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.

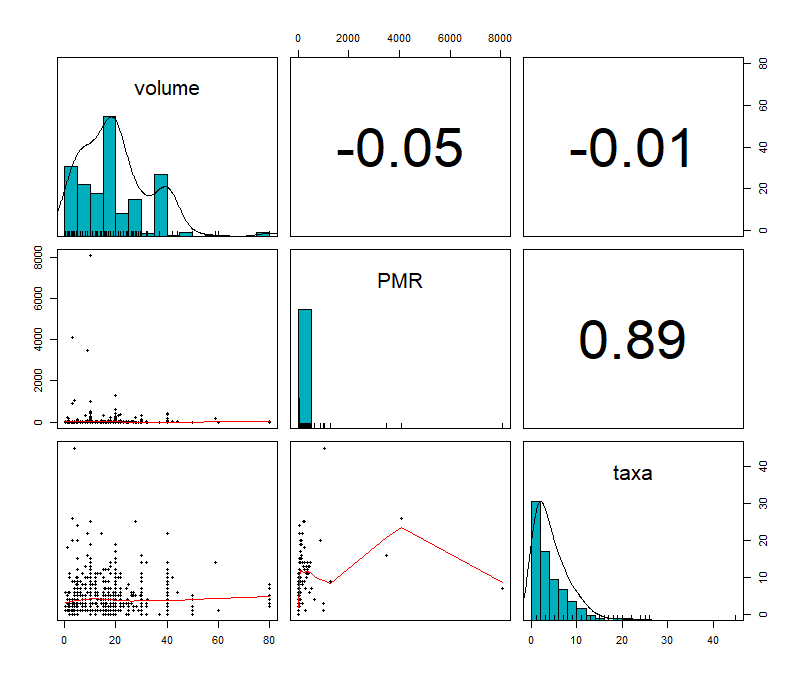
V ideálnom prípade predchádza samotnú analýzu teoretická príprava a formulácia dôležitých otázok, ktoré následne ovplyvňujú zber dát a vzorkovanie. Z archeobotanického hľadiska je však situácia zložitejšia, resp. opačná. V mnohých prípadoch však archeobotanik pracuje s tým, čo už niekto pred ním nejakým 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

Pmocou deskriptívnej štatiskty môže odhaliť základné vlastnosti dát.

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.

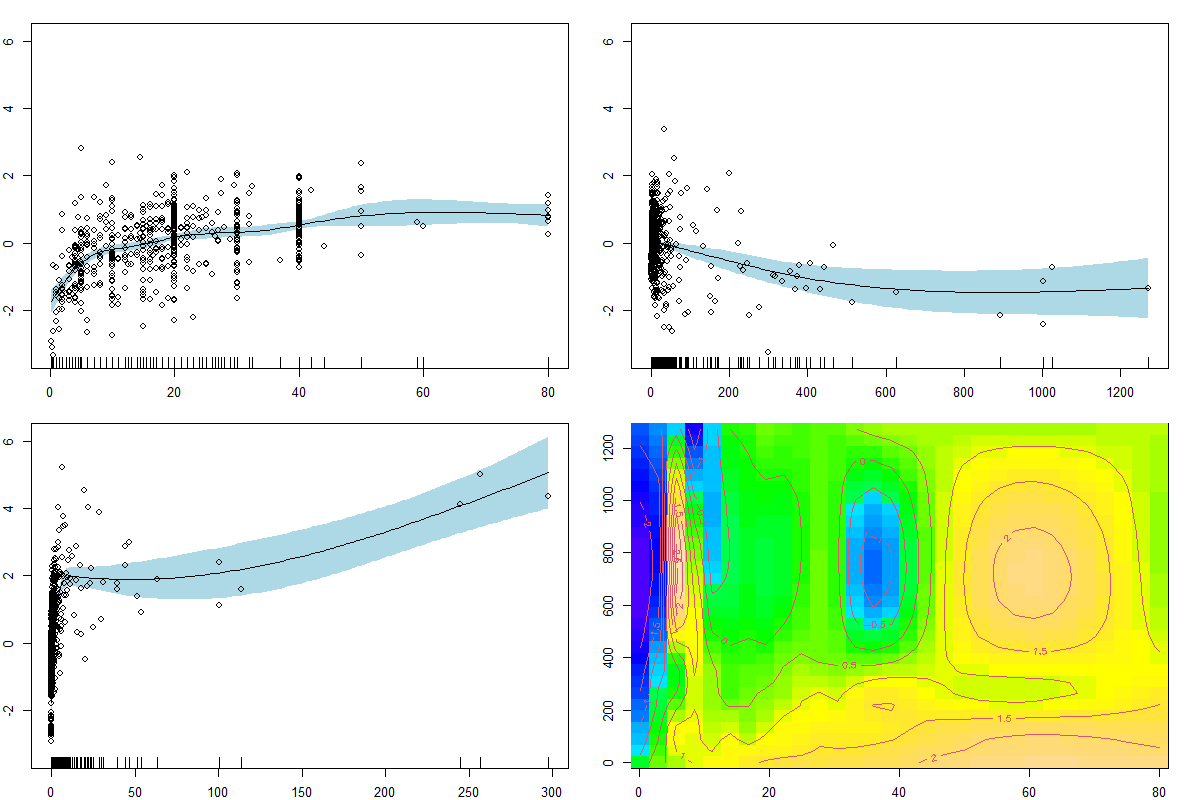
V ďalšom kroku bude potrebné nájsť matematický model (formu), ktorý by najlepšie reprezentoval spracovávané dáta, aby mohli byť ďalej štatisticky vyhodnotené (Orton 2000, 9-10; Orton 1980, 20-21). Keďže skúmané premenné (response variables) nespĺňali predpoklady normálnej distribúcie, využili sme tzv. 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 alebo odsrániť. Čitateľa odkazujeme najmä na práce od S. Wood (2017),  Alain. F. Zuur et. al. (200?), M.J Crawley (2007) a takisto množstvo užitočných informácii a príkladpv 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.

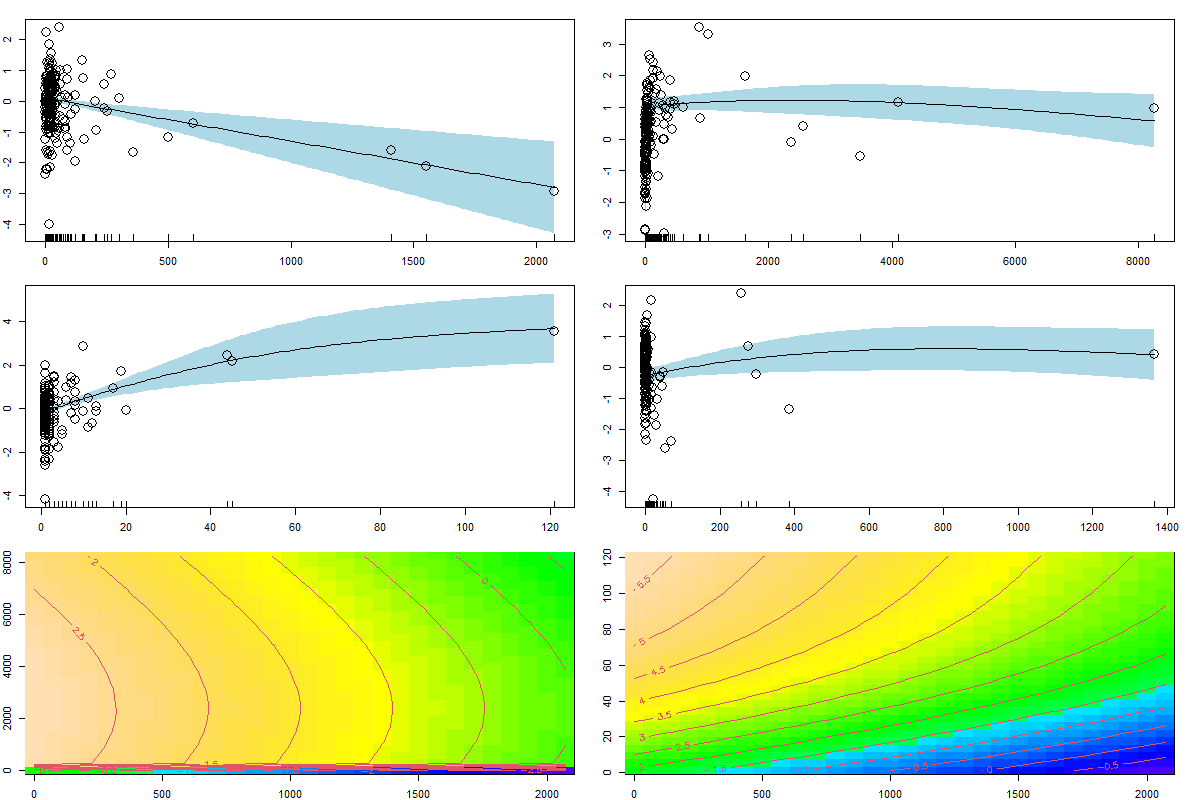
Model 1

Pri analyzovaní a vyhodnocovaní modelu 1 sme použili maticu m x n, kde m = 683 a n = 4 (volume, PMR, taxa, density). PMR sme zaokrúhlili na celé čísla smerom hore.



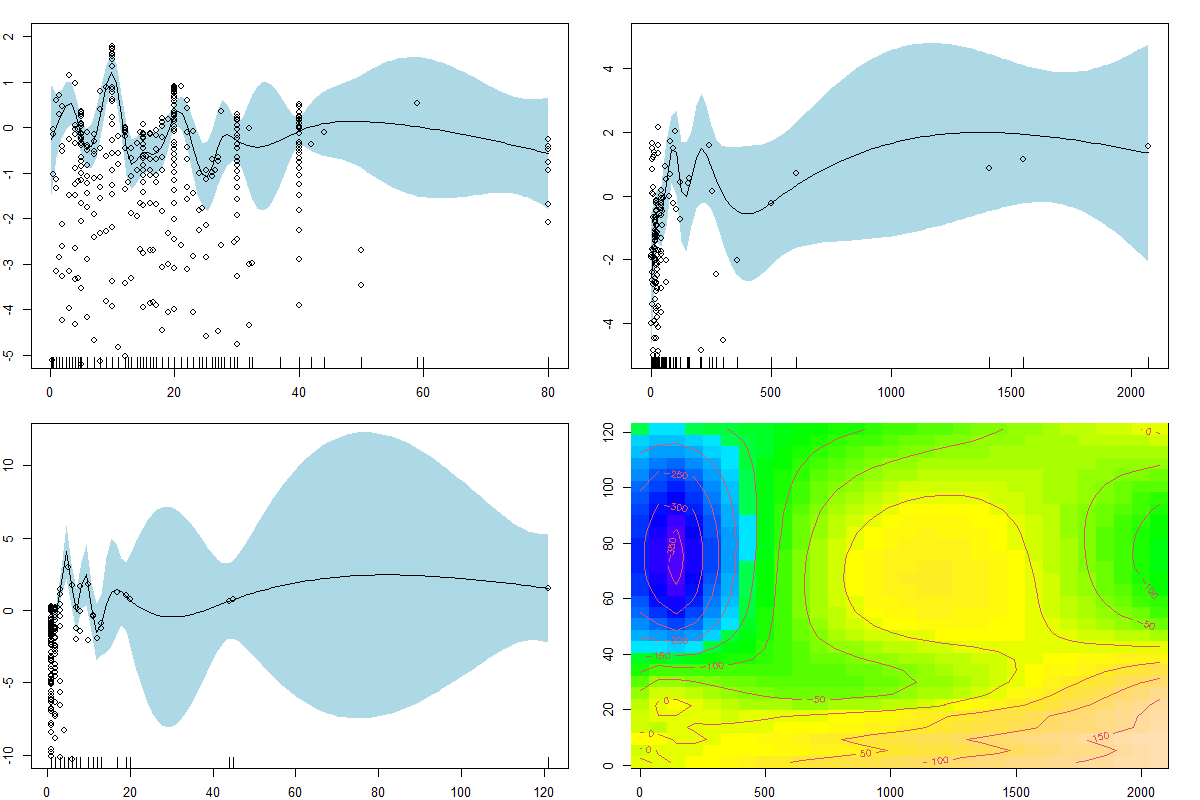
Model 1 zobrazuje vzťah (efekt) medzi počtom rastlinných druhom a jednotlivými prediktormi (objem, PMR, densiity, interakcia objem + RMZ) z hľadiska celého súboru vzoriek. Podľa grafu (graf) a údajov v tabuľke (tab, m1) majú všetky predikory efekt na výsledný model a sú okrem PMR sú štatisticky signifikantné. Súbor je do istej miery poznačený nedostatočným rozptylom (underdispersion), čo je často spôsobené, keď sú počty (counts) zhluknuté (Hilbe 2014, 136) a taktiež ich mohli ovplyvniť viaceré extrémne hodnoty do ktorých sa snaží model napasovať (Zuur et al. 2009, 225). V našom mali na to najväčší vplyv PMR, pričom z celkovej variability modelu 81% vysvetľujú práve PMR 76%. Ako bolo ukázané vyššie, PMR navyše vykazujú silnú koreláciu s druhmi. Spolu vzaté, PMR by boli vhodným kanditátom na odstránenie z modelu a ich separátnu analýzu. Napriek tomu si ich v modeli kvôli ilustrácii ponecháme, aby sme sledovali, ako sa bude model správať. GAM modely sú pomerne robustné, takže sa pokusíme zinterpetovať tento model s PMR berúc do úvahy jeho nedostatky. Objem (graf – A) má na počet druhov efekt a teda ich vzťah nie je lineárny. Najvyšší skok dosahuje približne pri 10 litrov a potom má slabo stúpajúci trend sprevádzaný vyššou mierou nespľahivlosti odhadu funkcie (interval spoľahlivosti na hladine 95% - modrá farba). Jeho vplyv na počet druhov pri vyšších množstvách nie je presvedčivý. Smotný objem (bez ostatných prediktorov) pritom vysvetľuje len 2,2% variability a jeho celková vypovedacia hodnota tak ostáva len orientačná. Vyššie spomínané PMR majú najväčší rozptyl, resp. najviac extrémnych hodnôt, čo sa spolu vyššie uvedeným podpísalo pod ich vypovedaciu hodnotu. Majú najnižší efekt a štatisticky nie sú významné. Graf – b však nazančuje, že vyššie množstvá odobratých objemov v tomto prípade nie sú garanciou vyšších počty druhov. Naopak hustota vykazuje pomerne silný efekt k počtu druhov a je tiež štatistický významná, ale jej interprtácia je komplikovanejšia, kedže je počítana pomocu PMR, takže jej vypovedacia hodnota je tiež len orientačná. Podľa grafu -c je vidieť výrazné (takmer lineárne) stúpanie niekde pri hustote cca 25 nasledovaný výrazným skokom, spôsobeným opäť extrémnymi hodnotami. Podobne ako pri PMR ani vyššie hustoty nie je jednoznačným indikátorom vyšších počtov druhov.- Ako sa toto dá intepretovať? S tým dobre korešpondujú výsledky na grafe d, kde je pomocou „ti“ funkcie[[5]](#footnote-5) zobrazená interakcie objemu a PMR z hľadiska počtu druhov. Na grafe d sa vyprofilovali približne tri hlavné efekt (žltá farba). Najväčší efekt je badať pri menších objemoch s vysokými počtami makrozvyškov, potom pri vyšších objemoch a vyšších počtoch makrozvyškov a nakoniec pri vysokých objemoch a nižších počtoch makrozvyškov. Ak to zhrnemie, tak tu nevidíme žiadny jasný trend alebo vzorec, ktorý uprednostňoval jeden prediktor pred druhým. Hlavným výsledkom modelu 1 je, že objem, PMR, hustoty majú slabú vypovedaciu hodnotu z hľadiska predikcie počtu druhov. To znamená, že vzťah medzi počtami druhov a prediktormi nie je linerárny a spoliehať sa tak na predpoklad, že s vyšším množstvom objemu / PMR by sme získali aj vyšší počet druhov v tomto prípade neopodstatnený.

Model 2

Pri modele 2 sme pracovali s upravnými dátami a snažili sme preskúmať vyššie vzťah medzi rastlinnými druhmi vzhľadom k objektom/kontextom. Pôvodnú maticu sme upravili ako n x m, kde n = 210 (objekty/kontexty) a m = 4 (druhy, PMR, objem, počet vzoriek). Tentokrát sme však už nerobili jeden model pre každú premennú, ale použili sme všetky naraz vrátane ich niektorých interakcií. Hodnotu gamma sme ponechali na odprúčanej hladine 1.14 (Wood 2017, starna, Zuur et al. 2009, strana) a znížili sme tak penalizáciu pre smooth funkcie, aby sme postihli hlavný trend.

V modeli 2 (graf XY) sme sa pozreli na na vzťah (efekt) medzi počtom druhom a jednotlivými premennými (objem, PMR, počet vzoriek, hustota, objem + PMR a objem + hustota). Tento model vysvetľuje 85% variability avšak hodnoty v tabuľkr Xy (m6) naznačujú, že skúmané premenné sa na tom podieľajú rôznou mierou. Napr. objem je síce štatistický signifikantý, ale jeho efekt na celkový výsledok v podsatate nulový, resp. jeho vzťah k druhom je lineárny (rozumej „plochý“ - flat). (graf XY – a). Klesajúci trend je spôsobený predovšetkým extrémnymi hodnotami[[6]](#footnote-6) a s tým stúpa miera nespoľahlivosti. Sám o sebe by objem vysvetľoval približne len 11,5 % variability s menším signifikantným a mierne nelineárnym efektom, avšak celkovo je možné zhodnotiť, že ide o slabý prediktor. Neprekvapivý výsledok priniesli PMR, ktoré potvrdzujú úzky vzťah k druhom. Štatisticky význame sú v tomto prípade počty vzoriek, ktoré sú zodpovedné približne za 21,4% variability. Na grafe XY – c vidieť mierny nelienárny efekt, nazančujúci, že počet vzoriek „pozitívne“ ovplyvňuje počet druhov. Hustota v tomto prípade nie je príliš signifikantná a korenšpoduje s ďalšímí pozorovanimi. Konktrétne ide posledné dva grafy (e,f), ktoré ilustrujú vzájomú interakciu PMR – objem a počet vzoriek – objem. Ich spoločný efekt na počty druhov je zanedbateľný a navyše nesignifikantý. Tento prípad ilustruje, že hlavný vplyv majú skôr jednotlivé premenné (resp. hlavné smooth funkcie) ako ich interakcie. V oboch prípadoch sa však pekne ukazuje, že s vyššími hodnotami objemu sa postupne stráca jeho efekt na počet druhov alebo PMR. Najlepšie si je to možné predstaviť na (grafe - f), kde počty druhov sú inklinujú hlavne k počtom odobratých vzoriek (žltá) ako množstvom objemu a teda ani veľké množstvá negarantujú vyššie počty objemov.

Model 3



V modeli 3 (graf XY - a) sme pracovali s rovnakou maticou ako v modeli 1, tentokrát sme však skúmali vzťah medzi objemom a počtom PMR. Štatisticky významný objem má efekt na PMR a podčiarkuje ich nelineárny vzťah. Z grafu možno vyčítať tri hlavné vrcholy – menšie vzorky okolo 5 litrov potom ca 10 litrové a ca 20 litrové vzorky, kde sa prejavuje najväčší efek objemu. S vyššími objemami stúpa aj miera nespoľahlivosti. Podobne ako pri počtoch druhoch, ani tu sa nedá spoliehať na predpoklad, že s vyššími objemami sa dostaví aj vyššie množstvo makrozvyškov.

Modely 4-6

Modely 4-6 sa počítali z rovnakej matice ako model 2. Pôvodne mali mali byť všetky premenné obsiahnuté v jednom modeli avšak extrémne hodnoty príliš „natiahli“ osu y a menšie zmeny unikali na úkor hlavného trendu. Inak boli výsledné hodnoty podobné či už v rámci jedného modelu alebo samostatne. Tiež sme znížili hodnotu gamma na 0.1 a zvýšili tak „wiggliness“ funkcie, aby sme graficky zachytili menšie nuansy modelu, ktoré by inak unikli.

Model 4 (graf XY-b, tab XY) zachytáva vzťah medzi objemom a PMR a dokladá pomerne vysoký efekt a nelineárny vzťah. PMR je najviac sú ovplyvnené menšími objemami, odľhahlé hodnoty – vysoké objemy - zvyšujú mieru nespoľahlivosti.

Model 5

Podobná situácia ako v modeli 4 je aj v modeli 5 (graf XY – c), ktorý tento krát zohľadňuje počet PMR z hľadiska počtu vzoriek. Model však trpí tým, že z 209 objektov/kontextov sa zo 131 odobrala len jedna vzorka, Napriek tomu model opäť povrdzuje silnú nelinearitu.

Model 6

Posledný model (Graf XY - d)reprezentuje interakciu medzi objemom a počtom odobraných vzoriek vzhľadom k PMR. Napriek tomu, že táto interakcia má signifikantý efekt na PMR a v podstate potvrdzuje výsledky modelov 4 a 5. Na prvý pohľad to vyzerá, že vyššie počty odobraných vzoriek nemajú žiadny vplyv na množstvo PMR (modrá zóna). Ak však vezmeme do úvahy vysoký počet objektov s jednou odobranou vzorkou, tak to dáva zmysel. Rovnako ako zväčšujúci sa efekt smerom k vyšším objemom pri nízkych počtoch vzorkách spôsobené odľahlými hodnotami.

Modely 3 až 6 spoločné jednodznačne poukazujú na nejednoznačnost

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)
5. Ide o tzv. tensor product smooth interaction funkciu (pozri Wood 2017, 187, 326; https://rdrr.io/cran/mgcv/man/gam.html). [↑](#footnote-ref-5)
6. Pre porovnanie, pri modelovaní pomocou generalized linear models za využitia negatívnej binomiálnej regresie (funkcia glm.nb z balíka MASS) existuje šanca, že ak by sme odobrali jednu jednotku objemu navyše, tak počet druhov sa **priemerne** zvýši približne o 1%. Aj v tomto prípade je objem veľmi slabý prediktor, ak sa na to pozeráme z hľadiska objektov/kontextov. [↑](#footnote-ref-6)