ACTA GEOGRAPHICA UNIVERSITATIS COMENIANAE, Vol. 61, 2017, No. 2, pp. 241-256

MATEMATICKÉ PRINCÍPY REGRESNEJ ANALÝZY V GEOGRAFICKEJ PERSPEKTÍVE

Tomáš Goga[1], Hana Bobáľová[2]

[1] *Slovenská akadémia vied, Geografický ústav, Bratislava, email: tomas.goga@savba.sk*

[2] *Univerzita Komenského v Bratislave, Prírodovedecká fakulta, Katedra kartografie, geoinformatiky a diaľkového prieskumu Zeme*

Abstract: The main objective of regression analysis is to examine and characterize the relationship between variables. Its task is to find a mathematical function called regression function respectively a regression model. This model has to be fitted the best it could to describe the course of dependence between variables. The goal of this paper is to help start-up scientists and students to understand the principles of statistical regression analysis. The paper provides a comprehensive overview about statistical evaluation process or about choosing the optimal regression function. It also represents the basic principles needed to reject regression models. The article does not deal with overview of available literature or with methods of data acquisition, but thoroughly describes methods of statistical analysis, pointing to the importance of these statistical methods and tests. We used the data created by author.

Key words: statistics, regression, analysis, model, function, residuals

1 ÚVOD

Štatistická analýza je v súčasnosti už nevyhnutným nástrojom pri interpretácii dát získaných počas geografického výskumu. Preto sa pri štatistickej analýze v geografickom výskume sa často stretávame s kvantitatívnym hodnotením dvoch a viac veličín. Často pritom rátame s hypotézou, že medzi veličinami existujú určité vzťahy tak, že zmena jednej premennej vyvolá zmenu inej premennej. Tieto veličiny bývajú vyjadrené funkčným vzťahom (1), pričom sú na základe uvedenej hypotézy štatisticky korelované (závislé):

$$y = f(x) \quad , \quad z = \varphi(y, x) \tag{1}$$

Preto uvedenú hypotézu môžeme analyzovať s využitím rôznych typov funkcií a ich konštánt, pričom na uvedenú analýzu je potrebné do procesu dodať dostatočné množstvo empiricky zistených údajov. Takýto druh riešenia kvantitatívneho hodnotenia sa nazýva regresnou analýzou. Jej cieľom je určiť štatistickú tesnosť empiricky získaných veličín (Bitterer, 2003).

Cieľom regresnej analýzy je preto hlbšie poznanie všeobecných vzťahov medzi veličinami, vniknutie do ich vnútorných súvislostí a konštrukcia vhodných regresných (matematických) modelov (Hindls, 1999).

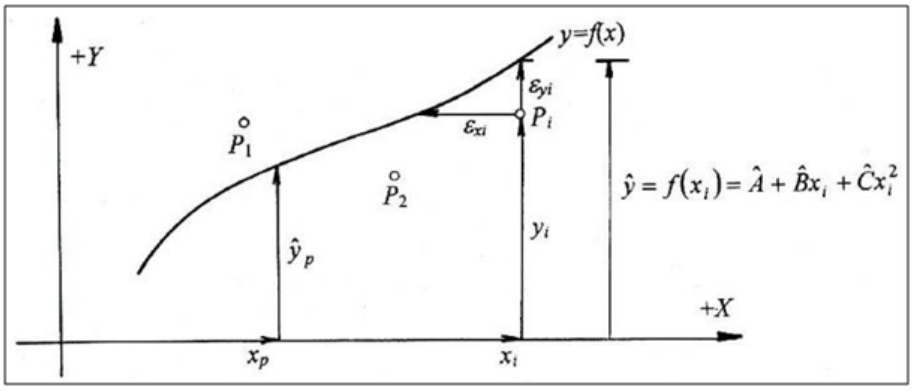
Cieľom príspevku je vysvetliť teoretický základ regresnej analýzy, poukázať na postup v prípade ich štatistického vyhodnocovania a výber najvhodnejšej regresnej funkcie. V geografickom výskume sa niekedy stretávame s problematikou nezamietania štatisticky nevýznamných modelov alebo regresných parametrov. Preto tento príspevok taktiež predstavuje základné princípy potrebné pre zamietnutie regresných modelov. Príspevok sa primárne nezaoberá prehľadom dostupnej literatúry ani metódami získavania použitých dát, ale dôkladne opisuje metódy štatistickej analýzy, pričom poukazuje na dôležitosť uvedených štatistických metód a testov.

2 METÓDY

Ak na základe vzťahu (1) nahradíme hodnoty X, Y reálnymi hodnotami x_i, resp. y_i, tak vyrovnávacia krivka $y = f(x)$ bude na základe vyššie uvedeného spojitá, a bude prechádzať medzi bodmi polygónu, ktorý je vytvorený reálnymi hodnotami x_i, resp. y_i (obr. 1). Odstupy bodov Pi od krivky ε_i nazývame reziduá alebo regresné chyby.

Vyjadrením priebehu javov odmeraných hodnôt závislej premennej y pri meniacich sa hodnotách argumentu x je nepravidelný rad bodov (tzv. empirický polygón). Cieľom regresnej analýzy je nájsť takú závislosť vyjadrenú vyrovnávacou krivkou, aby sa v ideálnom prípade primkla k empirickému polygónu. Výsledkom je regresná krivka.

Na základe matematickej štatistiky je potrebné zdôrazniť, že metódy regresnej analýzy budú mať pri obmedzenom splnení podmienok aj obmedzenú platnosť záverov. V konečnom dôsledku budú výsledky regresnej analýzy platiť iba na definovanom intervale $\in x_i$ ($i = 1, 2, \ldots n$) (Bitterer, 2003).



Obrázok 1 Regresná funkcia (Bitterer, 2003)

Regresná analýza je založená na empiricky získaných meraniach radu dvoch vzájomne nezávislých premenných. Výsledkom je rad dvojíc hodnôt, ktoré považujeme za dvojrozmerné veličiny, pričom predpokladáme, že aspoň jedna z týchto premenných je spojitá náhodná veličina (Bitterer, 2003).

Všeobecne možno regresné modely formulovať nasledovnými vzťahmi:

1. Základný model pre dáta x: model náhodného výberu – x je realizáciou náhodného výberu F_θ, $\theta \in \Theta$: $X = [X_1, X_2,\ldots, X_n]$, $E[X] = \mu\ (1, 1,\ldots, 1)^T$, $D[X] = \sigma^2 I_n$
2. Regresný model na Y_i, $i = 1,\ldots, n$:
   $y_i = g(x_i) + \varepsilon_i$ je ľubovoľná spojitá funkcia na R.
   A) g(.) je neparametrická funkcia – neparametrický model,
   B) g(.) je parametrická funkcia – parametrický model, kam patria klasické lineárne a nelineárne modely.

Hľadanie regresnej funkcie je podmienené vzťahom (2), kde β_j, $j = 0,1,\ldots, p$ sú regresné parametre:

$$Y = f(X, \beta_0, \beta_1, \ldots, \beta_p) = E(Y|X) \tag{2}$$

Regresné funkcie charakterizuje závislosť podmienených stredných hodnôt náhodnej veličiny Y na hodnotách náhodnej veličiny X (Katina, 2015).

Na základe vyššie uvedeného vzťahu (2) a matematických vzťahov môžeme vytvoriť viacero regresných predpisov a tým získať viacero typov regresných funkcií:

– Priamková regresia $Y = \beta_0 + \beta_1 X$

– Hyperbolická regresia $Y = \beta_0 + \frac{\beta_1}{X}$

– Logaritmická regresia $Y = \beta_0 + \beta_1 lnX$

– Parabolická regresia $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

– Polynomická regresia $Y = \beta_0 + \beta_1 X + \cdots, \beta_p X^p$

– Exponenciálna regresia $Y = \beta_0 \beta_1^X$

– Mocninová regresia $Y = \beta_0 X^{\beta_1}$

Pri regresnej analýze vychádzame z predpokladu, že každá hodnota y_i vysvetľovanej premennej Y je funkciou modelovej hodnoty η_i. Táto hodnota zahŕňa pôsobenie kontrolovaných vplyvov vysvetľujúcich premenných a hodnoty rušivej zložky ε_i. Rušivá zložka obsahuje pôsobenie nekontrolovaných, resp. menej podstatných alebo neuvažovaných vplyvov.

Pri využívaní regresných modelov sa predpokladá, že medzi zložkami η a ε platí súčtový vzťah (3), kde Y je skúmaná vysvetľovaná premenná, η matematicky modelovaná lineárna regresná funkcia a ε nepozorovateľná rušivá zložka (Herbák, 2005). Iné zdroje môžu zložku ε označovať aj ako tzv. rezíduum (Katina, 2015).

$$Y = \eta + \varepsilon \tag{3}$$

O rušivých zložkách sa zároveň predpokladá, že sú nezávislé a majú normálne rozdelenie s nulovými strednými hodnotami a rovnakými rozptylmi rovnými konštante σ^2 (stochastická premenná) (Hindls et al., 2005).

Najvhodnejšou metódou na získanie parametrov regresnej krivky je metóda najmenších štvorcov. Aproximácia tejto metódy vychádza z predpokladu, že súčet štvorcov zvyškových chýb modelu (reziduí) bude čo najmenší (4). Metóda ďalej pokračuje deriváciou podľa oboch koeficientov a následne riešením sústavy podľa Cramerovho pravidla (Neubauer, 2011).

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2 \rightarrow min \tag{4}$$

Interval spoľahlivosti pre regresný koeficient β_j odvodzujeme z náhodnej premennej v Studentovom rozdelení s n-2 stupňami voľnosti. Východiskom sú štatistiky (5):

$$t_j = \frac{(b_j - \beta_j)}{s(b_j)} \sim t(n-c) \quad \text{pre } j = 0, 1, \ldots, p \tag{5}$$

pričom b_j predstavuje bodový odhad parametra β_j, $s(b_j)$ predstavuje smerodajnú odchýlku tohto odhadu. Interval spoľahlivosti má následne tvar (6):

$$b_j - t_{1-\frac{\alpha}{2}}(n-c).s(b_j) < \beta_j < b_j + t_{1-\frac{\alpha}{2}}(n-c).s(b_j) \tag{6}$$

Ak ktorýkoľvek parameter pre tento interval obsahuje nulu, môžeme na predefinovanej hladine významnosti prehlásiť, že nulový parameter je pre model štatisticky nevýznamný.

Štatisticky významný parameter β_j je taký parameter, ktorý je nenulový. Preto pri teste významnosti regresných parametrov testujeme hypotézy (7) pri testovom kritériu podľa vzorca (5) a kritickom obore (8):

$$H_0: \beta_j = 0 \rightarrow H_A: \beta_j \neq 0 \tag{7}$$

$$W_\alpha: |t_j| \geq t_{1-\frac{\alpha}{2}}(n-c) \tag{8}$$

Pri teste významnosti celého regresného modelu platí (9):

$$S_Y = S_R + S_T \tag{9}$$

pričom S_Y predstavuje celkový súčet štvorcov definovaný vzťahom (10). S_R predstavuje súčet štvorcov reziduí definovaný vzťahom (11). S_T predstavuje teoretický súčet štvorcov definovaný vzťahom (12). Teoretický súčet štvorcov S_T je tá časť celkového súčtu štvorcov, ktorá je vysvetlená regresnou funkciou. Reziduálny súčet štvorcov S_R je tá časť celkového súčtu štvorcov, ktorá zvolenou regresnou funkciou vysvetlená nie je (Neubauer, 2011).

$$S_Y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n.s^2(y) \quad \text{, kde} \quad s^2(y) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{10}$$

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (n-c).s_R^2(y) \quad \text{, kde} \quad s_R^2(y) = \frac{1}{(n-c)}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{11}$$

$$S_T=\sum_{i=1}^{n}(\hat{y}_i-\hat{y})^2=n.s^2(\hat{y}) \quad \text{, kde} \quad s^2(\hat{y})=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\hat{y})^2 \tag{12}$$

Významnosť celého regresného modelu sa posudzuje pomocou celkového *F*-testu (14), pri hypotézach (13) a kritickom obore (15).

$$H_0: \quad \beta_0=k, \quad k\neq 0, \quad \beta_1=\beta_2=\cdots=\beta_j=0 \tag{13}$$

$$H_A: \quad \beta_j\neq 0 \quad \text{pre aspoň jedno } j = 1, 2,\ldots, p$$

Testovacím kritériom je štatistika:

$$F=\frac{\frac{S_T(y)}{c-1}}{\frac{S_R(y)}{n-c}}\sim F(c-1,n-c), \tag{14}$$

$$W_\alpha:|t_j|\geq t_{1-\frac{\alpha}{2}}(n-c) \tag{15}$$

pričom $c = p + 1$ je počet odhadovaných parametrov a W_α je kritický obor.

Pri hodnotení významnosti regresného modelu môžu podľa Herbáka (2005) nastať štyri rôzne situácie:

– *F*-test aj *t*-testy sú štatisticky významné, model sa považuje za vhodný, avšak nemusí to znamenať, že model je navrhnutý správne;
– *F*-test aj *t*-testy sú štatisticky nevýznamné, model sa považuje za nevhodný;
– *F*-test je štatisticky významný, ale niektoré *t*-testy sú štatisticky nevýznamné, model sa považuje za vhodný, avšak regresná funkcia je preparametrizovaná a vypúšťajú sa štatisticky nevýznamné regresné parametre, čím sa model stáva matematicky nelogickým;
– *F*-test je štatisticky významný, ale všetky *t*-testy sú nevýznamné, problém multikolinearity, formálne model vyhovuje, avšak ani jeden regresný parameter nie je štatisticky významný – lineárna závislosť medzi jednotlivými regresnými parametrami.

Vzťah medzi pozorovanými y_i a modelom predikovanými $\hat{y}_i$ hodnotami sa určuje viacerými spôsobmi. Jedným zo spôsobov je index determinácie R^2, v prípade lineárnej regresie nazývaný aj koeficient determinácie (16). Hodnota R^2 poukazuje na proporciu rozptylu odpovede (podiel vysvetlenej variability) (Katina, 2015).

$$R^2=\frac{S_T(y)}{S_Y(y)}, i_{yx}^2\in\langle 0,1\rangle \tag{16}$$

Čím viac sa hodnota R^2 blíži k 1, tým je skúmaná závislosť intenzívnejšia a opačne. Nižšia hodnota R^2 nemusí znamenať nízky stupeň závislosti medzi premennými, ale môže signalizovať iba chybnú voľbu regresnej funkcie.

Upravený index determinácie R_{adj}^2 (17) predstavuje upravenú hodnotu R^2 tak, aby bol tento index nezávislý na množstve použitých meraní (Katina, 2015).

$$R_{adj}^2=1-(1-i^2)\frac{n-1}{n-c} \tag{17}$$

Predpokladom každého regresného modelu podľa Katinu (2015) sú:

1. Stredná hodnota chybovej zložky je 0;
2. Správne špecifikovaný regresný model;
3. Chybová zložka má konštantný rozptyl;
4. Jednotlivé zložky chybového vektoru sú nekorelované;
5. Rozdelenie chýb je normálne.

Na základe vyššie uvedeného je zreteľné, že pri regresnej diagnostike zohrávajú dôležitú úlohu reziduály v absolútnej hodnote (18). Najefektívnejšou metódou pre regresnú diagnostiku s využitím týchto reziduálov je hodnotenie na základe grafov (Katina, 2015).

$$_{\text{abs}}r_i = |y_i - \hat{y}_i| \tag{18}$$

Podľa Katinu (2015) ide o tieto grafy:

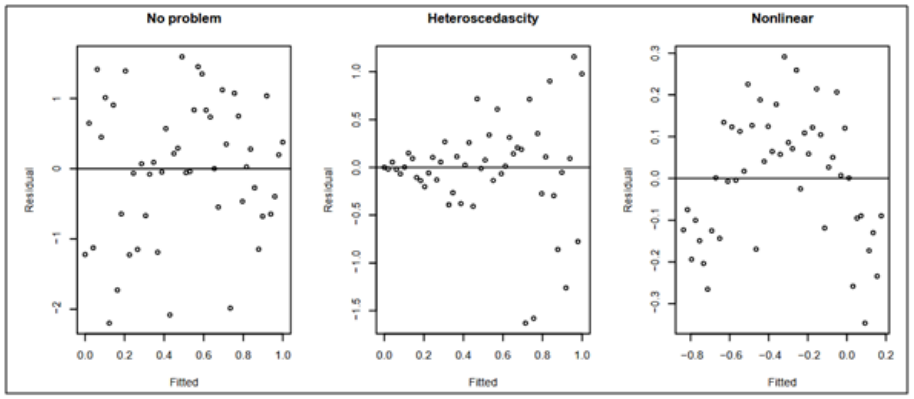
– Reziduály vs. fittované hodnoty – tvorba horizontálne kompaktného oblaku, používa sa na detekciu nelinearity, nerovností v odchýlkach;
– Reziduály vs. realizácie – systematická štruktúra indikuje nedodržanie konštantnosti rozptylu alebo nevhodnosť typu modelu;
– Q-Q graf reziduálov;
– Graf Cookových vzdialeností D_k pre každé pozorovanie – hodnoty D_k predstavujú vplyv k-teho pozorovania na odhadované regresné koeficienty. Definíciu týchto vzdialeností vyvinuli matematici Cook a Weisberg v 1982.

Najčastejšie využívaným grafom pre hodnotenie správnosti regresných modelov je graf reziduálov vs. fittovaných hodnôt. Tento graf zobrazuje modelom vypočítané (fittované) hodnoty podľa vzťahu (2) a hodnoty reziduálov vypočítané podľa vzťahu (18) v bodovom grafe závislostí (Massey University, 2010). Faraway (2009) označuje tento graf aj matematicky ako graf $\hat{e}$ vs. $\hat{y}$. Analyzovaný graf zobrazuje heteroskedasticitu (ak je rozptyl súboru je závislý na niektorom z parametrov, odchýlka bude nekonštantná) a nelinearitu (ak je detekovaná, naznačuje potrebné úpravy v modeli). Ak je všetko v danom grafe v poriadku, používateľ by mal byť schopný vizuálne hodnotiť konštantnosť odchýlky vo vertikálnom smere $(\hat{e})$ a rozptyl hodnôt by mal byť symetrický okolo 0 (Faraway, 2009).

Na príklade obr. 2 si môžeme všimnúť tri rôzne prípady, ktoré môžu nastať. Na základe obr. 2 vľavo môžeme povedať, že daný model je vhodný a nie je potrebný žiadny zásah užívateľa. Obr. 2 v strede vykazuje nekonštantný rozptyl a obr. 2 vpravo indikuje nelinearitu modelu, na základe ktorej by mal používateľ vykonať zmenu v štruktúre, napr. použiť iný typ funkcie (Faraway, 2009).

Podrobnejšie sa analýzou zmien, ktoré môže používateľ vykonať pri detekcii heteroskedasticity alebo nelinearity sa zaoberajú technologické príručky štatistických nástrojov alebo softvérov, ako napr. nástroj *Statial statistics toolbox* od Esri (2017) alebo softvér *R* Faraway (2002).

V rámci predkladanej práce ilustrujeme popísané princípy regresnej analýzy na príklade geografických údajov, ktoré sú súčasťou rozsiahlejšej štúdie pre analýzu poškodenia lesných porastov (Goga, 2017a, 2017b). V našom prípade údaje nezávislej premennej X predstavujú údaje z rastra (komponentu) optimalizovaného pre zvý-

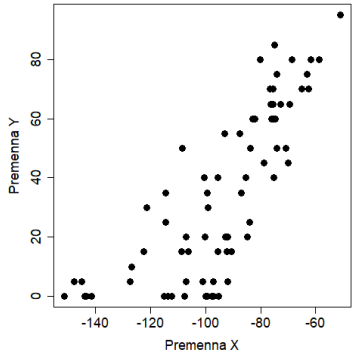


Obrázok 2 Grafy reziduálov vs. fittovaných hodnôt (Faraway, 2009)

raznenie poškodenia lesných porastov. Optimalizácia spočívala v ortogonálnej transformácii spektrálnych pásiem údajov leteckého snímkovania na základe *Gramm-Schmidt* transformácie (GST). Cieľom takejto ortogonálnej transformácie je nájsť optimálne transformačné koeficienty na klasifikáciu poškodenia pre zvolenú kombináciu kanálov. Matematický podklad pre túto transformáciu prehľadne spracoval Jackson (1983) a jeho upravenú verziu pre potreby odvodenia transformovaných komponentov pre odhad defoliácie porastov na úrovni Slovenska spracoval Bucha et al. (2014), resp. pre oblasť lužných lesov v oblasti VD Gabčíkova v práci Bucha a Slávik (2013).

Údaje závislej premennej *Y* predstavujú vizuálne interpretované údaje o defoliácii (odlistení) lesných porastov v zmysle výberového interpretačného kľúča vypracovaného Grossom (2000). Interpretácia prebehla na základe prípadovej štúdie *Fagus sylvatica*.

Takto špecifikované údaje predstavujúce premennú *X* a *Y* je potrebné prepojiť tak, aby predstavovali vstupné hodnoty pre výpočet regresného modelu. Párové hodnoty *X* a *Y* boli vizualizované v prostredí štatistického programovacieho jazyka R (obr. 3).



Obrázok 3 Párové hodnoty premennej *X* a premennej *Y*. Zdroj: vlastné výpočty

Výpočet aj všetky výstupy sú produkované v prostredí štatistického programovacieho jazyka R s využitím knižníc {stats}, {broom}, {minpack.lm}, {graphics}.

Všetky hypotézy vhodnosti regresných modelov verifikujeme na základe *t*-testov a *F*-testov na hladine *p*-hodnoty 0,05.

3 VÝSLEDKY

V rámci práce bolo vypočítaných celkovo 6 regresných modelov, z toho 2 z nich boli vytvorené pomocou nelineárnych funkcií a nástrojov. Postupne sa v tejto kapitole budeme zaoberať každým z nich, pričom poukážeme na vhodnosť / nevhodnosť použitého typu funkcie a vysvetlíme dôvody, prečo uvedený model prijímame ako vhodný, resp. zamietame ako nevhodný. Taktiež poukážeme na problematickú interpretáciu nelineárnych regresných modelov.

Priamková regresia

Na základe predpisu priamkovej regresnej funkcie bol odvodený regresný model (obr. 4). Štatistické parametre odvodeného priamkového regresného modelu sú uvedené v tab. 1. Všetky koeficienty (β) sú na základe *p*-hodnoty *t*-testu štatisticky významné. S využitím *p*-hodnoty *F*-testu považujeme regresný vzťah (19) za štatisticky významný

Daný model na základe koeficientu determinácie vysvetľuje približne 61 % rozdelenia. Interpretácia strednej chyby modelu vyjadruje, že výsledná hodnota premennej *Y* sa môže pohybovať v rozsahu ± 17,9 % od vypočítanej hodnoty *Y*. Koeficient korelácie $\left(\sqrt{r^2}\right)$ nadobúda hodnotu 0,83.

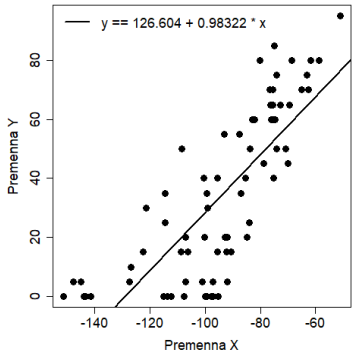
Napriek štatisticky významným parametrom z analyzovanej tabuľky model nepovažujeme za vhodne zvolený. Tento záver potvrdzuje analýza grafu reziduálnych vs. fittovaných hodnôt (obr. 5). Rozloženie výsledných reziduálov naznačuje, že daná priamková funkcia nie je vhodne zvolená. Z grafu predpokladáme, že model by mohla dostatočne vysvetľovať parabolická funkcia, keďže rozloženie mraku reziduálov má tvar paraboly.

$$y = 126{,}60 + 0{,}98322\,x \tag{19}$$

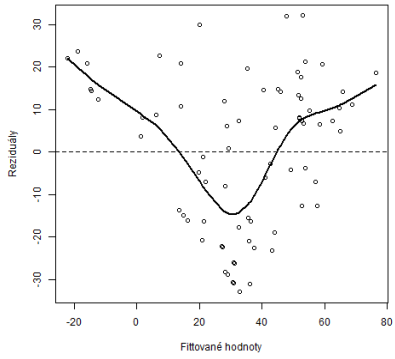
Tabuľka 1 Štatistické parametre priamkovej regresie

	názov	koeficienty	chyba strednej hodnoty	t-test	p-hodnota
PRIAMKOVÁ REGRESIA	β_0	126,60	8,80	14,39	0,00000
	β_1	0,98322	0,09	10,89	0,00000
	koeficient determinácie	upravený koeficient determinácie	stredná chyba regresnej funkcie	F	Významnosť F
	0,61242	0,60725	17,90	118,51	0,00000

Zdroj: vlastné výpočty



Obrázok 4 Priamková regresia – model. Zdroj: vlastné výpočty



Obrázok 5 Priamková regresia – graf reziduálnych vs. fittovaných hodnôt. Zdroj: vlastné výpočty

Parabolická regresia

Na základe predpisu parabolickej regresnej funkcie bol odvodený regresný model (obr. 6). Štatistické parametre odvodeného parabolického modelu sú uvedené v tab. 2. Všetky koeficienty (β) sú na základe p-hodnoty t-testu štatisticky významné. S využitím p-hodnoty F-testu považujeme regresný vzťah (20) za štatisticky významný.

$$y = 273{,}63 + 4{,}03862\,x + 0{,}01502\,x^2 \tag{20}$$

Tabuľka 2 Štatistické parametre parabolickej regresie

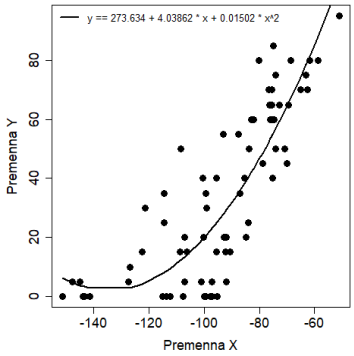
	názov	koeficienty	chyba strednej hodnoty	t-test	p-hodnota
PARABOLICKÁ REGRESIA	β_0	273,63	27,60	9,91	0,00000
	β_1	4,03862	0,56	7,24	0,00000
	β_2	0,01502	0,0027	5,53	0,00000
	koeficient determinácie	**upravený koeficient determinácie**	**stredná chyba regresnej funkcie**	**F**	**Významnosť F**
	0,72583	0,71842	15,16	97,95	0,00000

Zdroj: vlastné výpočty

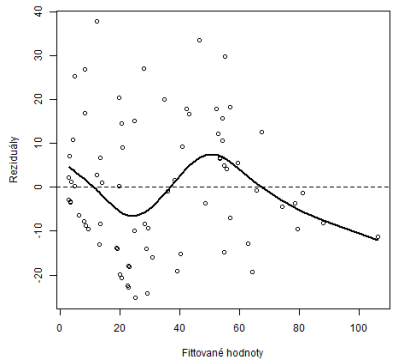
Daný model na základe koeficientu determinácie vysvetľuje približne 73 % rozdelenia. Koeficient korelácie $(\sqrt{r^2})$ na úrovni 0,83 naznačuje, že v rámci využitej regresnej funkcie je veľmi silný pozitívny korelačný vzťah. Pozitívny vzťah znamená, že s rastúcou hodnotou premennej X rastie tiež hodnota premennej Y. Interpretácia strednej hodnoty nám vraví, že výsledná hodnota premennej Y sa môže pohybovať v rozsahu ± 15,16 % od vypočítanej hodnoty Y.

Na základe analýzy grafu reziduálnych vs. fittovaných hodnôt (obr. 7) považujeme analyzovaný regresný model za vhodný, keďže mračno reziduálov vytvára približne kompaktný horizontálny tvar. Funkcia výsledných reziduálov síce naberá tvar

goniometrickej funkcie, avšak jej hodnoty sa pohybujú relatívne blízko nulovej línie, čo je z hľadiska štatistického vyhodnotenia markantný znak.



Obrázok 6 Parabolická regresia – model. Zdroj: vlastné výpočty



Obrázok 7 Parabolická regresia – graf reziduálnych vs. fittovaných hodnôt. Zdroj: vlastné výpočty

Polynomická regresia 3. stupňa

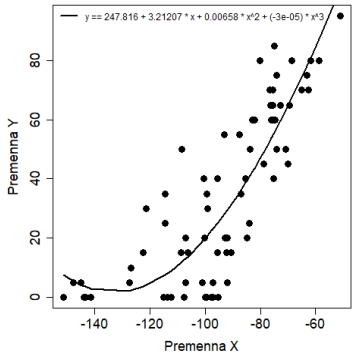
Na základe predpisu polynomickej regresnej funkcie 3. stupňa bol odvodený regresný model (obr. 8). Štatistické parametre odvodeného polynomického regresného modelu 3. stupňa sú uvedené v tab. 3. Na základe grafu fittovaných vs. reziduálnych hodnôt (obr. 9) považujeme model za vhodný, avšak viaceré koeficienty (β) sú na základe p-hodnoty t-testu štatisticky nevýznamné, v zmysle štatistickej metodiky preto dané koeficienty zamietame (21). Podľa p-hodnoty F-testu je model štatisticky významný, avšak vzhľadom na zamietanie viacerých koeficientov ovplyvňujúcich hodnotu nezávislej premennej považujeme model za nevhodný.

$$y = 247{,}82 + \sout{3{,}21207\,x} + \sout{0{,}00658\,x^2} - \sout{0{,}00003\,x^3} \tag{21}$$

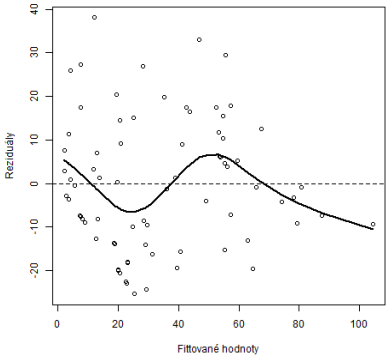
Tabuľka 3 Štatistické parametre polynomickej regresie 3. stupňa

POLYNOMICKÁ REGRESIA 3. STUPŇA	názov	koeficienty	chyba strednej hodnoty	t-test	p-hodnota
	β_0	247,82	107,32	2,31	0,02377
	β_1	3,21207	3,37	0,95	**~~0,34309~~**
	β_2	0,00658	0,034	0,19	**~~0,84700~~**
	β_3	-0,00003	0,00011	-0,25	**~~0,80403~~**
	koeficient determinácie	**upravený koeficient determinácie**	**stredná chyba regresnej funkcie**	**F**	**Významnosť F**
	0,72606	0,71480	15,25	64,49	0,00000

Zdroj: vlastné výpočty



Obrázok 8 Polynomická regresia 3. stupňa – model. Zdroj: vlastné výpočty



Obrázok 9 Polynomická regresia 3. stupňa – graf reziduálnych vs. fittovaných hodnôt. Zdroj: vlastné výpočty

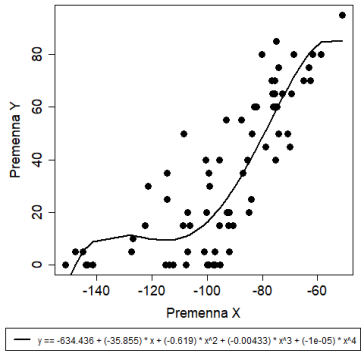
Matematický dôkaz na preukázanie nevhodnosti daného regresného modelu je možné potvrdiť, ak za nezávislú premennú X dosadíme hodnotu -100, pričom na základe odvodeného regresného modelu dosiahne hodnota premennej Y približne 20 %. Pri výpočte hodnoty premennej Y po zamietnutí štatisticky nevýznamných regresných koeficientov bude hodnota premennej Y odhadovaná na úrovni 247 %. Uvedený regresný model je preto po vylúčení zamietnutých regresných koeficientov z výpočtu matematicky nelogický.

Polynomická regresia 4. stupňa

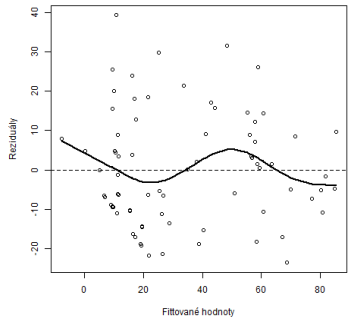
Na základe predpisu polynomickej regresnej funkcie 4. stupňa bol odvodený regresný model (obr. 10). Štatistické parametre odvodeného polynomického regresného modelu 4. stupňa sú uvedené v tab. 4. Funkcia reziduálov na grafe fittovaných vs. reziduálnych hodnôt (obr. 11) vykazuje takmer ideálny tvar. Je však potrebné konštatovať, že na základe p-hodnoty t-testu je regresná konštanta β_0 štatisticky nevýznamná, v zmysle štatistickej metodiky preto danú regresnú konštantu zamietame (22). Napriek zamietnutiu tejto konštanty považujeme podľa p-hodnoty F-testu model za štatisticky významný, avšak na základe opísaných skutočností považujeme model za nevhodný.

Matematický dôkaz na preukázanie nevhodnosti daného regresného modelu je možné potvrdiť, ak za nezávislú premennú X je dosadená hodnota –100, pričom premenná Y na základe odvodeného regresného modelu bude dosahovať hodnotu približne 16 %. Pri výpočte hodnoty premennej Y po zamietnutí uvedenej regresnej konštanty je jej hodnota odhadovaná na úrovni 654 %. Ďalším logickým argumentom pre zamietnutie uvedeného regresného modelu je fakt, že funkcia po dosiahnutí lokálnych maxím pri hodnotách premennej $X = -122$, resp. –54 naberá klesajúci charakter. Uvedený regresný model je preto z matematického a predikčného hľadiska nelogický.

$$y = -\cancel{634,44} - 35,85524\,x - 0,61945\,x^2 - 0,00433\,x^3 - 0,00001\,x^4 \qquad (22)$$



Obrázok 10 Polynomická regresia 4. stupňa – model. Zdroj: vlastné výpočty



Obrázok 11 Polynomická regresia 4. stupňa – graf reziduálnych vs. fittovaných hodnôt. Zdroj: vlastné výpočty

Tabuľka 4 Štatistické parametre polynomickej regresie 4. stupňa

	názov	koeficienty	chyba strednej hodnoty	t-test	p-hodnota
POLYNOMICKÁ REGRESIA 4. STUPŇA	β_0	-634,44	345,86	-1,83	~~0,07073~~
	β_1	-35,85524	14,97	-2,39	0,01924
	β_2	-0,61945	0,24	-2,62	0,01075
	β_3	-0,00433	0,00161	-2,68	0,00903
	β_4	-0,00001	0,000004	-2,67	0,00931
	koeficient determinácie	**upravený koeficient determinácie**	**stredná chyba regresnej funkcie**	**F**	**Významnosť F**
	0,75078	0,73694	14,65	54,23	0,00000

Zdroj: vlastné výpočty

Hyperbolická a exponenciálna regresia

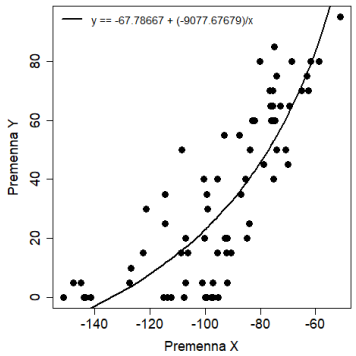
Na základe predpisu hyperbolickej regresnej funkcie bol odvodený hyperbolický regresný model (obr. 12). Štatistické parametre hyperbolického regresného modelu (23) sú uvedené v tab. 5. Podobným spôsobom bol odvodený exponenciálny regresný model (obr. 13). Štatistické parametre exponenciálneho regresného modelu (24) sú uvedené v tab. 6.

$$y = -67{,}79 - \frac{9077{,}68}{x} \tag{23}$$

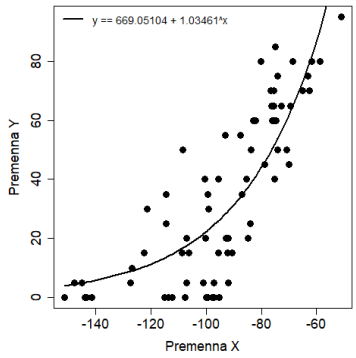
$$y = 669{,}05 * 1{,}03461^{x} \tag{24}$$

Pri interpretácii štatistických parametrov týchto regresných modelov je potrebné byť veľmi opatrný. Keďže v zmysle paradigmy pri štatistickej analýze ide o nelineárny regresný model, hodnoty indexu determinácie sa veľmi zložito interpretujú (Cameron, 1995; Spiess, 2010). Aplikácia výpočtu indexu determinácie môže podľa

uvedených autorov viesť k meraniu, ktoré bude ležať mimo požadovaného intervalu <0, 1>. Pre konkrétne nelineárne modely boli preto v minulosti vytvorené alternatívne súhrnné štatistické merania podobného typu ako index determinácie. Zaoberali sa tým viacerí autori, napr. Maddala (1983), Windmeijer (1995), Laitila (1993) a iní. V praxi sa najčastejšie využívajú metódy aproximácie indexu determinácie (McKelvey a Zavoina, 1975), alebo metódy AICc, prípadne BIC, ktoré bližšie popisuje Spiess (2010) a v rámci svojej práce poskytuje k danej téme aj množstvo veľmi rozsiahlych matematických, resp. štatistických prác (Akaike, 1973; Schwarz, 1978).



Obrázok 12 Hyperbolická regresia – model. Zdroj: vlastné výpočty



Obrázok 13 Exponenciálna regresia – model. Zdroj: vlastné výpočty

Tabuľka 5 Štatistické parametre hyperbolickej regresie

	názov	koeficienty	chyba strednej hodnoty	t-test	p-hodnota
HYPERBOLICKÁ REGRESIA	β_0	-67,79	7,82	-8,67	0,00000
	β_1	-9077,68	682,50	-13,30	0,00000
	koeficient determinácie		**stredná chyba regresnej funkcie**		
	0,70227		15,69		

Zdroj: vlastné výpočty

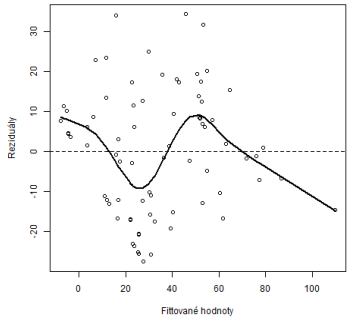
Tabuľka 6 Štatistické parametre exponenciálnej regresie

	názov	koeficienty	chyba strednej hodnoty	t-test	p-hodnota
EXPONENCIÁLNA REGRESIA	β_0	669,05	152,66	4,38	0,00004
	β_1	1,03461	0,003	324,93	0,00000
	koeficient determinácie		**stredná chyba regresnej funkcie**		
	0,70546		15,67		

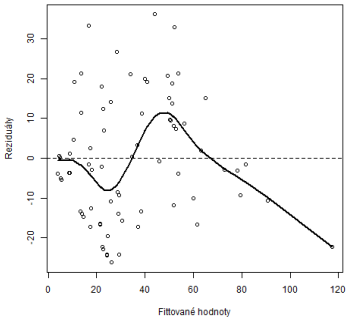
Zdroj: vlastné výpočty

V našom prípade bola hodnota indexu determinácie vypočítaná pomocou predikcie modelu s využitím metódy najmenších štvorcov. avšak je potrebné zdôrazniť, že vypočítané hodnoty najmenších štvorcov sa na základe vzťahu (9) nerovnajú. Táto hodnota je preto len informatívna a podľa Camerona (1995), resp. Spiessa (2010) je potrebné celkové vyhodnotenie hyperbolického, resp. exponenciálneho regresného modelu vykonávať aproximáciou indexu determinácie, resp. s využitím vyššie opísaných vhodných metód. Ak uvedené postupy neaplikujeme, je potrebné v zmysle Camerona (1995) a Spiessa (2010) vykonať vizuálne hodnotenie, na základe strednej chyby regresnej funkcie a grafu reziduálnych vs. fittovaných hodnôt (obr. 14, resp. obr. 15).

Na základe uvedených faktov hodnotíme analyzované regresné modely ako štatisticky a predikčne vhodné, keďže grafy reziduálnych vs. fittovaných hodnôt vytvárajú relatívne kompaktné mračná, pričom nedetekujeme heteroskedasticitu ani nelinearitu (mierne horšie je to u exponenciálnej regresnej funkcie) a zobrazené funkcie na základe stredných chýb regresných modelov relatívne kvalitne vysvetľujú rozdelenie sledovaných hodnôt.



Obrázok 14 Hyperbolická regresia – graf reziduálnych vs. fittovaných hodnôt. Zdroj: vlastné výpočty



Obrázok 15 Exponenciálna regresia – graf reziduálnych vs. fittovaných hodnôt. Zdroj: vlastné výpočty

4 DISKUSIA A ZÁVER

V príspevku sme prezentovali možnosti štatistickej analýzy dát, ktoré sme podrobili dôkladnej analýze. Jednotlivé použité regresné modely poskytli užívateľovi precízny prehľad o vzťahoch, ktoré sa medzi použitými dátami nachádzajú. Zároveň mu umožňujú predikovať, ako by sa mohol údajový model ďalej rozvíjať. Regresná analýza je preto veľmi často využívaná v geografickom modelovaní (vývoj počtu obyvateľstva, modelovanie pribúdania biomasy, predikcia rastu priemernej teploty a iné).

V predchádzajúcej kapitole si môže čitateľ všimnúť všetky základné štatistické problémy, ktoré sa v prípadnej štatistickej analýze môžu objaviť. Či už ide o nesprávnu distribúciu výsledných reziduálov, zamietanie regresného modelu na zá-

klade *p*-hodnôt *t*-testov alebo problematiku klesajúcej regresnej funkcie. Zároveň poukazuje na problematiku nelineárnych regresných modelov. Práve z tohto dôvodu je potrebné klásť dôraz na problematiku linearizácie regresných modelov, ktorá umožní užívateľovi využívať bežne používané štatistické hodnotenie aj nelineárnych regresných modelov.

Na základe vyššie uvedeného by sme pre prípadné modelovanie špecifického geografického problému defoliácie z optimalizovaného spektrálneho komponentu vybrali ako najvhodnejší práve regresný model práve parabolický regresný model. Uvedený model je lineárny a preto sa dá spoľahlivo vyhodnocovať s využitím základných štatistických parametrov. Ak by sme použité nelineárne regresné modely linearizovali a poskytli by rovnaké alebo lepšie štatistické výsledky ako parabolický model, tak by výsledný rozhodovací proces pre použitie toho správneho regresného modelu závisel práve od grafu distribúcie výsledných fittovaných reziduálov alebo analýzou systematickosti štruktúry realizácií uvedených reziduálov.

Uvedený príspevok poskytuje základný prehľad štatistickej regresnej analýzy a otvára priestor vedeckej obci pre rozšírenie poznatkov v sledovanej problematike. Zároveň práca poskytuje základný záchytný bod študentom alebo vedeckým pracovníkom pre základný, rozširujúci alebo inak špecifický výskum.

Literatúra

AKAIKE, H. 1973. Information theory and an extension of the maximum likelihood principle. In: *Proceedings of the International Symposium on Information Theory*, 2nd. Akademiai Kiado, Budapest, , pp. 267-281.

BITTERER, L. 2003. *Vyrovnávací počet II*. Vysokoškolská učebnica. Žilina: Žilinská univerzita.

BUCHA, T., et al. 2014. *Satelity v službách lesa*. Zvolen: SAP - Slovak Academic Press.

BUCHA, T., SLÁVIK, M. 2013. Improved methods of classification of multispectral aerial photographs: evaluation of floodplain forests in the inundation area of the Danube. *Folia Forestalia Polonica*, 55(2), pp. 58-71.

CAMERON, A. C, WINDMEIJER, F. A. G. 1995. *An R-squared measure of goodness of fit for some common nonlinear regression models*. [online] [citované 2017-04-24]. Dostupné na: <http://old.econ.ucdavis.edu/faculty/cameron/research/je97preprint.pdf>

ESRI. 2017. *Modeling Spatial Relationships toolset concepts*. [online] [citované 2017-04-24]. Dostupné na: <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/regression-analysis-basics.htm>

FARAWAY, J. J. 2002. *Practical Regression and Anova using R*. [online] [citované 2017-04-24]. Dostupné na: <https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>

FARAWAY, J. J. 2009. *Linear models with R*. New York: Taylor & Francis.

GOGA, T. 2017a. *Hodnotenie zdravotného stavu lesných porastov na podklade leteckých multispektrálnych snímok a lidarových údajov*. Diplomová práca. Bratislava: Univerzita Komenského v Bratislave, Prírodovedecká fakulta, Katedra kartografie, geoinformatiky a diaľkového prieskumu Zeme.

GOGA, T. 2017b. *Porovnanie pixelovo-orientovaného prístupu a objektovo-orientovaného prístupu pri tvorbe masky lesa s využitím leteckých multispektrálnych snímok a lidarových údajov*. In: Študentská vedecká konferencia PriF UK 2017, Zborník recenzovaných príspevkov, Bratislava, s. 1384-1389.

HERBÁK, P., HUSTOPECKÝ, J., MALÁ, I. 2005. *Vícerozmerné statistické metody 2*. Praha: Informatorium.

HINDLS, R., HRONOVÁ, S., NOVÁK, I. 1999. *Analýza dat v manažerském pojetí*. Praha: Grada Publishing.
JACKSON, R. D. 1983. Spectral indices in N-space. *Remote Sensing of Enviroment*, 13(5), pp. 409-421.
KATINA, S. 2015. *Aplikovaná štatistická inferencia*. Brno: MUNI Press.
LAITILA, T. 1993. A pseudo-R2 measure for limited and qualitative dependent variable models. *Journal of Econometrics*, 56, 3, pp. 341-355.
MADDALA, G. S. 1983. *Limited dependent and qualitative variables in econometrics*. Cambridge University Press: Cambridge.
MASSEY UNIVERSITY. 2010. *Fitted values and residuals*. [online] [citované 2017-04-24]. Dostupné na: <http://www-ist.massey.ac.nz/dstirlin/CAST/CAST/HleastSqrs/leastSqrs_c 3.html>
MCKELVEY, R. D., ZAVOINA, W. 1975. Statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, pp. 103-120.
NEUBAUER, J. 2011. *Regresní a korelacní analýza*. Studijné materiály. Brno: FEM OU.
SCHWARZ, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, 6, 2, pp. 461-464.
SPIESS, A. N., NEUMEYER, N. 2010. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacology*, 10(6).
WINDMEIJER, F. A. G. 1995. Goodness-of-fit measures in binary choice models. *Econometric Reviews*, 14, pp. 101-116.

Mathematical principles of regression analysis: a geographical perspective

Summary

This paper presents the possibilities of statistical data analysis. These data are thoroughly analysed. The used regression models have provided a clear overview for the user about the relationships that exists between data. They also allow him to predict how the model could be further developed. Regression analysis is very often used for geographical modelling (population development, modelling of biomass growth, prediction of temperature growth and others).

The reader may notice all the basic statistical problems that may arise during statistical analysis. We analyse misallocation of the resulting residuals, rejection of the regression model based on the *p*-values of the *t*-tests or the problem of decreasing of regression function. We also point to nonlinear regression models. Especially for this reason it is necessary to emphasize the problem of linearization of regression models. This aspect could allow user to use basically used statistical evaluations for nonlinear models.

Based on the analysis of the results we could chose the parabolic regression model as the most advantageous regression model for possible subsequent modeling of a specific geographic problem. This model is linear therefore could be reliably evaluated using basic statistical parameters. If nonlinear regression models have been linearizated they could provide the same or better results as parabolic model. Afterward the final decision-making process for correct regression model usage must be done by distribution diagram of fitted residuals or by the analysis of the systematical structure of the residuals realization.

Nevertheless, this paper provides a basic overview of the statistical regression analysis and opens up space for the scientific community to expand their knowledge in the subject matter. At the same time, work provides a basic clue for students, or scientists for basic, expanding or otherwise specific research.