Dataset - vzorky

Z dát sme odstránili tie vzorky, ktoré nemali informáciu o objeme. RMZ sme zaokrúhli nahor, aby sme každý fragment počítal ako jeden druh.

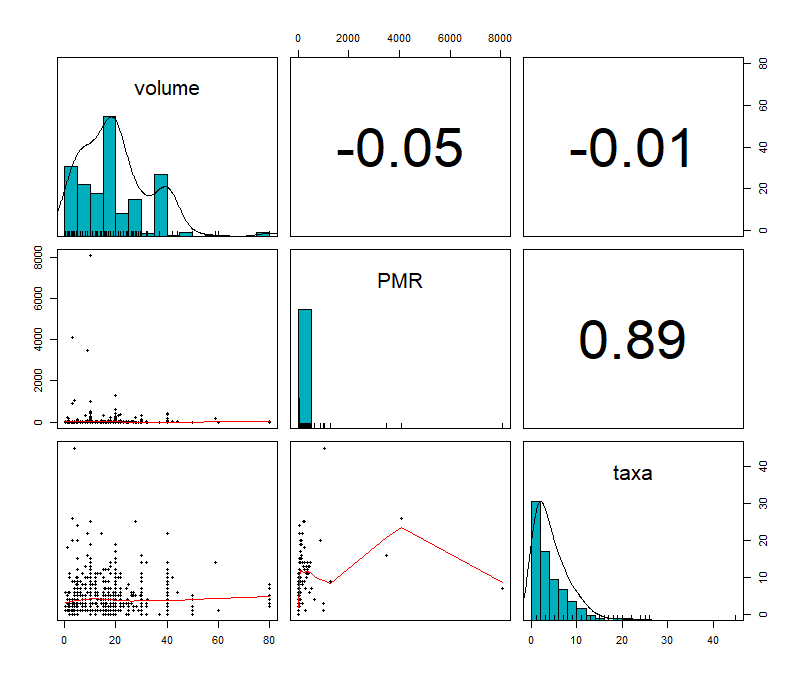
Na úvod sa pozrime základné štatistické vlastnosti dát[[1]](#footnote-1). Spolu sa odobralo 13757,78 litrov sedimentu v 686 vzorkách,  v ktorých sa identifikovalo 36723 RMZ a XY rastlinných druhov.

V ideálnom prípade predchádza samotnú analýzu teoretická príprava a formulácia dôležitých otázok, ktoré následne ovplyvňujú zber dát a vzorkovanie. Z archeobotanického hľadiska je však situácia zložitejšia, resp. opačná. V mnohých prípadoch archeobotanik pracuje s tým, čo už niekto pred ním nejakým spôsobom ovzorkoval a zozbieral a až potom môže skúmať dáta a formulovať otázky. To ho kladie pre náročnú úlohu ako čo najobjektívnešie interpretovať často nereprezentatíve (a subjektívne zozbierané) dáta. S takýmto druhom dát pracujeme aj v tomto článku, kde sme sa v rámci možností pokúsili vyťažiť čo najviac informácii z dostupných dát a začleniť ich do archeologického/archeobotanického kontextu. **Tu našu hypotézu**

Tabuľka 1 sem

Vysokú hodnoty šikmosti a špicatosti dokladajú aj údaje v stĺpcoch *skew.2SE* a *kurt.2SE*. Tie predstavujú šikmosť a špicatosť, vydelenú dvoma štandardnými odchýlkami. Ak je ich absolútna hodnota vyššia ako 1, tak sú na hladine významnosti 0,5%  štatisticky významné (Field/Miles/Field 2012, 17). Ako je vidieť, hodnoty sú vysoko nad 1 a teda šikmosť aj špicatosť jasne dominujú. Ďalšími indikátormi sú vysoké rozdiely medzi priemermi a rozptylmi, resp.štandardnými odchýlkami (var a std.dev stĺpce). Predovšetkým v prípade RMZ tento rozdiel ukazuje na tzv. nadmerný rozptyl (*overdispersion*) – častý jav najmä v ekologických a enviromentánlych štúdiách a môže mať rôzne príčiny (O’Hara/Kotze 2010; Zuur et al. 2009, 270-271; Martin et al. 2005; súhrne Hilbe 2011, kap. 4). Takisto hodnota *Shapiro-Wilk*testu (Shapiro-Wilk stĺpec) a jej p-value potvrdzujú „ne-normálnu“ distribúciu pre objem a RMZ. Pre lepšiu si môžeme dáta vizualizovať.

Vizuálne si môžeme prezrieť



Graf 1

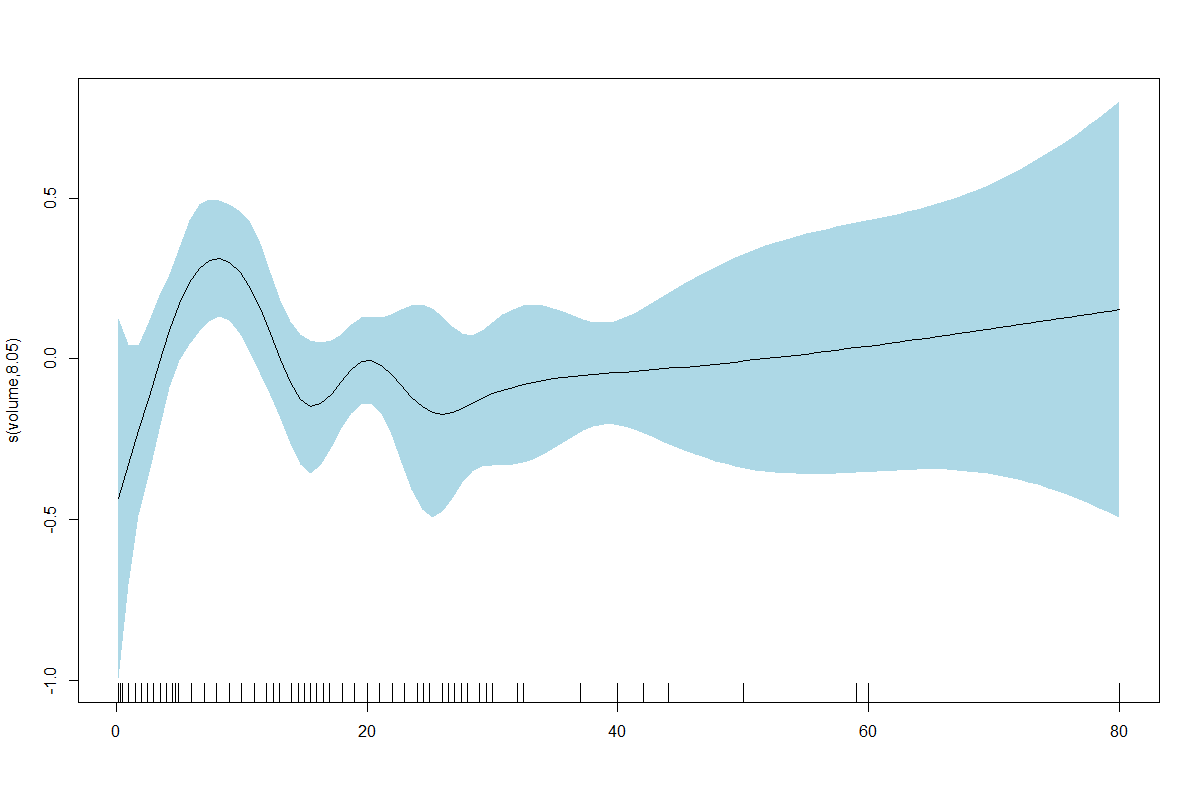
Graf 1 je vypočítaný zo surových dát a je možné na ňom vidieť hneď niekoľko vzájomných vzťahov medzi skúmanými premennými. Už na prvý pohľad je zrejmé, že normálnu distribúciu by sme tu hľadali márne. Diagonálne sú zobrazené histogrami a ich distribúcia/rozdelenie (pomocu jádroveho odhadu hustot – kernel density estimation, napr. Baxter 2015, 37-49). Objem má črty bimodálnej distribúcie, zatiaľ čo druhy sú viditeľne zošikmené doľava. Najextrémnejšie sú v tomto prípdade PMR, keďže na jednej strane obsahujú veľa nulových hodnôt, na druhej strane majú viacero extrémne vysokých hodnôt (pozri tab. 1), ktoré doslova natiahli osu x a teda akoby „stlačili“ hodnoty do jedného koša (bin)[[2]](#footnote-2). V ľavej dolnej časti sú zobrazené scatterploty spolu s trendovou spojnicou (LOESS krivka, k tomu napr. Zuur et al. 2009, 39-42). Medzi objemom a RMZ a objemom a druhmi nie je pozorovaná takmer žiadna závislosť. Naproti tomu medzi PMR a druhmi sa javí vysoká závislosť, ale výsledok podstatne skresľujú odľahlé hodnoty. V hornej časti sú hodnoty miery asociácie vypočítané pomocou Spearmanovho korelačného koeficientu[[3]](#footnote-3). Vysoká pozitívna korelácia je medzi PMR a druhmi, čo sa zdá byť na prvý pohľad logické, a teda, keď stúpa počet extraktovných RMZ, stúpa aj počet identifikovaných druhov. Samozrejme toto tvrdenie nie je kauzálne podmienené. Čo je však zaujímavé je v podstate nulová korelácia medzi objemom a RMZ/druhmi, čo korešponduje aj s výsledkami v scatterplotoch. To by naznačovalo, že množstvo odobraného sedimentu nezávisí s počtom nájdených RMZ/druhov. Toto tvrdenie sa javí na prvý pohľad ako kontraintuitívne a preskúmame ho neskôr.

V nasledujúcich riadkoch sa podrobnejšie pozrieme na vyššie popísane javy, nájsť ich príčiny a pokúsime sa pomocou štatistický modelouv zistiť, nakoľko sú tieto javy spracovateľné a reprezentatívne.

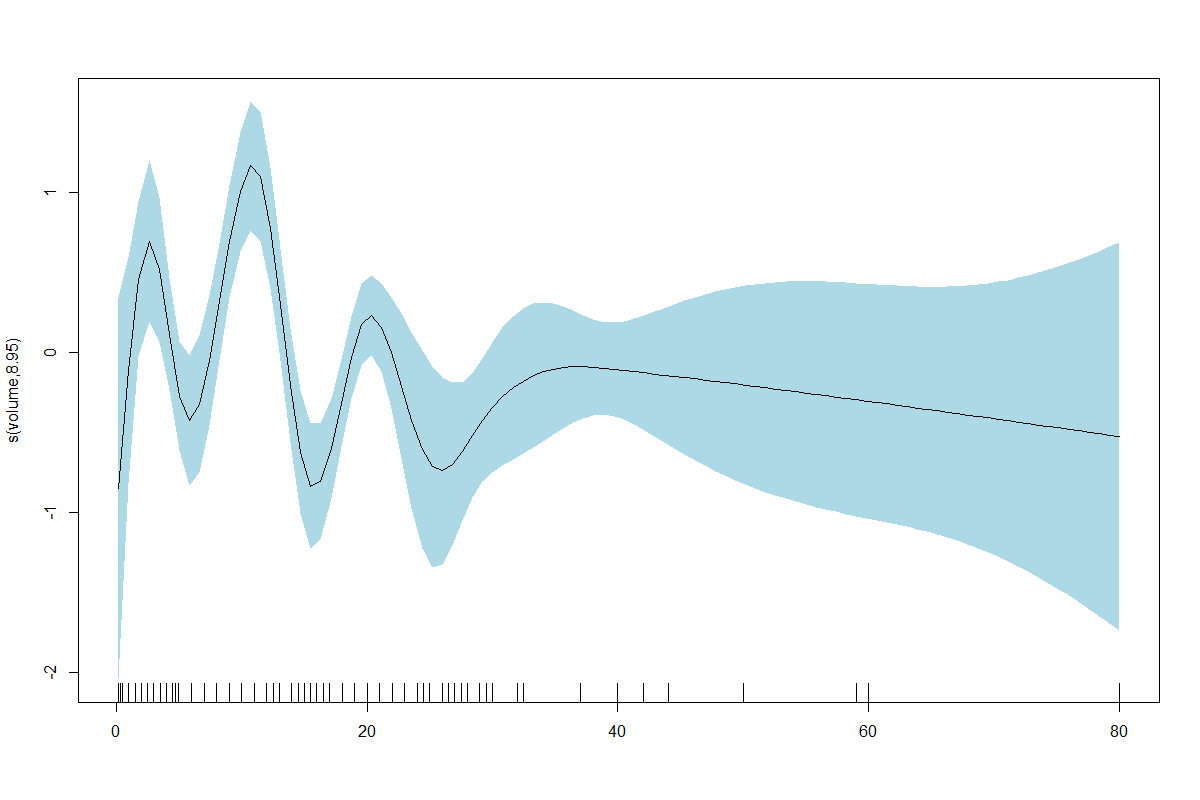
V ďalšom kroku bude potrebné nájsť matematický model (formu), ktorý by najlepšie reprezentoval spracovávané dáta, aby mohli byť ďalej štatisticky vyhodnotené (Orton 2000, 9-10; Orton 1980, 20-21). Keďže skúmané premenné (response variables) nespĺňali predpoklady normálnej distribúcie, využili sme tzv. negatívne-binomálne rozdelenie (*negative binomial distribution*), ktoré sa najviac približovalo naším dátam. Ide vlastne o rozšírenie Poissonovho rozdelenia a obsahuje extra parameter pre nadmerný rozptyl (Hilbe 2011, 253) a často sa použiva  ekologických  enviromentálnych štúdiách (napr. Zuur et al. 2009, 225-230).

Pri modelovaní sme použili tzv. Generalized additive models (GAM)[[4]](#footnote-4), ktoré sú vlastne rošírené Generalized linear models (GLM), avšak nepredpokladajú normálnu distribúciu ani linearitu, resp. parametrickú formu dát (Crawley 2007, 611). Hlavný rozdiel oproti lineárnym modelom je, že GAM namiesto *β* parametrov využívajú tzv. neparametrické smooth funkcie v zjednodušenej forme Yi = α + f (Xi) + εi kde εi ∼ N(0,σ 2), pričom Yi je závislá (response) premenná a f (Xi) je smooth funkcia (krivka) pre vysvetľujúcu premmenú (explanatory) – lepšie napísať. Čitateľa odkazujeme najmä na práce od S. Wood (2017),  Alain. F. Zuur et. al. (200?), M.J Crawley (2007) a takisto množstvo užitočných informácii je možné dohľadať na internete (<https://rdrr.io/cran/mgcv/> - toto pd čiaru)

V tejto časti predstavíme niekoľko modelov a stručne ich zanalyzujeme. Do analýzy zahrnieme aj modely, ktoré síce nie sú veľmí štatistický významé (ak vôbec), ale dobre ilustrujú spravovávaný súbor a postup vyhodnocovania. Chceme upozorniť čitateľa, že sa tu obmedzíme predovšetkým na štatistický popis, ktorý bude až neskôr doplnený archeologickú/archeobotanickú interpretáciu. Tak isto netreba zabúdať na kvalitu dát (pozri vyššie). Pred štatistickým modelovaním sme odstránili 3 vzorky, ktoré okrem toho, že predstavovali najextrémnejšie hodnoty, pochádzali z hromadných nálezov.

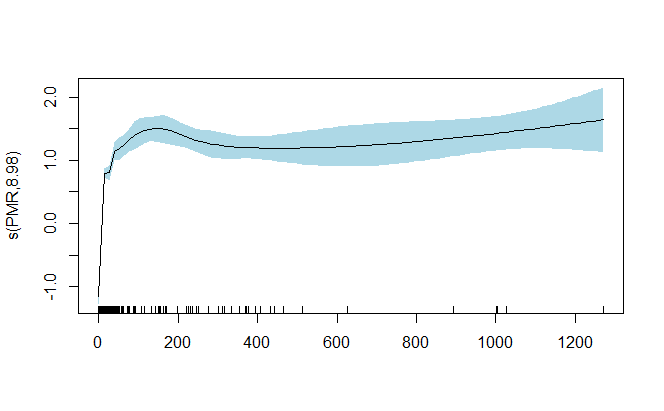
Model 1

Model 1 zobrazuje vzťah (efekt) medzi objemom a počtom rastlinných druhom z hľadiska celého súboru. Podľa grafu (graf) a údajov v tabuľke (tab, m1) nie je síce štatisticky signifikantný, každopádne vykazuje mierne znaky nelinearity a naznačuje, že objem má určitý vplyv na množstvo nájdených rastlinných druhov. Predovšetkým v prípade vzoriek o objeme okolo 10 a 20 litrov. Smerom k väčším vzorkám sa však efekt stráca a navyše výrazne stúpa miera neurčitosti (uncertainity), ktorá naznačuje, že odhad funkcie už nie je pri väčších vzorkách spoľahlivý (modrá zóna – napísať do popisku grafu, že ide o 95% inervaly spoľahlivost – confidence interval). Samotný model pritom vysvetľuje len 2,2% variability a jeho celková vypovedacia hodnota je len orientačná.

Model 2

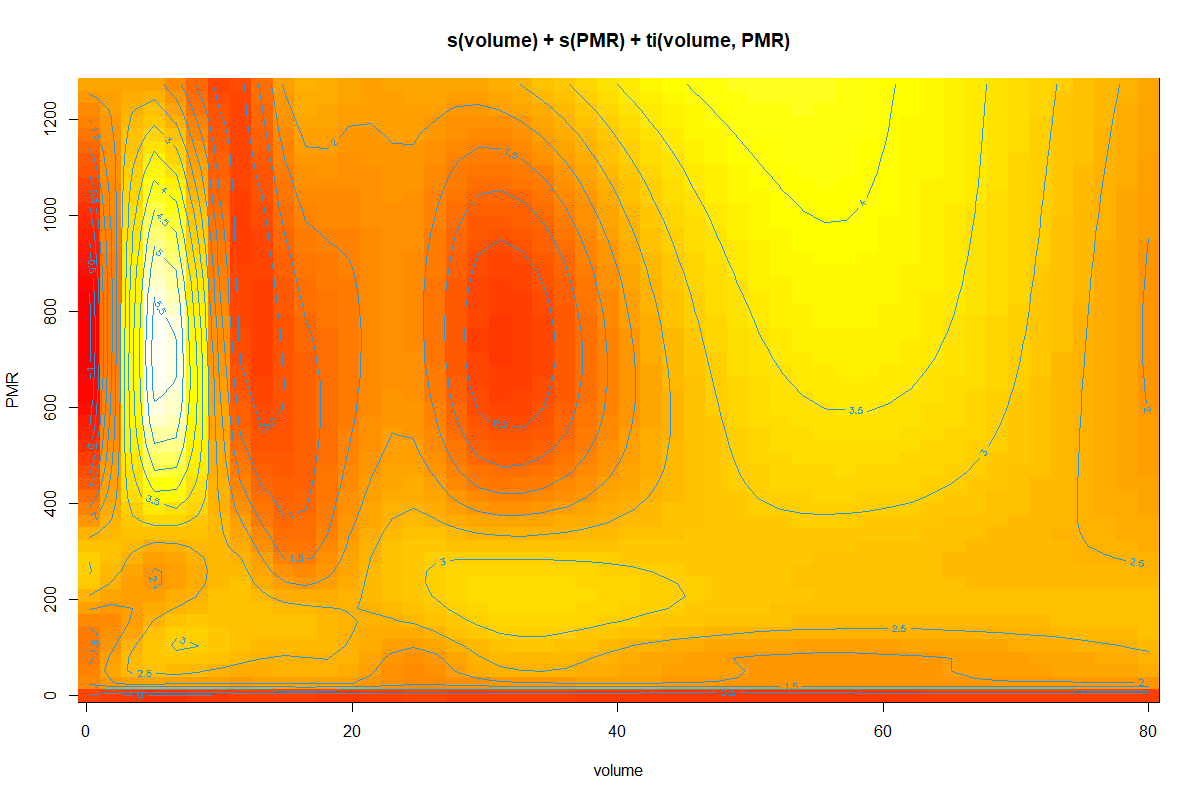
Model 2 zozbrazuje vzťah (efekt) medzi objemom a PMR. Vidieť tu jednoznačný efekt objemu na množstvo makrozvyškov . Opäť sa to prejavuje najmä pri vzorkách o objeme približne 10 a 20 litrov, plus tentokrát majú výraznejší efekt aj malé vzorky okolo 2l. Pri väčších vzorkách sa efekt a ako aj spoľahlivosť odhadu funkcie sráca rovnako ako v modeli 1. V porovnaním s modelom 1 je však tento model o niečo horší (AIC hodnoty v tab). Navyše je tu badateľný značný nadmerný rozptyl (tab. – dispersion), ktorý by mal v našom prípade dosiahnuť hodnotu maximálne 1.25 (Hilbe 2014, 82, 133). Existuje viacero dôvodov prečo sa nadmerný rozptyl objavuje (Hilbe, 2014, 82), V našom prípade môžu zohrávať úlohu chýbajúce prediktory, odľahlé hodnoty či väčšie množstvo núl (zero-inflation), ale nevylučujeme aj ďalšie dôvody. PMR napriek tomu vysvetľujú o trochu viac variability (7,7%).

Model 3



Model 3 predstavuje vzťah (efekt) medzi počtom druhov a počtom PMR a jednoznačne dokladá ich vzájomný vzťah (graf + tab) a takisto AIC hodnota indikuje , že ide o niečo kvalitnejší (better-fitted model). Navyše vysvetľuje približne 74% variability, čo je značný skok oproti predchádzajúcim dvom modelom. Model ukazuje mierne stúpajúci trend, indikujúci, že s počtom extraktovaných PMR stúpa aj počet nájdených druhov, pričom najväčšiu zmenu je možné pozorovať v intervale približne od 100 do 200 PMR. Oproti modelu dva má však nedostatočný rozptyl (underdispersion), čo pravdepodobne mohli ovplyvniť extrémne hodnoty (Zuur et al. 2009, 225). Prezentovaný výsledok – teda silnejší vzťah medzi druhmi a PMR - však nie je prekvapujúci a na čo poukazovali už vyššie načrtnuté analýzy.

Model 4

Do modelu 4 sme zahrnuli objem a PMR spolu pomocu tzv. product tensor smooth interaction funkcie (Wood 2017, 187, 326). Tzn. že sme sa pozreli na “čistý“ vzájomný vzťah (efekt) medzi objemom/PMR a počtom druhov. Grafický výstup je v podobe tvz. heat mapy. Výsledok korešponduje s modelmi vyššie a naznačuje, že spoločný efekt sa prejavuje predovšetkým pri vzorkách o objemoch približne 2 l, 10 -20 l, 25-45 l a s počtom makrozvyškom cca 400 a vyššie. Spoločná interakcia objemu a makrozvyškov priniesla aj väčšie percento vysvetlenej variablity a takisto o niečo kvalitnejší model (tab.) i keď odľahlé hodnoty trochu poddimenzovali rozptyl. Bežne sa pri analýze archeobotanických súborov pracuje s hustorou (pomer PMR/objem). Pri modelovaní s hustotou (tab. m5) sme dostali veľmi podobné, i keď o trochu slabšie, výsledky ako pri modele 4. Kvôli tomu a tiež kvôli zachovanie určitej jednotnosti s hustotu nebudeme naďalej pracovať, namiesto toho využijeme spomínaný product tensor ako indikátor spoločnej interakcie, ktorý sa ukazuje ako lepšia varianta.

Úmyslom bolo pozrieť sa, či sa tieto prediktory navzájom neovplyvňujú pri modelovaní.

Model 4

Model 3 (tab – m3) je výsledok dvoch smooth funkcií pre objem a PMR. Graficky

Hilbe (2014, 79-80) – P\_\_dist funkcia

Hilbe (2014, 136) – under dispersion – 210 – cela kap. 8

underdispersion, xii, 11, 12, 33, 48, 83, 124,  
136, 156, 212–214, 221, 245

V našom prípade nemôžeme tvrdiť, že výsledky skúmaného súboru sú všeobecne aplikovateľné, ale platia len na daný súbor-

1. Psych package, stat\_desc() [↑](#footnote-ref-1)
2. Ak by sme na základe týchto dát spravili jednoduchý lineárny model, zistili by sme, že tieto dáta neslňajú ani ďalšie predpoklady ako homogenita a sčasti aj nezávislosť [↑](#footnote-ref-2)
3. veľmi podobné výsledky sme dosiahli aj pri použití Kednall’s tau koeficientu. [↑](#footnote-ref-3)
4. Za použitia balíka (package mgcv) [↑](#footnote-ref-4)