for other problems, not just for windthrow but also for every tasks, where is needed likelihood prediction (where can we expect floods, eruptions and so on).

The Czech Republic is situated in the middle of Europe, therefore has been hit by extreme storm events many times. The most affected and the most predicative zone for simulation windthrow is concededly Šumava mountains. History of Šumava mountains show the most powerful storms between the years 1868 – 1878 and 1955-1962. In each period over 3 million cubic meter of harmed or broken trees [3]. In last decade are the biggest calamity caused by Kyrill 2007, which describe my datasets.

The windthrow risk is defined as the probability of a tree or stand being blown down by an extreme strong wind. It takes negative impact for the water regime, because forests naturally regulate the retention of watershed and flooding [21]. Three main methods can evaluate the windthrow risk: empirical (use a qualitative assessment for wind damage), mechanistic (for calculating critical wind speed and the probability of this wind occurring) and proper observe in the particular landscape. Hybrid mechanistic/empirical models can be represented by HWIND model [26] and also ForestGALES [20].

Methods and materials describe data gaining and creating dataset, depending determination and independent variables, data screening, statistical method of logistic regression using the R software and describing logistic regression also with other statistical and GIS software.

The data were obtained by aerial photographs, field inventory, digital elevation model, forest maps, and topographic maps. The windthrow areas were detected and identified by aerial photographs taken in 2007 (monitoring wind damage caused by hurricane Kyrill immediately after the disaster). Dataset used for wind damage classification included Regional Plans of Forest Development (RPFD),

data (typology map, map of forest vegetations levels, map of functional forest potential etc.) and forest stand map.

The classical regression are not suitable for prediction of windthrow, but logistic regression is great method for the prediction of discrete outcomes (binomial is damage (1) / undamaged (0)).

Logistic regression works with the method of maximum likelihood to estimate parameters and takes the model form [32]:

$$Y = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{e^{g(x)}}{1 + e^{-g(x)}}$$

$$\mathrm{g(x)} = ß_0 + ß_1\ x_1 + ß_2\ x_2 + \ldots + ß_p\ x_p$$

$$ß_0, \ldots, ß_p = \text{parameter estimates, and}$$

$$x_1, \ldots, x_p = \text{predictor variables.}$$

Logistic regression has proven to be a useful tool to estimate the probability of windthrow. It has been utilized by Valinger and Fridman (1999), Jalkanen and Matilla (2000), Canham et al. (2001), Mitchell et al. (2001), Peterson (2004), Lanquaye - Opoku and Mitchell (2005), and Scott and Mitchell (2005) to predict the probability of windthrow on an individual tree level [17]. This statistical method was therefore used in his study to assess windthrow risk and generate the probability of windthrow in the study territory [5].

Logistic regression should be shown with R software, but also sum up the possibility of analysing dataset with other statistical and geoinformatics software. R software [28] is open source mathematician program with graphical outcome, programmed with R and S language. For expanding basic configuration R software is possible set up packages on CRAN Mirrors. There is allowed input, store data, manipulate with sets and matrix and analyzing datasets. R software can be also used for data mining. Other statistic programs can be used for logistic regression for example SAS system (used with a company management´s database care, but also for science purpose), STATA software (Data Analysis and Satistical

Software, here is able manage data, analyse them, but graphical tools are worse than R software). The used GIS software is from ESRI family, ArcGIS Desktop 9.x with geostatistical analyst and licence ArcInfo. But all outcomes are from R software, because the aim was to show that statistical software can be also used for geoinformatics analysis and not just clear GIS software.

The dependent variables, present the wind damage, are usually dichotomous and can take the value 1 (a probability of success) or the value 0 (the probability of failure). Sets of dependant variables were then derived from the spatial layer containing attributes related to occurrence of windthrow. But the independent variables in logistic regression can take more than one form, and logistic regression makes no premise about the distribution of the independent variables. The final dataset contained 85 141 segments.

Logistic regression was performed in the R software. The potential independent (explanatory) variables include only the relevant variables. Pearson correlation coefficients identified variables to minimize multicolinearity. The suitable independent variables were obtained, the stepwise selection methods were chosen to aid in the formulation of a model. The stepwise result of variable selection elected this the most important parameters nadmor_vyska (ELEV, elevation), vek (VEK, age), zakmenení (ZAKM, crop density), vlhkost (VLHK, humidity of soil), hloubka (HLOUB, depth of rootage) and smrk_zast (ZAST_S,number of suprce in the territory).

Result in R software for windthrow hazard likelihood:

p = exp (-(-14.96 + ELEV * 0.008603 + VEK * 0.01961 + ZAKM * 0.04808 -VLHK * 0.1288 + HLOUB * 0. 3259 + ZAST_S * 0.004872)/(1 + exp(-14.96 + ELEV * 0.008603 + VEK * 0.01961 + ZAKM * 0.04808 -VLHK * 0.1288 + HLOUB * 0. 3259 + ZAST_S * 0.004872)

Conclusion the R software can give good graphical results, but not as good as GIS software, what can give also maps. The logit (log of the odds) was calculated and then converted to a probability. A large part (almost 60%) of study territory

show a low risk (green) of windthrow. The medium risk, one quarter of area, is colored with yellow, Rest of the territory (white) shows places with the highest probability for windthrow disaster.

Many natural and anthropogenic factors induce the probability of the windthrow. We should beware, that not only knowledge, how various factors relate, can lead us to the right results, but as important as this is know how factors are related to each other. Then we can understand how the factors affect the occurrence of windthrow disaster and choose the method for the best simulation of likelihood, when can be wind damage and where the probability is low.

The process of assessing wind damage hazard using free R software, but also other statistical and GIS software were described, if statistical tools for logistic regression are there.

Literatura

[1] DEMEK, J., et al. *Geomorfologie českých zemí.*Praha: Academia, 1965.

[2] JANOŠKA, Zbyněk.*FAKTOROVÁ ANALÝZA VÝSKYTU VYBRANÝCH BOTANICKÝCH A ZOOLOGICKÝCH TAXONŮ.*Olomouc, 2009. 64 s. Bakalářská práce. Universita Palackého.

[3] JELÍNEK, J.(1985). *Větrná a kůrovcová kalamita na Šumavě z let 1868 až 1878.*Ústav pro hospodářskou úpravu lesů, Brandýs nad Labem.

[4] JUŘIKOVSKÁ, L. (2009).*Geostatistika populární a hravá. Seminář při mezinárodním symposiu GIS Ostrava 2009.*. - Institut geoinformatiky, VŠB-TU Ostrava, Ostrava, 11 pp.

[5] KREJČÍ, Lukáš. *Hodnocení lesních území z hlediska výskytu větrných polomů pomocí GIS a logistické regrese* . GIS Ostrava 2010,Ústav geoinformačních technologií, Lesnická a dřevařská fakulta MZLU v Brně, s. 1-8.

[6] KREJČÍ, Lukáš. *he assesment of forest areas in term of windthrow occurrence using GIS.*Faculty of Forestand Wood Technology, Mendel University of Agriculture and Foresty,Brno, Czech republic. 2010, [cit. 2010-05-17].

[7] MELOUN, Milan; MILITKÝ, Jiří. *Satistická analýza experimentálních dat.*Vydání 2., upravené a rozšířené. Praha : Academia (Akademie věd České republiky), 2004. Korelace, s. 737-779. ISBN 80-200-1254-0.

[8] MELOUN, Milan; MILITKÝ, Jiří. *Statistická analýza experimentálních dat.*Vydání 2., upravené a rozšířené. Praha : Academia (Akademie věd České republiky), 2004. Lineární regresní modely, s. 565-734. ISBN 80-200-1254-0.

[9] REKTORYS, Karel, et al. *Přehled užité matematiky II.*sedmé vydání. Praha : Prometheus, 2003. 876 s. ISBN 80-7196-181-7, 2.díl.

[10] REKTORYS, Karel, et al. *Přehled užité matematiky II.* sedmé vydání. Praha : Prometheus, 2003. Matematická statistika, s. 683-742. ISBN 80-7196-181-7, 2.díl.

[11] QUITT, E. *Klimatické oblasti Československa.* Mapa měřítka 1:500 000, Brno: GgÚČSAV, 1971.

[12] Using ArcGIS Geostatistical Analyst.

[13] VICENA, I., JUHA, M., NOŢIČKA, S. *Větrné polomy a vývraty na území NP a CHKO Šumava v roce 2002, jejich příčiny a následky* In Aktuality šumavského výzkumu II, Sborník z konference. Vimperk: Správa NP a CHKO Šumava, 2004.

[14] ZVÁRA, Karel. *Regrese*Vydání první. Praha : MATFYZPRESS, 2008. 12 Logistická regrese, s. 171 - 186.

[15] AMMANN, R.A. Defibrotide for hepatic VOD in children: exact statistics can help! *Bone Marrow Transplantation* [online]. 2004, n. 34, [cit. 2010-05-16]. Dostupný z WWW: <http://www.cytel.com/Papers/2004.08%20Letter_nature.pdf >.

[16] Answers Corporation. *Answer.com* [online]. 2008 [cit. 2010-05-17]. SAS. Dostupné z WWW: <http://www.answers.com/topic/sas-system>.

[17] DRAKE, T. *Empirical Modeling of WindthrowOccurrence in Streamside Buffer Strips* [on-line].c2008 [cit. 2010-05-09]. <http://ir.library.oregonstate.edu/jspui/bitstream/1957/9544/1/Drake-_Thesis.pdf>

[18] ESRI. *ESRI*[online]. 2008 [cit. 2010-05-17]. ArcGIS. Dostupné z WWW <http://www.esri.com/software/arcgis/index.html>.

[19] ESRI Products Overview [online]. ©ESRI. [cit. 2010-04-05] Dostupné z WWW <http://www.esri.com/products/index.html>

[20] GARDINER, B. et al. *ForestGALES – A PC-based wind risk model for British Forest – USER'S GUIDE* online]. June 2004 [cit. 2010-05-8]. Dostupné z WWW: <http://www.forestry.gov.uk/pdf/ForestGALES2_manual_2004.pdf>

[21] GARDINER, B. et al. *Review of mechanistic modelling of wind damage risk to forests.*Forestry [on-line]. 2008, vol. 81, no. 3 [cit. 2010-05-16].Dostupné z WWW: <http://forestry.oxfordjournals.org/cgi/reprint/81/3/447>.

[22] MAITI, Tapabrata ; PRADHAN, Vivek. Bias Reduction and a Solution for Separation of Logistic Regression with Missing Covariates. *Cytel, Statistical Services Services*[online]. 2008, n., [cit. 2010-05-16]. Dostupný z WWW: < http://www.cytel.com/Papers/Pradhan_Bias_Reductn_Solutn-_Separatn_Logstc.pdf >.

[23] MEHTA, Cyrus R. ; PATEL, Nitin R. ; SENCHAUDHURI, Pralay. Efficient Monte Carlo Methods for Conditional Logistic Regression. *Cytel, Statistical Software Services* ,Services [online]. 1999, n.1, [cit. 2010-05-16]. Dostupný z WWW: < http://www.cytel.com/papers/EfficientMonteCarlo2000.pdf >.

[24] MITCHELL, S., OPOKU, N. *Windthrow Hazard Mapping using GIS, Canadian Forest Products TFL 30, McGregor Model Forest* [on-line]. 2004 [cit. 2010-05-16].Dostupné z WWW: <http://faculty.forestry.ubc.ca/mitchell/publications/mcg.pdf>.

[25] Národní park Šumava. *Správa národního parku a chraněné krajinné oblasti Šumava* [online]. 2008 [cit. 2010-05-16]. Základní údaje. Dostupné z WWW: <http://www.npsumava.cz/1261/sekce/zakladni-udaje/>.

[26] PELTOLA, H. et al. *A Mechanistic Model for Assessing the Risk of Wind and Snow Damage to Single Trees and Stands of Scots pine, Norway spruce and birch* [on-line]. 1999 [cit. 2010-04-28].Dostupné z WWW: <http://www.forest.joensuu.fi/storms/hwinkkuv.htm>

[27] Qweas.*Qweas*[online] 2005-2009 [cit. 2010-05-17]. Sotware guide. Dostupné z WWW: <http://www.qweas.com/downloads/education/mathematics/scr-stata-for-mac.jpg>.

[28] r-project.org. *What is R* [online]. 1997 [cit. 2010-05-10]. The R Project for Statistical Computing. Dostupné z WWW: <http://www.rproject.org/>.

[29] SAS Institute Ic. *SAS* online]. 2008 [cit. 2010-05-17]. Products and Solution. Dostupné z WWW: <http://www.sas.com/>.

[30] ŘEZANKOVÁ, Hana ; MAREK, Luboš Marek; VRABEC, Michal. *IASTAT*[online]. 2001 [cit. 2010-05-16]. INTERAKTIVNÍ UČEBNICE STATISTIKY. Dostupné z WWW: < http://iastat.vse.cz/>.

[31] StataCorp LP. *Data Analysis and Staistical Software* online]. 1996-2010 [cit. 2010-05-17]. Learn about STATA. Dostupné z WWW: <http://www.stata.com/stata11/>.

[32] Tree stability and climate [on-line]. 2009 [cit. 2010-05-16]. Dostupné z WWW: <http://www.forestresearch.gov.uk /fr/INFD639A92>

[33] UCLA Academic Technology Services. *Stat Computing*[online]. 2001 [cit. 2010-05-17]. Logistic Regression with Stata. Dostupné z WWW: <http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter1/statalog1.htm>.

[34] UCLA Academic Technology Services. *Statistical Computing Seminar*[online]. 2001 [cit. 2010-05-17]. Proc Logis-

tic and Logistic Regression Models. Dostupné z WWW: <http://www.ats.ucla.edu/stat/sas/seminars/sas_logistic/logistic1.htm>.

[35] ULBRICHOVÁ, I. et al.*Development of the spruce natural regeneration of mountain sites in the Sumava Mts. Journal of Forest Science* [on-line]. 2006,vol. 52, no. 10, p. 446-456 [cit. 2010-05-5].Dostupné z WWW: <http://www.cazv.cz/userfiles/File/JFS%052_446-456.pdf >

[36] Zpráva o stavu lesa a lesního hospodářství České republiky 2006 [on-line]. Brandýs nad Labem Ústav pro hospodářskou úpravu lesa, 2007 [cit. 2010-05-16].Dostupné z WWW: <http://www.uhul.cz/zelenazprava/>.

Přílohy

CD-ROM:

Textové přílohy:

- Bakalářské práce

Praktický výpočet(prakticky_vypocet):

- fin.csv
- fin_nah.csv
- final_nad.txt
- navod.txt
- polomy.txt
- xy_cord.txt

WWW stránky (www):

- images
- index.htm
- metody.html
- polomy.html
- prakticka.html
- style.css
- sw.html
- teorie.html