VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

FAKULTA INFORMATIKY A STATISTIKY

SHLUKOVÁ A REGRESNÍ ANALÝZA MIKROPANELOVÝCH DAT

Disertační práce

Doktorand: Ing. Lukáš Sobíšek

Školitel: doc. Ing. Iva Pecáková, CSc.

Studijní program: Kvantitativní metody v ekonomice

Studijní obor: Statistika

Praha, říjen 2016

Prohlášení

Prohlašuji, že jsem disertační práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze kterých jsem čerpal.

V Praze dne 31. října 2016

…….
podpis

Poděkování

Rád bych na tomto místě poděkoval své školitelce a vedoucí mé dizertační práce doc. Ing. Ivě Pecákové, CSc. za odborné vedení, cenné připomínky, rady a podněty a její čas strávený na konzultacích.

Dále bych chtěl poděkoval svým spolupracovníkům prof. Ing. Haně Řezankové, CSc., Ing. Karlovi Helmanovi, Ph.D., Mgr. Milanovi Baštovi, Ph.D., Ing. Jiřímu Procházkovi a Mgr. Márii Stachové, PhD. za cenné připomínky.

Abstrakt

Panelové studie se provádí především za účelem analýzy změn hodnot sledovaných proměnných v čase. V mikropanelovém výzkumu se sleduje velké množství objektů periodicky během relativně krátkého časového úseku (v řádu let). Počet opakovaných měření je v řádu jednotek. Tato práce se věnuje stávajícím přístupům k regresní a shlukové analýze mikropanelových dat. Jedním z přístupů k analýze mikropanelu je využití modifikovaných vícerozměrných statistických modelů pro průřezová data, které zohledňují korelaci měření pro daný objekt.

V práci jsou shrnuty dostupné nástroje pro regresní analýzu mikropanelových dat. Kromě rekapitulace známých a užívaných smíšených lineárních modelů pro normálně rozdělenou závisle proměnnou jsou stručně představeny nové přístupy pro analýzu vysvětlovaných proměnných s jiným než normálním rozdělením. Mezi ně patří například zobecněný lineární marginální model, zobecněný lineární model se smíšenými efekty a bayesovský přístup. Kromě popisu těchto modelů je uveden stručný přehled jejich implementace v systému R.

S regresními modely upravenými pro mikropanelová data je spjato úskalí v nejednoznačnosti odhadu jejich parametrů. V práci je navrženo, jak zpřesnit odhady pomocí shlukové analýzy. Proto jsou v práci popsány metody shlukové analýzy mikropanelových dat. Vzhledem k tomu, že nabídka metod je omezená, hlavním cílem práce bylo navrhnout vlastní dvoukrokový postup shlukování mikropanelových dat. V prvním kroku jsou transformována panelová data na statická pomocí skupiny navržených charakteristik dynamiky, které reprezentují různé vlastnosti časového vývoje sledované proměnné. Ve druhém kroku jsou shlukovány objekty konvenčními prostorovými technikami (aglomerativní shlukování a metoda *C*-průměrů) na základě matice nepodobnosti hodnot shlukovacích proměnných spočítaných v prvním kroku. Dalším cílem práce je zjistit, zda navržený postup shlukování vede ke zkvalitnění regresních modelů pro tento typ dat.

Pomocí simulační studie je porovnáván navržený shlukovací přístup s postupem aplikovaným v balíčku *kml* systému R a se shlukovacími charakteristikami, které navrhuje D'Urso (2004). V provedené studii dosáhla kombinace navržených shlukovacích proměnných lepších výsledků než používané skupiny shlukovacích proměnných. Dalším přínosem práce je skript napsaný pro jazyk R přiložený na CD. Tento skript je možno použít pro analýzu vlastních mikropanelových dat.

Klíčová slova: shluková analýza mikropanelových dat, regresní analýza mikropanelových dat, smíšený lineární model, Wardova metoda, metoda *C*-průměrů, charakteristiky dynamiky

Abstract

The main purpose of panel studies is to analyze changes in values of studied variables over time. In micro panel research, a large number of elements are periodically observed within the relatively short time period of just a few years. Moreover, the number of repeated measurements is small. This dissertation deals with contemporary approaches to the regression and the clustering analysis of micro panel data. One of the approaches to the micro panel analysis is to use multivariate statistical models originally designed for cross-sectional data and modify them in order to take into account the within-subject correlation.

The thesis summarizes available tools for the regression analysis of micro panel data. The known and currently used linear mixed effects models for a normally distributed dependent variable are recapitulated. Besides that, new approaches for analysis of a response variable with other than normal distribution are presented. These approaches include the generalized marginal linear model, the generalized linear mixed effects model and the Bayesian modelling approach. In addition to describing the aforementioned models, the paper also includes a brief overview of their implementation in the R software.

The difficulty with the regression models adjusted for micro panel data is the ambiguity of their parameters estimation. This thesis proposes a way to improve the estimations through the cluster analysis. For this reason, the thesis also contains a description of methods of the cluster analysis of micro panel data. Because supply of the methods is limited, the main goal of this paper is to devise its own two-step approach for clustering micro panel data. In the first step, the panel data are transformed into a static form using a set of proposed characteristics of dynamics. These characteristics represent different features of time course of the observed variables. In the second step, the elements are clustered by conventional spatial clustering techniques (agglomerative clustering and the *C*-means partitioning). The clustering is based on a dissimilarity matrix of the values of clustering variables calculated in the first step. Another goal of this paper is to find out whether the suggested procedure leads to an improvement in quality of the regression models for this type of data.

By means of a simulation study, the procedure drafted herein is compared to the procedure applied in the *kml* package of the R software, as well as to the clustering characteristics proposed by D'Urso (2004). The simulation study demonstrated better results of the proposed combination of clustering variables as compared to the other combinations currently used. A corresponding script written in the R-language represents another benefit of this paper. It is available on the attached CD and it can be used for analyses of reader's own micro panel data.

Keywords: clustering analysis of micro panel data, regression analysis of micro panel data, mixed effects model, Ward's clustering method, *C*-means clustering method, characteristics of dynamics

OBSAH

Význam použitých symbolů ... 9
Seznam zkratek ... 12
Úvod ... 14
1. Charakter dat pocházejících z panelových šetření ... 19
1.1 Panelový výzkum a jeho význam ... 19
1.2 Příklady panelových šetření v ekonomických a neekonomických oblastech ... 22
1.3 Výhody a nevýhody panelového výzkumu ... 26
2. Regresní analýza normálně rozdělené longitudinální vysvětlované proměnné ... 28
2.1 Dvoukroková analýza ... 29
2.2 Smíšený lineární model (LMM) ... 31
2.3 Marginální model (model populačních průměrů, MMM) ... 40
3. Nové přístupy v oblasti regresní analýzy panelových dat ... 43
3.1 Zobecněný lineární marginální model (GMMM) ... 44
3.2 Zobecněný lineární model se smíšenými efekty (GLMM) ... 45
3.3 Bayesovský přístup ... 49
4. Analýza longitudinálních dat v systému R ... 51
4.1 Frekventistické LMM balíčky ... 53
4.2 Marginální model v R ... 59
4.3 GLMM v R ... 60
4.4 Bayesovský přístup v R ... 63
5. Shlukování trajektorií ... 65
5.1 Obecný úvod – kategorizace metod shlukové analýzy ... 66
5.2 Popis vybraných shlukovacích metod ... 70
6. Návrh a ukázky využití shlukování individuálních trajektorií ... 78
6.1 Regresní model prostorových bloků ... 78
6.2 Identifikace odlehlých trajektorií ... 81
6.3 Profil shluků jako cíl analýzy ... 83
7. Charakteristiky dynamiky trajektorií mikropanelových dat ... 84
7.1 Návrh nových charakteristik dynamiky a jejich testování ... 84
7.1.1 Charakteristiky popisující dynamiku individuálních trajektorií ... 84
7.1.2 Porovnání metod založených na práci s původními daty a jejich transformacemi 89
7.2 Porovnání nejlepší metody a zavedených shlukovacích postupů ... 125
8. Vyhodnocení zpřesnění odhadů regresních modelů shlukovou analýzou ... 130
8.1 Data a metodologie ... 130
8.2 Výsledky shlukování a klasifikace ... 133
8.3 Výsledky regresní analýzy ... 135

Závěr 140
Literatura 144
Přílohy 157

Význam použitých symbolů

A	… počet shlukovacích proměnných. V kontextu této práce jsou objekty (individuální trajektorie) shlukovány na základě A charakteristik popisujících dynamiku vývoje individuální trajektorie
$b_{i1}, \ldots, b_{iQ}$	… značí náhodné efekty (regresní parametry) pro i-tý objekt a q-tou vysvětlující proměnnou Z_{iq}, kde $q = 1, 2, \ldots, Q$. Náhodný efekt b_{iq} je nepozorovaná normálně rozdělená náhodná veličina různá pro každá objekt
C	… počet shluků ve výběrovém souboru, tj. $c = 1, 2, \ldots, C$
c	… značí c-tý shluk
d_{cc}	… míra podobnosti založená na Pearsonově korelačním koeficientu
d_{ict}	… Euklidovská míra vzdálenosti
$\overline{\lvert diff_i \rvert}$	… průměrná absolutní trojúhelníková vzdálenost mezi dvěma sousedními měřeními pro i-tý objekt
$\overline{diff}_i$	… průměrná trojúhelníková vzdálenost mezi dvěma sousedními měřeními pro i-tý objekt
$diff_{it}$	… trojúhelníková vzdálenost dvou sousedních hodnot pro i-tý objekt
D_{it}	… časová vzdálenost t od počátku pro i-tý objekt (v letech)
$\mathbf{e}_i$	… vektor $\mathbf{e}_i$ obsahuje A hodnot shlukovacích proměnných e_{ia} pro i-tý objekt
e_{ia}	… hodnota a-té vývojové charakteristiky (shlukovací proměnné) pro i-tý objekt, kde $a = 1, 2, \ldots, A$
$\mathbf{h}_{ct}$	… c-tý centroid v čase t $\mathbf{h}_{ct} = (h_{c1t}, \ldots, h_{cjt}, \ldots, h_{cJt})'$
i	… značí i-tý objekt (statistickou jednotku)
J	… počet exogenních vysvětlujících proměnných
k	… k-tá iterace (viz kapitola 8)
$\overline{k_i}$	… průměrný koeficient růstu pro i-tý objekt
m	… konstanta kontrolující zmatenost (*fuzziness*) získaných shluků v modelu fuzzy C-průměrů
$\max \triangleleft$	… maximální úhel (v radiánech) mezi spojnicí okrajových měření a spojnicí vnitřního bodu a prvního měření
N	… počet objektů ve výběrovém souboru, tj. $i = 1, 2, \ldots, N$
n_c	… počet objektů ve shluku c
n_{ct}	… počet objektů v prostorovém bloku$_{ct}$

PRED	… náležitost do O shluků odhadnutá shlukovacím algoritmem, $PRED = \{l_1, \ldots, l_O\}$, kde l_o jsou odhadnuté shluky
Q	… počet náhodných efektů
R	… značí počet regresorů v regresním modelu
$\mathbf{R}$	… blokově diagonální kovarianční čtvercová matice náhodných chyb ε_{it} pro všech N objektů, tj. obsahuje N matic $\mathbf{R}_i$ na diagonále.
$\mathbf{R}_i$	… kovarianční čtvercová matice náhodných chyb ε_{it} o rozměrech $T \times T$ pro i-tý objekt
$\mathrm{sd}(\lvert diff_i \rvert)$	… směrodatná odchylka absolutních trojúhelníkových vzdáleností mezi dvěma sousedními měřeními pro i-tý objekt
$\mathrm{sd}(diff_i)$	… směrodatná odchylka trojúhelníkových vzdáleností mezi dvěma sousedními měřeními pro i-tý objekt
S_y	… celková variabilita proměnné Y
$S_{y,B}$	… vnitrobloková variabilita
$S_{y,W}$	… mezibloková variabilita
t	… pomocná časová proměnná, tj. $t = 1, 2, \ldots, T$
T	… počet měření v čase pro jeden objekt v případě stejného počtu opakování v čase pro různé objekty
T_i	... počet měření v čase pro objekt i
TRUE	... skutečná náležitost do C tříd, $TRUE = \{c_1, \ldots, c_C\}$, kde c_c jsou uvažované třídy (shluky)
u_{ict}	… stupeň náležitosti i-tého objektu do c-tého shluku v čase t
$\mathbf{V}_i$	… kovarianční matici vektoru hodnot vysvětlované proměnné $\mathbf{Y}_i$ pro objekt i
v_{it}	… náhodná složka REM a MEM s komponenty $\alpha_i + \lambda_t + \varepsilon_{it}$
$\mathbf{x}_{it}$	… vektor pozorování pro i-tý objekt v čase t $\mathbf{x}_{it} = (x_{i1t}, \ldots, x_{ijt}, \ldots, x_{iJt})'$
X_j	… značí j-tou fixní exogenní vysvětlující proměnnou, kde $j = 1, 2, \ldots, J$. Její efekt (regresní parametr) β_j je konstantní pro různé objekty
x_{jit}	… hodnota vysvětlující proměnné X_j pro i-tý objekt v čase t
$\mathbf{Y}_i$	… vektor hodnot vysvětlované proměnné Y pro objekt i
y_{it}	… označuje hodnotu vysvětlované proměnné u i-tého objektu v čase t
$\mathbf{Z}_i$	… je matice (plánu) hodnot exogenních Q vysvětlujících proměnných $z_{1it}, \ldots, z_{Qit}$

Z_q	… značí q-tou exogenní vysvětlující proměnnou, kde q = 1, 2, …, Q. Její efekt (regresní parametr) b_{iq} je náhodný (různý) pro každý objekt
z_{qit}	… hodnota vysvětlující proměnné Z_q pro i-tý objekt v čase t
α_i	… individuální efekt pro i-tý objekt. V FEM α_i je v čase fixní. V REM a MEM je α_i definován jako náhodná veličina (náhodný efekt)
$\boldsymbol{\beta}'$	… značí vektor pevných efektů (regresních parametrů) J vysvětlujících exogenních proměnných X_j
β_j	… pevný efekt (regresní parametr) pro X_j
ε_{it}	… náhodná chyba (teoretické reziduum) pro i-tý objekt v čase t v REM a MEM. V KLM a FEM značí náhodnou složku $\varepsilon_{it} = v_{it}$
η_{it}	… lineární prediktor (systematická část regresního modelu)
Λ	… relativní kovarianční faktor
λ	… parametr (střední hodnota) Poissonova rozdělení
λ_t	… náhodný efekt času t v REM a MEM
μ	… parametr regresního modelu vyjadřující společnou střední hodnotu pro všech N objektů, který je konstantní v čase
$\boldsymbol{\Sigma}$	… pozitivně semidefinitní kovarianční matice náhodných efektů o rozměrech $Q \times Q$ obsahující hodnoty rozptylů ψ_q^2 a kovariancí $\psi_{qq'}$ náhodných efektů
σ^2	… konstantní rozptyl náhodné složky ε_{it} (viz předpoklad KLM); $\sqrt{\sigma^2}$ v LMM nazýván škálový parametr
ψ_q^2	… rozptyl náhodného efektu b_{iq}
$\psi_{qq'}$	… kovariance mezi náhodnými efekty b_q, $b_{q'}$

Seznam zkratek

AIC	… Akaikeho informační kritérium
ANOVA	… jednofaktorová analýza rozptylu
BIC	… Schwarzovo Bayesovské informační kritérium
CCR	… míra správné klasifikace (Correct Classification Rate)
CLMM	… kumulativní spojovací smíšený model (Cummulative Link Mixed Model)
DQI	… index datové kvality (Data Quality Index)
FCM	… metoda fuzzy *C*-průměrů (Fuzzy *C*-Means)
FEM	… model s fixními (pevnými) efekty (Fixed Effects Model)
GAM	… zobecněný aditivní model (Generalized Additive Model)
GAMM	… zobecněný aditivní model se smíšenými efekty (Generalized Additive Mixed Model)
GEE	… Generalized Estimating Equations
GHQ	… Gaussova-Hermiteho kvadratura (Gauss-Hermite Quadrature)
GLM	… obecný lineární model (General Linear Model)
GLMM	… zobecněný lineární model se smíšenými efekty (Generalized Linear Mixed Model)
GMMM	… zobecněný lineární marginální model (Generalized Marginal Means Model)
HLM	… hierarchický lineární model (Hierarchical Linear Model)
ICC	… vnitroskupinový korelační koeficient (Intraclass Correlation)
IRLS	… iterativní obecná metoda vážených nejmenších čtverců (Iteratively Reweighted Generalized Least Squares)
Jac	… Jaccardův index
KLM	… klasický lineární model
LMM	… smíšený lineární model (Linear Mixed Model)
MCMC	… odhad pomocí simulace Markovova řetězce Monte Carlo (Markov Chain Monte Carlo)
MEM	… model se smíšenými efekty (Mixed Effects Model)
MSEE	… střední čtvercová chyba odhadu (Mean Squared Error of Estimate)
ML	… metoda maximální věrohodnoti (Maximum Likelihood)
MMM	… marginální model (Model populačních průměrů, Marginal Means Model)
MSE	… střední čtvercová chyba (Mean Squared Error)
NLNO	… nejlepší lineární nezkreslený odhad

OLS	… metoda nevážených nejmenších čtverců (Ordinary Least Square Method)
PQL	… pseudo- a penalizovaná kvazi-věrohodnost (Pseudo- and Penalized Quasi Likelihood)
REM	… model s náhodnými efekty (Random Effects Model)
REML	… metoda omezené maximální věrohodnosti (Restricted Maximum Likelihood)
RI	… Randův index
RIadj	… modifikovaný Randův index
RIM	… model s náhodným absolutním členem (Random Intercept Model)
RSM	… model s náhodným sklonem (Random Slope Model); rovněž se používá označení Individual Growth Curve Model
SE	… odhad střední chyby, neboli směrodatné chyby odhadu (Standard Error)
ZIGLMM	… zobecněný lineární model se smíšenými efekty s vlivnými nulami (Zero-Inflated Generalized Linear Mixed-Effects Model)
ZIP	… Poissonův regresní model s vlivnými nulami (Zero-Inflated Poisson regression)
ZLM	… zobecněný lineární model

Úvod

Panelový (nebo v biostatistice používaný termín longitudinální) výzkum sdružuje poznatky o objektech (statistických jednotkách) po delší časové období. Těmito objekty mohou být například pacienti, domácnosti, země, firmy atd. Sledované proměnné (náhodné veličiny) jsou pro každý objekt měřeny vícekrát v různých časových obdobích (okamžicích). Longitudinální data zachycují změny hodnot proměnných pro každý objekt v čase. Primární cíl longitudinální analýzy je deskriptivní analýza vývoje sledovaných proměnných a analýza závislosti mezi změnami sledovaných proměnných, případně analýza vztahu mezi změnami primárně sledovaných proměnných a k jednomu časovému okamžiku, často k bazickému období, naměřenými hodnotami jiných proměnných.

Panelový výzkum lze členit dle typu výběrového souboru na mikro- a makropanelový výzkum. V mikropanelovém (*short panel*) výzkumu se sleduje velké množství objektů N (běžně v řádech stovek, či tisíců). Objekty jsou periodicky pozorovány během relativně krátkého časového úseku (zhruba během 2 až 10 let) a počet opakovaných individuálních měření je v řádu jednotek. Makropanel (*long panel*) obsahuje menší počet objektů N (jednotky, desítky) a tyto objekty jsou navzájem porovnávány většinou ročně během dlouhého časového horizontu (délka studie je běžně 20–60 let). Tyto dlouhodobé studie slouží například k porovnání vývoje různých ukazatelů zemí.

V panelových studiích, kde je každý objekt měřen vícekrát v různých časových obdobích, lze očekávat korelaci měření pro daný objekt. Z tohoto důvodu nelze použít pro regresní analýzu klasický lineární regresní model nebo zobecněný lineární model. Regresní modelování panelových dat v sobě obsahuje prvky regresní analýzy pro prostorová (statická) data a analýzy časových řad. Z povahy dat plynou určité odlišnosti v analytickém přístupu k mikropanelu a makropanelu. Pro analýzu mikropanelu se využívají vícerozměrné statistické modely pro prostorová data upravené tak, aby zohledňovaly korelaci měření pro daný objekt, zatímco pro makropanel vychází statistická teorie z analýzy časových řad. Tato disertační práce se zabývá analýzou mikropanelových dat.

S regresními modely určenými pro mikropanelová data je spjato úskalí v nejednoznačnosti odhadu jejich parametrů. Regresní parametry a jejich variabilita jsou odhadovány iterativně numerickými optimalizačními metodami. Odhady parametrů modelu a odhady hodnot vysvětlované proměnné jsou nepřesné, pokud kovarianční struktura náhodných chyb (teoretických reziduí) není známá, nebo se v datech vyskytuje směs rozdělení. Shluková analýza je jednou z možností, jak lze odhady těchto regresních modelů zpřesnit. Metody shlukové analýzy dokáží

identifikovat podskupiny (shluky) objektů s homogenním vývojem sledované proměnné v čase (trajektorií). Individuální trajektorie popisuje tvar vývoje sledované proměnné v čase pro daný objekt.

Regresní analýza provedená pro každý shluk zvlášť může vést k optimálnímu a nezkreslenému odhadu parametrů s menší variabilitou na rozdíl od přímého použití vybraného regresního modelu na všechny objekty s různou kovarianční strukturou náhodných chyb. Realizace tohoto postupu znamená začlenit do modelu náležitost do shluku jako kategoriální pomocnou vysvětlující proměnnou a její interakce s ostatními vysvětlujícími proměnnými (např. s pomocnou časovou proměnnou *t*). Shluková analýza vývojových trajektorií opakovaně v čase zjišťovaných proměnných může dále sloužit k identifikaci odlehlých trajektorií. Kromě exploratorní role může shlukování trajektorií vystupovat jako primární cíl analýzy, např. Procháska a kol. (1991), Koestler a kol. (2014). Proto jsou v práci rekapitulovány metody shlukové analýzy mikropanelových dat.

Nabídka softwarových nástrojů pro shlukování mikropanelových dat je omezena. K dispozici je pouze jedna funkce, která je implementována v jazyce R v balíčku *kml*, autorem je Genolini a Falissard (2011). Funkce aplikuje metodu *C*-průměrů na panelová data. Tato metoda však není univerzálním řešením pro všechny aplikace, což ukazuje například implementace *kml* pro shlukování mikropanelových finančních ukazatelů podniku viz Stachová a Sobíšek (2016). Cílem uvedené analýzy byla klasifikace slovenských průmyslových podniků na základě vývoje vybraných finančních ukazatelů. Podniky byly shlukovány pomocí funkce *kml*, aplikované na roční hodnoty z let 2010–2013. Výsledek shlukové analýzy nebyl uspokojivý. Z tohoto důvodu je hlavním cílem této disertační práce navrhnout vlastní postup shlukování mikropanelových dat. Dalším cílem této práce je tento postup poskytnout veřejnosti v podobě programové implementace v jazyce R.

V práci je navrženo, jak využít výsledky shlukové analýzy k zpřesnění odhadů regresních modelů mikropanelových dat. Dalším cílem práce je zjistit, zda navržený postup shlukování vede k přesnějším odhadům regresních modelů pro tento typ dat.

První kapitola této práce vysvětluje problematiku panelových dat. Rekapituluje termíny z této oblasti a uvádí příklady panelových studií. Panelové studie (mikro a makro) jsou běžně užívány v neekonomických vědních oborech, např. klinické studie, genetika, epidemiologie, vzdělávání, politologie, ekologie. V ekonometrii se využívají makropanelové studie např. s cílem porovnat země z hlediska vývoje makro agregátů. Mikropanelové studie se používají

v marketingu k modelování dynamiky spotřebitelského chování. V jiných ekonomických oborech mikropanelové studie příliš rozšířeny nejsou.

Ve druhé kapitole práce jsou popsány regresní modely, které se používají pro regresní analýzu korelovaných dat s normálně rozdělenou vysvětlovanou proměnnou. Historicky se vyvíjely a paralelně se používají dva přístupy k modelování panelových dat: dvoukrokový přístup a modifikace obecného lineárního modelu nazývaná smíšený lineární model. Oba přístupy jsou zde vysvětleny a vzájemně porovnány z hlediska jejich výhod a nevýhod. Jak název smíšený lineární model napovídá, jedná se o lineární model, který umožňuje zahrnout směs dvou typů parametrů vysvětlujících proměnných: pevné (fixní) efekty a náhodné efekty. Definice těchto skupin se v literatuře různí. Odlišné přístupy jsou v této kapitole popsány s cílem vytvořit základ pro identifikaci a správnou klasifikaci typu vysvětlující proměnné. V kapitole jsou dále stručně popsány a vzájemně porovnány následující modely: modely s pevnými efekty, modely s náhodnými efekty, modely se smíšenými efekty, model s náhodným absolutním členem, model s náhodným sklonem, hierarchický lineární model a marginální model. Závěr kapitoly rekapituluje postupy odhadů jejich parametrů.

Modely, kterým je věnována druhá kapitola, předpokládají spojitou normálně rozdělenou vysvětlovanou proměnnou. Tento předpoklad není často realistický. V mnoha aplikacích je longitudinální vysvětlovaná proměnná diskrétní. Jako příklad lze uvést dvě aplikace z ekonomické oblasti: v marketingové aplikaci je respondentem preferovaná značka nominální vysvětlovaná proměnná a míra preference konkrétní značky ordinální proměnná. V oblasti bankovnictví je kriticky důležité správně vyhodnotit finanční zdraví firmy žádající o úvěr. Binární proměnná vyjadřuje solventnost firmy, neboli schopnost hradit své závazky vůči bance včas (ano/ne). Klasifikace firmy se v čase mění na základě vývoje finančních ukazatelů. V oblasti biostatistiky se například sleduje počet relapsů (znovuvzplanutí příznaků nemoci), což je ukázka diskrétní proměnné nabývající kladných hodnot.

Za účelem regresní analýzy vysvětlovaných proměnných pocházející z jiného než normálního rozdělení bylo nutné (a díky rozvoji výpočetní techniky i možné) rozvíjet a v praxi využívat nejen modifikace těchto přístupů (kapitola 2), ale i nové přístupy k modelování složitějších úloh. Třetí kapitola práce nabízí přehled nových a relativně krátkou dobu (od přelomu tisíciletí) používaných přístupů k modelování panelových dat, jmenovitě zobecněný lineární marginální model, zobecněný lineární model se smíšenými efekty, zobecněný lineární model se smíšenými efekty modifikovaný pro data s vysokou četností vlivných nulových hodnot, kumulativní spojovací model se smíšenými efekty a bayesovský přístup.

Čtvrtá kapitola práce shrnuje existující software pro využívání modelů regresní analýzy panelových dat popsaných v kapitolách 2 a 3. Kromě komerčních systémů (SAS, IBM SPSS Statistics, Stata, S+ a HLM) nabízejících regresní analýzu panelových dat je v současnosti i v této oblasti extenzivně rozvíjen jazyk R (R Core Team 2013). Mezi jeho hlavní konkurenční výhody patří bezplatné užívání, implementace nejnovějších poznatků z oblasti statistiky a možnost programovat si vlastní funkce s využitím funkcí ze zavedených balíčků. V kapitole 4 se dále zabývám nedořešenými otázkami a problémy v implementaci regresních modelů pro panelová data v jazyce R. Stejná metoda je často implementována ve více balíčcích. Tato část práce vymezuje rozdíly mezi nimi, jelikož volba správné funkce je zásadní pro odhad optimálního modelu. Kromě volby správné funkce je rovněž kriticky důležité správně nastavit parametry funkce. Výzkumníci z různých oblastí nechávají poměrně často automaticky argumenty funkcí na přednastavených hodnotách. V takovém případě nemusí funkce nalézt optimální řešení (odhad parametrů modelu). Cílem popisu funkcí v této kapitole je upozornit nezasvěcené výzkumníky na existenci těchto zásadních argumentů a možnosti volby jejich hodnot. Uživatelům zvyklým pracovat s jiným statistickým softwarem je doporučena relevantní literatura.

Kapitoly 1–4 shrnují, jaký typ návrhů (designů) panelových studií je možno v současnosti modelovat. Tato část práce porovnává metody regresní analýzy korelovaných dat z hlediska vhodnosti jejich použití pro daný výzkumný záměr při daném typu proměnných a rozměrech matice pozorování. Výše zmíněné kapitoly rovněž upozorňují na úskalí metod a aktuálně nedořešené problémy v programové implementaci jednotlivých modelů. Nabízí se zde aktuální, maximalistický výčet možností regresní analýzy panelových dat včetně nejnovějších přístupů za účelem poskytnout zájemci ucelenou výchozí informaci, která není v české ani zahraničí literatuře k dispozici v jednom zdroji. Z české literatury podává nejkomplexnější informaci o regresní analýze korelovaných dat Pekár a Brabec (2012) nebo Šmilauer (2007). Vývoj regresních metod je dynamický a tyto zdroje neobsahují nejnovější přístupy k modelování. Cílem těchto kapitol je shrnout dostupné nástroje pro regresní analýzu mikropanelových dat z důvodu, že ucelený přehled (včetně nových) regresních metod a jejich implementace do softwarových nástrojů v literatuře chybí.

V souvislosti s cílem této disertační práce hodnotí kapitola 5 stávající přístupy ke shlukování panelových dat. Metod pro shlukování makropanelových dat existuje relativně mnoho. Vycházejí z teorie modelování stacionárních časových řad. Přehled různých přístupů shlukové analýzy časových řad poskytuje Liao (2005). Při shlukování makropanelových dat, tedy dat

s velkým počtem opakovaných měření, je využíváno např. modelování náhodných (např. Gaussovských) procesů. Tento přístup tedy předpokládá, že jsou trajektorie generovány z určitého rozdělení. Avšak u trajektorií s malým počtem opakovaných měření v čase není modelový přístup použitelný. Několik možností shlukové analýzy trajektorií s malým počtem opakovaných měření (mikropanelových dat) je popsáno v této kapitole. Většina z těchto metod vychází ze shlukové analýzy prostorových dat. Konkrétně se používají hierarchické metody a metody rozkladu.

V šesté kapitole práce je pro shlukování individuálních trajektorií navržen dvoukrokový přístup. Postup je založen na transformaci původních longitudinálních dat na statická. Statická data jsou shlukována standardní shlukovou metodou. Transformace na statická data je provedena pomocí charakteristik, které reprezentují různé vlastnosti krátkého časového vývoje sledované proměnné. Ve druhém kroku jsou shlukovány jednotky konvenčními prostorovými technikami na základě matice nepodobnosti hodnot shlukovacích proměnných spočítaných v prvním kroku.

V sedmé kapitole hledám pomocí simulační studie nejvhodnější kombinaci charakteristik ze sedmi uvažovaných charakteristik. Simulační studie obsahuje osm různých scénářů. Scénáře se liší v předpokládaném rozdělení, ze kterého jsou shluky individuálních trajektorií generovány. Pomocí stejné studie je shlukovací schopnost navrženého postupu porovnána s postupem aplikovaným ve funkci *kml* a s charakteristikami, které používá ve svém fuzzy přístupu ke shlukování D'Urso (2004). Výsledky simulačního srovnání jsou shrnuty v této kapitole. Skript s navrženým shlukovacím přístupem, který je možno implementovat na vlastní aplikaci, je k dispozici na přiloženém CD.

V poslední kapitole je v souvislosti s dalším cílem práce testováno, zda navržený shlukovací postup zpřesní odhady regresních modelů. Testování je provedeno pomocí simulovaných dat, na která se aplikuje navržený shlukovací postup. Dále se odhadnutá náležitost do shluků začlení do odhadovaného modelu jako kategoriální pomocná vysvětlující proměnná. Zároveň je odhadnut populační regresní model na všechny objekty bez využití náležitosti do shluku. Oba modely (s náležitosti do shluku a bez ní) jsou porovnány pomocí stanovení průměrné chyby odhadu.

1. Charakter dat pocházejících z panelových šetření

Zvýšená dostupnost panelových dat z výběrových šetření domácností je jedním z nejzásadnějších pokroků v oblasti výzkumu společnosti za posledních třicet let.

Fitzgerald a kol. (1998)

1.1 Panelový výzkum a jeho význam

Panelový (nebo v biostatistice používaný termín *longitudinální*) výzkum sdružuje poznatky o pacientech, domácnostech, zemích, firmách atd. po delší časové období. Menard (2002) definuje panelový výzkum z hlediska typu zkoumaných dat a typu použité analýzy následovně:

- sledované proměnné jsou měřeny vícekrát (alespoň dvakrát) v různých časových periodách pro jednotlivé objekty pozorování,
- tyto objekty pozorování jsou během celého průzkumu identické nebo přinejmenším porovnatelné,
- analýza dat zahrnuje porovnání dat z hlediska času. Umožňuje tedy sledovat vývoj proměnné v čase pro daný objekt.

Na základě Menardovy definice panelového výzkumu lze rozlišit tři základní typy panelových studií:

- prospektivní panelová studie (*Prospective Panel Design*) – proměnné jsou opakovaně měřeny ve více obdobích T vždy na stejných objektech. V čase t jsou získány aktuálně platné hodnoty měřených proměnných pro tento časový okamžik, kde $t = 1, 2, \ldots, T$,
- retrospektivní panelová studie (*Retrospective Panel Design*) – v jeden okamžik t jsou měřeny hodnoty daného objektu nejen pro aktuální časový okamžik t, ale jsou zjišťovány i historické hodnoty za předchozí období. V extrémním případě může dojít k situaci, kdy jsou všechny hodnoty (aktuální a minulé) sledovaných proměnných sbírány pouze v jeden časový okamžik,
- opakovaná průřezová studie (*Repeated Cross-Sectional Design*) – oproti prospektivnímu či retrospektivnímu typu jsou proměnné měřeny na různých (ale porovnatelných) objektech. V panelu se tedy obměňují objekty pro různá období t. Data pro každé období t lze považovat za oddělenou prostorovou (průřezovou) studii (*Pure Cross-Sectional Data*).

Panelový průzkum umožňuje pozorovat oproti prostorovým datům kauzalitu. Je možné tedy podle Ruspiniho (2002) zachytit sociální, ekonomickou, kulturní nebo politickou změnu a jiné

sociální procesy na úrovni jednotek výzkumu. Odborníci mohou pomocí panelové studie sledovat vývoj postojů, charakteristik nebo chování. Panelová studie tak umožňuje nejen zkoumání trendů u sledované skupiny jedinců, ale také poskytuje informace o tom, jak se změnily názory, postoje a hodnoty jedinců během určitého období. Ambicióznější výzkumy se pokoušejí zjišťovat i možnou příčinu změn či zvratů, které nastaly v průběhu dotazování mezi jednotlivými časovými okamžiky *t*. Panelové průzkumy se začínají využívat od přelomu třicátých a čtyřicátých let minulého století zejména v amerických výzkumech veřejného mínění. Již od počátku byly spojeny se snahou zachytit a měřit změnu na úrovni jedinců – zachytit hrubou změnu. Za tuto dobu vznikla řada velkých národních panelových výzkumů, které se zaměřují na studium změny – změny příjmů, dynamiku zaměstnanosti, chudoby, exkluze, změny rodinného a ekonomického statusu či vliv sociálních politik na způsob života jedinců. Zvýšený zájem o zachycení změn a porozumění jim na úrovni jedinců v jednotlivých státech vedla výzkumníky v poslední době k tendenci vytvořit porovnatelné panelové soubory, viz Lechnerová (2009).

Panelový soubor lze členit dle typu výběrového souboru na mikropanel a makropanel:

- *Mikropanel* se vyznačuje tím, že se jedná o soubor čítající mnoho objektů (běžně v řádech stovek, či tisíců). Objekty jsou periodicky pozorovány během krátkého časového úseku (zhruba během 2 až 10 let). Jedná se například o průzkumy domácností, zdravotní vyšetření pacientů apod.
- *Makropanel* obsahuje menší počet objektů (jednotky, desítky) a tyto objekty jsou navzájem porovnávány většinou ročně během dlouhého časového horizontu (délka studie je běžně 20–60 let). Tyto dlouhodobé studie slouží nejčastěji k porovnání různých ukazatelů zemí.

V souvislosti s analýzou panelových dat se pro individuální časovou řadu (opakovaná měření v čase pro daný objekt) vžil termín individuální trajektorie nebo růstová křivka.

Na časové řady a prostorová data lze nahlížet jako na zvláštní formu panelových dat, kdy ve srovnání s panelovými daty jsou časové řady data o jednom nebo malém počtu (v řádu jednotek) objektů a velkém počtu časových období a prostorová data jsou naopak data, kdy máme velké množství objektů (reprezentativní výběrový soubor s dostatečným rozsahem) a pouze jeden časový okamžik. V panelových datech jsou opakovaná měření pro daný objekt zkorelována. Na základě empirického výzkumu v oblasti biologie a medicíny Fitzmaurice a kol. (2011) zjistili vlastnosti vzájemné párové lineární závislosti mezi měřeními pro daný objekt vyjádřené Pearsonovým korelačním koeficientem, tzv. vnitroobjektové (intraindividuální) korelace: (1) míra závislosti (hodnota Pearsonova korelačního koeficientu) mezi dvěma měřeními často

klesá (k nule) s rostoucím časovým odstupem, (2) korelace mezi měřeními je výjimečně nulová, přestože jsou od sebe vzdáleny mnoho let, (3) korelace mezi párem opakovaných pozorování naměřených během krátkého časového intervalu je zřídka úplná (rovna jedné v absolutní hodnotě), (4) korelace jsou většinou pozitivní. Na základě svých zkušeností zjistili, že existují tři zdroje variability, které mají potenciální dopad na korelaci mezi opakovanými měřeními pro daný objekt:

- Interindividuální (meziobjektová) heterogenita – rozdíly v úrovni absolutních hodnot sledovaných proměnných mezi objekty jsou přirozené a reflektují variabilitu dopadu vysvětlující proměnné na hodnoty vysvětlované proměnné (některé jednotky reagují ve vyšší míře než průměr a jiné konzistentně podprůměrně). Proto jsou si absolutní hodnoty opakovaných intraindividuálních měření bližší než interindividuální. Objekty se mohou lišit rovněž tvarem trajektorií.
- Intraindividuální (vnitroobjektová) heterogenita – u objektů (například u lidí) hodnoty mnoha proměnných fluktuují v relativně krátkém časovém intervalu. Fluktuaci způsobuje například teplota, světlo, roční sezóna, denní cyklus. Vlivem nejen těchto faktorů dochází ke změnám, např. pracovních výkonů zaměstnanců, nálad a názorů respondentů a změnám biologických proměnných (teplota, tlak, atd.).
- Chyba měření – přesnost měření se vyjadřuje koeficientem reliability (spolehlivosti).

Vnitroobjektová korelace má tendence snižovat se s rostoucím časovým odstupem z důvodu interindividuální biologické variability a intraindividuální variability ve tvaru trajektorií. Následkem intraindividuální heterogenity je korelace výjimečně nulová. Chyba měření způsobuje to, že korelace není úplná (rovna jedné v absolutní hodnotě).

Následkem ignorování výskytu vnitroobjektové korelace jsou zkreslené odhady výběrové směrodatné odchylky odhadů regresních parametrů. Z výše uvedených důvodů vyplývá, že je třeba k analýze panelových dat přistupovat specifickým způsobem. Modelování panelových dat v sobě obsahuje prvky regresní analýzy a analýzy časových řad. Z povahy dat se liší i analytický přístup k mikropanelu a makropanelu. Baltagi (2012) uvádí odlišné problémy, se kterými se panely potýkají. Pro makropanel, tj. jedná se o dlouhé časové řady, je nutno vypořádat se s nestacionaritou v časových řadách, strukturálními zlomy (*structural breaks*), v čase měnícím se efektem vysvětlující proměnné na vysvětlovanou proměnnou a kointegrací. V mikropanelu se kointegrace z důvodu malého počtu časových okamžiků T nevyskytuje. Naopak v makropanelu, například jde-li o agregované ukazatele za země, je třeba se vypořádat s vzájemnou zá-

vislostí mezi podobnými zeměmi (*cross-country dependence*). Z toho plyne, že kromě autokorelace hodnot ukazatele uvnitř každé časové řady (intraindividuální korelace) je nutno v regresním modelování zohlednit i korelaci mezi časovými řadami (interindividuální korelace) za účelem nezkresleného odhadu regresních parametrů. V mikropanelu se interindividuální korelace řeší méně často, jelikož objekty (např. domácnosti) jsou vybírány náhodně a je málo pravděpodobná vzájemná korelace. Lze tedy shrnout, že pro analýzu mikropanelu se využívají vícerozměrné statistické modely modifikované o možnost zahrnout vliv času, zatímco pro makropanel vychází statistická teorie spíše z analýzy časových řad. Dalším rozdílem je fakt, že mikropanelová data za jednotlivé objekty mohou být přesněji měřena než obdobné proměnné na makro úrovni. Zkreslení vyplývající z agregace více objektů může být redukováno, nebo eliminováno, pokud se v analýze závislosti vysvětlovaných proměnných na zvolených vysvětlujících proměnných zohlední působení doprovodných proměnných (*covariates*), které pokrývají relevantní informaci o vnitroskupinových a meziskupinových rozdílech mezi objekty z různých skupin (např. zemí).

1.2 Příklady panelových šetření v ekonomických a neekonomických oblastech

Panelová data jsou využívána v různých vědních oborech, např. v ekonomii, kdy výzkumníci modelují pomocí panelových dat vztah mezi lidským kapitálem a ekonomickým vývojem regionu. Mullin (2010) v kontextu teorie lidského kapitálu využívá makroekonomická data z let 1950–2000 za různé regiony USA a modeluje závislost mezi ekonomickým vývojem a investicemi do vzdělání. Mullin vychází mj. z empirické studie Baldwin a Borelli (2008), kteří modelují ekonomický růst (vysvětlovaná proměnná) na investici do lidského kapitálu a dalších vysvětlujících proměnných. Za tímto účelem aplikují hierarchické modely (viz kapitola 2) na makropanelová data. Jedním ze základních předpokladů teorie lidského kapitálu, ze kterého tyto studie vycházejí, je ochota jednotlivce investovat do vzdělání svůj čas (a tedy obětovat ušlé příjmy), dokud tato investice nepřevyšuje očekávaný vyšší budoucí finanční a nefinanční užitek, viz Becker (1993).

Statistický úřad USA (U.S. Census Bureau) sbírá každoročně od roku 1976 mikrodata o severoamerických firmách. Data z této databáze „Longitudinal Business Database“ vyhodnocuje Centrum ekonomických studií při Statistickém úřadu USA. Centrum má mandát zkoumat tato data za účelem získání nových vhledů na formování a růst hospodářství, podstatu a vývoj konkurenčního boje, vývoj na trhu práce, podstatu a vývoj hospodářských cyklů, zdroje produktivního hospodářského růstu a vztahy mezi bankovním a podnikohospodářským sektorem. Jedná se o obrovský panelový soubor, kdy úřad na svých webových stránkách (Center for economic

studies) uvádí, že má k dispozici data pro více než 8,5 milionů objektů za rok 2013. Více o databázi Jarmin a Miranda (2002).

V politické ekonomii se pomocí makropanelových dat řeší například otázka, zda ropa a přírodní bohatství brání zavedení demokracie a místo toho podporuje autoritativní vlády, více v Rossově studii (2012).

Často se panelová data uplatňují v oblasti financí, např. Boehmer (1990) pomocí panelových dat identifikuje determinanty solventnosti sekundárních tržních cen vládních syndikovaných půjček.

Panelové studie mají své místo i v managementu a marketingu. V marketingu výzkumníci zkoumají chování zákazníka ve vztahu ke značce (*brand positioning*) pomocí mikropanelových dat. Autoři Chintagunta (1994) a Elrod a Keane (1995) předpokládají invariantní vnímání atributů značky. O nákupu výrobku rozhodují jiné aspekty než značka a její atributy. Kannan (1994) pro svůj strukturální model trhu předpokládá významný efekt vyhledávání změn. Očekává, že zákazníci při volbě komoditního zboží mění značky z důvodu, že rádi zkoušejí nové věci (značky). Zmíněné modely preference značky mají společné to, že atributy značky nejsou rozhodující pro uskutečnění nákupu. Oproti těmto invariantním modelům Erdem (1996) ve své interní analýze struktury trhu zahrnuje do svého modelu dynamiku chování zákazníka při volbě značky. Erdemův model obsahuje heterogenitu v zákaznických preferencích a ve vnímání atributů značky. Givon (1984) odvozuje stochastický model volby značky jako nástroj pro segmentaci trhu. Někteří zákazníci jsou věrni značce, jiní značky střídají. Keane (1997) modeluje chování zákazníka při volbě značky pro často kupované výrobky. Prezentuje data z panelové studie společnosti Nielsen obsahující retailová data o nákupu kečupu různých značek sbírané čtečkou čárových kódů. Na základě těchto dat odhaduje pravděpodobnost nákupu stejné značky na základě předchozích zkušeností s touto značkou. Předpoklady různých tržních modelů testují Kahnová a kol. (1986) na rozsáhlém panelovém souboru a pomocí simulací.

V oblasti managementu (podnikové ekonomiky) byly panelové studie využity např. pro výzkum vztahu mezi individuálními pracovními výkony a celkovým výkonem celé společnosti. Deadrick a kol. (1997) zkoumali pracovní výkon určité skupiny pracovníků obsluhující pilu prvních 24 měsíců po nástupu do zaměstnání. Nalezli významné rozdíly mezi jednotlivými zaměstnanci jak v počáteční efektivitě (absolutní člen individuální trajektorie), tak ve změně výkonu v průběhu času (směrnice individuální lineární trajektorie). Další ukázkou z této oblasti je Childův výzkum (1974; 1975) týkající se manažerského vlivu na celkový výkon firmy. Výkon firmy (vysvětlovaná proměnná) posuzuje z hlediska tržeb, hodnoty aktiv a výnosů během

5leté periody. Kromě výkonnosti se v této oblasti pomocí panelových studií řeší další témata, např. bezpečností práce se zabývali Hofmann a Stetzer (1998), zákaznickou spokojenost pozorovali Jammieson a Griffin (1998) a zvyšování kvalifikace zaměstnanců zajímalo Eyringa a kol. (1993). Jak správně modelovat panelová data v oblasti podnikové ekonomiky popisují Hofmann a kol. (2000).

Častěji jsou panelová data užívána v neekonomických oblastech (medicína, vzdělávání, psychologie, politické vědy, ekologie, zoologie a další). Metody analýzy panelových dat jsou rozvíjeny zejména biostatistiky a výzkumníky z těchto neekonomických oborů. Správné aplikaci analýzy longitudinálních dat v medicíně je věnována například monografie Browna a Prescotta (2006). Chua a kol. (2015) porovnávají různé lineární modely se smíšenými efekty na základě mikropanelových dat a navrhuje vhodný model pro modelování atrofie mozku a vývoje lézí u pacientů s roztroušenou sklerózou. Analýzou panelových studií v oblasti vzdělávání se zabývá Goldsteinova monografie (1987). Výzkumníci v oblasti psychologie vycházejí ze Singerovy a Willettovy monografie (2003). Z politických věd bych rád zmínil zajímavý článek autorů Becka a Katze (1995), ve kterém autoři navrhují vlastní odhad směrodatných chyb parametrů regresního modelu vhodného pro analýzu panelových dat. Svůj přístup demonstrují na „sociálně demokratickém korporativistickém“ modelu. Na sociologické výzkumy je zaměřena Raudenbushova a Brykova monografie (2002). V sociologické studii se zabývají rozdíly mezi platy mužů a žen England a spol. (1988). V empirické studii predikují rozdíly mezi příjmy mladých mužů a žen pomocí Národního longitudinálního panelu. V oblasti ekologie patří mezi průkopníky moderní analýzy longitudinálních dat B. M. Bolker. Návod, jak správně používat jednu z nových metod (zobecněný lineární model se smíšenými efekty), popisují Bolker a spol. (2008). V České republice využívá v oblasti ekologie mikropanelová data Forbelská a kol. (2015). V této studii vyhodnocuje obsah těžkých kovů v sedimentech řeky Moravy. Správné užívání modelů určených pro analýzu panelových dat na příkladech ze zoologie nabízí Pekár a Brabec (2012).

S mikropanelovými datovými daty pracuji v rámci svého empirického výzkumu. Finanční mikropanelová data analyzuji ve spolupráci s kolegy z Univerzity Mateje Bela v Banské Bystrici. Máme k dispozici dva datové soubory slovenských průmyslových podniků, které obsahují údaje vybraných finančních výkazů z ročních účetních závěrek, jmenovitě rozvahy a výkazu zisku a ztráty. První soubor obsahuje údaje 190 společností z let 2002–2004. Druhý soubor zahrnuje roční data 6921 podniků z let 2009–2013. Na základě těchto dat odhadujeme predikční model finanční tísně podniku. Finanční tíseň podniku je situace, ve které není firma schopna

dostát svým finančním závazkům. Stanovit to, zda se nachází firma ve finačních problémech, se stalo předmětem zájmu mnoha studií počínaje známým Altmanovým Z–skóre modelem, viz Altman (1968). Několik podobných studií je založeno na statických klasifikačních modelech konstruovaných různými statistickými metodami, jako je například diskriminační analýza, logistická regrese a rozhodovací stormy, viz Balcean a Ooghe (2006) a Brezigar–Masten a Masten (2012). Novější studie kombinuje účetní, tržní a makroekonomická data, viz Tinoco a Wilson (2013). Avšak ve většině studií modely vychází z historických účetních dat a z nich dopočítaných finančních ukazatelů. Tyto studie předpokládají, že minulé hodnoty vhodně zvolených finančních a ekonomických ukazatelů dokážou určit finanční zdraví podniku v budoucnosti. Bohužel tento mikroekonomický přístup trpí známými nedostatky. V příspěvku Kráľa a kol. (2014b) jsme proto přišli s návrhem jednoduché metodiky, jak zakomponovat informace o změnách finančních ukazatelů do používaného statického přístupu k modelování finanční tísně podniku. Výsledek analýzy indikuje, že zahrnutí dynamiky do modelu může vést ke zlepšení predikční schopnosti klasifikátorů (modelů). Pro modelování jsme využili 3letá data a v dalším výzkumu jsme pak zjistili, že klasifikátor odhadnutý z delších longitudinálních dat (5letá data má podobnou, ale důvěryhodnější chybu klasifikace, Stachová a kol. (2015). Dále jsem s kolektivem autorů analyzoval jiný panelový soubor z oblasti pojišťovnictví. Cílem analýzy byla identifikace prediktorů a prediktivních modelování stornovosti životních smluv ze strany klienta. Výsledky a metodologie postupu jsou popsány v práci Stachové a Sobíška (2013).

V oblasti medicíny se věnuji výzkumu roztroušené sklerózy (RS) ve spolupráci s kolektivem autorů z 1. lékařské fakulty Univerzity Karlovy a jiných celosvětově uznávaných pracovišť (Melbourne Brain Centre). Náším hlavním cílem je identifikovat potenciální (klinické a paraklinické) biomarkery budoucího vývoje klinické a kognitivní míry disability pacientů s výskytem této nemoci. K dispozici máme dvě kohorty (mikropanelové soubory) pacientů. V prvním souboru ASA (Uher a kol., 2016) máme aktuálně 12-ti letá data pro 181 pacientů. Ve druhém souboru SET, viz Uher a kol. (2016b) sledujeme 220 pacientů 6 let. V této oblasti jsme publikovali například článek, viz Uher a kol. (2016), ve kterém jsme odhadovali spolehlivost klinických a rezonančních (MRI) parametrů predikovat vývoj klinické disability RS během 12 let. Naším dalším výzkumným záměrem je odhadnout predikční modely responze pacienta na konkrétní léčbu. Nasazení vhodné terapie na počátku nemoci je zásadní pro pozitivní modifikaci dlouhodobého vývoje RS. V dnešní době existuje několik kategorií léčby. Tyto modely by měly sloužit v klinické praxi jako podpůrný nástroj pro rozhodování o nasazení vhodné a efektivní léčby. Za tímto účelem využíváme tři různé mikropanely pacientů s RS.

1.3 Výhody a nevýhody panelového výzkumu

Nespornou výhodou panelového výzkumu je fakt, že oproti prostorovým datům a příliš agregovaným časovým řadám je možné analyzovat a testovat komplikovanější hypotézy dynamiky a změny v chování objektů. Panelová data poskytují více informací, díky nimž lze lépe hledat latentní vztahy mezi objekty a proměnnými.

Hsiao (2003) uvádí několik výhod, jež vyplývají z užívání panelových dat:

- Patří k nim např. přehled o individuální heterogenitě. V rámci panelového souboru lze očekávat, že porovnávané objekty (jednotlivci, firmy, státy nebo země) jsou heterogenní. Časové řady a prostorové studie, které individuální heterogenitu neberou v potaz, se vystavují dle Mounltona (1986; 1987) riziku získání zkreslených výsledků.
- Panelová data jsou vhodnější pro studium dynamiky změny. Lze tak například sledovat, s jakou dynamikou reaguje míra nezaměstnanosti na změnu hospodářské politiky. Oproti prostorovým datům, která nám poskytují pouze míru nezaměstnanosti v daném období, panelová studie obsahuje např. informaci, zda nezaměstnaný v předchozím období (t – 1) je stále nezaměstnaný i v období t po změně politiky atd.
- Panelová data umožňují lépe identifikovat a měřit efekty, které nelze jednoduše detekovat u prostorových dat a časových řad. Příklad je uveden ve studii Winkelmanna a Winkelmanna (1998), kdy v případě dotazování pomocí škály vnímá každý respondent hodnoticí škálu odlišně a je vhodnější parametry populačního modelu odhadovat pomocí změn pro každého jednotlivce než porovnávat změny mezi jednotlivci (přístup klasického lineárního modelu).
- Modely panelových dat nám umožňují konstruovat a testovat komplikovanější modely chování. Například Erdem (1996) vytváří dynamický model nákupních preferencí zákazníků a dochází k závěru, že předchozí statické modely vycházející z předpokladu ceteris paribus, které nezohledují vývoj zákaznických preferencí, vedou ke zkresleným závěrům.
- Z důvodu velkého počtu pozorování v čase má test jednotkových kořenů u makropanelových dat standardní asymptotické rozdělení na rozdíl od časových řad, kde se problém s nestandardním rozdělením vyskytuje.

Nevýhody panelových šetření spatřuje Hsiao (2003) zejména v následujících skutečnostech:

- Problémy spojené s datovým managementem a sběrem dat. Problémy popisují Kasprzyk a kol. (1989). V medicínském výzkumu jsem se setkal s retrospektivním dohledáváním

dat. Zpětně doplněná data již nemusí přesně odpovídat původní hodnotě. Retrospektivní dohledávání dat je bohužel běžný fenomén.

- Zkreslení způsobené chybou měření. Chyby měření („špatné odpovědi“) dle Kaltona (1989) mohou vzniknout z důvodu nepochopení dotazů, zapomětlivosti, záměrně zkreslené odpovědi respondenta, nevhodné volbě tazatele, chybě způsobené tazatelem. S chybou měření jsem se rovněž setkal při analýze longitudinálních měření objemu mozku pomocí magnetické rezonance. Relativně velká vnitroobjektová variabilita měření je způsobena z části fyziologickými změnami u pacienta. Faktorem může být například dehydratace pacienta nebo změna léčby.
- Problém selektivity: autoselekce (*self–selectivity*), nemožnost získat odpovědi od respondenta a to nejčastěji v první vlně (*nonresponse*), ale i nemožnost získat odpověď v následujících vlnách (*attrition*). S tímto problémem jsem se setkal rovněž v klinickém výzkumu, kdy ze studií vypadnou pacienti z různých důvodů. Pokud důvod odchodu ze studie může souviset se studií, je to třeba v analýze zohlednit.
- Krátká dimenze časové řady u mikropanelů. Pro zhodnocení efektu léku na modifikaci dlouhodobého vývoje nemoci nám někdy chybí dostatečně dlouhá doba sledování pacienta na léčbě.
- Průřezová závislost (*cross–section dependence*). Makropanely (zemí, regionů) s dlouhou časovou řadou, které opomíjejí vzájemnou meziskupinovou (mezi zeměmi) závislost, vedou ke zkresleným závěrům.

2. Regresní analýza normálně rozdělené longitudinální vysvětlované proměnné

U panelových dat nejsou splněny některé Gaussovy-Markovovy předpoklady pro klasický lineární model. Uvažujme klasický lineární regresní model (KLM) ve tvaru

$$y_{it} = \mu + \sum_{j=1}^{J} \beta_j x_{jit} + \varepsilon_{it}, \tag{2.1}$$

kde y_{it} označuje hodnotu vysvětlované proměnné Y pro i-tý objekt (i = 1, 2, …, N), v čase t (t = 1, 2, …, T), závislé na J exogenních (nenáhodných) vysvětlujících proměnných X_j (j = 1, 2, …, J). Hodnota j-té proměnné X_j pro i-tý objekt v čase t je značena x_{jit}. Pro náhodnou složku ε_{it} platí, že její rozdělení je normální, střední hodnota je nulová, tedy $E(\varepsilon_{it}) = 0$ pro každé měření it. Rozptyl náhodné složky $D(\varepsilon_{it}) = \sigma^2$ je konstantní pro každé pozorování a kovariance náhodné složky je nulová $Cov(\varepsilon_{it}, \varepsilon_{i't'}) = 0$ pro každé $i \neq i'$, nebo $t \neq t'$. Předpoklady modelu o náhodné složce lze pak zkráceně zapsat $\varepsilon_{it} \sim N(0;\sigma^2)$. Parametry modelu jsou μ a β_j, j = 1, 2, …, J. Regresní parametry β_j vyjadřují marginální a konstantní (pevný) efekt jednotlivých proměnných X_j na vysvětlovanou proměnou Y. Model předpokládá, že vypuštěné „individuální“ vlivy a efekty času, zahrnuté v parametru μ, nejsou korelovány s vysvětlujícími proměnnými. Lineární regresní model uvažuje pouze jedinou složku variability ε_{it}. Avšak v panelových datech se předpokládají různé zdroje variability blíže popsané v oddílu 1.1.

Dále lze očekávat korelaci (viz oddíl 1.1) u opakovaně naměřených pozorování u jednotlivců. Jejich pozorování jsou zkorelována, a tedy nejsou nezávisle identicky rozdělena (iid). Pozorování u jednotlivce jsou si podobnější než pozorování dvou jednotlivců. Tato korelace musí být v analýze zohledněna. Z těchto důvodů nelze pro odhadování vztahu mezi vysvětlovanou proměnnou Y a vysvětlující proměnou X (vysvětlujícími proměnnými X_j) použít klasickou analýzu rozptylu (ANOVA) nebo klasický lineární model (KLM). Důsledkem použití KLM na analýzu korelovaných dat jsou zkreslené odhady směrodatných chyb odhadů parametrů.

Hlavním cílem statistických modelů pro longitudinální data je zachycení individuálních změn vysvětlované proměnné v čase a vyjádření intraindividuální změny v závislosti na hodnotě vysvětlující proměnné (v případě jedné vysvětlující proměnné; Fitzmaurice a kol., 2011). V této kapitole popisuji různé regresní přístupy vhodné pro normálně rozdělenou longitudinální

vysvětlovanou proměnnou. Historicky se vyvíjely dva směry, jak přistupovat k analýze panelových dat s normálně rozdělenou vysvětlovanou proměnnou: dvoukroková analýza a smíšené lineární modely (*Linear Mixed Models*).

Ve dvoukrokové analýze se odhadují individuální modely pro každý objekt a na základě individuálních modelů se odhaduje populační model (tento přístup je blíže popsán v oddílu 2.1). Z důvodu udržení únosného počtu parametrů modelu byl navržen smíšený lineární model (viz oddíl 2.2), ve kterém se neodhaduje individuální model pro každý objekt zvlášť.

2.1 Dvoukroková analýza

V přístupu dvoukrokové analýzy (v práci Hoffmana a Rovina (2007) rovněž známé pod označením *Parameter-As-Outcome Model*, PAOM) nebo *slopes as outcomes* podle Lorcha a Myerse (1990) se korelace měření pro daný objekt řeší dvoukrokově. V prvním kroku se odhadnou individuální regresní modely pro každý i-tý objekt zvlášť, $i = 1, 2, \ldots, N$, kde N značí počet objektů ve studii. Ve druhém kroku (sdružená analýza, meziobjektová analýza) se vyhodnocuje analyzovaný parametr modelu (například průsečík, směrnice pro lineární model nebo kvadratický člen pro parabolu) pomocí N odhadnutých individuálních parametrů. Ve sdružené analýze se provádí testování například pomocí standardních statistik t nebo F.

V dnešní době výzkumníci, např. Moscatelli a kol. (2012), stále častěji preferují simultánní přístup, ve kterém jsou pro odhad regresních parametrů (pevných efektů) populačního modelu využity hodnoty všech N objektů najednou, před dvoukrokovým přístupem z následujících důvodů:

- zachování informace o objektově specifické směrodatné chybě a z toho plynoucí odhad variability pro náhodné a pevné efekty zvlášť;
- PAOM nepřenese do druhého kroku informaci o počtu opakování měření pro i-tý objekt;
- PAOM implicitně přikládá stejnou váhu každému objektu ve druhém kroku;
- u PAOM je obtížnější vyhodnotit kvalitu modelu a učinit správný výběr parametrů modelu, jelikož v prvním kroku je předpokládán stejný směr trajektorie pro každý objekt. Tento předpoklad bývá v reálných studiích často porušen.

PAOM se i přes uvedené výtky stále využívá například u Johnstona a kol. (2008) nebo u De Stefana a kol. (2015). Mezi důvody preference PAOM patří nejednoznačnost odhadu náhodných efektů (regresních parametrů) a nevhodnost využívat zavedená informační kritéria,

založená na věrohodnostní funkci např. Akaikeho informační kritérium (AIC), pro správnou volbu modelu (skupiny vysvětlujících proměnných), viz části 2.2 a 3.2.

Dvoukrokový přístup je preferován například v epidemiologii u analýzy výskytu nemocí, kde se výsledky jednotlivých (regionálních) studií liší z důvodu malých rozsahů výběru a regionální variability. Z těchto důvodů výzkumníci provádějí metaanalýzu a přezkoumávají závěry jednotlivých studií výskytu nemoci pomocí sdružených (*pooled*) dat získaných z více studií. Pro každou zemi se odhaduje specifický model zohledňující matoucí faktory (*confounders*). Simulace provedená v práci Stukel a kol. (2001) ukazuje lepší, nezkreslené odhady regresních parametrů a jejich středních chyb u dvoukrokového přístupu. Naproti tomu simultánní model nadhodnocuje odhady regresních parametrů a podhodnocuje střední chyby odhadů v případě výrazné heterogenity jednotlivých studií. Navíc, smíšený model nelze aplikovat v případě studií lišících se v počtu a v různé metodice měření a kódování vysvětlujících proměnných.

Stejní autoři dále upřednostnili dvoukrokový přístup před simultánním modelem s náhodným sklonem (růstový model), vysvětleném v oddílu 2.2, v jiné studii u Stukela a Demidenka (1997). Autoři tvrdí, že dvoukrokový přístup je relevantnější pro analýzu, ve které není výzkumným záměrem odhadovat regresní parametry růstové funkce, ale pozornost je soustředěna na analýzu efektu vysvětlujících proměnných na jiné růstové proměnné (například míra přírůstku pro kvadratický model, asymptotická úroveň atd.). Sledované proměnné jsou vyjádřeny jako lineární kombinace parametrů růstového modelu. Autoři navrhli dvoukrokový estimátor ekvivalentní iterativní obecné metodě vážených nejmenších čtverců (IRLS), který má asymptoticky optimální vlastnosti pro velký počet objektů. V prvním kroku odhadují individuální regresní modely klasickou metodou nejmenších čtverců (OLS). V druhém kroku odhadují pomocí modelu s náhodným sklonem sledované proměnné, vyjádřené jako lineární kombinace parametrů odhadnutých v prvním kroku. Dvoukrokový přístup vyšel v porovnání s modelem s náhodným sklonem lépe na simulovaných a reálných datech, kde odhadovali efekt vysvětlující proměnné (věku) na míru přírůstku (kvadratický člen).

Ve prospěch PAOM bych rád zdůraznil fakt, že poslední uvedenou nevýhodu PAOM (individuální trajektorie nelze často popsat stejným typem modelu) lze redukovat tím, že se v rámci exploratorní analýzy naleznou shluky (skupiny) objektů s podobným tvarem trajektorie vysvětlované proměnné pomocí vhodného shlukovacího algoritmu. Na každý shluk se provede PAOM zvlášť a výsledné modely pro jednotlivé shluky se navzájem porovnají. Shlukovacími algoritmy pro mikropanelová data se zabývají kapitoly 5 až 7.

2.2 Smíšený lineární model (LMM)

Obecný lineární model (*General Linear Model,* **GLM**) umožňuje explicitně modelovat širokou škálu kovariančních a korelačních struktur náhodné složky ε_{it}. Z tohoto důvodu se modifikace GLM používají pro analýzu panelových dat. Mezi regresní modely, které se používají pro analýzu panelových dat a jsou specifickou formou GLM, se řadí modely s pevnými (fixními) efekty (*Fixed Effects Models,* **FEM**), s náhodnými efekty (*Random Effects Models,* **REM**), se smíšenými efekty (*Mixed Effect Models,* **MEM**); souhrnně je můžeme zařadit do smíšených lineárních modelů (*Linear Mixed Models,* **LMM**). Efekt (vysvětlující proměnné na vysvětlovanou proměnnou) je ekvivalent regresního parametru používaný v oblasti regresního modelování panelových dat.

V ekonometrii a statistice je model s pevnými efekty statistický model, který považuje parametry všech vysvětlujících proměnných za nenáhodné[1], zatímco modely s náhodnými efekty a se smíšenými efekty vnímají všechny (REM) nebo některé (MEM) parametry vysvětlujících proměnných jako náhodné[2]. Tato definice pevných a náhodných efektů je však v rozporu s definicí, která je zavedena v biostatistice u Lairda a Wareho (1982); Diggla a kol. (2002); u Fitzmaurice a kol. (2011). Biostatistici definují pevné efekty jako regresní parametry vysvětlujících proměnných, jejichž hodnoty jsou sledovány primárně a mohou být opakovaně použity v případě opakování experimentu. Náhodné efekty jsou regresní parametry, jejichž hodnoty nejsou primárně sledovány a jsou vybírány náhodně z velkého množství hodnot. Pevné efekty nám tedy cíleně dělí základní soubor do sledovaných dílčích skupin (např. věkové kategorie) a je díky nim odhadován populační regresní parametr (pro danou věkovou kategorii). S pevnými efekty není spojena vnitroobjektová korelace. Náhodné efekty reprezentují obecnou variabilitu mezi objekty (např. individuální vlastnosti, schopnosti, znalosti jednotlivce). Přesná definice pevného a náhodného efektu je kontroverzní. Gelman (2005) porovnává pět rozdílných definic pevných a náhodných efektů:

- Pevné efekty jsou konstantní napříč objekty uvnitř skupiny (výběru) a náhodné efekty se liší, viz Kreft a De Leeuw (1998).
- Efekty jsou pevné v případě, že výzkumným záměrem je odhad jejich hodnot (efekt vysvětlující proměnné na vysvětlovanou proměnnou). Zatímco efekty jsou náhodné v případě, že nás zajímá pouze rozdělení hodnot těchto náhodných veličin (efektů). Na

[1] Nenáhodné parametry (efekty) jsou pevné (konstantní) pro všechny objekty.
[2] Náhodné parametry (efekty) nejsou pevné pro všechny objekty, ale jsou to náhodné veličiny.

základě této definice lze měnit status se změnou cíle odhadu (analýzy) při nezměněném návrhu studie a se stejným výběrem, viz Searle a kol. (1992).

- Efekt označíme jako pevný, pokud je k dispozici úplný (konečný) výběr hodnot x_{jit} vysvětlující proměnné X_j v základním souboru. Naopak pokud výběr obsahuje pouze malý počet hodnot z celého oboru hodnot vysvětlující proměnné, označíme efekt (regresní parametr pro tuto vysvětlující proměnnou) jako náhodný. Tato definice zavádí pojem úplný výběr hodnot, viz Green a Tukey (1960).
- Za předpokladu, že efekt je realizovaná hodnota náhodné veličiny, jedná se o náhodný efekt dle Lamotta (2014).
- Pevné efekty jsou odhadnuty pomocí maximální věrohodnosti (nebo metodou nejmenších čtverců). Náhodné efekty jsou odhadovány metodou *shrinkage* (poskytuje podle Snijderse a Boskera (2012) lineární nezkreslený odhad. Tato definice odpovídá ekonometrické definici. Neodpovídá však na otázku jak rozlišit pevný a náhodný efekt. Předchozí čtyři definice odpovídají biostatistické definici.

Dále se může status (pevný nebo náhodný) efektu v modelu lišit dle návrhu experimentu nebo dle kontextu. Oproti předchozím definicím se definice pevných a náhodných efektů významně liší také v bayesovské statistice. Pevné efekty jsou zde odhadovány podmíněně na základě variability z celého (a nekonečného) oboru hodnot. Náhodné efekty jsou odhadovány podmíněně z oboru hodnot posteriorního rozdělení.

LMM se používá nejen pro analýzu panelových dat (*repeated measures design*), ale i pro jiný návrh studie, a to pro studie sdružených dat (*clustered data* nebo *nested design*). Sdružená data odpovídají studii, ve které jsou měřeny náhodně vybrané objekty náležející do určité skupiny (*cluster, block, site*), např. data jsou náhodně sbírána u studentů v jednotlivých třídách. Náležitost do třídy je vnímána jako pomocná proměnná, která není v popředí zájmu výzkumu, avšak je třeba ji brát v potaz při analýze vztahu mezi Y a X.

Modely s pevnými efekty (FEM), modely s náhodnými efekty (REM) a modely se smíšenými efekty (MEM) jsou stručně popsány v následujících částech tohoto oddílu. Dále se zde nachází přehled metod odhadu parametrů modelu a je vysvětlen pojem hierarchický lineární model. Více o využití jednotlivých typů LMM pro různé druhy návrhu experimentu viz Chamberlain (1982), Raudenbush a Bryk (2002), Pekár a Brabec (2012).

Modely s pevnými efekty (FEM)

Model s pevnými efekty pro i-tý objekt v čase t, který lze zapsat ve tvaru

$$y_{it} = \mu + \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + \varepsilon_{it}, \tag{2.2}$$

obsahuje nepozorované specifické pevné (fixní) efekty α_i, které jsou nezávislé na čase (v čase fixní) a obsahují v sobě individuální efekty každého i-tého objektu. Individuální efekty α_i reprezentují odchylku i-tého pozorování od společné střední hodnoty μ. Předpokládá se, že heterogenita individuálních pevných efektů α_i je v čase konstantní a mohou být korelovány s hodnotami $\mathbf{x}_{it}$ vysvětlujících proměnných $\mathbf{X}$, kde $\mathbf{x}_{it} = (x_{1it}, \ldots, x_{Jit})'$ značí vektor pozorování vysvětlujících proměnných pro i-tý objekt v čase t. Přesto musí být striktně exogenní. Vektor regresních parametrů $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_j, \ldots, \beta_J)$ obsahuje J pevných efektů, které kvantifikují v čase konstantní působení j-té vysvětlující proměnné X_j na vysvětlovanou proměnnou Y. Náhodná složka ε_{it} reprezentuje efekty nevýznamných proměnných příznačných i-tým pozorováním. Náhodná složka ε_{it} má stejné vlastnosti jako v KLM. Pochází z nezávisle identického (normálního) rozdělení s nulovou střední hodnotou a konstantním rozptylem, $\varepsilon_{it} \sim$ iid $N(0;\sigma^2 I)$, kde I je jednotková matice. Zároveň je náhodná složka nekorelovaná s vektorem hodnot vysvětlujících proměnných $\mathbf{x}_{it}$. Vektor hodnot náhodné složky (teoretických reziduí) pro i-tý objekt $\boldsymbol{\varepsilon}_i'$ obsahuje T teoretických reziduí $\boldsymbol{\varepsilon}_i' = (\varepsilon_{i1}, \ldots, \varepsilon_{it}, \ldots \varepsilon_{iT})'$. Autokovariance vektorů teoretických reziduí pro různé objekty se rovná $E(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_{i'}') = 0,$ kde $i \neq i'$. Odhady parametrů μ, α_i a $\boldsymbol{\beta}$ obdržíme pomocí metody nevážených nejmenších čtverců (*Ordinary Least Square Method* – OLS). Model s pevnými efekty je také označován jako model analýzy kovariance (*Analysis of Covariance*). Bližší popis modelu lze nalézt například v Novákově práci (2007).

Modely s náhodnými efekty (REM)

Lineární prediktor η_{it}[3] REM obsahuje pouze náhodné efekty na rozdíl od FEM, který obsahoval pouze pevné efekty. Žádný z náhodných efektů (regresních parametrů) není konstantní pro všechny jednotky, ale jsou to náhodné veličiny, které mají normální rozdělení s nulovou střední hodnotou a konečný konstantní rozptyl. REM umožňuje modelovat individuální efekty. West a kol. (2014) uvádí, že první známá formulace jednorozměrného modelu s náhodným efektem (bez pevného efektu) patří Airymu (1861). Tento model byl později objasněn Scheffém v roce 1956. Airy provedl několik teleskopických měření během jedné noci. Experiment opakoval několik nocí a analyzoval data s využitím rozkladu rozptylu na náhodný efekt noci a na reziduální variabilitu uvnitř noci. Tento příklad demonstruje podstatu REM. Celková náhodná složka bude

[3] *Lineární prediktor η_{it} je systematická část regresního modelu. V KLM se rovná odhadované střední hodnotě Y_{it}, tj. $E(Y_i) = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$. V zobecněných lineárních modelech (ZLM) se η_i rovná transformované střední hodnotě, tj. $g(E(Y_i)) = \eta_i = \mathbf{X}_i\boldsymbol{\beta}$, kde $g(.)$ je hladká, striktně monotónní funkce, tzv. spojovací funkce (linkovací, transformační), která propojuje neznámé střední hodnoty Y a vysvětlující proměnné* **X**.

dále (pro REM a MEM) značena v_{it}. REM rozkládá náhodnou složku na náhodnou chybu ε_{it} a náhodné efekty (zde jedna proměnná efekt noci). Náhodná chyba ε_{it} věcně reprezentuje totéž, co náhodná složka ε_{it} v KLM a FEM. Na rozdíl od FEM není individuální efekt α_i v čase fixní, ale je považován za náhodnou veličinu, která má stejné vlastnosti jako náhodná chyba ε_{it}. Navíc zde rozlišujeme třetí skupinu náhodných efektů λ_t, jež jsou vlastní specifickým časovým úsekům a mají podobný dopad na chování jednotlivých objektů. Náhodnou složku v_{it} modelu REM lze zapsat pomocí komponent

$$v_{it} = \alpha_i + \lambda_t + \varepsilon_{it}, \tag{2.3}$$

přitom předpokládáme, že $v_{it} \sim$ iid N(0; σ_v^2). Pomocí dekompozice náhodné složky v_{it} lze vysvětlit variabilitu proměnné Y. Rozptyl Y je potom součtem rozptylů jednotlivých komponent náhodné složky v_{it}, tedy

$$\sigma_Y^2 = \sigma_v^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_\varepsilon^2. \tag{2.4}$$

Pro jednotlivé náhodné komponenty platí

$$E(\alpha_i) = E(\lambda_t) = E(\varepsilon_{it}) = 0, \tag{2.5}$$

$$E(\alpha_i \lambda_t) = E(\alpha_i \varepsilon_{it}) = E(\lambda_t \varepsilon_{it}) = 0, \tag{2.6}$$

$$E(\alpha_i \alpha_i) = \sigma_\alpha^2, E(\alpha_i \alpha_{i'}) = 0 \text{ pro } i \neq i', \tag{2.7}$$

$$E(\lambda_t \lambda_t) = \sigma_\lambda^2, E(\lambda_t \lambda_{t'}) = 0 \text{ pro } t \neq t', \tag{2.8}$$

$$E(\varepsilon_{it} \varepsilon_{it}) = \sigma_\varepsilon^2, E(\varepsilon_{it} \varepsilon_{i't'}) = 0 \text{ pro } i \neq i', \text{ nebo } t \neq t'. \tag{2.9}$$

Ze vzorců 2.5 až 2.9 je zřejmé, že kromě vzájemné lineární nezávislosti náhodných efektů α_i a λ_t a náhodné chyby ε_{it} je předpokládána jejich nezávislost přes různé objekty. Dále je předpokládána normalita těchto komponent. Jednotlivé náhodné efekty jsou individuální odchylky od konstantní střední hodnoty μ a REM lze zapsat

$$y_{it} = \mu + v_{it}. \tag{2.10}$$

REM se používá v případě, že výzkumným zájmem je vysvětlit zdroje variability v datech.

Modely se smíšenými efekty (MEM)

Předchozí modely 2.2 a 2.10 uvažují buď pevné, nebo náhodné efekty. Lineární prediktor modelu MEM obsahuje alespoň jeden pevný a jeden smíšený efekt

$$y_{it} = \mu + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + \lambda_t + \varepsilon_{it}, \tag{2.11}$$

kde $\mathbf{x}_{it}$ značí vektor hodnot J vysvětlujících exogenních proměnných. Vektor $\boldsymbol{\beta}'$ obsahuje J pevných efektů (regresních parametrů), jejichž hodnoty jsou konstantní pro různé objekty, a je nezávislý na náhodné složce v_{it} viz vzorec 2.3. Vysvětlení variability zařazením náhodného efektu (náhodných efektů) λ_t, α_i do MEM umožňuje na rozdíl od FEM extrapolovat hodnoty pro objekty nezařazené do výběru. Laird a Ware (1982) popisují teorii k jednokrokovému odhadu modelu s náhodným efektem. Do té doby byl model s náhodným parametrem odhadován dvoukrokově, viz oddíl 2.1. Model (2.11) se zapíše pro i-tý objekt maticově pomocí Lairdova a Wareho vzorce

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{2.12}$$

$\mathbf{X}_i$ typu $(T_i \times J)$ a $\mathbf{Z}_i$ typu $(T_i \times Q)$ jsou matice plánu vysvětlujících proměnných, kde T_i značí počet měření (opakovaných v čase) a Q počet vysvětlujících proměnných Z_q pro objekt i. Vektor $\mathbf{Y}_i = (y_{i1}, \ldots, y_{iT_i})'$ obsahuje hodnoty vysvětlované proměnné Y pro i. Vektor hodnot pevných efektů $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$ typu $(J + 1) \times 1$ obsahuje konstantní efekty (regresní parametry) J vysvětlujících proměnných X_j.

Vektor náhodných efektů $\mathbf{b}_i = (b_{i1}, \ldots, b_{iQ})'$ typu $(Q + 1) \times 1$ obsahuje Q hodnot pevných efektů Q vysvětlujících proměnných Z_q. Náhodné efekty b_{iq} pro i-tý objekt jsou náhodné veličiny, které se řídí normálním rozdělením s nulovou střední hodnotou a konečným konstantním rozptylem ψ_q^2. Náhodné efekty odhadují odchylky od pevných efektů pro i-tý objekt. Vlastnosti vektoru b_i lze zapsat souhrnně jako $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0},\boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je pozitivně semidefinitní kovarianční matice náhodných efektů o rozměrech $Q \times Q$ obsahující hodnoty rozptylů jednotlivých náhodných efektů ψ_q^2 na diagonále a hodnoty kovariancí mezi jednotlivými náhodnými efekty $\psi_{qq'}$ mimo diagonálu. Hodnoty rozptylů ψ_q^2 a kovariance $\psi_{qq'}$ mezi náhodnými efekty b_q, $b_{q'}$ jsou předpokládány konstantní mezi objekty.

Náhodná chyba ε_{it} pro i-tý objekt se řídí vícerozměrným normálním rozdělením. Náhodné chyby pro i-tý objekt mají následující vlastnosti $\varepsilon_{it} \sim \mathrm{N}(0,\sigma^2\lambda_{itt})$, $\mathrm{Cov}(\varepsilon_{it},\varepsilon_{it'}) = \sigma^2\lambda_{itt'}$. Vlastnosti vektoru hodnot náhodných chyb pro i-tý objekt $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{it}, \ldots, \varepsilon_{iT})'$ lze shrnout jako $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}; \mathbf{R}_i)$, kde $\mathbf{R}_i$ značí obecnou kovarianční čtvercovou matici náhodných chyb (teoretických reziduí) o rozměrech $T \times T$ pro i-tý objekt, která obsahuje rozptyly jednotlivých náhodných chyb ε_{it} na diagonále, a mimo diagonálu se nacházejí kovariance mezi jednotlivými náhodnými chybami $\varepsilon_{it},\varepsilon_{it'}$, kde $t \neq t'$. Korelace mezi vektory hodnot náhodných efektů $\mathbf{b}_i$ a náhodných chyb $\boldsymbol{\varepsilon}_i$ je nulová, zkráceně $\mathrm{cor}(\mathbf{b}_i, \boldsymbol{\varepsilon}_i) = \mathbf{0}$. Dále je v KLM předpokládána nezávislost

mezi různými objekty neboli $\mathrm{cor}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}) = \mathbf{0}$, kde $i \neq i'$. Zavedením náhodných efektů modelujeme varianční strukturu vektoru hodnot vysvětlované proměnné $\mathbf{Y}_i$. Kovarianční matici pozorování i-tého objektu (tedy kovarianční matici vektoru $\mathbf{Y}_i$) značenou jako $\mathbf{V}_i$ můžeme vyjádřit $\mathbf{V}_i = \mathrm{Var}(\mathbf{Y}_i) = \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}'_i + \mathbf{R}_i$.

V porovnání s analýzou kompletních dat (*complete-case analysis*), pro kterou lze využít FEM, si MEM poradí s chybějícími hodnotami měření a nevyřadí statistické jednotky (dále objekty) s několika náhodně chybějícími pozorováními, viz Little a Rubin (2002). Další výhodou je schopnost vypořádat se s časovou nepravidelností opakovaného měření. V takovém případě hovoříme o panelu jako o nevybalancovaných datech (*unbalanced data*). Naproti tomu longitudinální data se nazývají v čase vybalancovaná (*balanced data*), pokud má každý objekt stejný počet opakovaných měření v daném časovém období. Dále lze do MEM zahrnout vysvětlující proměnné s proměnlivými hodnotami v čase (*time-varying covariates*).

V neposlední řadě lze kovarianční matici náhodných chyb $\mathbf{R}_i$ parametrizovat. Volbou kovarianční struktury náhodných chyb ε_{it} model umožňuje modelovat prostorovou a časovou korelaci reziduí. Každá kovarianční struktura je vhodná pro různé typy panelů. Jennrich a Schluchter (1986) popisují různé typy kovariančních struktur a volbu správné z nich v analýze panelových dat pro nevybalancované návrhy nebo data s nekompletními záznamy. Nejnovější vývoj statistického softwaru umožňuje modelovat kovarianční strukturu pro komplikovaný *crossed design* obsahující křížové náhodné efekty (*crossed random effect*). Jedná se o náhodný efekt, jehož jedna kategorie se může vyskytovat současně ve spojení s více než jednou kategorií jiného náhodného efektu. Příkladem křížového náhodného efektu může být kombinace prostorových a časových náhodných efektů. Jde tedy o návrh, ve kterém se vyskytuje náhodná variabilita mezi časovými bloky (náhodný efekt času) a náhodná variabilita mezi prostorovými bloky (individuální náhodný efekt). A zároveň je zřejmý konzistentní efekt času na všechny objekty (prostorové bloky) a naopak individuální variabilita působí na časové bloky.

Pro longitudinální data lze rozlišovat dva základní typy LMM (vzorec 2.11): model s náhodným absolutním členem a model s náhodným sklonem a kombinovat tyto dva typy do jednoho modelu. Pro zdůraznění sledování vývoje vysvětlovaných proměnných v čase se tyto LMM nazývají růstové modely.

Model s náhodným absolutním členem (*Random Intercept Model*, **RIM**) vysvětluje (a tudíž předpokládá možnost výskytu) variabilitu výchozích hodnot (*Baseline Values*) y_{it} naměřených v čase $t = 1$ mezi objekty i, i'. Náhodný efekt v RIM je absolutní člen, tedy průsečík s osou

y (výchozí hodnoty). Namísto klasického odhadu korelace známého z KLM využívá RIM rozdíly mezi individuálními průsečíky (pro jednotlivé objekty i). Rozkládá celkovou variabilitu náhodné složky (individuálních průsečíků) σ_v^2, a tedy i variabilitu vysvětlované proměnné Y na variabilitu náhodného efektu (absolutního členu) σ_α^2, která kvantifikuje rozptyl individuálních absolutních členů kolem průměrné hodnoty absolutního členu, a na reziduální složku σ_ε^2 vyjadřující normálně rozdělené nevysvětlené, náhodné kolísání hodnot v čase a mezi objekty. Z rozkladu variability (2.4) lze dopočítat specifický výstup pro RIM, kterým je odhad vnitroskupinového korelačního koeficientu (*Intraclass Correlation*, **ICC**), který měří korelaci mezi pozorováními na stejném objektu jako

$$\mathrm{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}. \tag{2.13}$$

ICC je míra podobnosti kvantifikující sílu lineární závislosti mezi měřeními uvnitř objektu. Oproti Pearsonovu korelačnímu koeficientu nabývá pouze kladných hodnot v intervalu od $\langle 0; 1 \rangle$. Hodnota ICC blízká 1 říká, že jsou si hodnoty stejného objektu podobnější než hodnoty různých objektů, viz Pekár a Brabec (2012). U RIM je tedy nutno vnímat rezidua jinak, než jsme zvyklí např. z metod ANOVA či KLM.

Odhady pevných efektů $\boldsymbol{\beta}'$ kvantifikují efekt vysvětlujících proměnných na hodnotu vysvětlované proměnné. RIM je model (2.11) bez členu λ_t, tj.

$$y_{it} = \mu + \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + \varepsilon_{it}. \tag{2.14}$$

Model s náhodným sklonem (*Random Slope Model;* **RSM**) má ještě další cíl. Odhaduje společný lineární trend pro všechny objekty a náhodné průsečíky a směrnice pro jednotlivé objekty. RSM zohledňuje odlišný individuální sklon (trajektorii) vysvětlované proměnné v čase nebo pro každou vysvětlující proměnnou. V tomto případě je náhodným efektem individuální sklon trajektorie. V modelu lze odlišit, zda má na individuální sklon a jeho odchylku od populačního sklonu vliv buď časová proměnná t, anebo vysvětlující proměnná (proměnné) $\mathbf{X}_i$. Náhodný efekt (t) je symbolizován pomocí λ_t v modelu (2.11). RSM je model (2.11) bez členu α_i, tj.

$$y_{it} = \mu + \boldsymbol{\beta}'\mathbf{x}_{it} + \lambda_t + \varepsilon_{it}. \tag{2.15}$$

Model umožňuje modelovat různé tvary individuálních a populačních trajektorií pomocí funkce $g(\lambda_t)$ pro λ_t. Lze využít například kvadratickou, exponenciální nebo logaritmickou funkci. RSM je vhodné použít v případě, že výzkumníka zajímá efekt vysvětlující proměnné na

změnu sklonu v čase. Avšak v mnoha případech je RIM dostačující model zohledňující mezi-objektové rozdíly.

Hierarchický lineární model (HLM)

Lindley a Smith představují hierarchický lineární model (HLM) v roce 1972. MEM je rovněž dle Raudenbushe a Bryka (2002) nazýván hierarchický lineární model. Tento název vystihuje hierarchické uspořádání dat. HLM je LMM s přesně vymezenou víceúrovňovou (dvě nebo více úrovní) strukturou. Ve dvouúrovňovém modelu pro longitudinální data jsou jednotlivé objekty hierarchicky vyšší jednotky (úroveň 2) a na řádově nižší úrovni 1 se nachází jednotlivá pozorování vyšších jednotek, tedy jednotlivá měření objektů.

Víceúrovňové (strukturované) vnímání dat umožňuje v LMM modelu zohlednit lišící se vztah (efekt) mezi vysvětlující a vysvětlovanou proměnnou na jednotlivých úrovních. Howard (2015) ilustruje odlišný vztah mezi proměnnými na různých úrovních na příkladu, ve kterém X značí příjem jednotlivce (objektu) a sociální postoj je vysvětlovaná proměnná Y. Například vysoký příjem (x_{it}) v čase t může ovlivňovat odlišně (mít jiný efekt na) sociální postoj i-tého jednotlivce (y_{it}), pro kterého je tento příjem nadstandardně vysoký, než obecný postoj jednotlivců, kteří mají nadstandardně vysoké příjmy. V případě očekávání odlišných vlivů vysvětlující proměnné na vysvětlovanou se rozlišují (a odhaduje) vnitřní a vnější efekt vysvětlující proměnné (*within- and between- effects*). Vnitřní efekt je spojen s úrovní 1 a tedy s otázkou: „Jaký vliv má nadstandardně (ve srovnání s průměrným příjmem i-tého jednotlivce) vysoký příjem na sociální postoj i-tého jednotlivce?“ Vnější efekt vyjadřuje vztah mezi proměnnými na druhé úrovní, tj. odpovídá na to, jaký je vztah mezi příjmovou skupinou (průměrnou mzdou i-tého jednotlivce) a sociálním postojem. Vnitřní a vnější efekty jsou zohledněny ve dvou modifikacích MEM: Within-Between Random Effects model a Mundlakův model (1978).

Více o HLM lze nalézt v práci autorů Raudenbushe a Bryka (2002) a Bella a kol. (2016).

Odhady parametrů LMM

Odhady regresních parametrů LMM jsou založeny na metodě maximální věrohodnosti (Maximum likelihood, **ML**) a omezené metodě maximální věrohodnosti (Restricted Maximum Likelihood, **REML**).

Hartley a Rao (1967) ukazují, že jednoznačné odhady komponent variability, viz vzorec 2.4, mohou být získány metodou maximální věrohodnosti (ML). Používají rovnice, které vycházejí z maticového zápisu smíšených modelů. Později však bylo zjištěno Searlem a kol. (1992), že tyto odhady komponent jsou zkreslené (podhodnocené), jelikož metoda předpokládá,

že pevné efekty jsou známy a nejsou odhadovány z dat. Maximálně věrohodný odhad je získán jako hodnota parametru (regresního parametru), ve které logaritmus věrohodnostní funkce nabývá maxima. Na rozdíl od situace v KLM, kde je možné nalézt optimum explicitně a relativně snadno, není pro LMM ani marginální model (viz dále, kapitola 2.3) explicitní řešení k dispozici. Z tohoto důvodu se rovněž využívají numerické optimalizační metody iterativního charakteru například REML. REML vychází z ML a opravuje ji formou redukce stupňů volnosti, která je nutná z důvodu, že pevné efekty jsou odhadovány ze stejných dat jako parametry související s rozptyly a korelacemi dle Pekára a Brabce (2012). Metoda omezené maximální věrohodnosti (REML) je uvedena Pattersonem & Thompsonem (1971) pro odhad komponent variability (bez předpokladu známých pevných efektů) v GLM pro nevybalancovaná data. Metody založené na věrohodnosti se rozvíjejí pomalu z důvodu výpočetní náročnosti. Poprvé aplikuje REML na panelová data Verbyla (1990). Až Searle (1992) popisuje intervaly spolehlivosti pro odhad komponent variability v LMM s jedním náhodným faktorem.

Z výše popsaných důvodů by se metodou ML měly odhadovat pouze parametry FEM, zatímco náhodné efekty v REM a MEM je vhodné odhadovat iterativně metodou REML. Odhad náhodných efektů pomocí algoritmu REML umožňuje vypořádat se s vnitroobjektovou korelací, viz Rasbash (2010).

V případě nevybalancovaného návrhu (za předpokladu, že hodnoty chybějí zcela náhodně) je nutné využít iterativní přístup k maximálně věrohodnému odhadu pevných efektů a komponent variability σ_v^2, např. iterativní obecnou metodu vážených nejmenších čtverců (*Iteratively Reweighted Generalized Least Squares*, **IRLS**), s kterou přichází Goldstein (1986). IRLS odhaduje regresní parametry metodou ML za předpokladu známé kovarianční matice $\mathbf{V}_i$ vysvětlované proměnné, z čehož vyplývá předpoklad známé kovarianční matice náhodných chyb $\mathbf{R}_i$.

V IRLS se přiřazují různé váhy w (diagonálním nebo nediagonálním) prvkům obecné kovarianční matice náhodných chyb $\mathbf{R}_i$ pro i-tý objekt oproti OLS (využívaném pro odhady např. v KLM a FEM kap. 2.2.1), kde $\boldsymbol{\varepsilon}_i \sim \text{iid N}(\mathbf{0}; \sigma^2\mathbf{I})$. Kovarianční matici vysvětlované proměnné $\mathbf{V}$ a náhodných chyb $\mathbf{R}$ pro všech N objektů lze maticově zapsat jako

$$\mathbf{V} = \mathbf{R} = \sigma^2 \mathbf{W}^{-1}, \tag{2.16}$$

kde je směrodatná odchylka σ, která je v této problematice nazývána škálový parametr (parametr měřítka). Matice $\mathbf{R}$ je blokově diagonální a obsahuje N matic $\mathbf{R}_i$ na diagonále. Mimo diagonálu jsou nulové hodnoty, které odpovídají nezávislosti mezi objekty. Prvky w_i inverzní diagonální matice apriorních vah $\mathbf{W}^{-1}$ mohou být stanoveny např. na základě apriorní informace nebo mohou být funkcí jedné nebo více vysvětlujících proměnných, například $w_i = (\gamma_0 + \gamma_1 x_{1i})$.

Iterativní postup odhadu regresních parametrů metodou IRLS lze pro tento ilustrativní příklad shrnout:

1. Nejprve se volí počáteční hodnoty w_i pro každý objekt, například $w_i = 1$.
2. Střední hodnoty pevných efektů jsou odhadnuty s využitím počátečních hodnot w_i.
3. Použijí se Pearsonova rezidua (spočítané na základě odhadu modelu z druhého kroku, zkráceně *res*) k odhadu parametrů γ_0, γ_1 jako $res^2 = \gamma_0 + \gamma_1 X_1$).
4. Přepočítají se váhy w_i. Přepočítané váhy jsou použity pro odhad náhodných efektů a následně slouží k úpravě odhadů středních hodnot a středních chyb pevných efektů.

Uvedený postup lze opakovat do splnění vybraného kritéria (například doba výpočtu). Ve všech krocích 1 – 4 IRLS jsou parametry odhadovány metodou ML nebo REML. Více o tomto postupu lze nalézt v práci autorů Rupperta a Carrolla (1988).

2.3 Marginální model (model populačních průměrů, MMM)

Uvést a porovnat běžné přístupy pro regresní modelování normálně rozdělené vysvětlované proměnné jsem si stanovil jako jeden z cílů kapitoly 2. Z tohoto důvodu stručně popisuji ještě marginální model (*Marginal Means Model*, **MMM**) v závěrečném oddíle 2.3 této kapitoly, jelikož patří do základní analytické výbavy pro longitudinální data. Marginální model lze vnímat jako přechod z metody ANOVA určené pro opakovaná měření (FEM) do MEM. Softwarový výstup marginálního modelu bývá intuitivní a podobný výstupu z metody ANOVA. Marginální model lze zapsat ve tvaru

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{2.17}$$

ve kterém platí stejné předpoklady jako v MEM pro vektor náhodných chyb $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}; \mathbf{R}_i)$ a $\mathrm{cor}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}) = \mathbf{0}$, kde $i \neq i'$. Oproti MEM (vzorec 2.13) marginální model neobsahuje náhodné efekty (člen $\mathbf{Z}_i\mathbf{b}_i$), tudíž nedokáže vysvětlit variabilitu náhodné složky pomocí různých (varianěních) komponent. Stejně jako v FEM marginální model obsahuje pouze pevné efekty, avšak ve srovnání s tímto modelem zohledňuje unikátní vnitroobjektovou korelaci. Parametry marginálního modelu jsou odhadovány pomocí IRLS, viz oddíl 2.2. MMM má podobné výhody pro modelování dat jako MEM a lze ho použít jako alternativní přístup k MEM. V mnoha případech marginální model lépe modeluje data než MEM. Avšak v marginálním modelu se nevyskytuje náhodný efekt, a nelze jej tudíž modelovat. Další rozdíl mezi porovnávanými přístupy je v tom, jak pracují s korelací mezi objekty. Marginální model odhaduje korelace přímo. MEM

korelace přímo nepočítá a zaměřuje se na to, jak moc se mezi sebou liší výchozí hodnoty y_{it} naměřené mezi objekty *i*, *i'*, viz podkapitola 2.2.

Z důvodu nižší výpočetní náročnosti (neodhadují se náhodné efekty) je dostačující marginální model a bývá preferován před MEM v situaci, kdy není výzkumným záměrem odhadnout komponenty náhodné složky (náhodná variabilita mezi objekty a vnitroobjektová korelace) a cílem je odhadnout správně populační regresní parametry se zohledněním vnitroobjektové korelace a heteroskedasticity.

Více o marginálním modelu například v práci Verbekeho a Molenberghse (2000) a Pekára a Brabce (2012).

V oddílech 2.1 a 2.2 jsem popsal dva odlišné přístupy (dvoukrokový a simultánní). V LMM se odhadují náhodné efekty namísto odhadování regresních parametrů pro každý objekt (viz dvoukroková analýza). Náhodný efekt vyjadřuje variabilitu kolísání individuálních parametrů kolem skupinového (populačního) parametru. Například pro model s náhodným sklonem s předpokladem lineárního vývoje v čase se odhadují dva pevné a dva náhodné efekty. Pevné efekty (β_0, β_1) vyjadřují skupinový průsečík a směrnici. Náhodný efekt pro průsečík vyjadřuje variabilitu individuálních průsečíků kolem skupinového průsečíku β_0. Průměrná variabilita individuálních průsečíků okolo skupinového je nulová. Náhodný efekt pro směrnici určuje míru kolísání individuálních směrnic okolo skupinové směrnice β_1.

Oba směry mají své výhody a nevýhody. Více výhod lze zřejmě spatřovat ve smíšených lineárních modelech (LMM). V případě malého počtu měření (v řádu jednotek) pro pozorovaný objekt nelze očekávat kvalitní odhad regresních parametrů individuálního KLM. Individuální odhad regresních parametrů pro každý objekt zvlášť je prvním krokem v diskutovaném dvoukrokovém přístupu. Velkou výhodou LMM oproti dvoukrokovému přístupu je, že díky simultánní analýze informací o všech objektech získáme přesnější odhad individuálních parametrů vyjádřených pomocí náhodných efektů. Sdružený odhad náhodných efektů je nejlepší lineární nezkreslený odhad (NLNO), viz Pinheiro a Bates (2000).

Nevýhodu LMM vnímám zejména v nejednoznačnosti odhadu parametrů modelu. Parametry jsou odhadovány iterativně a odhad nemusí konvergovat ke svému maximu věrohodnostní funkce. Intervalové odhady (založené na teorii normálního rozdělení) náhodných efektů je obecně (ve všech softwarech) problematické odhadnout. Nejsou součástí standardních výstupů výpočetních programů, popsaných viz kapitola 4. V současné době začali tvůrci nového softwarovu využívat simulaci Markovova řetězce Monte Carlo ke konstrukci intervalových odhadů náhodných efektů. Avšak tento bayesovský přístup intervalového odhadu není obecně (frekventisty) přijímán. Další nevýhodou je výpočetní čas, který např. v případě analýzy datových souborů velkého rozsahu (od tisíců záznamů), kdy navíc předpokládám nestrukturovanou (obecnou) kovarianční matici náhodných chyb (*unstructured covariance matrix*), může výpočet trvat hodiny nebo dny, nebo vůbec nebude z důvodu výpočetní náročnosti spočítána (ani v současné době s výkonnými počítači). Více o porovnání dvoukrokové analýzy a smíšených lineárních modelů Rosenberg (1973).

3. Nové přístupy v oblasti regresní analýzy panelových dat

S rozvojem výpočetní techniky se začaly objevovat a v praxi využívat modifikace metod popsaných v kapitole 2 a vznikají nové přístupy k modelování složitějších úloh, kdy je porušen například předpoklad normality vysvětlované proměnné.

V současné době se rozšiřují LMM na neparametrické a semiparametrické panelové regresní modely. Prvopočátky neparametrických a semiparametrických modelů pro panelová data lze přiřadit autorům Ullahovi a Royovi (1988). Rozšířením modelu náhodných efektů na neparametrický regresní model se zabývá Wang (2003) a na semiparametrický částečně lineární model Wang a kol. (2005). Neparametrickým a semiparametrickým částečně lineárním modelům pevných efektů se věnují Li a Racine (2007). Nelineární modely pro opakovaná měření popisuje Davidian a Giltinan (1995).

Pro popis dat, která nepochází z normálního rozdělení (náhodné efekty, vysvětlovaná proměnná), ale z jiného rozdělení exponenciálního typu se využívají dva typy modelů: zobecněné marginální modely (oddíl 3.1) a zobecněné lineární modely se smíšenými efekty (oddíl 3.2). Alternativním přístupem pro obecně rozdělené proměnné jsou bayesovské modely (oddíl 3.3).

O'Hara a Kotze (2010) preferují aplikovat zobecněné marginální modely nebo zobecněné lineární modely se smíšenými efekty v případě, že vysvětlovaná proměnná se řídí rozdělením z exponenciální rodiny, např. je typu: binární (binomická, tj. náhodný jev nastal/nenastal), diskrétní vyjadřující absolutní četnost výskytu (*count*), nebo relativní četnost (poměr) namísto postupu, kdy je tento typ proměnné transformován (logaritmická, odmocninová transformace atd.) na normálně rozdělenou proměnnou a na transformovanou vysvětlovanou proměnnou je použit MEM. Tito autoři pomocí simulační studie ukazují, že poissonovský nebo negativně binomický model pro diskrétní proměnnou vyjadřující absolutní četnost odhaduje regresní parametry s výrazně menším zkreslením než aplikovat MEM na transformovanou vysvětlovanou proměnnou. Navíc se zobecněné modely vypořádají i s nulovými hodnotami na rozdíl od logaritmické transformace.

Přestože jsou tři výše jmenované metody (popsané v oddílech 3.1, 3.2 a 3.3) dobře známé již delší dobu (od 90. let minulého století), začaly se pro jejich výpočetní náročnost implementovat do statistických programových systémů až v posledních letech. Použitelnost těchto metod se v současné době ověřuje na reálných studiích. V tuto chvíli nepanuje mezi uživateli a ani mezi tvůrci programů konsenzus, která z metod je lepší pro konkrétní typ studie a často je do-

poručováno vyzkoušet a porovnat odhady získané z více metod. Opakovaně se vyskytující specifické úlohy vedou k dynamickému vývoji úzce profilovaných modifikací a kombinací těchto tří metod. Například mezi nejnovější modifikace zobecněných lineárních modelů se smíšenými efekty, jejichž základní charakteristiky ve své práci zmiňuji, patří například: zobecněný lineární model se smíšenými efekty s vlivnými nulami, kumulativní spojovací model a zobecněný aditivní model se smíšenými efekty.

Používáním těchto metod se postupně odstraňují nedostatky v implementaci těchto metod v programech. Problematická je zejména oblast odhadu regresních parametrů modelu pro velké anebo specifické datové struktury. Tento problém by mohl být vyřešen v řadě případů pomocí shlukové analýzy, která by v exploratorní fázi (před regresní analýzou) pomohla identifikovat objekty s podobnou kovarianční strukturou vysvětlované proměnné. Náležitost do shluku může vstoupit do regresní analýzy například formou kategoriální pomocné vysvětlující proměnné. O shlukové analýze mikropanelových dat a jejím využití nejen v regresní analýze pojednávají kapitoly 5 - 8.

3.1 Zobecněný lineární marginální model (GMMM)

Zobecněný marginální model (*Generalized Marginal Means Model*, **GMMM**) lze použít k odhadu modelu populačních průměrů s ohledem na vnitroobjektovou korelaci bez plné specifikace spojitého rozdělení vysvětlované proměnné. Stejně jako v marginálním modelu (MMM) se odhadují populační průměry a nevyskytuje se v něm náhodný efekt. Oproti marginálnímu modelu se může řídit vysvětlovaná proměnná některým rozdělením z větší škály rozdělení pocházející z exponenciální rodiny (například Poissonovo, binomické, negativně binomické, normální, gamma atd.).

Odhady parametrů a jejich střední chyby jsou provedeny metodou **GEE** (Generalized Estimating Equations), která je rozšířením zobecněných lineárních regresních modelů (**ZLM**) na korelovaná data (výskyt vnitroobjektové korelace). Stejně jako ZLM využívá GEE rozdělení exponenciální třídy a spojovací (transformační, *link*) funkce, avšak je možné aplikovat různé korelační struktury. Bližší popis metody lze nalézt ve článku Asara a Ilka (2013).

GEE je iterační metoda odhadu regresních parametrů, která rozšiřuje ZLM o předpoklad specifické korelační struktury náhodných chyb. GEE je zobecněním metody IRLS, viz stručný popis iteračního postupu v oddílu 2.2, pro vysvětlovanou proměnnou, která se řídí některým rozdělením z exponenciální rodiny, více k GEE autoři Liang a Zeger (1986), Prentice a Zhao (1991), Zuur a kol. (2009), Fitzmaurice a kol. (2011).

3.2 Zobecněný lineární model se smíšenými efekty (GLMM)

Zobecněný lineární model se smíšenými efekty (*Generalized Linear Mixed Effects Models*, **GLMM**) kombinuje vlastnosti dvou statistických metod: LMM a zobecněného lineárního modelu (ZLM). ZLM s využitím spojovací funkce umožňuje modelovat jinak než normálně rozdělenou vysvětlovanou proměnnou, která se řídí rozdělením náležejícím do exponenciální rodiny. Současně GLMM umožňuje zahrnovat více zdrojů variability (náhodné efekty a náhodné chyby) díky kombinaci ZLM s LMM. Díky odhadu náhodných efektů lze extrapolovat hodnoty pro nové objekty nezařazené do výběru. Náhodný efekt lze do GLMM zahrnout díky LMM, což je zásadní praktický rozdíl oproti GMMM (viz oddíl 3.1).

V GLMM se odhadují 2 skupiny efektů (regresních parametrů), jmenovitě: pevné a náhodné efekty, stejně jako v LMM.

Odhady efektů GLMM

K odhadu efektů (regresních parametrů) GLMM lze využít metodu ML. Její používání je omezené na malý počet uvažovaných náhodných efektů z důvodu výpočetní náročnosti, kdy se hledá maximum logaritmu věrohodnostní funkci pomocí derivace přes všechny možné hodnoty náhodných efektů. Z vlastní zkušenosti vím, že LM nefunguje (neodhadne efekty) již pro relativně malý soubor (v řádu stovek objektů) v případě zahrnutí tří a více náhodných efektů do modelu. Častěji se efekty odhadují pomocí metody REML. Statistici navrhli čtyři metody aproximace věrohodnosti (REML) s cílem umožnit odhadovat složitější GLMM. Zásadní rozdíl mezi ML a REML je v tom, že ML pro odhad směrodatné odchylky náhodného efektu předpokládá precizní, spolehlivý odhad pevných efektů. ML podhodnocuje směrodatnou odchylku náhodného efektu. ML je užitečná pro porovnání modelů s různými pevnými efekty. Naproti tomu REML předpokládá a průměruje do určité míry nejistotu v odhadech pevných efektů, viz Pinheiro a Bates (2000) a Littell a kol. (2006). Ve statistických programech bývají implementovány ML, níže popsané tři REML techniky odhadu a metoda odhadu založená na Bayesovském teorému (oddíl 3.3), viz Bolker a kol. (2008):

- Pseudo- a penalizovaná kvazi-věrohodnost (*pseudo- and penalized quasilikelihood*, PQL) je nejjednodušší aproximace ML. Bohužel PQL zkresluje odhady efektů, pokud jsou směrodatné odchylky náhodných efektů velké. Toto platí zejména pro binární data. Dále PQL zkresluje odhady efektů pro vysvětlovanou proměnnou pocházející z Poissonova rozdělení s malým průměrným počtem pozorování ($\lambda < 5$) pro danou kategorii diskrétního, pevného faktoru. Špatné výsledky dosahuje i pro binární data s malým (< 5) očekávaným počtem úspěchů a zároveň malým počtem neúspěchů pro danou kombinaci

faktorů, viz Breslow (2003). Další nevýhodnou PQL je výpočet kvazi-věrohodnosti místo skutečné věrohodnosti. To znemožňuje provádět úsudky (testování hypotéz o parametrech, použití kritéria AIC pro výběr modelu) založené na věrohodnosti. Více v Schallově (1991) a Breslowově studii (2003).

- Laplaceova aproximace je přesnější než PQL. Aproximuje skutečnou věrohodnost a umožňuje provádět úsudky založené na ML. Ve srovnání s PQL je výpočet pomalejší, více viz Raudenbush a kol. (2000).
- Gaussova-Hermiteova kvadratura (*Gauss-Hermite quadrature*, GHQ) je dokonce přesnější než Laplaceova aproximace. GHQ opakuje iterace vícekrát na rozdíl od Laplaceovy aproximace, která odhaduje efekty na základě jedné iterace. Bohužel rychlost výpočtu GHQ se s rostoucím počtem náhodných efektů zpomaluje a není doporučováno zařadit do modelu více než jeden náhodný efekt. Při zařazení třech a více náhodných efektů do modelu výpočet nekonverguje k maximu věrohodnostní funkce, více viz Pinheiro a Chao (2006).

Pokud je model přeparametrizovaný, techniky odhadu nekonvergují ke globálnímu maximu logaritmu věrohodnostní funkce. V přípravné fázi experimentu je na to třeba myslet a naměřit dostatečně velký výběrový vzorek (> 5–6 úrovní, jmenovitě počet objektů nebo časových okamžiků, pro každý náhodný efekt a > 10–20 měření pro každou úroveň kategoriálního pevného efektu). Zájemci o GLMM naleznou aplikace a matematický popis modelu v pracích Diggla a kol. (2002); Molenberghse a Verbekeho (2005); McCullocha, (2008); Fitzmaurice a kol. (2011).

GLMM pro data s vysokou četností vlivných nulových hodnot (ZIGLMM)

Pro analýzu dat s pevnými a smíšenými efekty a vysvětlovanou proměnnou pocházející z Poissonova rozdělení s častým výskytem nulových hodnot byl odvozen specifický zobecněný lineární model se smíšenými efekty s vlivnými nulami v anglické literatuře nazývaný *zero-inflated generalized linear mixed-effects model,* zkráceně **ZIGLMM**. Tento model v sobě zahrnuje prvky LMM, GLMM a Poissonův regresní model s vlivnými nulami (*zero-inflated Poisson regression,* ZIP).

Model ZIP byl navržen pro modelování vysvětlované proměnné s vlivným (častým) výskytem nulových hodnot. Vlivný výskyt nulových hodnot znamená, že se nuly vyskytují častěji, než dokáže být vysvětleno Poissonovým rozdělením. Vysvětlovaná proměnná se řídí směsí rozdělení. ZIP předpokládá s pravděpodobností p výskyt pouze nulových hodnot a s pravděpodobností $1 - p$ Poissonovo rozdělení s parametrem λ. Výskyt nul se řídí binomickým rozdělením

a počet výskytů ostatních hodnot větších než nula je generován Poissonovským procesem. Odhady parametrů modelu jsou získány metodou ML. Metodu navrhla Lambert (1992).

Myšlenka ZIP byla modifikována i pro jiná rozdělení: negativně-binomický model s vlivnými nulami (*zero-inflated negative binomial model*), který navrhl Greene (1994) a binomický model s vlivnými nulami (*zero-inflated binomial model*) navržený Hallem (2000), který je určen pro zprava omezené hodnoty Y. Rozšíření ZIP na panelová data zprava omezené (binomický model), či neomezené (poissonovský model) a zároveň obsahující smíšené efekty (ZIGLMM) odvodil Hall (2000).

ZIGLMM s parametry p a λ závislými na vysvětlujících proměnných popisují Bolker a kol. (2012a). V tomto modelu je využita logaritmická transformační funkce pro Poissonovo rozdělení a logitová transformační funkce pro binomické rozdělení. Hodnotu y_{it} lze vyjádřit

$$y_{it} \sim \mathrm{ZIPois}(\lambda = \exp(\eta_{it}), p = \mathrm{logistic}(z_i)), \tag{3.1}$$

kde ZIPois značí Poissonovo rozdělení s nulovými hodnotami a logistic(z_i) = $(1 + \exp(-z_i))^{-1}$. Pravděpodobnost, že $y_{it} = 0$ je rovna $p + (1 - p)\exp(-\lambda)$ a pro $y_{it} > 0$ je pravděpodobnost nastání $(1 - p)\mathrm{Pois}(X, \lambda)$. Parametr z_i vyjadřující vliv nul je nastaven jako konstanta. Lineární prediktor η_{it} má identické vyjádření jako y_{it} ve vzorci (2.12).

Kumulativní spojovací model se smíšenými efekty (CLMM)

Ordinální vysvětlovanou proměnnou je zvykem odhadovat pomocí LMM po přeškálování všech hodnot y na z-skóre. Škálováním je získáno z-scóre z_{it} = (y_{it} – průměr(y_i))/směrodatná odchylka(y_i), které se odhaduje metodou LMM. Nevýhodou tohoto přístupu je redukce variability a změna původních dat.

Relativně novou možností modelování ordinální proměnné je specifická forma GLMM nazvaná kumulativní spojovací model se smíšenými efekty (*Cummulative Link Mixed Model*, CLMM). V CLMM se využívají ordinální probitové modely se smíšenými efekty nebo ordinální logitové modely se smíšenými efekty popsanými Hedekerem a Gibbonsem (1994). Obecně platí, že probitový model je vhodnější než logitový v případě velkého výskytu extrémních pozorování. Důvodem je rozdělení Y s tlustšími konci (platykurtické rozdělení) ve srovnání s logitovým rozdělením. CLMM je vhodný v případě, že vzdálenosti mezi kategoriemi nejsou stejné, např. vzdálenost mezi pořadovými hodnotami 1 a 3 je jiná než vzdálenost mezi pořadovými hodnotami 2 a 4. CLMM ve své monografii popisuje Agresti (2010).

Zobecněný aditivní model se smíšenými efekty (GAMM)

Zobecněné aditivní regresní modely (*Generalized Additive Models,* **GAM**) pro prostorová data (dodržen předpoklad homoskedasticity a vzájemné nezávislosti hodnot *Y*) jsou navrženy k získání nelineárního odhadu vztahu mezi vysvětlovanou proměnnou *Y*, která se řídí některým z rozdělení exponenciálního typu a vysvětlujícími proměnnými **X**. Lineární prediktor zobecněného aditivního modelu obsahuje dvě složky. První složka $\mathbf{X}_i\boldsymbol{\beta}$ lineárního prediktoru je identická se ZLM. Druhá část, tzv. vyhlazovací složka (*smooth term*), lineárního prediktoru obsahuje součet *R* regresorů $f_r(x_r)$[4]. Pro *i*-tý objekt lze zobecněný aditivní model (GAM) zapsat

$$g\left(E\left(Y_i\right)\right)=\mathbf{X}_i\boldsymbol{\beta}+f_1\left(x_{1i}\right)+f_2\left(x_{2i}\right)+f_3\left(x_{1i},x_{2i}\right)+...+f_R\left(x_{Ri}\right), \tag{3.2}$$

kde *g* značí spojovací funkci, $E(Y_i)$ je odhad střední hodnoty vysvětlované proměnné *Y*, která se řídí rozdělením z exponenciální rodiny, pro *i*-tý objekt. První část $\mathbf{X}_i\boldsymbol{\beta}$ lineárního prediktoru se skládá ze součinu vektoru pevných efektů $\boldsymbol{\beta}$ a *i*-tého řádku matice plánu $\mathbf{X}_i$. Oproti ZLM však model obsahuje *R* regresorů $f_r(x_r)$, které nemusí nabývat parametrické formy (např. polynom), ale může být specifikován neparametricky, nebo semiparametricky pomocí vyhlazovací funkce (*smooth function*). Aditivní model je specifickou formou zobecněného aditivního modelu, ve kterém *g* je identita a *Y* se řídí normálním rozdělením. Autory aditivních regresních modelů jsou Hastie a Tibshirani (1990).

Lin a Zhang (1999) rozšířili model 3.2 o náhodné efekty, stejný člen $\mathbf{Z}_i\mathbf{b}_i$ jako v LMM ve vzorci 2.12, a navrhli zobecněné aditivní modely se smíšenými efekty (*Generalized additive mixed models,* **GAMM**) pro korelovaná data. GAMM lze pro *i*-týobjekt zapsat podle Wooda (2006b) následovně:

$$g\left(E\left(Y_i|\mathbf{b}_i\right)\right)=\mathbf{X}_i\boldsymbol{\beta}+w_{1i}\,f_1\left(x_{1i}\right)+w_{2i}\,f_2\left(x_{2i}\right)+w_{2i}\,f_3\left(x_{1i},x_{2i}\right)+...+\mathbf{Z}_i\mathbf{b}_i, \tag{3.3}$$

kde w_{ri} jsou váhy jednotlivých regresorů zadané jako spojité hodnoty, dummy proměnné nebo často hodnoty 1; K odhadu neparametrických vyhlazovacích funkcí obsažených ve vyhlazovací složce $\sum_{j=1}^{J} w_{ji}f_j\left(x_{ji}\right)$ slouží tzv. vyhlazovací spline. Výzkumník má na výběr několik typů vyhlazovacích spline například penalizovaný regresní spline (*penalized regression splines*), regresní křivka s fixními stupni volnosti, *thin plate regression spline*, Duchonův spline, *P*- spline, sférický spline atd. Kromě typu splinu je třeba správně zvolit hodnotu vyhlazovacího parametru (*smoothing parameter*), který určuje stupeň vyhlazení, více viz Wood (2006a).

[4] Regresor je známá funkce vysvětlujících proměnných, která neobsahuje žádné další parametry.

3.3 Bayesovský přístup

Bayesovský přístup je alternativním a zcela odlišným přístupem pro modelování korelovaných dat. Je řešením v situaci, kdy nelze aplikovat LMM, MMM, GMMM ani GLMM, jelikož vysvětlovaná proměnná nepochází z exponenciální rodiny rozdělení. Používá se rovněž jako alternativní metoda odhadu regresních parametrů GLMM k metodám odhadu ML, PQL, Laplaceova aproximace a GHQ.

Jak název napovídá, tento přístup je založený na bayesovském modelování. Oproti frekventistickému (klasickému) přístupu (ke statistice) při modelování kombinuje data s objektivní apriorní informací o rozdělení hodnot parametru za účelem odvození posteriorních pravděpodobností různých modelů nebo hodnot parametru. Jedná se o pseudo-zobecněný lineární model se smíšenými efekty, jehož regresní parametry jsou odhadovány pomocí Monte Carlo simulace Markovova řetězce (Markov Chain Monte Carlo, **MCMC**). Metody odhadu popsané v předchozí sekci 3.2 hledají maximum logaritmu věrohodnostní funkce napříč celý obor hodnot náhodných efektů. MCMC generuje řetězec náhodných výběrů z pravděpodobnostního rozdělení hodnot pevných a náhodných efektů. K odhadu využívá apriorní informaci o rozdělení parametrů. Například ve funkci *MCMCglmm* v programu R (více v oddílu 4.4) je předpokládáno (naprogramováno) normální rozdělení pro pevné efekty a inverzní Wishartovo rozdělení pro náhodné efekty. V případě neznalosti apriorní informace lze nastavit velmi slabý prior (hodnoty parametrů předpokládaného apriorního rozdělení) bez informační hodnoty. Odhady jsou založeny na popisných statistikách (průměr, modus, kvantily atd.) posteriorního rozdělení. Oproti metodám uvedeným v části 3.2 umožňuje odhadovat model s větším počtem náhodných efektů, pokud je k dispozici velký datový soubor. Více informací je k dispozici například ve studiích Gilkse a kol. (1996) a Bolker a kol. (2008).

V posledních letech statistici sahají stále častěji po tomto nástroji z důvodu rychlého rozvoje výpočetní techniky, která umožňuje provádět složité a náročné výpočty, jež jsou u tohoto typu modelování zapotřebí při odhadování parametrů modelu pomocí simulace. Nevýhodou MCMC je délka výpočtu a technické detaily, které je třeba správně stanovit: výběr vhodného prioru popisuje Berger (2006), výběr efektivního algoritmu pro velké soubory řeší Carlin (2006) a správným stanovením dostačující délky běhu algoritmu, který konverguje ke správným hodnotám, se věnují Brooks a Gelman (1998) a Paap (2002).

Využití bayesovského přístupu přesahuje hranice odhadu parametrů modelu. Aplikuje se například v kombinaci s frekventistickým modelem GLMM:

- Baayen a kol. (2008) využili MCMC k výpočtu intervalových odhadů náhodných efektů odhadnutých frekventistickými metodami, popsanými v části 3.2. Upravený MCMC algoritmus provádí výběry z posteriorního rozdělení náhodných efektů (nepředpokládá informativní prior). Výhodou kombinace dvou statistických přístupů je výpočetní efektivita frekventistického přístupu a vyhnutí se obtíží s odhadem stupňů volnosti pro F-testy,
- Gelman a kol. (2008) navrhují bayesovské řešení, jak se vypořádat s výskytem kompletní nebo částečné (kvazi-kompletní) separace u korelovaných dat s binomickou vysvětlovanou proměnnou. Kompletní separace je jev, který značí přítomnost jedné nebo více vysvětlujících proměnných, které dokonale (přesně) predikují vysvětlovanou proměnnou. To znamená, že alespoň pro jednu kategorii (hodnotu) vysvětlující proměnné X nabývá vysvětlovaná binární proměnná stejnou hodnotu pro všechny objekty. U částečné (kvazi-kompletní) separace dokonalá predikce platí pro podskupinu objektů. Výskyt separace v datech způsobuje vychýlenost odhadů pevných efektů získaných metodou maximální věrohodnosti. Navíc jsou jejich SE extrémně velké hodnoty (v extrémních případech limitně blízké nekonečnu). Extrémní hodnoty odhadů vedou ke kolapsu ML algoritmů odhadu. U nekorelovaných dat frekventisti řeší výskyt separace například pomocí úpravy věrohodnostní funkce. Ditrich (2015) shrnuje tři metody úpravy věrohodnostní funkce: exaktní logistická regrese, skrytá logistická regrese, nebo penalizovaná logistická regrese. U penalizované logistické regrese se odhady pevných efektů odhadují Firthovým algoritmem penalizované ML, viz Firth (1993). Stanovení apriorní informace o rozdělení pevných efektů, a tedy využití bayesovského přístupu, umožňuje řešit problém separace u korelovaných dat, bez nutnosti vyřadit vysvětlující proměnnou z modelu. Hodnoty všech nebinárních vysvětlujících proměnných $\mathbf{X}$ navrhují Gelman a kol. (2008) či Abrahantes a Aerts (2012) transformovat (škálovat) na hodnoty Cauchyho t-rozdělení. Gelman a kol. (2008) doporučují použít t-rozdělení centrované okolo 0 s hodnotou škálového parametru σ (vzorec 2.16) rovnou 2,5 a dále centrovat hodnoty X okolo střední hodnoty odhadu pevného efektu se škálovým parametrem rovným dvojnásobku SE odhadu pevného efektu. Pevný efekt a jeho SE lze odhadnout Firthovou metodou penalizované maximální věrohodnosti s ignorováním náhodných efektů.

4. Analýza longitudinálních dat v systému R

V této kapitole je uveden stručný přehled aktuálních softwarových nástrojů pro využití LMM i nových přístupů k analýze mikropanelových dat, včetně relevantních odkazů na literaturu (učebnice, manuály).

V dnešní době si může uživatel (statistik, výzkumník) vybrat ze široké škály statistického softwaru nabízejícího výpočet regresních modelů, popsaných v kapitolách 2 a 3, například v programech R, SAS, IBM SPSS Statistics, Stata, S+ a HLM. Základní historické milníky pro implementaci smíšených lineárních modelů do statistických programů shrnuje West a kol. (2014):

1982: Bryk a Raudenbush vydávají první počítačový program HLM, viz Bryk a kol. (1988),

1988: Schluchter a Jennrich přicházejí se softwarem BMDP5 určeným pro modelování nevybalancovaných opakovaných měření,

1992: SAS uvádí Proc Mixed jako součást analytického balíčku SAS/STAT,

1995: StataCorp spouští Stata Release 5, který nabízí proceduru xtreg k odhadu modelu s náhodnými efekty spojenými s jedním náhodným faktorem. Release 5 dále nabízí proceduru xtgee pro modelování panelových dat,

1998: Bates a Pinheiro vytvářejí funkci *lme()* v systému R pro odhad smíšených lineárních modelů,

2001: Rabe-Hesketh a kol. napsali pro software Stata příkaz gllamm k odhadu LMM.

SPSS vydává první verzi MIXED procedury v SPSS 11.0,

2005: Stata připravila obecný LMM příkaz xtmixed, který je součástí Stata Release 9.

Bates přichází s *lmer()* funkcí pro systém R.

Porovnání nejužívanějších funkcí (nejen R) pro modelování LMM shrnuje tabulka P1 (v příloze) z hlediska použité metody pro odhad parametrů, možnosti provedení a výběru celkového testu o modelu a individuální testy o parametrech, možnosti a výběru intervalových odhadů, volby struktury náhodných efektů a struktury náhodných chyb.

Největší nabídka různých (včetně nejnovějších) metod (viz kapitoly 2 a 3) je k dispozici ve statistickém systému R.

R (www.r-project.org) je jazyk a prostředí zaměřené na implementaci statistických a grafických metod pro analýzu dat. Jedná se o open-source program, který může být zdarma stažen z výše uvedených stránek a užíván v souladu s GNU General Public License. R se liší oproti jiným statistickým programům v tom, že jsou v něm implementovány některé koncepty objek-

tově-orientovaného (O-O) programování, a to třídy objektů a metody operující nad těmito třídami. V R jsou implementovány dva O-O systémy, S3 a S4. Systém S3 je méně formální implementací O-O oproti S4, který přísně zohledňuje O-O koncept. V obou systémech jsou koncepty O-O implementovány pomocí speciálního typu funkcí nazývaných generické funkce. Generická funkce umožňuje aplikovat vhodnou metodu na objekt dle třídy objektu. O-O programování je vhodné pro statistické modelování, jelikož modely mohou být rozděleny do oddělitelných komponent, jako jsou například data, struktura průměrů, varianční funkce[5] atd. Navíc komponenta definovaná pro jeden typ modelů může být použita jako stavební kámen pro jiný typ modelů, viz Galecki a Burzykowski (2013).

Uživatelům jiného software (než R) lze doporučit aplikaci LMM v:

- SAS: MIXED procedura aplikovaná na data z oblasti vzdělávání, viz Singer (1998), monografie Singera a Willetta (2003) ukazující software na datech z oblasti vzdělávání, sociologie a biostatistiky, monografie Singera a Willetta (2003) s příklady z oblastí biomedicíny, veřejného zdraví, vzdělávání, managementu. Pro seznámení se s PROC MIXED doporučuji práce SAS Institutu (1992); Latoura a kol. (1994); Hoxa (1995); SAS Institutu (1996); Littella a kol. (2006).
- SPSS: pro zájemce, kteří se chtějí naučit v SPSS aplikovat a interpretovat výsledky LMM doporučuji online Seltmanovu učebnici (2014), monografii Singera a Willetta (2003),
- Stata: monografie Singera a Willetta (2003).
- HLM: manuál Bryka a kol. (1988); Bryk a kol. (1996), monografie Singera a Willetta (2003).
- MLwiN: manuál Prossera a kol. (1996), monografie Singera a Willetta (2003) a ukázkový příklad aplikace Leckieho a Bella (2013).

V HLM a MLwiN jsou naprogramovány modely užívané empirickými výzkumníky a statistiky z oblasti vzdělání, sociologie a behaviorální psychologie. Oproti tomu PROC MIXED v SAS byl primárně naprogramován pro výzkum v oblasti zemědělství a fyziky. Výzkumník by tedy měl volit software (model) dle oblasti analýzy, viz Ferron (1997).

Tato kapitola stručně představuje balíčky napsané pro systém R umožňující modelovat LMM i používat nové přístupy k modelování mikropanelových dat (viz kapitola 3). V R je mnoho balíčků obsahujících funkce pro modelování LMM a GLMM, jmenovitě *amer*, *arm*,

[5] Varianční (rozptylová) funkce vyjadřuje vztah mezi střední hodnotou a rozptylem vysvětlované proměnné *Y*. Zároveň jednoznačně identifikuje rozdělení z exponenciální rodiny, viz Branda (2013).

gamm, gamm4, GLMMarp, glmmAK, glmmBUGS, heavy, HGLMMM, lme4.0, lmec, lmm, longRPart, MASS, MCMCglmm, nlme, PSM a *pedigreemm*. Porovnání vybraných charakteristik pro nejužívanější funkce (nejen R) shrnuje pro LMM tabulka P1 (v příloze).

V části 4.1 jsou popsány nejužívanější funkce systému modelující LMM klasickým (frekventistickým) přístupem. Balíčky pro marginální model shrnuje část 4.2. Přehled GLMM funkcí čtenář nalezne v části 4.3. Porovnání (nejen v R) základních rozdílů pro GLMM funkce shrnuje tabulka P2 (v příloze). Konečně bayesovský přístup k modelování LMM v R shrnuje část 4.4. Správné nastavení argumentů naprogramované funkce, která vybraný typ modelu odhaduje, je zásadní pro nalezení optimálního řešení (maxima věrohodnostní funkce).

V popisu funkcí se zaměřuji na argumenty funkcí, které jsou zásadní pro správný odhad parametrů modelu, avšak bývají v praxi často ignorovány a ponechány nevhodně na přednastavené hodnotě. V dalším textu je pomocí dvojtečky vždy vyznačena knihovna R (před dvojtečkou) a její vybraná funkce (za dvojtečkou) pro modelování panelových dat, tedy *balíček:funkce()*, případně je samostatně kurzívou označen balíček a funkce je označena kurzívou a prázdnou závorkou.

4.1 Frekventistické LMM balíčky

V této části se věnuji základním funkcím pro modelování *lme()* a *lmer()* z nejčastěji užívaných balíčkům *nlme* a *lme4*, které obsahují řadu funkcí pro modelování frekventistického přístupu k LMM. Balíčky *nlme* a *lme4* jsou nejužívanější v praxi a nejčastěji popisovány v pramenech například, viz Galecki a Burzykowski (2013). Důvodem jejich hojného užívání je rozsáhlá nabídka funkcí umožňující modelovat širokou škálu LMM. Jejich další výhodou je to, že obsahují různé praktické nástroje vhodné například pro diagnostiku. Navíc tyto dva balíčky patří mezi základ, z něhož vychází (pracují s nimi) mnohé z výše uvedených balíčků. Tyto dva balíčky patří k nejstarším a neustále zdokonalovaným nástrojům LMM v R a lze se tedy s větší důvěrou spolehnout na jejich syntax.

nlme

Funkce *nlme:lme()* využívá Laird-Wareho vzorec (2.12). Přestože je v jazyce S-PLUS (balíček *nlme3*) a R (*lme*) aplikován stejný model, jeho parametry odhadnuté metodou REML se mohou výrazně lišit, což na příkladech demonstruje Fox (2002), kdy v jednom případě se liší odhad rozptylu náhodného efektu ψ_q^2 a kovariance mezi náhodnými efekty $\psi_{qq'}$. Důvodem je odlišný

optimizátor[6] v programech. Optimizátory jsou použity z důvodu výpočetní náročnosti odhadu prvků kovarianční matice náhodných efektů **Σ** pomocí maximalizace (derivace) věrohodnostní funkce. V *nlme* lze pomocí funkce *lmeControl()* volit optimizátor. Optimizátor se volí pomocí nabídky *opt*. K dispozici je „nlminb“ (nastaven jako default) nebo „optim“.

„Nlminb“ je neomezený nebo krabicově omezený (omezení hodnot prvků vektoru odhadovaných parametrů shora nebo zdola) optimizátor využívající optimalizační metodu PORT. Metoda PORT je optimalizace převzatá z jazyka S+ napsaná v jazyce Fortran. Metodu PORT detailně popisuje Fox a kol. (1978).

Pro optimizátor „optim“ si lze vybrat optimalizační metodu. Přednastavená simplexová metoda „Nelder-Mead”, viz Nelder a Mead (1965) využívá pouze funkční hodnoty (nevyžaduje transformovat hodnoty). Jedná se o robustní, heuristický algoritmus, který je bohužel výpočetně pomalý. Alternativní metody „BFGS ”, „CG”, „L-BFGS-B”, „SANN” a „Brent” lze vybrat pomocí nabídky *optimMethod*.

- Numerická kvazi-Newton iterativní metoda „BFGS” rovněž známá jako algoritmus variabilní metriky (*variable metric algorithm*) využívá funkční hodnoty a gradienty. Je vhodná pro řešení neomezené nelineární optimalizace a odhady modelů s velkým počtem parametrů. Podmínkou nalezení optimálního řešení je nulová hodnota gradientu. O autorství této metody se dělí čtyři autoři: Broyden (1970), Fletcher (1970), Goldfarb (1970), Shanno (1970).
- Metoda sdružených gradientů „CG“ je modifikace „BFGS“. Metodu navrhli Fletcher a Reeves (1964). „CG“ je méně robustní než „BFGS“, ale z důvodu neukládání matic může konvergovat pro větší datové soubory a komplikovanější optimalizační metody. Tato metoda dosahuje nejhorších výsledků ve srovnání s ostatními uvedenými metodami dle Nashe (2014).
- Metoda „L-BFGS-B“ je modifikovaná „L-BFGS“, což je „BFGS“ algoritmus s limitací využité paměti vhodný pro úlohy s velkým počtem proměnných (>1000). Autor Byrd a kol. (1995) přiznává občasná selhání metody. Metody „BFGS“, „CG“ a „L-BFGS-B“ popisuje Nocedal a Wright (2006).
- Metoda „SANN” je založena na simulovaném „žíhání“, viz Belisle (1992). Tento postup náleží do skupiny stochastických globálních optimalizačních metod. Nevyžaduje deri-

[6]Optimizátor (*optimizer*) využívá vybraný numerický optimalizační algoritmus k řešení klasického optimalizačního problému bez omezení pro nelineární funkci. V kontextu této práce se využívá k odhadu regresních parametrů (tj. k nalezení maxima logaritmu věrohodnostní funkce) bez pomoci první derivace funkce.

vace a využívá funkčních hodnot. Lze ji využít pro nediferencovatelné funkce. Ve srovnání s ostatními metodami je relativně pomalá a neobsahuje časové ukončení iterace v případě problémů s nalezením optima.

- Metoda „Brent" řeší pouze úlohy v jedné dimenzi. Jedná se o komplement použitelný v kombinaci s jinými metodami.

Argument *correlation (*zkráceně lze použít *cor)* funkce *lme()* umožňuje vybrat některou z předdefinovaných kovariančních struktur náhodných chyb (volba kovarianční matice náhodných chyb $\mathbf{R}_i$ popsané v oddílu 2.2) a modelovat korelaci mezi náhodnými chybami pro stejný objekt *i*. Struktury lze členit dle lokalizace korelace na prostorové a časové:

Prostorové:

- *corExp* – exponenciální prostorová korelace,
- *corGaus* – gaussovská,
- *corLin* – lineární,
- *corRatio* – racionální kvadratická,
- *corSpher* – sférická.

Časové:

- *corCompSymm* – zaměnitelná rovnoměrná struktura (*a compound-symmetry structure corresponding to uniform correlation*), kdy se odhaduje pouze jeden konstantní korelační koeficient pro všechny dvojice měření uvnitř objektu.
- *corSymm* – obecná saturovaná struktura. Jedná se o nejsložitější strukturu s minimálním omezením a možností modelovat jakoukoliv validní korelační strukturu. Jediný požadavek na matici $\mathbf{R}_i$ je symetrická, pozitivně definitní matice s jedničkami na diagonále. Odhad všech prvků korelační matice bez zadání informace o jejich vzájemném vztahu (struktuře) může vést k výpočetním problémům.
- *corAR1* – autoregrese 1. řádu s diskrétně běžícím časem (pozorování jsou provedena v diskrétních časových intervalech).
- *corCAR1* – autoregrese 1. řádu se spojitě běžícím časem; tuto strukturu lze využít pro longitudinální data s nepravidelným a malým počtem opakování měření v čase, pro *i*-tý objekt.
- *corARMA* – ARMA model (model autoregrese a klouzavých průměrů v diskrétním čase).

Detailnější popis korelačních struktur náhodných chyb se nachází v monografii Pekára a Brabce (2012). Volba správné struktury vede k odhadům s menší variabilitou (odhad směrodatné chyby odhadu (*SE*)).

Popis dostupných vnitroobjektových korelačních struktur (volba struktury přes argument *correlation*) zájemce nalezne například v Box a Jenkins (1994); Venables a Ripley (2002); Littell a kol. (2006).

Argument *weights* funkce *lme* umožňuje vybrat některou z předdefinovaných varianční funkcí a modelovat heteroskedasticitu náhodných chyb pro stejný *i*-tý objekt, viz Pekár a Brabec (2012):

- *varFixed()* – rozptyl je úměrný hodnotě zadané spojité kovariátě *X*, tedy $\mathrm{Var}(\varepsilon_{it}) = \sigma^2 x_{it}$,
- *varIdent()* – rozptyl je konstantní pro všechna pozorování v *k*-té úrovni kategoriální proměnné (stratu) X_j a platí $\sigma_k^2 = \sigma^2 \delta_k^2, \quad \text{kde } \delta_1^2 = 1$,
- *varExp()* – rozptyl je exponenciální funkcí kovariáty X_j,
- *varPower()* – rozptyl je mocninnou funkcí kovariáty X_j,
- *varConstPower()* – rozptyl je dán součtem konstanty a mocninné funkce kovariáty X_j,
- *varComb()* – umožňuje kombinovat různé varianční funkce.

V případě, že není specifikována struktura, modeluje se kovarianční struktura s konstantním rozptylem pro všech T_i hodnot pro objekt *i*.

Kromě LMM umožňuje *nlme* odhadovat nelineární smíšené modely pomocí funkce *nlme()*. Autoři balíčku aplikují v *nlme:nlme()* nelineární smíšený model popsaný v Lindstrom a Bates (1990).

Systém O-O S3 je použit v balíčku *nlme* na rozdíl od *lme4*, kde je aplikován S4. Více o *nlme* v Pinheiro a Bates (2000). Manuál od Pinheira a kol. (2016) k *nlme* je dostupný online a jsou v něm popsány všechny dostupné funkce.

lme4

Stejně jako v balíčku *nlme* lze v balíčku *lme4* odhadovat LMM funkcí *lmer()* a nelineární smíšené modely funkcí *nlmer()*. Oproti balíčku *nlme* je v tomto balíčku funkce *glmer()* sloužící k odhadování GLMM. Knihovna *lme4* nahrazuje svoji starší verzi *lme4.0*, která byla vyvíjena a používána do roku 2011.

Optimizátory „*Nelder-Mead*“, nebo „*bobyqa*“ lze volit pro optimalizaci funkce *lmer()* a *glmer()* pomocí funkce *lmerControl()* a *glmerControl()*. „*Bobyqa*“ umožňuje krabicové omezení

hodnot parametrů (zdola a shora). Minimalizuje funkci mnoha proměnných s využitím metody důvěryhodné oblasti (*trust region method*). Algoritmus je popsán v Powell (2009).

Další rozdíly mezi *lme4* a *nlme* kromě výše zmíněných jmenovitě: jiný systém práce s objekty a modely (O-O), jiný optimizátor, funkce pro jiné modely než LMM shrnuje Bates a kol. (2016) v manuálu k *lme4*:

- *lme4* používá moderní, efektivní metody lineární algebry implementované v balíčku *Eigen*.
- *lme4* využívá referenční třídy za účelem zefektivnění práce s velkými objekty a zrychlení výpočtu.
- V současné době je v *lme4* implementováno méně nástrojů (než v *nlme*) pro modelování heteroskedasticity a korelace náhodných chyb. Nedisponuje předdefinovanými kovariančními a korelačními strukturami.
- *lme4* je vybaven nástroji pro profilování věrohodnosti a parametrický bootstrapping.
- *lme4* je naprogramován tak, aby byl více modulární než *nlme*. Tím je myšleno to, že vývojáři balíčků a koncoví uživatelé mají možnost využívat jednotlivé komponenty funkce a modifikovat či implementovat je do svých specifických aplikací LMM. Další výhodu lze spatřovat ve větší flexibilitě specifikovat různé optimalizační funkce pro odhad náhodných efektů.
- V odhadu kovarianční matice o rozměrech $q \times q$ náhodných efektů $\mathbf{\Sigma}$ metodou IRLS je v *lme()* implementován výpočetní vzorec $\mathbf{\Sigma} = \sigma^2\mathbf{W}^{-1}$. Oproti tomu je v *lmer()* vyjádřena $\mathbf{\Sigma} = \sigma^2\mathbf{\Lambda}\mathbf{\Lambda}^{\mathbf{T}}$ s využitím relativního kovariančního faktoru $\mathbf{\Lambda}$, což je matice o rozměrech $q \times q$, závislá na parametru varianční komponenty, která generuje $\mathbf{\Sigma}$. Díky tomu konverguje (k maximu věrohodnostní funkce) *lmer()* na rozdíl od *lme()* v případě rovnoměrného povrchu bez zjevných modů.
- Rozdíl je i v intervalových odhadech náhodných efektů. Směrodatná chyba (SE) náhodných efektů je v *nlme* počítána asymptoticky. Funkce *intervals()* spočítá Waldovy intervalové odhady náhodných efektů na transformované škále, logaritmická transformace pro rozptyly a transformace tanh(x)/2 pro korelační parametry, kde tanh(x) je hyperbolická tangens funkce x. *Lme4* odhaduje SE pomocí bootstrapingu. Waldovy odhady lze získat i pro *lme4* nepřímo pomocí interní varianční funkce (*devfun2*), či delta metodou. Jak postupovat lze nalézt v Bolkerově online návodu (2016).

Technický popis výpočtu LMM v *lme4* a detailní popis výše uvedených rozdílů mezi balíčky uvádí Bates a kol. (2015a).

K *lme4* existuje komplementární balíček *LMERConvenienceFunctions*, který obsahuje nástroje pro funkce *lmer()* a *glmer()*. V balíčku jsou funkce sloužící k:

- Testování významnosti pevného efektu v modelu, neboli k výběru vhodného modelu z množiny modelů s různými pevnými efekty zpětnou metodou (*backward selection*). Pro zpětnou selekci jsou k dispozici dvě funkce. Funkce *bfFixefLMER_F.fnc()* zpětnou metodou vybírá model dle *p*-hodnot ANOVA F-testu nebo testu věrohodnostním poměrem, pro který jsou k dispozici informační kritéria AIC, BIC (Schwarzovo Bayesovské informační kritérium), (rel.likAIC a rel.likBIC). Nelze však použít pro *glmer()*. Pro *glmer()* lze aplikovat metodu zpětného výběru přes funkci *bfFixefLMER_t.fnc()*. V této funkci se vybírá model dle absolutní hodnoty testového kritéria Waldova t-testu nebo testu věrohodnostním poměrem. Dle Pinheira a Batese (2000) není doporučeno používat test věrohodnostním poměrem pro testování zařazení pevného efektu do GLMM v případě malých a středně velkých výběrů. Jeho použití pro testování pevných efektů v GLMM je adekvátní pouze v případě, že poměr velikosti výběru k počtu testovaných pevných parametrů (počet úrovní pevných efektů pro kategoriální X_j) a zároveň je poměr rozsahu výběru k počtu úrovní náhodného efektu (objekt, T) velký.
- Testování signifikance náhodného efektu pomocí testu věrohodnostním poměrem, který porovnává model bez a s testovaným náhodným efektem z hlediska shody odhadnutých a empirických hodnot. Pokud je komplexnější model (s náhodným efektem) signifikantně lepší, je zachován náhodný efekt v modelu. Lze využít i pro *glmer()*.
- Zobrazení diagnostických grafů funkcí *mcp.fnc()*.
- Post-hoc analýze modelu funkcí *mcposthoc.fnc()*, která využívá balíček *parallel*. Post-hoc analýza je provedena pomocí MCMM. Tato analýza odhaduje SE pevných efektů a odhaduje P-hodnoty (dolní a horní mez) dílčích *t*-testů o parametrech.
- Odhadu horní a dolní meze intervalu spolehlivosti pro P-hodnoty u ANOVA funkcí *pamer.fnc()*. Zatím nefunguje pro *glmer()*.
- Automatickému odstranění extrémních měření uvnitř objektu pomocí funkce *perSubjectTrim.fnc()*.
- Grafickému zobrazení Kernelových odhadů hustot dvou číselných proměnných funkcí *plotDensity3d.fnc()*.
- Grafickému zobrazení dílčích efektů jednotlivých prediktorů funkcí *plotLMER.fnc()* a *plotLMER3d.fnc()*.

Více o *LMERConvenienceFunctions* v manuálu Tremblaye (2015).

Dalším zajímavým komplementárním balíčkem k *lme4* je balíček *lmerTest*, který umožňuje provádět různé druhy testů u modelů LMM a GLMM implementovaných v *lme4*. Mezi tyto testy patří *F* test typu I až III, individuální *t*-testy o pevných efektech založené na testovém kritériu *F* se Satterthwaiteovu nebo Kenwardovo-Rogerovo aproximací počtu stupňů volnosti ve jmenovateli a test poměrem věrohodnosti pro náhodné efekty. Balíček dále umožňuje konstruovat intervalové odhady pevných efektů. Více o balíčku v Kuznetsova a kol. (2016).

Popis rozdílů mezi *nlme* a *lme4* jsem publikoval s kolektivem autorů, viz Sobíšek a kol. (2015).

4.2 Marginální model v R

V této části jsou shrnuty balíčky, které obsahují funkce pro modelování MMM a GMMM, viz oddíl 2.3 a 3.1.

nlme

V balíčku *nlme* lze odhadovat marginální model pomocí funkce *gls()*. Předdefinovanou korelační strukturu náhodných chyb lze volit stejně jako u *lme()* funkce pomocí argumentu *correlation* a varianční funkci zadáním argumentu *weights*. Více o parametrizaci v manuálu k *nlme*, viz Pinheiro a kol. (2016).

mmm

Relativně nový balíček z roku 2014 umožňuje odhadovat vícerozměrné marginální zobecněné lineární modely (GMMM) vhodné pro simultánní analýzu více (alespoň dvou) vysvětlovaných proměnných, které mohou být spojité (normálně nebo gamma rozdělené), diskrétní (binární - binomické rozdělení, absolutní nebo relativní četnosti - Poissonovo rozdělení)). Odhady parametrů a jejich směrodatných chyb jsou provedeny metodou GEE, kterou popisuji v oddílu 3.1. Zájemce nalezne popis balíčku v online manuálu Asara a Ilka (2014) a v článku Asara a Ilka (2013).

geepack

V balíčku *geepack* ve funkci *geeglm()* je implementován GEE k odhadu marginálních zobecněných modelů pro vysvětlované proměnné pocházející z exponenciálního rozdělení. Stejně jako v marginálním modelu se odhadují populační průměry s ohledem na korelaci uvnitř objektu bez plné specifikace spojitého rozdělení. K dispozici je pět korelačních struktur pro vnitroobjektová rezidua:

- *independence* – předpokládá „nezávislá“ měření uvnitř objektu a je vhodná v situaci, kdy nedokážeme tvar závislosti popsat,
- *exchangeable* – předpokládá konstantní korelaci a odhaduje tedy pouze jeden korelační koeficient pro *i*-tý objekt,
- *unstructured* – obecná struktura (viz *corSymm* z *nlme*) umožňuje modelovat libovolnou korelaci pro každou dvojici měření pro *i*-tý objekt,
- *userdefined* – uživatel může nastavit specifickou hodnotu korelace v případě apriorní znalosti o korelační struktuře,
- *ar1* – autokorelace prvního řádu, viz *ar1* z *nlme*.

Autory balíčku jsou Højsgaard a kol. (2006).

4.3 GLMM v R

Bolker (2008) ve svém přehledu upozorňuje na časté nevhodné užívání GLMM. U 58 % (311/537) článků (publikovaných v letech 2005–2008, nalezeno přes Google Scholar) z oblasti ekologie a evoluce nalezl chybnou aplikaci GLMM (volba efektů, volba metody odhadu a jiné). Aktuálním problémem je nalezení vhodného informačního kritéria nebo postupu, který pomůže s volbou vhodné kovarianční struktury. Odpůrci navrhovaných tzv. kvazi kritérií (např. kvazi AIC) oponují, že kritéria implementovaná ve funkcích *glm()*, *glmmPQL()*, aj. založená na věrohodnosti nejsou vhodná pro GLMM, které využívají iterativní algoritmy odhadu parametrů (např. REML). Řešení nabízí Bolker (2014).

Porovnání nejužívanějších funkcí (nejen R) shrnuje tabulka P2 pro GLMM z hlediska použité metody pro odhad parametrů, inferenčních odhadů a typu modelu, který může být odhadován.

GLMM, GAMM a ZIGLMM modely v R

Zobecněné lineární modely jsou implementovány v následujících balíčcích programu R:

- *MASS* – obsahuje funkci *glmmPQL()* pro odhad GLMM metodou PQL s možností zahrnout více náhodných efektů, více viz Venables a Ripley (2002).
- *lme4* – obsahuje funkci *glmer()*, která umožňuje odhadovat regresní parametry dvěma metodami Laplaceho aproximací nebo Gaussovou-Hermitovou kvadraturou. Druhou metodu lze použít pouze pro model s jedním náhodným efektem. K dispozici jsou přednastavené (doporučené) kombinace rozdělení vysvětlované proměnné a spojovacích funkcí, stejné jako ve funkci *glm* určené pro GLM modely, a to: binomické (spojovací

funkce = „logit“, tj. $g(E(Y_i)) = \log(\mu/(1 - \mu))$, kde μ je střední hodnota normálního rozdělení), normální („identita“, tj. $g(E(Y_i)) = \mu$), gamma („inverzní“, tj. $1/\mu$), inverzní Gaussovo (Waldovo) ($1/\mu^2$), Poissonovo („log“, tj. $\log(\mu)$), kvazi (toto „rozdělení“ umožňuje vybrat z velkého počtu spojovacích a variančních funkcí. Pro ostatní rozdělení je varianční funkce dána rozdělením. Spojovací funkce je přednastavena na „identitu“ a varianční funkce je „rozdělení“), kvazi binomické („logit“) a kvazi Poissonovo („log“). Kvazi binomické a kvazi Poissonovo rozdělení se liší od binomického a Poissonova pouze v tom, že hodnota disperzního parametru[7] není konstantní (rovna jedné), ale je odhadnuta. Tato rozdělení lze použít v případě výskytu overdisperze (nadměrného rozptylu). Gelnarová (2007) popisuje overdisperzi jako situaci, kdy je skutečný rozptyl pozorovaných dat větší než rozptyl predikovaný modelem. Odhady rozptylů získaných na základě modelu jsou tedy podhodnocené a predikční intervaly pro odhad hodnot vysvětlované proměnné příliž úzké. V případě overdisperze u Poissonova rozdělení je porušen předpoklad shody střední hodnoty a rozptylu. Overdisperze může být způsobena přítomností vlivných shluků v populaci (např. rodiny, domácnosti, sídla, etnické skupiny, oblasti), které se odlišují od běžné sledované populace, tedy je-li sledovaná populace heterogenní směsí několika populací. Jedním z řešení problémů overdisperze je vynásobení odhadu rozptylu disperzním parametrem. Popis balíčku *lme4* a odkazy na něj se nacházejí v oddílu 4.1.

- *glmmADMB* – je balíček postavený na open-source programu AD Model Builder s nelineárním enginem pro odhad GLMM přes funkci *glmmadmb()*. Rovněž lze odhadovat ZIGLMM (oddíl 3.2) pomocí funkce *glmmadmb()* nebo balíčku *MCMCglmm* (oddíl 4.4). Postup využití tohoto balíčku pro modelování GLMM popisují Bolker a kol. (2012a). Vysvětlovaná proměnná se může řídit rozděleními: Poissonovo, binomické, negativně binomické, gamma, beta, cenzorované Poissonovo a negativně binomické. Spojovací funkci lze volit z následující nabídky: log, logit, probit, cloglog, inverzní, identita. Dále lze provádět inferenční odhady MCMC přístupem. Autoři Bolker a kol. (2012b) objasňují, jak správně používat *glmmADMB*, v online tutoriálu,
- *mgcv* – obsahuje funkce *gamm()* a *gam()*. Funkce *gamm()* slouží k odhadu GLMM pomocí zobecněných aditivních modelů se smíšenými efekty (GAMM) popsanými v oddílu 3.2. Pro odhad aditivního modelu s náhodným efektem pro normálně rozdělenou

[7] Disperzní parametr je koeficient určující variabilitu vysvětlované proměnné. Rozptyl Y je dán součinem varianční funkce a (funkcí) disperzního parametru.

vysvětlovanou proměnnou (AMM) využívá *gamm()*metody odhadu implementované v *lme()*. Pro odhad GAMM (vysvětlované proměnné se řídí jiným než normálním rozdělením z exponenciální rodiny) si *gamm()* volá *glmmPQL()*. Pro spline sahá *gamm()* do argumentů funkce *gam()*. Funkce *gam()* slouží k odhadu zobecněných aditivních modelů bez náhodných efektů. Šmilauerův (2007) český tutoriál ke *gam()* je online. Struktura náhodných efektů a chyb je vzata z *lme()*. Woodův manuál (2016) k balíčku *mgcv* je rovněž k dispozici.

- *gamm4* – obsahuje funkci *gamm4()*, která slouží stejně jako *gamm()* k odhadu GLMM. Rozdíl je v tom, že využívá algoritmus odhadu parametrů implementovaných v *lme4*; konkrétně pro normálně rozdělenou chybovou složku využívá *lmer()* a *glmer()* pro chyby pocházející z exponenciální rodiny rozdělení. *Gamm4()* je numericky robustní ve srovnání s *gamm()*. V balíčku jsou využity jiné REML metody než PQL, což přináší lepší odhady pro binární a multinomická data s malým počtem obměn. S tímto typem *Y* má PQL a tedy *gamm*() problém konvergovat. Nevýhodou *gamm4()* oproti *gamm()* je nemožnost využít korelační struktury z *nlme* a také si nedokáže poradit s vícerozměrnými spline. Více ve Woodově manuálu k balíčku (2015).
- *dhglm* – odhaduje dvojitý hierarchický zobecněný lineární model (*double hierarchical generalized linear models, DHGLMs*). DHGLM modeluje obě komponenty náhodných efektů: střední hodnotu a variabilitu. Vysvětlovaná proměnná se může řídit rozděleními: Gaussovo, gamma, inverzní-gamma nebo beta. Poradí si s křížovým i hierarchickým návrhem. Pevné efekty mohou být odhadnuty maximalizací *h*-věrohodnosti, viz Lee a kol. (2006) nebo Laplaceovou aproximací prvního řádu. Více o DHGLM viz Lee a Nelder (2006) a manuál k balíčku, viz Lee a Noh (2015).

CLMM modely v R

Balíček *ordinal* je určen pro odhad CLMM (oddíl 3.2). Dovoluje odhadovat model s více smíšenými efekty (křížovými, hierarchickými nebo oběma typy najednou). CLMM lze odhadovat REML metodami Laplaceovy aproximace nebo Gaussovy-Hermiteovy kvadratury (pro jeden náhodný efekt) popsanými v oddílu 3.2. K dispozici jsou dvě funkce *clmm()* a *clmm2()*. *Clmm()* je novější implementace CLMM a do budoucna má zcela nahradit *clmm2()*. Nevýhodou *clmm2()* je například možnost odhadovat pouze jeden náhodný efekt.

Christensův manuál (2015a) k *ordinal* a tutoriál (2015b) pro použití funkce *clmm2()* lze najít online.

Ordinální proměnnou lze dále modelovat pomocí funkce *MCMCglmm()* nebo přes funkci *lmer()* po přeškálování všech hodnot *y* na *z*-skóre (oddíl 3.2).

4.4 Bayesovský přístup v R

Bayesovský přístup k modelování korelovaných dat je stručně popsán v části 3.3. Zde představím tři balíčky systému R pro bayesovské modelování LMM a GLMM, jmenovitě: *blme*, *MCMCglmm* a *brms*.

blme

Balíček *blme* rozšiřuje možnosti balíčku *lme4*. Oproti *lme4* dovoluje provádět posteriorní inferenční odhady a zadávat apriorní informaci o rozdělení regresních parametrů argumenty *fixedef.prior*, *cov.prior*, *resid.prior*. *Blme* obsahuje funkci *blmer()* pro odhad LMM a *bglmer()* pro odhad GLMM.

Funkce *blmer()* a *bglmer()* mají identickou syntaxi jako funkce *lme4:lmer()* a *lme4:glmer()*. *Bglmer()*umožňuje vypořádat se s úplnou separací, zadáním apriorní informace o rozdělení pevného efektu pomocí argumentu *fixedef.prior*. Kromě *t*-rozdělení navrhovaného v části 3.3 lze stanovit normální rozdělení pevných efektů.

Argumentem *cov.prior* lze nastavit apriorní informaci o rozdělení kovarianční matice náhodných efektů, případně o dalších regresních parametrech vyskytujících se v modelu za znakem „/“. Lze specifikovat vlastní rozdělení (*custom*) náhodných efektů. Podporovány jsou dále rozdělení gamma, inverzní gamma, Wishartovo a inverzní Wishartovo.

Argument *resid.prior* specifikuje rozdělení rozptylu náhodných chyb. Aktuálně je na výběr gamma, inverzní gamma a bodové (point) rozdělení.

Více info o *blme* ve článku Chunga a kol. (2013) a v online manuálu k balíčku od Dorieho (2015).

MCMCglmm

MCMCglmm odhaduje parametry GLMM technikou MCMC, Hadfield (2010) stejnojmennou funkcí *MCMCglmm()*. Vysvětlovaná proměnná může pocházet z rozdělení normálního, exponenciálního, Poissonova, Poissonova s častým výskytem nulových hodnot (ZIPois, oddíl 3.2), binomického a multinomického. Data mohou být cenzurována (zleva, zprava, intervalově). Lze provádět vícerozměrné modelování a simultánně modelovat více vysvětlovaných proměnných, které pocházejí z různých rozdělení. *MCMCglmm()* pracuje s různými kovariančními strukturami náhodných chyb a efektů (heterogenní, obecná, RSM). Autor balíčku v Hadfieldově

článku (2010) porovnává výpočetní efektivitu algoritmu jeho funkce s MCMC algoritmem (implementovaným v softwaru WinBUGS). V tomto testu konvergovala funkce *MCMCglmm()* 120krát rychleji k optimu (maximu věrohodnostní funkce).

brms

Komplexní balíček *brms* byl publikován nedávno v roce 2015. Bayesovské regresní modelování v tomto balíčku využívá open source programovací jazyk Stan (http://mc-stan.org/), který je zaměřený na implementaci bayesovské statistické inference. *Brms* dovoluje modelovat vysvětlovanou proměnnou pocházející z exponenciální rodiny rozdělení včetně Poissonova a negativně binomického s častým výskytem nulových hodnot. Oproti jiným balíčkům však umožňuje:

- uvažovat jiná rozdělení vysvětlované proměnné: Weibullovo, gamma, exponenciální a inverzní gaussovské rozdělení a s využitím logaritmické spojovací funkce provádět regresní analýzu přežití,
- uvažovat pro *Y* kumulativní rozdělení (kumulativní, kontinuální poměr (*contiuation ratio*), zastavený poměr (*stopping ratio*) a přilehlé rozdělení (*adjecent category*)) a provádět ordinální regresní analýzu.

K dispozici jsou tyto korelační struktury náhodných chyb: AR(*p*), MA(*q*), ARMA(*p*,*q*) a ARR(*r*). Funkce *brm()* pro odhad modelu má podobnou syntax jako balíček *lme4*. Více o balíčku v manuálu (Buerkner, 2016).

5. Shlukování trajektorií

V kapitole 2 upozorňuji na problém s odhadem náhodných efektů LMM, zejména pro datový soubor s obecnou kovarianční maticí náhodných chyb **R**. Náhodné efekty jsou odhadovány vybranou iterativní optimalizační metodou. Takový odhad nemusí být nezkreslený, jelikož optimalizační algoritmus nemusí nalézt globální maximum věrohodnostní funkce, ale nalezne pouze lokální maximum věrohodnostní funkce. Další problém jsou nepřesné intervalové odhady pevných efektů pro datové soubory s velkým počtem objektů a obecnou kovarianční maticí **R**. Tento problém může nastat i u jednoduchých modelů s jednou vysvětlující proměnnou X, například pokud je cílem regresní analýzy odhadnout model, který dokáže s požadovanou přesností predikovat změny vysvětlované proměnné Y v čase (hodnoty y_{it}) pro i-tý objekt. V situaci, kdy vývoj vysvětlované proměnné v čase je heterogenní mezi objekty, tzn. funkční závislost Y na čase se liší, populační bodový odhad pevného efektu času na změnu Y nebude nezkreslený pro většinu objektů a intervalový odhad bude nepřesný (SE pevných efektů budou příliš široké). Dalším problémem je výpočetní čas pro velký počet objektů a složitější modely s větším počtem vysvětlujících proměnných X_j a Z_q. Odhad modelu může trvat hodiny nebo dny nebo vůbec nebude z důvodu výpočetní náročnosti dokončen. Identické problémy při odhadu regresních parametrů se týkají i GLMM (kapitola 3).

Problémy s odhady parametrů mohou být řešeny v řadě případů pomocí shlukové analýzy, která by v exploratorní fázi (před regresní analýzou) pomohla identifikovat objekty s podobnou kovarianční strukturou vysvětlované proměnné. Náležitost do shluku může vstoupit v dalším kroku do regresního modelu například formou kategoriální pomocné vysvětlující proměnné a její interakce s ostatními vysvětlujícími proměnnými (např. s pomocnou časovou proměnnou t). V kapitole 8 testuji pomocí simulovaných dat, zda tento postup zpřesní odhady regresních modelů. Další možností využití získané náležitosti do shluku pro zpřesnění intervalových odhadů pevných efektů popisuji v oddílu 6.1.

V této kapitole se zabývám vybranými shlukovacími metodami určenými pro analýzu prostorových dat. Dále shrnuji jejich eventuální modifikace pro mikropanelová data. Pro shlukování mikropanelových dat se nejčastěji aplikují hierarchické shlukování a metody rozkladu (C-průměrů) popsané v oddílu 5.2. Často se však tyto metody aplikují přímo na původní data bez jakékoliv modifikace, která by zohledňovala časový rozměr dat. Z novějších přístupů se pro shlukování časových trajektorií používají metody založené na modelu (modelový přístup), blíže specifikované v oddílu 5.1. Tyto metody je však vhodné používat pro makropanelová data s velkým počtem měření v čase. Výhodou metody C-průměrů oproti modelovému přístupu je to, že

nepředpokládá normalitu nebo jiné parametrické předpoklady o shlucích (třídách), což lze kvitovat v situaci, kdy nemáme apriorní informaci o rozdělení. Metody rozkladu navíc nepředpokládají žádné rozdělení individuálních trajektorií a nezávisí na časové škále. Z těchto důvodů modelový přístup neuvažuji ve zbytku této práce, která je zaměřena na analýzu mikropanelových dat.

Před několika lety jsem se spoluautory navzájem porovnal na reálných datech různé metody shlukování primárně určené pro průřezová data, viz Mackovičová a kol. (2011); Řezanková a Sobíšek (2011); Sobíšek a kol. (2011); Sobíšek a Řezanková (2011). Tyto metody nejsou užívány ani modifikovány pro shlukování mikropanelových dat. Zřejmě z důvodu nízkého povědomí o těchto možnostech shlukování mezi výzkumníky, kteří se nezabývají shlukovou analýzou, volí výzkumníci nejčastěji nejznámější (a jednoduché) metody: hierarchické shlukování a metodu *C*-průměrů. Do budoucna plánuji posoudit použitelnost méně známých metod, které dosáhly dobrých výsledků pro mnou analyzovaná průřezová data, avšak netradičních pro shlukování mikropanelových dat. V této práci tyto metody stručně charakterizuji z důvodu rozšíření informovanosti o jejich existenci.

5.1 Obecný úvod – kategorizace metod shlukové analýzy

Shluková analýza se řadí mezi klasifikační metody. Ty se člení na metody učení s učitelem (*supervised learning*) a bez učitele (*unsupervised learning*). Úlohou klasifikačních metod je vytvoření modelu (pravidla), který umožní zařadit objekt do správné skupiny (třídy) na základě hodnot vysvětlujících proměnných. U první skupiny klasifikačních metod (učení s učitelem) je pro výběrový soubor známá příslušnost ke skupinám (tedy hodnotami vysvětlované proměnné). Často se užívá logistická regrese, což je specifický případ zobecněného lineárního modelu. Mezi starší a osvědčené metody patří diskriminační analýza. U druhé skupiny klasifikačních metod neznáme náležitost do skupin (hodnoty vysvětlované proměnné). Tyto metody jsou navrženy za účelem průzkumu dat. Slouží k identifikaci a popisu předem neznámých shluků (skupin), které jsou vnitřně homogenní a navenek heterogenní. Mezi tyto metody se řadí například metoda hlavních komponent nebo různé shlukovací algoritmy, jako například hierarchické shlukování, metoda *C*-průměrů, fuzzy algoritmy, algoritmy inspirované biologickými organismy (samoorganizující se mapy, algoritmus neuronového plynu). K novějším přístupům patří například modely latentních tříd. Neexistuje jednotný návod pro volbu metody, jelikož každá z metod má svoje výhody a omezení, a je ji třeba pečlivě volit na základě charakteristiky ana-

lyzovaného souboru. K dispozici je mnoho monografií popisujících (nejen) výše uvedené metody, např. Breiman a kol. (1984); Hastie a kol. (2001); Hagenaars a McCutcheon (2002); Congdon (2006); Härdle a Simar (2007), v češtině např. Hebák a kol. (2007).

Pro shlukovou analýzu (učení bez učitele) je důležité vybrat vhodný shlukovací algoritmus, míru podobnosti/nepodobnosti a kritérium pro hodnocení výsledků shlukování. Shlukovací algoritmy pro časové řady vycházejí z existujících algoritmů pro shlukování prostorových dat. Modifikují uvedené přístupy tak, aby zohlednily charakteristiky časových řad. Často se neupravuje samotný algoritmus, ale výpočet měr podobnosti mezi objekty. V závislosti na typu proměnných a charakteristice časových řad se vybírá vhodná míra podobnosti a metoda. Neexistuje univerzální postup a volba metody závisí na konkrétní úloze. Dalším přístupem je transformace časových řad na prostorová data a na ta se aplikují zavedené metody.

Shlukovací metody lze členit do dvou skupin dle výsledku shlukování na:

- disjunktní – objekt je zařazen jednoznačně právě do jednoho shluku,
- fuzzy – pro daný objekt je míra příslušnosti k danému shluku vyjádřena reálným číslem z intervalu 0 až 1. Součet měr příslušnosti ke každému shluku pro daný objekt je roven jedné.

Řezanková a kol. (2009) kategorizují shlukovací metody prostorových dat do tří skupin:

- Metody rozkladu (partitioning) – rozkládají data na *C* částí. Každá část reprezentuje shluk obsahující alespoň jeden objekt. Z disjunktních metod se do této skupiny řadí metody *C*-průměrů, *C*-medoidů, *C*-modů, *C*-histogramů; z fuzzy postupů metody fuzzy *C*-průměrů či fuzzy *C*-medoidů. Modifikaci *C*-medoidů pro velké datové soubory popisujeme a aplikujeme na reálných datech spolu s kolektivem autorů, viz Mackovičová a kol. (2011). Jedná se o algoritmus CLARA, který v sobě zahrnuje PAM algoritmus. Využívá přitom minimalizaci součtu vzdáleností mezi každým objektem ve shluku a příslušným reprezentantem shluku, přičemž reprezentantem může být pouze některý z objektů shluku (na rozdíl od metody *C*-průměrů, kde se může stát reprezentantem jakýkoliv jiný objekt).
- Metody hierarchické – shlukují objekty do hierarchického stromu shluků. Řadí se mezi ně aglomerativní a divizivní shlukování.
- Novější přístupy – lze dále podle Hana a Kambera (2001) členit na 3 skupiny:

  1. metody založené na hustotě – jejich hlavní myšlenka tkví v tom, že shluk se zvětšuje do té doby, než hustota (počet objektů nebo datových bodů v okolí) nepřekročí

stanovený limit. Příkladem je algoritmus OPTICS, viz Ankerst a kol. (1999) nebo algoritmus určený pro velké datové množiny DBSCAN. Tento algoritmus popisujeme a porovnáváme shlukovací schopnost tohoto algoritmu s algoritmem CLARA na reálných datech, viz Mackovičová a kol. (2011). Velkou výhodou metod založených na hustotě oproti metodám rozkladu je schopnost rozpoznat shluky různých tvarů.

2. metody založené na mřížce – rozdělují datový prostor do konečného počtu pravoúhlých buněk, které tvoří mřížkovou strukturu. Všechny shlukovací operace jsou prováděny na této struktuře. Příkladem je algoritmus STING autora Wanga a kol. (1997).

3. metody založené na modelu – předpokládají existenci modelu pro každý shluk. Modelový přístup lze členit na dvě skupiny: statistický přístup a neuronové sítě. Řezanková a kol. (2009) do této kategorie navíc řadí metody pro shlukování podprostorů určené pro datové soubory s velkým počtem proměnných.

Modelový statistický (pravděpodobnostní) přístup modeluje data jako směs podmíněných nezávislých latentních tříd. Patří do skupiny modelů směsí rozdělení (finite mixture models, gaussovské lineární procesy). Statistický přístup předpokládá, že existuje relativně malý počet nepozorovaných (skrytých) proměnných C_c, kde $c = 1, 2, \ldots, C$. Tyto latentní proměnné stojí v pozadí vzájemných vztahů mezi sledovanými proměnnými X_j, $j = 1, 2, ..., J$, kde C_c se řídí předpokládaným rozdělením. Odhaduje se pravděpodobnost náležitosti jednotlivých objektů do C tříd. Modely směsí se od sebe vzájemně liší v požadavcích na typ a pravděpodobnostní rozdělení pozorovaných a skrytých proměnných. Tyto modely předpokládají podmíněnou nezávislost pozorování uvnitř jednotlivých tříd. Více o modelech pojednává například Vermunt a Magidson (2002), Gelman (2004). Naše vlastní aplikace modelů směsí na prostorová data a porovnání jejich klasifikačních schopností s jinými přístupy jsou obsaženy v publikacích Sobíška a Řezankové (2011) a Řezankové a Sobíška (2011).

Přístup neuronových sítí jsme rovněž aplikovali na reálná data, viz Sobíšek a kol. (2011). Konkrétně jsme použili metody neuronový plyn a samoorganizující se mapy (SOM). Tyto mapy jsou podle jejich autora nazývány též Kohonenovy mapy, viz Kohonen (2001). Využívají se k transformaci nelineárních vztahů uvnitř trénovací datové množiny prostřednictvím uzlů (neuronů) na méně rozměrný (1- nebo 2- rozměrný) prostor $\mathbf{R}^n \rightarrow \mathbf{R}^2$. Každý neuron z množiny neuronů C_c, $c \in \{1, ..., C\}$ je reprezentován d-rozměrným váhovým vektorem (nazývaným prototyp nebo referenční vektor):

$\mathbf{c}_c = [c_1, ..., c_d]$, kde d je rozměr vstupního prostoru ($\mathbf{R}^d$) a dolní index c značí c-tý neuron, $c = 1, ..., C$. Vztahy mezi neurony jsou charakterizovány svým okolím, které tvoří strukturu mapy (topologii a tvar). Oproti SOM algoritmus neuronového plynu, viz Martinetz a kol. (1993) v sobě nezahrnuje fixní topologii mřížky, ale obsahuje seznam tzv. „pořadí sousedů“.

Výše uvedené metody jsou blíže popsány v monografiích Řezankové a kol. (2009) a Hana a Kambera (2001). Z široké nabídky výše popsaných metod se dosud užívají pouze některé z nich pro mikropanelová data.

Přehled přístupů ke shlukování časových řad a kategorizaci do tří skupin dle použití původních dat uvádí Liao (2005):

- Metody založené na modelu – s původními daty se pracuje nepřímo přes modely odhadnuté z původních dat. Původní hodnoty (časové řady) jsou transformovány na parametry modelu. Předpokládá se, že časové řady jsou generovány nějakým modelem nebo směsí latentních pravděpodobnostních rozdělení. Modifikace gaussovských procesů pro shlukování longitudinálních dat formulují Koestler a kol. (2014); De la Cruz-Mesía a kol. (2008). Tyto nelineární hierarchické modely směsí oproti klasickým modelům směsí v sobě zahrnují náhodné efekty, díky čemuž se vypořádají s vnitroobjektovou autokorelací. Modifikovaný modelový přístup (modely ARMA, hierarchické klouzavé modely, gaussovské procesy, AR, Markovův řetězec atd.) je použitelný pro shlukování dostatečně dlouhých stacionárních časových řad. Různé modely směsí rozdělení a hierarchický model popisuje a pomocí simulace porovnává například Pešout (2010).
- Metody založené na práci s původními daty – původní hodnoty jsou shlukovány přímo. Hlavní modifikace metod pro prostorová data spočívá v úpravách měr vzdálenosti/podobnosti, aby reflektovaly časovou dimenzi dat. Takovou úpravu euklidovské míry vzdálenosti navrhují Lombardo a Falcone (2011).
- Metody založené na práci s transformacemi původních hodnot – shlukovány jsou charakteristiky odvozené z původních hodnot. Časové řady jsou transformovány na vektor hodnot nižší dimenze. Poté se na tento vektor charakteristik aplikuje konvenční algoritmus. Transformace vhodné pro časové řady s velkým počtem opakování v čase (makropanel) shrnuje Liao (2005). Jedná se například o Haar wavelet decomposition, sequence cluster refinement algorithm, perceptually important point identification algorithm. Další strukturální charakteristiky navrhují Wang a kol. (2006). Navrhované charakteris-

tiky reprezentují následující vlastnosti makropanelových časových řad: trend, sezónnost, periodicita, sériová korelace, šikmost, špičatost, chaos, nelinearita a vlastní podobnost (self-similarity). Autoři testují shlukovací schopnost těchto charakteristik v kombinaci s hierarchickým shlukováním a SOM. Wang a kol. (2006) tvrdí, že dokáží efektivně shlukovat dlouhodobé časové řady, tj. data s vysokou dimenzí, pomocí tohoto přístupu (transformace původních hodnot) na rozdíl od ostatních přístupů.

V následující části blíže popisuji metody hierarchického shlukování, *C*-průměrů a fuzzy *C*-průměrů. Metody hierarchického shlukování a *C*-průměrů jsou kombinovány s originálním návrhem charakteristik, které reprezentují tvar trajektorie, pro mikropanelová data. Svůj návrh výpočtu charakteristik a testování jejich použitelnosti v kombinaci s různými metodami shlukování popisuji v kapitole 7.

5.2 Popis vybraných shlukovacích metod

V této části se zabývám vybranými shlukovacími metodami určenými pro analýzu prostorových dat. Dále shrnuji jejich převzaté modifikace pro mikropanelová data. Pro shlukování mikropanelových dat je vhodné modifikovat z hierarchického shlukování aglomerativní shlukování a metodu *C*-průměrů z metod rozkladu. Z novějších přístupů se pro shlukování časových trajektorií používá modelový přístup. Tento přístup je však vhodné používat pro makropanelová data s velkým počtem měření v čase. Výhodou metody *C*-průměrů oproti modelovému přístupu je to, že nepředpokládá normalitu nebo jiné parametrické předpoklady o shlucích (třídách), což lze kvitovat v situaci, kdy nemáme apriorní informaci o rozdělení. Metody rozkladu navíc nepředpokládají žádné rozdělení individuálních trajektorií a nezávisí na časové škále. Z těchto důvodů modelový přístup neuvažuji ve zbytku této práce, která je zaměřená na analýzu mikropanelových dat.

Hierarchické shlukování

Hierarchické shlukování seskupuje objekty (trajektorie) do hierarchického stromu shluků. Rozlišují se dva přístupy aglomerativní a divizivní na základě strategie směru shlukování buď „zdola nahoru“, nebo „shora dolů“. Populárnější aglomerativní přístup jsem vybral pro shlukování trajektorií a aplikoval ho pro shlukování odvozených charakteristik dynamiky. V aglomerativním přístupu je na začátku každý objekt zařazen do samostatného shluku. Tyto atomické shluky se postupně spojují do větších a větších shluků, dokud nejsou všechny objekty v jednom shluku, nebo není splněna ukončovací podmínka. Shluky jsou spojovány na základě aglomera-

tivních algoritmů, kterými jsou například metoda průměrné vazby pro mezishlukové vzdálenosti shlukování, metoda průměrné vazby pro vnitroshlukové vzdálenosti, metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, centroidní metoda a Wardova metoda. Algoritmy blíže popisují například Řezanková a kol. (2009). Z těchto metod jsem vybral pro shlukování trajektorií (v kapitole 7) Wardovu metodu a metodu nejvzdálenějšího souseda. Obě metody byly počítány pomocí funkce *hclust()* v systému R. Důvodem jejich volby jsou vlastnosti popsány níže:

- Wardova metoda – kritériem pro spojování shluků je přírůstek celkového vnitroskupinového součtu čtverců odchylek pozorování od průměru shluku. Protože provádíme spojení zajišťující minimalizaci kritéria, má Wardova metoda tendenci odstraňovat malé shluky, tedy tvořit shluky zhruba shodné velikosti, což je často vítaná vlastnost, viz Hebák a kol. (2007).
- Metoda nejvzdálenějšího souseda (úplné vazby) – definuje vzdálenost mezi shluky jako maximální vzdálenost mezi jejich členy. V každém kroku se spojují dva nejbližší shluky. Výhodou metody je tvorba vyrovnaných, tj. vybalancovaných z hlediska počtu zařazených objektů, shluků a je méně citlivá na šum. Limitací metody je tendence rušit velké shluky a všechny shluky mají tendenci být stejně velké. Malé shluky jsou tedy spojovány s většími.

Vhodnost volby těchto metod pro shlukování mikropanelových dat potvrzují Ferreira a Hitchcock (2009). Srovnávají více metod pomocí simulace a na reálných funkčních datech (trajektoriích). Wardova metoda a metoda nejvzdálenějšího souseda v této studii vychází nejlépe.

Pro vyjádření vztahu míry podobnosti se používají u hierarchického shlukování míry vzdálenosti. Mezi nejznámější patří euklidovská, vážená euklidovská, čtvercová euklidovská, manhattanská, Čebyšeova, Minkowského, Lanceh-Williamsova. Výpočetní vzorce výše uvedených a obecně známých metod shlukování a měr vzdáleností uvádí Řezanková a kol. (2009).

Pro testování shlukovací schopnosti sktrukturálních charakteristik, které popisují dynamiku v čase, jsem vybral z konvenčních metod základní euklidovskou vzdálenost, Mahalanobisovu vzdálenost a míru podobnosti založenou na Pearsonovu korelačním koeficientu. Euklidovská vzdálenost je standardní míra, s kterou porovnávám další metriky. Míru založenou na korelačním koeficientu jsem vybral proto, že detekuje lépe rozdíly ve tvarech trajektorií než jiné míry podobnosti dle Grabustse (2011). Jako lepší než euklidovskou vzdálenost ve smyslu shlukovacích schopností trajektorií hodnotí tuto míru například Amit a kol. (2009). U Mahalanobisovy

vzdálenosti se využívá inverzní kovarianční matice, čímž se odstraní závislost na měřicích jednotkách a potlačí se nevýhoda plynoucí z nadměrného vlivu korelovaných proměnných, viz Hebák a kol. (2007), což je případ měr dynamiky a důvod jejího zařazení do srovnávací simulace v kapitole 7. Euklidovská matice vzdálenosti byla spočítána pomocí funkce *dist()*, Mahalanobisova matice vzdálenosti přes funkci *distance()* z balíčku *ecodist*, viz Goslee a Urban (2007) a míra založená na korelačním koeficientu funkcí *Dist()* z balíčku, viz *amap* Lucas (2015).

Pro shlukování časových řad byly navrženy speciální míry podobnosti/vzdálenosti, jako například:

- míra vzdálenosti pro krátké časové řady (STS) – počítá se jako součet čtvercových diferencí směrnic mezi trajektoriemi, viz Grabusts (2011),
- vzdálenost dynamického zvlnění času (*Dynamic time warping distance*) – je zobecněním klasického algoritmu pro porovnání sekvencí diskrétních a spojitých hodnot,
- Kullbackova-Lieblerova vzdálenost – porovnává pravděpodobnostní rozdělení Markovových řetězců, index nepodobnosti (založený na korelační funkci (*cross-correlation function*) mezi dvěma časovými řadami atd., viz Liao (2005),
- vzdálenost založená na hustotě pravděpodobnosti pro data s chybami, kterou navrhli Kumar a kol. (2002),
- *J* divergence a symetrická Chernoffova informační divergence, již odhaduje Kakizawa a kol. (1998).

Kromě konvenčních metod (euklidovské a Mahalanobisovy vzdálenosti) bude použita v simulaci (kapitola 7) míra podobnosti založená na Pearsonovu korelačním koeficientu d_{cc}:

$$d_{cc}(\mathbf{e}_i, \mathbf{e}_{i'}) = 1 - \frac{\sum_{a=1}^{A} (e_{ia} - \bar{e}_i)(e_{i'a} - \bar{e}_{i'})}{\sqrt{\sum_{a=1}^{A} (e_{ia} - \bar{e}_i)^2 \sum_{a=1}^{A} (e_{i'a} - \bar{e}_{i'})^2}}, \tag{5.1}$$

kde $i \neq i'$. Hodnota a-té shlukovací proměnné pro i-tý objekt, kde $a = 1, 2, \ldots, A$ se zde značí e_{ia}. V kontextu této práce (kapitola 7) jsou objekty (individuální trajektorie) shlukovány na základě A charakteristik popisujících dynamiku vývoje individuální trajektorie. Vektor $\mathbf{e}_i$ obsahuje A hodnot shlukovacích proměnných e_{ia} pro i-tý objekt. Průměr hodnot všech A shlukovacích proměnných pro i-tý objekt se počítá jako $\bar{e}_i = \frac{1}{A}\sum_{a=1}^{A} e_{ia}$.

Metoda *C*-průměrů a fuzzy *C*-průměrů

Metoda *C*-průměrů

Shlukovací metoda *C*-průměrů, viz MacQueen (1967) iterativně minimalizuje účelovou funkci. Pro časový okamžik *t*, *N* objektů a za předpokladu *C* shluků lze model *C*-průměrů (účelovou funkci) zapsat jako

$$\min \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ict} d_{ict}^2, \tag{5.2}$$

kde u_{ict} je stupeň náležitosti *i*-tého objektu do *c*-tého shluku v čase *t* s podmínkami:

$$\sum_{c=1}^{C} u_{ict} = 1, \forall i, t, \tag{5.3}$$

$$\forall u_{ict} : u_{ict} = \begin{cases} 1 & \|\mathbf{x}_{it} - \mathbf{h}_{ct}\| = \arg\min_i \|\mathbf{x}_{it} - \mathbf{h}_{ct}\| \\ 0 & jinde \end{cases}.$$

Nejčastěji se volí euklidovská míra vzdálenosti $d_{ict} = \|\mathbf{x}_{it} - \mathbf{h}_{ct}\|$ mezi *i*-tým vektorem pozorování $\mathbf{x}_{it} = (x_{i1t}, \ldots, x_{ijt}, \ldots, x_{iJt})'$ a *c*-tým centroidem $\mathbf{h}_{ct} = (h_{c1t}, \ldots, h_{cjt}, \ldots, h_{cJt})'$ v čase *t*.

Tento model je aplikován v balíčku *kml* systému R určeném pro shlukování trajektorií. Algoritmus se aplikuje přímo na původní naměřené hodnoty, tj. neprovádí se žádná transformace. V balíčku lze volit konvenční míry vzdálenosti (euklidovská a manhattanská), nebo míry specifické pro panelová data: Frechetova míra vzdálenosti, viz Chouakria a Nagabhushan (2007) nebo vzdálenost dynamického zvlnění času. *Kml* nabízí také několik technik pro vypořádání se s chybějícími hodnotami: lineární interpolaci, LOCF (*Last Occurance Carried Forward*) – doplní se předchozí známá hodnota, FOCB (*First Occurance Carried Backward*) – doplní se první následující známá hodnota. Kromě těchto technik se k doplnění chybějících hodnot využívá například metoda B-splines, viz Ferreira a Hitchcock (2009). Více o *kml* ve článku Genoliniho a Falissarda (2011) a online Genoliniho manuálu (2016).

Klasickou metodu *C*-průměrů lze v systému R aplikovat na prostorová data pomocí funkce *kmeans()*. Standardně je nastaven algoritmus popsaný ve studii Hartigana a Wonga (1979). Tuto funkci a algoritmus jsem použil pro shlukování trajektorií v kapitole 7. Pomocí argumentu *algorithm* lze vybrat z dalších algoritmů: Lloydův, Forgyho a MacQueenův.

Metoda fuzzy *C*-průměrů (Fuzzy *C*-Means, **FCM)**

Fuzzy metoda *C*-průměrů, viz Bezdek (1981) iterativně minimalizuje účelovou funkci. Pro časový okamžik *t*, *N* objektů a za předpokladu *C* shluků lze model *C*-průměrů (účelovou funkci) zapsat jako

$$\min \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ict}^m d_{ict}^2, \tag{5.4}$$

kde pro u_{ict} platí $\sum_{c=1}^{C} u_{ict} = 1, \forall i,t,\ u_{ict} \geq 0, \forall i,c,t$ a $m > 1$ je konstanta, která vyjadřuje stupeň překrývání (fuzziness) získaných shluků. Pro dané t je iterativním řešením (5.4)

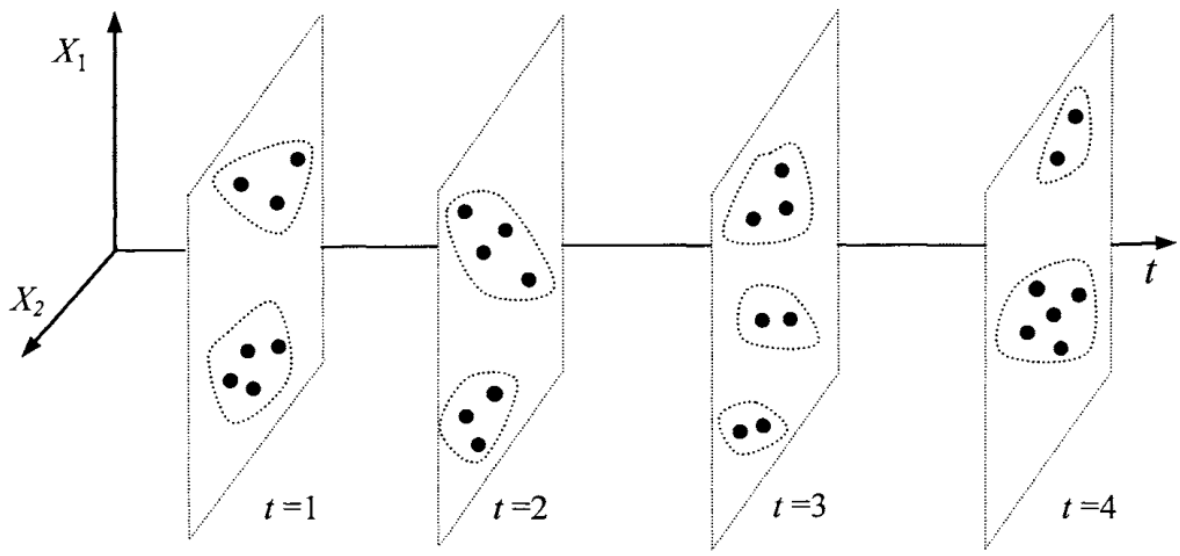
$$u_{ict} = \frac{1}{\sum_{c'=1}^{C} \left(\frac{d_{ict}}{d_{ic't}} \right)^{\frac{2}{m-1}}}, \tag{5.5}$$

$$\mathbf{h}_{ct} = \frac{\sum_{i=1}^{N} u_{ict}^{m} \mathbf{x}_{it}}{\sum_{i=1}^{N} u_{ict}^{m}}. \tag{5.6}$$

Modely fuzzy C-průměrů modifikované pro shlukování trajektorií navrhuje D'Urso (2004), který rozlišuje dva přístupy ke shlukování longitudinálních dat: prostorový (průřezový) a longitudinální.

Prostorový přístup

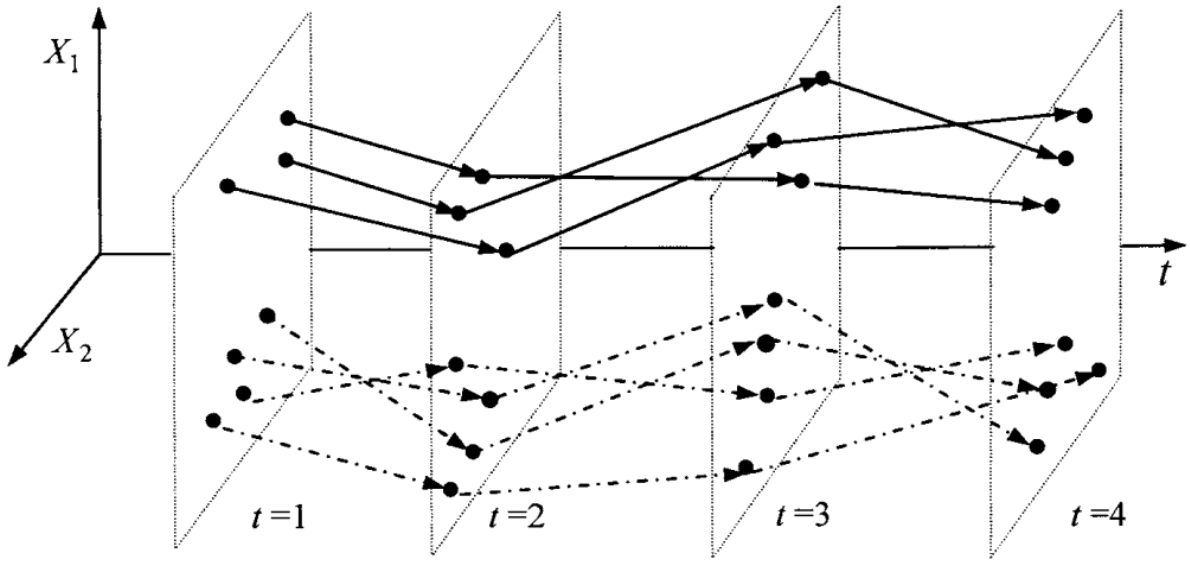
Klasifikace se provádí na základě modelu 5.4 pro každý t-tý časový okamžik zvlášť. Tento přístup je zajímavý v případech, kdy je kladen důraz na momentální (*instantaneous*) stav dat. Výsledných T klasifikací (skupin shluků) se vzájemně porovnává. Studuje se chování a vývoj T klasifikátorů. Zásadní problém prostorového přístupu je charakterizovat a vzájemně porovnat shluky získané z T okamžitých klasifikací. Tento problém se v literatuře vyskytuje pod pojmem *cluster matching problem*, viz Sato a kol. (1997) a Bezdek a kol. (1999). Obrázek 5.1 ilustruje prostorový přístup pro 7 objektů, 2 shlukovací proměnné (X_1, X_2) a 4 časové okamžiky.



Obrázek 5.1 Ilustrace prostorového shlukování v prostoru $\mathbf{R}^{J+1}$ pro N = 7, A = 2 a T = 4.
zdroj: D'Urso (2004).

Longitudinální přístup

Shlukují se časové trajektorie objektů na základě okamžikových a vývojových charakteristik trajektorií. Výsledkem je jedna klasifikace trajektorií. Obrázek 5.2 ilustruje longitudinální přístup pro 7 objektů, 2 shlukovací proměnné a 4 časové okamžiky.



Obrázek 5.2 Ilustrace longitudinálního shlukování v prostoru $\mathbf{R}^{J+1}$ pro $N = 7$, $A = 2$ a $T = 4$. *zdroj: D'Urso (2004).*

D'Urso (2004) navrhuje několik dynamických modelů fuzzy *C*-průměrů a kategorizuje je do 2 skupin: jednokrokový a dvoukrokový přístup. V jednokrokovém přístupu jsou trajektorie shlukovány přímo na základě specifických měr nepodobnosti. Specifickou mírou vzdálenosti je vážená euklidovská vzdálenost ${}_w d_{ict} = w_t \|\mathbf{x}_{it} - \mathbf{h}_{ct}\|$ a okamžitá (poziční) míra nepodobnosti d_{ic}^2 je navržena jako

$$d_{ic}^2 = \sum_{t=1}^{T} {}_w d_{ict}^2 = \sum_{t=1}^{T} \left(w_t \|\mathbf{x}_{it} - \mathbf{h}_{ct}\| \right)^2, \tag{5.7}$$

kde váhy w_t jsou dány pro každé t a platí pro ně $w_t = \sum_{t=1}^{T} w_t = 1,\ w_t \geq 0.$ Váhy mohou být stanoveny fixně na základě apriorní informace, nebo zahrnuty jako další odhadovaný parametr třírozměrného rozšíření modelu fuzzy *C*-průměrů, viz Bezděk (1981):

$$\min : \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic}^m d_{ic}^2 = \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic}^m \sum_{t=1}^{T} \left(w_t \|\mathbf{x}_{it} - \mathbf{h}_{ct}\| \right)^2 \tag{5.8}$$

s omezeními $\sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0$ a $\sum_{t=1}^{T} w_t = 1,\ w_t \geq 0.$ Tento model D'Urso (2004) nazývá *okamžitý (poziční) fuzzy model C-průměrů*. Řešením minimalizačního problému s omezením (5.8) je

$$u_{ic} = \frac{1}{\sum_{\acute{c}=1}^{C}\left[\frac{\sum_{t=1}^{T}\left(w_t d_{ict}\right)^2}{\sum_{t-1}^{T}\left(w_t d_{i\acute{c}\acute{c}}\right)^2}\right]^{\frac{1}{m-1}}}, w_t = \frac{1}{\sum_{t'=1}^{T}\left[\frac{\sum_{i=1}^{N}\sum_{c=1}^{C}u_{ic}^m d_{ict}^2}{\sum_{i=1}^{N}\sum_{c=1}^{C}u_{ic}^m d_{ict'}^2}\right]}, \mathbf{h}_{ct} = \frac{\sum_{i=1}^{N}u_{ic}^m \mathbf{x}_{it}}{\sum_{i=1}^{N}u_{ic}^m}. \tag{5.9}$$

Iterativní algoritmus odhadu modelu 5.8 lze shrnout do posloupnosti 4 kroků:

1. pro fixní m a C je odhadnuta počáteční matice příslušnosti ke shlukům $\mathbf{U}^{(0)}$ o rozměrech $I \times C$,
2. pro $\mathbf{U}^{(0)}$ jsou spočteny počáteční centriody $\mathbf{h}_{ct}^{(0)}$ a váhy $w_t^{(0)}$,
3. je spočtena nová matice $\mathbf{U}$ a procedura je iterována,
4. porovnají se matice $\mathbf{U}^{(p)}$ a $\mathbf{U}^{(p+1)}$ pomocí maticové normy $\left\|\mathbf{U}^{(p+1)} - \mathbf{U}^{(p)}\right\| < \varepsilon$, ε je libovolně malé kladné číslo a p značí p-tou iteraci. V případě platnosti nerovnice je algoritmus ukončen, jinak se vrací ke kroku 2.

Okamžitý model nebere v potaz vývojové charakteristiky trajektorií. Za účelem využití informace o vývoji trajektorií jsou navrženy dvě míry nepodobnosti: míra rychlosti (*velocity-dissimilarity*) a míra zrychlení (*acceleration-dissimilarity*).

Míra rychlosti ${}_v d_{ic}^2$ měří pro i-tý objekt euklidovskou vzdálenost 1. diferencí dvou sousedních $[t-1, t]$ vektorů naměřených hodnot a centroidů:

$$\begin{aligned} {}_v d_{ic}^2 &= \sum_{t=2}^{T} {}_{wv} d_{ict}^2 = \sum_{t=2}^{T}\left({}_v w_t \left\|\mathbf{v}_{it} - \mathbf{v}_{ct}\right\|\right)^2, \\ {}_v d_{ict} &= \left\|\mathbf{v}_{it} - \mathbf{v}_{ct}\right\| = \left\|\left(\mathbf{x}_{it} - \mathbf{x}_{it-1}\right) - \left(\mathbf{h}_{ct} - \mathbf{h}_{ct-1}\right)\right\|. \end{aligned} \tag{5.10}$$

Vektor 1. diferencí $\mathbf{v}_{it} = (\mathbf{x}_{i2} - \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it} - \mathbf{x}_{it\text{-}1}, \ldots, \mathbf{x}_{iT} - \mathbf{x}_{iT\text{-}1})'$ obsahuje $T-1$ párů 1. diferencí naměřených hodnot v časovém intervalu $[t-1, t]$. Prvky $\mathbf{v}_{it}$ vyjadřují sklon mezi dvěma sousedními body.

Míra zrychlení ${}_a d_{ic}^2$ měří pro i-tý objekt euklidovskou vzdálenost mezi vektorem 2. diferencí dvou sousedních $[t-2, t]$ vektorů naměřených hodnot a centroidů:

$$\begin{aligned} {}_a d_{ic}^2 &= \sum_{t=3}^{T} {}_{wa} d_{ict}^2 = \sum_{t=3}^{T}\left({}_a w_t \left\|\mathbf{a}_{it} - \mathbf{a}_{ct}\right\|\right)^2, \\ {}_a d_{ict} &= \left\|\mathbf{a}_{it} - \mathbf{a}_{ct}\right\| = \frac{1}{2}\left\|\left(\mathbf{v}_{it} - \mathbf{v}_{it-1}\right) - \left(\mathbf{v}_{ct} - \mathbf{v}_{ct-1}\right)\right\|. \end{aligned} \tag{5.11}$$

Vektor 2. diferencí $\mathbf{a}_{it}$ obsahuje následující prvky $\mathbf{a}_{it} = (\mathbf{v}_{i2} - \mathbf{v}_{i1}, \ldots, \mathbf{v}_{it} - \mathbf{v}_{it\text{-}1}, \ldots, \mathbf{v}_{iT} - \mathbf{v}_{iT\text{-}1})'$.

D'Urso (2004) navrhuje dva dynamické modely, které vycházejí z *okamžitého (pozičního) fuzzy modelu C-průměrů* (5.8). Pro konstrukci *modelu rychlosti fuzzy C-průměrů* se ve vzorci 5.8 nahrazuje ${}_v d_{ic}^2$ za d_{ic}^2. Analogicky v *modelu zrychlení fuzzy C-průměrů* se ve vzorci 5.8 nahrazuje ${}_a d_{ic}^2$ za d_{ic}^2.

Popsané modely lze odhadovat zvlášť nebo simultánně pomocí *simultánního modelu fuzzy C-průměrů*

$$\min : \sum_{i=1}^{N} \sum_{c=1}^{C} u_{ic}^{m} \sum_{i=1}^{N} \sum_{t} \left(w_{st} d_{icst} \right)^2, \tag{5.12}$$

kde w_{st} indikuje váhy v čase t pro vzdálenosti d_{icst} a kde s-tá vzdálenost, s = 1, 2, 3, je definována následovně:

$$\begin{aligned} &s=1 \;\; d_{ic1t} = \left\| \mathbf{x}_{it} - \mathbf{h}_{ct} \right\| \;\; i=1,2,...,N; \;\; c=1,2,...,C; \;\; t=1,2,...,T, \\ &s=2 \;\; d_{ic2t} = \left\| \mathbf{v}_{it} - \mathbf{v}_{ct} \right\| \;\; i=1,2,...,N; \;\; c=1,2,...,C; \;\; t=2,3,...,T, \\ &s=3 \;\; d_{ic3t} = \left\| \mathbf{a}_{it} - \mathbf{a}_{ct} \right\| \;\; i=1,2,...,N; \;\; c=1,2,...,C; \;\; t=3,4,...,T. \end{aligned} \tag{5.13}$$

V dvoukrokovém přístupu se v prvním kroku počítá N matic nepodobnosti. Počítá se vzdálenost mezi každým párem J proměnných v T časových okamžicích pro každý i-tý objekt. Ve druhém kroku se shlukuje N matic nepodobnosti konvenční shlukovací metodou (fuzzy C-průměrů).

Shlukovací schopnost metody C-průměrů a fuzzy C-průměrů jsme ověřovali v článku Sobíška a kol. (2011). Tyto metody jsme aplikovali na zákaznická data pojišťovny působící na českém trhu. Cílem analýzy byla marketingová segmentace pojištěných vozidel. Kvalita shlukování vybraných metod byla porovnána na základě dvou kritérií. Jako první kritérium jsme vybrali často volenou chybu kvantování, což je průměrná vzdálenost od každého datového vektoru. Největší chybu vykázala metoda SOM. S nejmenší chybou bylo vyhodnoceno shlukování metodou C-průměrů, druhá nejlépe hodnocená metoda je FCM a třetí je NG. Dále jsme metody hodnotili na základě interpretace nalezených shluků. Shluky identifikované pomocí metody NG odpovídají nejlépe rozdělení reálných dat. Vytvořené shluky jsou navzájem heterogenní a mají faktický význam.

6. Návrh a ukázky využití shlukování individuálních trajektorií

Existují různé důvody pro shlukování objektů dle tvaru dynamického vývoje sledované proměnné. Tato kapitola popisuje možnosti využití shlukování objektů založené na tvaru trajektorií. Kapitola 5 shrnuje problémy iterativního odhadu efektů (pevných a náhodných) LMM a GLMM a jsou navrženy dva postupy řešení těchto problémů s využitím shlukové analýzy trajektorií. Jedna možnost je přidat náležitost do shluku jako další vysvětlující proměnnou do LMM/GLMM a dále provést standardně odhady efektů ML/REML. V kapitole 8 je tento návrh testován pomocí simulace. V oddílu 6.1 je navržen druhý postup, ve kterém se kovarianční a korelační struktura náhodných chyb neodhaduje iterativně, ale vyjadřuje se pomocí strukturálních charakteristik trajektorií popisujících intraindividuální dynamiku. Navržený dvoukrokový přístup je pojmenován jako *blokový prostorový regresní model* (*spatial-blocks regression model*). Využití tohoto dvoukrokového přístupu dovoluje odhadovat závislost normálně rozdělené vysvětlované proměnné na vysvětlujících proměnných **X** metodou OLS.Dále lze výsledek shlukování trajektorií použít k identifikaci odlehlých trajektorií nebo jejich shluků. Odlehlé trajektorie lze z prediktivní části analýzy vyloučit nebo jim přiřadit menší váhu. Této možnosti využití shlukování se věnuje oddíl 6.2. Dále je v tomto oddíle představen index datové kvality (*Data Quality Index, DQI*) jako nástroj pro identifikaci odlehlých trajektorií. DQI je založený na míře vzdálenosti. DQI jsem navrhl s kolektivem autorů, viz Kráľ a kol. (2014a).

Kapitolu 6 uzavírá oddíl 6.3, ve kterém jsou uvedeny příklady toho, že shlukování trajektorií nemusí být vždy exploratorní první krok předcházející regresní analýze dat, ale může vystupovat rovněž jako primární cíl analýzy.

6.1 Regresní model prostorových bloků

Odhady regresních parametrů LMM a GLMM jsou získány pomocí vybrané nelineární optimalizační metody (REML), viz oddíly 2.2 a 3.2. Metody odhadu parametrů stejného typu modelu implementovány v různých balíčcích jazyka R se mezi sebou mohou výrazně lišit, viz kapitola 4, a tedy i odhady parametrů a jejich SE se běžně neshodují. Na diverzitu odhadů upozorňuje například Fox (2002). Volba vhodné metody není dle Nashe (2014) vždy jednoznačná. Intervalové odhady (založené na teorii normálního rozdělení) variančních komponent je obecně (ve všech softwarových produktech) problematické odhadnout a nejsou často věrohodné. V praxi stále častěji kombinují výzkumníci frekventistický a bayesovský přístup za účelem získat intervalové odhady parametrů GLMM, viz oddíl 3.3. Vedou se však minimálně na akademické úrovni diskuze o tom, zda lze kombinovat dvě paradigmata. Pro velké datové soubory u

komplexních modelů (tedy při zahrnutí většího počtu náhodných efektů pro LMM a GLMM, nebo většího počtu pevných efektů do GLMM) hrozí, že iterativní odhad parametrů může být zkreslený, viz Bates a kol. (2015b) nebo bude model přeparametrizovaný. Správná volba kovarianční struktury náhodných chyb je kriticky důležitá a vede ke zrychlení výpočetního času. V případě nejednoznačnosti struktury se nabízí použít obecnou strukturu, která však pro velké datové soubory nenalezne optimální řešení.

Další nevýhodou simultánního jednokrokového modelování (LMM, GLMM) je výpočetní čas. V případě analýzy velkých dat (od tisíců záznamů) kdy navíc odhaduji obecnou kovarianční matici náhodných chyb, může výpočet trvat hodiny/dny, nebo vůbec nebude z důvodu výpočetní náročnosti spočítán (ani v současné době s výkonnými počítači). Správné zařazení proměnné do modelu (jako X_j nebo Z_j) je často diskutabilní. Problém vychází z nejednoznačnosti definice pevných a náhodných efektů, viz kapitola 2.2.

Lze shrnout, že výzkumník musí při modelování panelových dat správně:

- vybrat vhodný model,
- rozhodnout o formě zařazení proměnných do modelu,
- vybrat metodu odhadu efektů,
- vybrat kovarianční (korelační) strukturu náhodných chyb. Navíc v případě, že je uvažován více než jeden náhodný efekt, lze volit i kovarianční strukturu náhodných efektů.

Výše uvedené problémy jednokrokového modelování mne inspirovaly k návrhu vlastního alternativního přístupu k modelování mikropanelových dat. Cílem je navrhnout alternativní přístup, který zohledňuje kovarianční a korelační strukturu náhodných chyb (bez zadávání typu kovarianční struktury) a regresní parametry neodhaduje iterativně, tj. využívá OLS nebo ML. Podstatu navrhovaného dvoukrokového přístupu lze nazvat jako *regresní model prostorových bloků*.

V *prvním kroku* se spočítají vývojové charakteristiky sledované proměnné pro každou i-tou individuální trajektorii navrženou v oddílu 7.1. Na základě těchto měr dynamiky jsou objekty klasifikovány vhodnou shlukovou metodou do C homogenních shluků. Uvnitř shluků se nacházejí objekty s homogenní kovarianční a korelační strukturou. Mezi shluky je významná heterogenita tvaru trajektorií. Soubor měření se následně rozdělí do $C \cdot T$ dílčích podsouborů, které lze nazvat jako *prostorové bloky*. Blok_{ct} obsahuje měření provedená v čase t náležející objektům klasifikovaných do c-tého shluku, kde $c = 1, 2, \ldots, C$. První krok lze chápat jako transformaci panelových dat na prostorová data.

Ve *druhém kroku* se odhaduje efekt vysvětlujících proměnných **X** na střední hodnotu vysvětlované proměnné metodou KLM dle vzorce 2.1 pro každý prostorový blok. Výsledkem je $C{\cdot}T$ vektorů $\widehat{\boldsymbol{\beta}_{ct}} = \left(\widehat{\beta_{ct0}}, \widehat{\beta_{ct1}}, \dots, \widehat{\beta_{ctK}}\right)$, kdy $\widehat{\boldsymbol{\beta}_{ct}}$ obsahuje $K + 1$ odhadnutých regresních parametrů pro blok$_{ct}$. Pro n_{ct} náhodných chyb $\boldsymbol{\varepsilon}_{ct} = (\varepsilon_{ct1}, \dots \varepsilon_{cti}, \dots, \varepsilon_{ctN})$ uvnitř bloku$_{ct}$ lze předpokládat stejné vlastnosti jako u náhodné složky $\boldsymbol{\varepsilon}$ v KLM jmenovitě normalitu, homoskedasticitu a nezávislost náhodné složky. Počet měření v bloku$_{ct}$ značí n_{ct}.

Pro každý parametr β_j získáme až $C{\cdot}T$ odhadů a tedy i představu o jeho empirické distribuci. Méně než $C{\cdot}T$ odhadů získáme v případě, že v alespoň jednom bloku máme nedostatečný počet měření nebo daný parametr v některých blocích neodhadujeme. Je možno porovnávat odhady pevných efektů jednotlivých bloků s reprezentativním odhadem (průměr nebo medián) a identifikovat extrémní odhady. Opakovaný výskyt extrémních odhadů pro bloky ve stejném čase *t* označují časový okamžik, ve kterém je efekt vysvětlující proměnné odlišný oproti jiným obdobím. Rovněž opakovaný výskyt extrémních odhadů pro bloky pocházející z *c*-tého shluku poskytuje informaci o odlišném efektu pro tuto skupinu objektů z tohoto shluku. Tuto informaci z LMM nezískáme. Další výhodou odhadu pro každý blok je možnost modelovat různé typy funkcí (polynomy *n*-tého stupně) pro každý blok. Toto je výhoda i ve srovnání s PAOM (oddíl 2.1), kdy nelze odhadovat individuální modely s větším počtem parametrů v mikropanelech s malým počtem opakovaných měření (v řádu jednotek).

Celkovou variabilitu S_y lze rozdělit na vnitroblokovou $S_{y,B}$ a meziblokovou $S_{y,W}$ následovně:

$$
\begin{aligned}
S_y &= \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \bar{y}\right)^2 = S_{y,B} + S_{y,W}; \\
S_{y,B} &= \sum_{c=1}^{C}\sum_{t=1}^{T}\left(\bar{y}_{ct} - \bar{y}\right)^2 n_{ct}, \\
S_{y,W} &= \sum_{c=1}^{C}\sum_{t=1}^{T}\sum_{i=1}^{n_{ct}}\left(y_{ict} - \bar{y}_{ct}\right)^2.
\end{aligned}
\tag{6.1}
$$

Jedná se o analogický rozklad čtverců jako u klasické jednofaktorové metody ANOVA. Tento rozklad čtverců odpovídá na stejnou otázku jako odhady smíšených efektů (např. pro časovou proměnnou *t*) v LMM nebo GLMM, kdy analytik porovnává odhad variability smíšeného efektu (například variabilita individuálních sklonů nebo průsečíků) s odhadem variability náhodných chyb. V tomto případě se porovnává $S_{y,B}$ reprezentující variabilitu tvaru trajektorií a $S_{y,W}$ vyjadřující individuální variabilitu. Vnitroskupinový korelační koeficient ICC (vzorec 2.13) lze pak zapsat jako ICC=$S_{y,B}/S_y$.

Na druhou stranu nevýhodu navrženého přístupu spatřuji v tom, že navržené charakteristiky počítané v prvním kroku nemusí pro některé složité tvary trajektorií popsat dostatečně jejich tvar. Dále hrozí riziko, že nebyl určen správný počet shluků a trajektorie uvnitř shluku nejsou homogenní.

Blokový přístup lze aplikovat i pro vysvětlovanou proměnnou pocházející z exponenciálního rozdělení namísto GLMM. V tomto případě se v druhém kroku využije GLM a ML nebo REML odhad.

6.2 Identifikace odlehlých trajektorií

Charakteristiky popisující dynamiku individuální trajektorie definované v kapitole 7, případně výsledek shlukování na základě těchto charakteristik, lze použít k identifikaci odlehlých trajektorií nebo shluků. Výskyt odlehlých trajektorií může významně zkreslit odhad parametrů a jejich středních chyb. Míru zkreslení ukazují Williams a kol. (2015) na reálných a simulovaných datech při použití nelineárního smíšeného modelu.

Odhad regresních modelů komplikuje mj. výskyt uměle vytvořených odlehlých trajektorií. Ty se objevují například v situaci, ve které je hlavní úlohou modelovat vývoj anualizované relativní změny vysvětlované proměnné (například objem mozku) v čase, kdy relativní změny nejsou měřeny primárně, ale jsou odvozeny ze dvou sousedních naměřených hodnot. Anualizovaná relativní změna Y (APC, *Annualized Percentage Change*) pro i-tý objekt v čase t se spočítá $\mathrm{APC}_{\mathrm{it}} = \left[\left(y_{it} - y_{i(t-1)}\right) / y_{i(t-1)}\right] / \left(d_{it}/365\right)$, kde d_{it} značí počet dní mezi t a (t-1) pro objekt i. Při anualizaci relativních změn počítaných ze sousedních absolutních hodnot od sebe vzdálených méně než jeden rok, tj. $d_{it} < 365$, dochází k „inflačnímu (multiplikačnímu) efektu“. Skutečná anualizovaná relativní změna je uměle nadhodnocena. V případě častého výskytu extrémních hodnot jsou regresní odhady nepřesné. Takový návrh studie není ideální. Avšak například v epidemiologické praxi často není realizovatelný i lepší návrh studie.

Odlehlé individuální trajektorie či jejich shluky lze zkoumat:

- Nezávisle na většině trajektorií a vyřadit je z regresního modelování pomocí LMM nebo blokového přístupu (oddíl 6.1) v případě velkého výběru při zachování reprezentativnosti vzorku.
- Analyzovat je simultánně s ostatními trajektoriemi pomocí vybraného regresního přístupu v případě studie s minimálním rozsahem výběru, kdy každý vyřazený objekt má významný negativní dopad na snížení statistické síly odhadů. Těmto objektům se přiřadí

menší váha při odhadování regresních parametrů. Rozdělení vhodné spojité charakteristiky dynamiky lze využít pro stanovení struktury vah individuálních trajektorií. V případě normálně rozdělené charakteristiky lze využít například princip 6σ, kdy největší váhu má střední trajektorie. Váha trajektorií se snižuje, čím jsou vzdálenější od střední trajektorie. Trajektorie vzdálené více než $\pm 3\sigma$ od střední hodnoty lze označit jako odlehlé trajektorie, kterým se určí minimální váha, nebo se z další analýzy vyřadí.

Datovou kvalitu, která může pomoci při identifikaci extrémních pozorování, zkoumáme v práci (Kráľ a kol., 2014a). Navrhujeme zde popis datové kvality sledovaných proměnných a individuálních případů. Navržený popis založený na míře vzdálenosti umožnuje definovat datovou kvalitu individuálních trajektorií. Datová kvalita je měřena jako vzdálenost objektů od ideálního objektu. Metrika měřící datovou kvalitu je pojmenována jako index datové kvality (*DQI*). DQI lze využít pro stanovení váhy individuální trajektorie. Čím je trajektorie vzdálenější od ideálního vývoje (objektu), tím je hodnota DQI této trajektorie nižší a získává menší váhu v dalším modelování. DQI nabývá hodnot z intervalu <0; 1>. Ideální trajektorie má DQI roven 1. Naopak DQI se rovná 0 pro trajektorii s nulovou datovou kvalitou. Ideální vývoj má trajektorie s předpokládaným tvarem. V případě neexistence teoretického ideálního tvaru lze volit ideální případ empiricky, například vzít průměrný vývoj. Navržený popis datové kvality je matematicky ekvivalentní metodě *hesitant fuzzy sets*, viz Zeshui a Meimei (2011). Ve článku autorů Kráľa a kol. (2014a) jsme zkoumali dvě míry vzdálenosti: normalizovanou Hammingovu vzdálenost a normalizovanou euklidovskou vzdálenost.

Možná aplikace DQI je ve výše zmiňovaném článku ukázána na reálných panelových datech z pojišťovnictví. Zabývali jsme se prediktivním modelováním stornovosti životních smluv ze strany klienta. Ve studii Stachové a Sobíška (2013) je předčasné ukončení pojistné smlouvy modelováno pomocí podmíněných stromových struktur. Pro pojišťovny je strategicky důležitá znalost prediktorů a identifikace investičních a kapitálových životních smluv do dvou let od uzavření smlouvy, u kterých je vysoká pravděpodobnost předčasného ukončení během jejího budoucího životního cyklu. Z podmíněných stromových struktur byly odhadnuty modely stornovosti v systému R pomocí algoritmu podmíněný náhodný strom, viz Strasser a Weber (1999) a podmíněný náhodný les, viz Horthorn a kol. (2006). Podmíněný strom byl odhadnut funkcí *ctree()* z balíčku *party* a podmíněný les funkcí *cforest()* ze stejného balíčku. Tyto algoritmy jsme porovnali s ostatními často používanými stromovými strukturami klasifikační strom a náhodný les.

Tato reálná data obsahují nezávislé potenciální prediktory a objekty s různou kvalitou dat. Právě různá kvalita uvnitř datového souboru vedla kolektiv autorů, viz Kráľ a kol. (2014a) k návrhu DQI a doporučení zahrnout ho do prediktivního modelování. K citovanému článku je přiložen ukázkový skript pro výpočet DQI v jazyce R.

6.3 Profil shluků jako cíl analýzy

V předchozích oddílech 6.1 a 6.2 je shlukování objektů na základě tvaru trajektorie dílčí, exploratorní úlohou. Výsledky jsou použity jako vstup pro hlavní cíl analýzy: bloková prostorová regresní metoda, hledání extrémních pozorování. Shlukování trajektorií může ovšem vystupovat rovněž jako primární cíl analýzy, například v ekonomii je důležité klasifikovat země dle vývoje ekonomických agregátů. Ve financích je užitečné získat klasifikaci společností kótovaných na burze dle časového vývoje výnosu akcií.

Procháska a kol. (1991) sledují dva roky panelový soubor kuřáků. U každého objektu mají k dispozici pět měření a sledují jeho změny nálad a objektivních pocitů na téma přestat kouřit. Mezi sledovanými proměnnými jsou pokušení kouřit, vnímaná pozitiva a negativa pro skončení s kouřením atd. Cílem je získat reprezentativní profily objektů dle dynamiky změn názorů na ukončení kouřit. Kombinují prostorovou analýzu bazických hodnot a longitudinální analýzu změn. Získali čtrnáct profilů, které doporučují zohlednit při stanovení odvykacích léčebných programů.

Aktuálně se shlukování trajektorií stává primárním cílem některých analýz v oblasti epidemiologie. Koestler a kol. (2014) zkoumají dynamický vývoj longitudinálních genových dat za účelem získání informace o kauzálním vztahu mezi časovým vývojem genových změn jednotlivce a progresí nemoci. Výsledkem shlukování objektů na základě trajektorie vývoje sledované množiny genů jsou profily pacientů, jež mohou vést k získání fenotypů nemoci.

V oblasti ekonomické analýzy kriminality ekonometři modelují vztah mezi kriminalitou a ekonomickou aktivitou regionu (země). Modely popisují například vliv míry a různých aspektů kriminální aktivity na ekonomické ukazatele a zaměstnanost. Lombardo a Falcone (2011) si kladou za cíl prozkoumat vztah mezi mírou kriminality a socio-ekonomickými ukazateli. K dispozici mají 24 indikátorů charakterizujících kriminalitu, demografii, hospodářský výkon a vzdělání ve 103 italských regionech (dle NUTS3) z let 1999–2009. Provádějí shlukovou analýzu tohoto mikropanelu, aby prozkoumali, zda existují shluky podobných regionů z hlediska sledovaných charakteristik. Mikropanel využívají, aby zohlednili prostorově a časově měnící se vztah mezi kriminalitou a ekonomickými indikátory. Nalezli 7 shluků.

7. Charakteristiky dynamiky trajektorií mikropanelových dat

Pro shlukování mikropanelových dat a nestacionárních časových řad, které nelze transformovat na stacionární, jsou použitelné hierarchické metody nebo metody rozkladu. Tyto dva přístupy kombinuji s transformacemi původních hodnot, které reprezentují dynamiku časové řady, za účelem ověřit schopnost shlukovat mikropanelová data. Pro shlukování individuálních trajektorií navrhuji přistupovat dvoukrokově. V prvním kroku se spočítají pro každý objekt *i* hodnoty navržených charakteristik, které transformují původní longitudinální data na méně rozměrná statická data. Ve druhém kroku jsou shlukovány objekty konvenčními prostorovými technikami na základě matice nepodobnosti hodnot shlukovacích proměnných spočítaných v prvním kroku. V oddílu 7.1 definuji vlastní charakteristiky transformace mikropanelových dat na prostorová data, které popisují dynamiku individuálních trajektorií. Pomocí simulační studie je posouzena jejich shlukovací schopnost. V oddílu 7.2 je porovnána shlukovací schopnost charakteristik navržených v oddílu 7.1 a charakteristik aplikovaných u zavedených postupů, které popisuji v oddílu 6.1.

7.1 Návrh nových charakteristik dynamiky a jejich testování

V oddílu 7.1.1 popisuji navržené charakteristiky dynamiky. Dále v oddílu 7.1.2 porovnávám pomocí simulační studie shlukovací schopnost různých kombinací těchto derivátů v kombinaci s hierarchickým shlukováním a metodou *C*-průměrů.

7.1.1 Charakteristiky popisující dynamiku individuálních trajektorií

V této části popisuji první krok navrhovaného dvoukrokového přístupu ke shlukování mikropanelových trajektorií. V prvním kroku se počítají hodnoty navržených shlukovacích proměnných. Z těchto hodnot se počítá matice nepodobnosti, na kterou se aplikuje shlukovací algoritmus ve druhém kroku. Shlukovací proměnné jsou v dalším textu označeny jako e_a, kde $a = 1, 2, \ldots, A$. Hodnota *a*-té shlukovací proměnné pro *i*-tý objekt se značí e_{ia}. Vektor $\mathbf{e}_i = (e_{i1}, \ldots, e_{ia}, \ldots, e_{iA})'$ obsahuje *A* hodnot shlukovacích proměnných e_{ia} pro *i*-tý objekt. Jedná se o charakteristiky, které popisují ze strukturálního hlediska dynamiku individuální trajektorie. Mezi navrhované charakteristiky patří:

1. Průměrná trojúhelníková vzdálenost mezi dvěma sousedními měřeními. Pro *i*-tý objekt se tato průměrná hodnota značená $\overline{diff}_i$ spočítá jako průměr všech trojúhelníkových diferencí mezi dvěma časově sousedními body $diff_{it}$ následovně:

$$\overline{diff_i} = \frac{\sum_{t=2}^{T} diff_{it}}{T-1} = \frac{\sum_{t=2}^{T} 1/2\left(\frac{y_{it} - y_{i(t-1)}}{D_{it} - D_{i(t-1)}}\right)}{T-1}, \tag{7.1}$$

kde y_{it} je hodnota sledované proměnné pro i-tý objekt v čase t. D_{it} značí časovou vzdálenost časového okamžiku t od počátku (t = 1) pro i-tý objekt (například v letech) a T je počet pozorování i-tého objektu.

2. Výběrová směrodatná odchylka trojúhelníkových vzdáleností mezi dvěma sousedními měřeními. Pro i-tý objekt se charakteristika variability značená sd($diff_i$) spočítá jako

$$\mathrm{sd}(diff_i) = \frac{\sum_{t=2}^{T}\left(diff_{it} - \overline{diff_i}\right)^2}{T-2}. \tag{7.2}$$

3. Průměrná absolutní trojúhelníková vzdálenost mezi dvěma sousedními měřeními. Pro i-tý objekt se tato průměrná hodnota značená $\overline{|diff_i|}$ spočítá jako

$$\overline{|diff_i|} = \frac{\sum_{t=2}^{T}|diff_{it}|}{T-1}. \tag{7.3}$$

Oproti (7.1) se průměrují absolutní hodnoty trojúhelníkových vzdáleností mezi y_{it} a $y_{i(t\text{-}1)}$.

4. Výběrová směrodatná odchylka absolutních trojúhelníkových vzdáleností mezi dvěma sousedními měřeními. Pro i-tý objekt se směrodatná odchylka značená $\mathrm{sd}(|diff_i|)$ spočítá jako

$$\mathrm{sd}(|diff_i|) = \frac{\sum_{t=2}^{T}\left(|diff_{it}| - \overline{|diff_i|}\right)^2}{T-2}. \tag{7.4}$$

5. Průměrný koeficient růstu $\overline{k_i}$. Měří průměrnou relativní změnu sledované proměnné Y a počítá se pro i-tý objekt jako

$$\overline{k_i} = \sqrt[n_i-1]{\frac{y_{iT}}{y_{i1}}} - 1 = \sqrt[n_i-1]{k_{i2}k_{i3}\ldots k_{it}\ldots k_{iT}} - 1, \tag{7.5}$$

kde $k_{it} = y_{it}/y_{i(t\text{-}1)}$ pro t = 2, …, T značí koeficient růstu mezi t a t – 1 hodnotou Y.

6. Relativní počet kladných změn k počtu negativních změn, značeno %pos (nebo rel.pos v přiloženém skriptu jazyka R)

$$\%pos = \frac{\mathrm{pocet}\, k_{it} \geq 1}{\mathrm{pocet}\, k_{it} < 1},\ t = 2,\ \ldots,\ T; \tag{7.6}$$

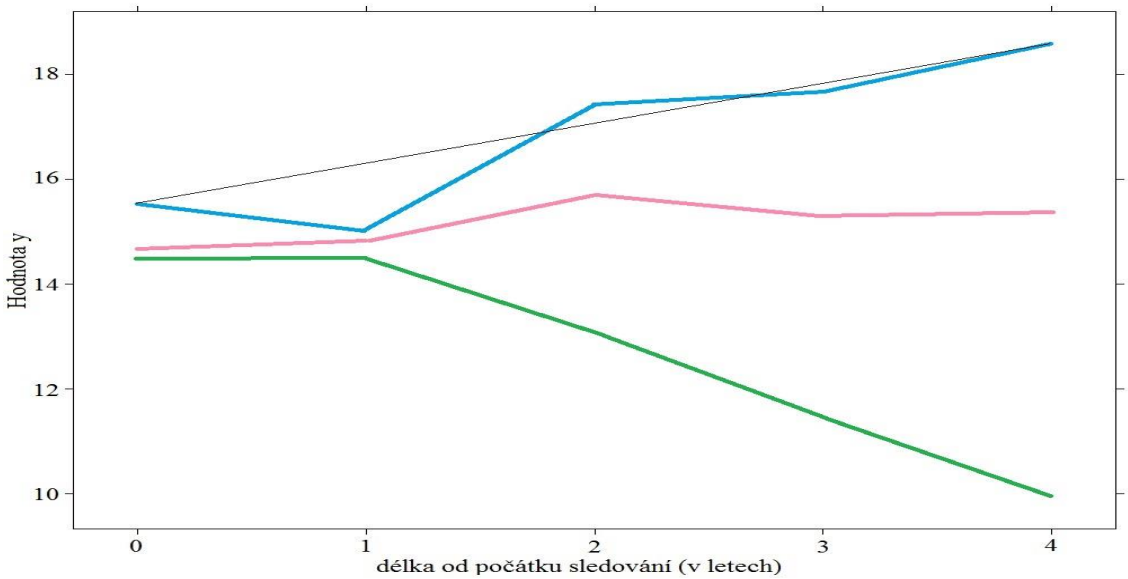
v případě, že je počet negativních změn (jmenovatel) nulový, nulová hodnota ve jmenovateli se nahrazuje hodnotou 0,1.

7. Hodnota maximálního úhlu (v radiánech) mezi spojnicí okrajových měření [$t = 1, y_{i1}$] a [$t = T, y_{iT}$] a spojnicí vnitřního bodu [$t = t, y_{iT}$], kde platí $t > 1$ a $t < T$, a prvního měření [$t = 1, y_{i1}$]. Výpočet úhlu v radiánech pro vnitřní bod t, kde $t > 1$ a $t < T$, i-tého objektu lze spočítat následovně:

$$\sphericalangle T1t = \arccos\left(\frac{[(D_{iT} - D_{i1})\cdot(D_{it} - D_{i1})] + [(y_{iT} - y_{i1})\cdot(y_{it} - y_{i1})]}{\sqrt{(D_{iT} - D_{i1})^2 + (y_{iT} - y_{i1})^2}\cdot\sqrt{(D_{it} - D_{i1})^2 + (y_{it} - y_{i1})^2}}\right), \qquad (7.7)$$

$t = 2, \ldots, T-1$. Z $T-2$ spočtených úhlů $\sphericalangle T1t$ se jako shlukovací proměnná vybere maximální úhel v radiánech. Vybranému maximálnímu úhlu v radiánech se přiřadí záporné znaménko v případě, že směrnice přímky procházející okrajovými měřeními je větší než směrnice přímky procházející vybraným vnitřním t a prvním měřením. Maximální úhel zůstane kladný v opačném případě, tj. směrnice přímky procházející okrajovými měřeními je menší než směrnice přímky procházející t a prvním měřením. Značení této shlukovací proměnné v dalším textu je $\max \sphericalangle$.

Následující příklad ilustruje transformaci původních hodnot Y na shlukovací proměnné. Původní hodnoty jsou zadány v dlouhém formátu a shlukovací proměnné v širokém formátu. V příkladu jsou tři objekty (trajektorie). Každý objekt je měřen pravidelně každý rok po dobu pěti let. Naměřené hodnoty ilustruje obrázek 7.1.



Obrázek 7.1 Ilustrace 3 individuálních trajektorií (objektů). Trajektorie objektu $i = 1$ je znázorněna modře, objekt 2 znázorňuje růžová křivka a objekt 3 zelená.

V tabulce 7.1 jsou kromě naměřených hodnot y_{it} uvedeny hodnoty pomocných výpočtů pro stanovení hodnot shlukovacích proměnných. Mezi pomocné hodnoty patří $diff_{it}$, $|diff_{it}|$, k_{it}, binární proměnná pozitivní změna, která nabývá 1 pro $k_{it} \geq 1$ a 0 pro $k_{it} < 1$. Hodnota „NA“ značí prázdné pole. Sloupec *i* identifikuje objekt. Tabulka v dlouhém formátu (jeden řádek = jedno měření) odpovídá standardnímu zadání longitudinálních dat do systému R pro další typy analýz (například funkce *lme(), lmer()*).

Tabulka 7.1 Spočtené hodnoty tří ukázkových trajektorií

i	D_{it}	t	y_{it}	$y_{i(t-1)}$	$diff_{it}$	$\|diff_{it}\|$	k_{it}	Pozitivní změna
1	0,00	1	15,54	NA	NA	NA	NA	NA
1	1,00	2	15,03	15,54	-0,26	0,26	0,97	0
1	2,00	3	17,42	15,03	1,20	1,20	1,16	1
1	3,00	4	17,67	17,42	0,13	0,13	1,01	1
1	4,00	5	18,59	17,67	0,46	0,46	1,05	1
2	0,00	1	14,67	18,59	NA	NA	NA	NA
2	1,00	2	14,84	14,67	0,08	0,08	1,01	1
2	2,00	3	15,71	14,84	0,44	0,44	1,06	1
2	3,00	4	15,32	15,71	-0,20	0,20	0,98	0
2	4,00	5	15,38	15,32	0,03	0,03	1,00	1
3	0,00	1	14,49	15,38	NA	NA	NA	NA
3	1,00	2	14,49	14,49	0,00	0,00	1,00	1
3	2,00	3	13,09	14,49	-0,70	0,70	0,90	0
3	3,00	4	11,48	13,09	-0,80	0,80	0,88	0
3	4,00	5	9,94	11,48	-0,77	0,77	0,87	0

V tabulce 7.2 jsou spočteny hodnoty shlukovacích proměnných dle vzorců (7.1) až (7.7). Tak například pro první objekt (i = 1)

$$\overline{diff}_i = \frac{(-0{,}26 + 1{,}20 + 0{,}13 + 0{,}46)}{4} = 0{,}38,$$

$$\mathrm{sd}\left(diff_i\right) = \frac{\left(-0{,}26 - 0{,}38\right)^2 + \left(1{,}20 - 0{,}38\right)^2 + \left(0{,}13 - 0{,}38\right)^2 + \left(0{,}46 - 0{,}38\right)^2}{3} = 0{,}62,$$

$$\left|\overline{diff}_i\right| = \frac{(0{,}26 + 1{,}20 + 013 + 0{,}46)}{4} = 0{,}51,$$

$$\mathrm{sd}\left(\left|diff_i\right|\right) = \frac{\left(0{,}26 - 0{,}38\right)^2 + \left(1{,}20 - 0{,}38\right)^2 + \left(0{,}13 - 0{,}38\right)^2 + \left(0{,}46 - 0{,}38\right)^2}{3} = 0{,}50,$$

$$\overline{k_i} = \sqrt[4]{\frac{18{,}59}{15{,}54}} - 1 = 0{,}05,$$

%pos=3/1=3,00,

$$\max \sphericalangle = \arccos\left(\frac{[(D_{i5}-D_{i1})\cdot(D_{i2}-D_{i1})]+[(y_{i5}-y_{i1})\cdot(y_{i2}-y_{i1})]}{\sqrt{(D_{i5}-D_{i1})^2+(y_{i5}-y_{i1})^2}\cdot\sqrt{(D_{i2}-D_{i1})^2+(y_{i2}-y_{i1})^2}} \right) =$$

$$= \arccos\left(\frac{[(4-0)\cdot(1-0)]+[(18{,}59-15{,}54)\cdot(15{,}03-15{,}54)]}{\sqrt{4^2+(18{,}59-15{,}54)^2}\cdot\sqrt{1^2+(15{,}03-15{,}54)^2}} \right) = 1{,}13.$$

Úhel je počítán pro $t = 2$, tedy jde o úhel mezi spojnicí 1. a 5. měření a spojnicí 1. a 2. měření v čase. Výpočet potvrzuje, co je patrné rovněž z obrázku 7.1, že právě $t = 2$ svírá největší úhel se spojnicí 1. a 5. měření (v obrázku vyznačena černou slabou čárou). Rovněž lze z obrázku vyčíst, že spojnice 1. a 2. bodu se napravo od $t = 1$ nachází pod spojnicí 1. a 5. měření. Sklon této spojnice je menší (–0,51) než sklon spojnice 1. a 5. měření (0,76). Z tohoto důvodu je hodnota $\max \sphericalangle$ záporná a rovna –1,13.

Tabulka 7.2 Navržené shlukovací proměnné reprezentující dynamiku 3 ukázkových trajektorií

i	$\overline{diff}_i$	sd($diff_i$)	$\overline{\lvert diff_i \rvert}$	sd($\lvert diff_i \rvert$)	$\overline{k_i}$	%pos	max ∢
1	0,38	0,62	0,51	0,48	0,05	3,00	–1,13
2	0,09	0,26	0,19	0,18	0,01	3,00	0,30
3	–0,57	0,38	0,57	0,38	–0,09	0,33	0,85

Kapitola 6 poukazuje na důležitost shlukování trajektorií. Výsledné přiřazení do shluků je buď primárním cílem analýzy, nebo přípravným prvním krokem pro další analýzu (například regresní). V tomto smyslu lze na shluky nahlížet jako na podskupiny (straty) a provádět regresní modelování uvnitř každé podskupiny (pomocí LMM nebo GLMM). V případě dostatečného počtu měření uvnitř shluku lze každou podskupinu dále rozdělit do T prostorových bloků a v každé podskupině/bloku modelovat regresní vztah mezi sledovanými proměnnými (pomocí LM). V kontextu tohoto potenciálního využití shluků vektor $\mathbf{e}_i$ reprezentuje korelační a kovarianční strukturu náhodných efektů. Vektor vývojových charakteristik lze chápat jako vektor náhodných efektů z LMM, jehož prvky nejsou odhadovány metodou REML, ale jedná se o funkce empirických hodnot $\mathbf{y}_i$.

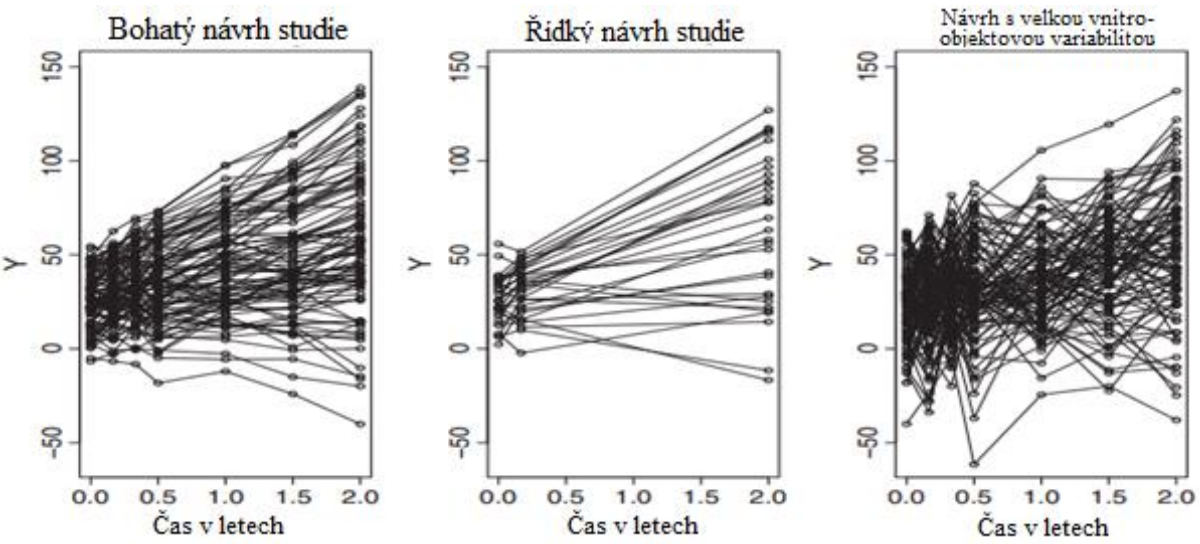
Uvedené původní charakteristiky jsem poprvé prezentoval na konferenci AS2015, viz Sobíšek a Stachová (2015). V tomto příspěvku byla náležitost do shluků dle těchto charakteristik zařazena do návrhu kompozitní definice finančního zdraví podniku.

7.1.2 Porovnání metod založených na práci s původními daty a jejich transformacemi

Pomocí simulační studie posuzuji v tomto oddílu schopnost různých kombinací navržených charakteristik dynamiky dostatečně popsat dynamiku trajektorií. Pokud charakteristiky dobře popisují individuální vývoj v čase, lze je použít ke shlukování trajektorií. Na tyto kombinace implementuji shlukovací algoritmy aglomerativní shlukování a metodu *C*-průměrů. Tyto shlukovací algoritmy jsou porovnány. Shlukovací schopnost za předpokladu známé náležitosti do shluků vyhodnocuji pomocí čtyř různých kritérií, kterými jsou Randův index, modifikovaný Randův index, Jaccardův koeficient a míra správné klasifikace (*correct classification rate, CCR*).

Popis simulačních scénářů

Navrhl jsem osm vybalancovaných simulačních scénářů. Liší se mezi sebou počtem shluků C a předpokládaným rozdělením jednotlivých shluků. Pro každý scénář platí, že je v každém shluku 100 trajektorií ($n_c = 100$ pro každé c) a každý objekt je měřen pětkrát ($T = 5$ pro všechna i) v pravidelných časových intervalech délky 1, tj. $t = 1, 2, 3, 4, 5$. Zvolený počet objektů ve shluku a počet měření T odpovídá velikosti výběrového souboru především v klinických studiích. Naopak návrh studie použité pro zkoumání vývoje Parkinsonovy choroby, viz Thai a kol. (2013), kdy v jedné skupině pacientů (*treatment arm*), tedy v jednom shluku, je 30 objektů a $T = 3$, je považován za řídký (*sparse design*). I řídký návrh je však pro některé úlohy dostačující. Návrh s $n_c = 100$ a $T = 7$ je považován za bohatý na počet měření (*rich design*). Autor se zabývá ještě třetím typem návrhu, kde se vyskytuje velká vnitroobjektová variabilita. Obrázek 7.2 ilustruje tyto příklady návrhů.



Obrázek 7.2 Ukázka tří typických návrhů studie z oblasti medicínské statistiky. (*zdroj: Thai, H.T.,P. a kol. (2013)*)

Některé dále uvedené scénáře (1, 3 a 5) jsou spíše teoretické s minimální vnitroobjektovou variabilitou a výraznou mezishlukovou heterogenitou tvaru trajektorií. Očekávám, že u těchto scénářů by vhodně navržený shlukovací postup (zvolené proměnné a metoda) měl dosáhnout maximální úspěšnosti klasifikace. Zbývající scénáře odpovídají reálným empirickým datům z hlediska vnitroobjektové variability a generických funkcí. Scénáře 7 a 8 jsou nejkomplexnější, s vysokým počtem vygenerovaných shluků (C = 9). Tyto scénáře jsou zatěžkávacím testem pro ověřované shlukovací postupy.

Scénář 1

Obsahuje 3 shluky (n_c = 100, pro každé c, N = 300, n_i = 5 pro každé i). Trajektorie jsou lineární (pozitivní trend, konstanta a negativní trend) s velmi malou vnitroobjektovou variabilitou.

Trajektorie (hodnoty y_{it} pro i = 1, …, 100 a D_{it} = 0, 1, 2, 3, 4) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it},\ \ \beta_0 \sim \mathrm{N}(10; 0{,}1^2), \beta_1 \sim \mathrm{U}(0{,}9; 1),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}05^2),$$

shluk 2:

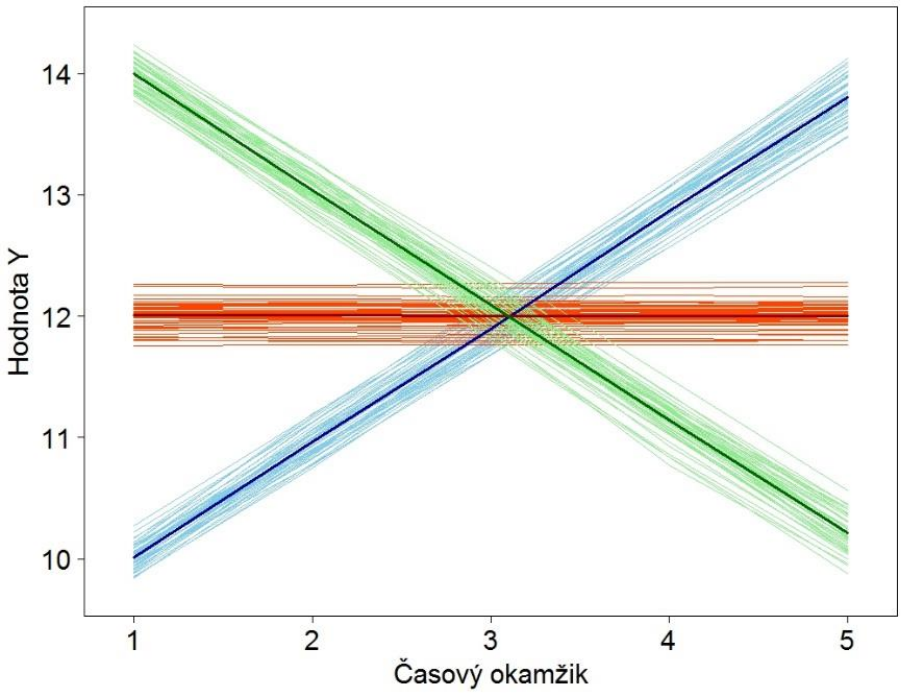
$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it},\quad \beta_0 \sim \mathrm{N}(12; 0{,}1^2), \beta_1 \sim \mathrm{U}(-0{,}01; 0{,}01),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}005^2),$$

shluk 3:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it},\ \ \beta_0 \sim \mathrm{N}(14; 0{,}1^2), \beta_1 \sim \mathrm{U}(-1; -0{,}9),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}05^2),$$

kde N(μ; σ^2) značí normální rozdělení s parametry μ (střední hodnota) a σ^2 (rozptyl). U(a,b) je zkráceně rovnoměrné rozdělení s parametry a (minimální hodnota) a b (maximální hodnota).

Časový okamžik $t = 1$ odpovídá měření k počátku, a tedy vzdálenost od počátku $D_{it} = 0$ atd., pro $t = 5$ platí $D_{i5} = 4$. Tento scénář považuji pouze za modelový vzhledem k téměř žádné vnitroobjektové variabilitě, jednoznačné heterogenitě[8] tvaru trajektorií mezi shluky a homogenitě tvaru uvnitř shluků. U vhodně navržených proměnných a metod shlukování očekávám 100% úspěšnost správné klasifikace. Scénář 1 ilustruje obrázek 7.3. Na časové ose je uvedeno 5 časových okamžiků $t = 1, \ldots, 5$. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. Trajektorie náležející do shluku 1 jsou nakresleny modře, trajektorie ve shluku 2 mají barvu oranžovo-červenou a třetí shluk je zelený. Tučné čáry zobrazují reprezentativní trajektorii daného shluku. Jedná se o spojnici průměrných hodnot pro jednotlivá t.



Obrázek 7.3 Ilustrace scénáře 1.

Scénář 2

Obsahuje 3 shluky ($n_c = 100$, pro každé c, $N = 300$, $T = 5$ pro každé i). Trajektorie jsou lineární (pozitivní trend, konstanta a negativní trend) s velkou vnitroobjektovou variabilitou.

Trajektorie (hodnoty y_{it} pro $i = 1, \ldots, 100$ a $t = 1, \ldots, 5$) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}6; 1{,}4),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}5^2),$$

[8] Heterogenitou a homogenitou tvaru trajektorií je bráno do jaké míry se překrývají objekty mezi shluky a uvnitř shluků. Jednoznačná heterogenita tvaru trajektorií mezi shluky znamená, že není překryv objektů z různých shluků. Homogenita tvaru uvnitř shluku značí, že objekty (trajektorie) uvnitř shluku jsou si podobné.

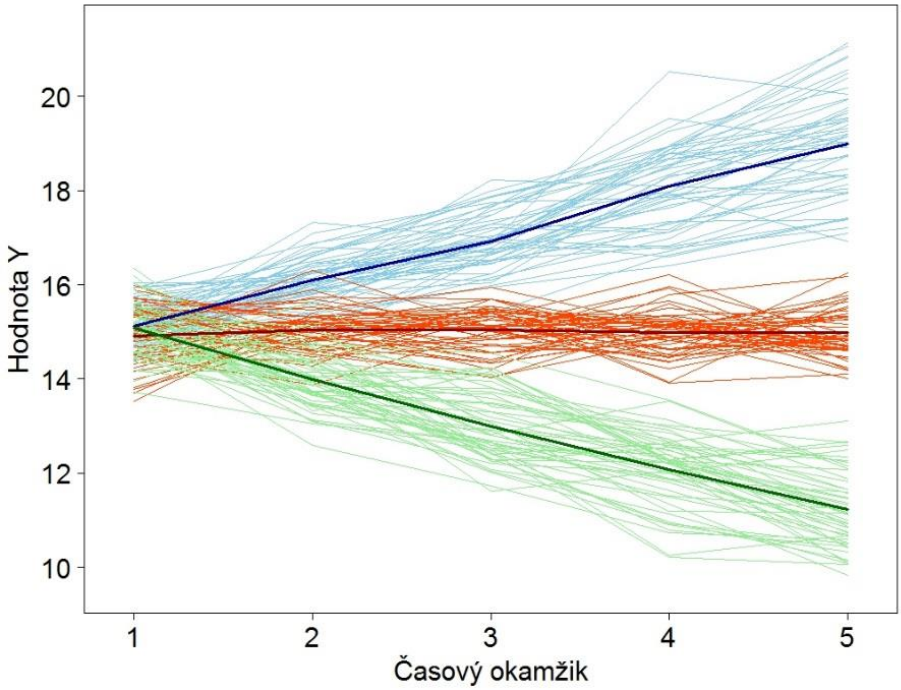
shluk 2:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(-0{,}05; 0{,}05),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}5^2),$$

shluk 3:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(-1{,}4; -0{,}6),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}5^2).$$

Scénář 2 odpovídá empirickým datům lépe než scénář 1 z důvodu výskytu vnitroobjektové variability. Scénář ilustruje obrázek 7.4. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. Z obrázku je patrná velká heterogenita mezi shluky a u vhodně navržených charakteristik lze očekávat vysokou klasifikační schopnost blízkou 100%.



Obrázek 7.4 Ilustrace scénáře 2.

Scénář 3

Obsahuje 4 shluky (n_c = 100, pro každé c, N = 400, T = 5 pro každé i). Trajektorie jsou nelineární s minimální vnitroobjektovou variabilitou. Jedná se o skupinu 2 kvadratických a 2 logaritmických funkcí, které zajišťují rostoucí a klesající průběh s konvexním a konkávním tvarem.

Trajektorie (hodnoty y_{it} pro i = 1, …, 100 a t = 1, …, 5) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}25; 0{,}30),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2),$$

shluk 2:

$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5), \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$

$\varepsilon_{it} \sim \mathrm{N}(0;0{,}25^2),$

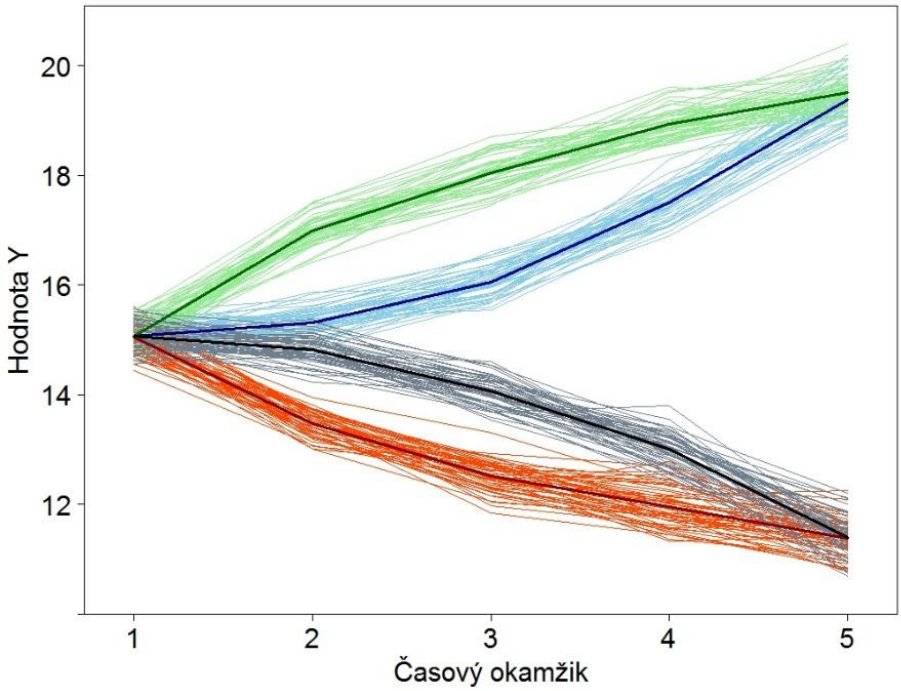
shluk 3:

$y_{it} = \beta_0 + \log_{\beta_1}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(1{,}4; 1{,}45), \varepsilon_{it} \sim \mathrm{N}(0;0{,}25^2),$

shluk 4:

$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(-0{,}25; -0{,}20), \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2).$

Scénář 3 ilustruje obrázek 7.5. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. Oproti předchozím scénářům je v grafu jeden shluk navíc (čtvrtý shluk) vyjádřen šedou barvou. Z obrázku je zjevná velká heterogenita naměřených hodnot mezi shluky v časovém intervalu mezi $t = 2$ a $t = 4$. Toto prostřední časové období je z hlediska tvaru trajektorií více homogenní (podobnost tvaru vývoje mezi shluky 2 a 4, 1 a 3) než první a pátý rok sledování. V prvním a posledním roce měření se všechny 4 shluky navzájem významně liší v tempu nárůstu či poklesu. Klasifikační schopnost lze očekávat nižší než pro scénář 1. Důvodem je větší počet shluků a měnící se míra disperze tvaru trajektorií v různých časových okamžicích. Z důvodu nízké vnitroobjektové variability lze u vhodného přístupu ke shlukování očekávat vysokou míru klasifikace blízkou 100%.



Obrázek 7.5 Ilustrace scénáře 3.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.
Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorii pro daný shluk.

Scénář 4

Obsahuje 4 shluky (n_c = 100, pro každé c, N = 400, T = 5 pro každé i). Trajektorie jsou nelineární s velkou vnitroobjektovou variabilitou. Jedná se o skupinu 2 kvadratických a 2 logaritmických funkcí, které zajišťují rostoucí a klesající průběh s konvexním a konkávním tvarem. Oproti scénáři 3 se liší ve velikosti variability náhodných chyb.

Trajektorie (hodnoty y_{it} pro i = 1, …, 100 a t = 1, …, 5) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(0{,}25; 0{,}30), \varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2),$$

shluk 2:

$$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5), \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$$
$$\varepsilon_{it} \sim \mathrm{N}(0;0{,}6^2),$$

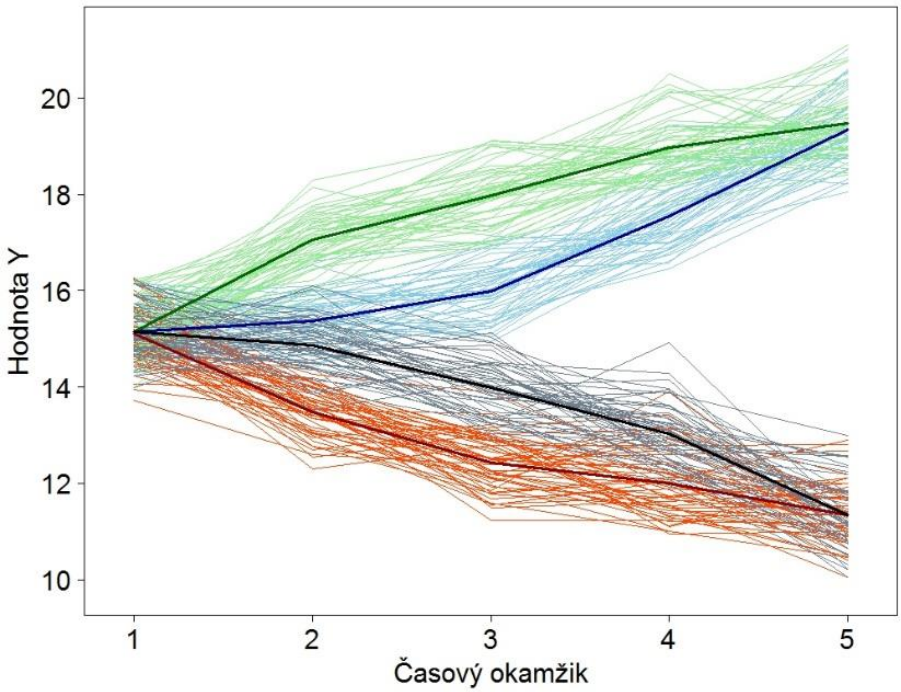
shluk 3:

$$y_{it} = \beta_0 + \log_{\beta_1}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(1{,}4; 1{,}45), \varepsilon_{it} \sim \mathrm{N}(0;0{,}6^2),$$

shluk 4:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(-0{,}25; -0{,}20), \varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2).$$

Scénář 4 ilustruje obrázek 7.6. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. Oproti scénáři 3 je menší heterogenita naměřených hodnot mezi shluky v časovém intervalu mezi t = 2 a t = 4. V tomto intervalu dochází v menší míře k průniku trajektorií ze shluků 1 a 3, 2 a 4. Toto prostřední časové období je z hlediska tvaru trajektorií více homogenní (podobnost tvaru vývoje mezi shluky 2 a 4, 1 a 3) než první a pátý rok sledování. V prvním a posledním roce měření se všechny 4 shluky navzájem významně liší v tempu nárůstu či poklesu. Klasifikační schopnost lze očekávat nižší než pro scénář 3. Důvodem je větší vnitroobjektová variabilita, která způsobuje, že u části případů není jednoznačná náležitost do shluku na základě vygenerovaného rozdělení.



Obrázek 7.6 Ilustrace scénáře 4.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.
Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorii pro daný shluk.

Scénář 5

Obsahuje 4 shluky (n_c = 100, pro každé *c*, *N* = 400, *T* = 5 pro každé *i*). Trajektorie jsou nelineární s minimální vnitroobjektovou variabilitou. Jedná se o skupinu 2 kvadratických a 2 logaritmických funkcí, které zajišťují rostoucí a klesající průběh s konvexním a konkávním tvarem.

Trajektorie (hodnoty y_{it} pro *i* = 1, …, 100 a *t* = 1, …, 5) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 (D_{it} - 2)^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(11{,}5; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}7; 0{,}9),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2),$$

shluk 2:

$$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5),\ \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$$
$$\varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2),$$

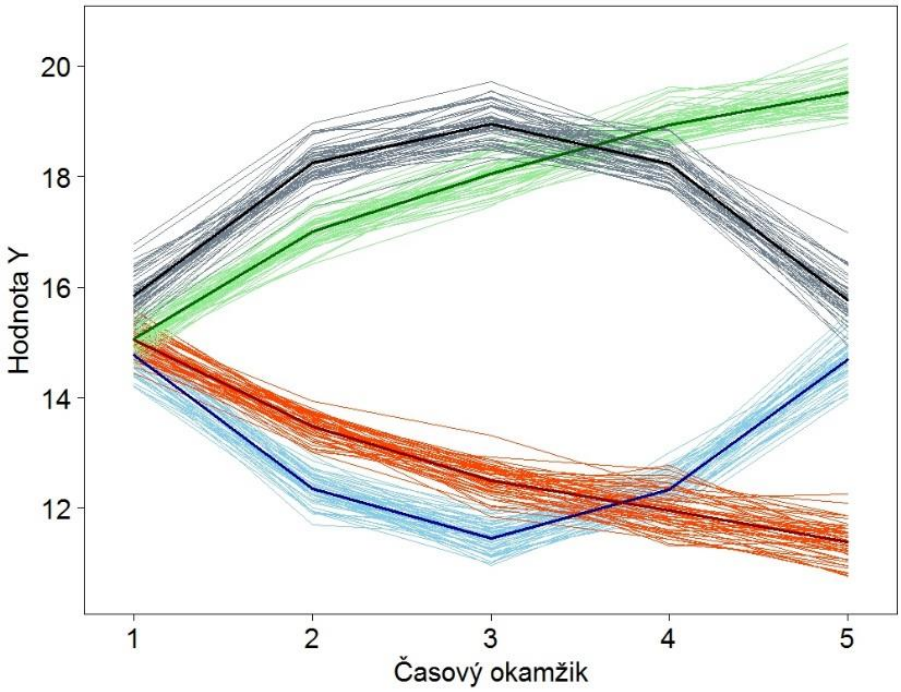
shluk 3:

$$y_{it} = \beta_0 + \log_{\beta_1}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(1{,}4; 1{,}45),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2),$$

shluk 4:

$$y_{it} = \beta_0 - \beta_1 (D_{it} - 2)^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(19; 0{,}2^2),\ \beta_1 \sim \mathrm{U}(0{,}7; 0{,}9),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2).$$

Scénář 5 ilustruje obrázek 7.7. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. Oproti scénáři 3 se liší v tom, že vývoj dvojice shluků (1 a 2, 3 a 4) je do $t = 3$ podobný, potom však dochází k zásadnímu odklonu. Při nízké vnitroobjektové variabilitě a zásadní odlišnosti jednotlivých shluků je očekávána vysoká klasifikační schopnost blízká 100% úspěšnosti správného zařazení.



Obrázek 7.7 Ilustrace scénáře 5.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.
Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorii pro daný shluk.

Scénář 6

Obsahuje 4 shluky ($n_c = 100$, pro každé c, $N = 400$, $T = 5$ pro každé i). Trajektorie jsou nelineární s minimální vnitroobjektovou variabilitou. Jedná se o skupinu 2 kvadratických a 2 logaritmických funkcí, které zajišťují rostoucí a klesající průběh s konvexním a konkávním tvarem.

Trajektorie (hodnoty y_{it} pro $i = 1, \ldots, 100$ a $t = 1, \ldots, 5$) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 (D_{it} - 2)^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(11{,}5; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}7; 0{,}9),\ \varepsilon_{it} \sim \mathrm{N}(0; 1{,}0^2),$$

shluk 2:

$$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5),\ \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$$

$$\varepsilon_{it} \sim \mathrm{N}(0; 1{,}0^2),$$

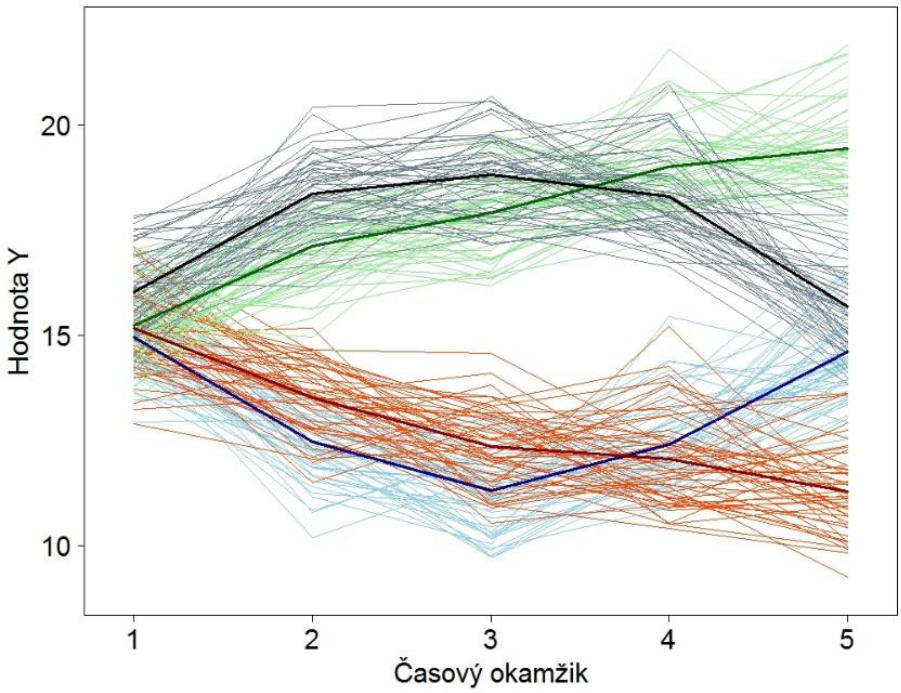
shluk 3:

$$y_{it} = \beta_0 + \log_{\beta_1}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15;0{,}1^2), \beta_1 \sim \mathrm{U}(1{,}4;1{,}45), \varepsilon_{it} \sim \mathrm{N}(0;1{,}0^2),$$

shluk 4:

$$y_{it} = \beta_0 - \beta_1(D_{it} - 2)^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(19;0{,}2^2), \beta_1 \sim \mathrm{U}(0{,}7;0{,}9), \varepsilon_{it} \sim \mathrm{N}(0;1{,}0^2).$$

Funkce jsou shodné se scénářem 5. Rozdíl je ve vnitroobjektové variabilitě, která je ve scénáři 6 výrazně větší. Velká variabilita způsobuje překryv významné části objektů do t = 3 u 1. a 2. shluku a mezi 3. a 4. shlukem. Tento fakt může vyústit ve snížení klasifikační schopnosti algoritmů a proměnných. Scénář 6 ilustruje obrázek 7.8. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti.



Obrázek 7.8 Ilustrace scénáře 6.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.
Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorii pro daný shluk.

Scénář 7

Obsahuje 9 shluků (n_c = 100, pro každé c, N = 900, T = 5 pro každé i). Jedná se o obecnou strukturu. Scénář kombinuje 3 lineární a 6 nelineárních funkcí (kvadratické, logaritmické a Gompertzova funkce). Vnitroobjektová variabilita je malá.

Trajektorie (hodnoty y_{it} pro i = 1, …, 100 a t = 1, …, 5) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15{,}0; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}25; 0{,}30),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2),$$

shluk 2:

$$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5),\ \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$$
$$\varepsilon_{it} \sim \mathrm{N}(0;0{,}25^2),$$

shluk 3:

$$y_{it} = \beta_0 + \log_{\beta_1}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(1{,}4;1{,}45),\ \varepsilon_{it} \sim \mathrm{N}(0;0{,}25^2),$$

shluk 4:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(-0{,}25; -0{,}20),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}25^2),$$

shluk 5:

$$y_{it} = \beta_0 + a \cdot e^{-b \cdot e^{-D_{it}}} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ a \sim \mathrm{N}(5{,}5; 0{,}2^2),\ b \sim \mathrm{U}(9;10),$$
$$\varepsilon_{it} \sim \mathrm{N}(0; 0{,}15^2),$$

shluk 6:

$$y_{it} = \beta_0 + a \cdot e^{-b \cdot e^{-D_{it}}} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ a \sim \mathrm{N}(-5{,}5; 0{,}2^2),\ b \sim \mathrm{U}(9;10),$$
$$\varepsilon_{it} \sim \mathrm{N}(0; 0{,}15^2),$$

shluk 7:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}95;1{,}05),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}05^2),$$

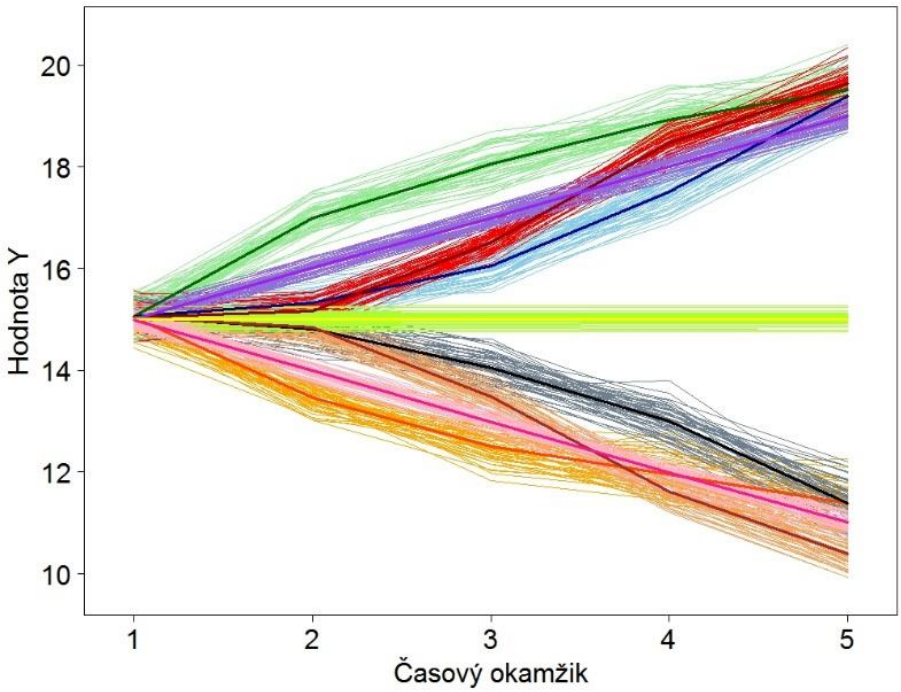
shluk 8:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(-0{,}01; 0{,}01),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}005^2),$$

shluk 9:

$$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(-1{,}05; -0{,}95),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}05^2).$$

Scénář 7 ilustruje obrázek 7.9. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. I přes nízkou vnitroobjektovou variabilitu lze očekávat výrazně větší klasifikační chybu než u předchozích případů z důvodu nižší mezishlukové variability a výrazně většího počtu shluků.



Obrázek 7.9 Ilustrace scénáře 7.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžová křivka, objekty z c = 3 zelená, objekty z c = 4 šedá, objekty z c = 5 červená, objekty z c = 6 hnědá, objekty z c = 7 fialová, objekty z c = 8 zeleno-žlutá a objekty z c = 9 růžová.
Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorii pro daný shluk.

Scénář 8

Obsahuje 9 shluků (n_c = 100, pro každé c, N = 900, T = 5 pro každé i). Jedná se o obecnou strukturu. Scénář 8 se liší oproti scénáři 7 ve variabilitě náhodných chyb, která je zde větší.

Trajektorie (hodnoty y_{it} pro i = 1, …, 100 a t = 1, …, 5) byly vygenerovány pomocí následujících funkcí:

shluk 1:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15{,}0; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(0{,}25; 0{,}30),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2),$$

shluk 2:

$$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5),\ \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$$
$$\varepsilon_{it} \sim \mathrm{N}(0;0{,}6^2),$$

shluk 3:

$$y_{it} = \beta_0 + \log_{\beta_1}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(1{,}4;1{,}45),\ \varepsilon_{it} \sim \mathrm{N}(0;0{,}6^2),$$

shluk 4:

$$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2),\ \beta_1 \sim \mathrm{U}(-0{,}25; -0{,}20),\ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2),$$

shluk 5:

$y_{it} = \beta_0 + a \cdot e^{-b \cdot e^{-D_{it}}} + \varepsilon_{it}$, $\beta_0 \sim \mathrm{N}(15; 0{,}1^2)$, $a \sim \mathrm{N}(5{,}5; 0{,}2^2)$, $b \sim \mathrm{U}(9;10)$, $\varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2)$,

shluk 6:

$y_{it} = \beta_0 + a \cdot e^{-b \cdot e^{-D_{it}}} + \varepsilon_{it}$, $\beta_0 \sim \mathrm{N}(15; 0{,}1^2)$, $a \sim \mathrm{N}(-5{,}5; 0{,}2^2)$, $b \sim \mathrm{U}(9;10)$, $\varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2)$,

shluk 7:

$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}$, $\beta_0 \sim \mathrm{N}(15; 0{,}1^2)$, $\beta_1 \sim \mathrm{U}(0{,}95;1{,}05)$, $\varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2)$,

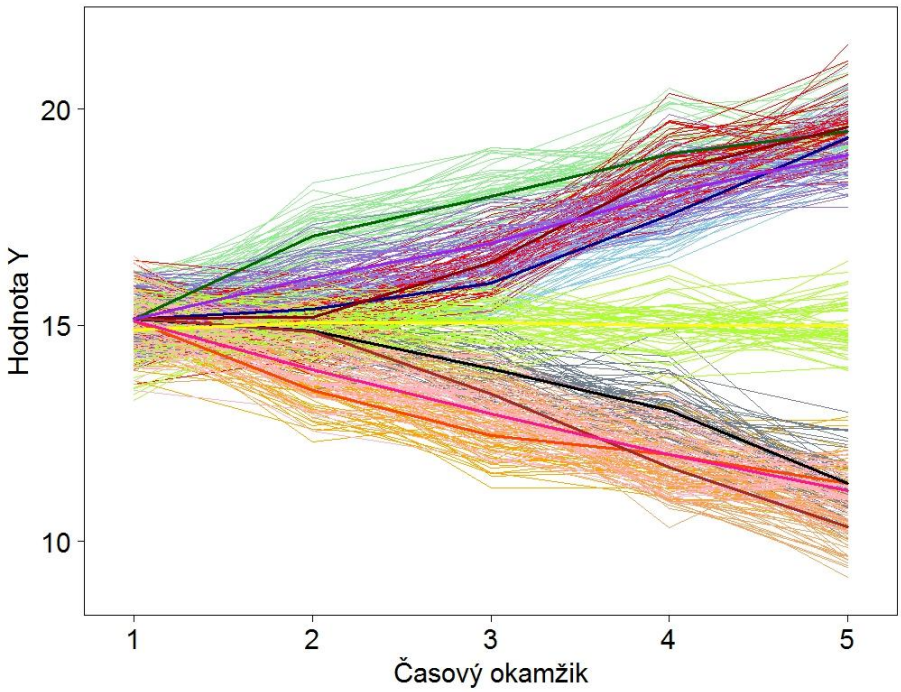
shluk 8:

$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}$, $\beta_0 \sim \mathrm{N}(15; 0{,}1^2)$, $\beta_1 \sim \mathrm{U}(-0{,}01; 0{,}01)$, $\varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2)$,

shluk 9:

$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}$, $\beta_0 \sim \mathrm{N}(15; 0{,}1^2)$, $\beta_1 \sim \mathrm{U}(-1{,}05; -0{,}95)$, $\varepsilon_{it} \sim \mathrm{N}(0; 0{,}6^2)$.

Scénář 8 ilustruje obrázek 7.10. Na obrázku je zobrazeno pro každý shluk pouze 50 trajektorií z důvodu přehlednosti. Vyšší (ve srovnání se scénářem 7) vnitroobjektová variabilita způsobuje malou heterogenitu mezi některými shluky. Lze očekávat největší klasifikační chybu ze všech scénářů. Tento scénář vnímám jako zatěžkávací test. Získáme představu o minimální úrovni klasifikační schopnosti shlukovacích postupů pro složité obecné datové struktury, které se řídí směsí různých rozdělení, a zároveň se v datech vyskytuje velká vnitroobjektová variabilita.



Obrázek 7.10 Ilustrace scénáře 8.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžová křivka, objekty z c = 3 zelená, objekty z c = 4 šedá, objekty z c = 5 červená, objekty z c = 6 hnědá, objekty z c = 7 fialová, objekty z c = 8 zeleno-žlutá a objekty z c = 9 růžová.
Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorii pro daný shluk.

Shlukovací postupy

Na každý z osmi výše popsaných scénářů bylo aplikováno několik shlukovacích postupů. Shlukovací postupy se liší v použité kombinaci navržených charakteristik a shlukovacích algoritmů. Porovnává se šest kombinací shlukovacích proměnných. První kombinace (s1) obsahuje všechny navrhované charakteristiky popsané v oddílu 7.1. Ostatní kombinace (s2 až s5) jsou podmnožinou s1. Kombinace se skládají z těchto proměnných:

- s1 – $\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$), max $\lhd$, $\overline{k_i}$ a %pos,
- s2 – $\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$), max $\lhd$ a $\overline{k_i}$,
- s3 – $\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$) a max $\lhd$,
- s4 – $\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$) a $\overline{k_i}$,
- s5 – $\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$) a %pos,
- s6 – max $\lhd$, $\overline{k_i}$ a %pos.

Kombinace s2 až s6 jsem vybral z množiny většího množství různých kombinací na základě pilotní analýzy, jejíž výsledky jsem prezentoval na konferenci AS2015, viz Sobíšek a Stachová

(2015). Všechny hodnoty shlukovacích proměnných e_a, vyjma max ◁ (úhlu v radiánech), jsem znormoval na *z*-skóre dle $z_{ia} = \left(e_{ia} - \overline{e_a}\right) / \mathrm{sd}(e_a)$, e_{ia} je hodnotá *a*-té shlukovací proměnné pro *i*-tý objekt, $\overline{e_a}$ je průměrná hodnota proměnná e_a a sd(e_a) je její směrodatná odchylka. Na *z*-skóre těchto proměnných aplikuji postupně 7 shlukovacích algoritmů:

- a1 – aglomerativní shlukování; algoritmus úplné vazby (úplná vazba); euklidovská matice vzdáleností (euklidovská),
- a2 – aglomerativní shlukování; algoritmus Wardův (Ward); euklidovská;
- a3 – aglomerativní shlukování; úplná vazba, Mahalanobisova matice vzdáleností (Mahalanobisova);
- a4 – aglomerativní shlukování; Ward, Mahalanobisova;
- a5 – aglomerativní shlukování; úplná vazba, míra podobnosti založená na Pearsonovu korelačním koeficientu (Pearson);
- a6 – aglomerativní shlukování; Ward, Pearson;
- a7 – metoda *C*-průměrů.

Za účelem porovnání postupů z hlediska schopnosti klasifikovat trajektorie do správných shluků jsem využil 4 různá hodnoticí kritéria: Randův index, modifikovaný Randův index, Jaccardův koeficient a míru správné klasifikace (*correct classification rate, CCR*). Tyto koeficienty kvantifikují, do jaké míry se shodují odhadnuté a skutečné (známé) příslušnosti objektů ke shlukům. Skutečnou náležitost do *C* tříd lze zapsat jako *TRUE* = $\{c_1, \ldots, c_C\}$, kde c_c jsou uvažované třídy (shluky). Výsledek shlukové analýzy, tedy náležitost do *O* shluků odhadnutou shlukovacím algoritmem označme *PRED* = $\{l_1, \ldots, l_O\}$, kde l_o jsou odhadnuté shluky. V provedené simulaci platí $C = O$.

Randův index (*RI*) (Rand, 1971) slouží jako míra konkordance mezi skutečnou strukturou shluků a strukturou odhadnutou daným shlukovacím algoritmem. Spočítá se následovně

$$RI = \frac{a+d}{a+b+c+d}, \tag{7.8}$$

kde počet párů objektů správně zařazených do stejné třídy v *TRUE* a stejného shluku v *PRED* se značí *a*. Nechť *b* určuje počet párů objektů ve stejné třídě *TRUE*, ale v jiném shluku v *PRED*, *c* je počet párů objektů ve stejném shluku v *PRED*, ale v jiné třídě v *TRUE* a *d* značí počet párů objektů v jiných třídách a jiných shlucích v *PRED* a *TRUE*. *RI* nabývá hodnot od 0 do 1, kdy 1 značí perfektní shodu (konkordanci) *PRED* a *TRUE*. Střední hodnota *RI* není konstantní pro dvě náhodné klasifikace (*TRUE* a *PRED*). Tento problém řeší modifikovaný RI, pro který je jeho střední hodnota konstantní. RI je upraven o tuto střední hodnotu.

Modifikovaný Randův index (*RIadj*), viz Hubert a Arabie (1985) předpokládá zobecněné hypergeometrické rozdělení jako model náhodnosti, tj. *PRED* a *TRUE* jsou brány náhodně tak, aby počet objektů ve třídách a shlucích byl fixní. *RIadj* nabývá rovněž hodnot od 0 do 1 a počítá se následovně:

$$RIadj = \frac{RI - E(RI)}{\max(RI) - E(RI)} = \frac{a - \left(\frac{(b+a)\cdot(c+a)}{a+b+c+d}\right)}{\frac{(b+c+2a)}{2} - \left(\frac{(b+a)\cdot(c+a)}{a+b+c+d}\right)}, \tag{7.9}$$

max(*RI*) je maximální hodnota indexu a *E*(*RI*) je očekávaná hodnota RI. Bližší popis s ukázkou výpočtu *RIadj* ukazují Yeung a Ruzzo (2001).

Jaccardův koeficient Jac, viz Hebák a kol. (2007); Ramey (2015), který je definován jako

$$JAC = \frac{a}{a+b+c}, \tag{7.10}$$

nabývá hodnot 0 až 1.

Konečně míra správné klasifikace (correct classification rate, CCR) je definována jako

$$CCR = \frac{a}{a+b+c+d} \tag{7.11}$$

a nabývá hodnot od 0 do 1.

RI použili Ferreira a Hitchcock (2009) ve své simulační studii pro porovnání algoritmů shlukování trajektorií. Přednost *RIadj* před *RI* dali ve své studii Koestler a kol. (2014) a *CCR* je užito pro validaci algoritmu, který je obsažen v balíčku *kml* Genoliniho a Falissarda (2011). Pro výpočet *RI* a *Jac* jsem ve své simulaci použil funkci *cluster_similarity()* v balíčku *clusteval* Rameye (2015). *RIadj* byl počítán pomocí funkce *adjustedRandIndex()* z balíčku *mclust* Fraleye a kol. (2012).

Výsledky shlukování

Velké množství testovaných postupů (6 kombinací shlukovacích proměnných × 7 shlukovacích algoritmů = 42) a scénářů (8) vede k dlouhému výpočetnímu času. U postupů, kde se počítá Mahalanobisova míra vzdálenosti, je výpočetní čas extrémně dlouhý. Z tohoto důvodu jsem experiment zopakoval pouze stokrát pro každý testovaný postup a scénář. Získaných 100 hodnot pro každé kritérium je prezentováno v tabulkách níže souhrnně pomocí průměru a dolního kvartilu.

V této části jsou uvedeny pro každý scénář výsledky shlukování. V tabulkách jsou uvedeny průměrné hodnoty a dolní kvartily použitých hodnotících kritérií. Tyto charakteristiky polohy slouží k porovnání úspěšnosti shlukovacích postupů. Kvantily jsou robustnější (vhodnější pro

popis rozdělení četností v případě nesymetrického rozdělení) charakteristiky polohy než průměr. Dolní kvartil stanovuje dolní („kritickou“) hodnotu, pod kterou je nízká šance (25 ze 100) očekávaného výskytu hodnoticího kritéria.

Pro každý scénář je graficky ilustrována úspěšnost klasifikace pro nejlepší shlukovací postup (kombinace proměnných a metod). Zobrazený výsledek pochází z jedné replikace. Pro přehlednost je zobrazen výsledek z replikace, ve které bylo nasimulováno 50 individuálních trajektorií v každém objektu (místo 100).

Scénář 1

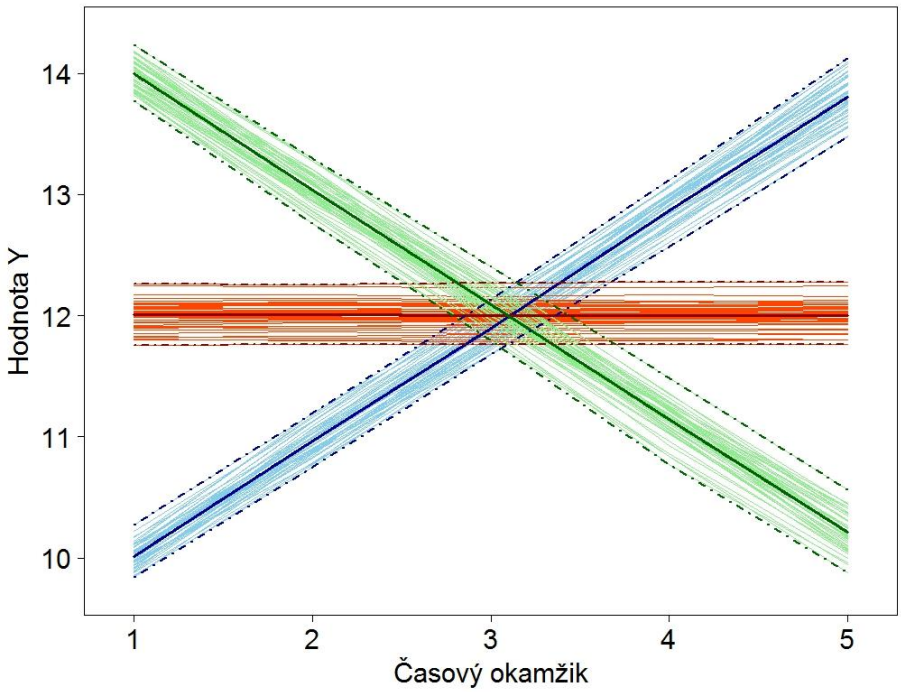
Tabulka 7.3 obsahuje deskriptivní statistiku čtyř hodnoticích kritérií. Tabulka je rozdělena do 8 bloků. Sedm bloků obsahuje statistiky pro jednotlivé shlukovací metody (a1 – a7). Poslední (pravý dolní) blok tabulky obsahuje minimální, maximální hodnoty a variační rozpětí hodnoticích kritérií ze všech 42 hodnot. Prvních 5 kombinací proměnných (s1 až s5) správně se 100% úspěšností zařadilo objekty do *TRUE* tříd, pokud se na ně aplikovaly metody a2 (aglomerativní metoda úplné vazby a Mahalanobisova vzdálenost) a a6 (aglomerativní algoritmus – Wardova metoda, míra podobnosti založená na Pearsonovu korelačním koeficientu d_{cc} (Pearson), vzorec 5.1). Metoda úplné vazby v kombinaci s Pearsonovou mírou podobnosti (a5) dosáhla 100% úspěšnosti pro kombinace s1 a s5. Nejúspěšnější kombinací pro tento scénář je s1 ($\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$), max $\triangleleft$, $\overline{k_i}$ a %pos) a s5 ($\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$), %pos), která v kombinaci s nejvíce algoritmy (a2, a5, a6) dosáhla 100% výsledek (hodnoty všech 4 hodnoticích kritérií jsou rovny jedné). Z tabulky 7.3 lze vyčíst, že metoda *C*-průměrů v kombinaci s s4 a s5 zařadila správně objekty ve více než 75% iterací, viz dolní kvartil roven 1, avšak v některých iteracích (průměr < 1) dopadl tento shlukovací postup špatně. Obrázek 7.11 porovnává reprezentanty skutečných tříd (*TRUE*) a odhadnutých shluků (*PRED*) pro maximálně úspěšné postupy (kombinace proměnných a shluková metoda). Reprezentativní hodnota skutečné třídy je průměrná hodnota pro dané t počítaná ze všech objektů ve shluku c_c, kde $c = 1, \ldots, C$. Na grafu je průměrná (reprezentativní) trajektorie vynesena jako tučná spojnice těchto průměrů přes všech T měření. Reprezentativní trajektorie pro odhadnuté náležitosti do shluku l_o je počítána stejně a je znázorněna jako tučná přerušovaná trajektorie. Slabou, přerušovanou spojnicí jsou vyznačeny minimální a maximální hodnoty pro odhadnutý shluk. Ve scénáři 1 došlo k perfektní klasifikaci pro více metod a proměnných. Obrázek 7.11 je totožný pro tyto maximálně úspěšné scénáře. Dokonalá shoda se projevuje v tom, že spojnice reprezentativní trajektorie pro skutečné třídy a pro odhadnuté shluky jsou identické. Rovněž minimální

a maximální odhadnuté hodnoty kopírují očekávané minimální a maximální hodnoty pro všechny 4 třídy.

Tabulka 7.3 Hodnoticí kritéria pro scénář 1

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,60; 0,52	0,60; 0,49	0,26; 0,08	0,41; 0,32	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s2	0,56; 0,51	0,56; 0,49	0,19; 0,08	0,37; 0,32	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s3	0,53; 0,48	0,57; 0,51	0,17; 0,08	0,35; 0,32	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s4	0,57; 0,51	0,58; 0,50	0,21; 0,09	0,38; 0,32	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s5	0,56; 0,50	0,56; 0,49	0,18; 0,08	0,37; 0,32	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s6	0,72; 0,71	0,78; 0,78	0,56; 0,56	0,58; 0,58	0,95; 0,95	0,94; 0,95	0,87; 0,95	0,84; 0,95
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,42; 0,39	0,46; 0,41	0,04; 0,01	0,30; 0,29	0,68; 0,61	0,79; 0,75	0,58; 0,50	0,61; 0,54
s2	0,45; 0,40	0,49; 0,44	0,06; 0,02	0,31; 0,30	0,94; 0,94	0,94; 0,93	0,88; 0,84	0,86; 0,81
s3	0,42; 0,39	0,45; 0,42	0,04; 0,01	0,31; 0,30	0,87; 0,73	0,90; 0,79	0,78; 0,53	0,77; 0,54
s4	0,45; 0,41	0,50; 0,44	0,07; 0,02	0,30; 0,29	0,98; 0,98	0,98; 0,97	0,95; 0,94	0,94; 0,92
s5	0,42; 0,38	0,45; 0,40	0,03; 0,01	0,30; 0,30	0,84; 0,63	0,87; 0,75	0,74; 0,52	0,74; 0,56
s6	0,51; 0,46	0,58; 0,52	0,16; 0,08	0,33; 0,29	0,95; 0,94	0,94; 0,93	0,86; 0,84	0,83; 0,81
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s2	0,99; **1,00**	0,99; **1,00**	0,97; **1,00**	0,97; **1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s3	0,86; 0,75	0,87; 0,76	0,71; 0,48	0,70; 0,49	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s4	0,98; **1,00**	0,98; **1,00**	0,96; **1,00**	0,96; **1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s5	**1,00; 1,00**	**1,00; 1,00**	0,99; **1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s6	0,95; 0,94	0,94; 0,93	0,87; 0,84	0,84; 0,81	0,95; 0,94	0,94; 0,93	0,86; 0,84	0,84; 0,81
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,84; 0,57	0,90; 0,73	0,81; 0,47	0,82; 0,52	minimum			
s2	0,86; 0,59	0,92; 0,74	0,83; 0,48	0,85; 0,53	0,42; 0,38	0,45; 0,38	0,03; 0,08	0,30; 0,32
s3	0,86; 0,58	0,92; 0,73	0,83; 0,47	0,85; 0,52	maximum			
s4	0,89; **1,00**	0,93; **1,00**	0,87; **1,00**	0,88; **1,00**	1,00; 1,00	1,00; 1,00	1,00; 1,00	1,00; 1,00
s5	0,90; **1,00**	0,94; **1,00**	0,88; **1,00**	0,89; **1,00**	rozpětí (maximum - minimum)			
s6	0,86; 0,57	0,91; 0,73	0,83; 0,47	0,84; 0,52	0,58; 0,62	0,55; 0,62	0,97; 0,92	0,70; 0,68

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.11 Ilustrace úspěšnosti nejlepších shlukovacích postupů pro scénář 1. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka a objekty z c = 3 zelená.

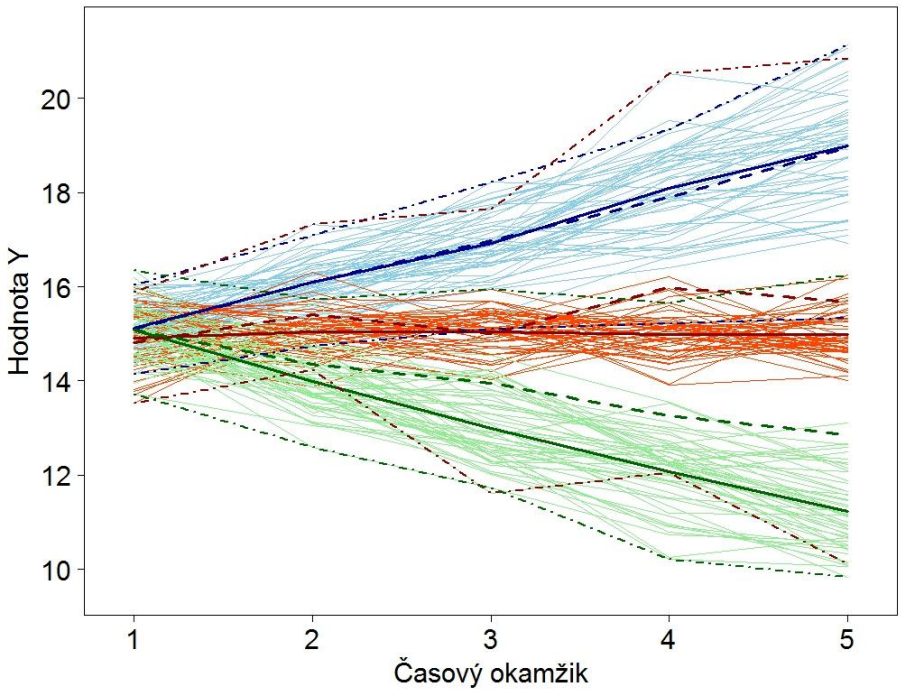
Scénář 2

U scénáře s větší variabilitou poklesla klasifikační schopnost. Nejlépe si vedla kombinace proměnných s2 ($\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd$(|diff_i|)$, max ◁ a $\overline{k_i}$), která v kombinaci s metodou *C*-průměrů (a7) dosáhla nejlepších výsledků (CCR: 0,89; **0,93**, *RI*: **0,90**; **0,91**, *RIadj*: **0,78**; **0,80**, *Jac*: **0,77**; **0,77**). Pro průměrnou hodnotu *CCR* dosáhla tato kombinace s2 maximálních výsledků ještě v kombinaci s Wardovou metodou s d_{cc} (a6) a s metodou úplné vazby a Mahalanobisovou vzdáleností (a2). Metody a1 (aglomerativní metoda, úplná vazba a euklidovská vzdálenost) a a3 (aglomerativní metoda, úplná vazba a Mahalanobisova vzdálenost) nedopadly dobře. Více v tabulce 7.4. Obrázek 7.12 porovnává reprezentanty skutečných tříd (*TRUE*) a odhadnutých shluků (*PRED*) pro nejúspěšnější postup (kombinace proměnných a shluková metoda) a7 s s2. Z obrázku je vidět, kde došlo k chybné klasifikaci. Odhadnuté shluky l_2 (červená přerušovaná spojnice) a l_3 (zelená přerušovaná spojnice) jsou částečně smíchány dohromady. Z obrázku je patrné, že se nejméně podařilo klasifikovat správně do třídy c_3. Pro ni lze vidět největší disperzi mezi TRUE (zelená tučná) a PRED (zelená přerušovaná) reprezentantem.

Tabulka 7.4 Hodnoticí kritéria pro scénář 2

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,58; 0,52	0,61; 0,55	0,21; 0,11	0,35; 0,31	0,87; 0,85	0,84; 0,82	0,65; 0,62	0,63; 0,60
s2	0,56; 0,50	0,60; 0,54	0,19; 0,10	0,34; 0,30	**0,90**; 0,86	0,88; 0,84	0,74; 0,65	0,71; 0,62
s3	0,49; 0,43	0,55; 0,52	0,09; 0,04	0,28; 0,26	0,78; 0,72	0,78; 0,72	0,52; 0,39	0,53; 0,44
s4	0,56; 0,49	0,59; 0,54	0,18; 0,08	0,34; 0,31	0,89; 0,85	0,87; 0,84	0,72; 0,64	0,70; 0,61
s5	0,52; 0,47	0,58; 0,53	0,13; 0,07	0,30; 0,28	0,74; 0,69	0,73; 0,68	0,41; 0,32	0,45; 0,39
s6	0,80; 0,73	0,81; 0,75	0,58; 0,45	0,58; 0,48	0,88; 0,86	0,86; 0,85	0,68; 0,66	0,66; 0,64
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,44; 0,41	0,48; 0,44	0,05; 0,02	0,30; 0,28	0,81; 0,80	0,79; 0,76	0,54; 0,48	0,54; 0,50
s2	0,43; 0,40	0,46; 0,42	0,04; 0,02	0,30; 0,29	0,85; 0,83	0,83; 0,79	0,63; 0,55	0,61; 0,55
s3	0,43; 0,39	0,49; 0,46	0,05; 0,01	0,29; 0,27	0,86; 0,84	0,84; 0,81	0,64; 0,57	0,63; 0,56
s4	0,43; 0,40	0,48; 0,44	0,05; 0,01	0,30; 0,28	0,88; 0,86	0,86; 0,83	0,69; 0,63	0,67; 0,60
s5	0,46; 0,42	0,51; 0,47	0,07; 0,03	0,30; 0,28	0,83; 0,82	0,80; 0,78	0,57; 0,52	0,57; 0,52
s6	0,58; 0,49	0,67; 0,66	0,31; 0,25	0,39; 0,34	0,81; 0,69	0,80; 0,69	0,55; 0,32	0,56; 0,39
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,79; 0,71	0,81; 0,75	0,57; 0,45	0,57; 0,48	0,87; 0,86	0,85; 0,84	0,67; 0,64	0,65; 0,62
s2	0,78; 0,70	0,80; 0,76	0,57; 0,47	0,57; 0,49	**0,90**; 0,87	0,88; 0,86	0,74; 0,68	0,72; 0,65
s3	0,71; 0,63	0,74; 0,67	0,42; 0,28	0,46; 0,36	0,81; 0,75	0,81; 0,77	0,58; 0,49	0,57; 0,50
s4	0,77; 0,70	0,80; 0,76	0,55; 0,46	0,55; 0,48	**0,90**; 0,86	0,88; 0,84	0,73; 0,65	0,71; 0,63
s5	0,67; 0,62	0,71; 0,68	0,36; 0,29	0,42; 0,37	0,77; 0,72	0,75; 0,69	0,46; 0,34	0,48; 0,40
s6	0,78; 0,72	0,79; 0,74	0,54; 0,44	0,54; 0,47	0,86; 0,85	0,85; 0,84	0,66; 0,64	0,64; 0,62
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,87; 0,72	0,89; 0,77	0,75; 0,51	0,74; 0,52	minimum			
s2	0,89; **0,93**	**0,90; 0,91**	**0,78; 0,80**	**0,77; 0,77**	0,43; 0,39	0,46; 0,42	0,04; 0,01	0,28; 0,26
s3	0,88; 0,75	0,89; 0,78	0,77; 0,52	0,75; 0,53	maximum			
s4	0,88; 0,73	0,89; 0,78	0,77; 0,52	0,75; 0,54	0,90; 0,93	0,90; 0,91	0,78; 0,80	0,77; 0,77
s5	0,88; 0,73	0,89; 0,77	0,77; 0,51	0,75; 0,53	rozpětí (maximum - minimum)			
s6	0,86; 0,71	0,88; 0,77	0,74; 0,50	0,73; 0,52	0,47; 0,54	0,44; 0,49	0,74; 0,79	0,49; 0,51

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.12 Ilustrace úspěšnosti nejlepšího shlukovacího postupu a7+s2 pro scénář 2. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka a objekty z c = 3 zelená. Tučně zvýrazněné spojnice znázorňují reprezentativní trajektorie pro očekávané třídy.

Scénář 3

U scénáře s větším počtem shluků (C = 4) a nelineárním průběhem trajektorií se zanedbatelnou vnitroobjektovou variabilitou poklesla klasifikační schopnost. Nejlépe si vedla kombinace proměnných s6 ($\max \triangleleft$, $\overline{k_i}$ a %pos), která dosáhla nejvyšších hodnot dolních kvartilů pro všechna 4 hodnoticí kritéria v kombinaci s metodou a6. V průměrných hodnotách hodnoticích kritérií dopadly nejlépe kombinace s3 ($\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd$(|diff_i|)$ a $\max \triangleleft$) a s4 ($\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd$(|diff_i|)$ a $\overline{k_i}$) po aplikaci metody a7 (C-průměrů). Velmi podobných průměrů dosáhla kombinace s6 ($\max \triangleleft$, $\overline{k_i}$ a %pos) s a7. Metody a1 (aglomerativní metoda, úplná vazba a euklidovská vzdálenost) a a3 (aglomerativní metoda, úplná vazba a Mahalanobisova vzdálenost) dopadly opět nejhůře (stejně jako u předchozích scénářů). Více v tabulce 7.5. Obrázek 7.13 porovnává reprezentanty skutečných tříd (*TRUE*) a odhadnutých shluků (*PRED*) pro nejúspěšnější postup a6 s s6. Z obrázku je vidět, kde došlo k chybné klasifikaci. Náležitost do shluků c_2 (oranžové trajektorii) a c_4 (šedé trajektorie) se zdá být odhadnuta v pořádku. Minimální, maximální a průměrné hodnoty jsou identické z hlediska minimálních, maximálních a průměrných hodnot s odhadnutými shluky l_2 (červená tučná přerušovaná spojnice reprezentující odhadnutou

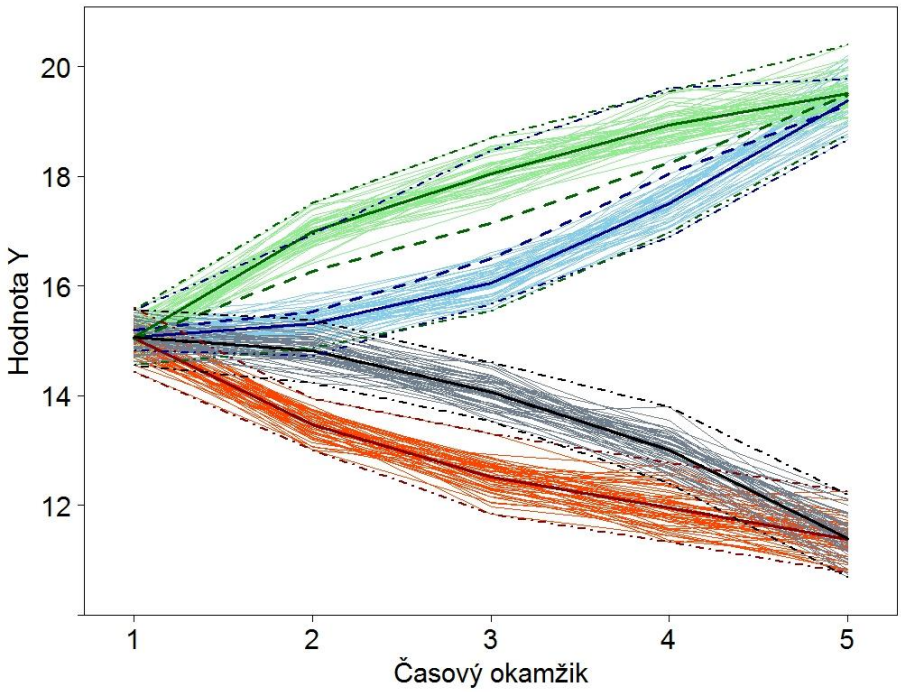
průměrnou trajektorii je identická s červenou tučnou spojnicí; červené slabé přerušované spojnice, které reprezentují minimální a maximální hodnoty v l_2 kopírují minimální a maximální oranžové trajektorie (hodnoty) v c_2) a l_4 (černá tučná přerušovaná spojnice reprezentující odhadnutou průměrnou trajektorii je identická s černou tučnou spojnicí; černé slabé přerušované spojnice, které reprezentují minimální a maximální hodnoty v l_2 kopírují minimální a maximální šedé trajektorie (hodnoty) v c_4).

Odhadnuté shluky l_1 (modrá přerušovaná tučná a slabá spojnice) a l_3 (zelená přerušovaná tučná a slabá spojnice) nekopírují přesně skutečné minimální, maximální a průměrné (tučné plné spojnice) hodnot v c_1 a c_3. U těchto dvou shluků dochází k částečné chybné záměně náležitosti trajektorií.

Tabulka 7.5 Hodnoticí kritéria pro scénář 3

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,55; 0,51	0,73; 0,72	0,33; 0,30	0,35; 0,32	0,61; 0,59	0,77; 0,76	0,39; 0,37	0,37; 0,36
s2	0,54; 0,50	0,72; 0,70	0,32; 0,27	0,34; 0,31	0,61; 0,59	0,76; 0,76	0,39; 0,37	0,38; 0,37
s3	0,49; 0,44	0,68; 0,63	0,22; 0,13	0,29; 0,24	0,62; 0,59	0,77; 0,76	0,39; 0,38	0,38; 0,37
s4	0,51; 0,46	0,70; 0,67	0,28; 0,20	0,32; 0,27	0,60; 0,58	0,76; 0,76	0,38; 0,37	0,37; 0,36
s5	0,52; 0,49	0,71; 0,69	0,28; 0,23	0,32; 0,29	0,59; 0,58	0,76; 0,75	0,37; 0,36	0,36; 0,35
s6	0,66; 0,55	0,81; 0,76	0,56; 0,48	0,53; 0,47	0,79; 0,78	0,88; **0,87**	0,70; **0,69**	0,64; **0,63**
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,34; 0,31	0,45; 0,35	0,04; 0,01	0,23; 0,22	0,56; 0,55	0,75; 0,75	0,45; 0,44	0,45; 0,44
s2	0,35; 0,32	0,51; 0,47	0,05; 0,02	0,23; 0,21	0,68; 0,65	0,81; 0,78	0,54; 0,49	0,51; 0,47
s3	0,35; 0,32	0,50; 0,47	0,04; 0,02	0,22; 0,21	0,68; 0,65	0,81; 0,79	0,55; 0,50	0,52; 0,47
s4	0,34; 0,31	0,54; 0,51	0,03; 0,01	0,21; 0,19	0,52; 0,49	0,74; 0,74	0,38; 0,38	0,39; 0,38
s5	0,34; 0,30	0,49; 0,44	0,03; 0,01	0,22; 0,20	0,56; 0,55	0,75; 0,75	0,45; 0,45	0,45; 0,44
s6	0,51; 0,45	0,60; 0,48	0,27; 0,09	0,36; 0,26	0,79; 0,78	0,88; **0,87**	0,70; **0,69**	0,64; **0,63**
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,60; 0,58	0,76; 0,76	0,39; 0,37	0,38; 0,36	0,62; 0,60	0,77; 0,76	0,39; 0,38	0,38; 0,37
s2	0,60; 0,58	0,76; 0,76	0,39; 0,37	0,38; 0,36	0,62; 0,60	0,77; 0,76	0,39; 0,38	0,38; 0,37
s3	0,60; 0,58	0,76; 0,76	0,39; 0,37	0,38; 0,37	0,63; 0,60	0,77; 0,76	0,40; 0,38	0,38; 0,37
s4	0,59; 0,58	0,76; 0,76	0,38; 0,37	0,37; 0,36	0,60; 0,59	0,76; 0,76	0,38; 0,37	0,37; 0,36
s5	0,58; 0,57	0,76; 0,75	0,37; 0,36	0,37; 0,36	0,59; 0,58	0,76; 0,76	0,37; 0,36	0,37; 0,36
s6	0,65; 0,55	0,80; 0,76	0,56; 0,48	0,53; 0,47	0,79; **0,79**	0,88; **0,87**	0,70; **0,69**	0,64; **0,63**
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,77; 0,62	0,89; 0,83	0,74; 0,59	0,71; 0,54	minimum			
s2	0,77; 0,61	0,89; 0,82	0,74; 0,58	0,70; 0,54	0,34; 0,30	0,45; 0,35	0,03; 0,01	0,21; 0,19
s3	**0,80**; 0,62	**0,90**; 0,83	**0,76**; 0,59	**0,73**; 0,55	maximum			
s4	**0,80**; 0,62	**0,90**; 0,83	**0,76**; 0,59	**0,73**; 0,54	0,80; 0,79	0,90; 0,87	0,76; 0,69	0,73; 0,63
s5	0,78; 0,61	0,89; 0,82	0,74; 0,58	0,71; 0,54	rozpětí (maximum - minimum)			
s6	0,79; 0,63	**0,90**; 0,83	0,75; 0,58	0,72; 0,54	0,46; 0,49	0,45; 0,52	0,73; 0,68	0,52; 0,44

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.13 Ilustrace úspěšnosti nejlepšího shlukovacího postupu a6+s6 pro scénář 3. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.

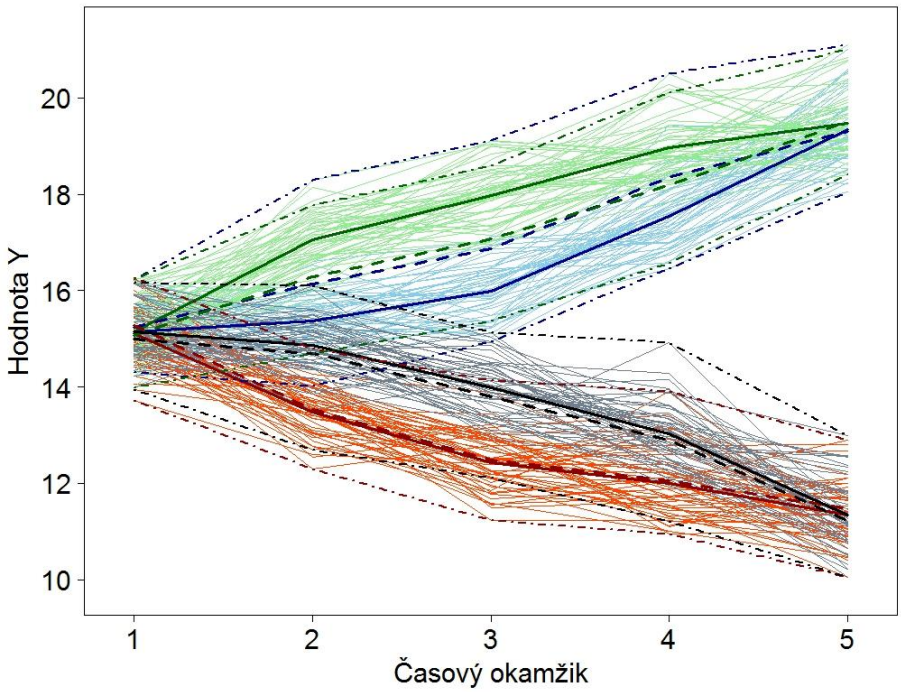
Scénář 4

U scénáře 4 s výskytem větší vnitroobjektové variability poklesla dle očekávání klasifikační schopnost ve srovnání se scénářem 3. Nejlépe si vedla kombinace proměnných s6 ($\max \triangleleft$, $\overline{k_i}$ a %pos). Nejlepší výsledek kombinace s6 dosáhla s Wardovou metodou a euklidovskou vzdáleností (a2) pro všechna hodnoticí kritéria (průměr i dolní kvartil). Obdobných výsledků dosáhla s Wardovou metodou a Pearsonovou mírou nepodobnosti (a6). Z ostatních kombinací proměnných nejlépe dopadla s2 v kombinaci s a7 (*C*-průměrů). Výsledek pro metodu *C*-průměrů (a7) je podobný pro všechny kombinace s1–s6. Metody a1 (metoda úplné vazby a euklidovská vzdálenost) a a3 (metoda úplné vazby a Mahalanobisova vzdálenost) dopadly opět nejhůře. Více v tabulce 7.6. Obrázek 7.14 porovnává reprezentanty skutečných tříd (*TRUE*) a odhadnutých shluků (*PRED*) pro nejúspěšnější postup a2 s s6. Z obrázku lze vyčíst, že reprezentanti odhadnutých shluků l_2 (přerušovaná červená spojnice) a l_4 (přerušovaná černá spojnice) jsou téměř totožní se zástupci skutečných tříd c_2 a c_4 (plná červená a černá spojnice). Reprezentanti odhadnutých shluků l_1 (přerušovaná modrá) a l_3 (přerušovaná zelená) jsou téměř identické a jsou v nich namíchány trajektorie z c_1 a c_3.

Tabulka 7.6 Hodnoticí kritéria pro scénář 4

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,43; 0,37	0,63; 0,57	0,17; 0,07	0,27; 0,22	0,55; 0,54	0,75; 0,75	0,35; 0,34	0,35; 0,34
s2	0,41; 0,36	0,62; 0,55	0,14; 0,04	0,25; 0,21	0,55; 0,54	0,75; 0,75	0,36; 0,35	0,36; 0,34
s3	0,35; 0,32	0,56; 0,54	0,05; 0,01	0,21; 0,19	0,55; 0,53	0,75; 0,75	0,35; 0,34	0,35; 0,34
s4	0,41; 0,36	0,63; 0,57	0,15; 0,06	0,25; 0,22	0,53; 0,52	0,75; 0,75	0,35; 0,34	0,35; 0,34
s5	0,36; 0,32	0,58; 0,55	0,07; 0,02	0,21; 0,19	0,52; 0,51	0,74; 0,73	0,31; 0,30	0,32; 0,32
s6	0,61; 0,51	0,78; 0,77	0,46; 0,43	0,43; 0,41	**0,71; 0,69**	**0,82; 0,81**	**0,52; 0,50**	**0,47; 0,46**
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,34; 0,31	0,51; 0,47	0,03; 0,01	0,21; 0,20	0,56; 0,52	0,75; 0,74	0,37; 0,34	0,36; 0,35
s2	0,34; 0,31	0,53; 0,51	0,03; 0,01	0,21; 0,19	0,60; 0,55	0,75; 0,74	0,38; 0,34	0,38; 0,35
s3	0,34; 0,32	0,56; 0,53	0,04; 0,01	0,20; 0,19	0,57; 0,52	0,74; 0,73	0,36; 0,31	0,37; 0,34
s4	0,32; 0,30	0,53; 0,50	0,02; 0,00	0,20; 0,18	0,48; 0,46	0,72; 0,72	0,31; 0,28	0,33; 0,31
s5	0,32; 0,30	0,56; 0,53	0,03; 0,01	0,19; 0,18	0,51; 0,50	0,74; 0,74	0,35; 0,32	0,35; 0,33
s6	0,57; 0,54	0,65; 0,59	0,23; 0,14	0,31; 0,26	0,65; 0,63	0,80; 0,78	0,48; 0,44	0,44; 0,41
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,54; 0,53	0,75; 0,75	0,35; 0,34	0,35; 0,34	0,55; 0,54	0,75; 0,75	0,35; 0,34	0,34; 0,34
s2	0,55; 0,53	0,75; 0,75	0,35; 0,34	0,35; 0,34	0,56; 0,54	0,75; 0,75	0,36; 0,34	0,35; 0,34
s3	0,54; 0,52	0,75; 0,75	0,35; 0,34	0,35; 0,34	0,55; 0,53	0,75; 0,75	0,35; 0,34	0,35; 0,34
s4	0,53; 0,52	0,75; 0,75	0,35; 0,34	0,35; 0,34	0,54; 0,53	0,75; 0,75	0,35; 0,34	0,35; 0,34
s5	0,51; 0,50	0,74; 0,73	0,31; 0,29	0,32; 0,31	0,54; 0,53	0,75; 0,75	0,33; 0,32	0,33; 0,33
s6	0,63; 0,53	0,80; 0,78	0,49; 0,46	0,46; 0,44	0,69; **0,69**	0,81; **0,81**	0,51; 0,49	0,46; 0,45
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,60; 0,55	0,77; 0,76	0,44; 0,40	0,42; 0,39	minimum			
s2	0,60; 0,54	0,78; 0,76	0,44; 0,42	0,44; 0,41	0,32; 0,30	0,51; 0,47	0,02; 0,00	0,19; 0,18
s3	0,60; 0,54	0,78; 0,76	0,44; 0,41	0,42; 0,40	maximum			
s4	0,61; 0,55	0,78; 0,76	0,44; 0,41	0,42; 0,40	0,71; 0,69	0,82; 0,81	0,52; 0,50	0,47; 0,46
s5	0,60; 0,53	0,77; 0,76	0,44; 0,41	0,42; 0,40	rozpětí (maximum - minimum)			
s6	0,60; 0,53	0,77; 0,76	0,44; 0,41	0,42; 0,40	0,39; 0,39	0,31; 0,34	0,50; 0,50	0,28; 0,28

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.14 Ilustrace úspěšnosti nejlepšího shlukovacího postupu a2+s6 pro scénář 4. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.

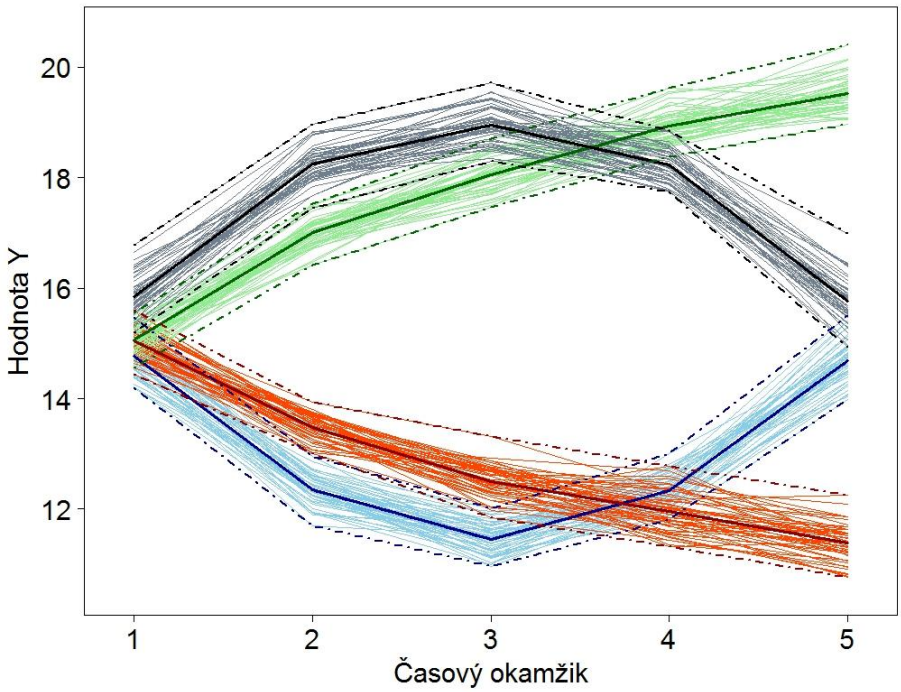
Scénář 5

U scénáře 5, s minimální vnitroobjektovou variabilitou, čtyřmi shluky s nelineárním průběhem, dosáhlo 100% úspěšnosti více shlukovacích algoritmů. Maximální úspěšnosti dosáhly metody a2, a4, a5 a a6 a kombinace s1, s2, s3 a s6. Nejúspěšnější kombinací pro tento scénář je s3, která v kombinaci s nejvíce algoritmy (a2, a4, a6) dosáhla 100% výsledek. Více v tabulce 7.7. Obrázek 7.15 porovnává reprezentanty skutečných tříd (TRUE) a odhadnutých shluků (PRED) pro maximálně úspěšné postupy. Na obrázku je vidět, že jsou očekávané třídy a odhadnuté shluky totožné.

Tabulka 7.7 Hodnoticí kritéria pro scénář 5

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,79; 0,77	0,88; 0,87	0,68; 0,67	0,62; 0,60	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s2	0,79; 0,77	0,88; 0,87	0,68; 0,67	0,61; 0,60	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s3	0,75; 0,71	0,85; 0,82	0,61; 0,55	0,56; 0,50	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s4	0,75; 0,75	0,87; 0,87	0,65; 0,67	0,59; 0,60	0,76; 0,76	0,88; 0,87	0,67; 0,67	0,61; 0,60
s5	0,74; 0,74	0,86; 0,85	0,63; 0,60	0,57; 0,54	0,75; 0,75	0,86; 0,86	0,64; 0,63	0,58; 0,57
s6	0,99; 0,98	0,99; 0,98	0,97; 0,95	0,95; 0,93	0,99; 0,99	0,99; 0,99	0,97; 0,96	0,96; 0,94
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,39; 0,35	0,55; 0,52	0,08; 0,04	0,23; 0,22	0,98; 0,98	0,99; 0,98	0,96; 0,95	0,94; 0,93
s2	0,43; 0,39	0,62; 0,59	0,13; 0,08	0,24; 0,22	**1,00; 1,00**	**1,00; 1,00**	0,99; **1,00**	0,99; **1,00**
s3	0,47; 0,41	0,66; 0,62	0,19; 0,10	0,27; 0,23	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s4	0,40; 0,35	0,61; 0,57	0,12; 0,06	0,24; 0,22	0,76; 0,75	0,87; 0,87	0,66; 0,66	0,60; 0,60
s5	0,38; 0,34	0,58; 0,55	0,08; 0,03	0,22; 0,20	0,73; 0,73	0,86; 0,87	0,68; 0,69	0,63; 0,64
s6	0,90; 0,76	0,94; 0,88	0,86; 0,71	0,83; 0,66	0,99; 0,98	0,99; 0,98	0,97; 0,95	0,95; 0,93
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,97; **1,00**	0,98; **1,00**	0,95; **1,00**	0,94; **1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s2	**1,00; 1,00**	**1,00; 1,00**	0,99; **1,00**	0,99; **1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s3	**1,00; 1,00**	**1,00; 1,00**	0,99; **1,00**	0,99; **1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
s4	0,76; 0,75	0,87; 0,87	0,67; 0,66	0,61; 0,60	0,76; 0,75	0,87; 0,87	0,67; 0,67	0,61; 0,60
s5	0,76; 0,75	0,87; 0,87	0,67; 0,66	0,60; 0,59	0,76; 0,75	0,87; 0,87	0,66; 0,65	0,60; 0,58
s6	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,81; 0,66	0,91; 0,85	0,8; 0,63	0,77; 0,59	minimum			
s2	0,82; 0,66	0,92; 0,85	0,81; 0,63	0,78; 0,59	0,38; 0,34	0,55; 0,52	0,08; 0,03	0,22; 0,20
s3	0,86; 0,67	0,94; 0,85	0,85; 0,64	0,83; 0,59	maximum			
s4	0,85; 0,66	0,93; 0,85	0,84; 0,63	0,82; 0,59	1,00; 1,00	1,00; 1,00	1,00; 1,00	1,00; 1,00
s5	0,8; 0,66	0,91; 0,85	0,79; 0,63	0,76; 0,59	rozpětí (maximum - minimum)			
s6	0,82; 0,66	0,92; 0,85	0,81; 0,63	0,79; 0,59	0,62; 0,66	0,45; 0,48	0,92; 0,97	0,78; 0,80

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.15 Ilustrace úspěšnosti nejlepších shlukovacích postupů pro scénář 5. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.

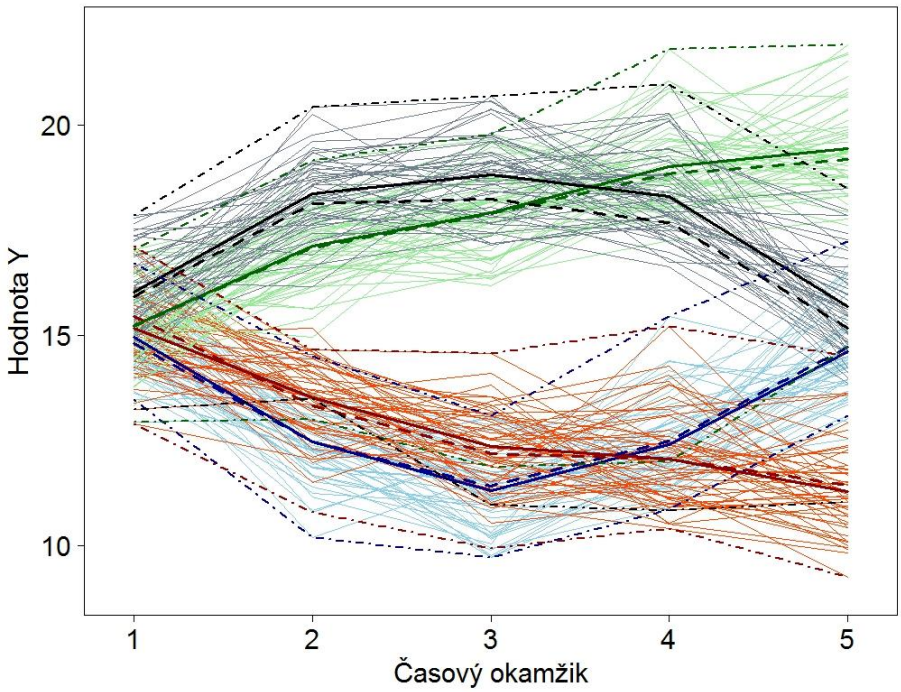
Scénář 6

Ve scénáři 6 pochází trajektorie ze stejného rozdělení jako scénář 5, ale s relativně velkou vnitroobjektovou variabilitou. Maximální klasifikace dosáhla množina proměnných s6 s metodou a6 ve všech 4 hodnoticích kritérií. Více v tabulce 7.8. Obrázek 7.16 porovnává reprezentanty skutečných tříd (TRUE) a odhadnutých shluků (PRED) pro nejlepší postup (a6 s s6). Na obrázku je vidět, že průměrné trajektorie očekávaných tříd a odhadnutých shluků jsou si blízké. Minima a maxima odhadnutých shluků nekopírují minimální a maximální trajektorie očekávaných tříd. Z toho lze usuzovat, že došlo k chybné klasifikaci v krajních případech.

Tabulka 7.8 Hodnoticí kritéria pro scénář 6

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,46; 0,41	0,61; 0,57	0,15; 0,08	0,26; 0,23	0,59; 0,55	0,73; 0,71	0,35; 0,30	0,37; 0,32
s2	0,53; 0,47	0,68; 0,64	0,23; 0,17	0,30; 0,26	0,77; 0,71	0,82; 0,78	0,53; 0,45	0,49; 0,42
s3	0,41; 0,37	0,59; 0,56	0,09; 0,06	0,23; 0,21	0,72; 0,68	0,79; 0,76	0,45; 0,40	0,43; 0,39
s4	0,51; 0,47	0,67; 0,65	0,23; 0,18	0,30; 0,26	0,62; 0,60	0,76; 0,75	0,38; 0,34	0,37; 0,34
s5	0,41; 0,39	0,57; 0,54	0,08; 0,05	0,23; 0,21	0,45; 0,40	0,65; 0,63	0,17; 0,11	0,26; 0,22
s6	0,62; 0,58	0,76; 0,75	0,43; 0,37	0,42; 0,37	0,72; 0,65	0,80; 0,79	0,52; 0,49	0,49; 0,46
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,38; 0,35	0,51; 0,46	0,07; 0,03	0,23; 0,22	0,69; 0,64	0,80; 0,79	0,52; 0,48	0,49; 0,46
s2	0,38; 0,35	0,52; 0,48	0,06; 0,03	0,23; 0,22	0,86; 0,83	0,87; 0,85	0,67; 0,60	0,62; 0,55
s3	0,42; 0,38	0,59; 0,56	0,11; 0,07	0,23; 0,22	0,81; 0,78	0,83; 0,8	0,57; 0,50	0,52; 0,46
s4	0,35; 0,32	0,51; 0,46	0,05; 0,01	0,23; 0,21	0,65; 0,62	0,77; 0,74	0,41; 0,34	0,40; 0,35
s5	0,40; 0,36	0,54; 0,5	0,08; 0,04	0,24; 0,22	0,53; 0,50	0,65; 0,62	0,27; 0,22	0,33; 0,31
s6	0,61; 0,58	0,76; 0,75	0,42; 0,39	0,41; 0,38	0,73; 0,71	0,80; 0,79	0,52; 0,49	0,49; 0,46
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,68; 0,61	0,78; 0,76	0,43; 0,36	0,41; 0,36	0,77; 0,75	0,82; 0,80	0,53; 0,47	0,48; 0,44
s2	0,70; 0,62	0,79; 0,76	0,45; 0,37	0,43; 0,36	0,79; 0,75	0,83; 0,80	0,55; 0,48	0,50; 0,44
s3	0,65; 0,53	0,77; 0,73	0,40; 0,30	0,39; 0,32	0,74; 0,72	0,80; 0,77	0,48; 0,41	0,44; 0,39
s4	0,56; 0,53	0,73; 0,72	0,30; 0,26	0,32; 0,29	0,58; 0,55	0,75; 0,73	0,33; 0,30	0,34; 0,32
s5	0,54; 0,52	0,71; 0,69	0,26; 0,22	0,30; 0,27	0,56; 0,53	0,73; 0,71	0,29; 0,25	0,31; 0,29
s6	0,77; 0,70	0,85; 0,80	0,6; 0,49	0,55; 0,45	**0,89; 0,87**	**0,90; 0,88**	**0,73; 0,68**	**0,66; 0,62**
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,80; 0,65	0,87; 0,83	0,68; 0,58	0,62; 0,54	minimum			
s2	0,80; 0,67	0,87; 0,83	0,68; 0,58	0,62; 0,54	0,35; 0,32	0,51; 0,46	0,05; 0,01	0,23; 0,21
s3	0,79; 0,65	0,87; 0,82	0,68; 0,57	0,62; 0,53	maximum			
s4	0,79; 0,65	0,87; 0,82	0,67; 0,56	0,62; 0,52	0,89; 0,87	0,90; 0,88	0,73; 0,68	0,66; 0,62
s5	0,80; 0,67	0,88; 0,83	0,68; 0,59	0,63; 0,55	rozpětí (maximum - minimum)			
s6	0,79; 0,65	0,87; 0,82	0,67; 0,58	0,62; 0,53	0,54; 0,55	0,39; 0,42	0,68; 0,67	0,43; 0,41

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.16 Ilustrace úspěšnosti nejlepšího postupu (a6+s6) pro scénář 6. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžovo-červená křivka, objekty z c = 3 zelená a objekty z c = 4 šedá.

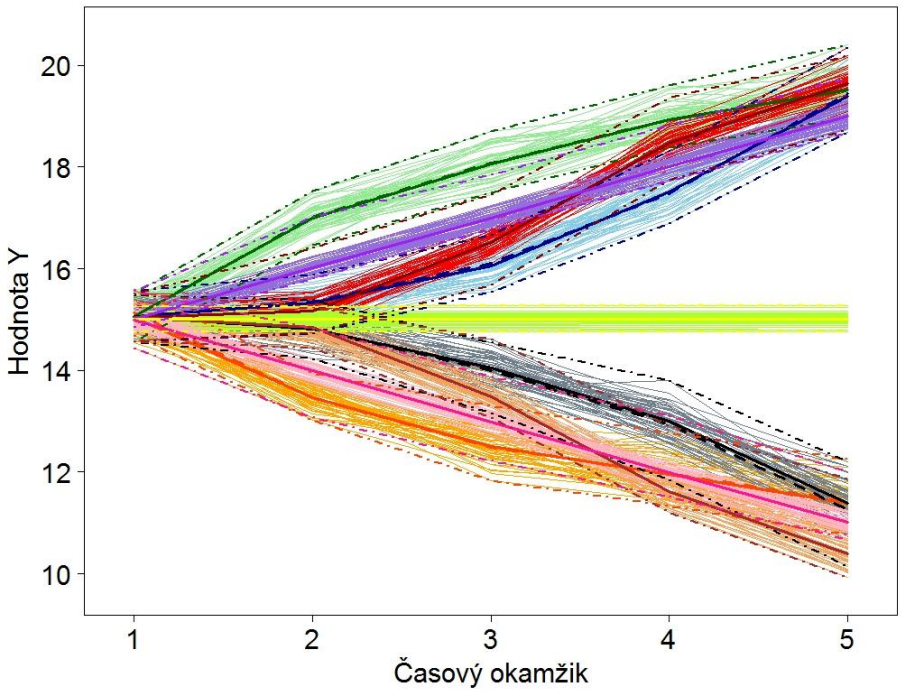
Scénář 7

Scénář 7 simuluje 9 shluků s lineárním (3krát) a nelineárním vývojem (6krát) s minimální vnitroobjektovou variabilitou. Maximální úspěšnosti dosáhla kombinace proměnných s3 ($\overline{diff}_i$, sd($diff_i$), $\overline{|diff_i|}$, sd($|diff_i|$) a $\max \triangleleft$) s metodou *C*-průměrů (a7) a Wardovou metodou s Mahalanobisovou vzdáleností (a4). Nejúspěšnější metodou je *C*-průměrů (a7). Pro všechny množiny (s1–s6) dosáhla podobných výsledků jako s s3. Více v tabulce 7.9. Obrázek 7.17 porovnává reprezentanty skutečných tříd (TRUE) a odhadnutých shluků (PRED) pro nejlepší postup (a7 s s3). Na obrázku je vidět, že průměrné trajektorie očekávaných tříd a odhadnutých shluků jsou si blízké. Minima a maxima odhadnutých shluků nekopírují minimální a maximální trajektorie očekávaných tříd. Z toho lze usuzovat, že došlo k chybné klasifikaci v malé míře.

Tabulka 7.9 Hodnoticí kritéria pro scénář 7

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,63; 0,61	0,90; 0,89	0,53; 0,51	0,42; 0,40	0,68; 0,66	0,91; 0,90	0,60; 0,55	0,48; 0,43
s2	0,64; 0,63	0,90; 0,89	0,54; 0,53	0,43; 0,41	0,74; 0,70	0,93; 0,92	0,66; 0,63	0,54; 0,51
s3	0,60; 0,55	0,89; 0,88	0,51; 0,49	0,41; 0,38	0,73; 0,69	0,93; 0,92	0,66; 0,63	0,54; 0,50
s4	0,64; 0,63	0,90; 0,90	0,54; 0,52	0,42; 0,41	0,67; 0,66	0,91; 0,91	0,57; 0,55	0,44; 0,43
s5	0,62; 0,61	0,90; 0,89	0,53; 0,51	0,41; 0,40	0,63; 0,62	0,90; 0,90	0,54; 0,52	0,42; 0,41
s6	0,57; 0,55	0,89; 0,88	0,57; 0,54	0,46; 0,43	0,75; 0,73	0,94; 0,93	0,72; 0,69	0,60; 0,57
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,28; 0,26	0,57; 0,50	0,10; 0,05	0,15; 0,12	0,64; 0,61	0,91; 0,89	0,59; 0,56	0,48; 0,45
s2	0,34; 0,30	0,67; 0,64	0,18; 0,12	0,19; 0,15	0,76; 0,74	0,94; 0,93	0,71; 0,67	0,59; 0,55
s3	0,42; 0,40	0,78; 0,75	0,32; 0,27	0,27; 0,24	0,78; **0,77**	**0,95; 0,94**	0,75; **0,75**	0,64; **0,63**
s4	0,35; 0,31	0,75; 0,73	0,24; 0,21	0,22; 0,20	0,62; 0,61	0,90; 0,89	0,55; 0,53	0,43; 0,42
s5	0,30; 0,27	0,66; 0,62	0,17; 0,14	0,19; 0,17	0,57; 0,56	0,87; 0,86	0,50; 0,50	0,40; 0,40
s6	0,56; 0,55	0,89; 0,88	0,54; 0,52	0,43; 0,41	0,63; 0,57	0,91; 0,90	0,60; 0,57	0,49; 0,46
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,63; 0,62	0,90; 0,89	0,54; 0,52	0,43; 0,41	0,68; 0,66	0,92; 0,91	0,62; 0,59	0,50; 0,47
s2	0,64; 0,62	0,90; 0,90	0,55; 0,54	0,44; 0,42	0,72; 0,70	0,93; 0,93	0,69; 0,67	0,57; 0,55
s3	0,61; 0,57	0,89; 0,88	0,53; 0,51	0,42; 0,40	0,72; 0,71	0,93; 0,93	0,70; 0,69	0,58; 0,57
s4	0,62; 0,60	0,89; 0,89	0,53; 0,52	0,41; 0,40	0,64; 0,63	0,90; 0,90	0,55; 0,53	0,43; 0,42
s5	0,62; 0,60	0,90; 0,89	0,52; 0,51	0,41; 0,40	0,64; 0,63	0,90; 0,90	0,54; 0,53	0,43; 0,42
s6	0,58; 0,56	0,90; 0,89	0,57; 0,55	0,46; 0,44	0,63; 0,57	0,91; 0,90	0,63; 0,58	0,51; 0,46
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,77; 0,69	0,94; 0,93	0,74; 0,69	0,63; 0,58	minimum			
s2	0,75; 0,67	0,94; 0,93	0,73; 0,68	0,62; 0,56	0,28; 0,26	0,57; 0,50	0,10; 0,05	0,15; 0,12
s3	**0,79**; 0,71	**0,95; 0,94**	**0,76**; 0,72	**0,65**; 0,60	maximum			
s4	0,76; 0,69	0,94; 0,93	0,74; 0,69	0,63; 0,57	0,79; 0,77	0,95; 0,94	0,76; 0,75	0,65; 0,63
s5	0,76; 0,68	0,94; 0,93	0,73; 0,69	0,63; 0,57	rozpětí (maximum - minimum)			
s6	0,77; 0,69	0,94; 0,93	0,74; 0,69	0,64; 0,57	0,51; 0,51	0,38; 0,44	0,66; 0,7	0,50; 0,51

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.17 Ilustrace úspěšnosti nejlepšího postupu (a7+s3) pro scénář 7. Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžová křivka, objekty z c = 3 zelená, objekty z c = 4 šedá, objekty z c = 5 červená, objekty z c = 6 hnědá, objekty z c = 7 fialová, objekty z c = 8 zeleno–žlutá a objekty z c = 9 růžová.

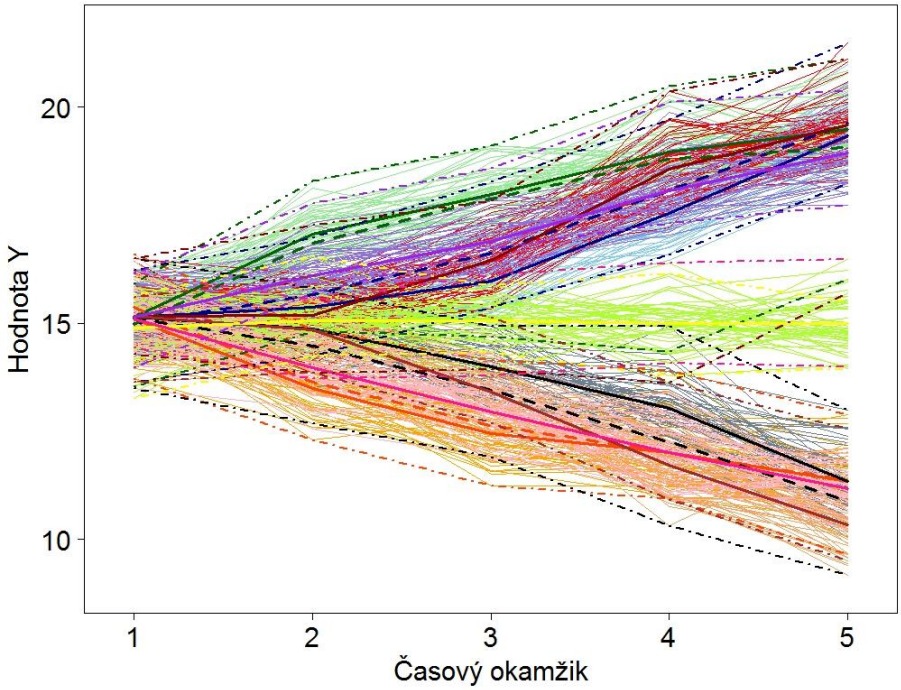
Scénář 8

Scénář 8 komplikuje scénář 7 velkou vnitroobjektovou variabilitou. V tomto scénáři dochází k výraznému pronikání trajektorií do jiných shluků. Tento šum vede obecně u všech postupů ke snížení klasifikační schopnosti. Maximální úspěšnosti dosáhla kombinace s6 po aplikaci metody a2. Podobných výsledků dosáhla s6 v kombinaci s metodou a4 a a6. Více v tabulce 7.10. Obrázek 7.18 porovnává reprezentanty skutečných tříd (TRUE) a odhadnutých shluků (PRED) pro nejlepší postup (a2 s s6). Na obrázku je vidět, že průměrné trajektorie očekávaných tříd a odhadnutých shluků jsou identické pro 5 tříd. Drobné odchylky v průměrech jsou mezi třídou a shlukem u c_2 a l_2 (oranžová), c_3 a l_3 (zelená). Větší odchylky lze vnímat u c_1 a l_1 (modrá), c_4 a l_4 (černá). Minima a maxima odhadnutých shluků nekopírují minimální a maximální trajektorie očekávaných tříd.

Tabulka 7.10 Hodnoticí kritéria pro scénář 8

	aglom., úplná, euklid. (a1)				aglom., Ward., euklid. (a2)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,32; 0,31	0,79; 0,78	0,18; 0,15	0,17; 0,16	0,37; 0,35	0,84; 0,83	0,22; 0,20	0,19; 0,17
s2	0,32; 0,30	0,78; 0,77	0,18; 0,16	0,17; 0,16	0,38; 0,37	0,84; 0,83	0,23; 0,22	0,19; 0,18
s3	0,27; 0,25	0,74; 0,73	0,10; 0,06	0,13; 0,11	0,36; 0,35	0,83; 0,83	0,21; 0,20	0,18; 0,17
s4	0,31; 0,29	0,78; 0,77	0,16; 0,13	0,16; 0,14	0,36; 0,34	0,84; 0,83	0,22; 0,21	0,19; 0,18
s5	0,28; 0,27	0,77; 0,75	0,11; 0,09	0,13; 0,12	0,35; 0,33	0,83; 0,83	0,20; 0,18	0,17; 0,16
s6	0,41; 0,39	0,84; 0,83	0,28; 0,27	0,23; 0,22	**0,44; 0,43**	**0,86; 0,85**	**0,31; 0,31**	**0,25; 0,24**
	aglom., úplná, Mahalanobis (a3)				aglom., Ward, Mahalanobis (a4)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,27; 0,25	0,72; 0,69	0,08; 0,06	0,12; 0,11	0,41; 0,39	0,84; 0,84	0,26; 0,24	0,21; 0,20
s2	0,27; 0,25	0,71; 0,69	0,08; 0,05	0,12; 0,11	0,41; 0,39	0,84; 0,83	0,26; 0,25	0,21; 0,20
s3	0,29; 0,27	0,75; 0,73	0,11; 0,08	0,14; 0,12	0,41; 0,39	0,84; 0,83	0,25; 0,23	0,21; 0,20
s4	0,26; 0,23	0,75; 0,73	0,09; 0,06	0,12; 0,11	0,37; 0,36	0,84; 0,83	0,24; 0,22	0,20; 0,19
s5	0,28; 0,26	0,77; 0,76	0,12; 0,09	0,14; 0,12	0,36; 0,35	0,84; 0,83	0,22; 0,21	0,19; 0,18
s6	0,39; 0,38	0,84; 0,83	0,26; 0,25	0,22; 0,21	0,42; 0,41	0,85; 0,85	0,30; 0,29	0,24; 0,23
	aglom., úplná, Pearson (a5)				aglom., Ward, Pearson (a6)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,37; 0,36	0,83; 0,82	0,23; 0,22	0,19; 0,19	0,38; 0,36	0,84; 0,84	0,23; 0,22	0,19; 0,19
s2	0,37; 0,35	0,82; 0,82	0,23; 0,22	0,20; 0,19	0,39; 0,38	0,84; 0,84	0,24; 0,23	0,20; 0,19
s3	0,36; 0,35	0,83; 0,82	0,22; 0,20	0,19; 0,18	0,39; 0,37	0,84; 0,84	0,23; 0,22	0,19; 0,18
s4	0,36; 0,35	0,83; 0,82	0,23; 0,22	0,19; 0,19	0,38; 0,37	0,84; 0,84	0,23; 0,22	0,19; 0,19
s5	0,34; 0,33	0,83; 0,82	0,20; 0,19	0,18; 0,17	0,36; 0,35	0,84; 0,83	0,22; 0,21	0,18; 0,18
s6	0,40; 0,38	0,83; 0,82	0,28; 0,27	0,23; 0,22	0,42; 0,40	0,84; 0,84	0,30; 0,29	0,24; 0,23
	metoda *C*-průměrů (a7)				Souhrnné statistiky			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s1	0,42; 0,40	0,84; 0,84	0,26; 0,25	0,21; 0,21	minimum			
s2	0,42; 0,40	0,84; 0,84	0,26; 0,25	0,21; 0,21	0,26; 0,23	0,71; 0,69	0,08; 0,05	0,12; 0,11
s3	0,42; 0,40	0,84; 0,84	0,26; 0,25	0,21; 0,21	maximum			
s4	0,42; 0,39	0,84; 0,84	0,26; 0,25	0,21; 0,20	0,44; 0,43	0,86; 0,85	0,31; 0,31	0,25; 0,24
s5	0,42; 0,40	0,84; 0,84	0,26; 0,25	0,21; 0,21	rozpětí (maximum - minimum)			
s6	0,42; 0,40	0,84; 0,84	0,26; 0,25	0,21; 0,21	0,18; 0,20	0,15; 0,16	0,23; 0,26	0,13; 0,13

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil), které korespondují s hodnotami v pravém dolním bloku tabulky.



Obrázek 7.18 Ilustrace úspěšnosti nejlepšího postupu (a2+s6) pro scénář 8.
Trajektorie objektů z c = 1 je znázorněna modře, objekty z c = 2 znázorňuje oranžová křivka, objekty z c = 3 zelená, objekty z c = 4 šedá, objekty z c = 5 červená, objekty z c = 6 hnědá, objekty z c = 7 fialová, objekty z c = 8 zeleno-žlutá a objekty z c = 9 růžová.

Shrnutí testování shlukovacích postupů

Z 8 testovaných scénářů bylo dosaženo maximální úspěšnosti správné klasifikace do očekávaných tříd u scénářů 1 a 5. Jedná se o scénáře s minimální vnitroobjektovou variabilitou. U scénáře 1 se 3 shluky definovanými 3 lineárními funkcemi byla maximální úspěšnost očekávána. Scénář 5 obsahuje 4 shluky nelineárního tvaru rovněž s minimální vnitroobjektovou variabilitou. S nárůstem počtu shluků a zvětšením vnitroobjektové variability u ostatních scénářů se zvyšuje heterogenita mezi shluky, která vede k poklesu přesnosti klasifikace.

V předchozí části „Výsledky shlukování“ tohoto oddílu byly individuálně hodnoceny výsledky shlukovacích postupů v každém scénáři. Pro účely simultánního vyhodnocení a výběru nejlepšího shlukovacího postupu ve všech osmi scénářích (S = 8, kde S značí počet scénářů) jsem navrhl následující dvoubodový skórovací systém založený na průměrech a dolních kvartilech 4 hodnoticích kritérií, viz tabulky 7.3 – 7.10. V každém scénáři jsou každému (ze 42) shlukovacímu postupu přiděleny dvě hodnoty bodů.

První počet bodů, $body1_{pr,sp,s}$, se spočítá pro sp-tý shlukovací postup, kde sp = 1, 2, …, 42, v s-tém scénáři, kde s = 1, 2, …, 8, z průměrných hodnot (pr) čtyř hodnoticích kritérií následovně

$$\begin{aligned} body1_{pr,sp,s} &= body\left(CCR_{pr,sp,s}\right)+body\left(RI_{pr,sp,s}\right)+body\left(RIadj_{pr,sp,s}\right)+body\left(Jac_{pr,sp,s}\right) \\ &= \frac{CCR_{pr,sp,s}-\min\left(CCR_{pr,s}\right)}{\max\left(CCR_{pr,s}\right)-\min\left(CCR_{pr,s}\right)}+\frac{RI_{pr,sp,s}-\min\left(RI_{pr,s}\right)}{\max\left(RI_{pr,s}\right)-\min\left(RI_{pr,s}\right)}+ \\ &+\frac{RIadj_{pr,sp,s}-\min\left(RIadj_{pr,s}\right)}{\max\left(RIadj_{pr,s}\right)-\min\left(RIadj_{pr,s}\right)}+\frac{Jac_{pr,sp,s}-\min\left(Jac_{pr,s}\right)}{\max\left(Jac_{pr,s}\right)-\min\left(Jac_{pr,s}\right)}, \end{aligned} \tag{7.12}$$

$CCR_{pr,sp,s}$ je průměrná hodnota CCR v s-tém scénáři pro sp-tý shlukovací postup. Minimální průměrná hodnota CCR ze 42 shlukovacích postupů v s-tém scénáři se značí min($CCR_{pr,s}$) a maximální max($CCR_{pr,s}$). Analogicky se počítají sčítance za zbývající tři hodnoticí kritéria (RI, $RIadj$, Jac) pro sp-tý postup. Celkový počet bodů $body1_{pr,sp,s}$ za průměrné hodnoty hodnoticích kritérií je konečný součet čtyř sčítanců. Ve jmenovateli jednotlivých sčítanců je variační rozpětí průměrných hodnot pro s-tý scénář. Počet dílčích bodů: $body(CCR_{pr,sp,s})$, $body(RI_{pr,sp,s})$, $body(RIadj_{pr,sp,s})$, $body(Jac_{pr,sp,s})$ se nachází v intervalu $\langle 0;1 \rangle$. Nula dílčích bodů, např. $body(CCR_{pr,sp,s}) = 0$, obdrží sp-tý postup, který dosáhl nejhorší průměrné hodnoty CCR ze 42 uvažovaných hodnot, tj. v případě, že $CCR_{pr,sp,s} = \min(CCR_{pr,s})$. Naopak jeden dílčí bod, např.

$CCR_{pr,sp,s}$ = 1, obdrží *sp*-tý postup, který dosáhl nejvyšší průměrné hodnoty *CCR* ze 42 uvažovaných hodnot, tj. v případě, že $CCR_{pr,sp,s} = \max(CCR_{pr,s})$.

Druhý počet bodů, $body2_{q25,sp,s}$, se spočítá pro *sp*-tý shlukovací postup z hodnot dolních kvartilů (*q*25) čtyř hodnoticích kritérií stejně jako $body1_{pr,sp,s}$, tedy

$$\begin{aligned} body2_{q25,sp,s} &= body\left(CCR_{q25,sp,s}\right) + body\left(RI_{q25,sp,s}\right) + \\ &+ body\left(RIadj_{q25,sp,s}\right) + body\left(Jac_{q25,sp,s}\right) = \\ &= \frac{CCR_{q25,sp,s} - \min\left(CCR_{q25,s}\right)}{\max\left(CCR_{q25,s}\right) - \min\left(CCR_{q25,s}\right)} + \frac{RI_{q25,sp,s} - \min\left(RI_{q25,s}\right)}{\max\left(RI_{q25,s}\right) - \min\left(RI_{q25,s}\right)} + \\ &+ \frac{RIadj_{q25,sp,s} - \min\left(RIadj_{q25,s}\right)}{\max\left(RIadj_{q25,s}\right) - \min\left(RIadj_{q25,s}\right)} + \frac{Jac_{q25,sp,s} - \min\left(Jac_{q25,s}\right)}{\max\left(Jac_{q25,s}\right) - \min\left(Jac_{q25,s}\right)}. \end{aligned} \tag{7.13}$$

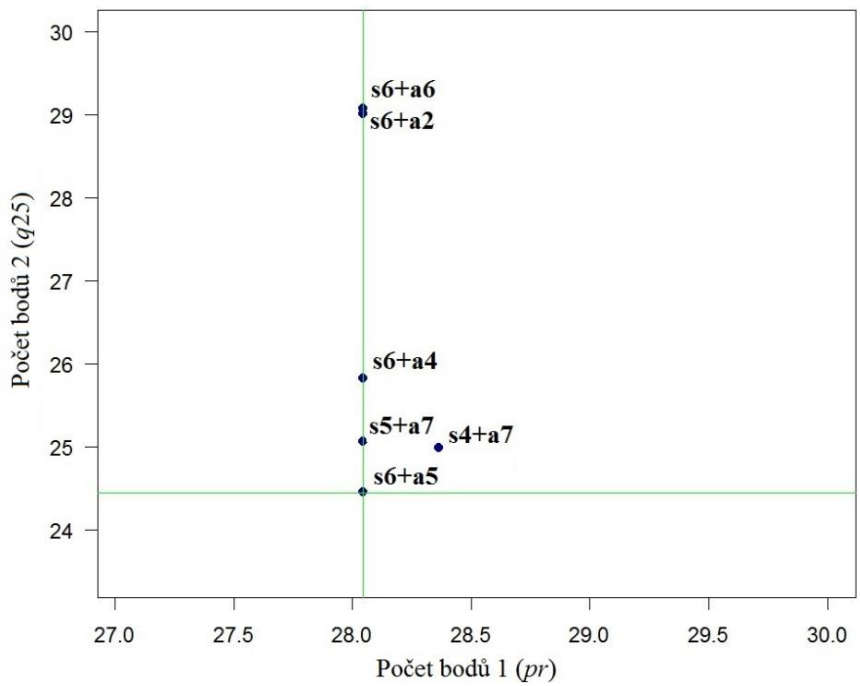
Celkový počet bodů 1 pro *sp*-tý shlukovací postup za všechny scénáře je dán jejich součtem $\sum_{s=1}^{8} body1_{pr,sp,s}$ a celkový počet bodů 2 pro *sp*-tý shlukovací postup za všechny scénáře je dán jejich součtem $\sum_{s=1}^{8} body2_{q25,sp,s}$.

Všechny hodnoty k výpočtu 7.12 a 7.13 se nachází v tabulkách 7.3 – 7.10. Minima ($\min(CCR_{pr,s})$, $\min(CCR_{q25,s})$), maxima ($\max(CCR_{pr,s})$, $\max(CCR_{q25,s})$) a jejich rozpětí pro daný scénář se nalézají v pravém dolním rohu těchto tabulek. Celkový počet získaných bodů 1 a 2 za všechny shlukovací scénáře pro každý shlukovací postup shrnuje tabulka 7.12. Postupy jsou seřazeny dle součtu bodů 1 a 2. Tabulka P3 (v příloze) obsahuje počet získaných bodů v jednotlivých scénářích. Obrázek 7.19 zobrazuje počet bodů 1 a 2 pro shlukovací postupy, které dosáhly v obou skupinách bodů vysoké hodnoty - větší než horní kvartil. Z tabulky 7.12 a obrázku 7.19 se jeví jako nejvhodnější pro shlukování trajektorií pro vybalancovaný návrh dva postupy s6+a6 a s6+a2.

Lze shrnout, že nejlépe dopadla Wardova metoda s mírou nepodobnosti založené na Pearsonovu korelačním koeficientu (a6) a Wardova metoda s euklidovskou vzdáleností (a2) v kombinaci se shlukovacími proměnnými s6: max $\triangleleft$, $\overline{k_i}$ a %pos.

Tabulka 7.11 Počet získaných bodů za všechny scénáře

Shlukovací postup	Součet bodů 1 (*pr*)	Součet bodů 2 (*q*25)	Součet bodů celkem	Shlukovací postup	Součet bodů 1 (*pr*)	Součet bodů 2 (*q*25)	Součet bodů celkem
s6+a6	28,04	29,08	57,12	s4+a2	24,64	22,20	46,84
s6+a2	28,04	29,00	57,05	s4+a4	24,64	21,35	45,99
s6+a4	28,04	25,82	53,87	s3+a5	26,02	19,27	45,29
s4+a7	28,36	24,99	53,35	s4+a5	24,64	20,61	45,25
s5+a7	28,04	25,06	53,11	s4+a6	22,09	22,20	44,29
s6+a5	28,04	24,45	52,50	s6+a3	28,04	14,48	42,52
s2+a6	26,02	25,81	51,83	s1+a6	12,69	25,01	37,70
s2+a4	25,61	26,14	51,75	s5+a1	28,36	9,15	37,51
s2+a2	25,61	25,13	50,74	s1+a2	12,69	23,36	36,05
s3+a4	26,02	24,47	50,50	s4+a1	24,64	11,03	35,67
s6+a7	27,82	22,57	50,38	s1+a5	12,69	22,81	35,50
s6+a1	28,04	22,15	50,20	s1+a7	12,69	22,67	35,37
s2+a7	26,02	24,17	50,19	s1+a4	12,69	22,00	34,69
s3+a2	26,02	23,22	49,24	s3+a1	26,02	7,66	33,68
s3+a6	24,64	24,42	49,06	s5+a3	28,36	2,25	30,61
s2+a5	25,61	22,97	48,58	s3+a3	26,02	3,59	29,61
s5+a6	28,36	20,08	48,45	s2+a3	25,61	1,79	27,40
s3+a7	24,64	22,99	47,63	s4+a3	24,64	2,10	26,74
s5+a5	28,36	18,87	47,24	s1+a1	12,69	11,46	24,15
s5+a2	28,36	18,56	46,92	s2+a1	12,51	11,50	24,01
s5+a4	28,36	18,52	46,88	s1+a3	12,69	0,63	13,32



Obrázek 7.19 Počet bodů 1 a 2 pro nejúspěšnější shlukovací scénáře. Zelené čáry značí horní kvartil pro počet bodů 1 a 2.

7.2 Porovnání nejlepší metody a zavedených shlukovacích postupů

V tomto oddílu porovnám shlukovací schopnost nejlepší kombinace mnou navržených proměnných s6 s proměnnými, které používají jiní autoři (D'Urso, 2004; Genolini a Falissard, 2011).

V oddílu 5.2 jsou popsány dynamické fuzzy *C*-modely určené pro shlukování trajektorií (D'Urso, 2004). Tyto modely využívají pro shlukování namísto původních hodnot $\mathbf{x}_{it}$ transformované vektory hodnot $\mathbf{v}_{it}$ a $\mathbf{a}_{it}$, které sdružují informaci o tempu růstu (rychlosti) a zrychlení. Prvky vektoru $\mathbf{v}_{it}$ jsou první absolutní diference mezi měřeními. Prvky $\mathbf{v}_{it}$ vyjadřují sklon lineární funkce pro každý pár 2 sousedních hodnot $[t, t-1]$. Prvky $\mathbf{a}_{it}$ jsou druhé diference původních hodnot a vyjadřují zrychlení (konvexitu/konkávnost) pro každý pár časového intervalu $[t, t-2]$. Tyto charakteristiky (první a druhé diference) jsou porovnány s navrhovanými proměnnými.

Rkový balíček *kml*, Genoliniho a Falissarda (2011) aplikuje metodu *C*-průměrů přímo na původní naměřené hodnoty y_{it} pro *i*-tý objekt v čase *t*. Pracuje tedy s *N* vektory $\mathbf{y}_i = (y_{i1}, \ldots, x_{it}, \ldots, x_{iT})'$. V *kml* se shlukuje pomocí funkce *kml()*, ve které je implementován stejný algoritmus jako výchozí postup ve funkci *kmeans()*, jmenovitě viz Hartigan a Wong

(1979). Původní hodnoty $\mathbf{y}_i$ beru jako další alternativu shlukovacích proměnných a porovnávám jejich shlukovací schopnost s kombinací s6.

Kombinace proměnných navržené různými autory jsou porovnány separátně. Dále je kombinuji dohromady s s6 s cílem posoudit, zda nezískám lepší klasifikační schopnost pro kombinaci různých skupin proměnných. Porovnávané kombinace proměnných jsou:

- s6 – max $\triangleleft$, $\overline{k_i}$ a %pos,
- s7 – $\mathbf{v}_{it}$ a $\mathbf{a}_{it}$,
- s8 – $\mathbf{y}_i$,
- s9 – kombinace s6 a s7,
- s10 – kombinace s6 a s8,
- s11 – kombinace s6, s7 a s8.

Na uvedené znormované hodnoty kombinací proměnných (s6, až s14) aplikuji nejúspěšnější metody z oddílu 7.1 aglomerativního shlukování s Wardovým algoritmem a mírou nepodobnosti založené na Pearsonově korelačním koeficientu (a6) a metodu *C*-průměrů (a7). Metoda a6 dopadla v kombinaci s navrženou kombinaci s6 nejlépe. Metodu *C*-průměrů jsem zařadil, jelikož je doporučována autory D'Ursem (2004) a Genolinim a Falissardem (2011). Shlukovací postup s8+a7 odpovídá postupu aplikovanému v balíčku *kml*. Tabulka 7.12 prezentuje průměry a dolní kvartily hodnoticích kritérií. Uvedené statistiky jsou výsledkem 100 opakování (replikací).

V prvním scénáři dopadla nejlépe kombinace s10 s metodou *C*-průměrů dle všech hodnoticích kritérií vyjma průměru pro CCR. Alespoň v 75 ze 100 dosáhla tato kombinace maximální úspěšnosti (hodnota horního kvartilu je rovna jedné) pro všechny 4 hodnoticí kritéria. S10 je kombinace s6 a s8 ($\mathbf{y}_i$). Kombinace proměnných s6, navržená v kapitole 7, dopadla lépe než kombinace shlukovacích proměnných s7 a s8, které doporučují použít jiní, výše zmiňovaní, autoři. Ve scénáři 2 rovněž dopadla nejlépe kombinace s10 avšak při aplikaci Wardova algoritmu (a6). Samostatně dopadla s6 lépe než s7 a s8 s použitím a6 a srovnatelně při aplikaci a7. Ve scénáři 3 dosáhla s10 s a6 maximálních nebo téměř maximálních (*RIadj* = 0,99; *Jac* = 0,99) výsledků klasifikace. Ve scénáři 4 nejlépe klasifikuje trajektorie do správných shluků s6 s a6. Ve scénáři 5 dosáhla maximálního úspěchu metoda a6 pro všechny kombinace proměnných. Metoda *C*-průměrů nedosáhla pro žádnou kombinaci proměnných maximální úspěšnosti klasifikace. Ve scénáři 6 se umístila nejlépe kombinace s10 s a6. V předposledním scénáři 7 dosáhl nejvyšších průměrných hodnot hodnoticích kritérií a jejich dolních kvartilů shlukovací postup s7 s a6. V posledním scénáři 8 dopadla nejlépe kombinace s6 s a6.

Tabulka 7.12 Porovnání nejúspěšnější kombinace a zavedených kombinací shlukovacích proměnných

Scénář 1	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	**0,95**; 0,94	**0,94**; 0,93	0,86; 0,84	0,84; 0,81	0,86; 0,57	0,91; 0,73	0,83; 0,47	0,84; 0,52
s7	0,80; 0,71	0,82; 0,77	0,60; 0,50	0,59; 0,50	0,88; 0,60	0,92; 0,74	0,85; 0,49	0,86; 0,53
s8	0,84; 0,83	0,84; 0,83	0,65; 0,64	0,63; 0,62	0,87; 0,59	0,92; 0,74	0,84; 0,48	0,86; 0,53
s9	**0,95**; 0,95	**0,94**; 0,93	**0,87**; 0,85	0,84; 0,82	0,86; 0,59	0,92; 0,74	0,83; 0,48	0,85; 0,52
s10	**0,95**; 0,95	**0,94**; 0,93	**0,87**; 0,85	0,84; 0,82	0,90; **1,00**	**0,94**; **1,00**	**0,87**; **1,00**	**0,88**; **1,00**
s11	**0,95**; 0,95	**0,94**; 0,93	**0,87**; 0,85	0,84; 0,82	0,86; 0,58	0,91; 0,73	0,83; 0,47	0,84; 0,52
Scénář 2	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	0,86; 0,85	0,85; 0,84	0,66; 0,64	0,64; 0,62	0,86; 0,71	0,88; 0,77	0,74; 0,50	0,73; 0,52
s7	0,63; 0,59	0,69; 0,67	0,31; 0,26	0,38; 0,34	0,86; 0,71	0,88; 0,77	0,74; 0,51	0,72; 0,52
s8	0,79; 0,77	0,80; 0,78	0,56; 0,52	0,56; 0,53	0,87; 0,71	0,89; 0,78	0,75; 0,52	0,74; 0,52
s9	0,70; 0,64	0,71; 0,66	0,37; 0,26	0,42; 0,35	0,89; 0,87	0,90; 0,87	0,77; 0,71	0,76; 0,69
s10	**0,94; 0,89**	**0,93; 0,88**	**0,85; 0,73**	**0,83; 0,70**	0,88; 0,74	0,90; 0,78	0,77; 0,52	0,76; 0,53
s11	0,82; 0,78	0,82; 0,79	0,60; 0,53	0,59; 0,53	0,87; 0,70	0,89; 0,77	0,75; 0,51	0,74; 0,52
Scénář 3	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	0,79; 0,79	0,88; 0,87	0,70; 0,69	0,64; 0,63	0,79; 0,63	0,90; 0,83	0,75; 0,58	0,72; 0,54
s7	0,98; 0,99	0,98; 0,99	0,96; 0,97	0,94; 0,96	0,77; 0,62	0,89; 0,82	0,73; 0,58	0,70; 0,54
s8	0,59; 0,53	0,76; 0,75	0,39; 0,35	0,38; 0,35	0,78; 0,62	0,90; 0,83	0,75; 0,59	0,71; 0,55
s9	0,99; 0,99	0,99; 0,99	0,97; 0,97	0,96; 0,96	0,81; 0,63	0,91; 0,83	0,77; 0,60	0,74; 0,55
s10	0,68; 0,60	0,82; 0,76	0,55; 0,42	0,52; 0,41	0,78; 0,62	0,89; 0,82	0,74; 0,58	0,71; 0,54
s11	**1,00; 1,00**	**1,00; 1,00**	**0,99; 0,99**	**0,99; 0,99**	0,82; 0,63	0,91; 0,83	0,78; 0,59	0,75; 0,54
Scénář 4	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	**0,69; 0,69**	**0,81; 0,81**	**0,51; 0,49**	**0,46; 0,45**	0,60; 0,53	0,77; 0,76	0,44; 0,41	0,42; 0,40
s7	0,55; 0,53	0,76; 0,75	0,36; 0,34	0,35; 0,34	0,60; 0,52	0,77; 0,76	0,44; 0,41	0,42; 0,40
s8	0,57; 0,54	0,76; 0,75	0,37; 0,35	0,37; 0,35	0,58; 0,51	0,77; 0,76	0,44; 0,41	0,42; 0,40
s9	0,56; 0,54	0,76; 0,75	0,36; 0,35	0,36; 0,34	0,59; 0,53	0,77; 0,76	0,43; 0,41	0,42; 0,40
s10	0,58; 0,52	0,77; 0,76	0,41; 0,35	0,40; 0,35	0,60; 0,52	0,78; 0,76	0,44; 0,41	0,43; 0,40
s11	0,63; 0,59	0,78; 0,76	0,43; 0,39	0,41; 0,38	0,59; 0,52	0,77; 0,76	0,43; 0,41	0,42; 0,40
Scénář 5	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	0,82; 0,66	0,92; 0,85	0,81; 0,63	0,79; 0,59
s7	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	0,80; 0,66	0,91; 0,85	0,78; 0,63	0,75; 0,59
s8	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	0,83; 0,66	0,92; 0,85	0,82; 0,64	0,80; 0,59
s9	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	0,85; 0,66	0,93; 0,85	0,84; 0,63	0,82; 0,59
s10	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	0,86; 0,67	0,94; 0,85	0,85; 0,64	0,83; 0,59
s11	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	**1,00; 1,00**	0,82; 0,66	0,92; 0,85	0,81; 0,63	0,78; 0,59

Scénář 6	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	0,89; 0,87	0,90; 0,88	0,73; 0,68	0,66; 0,62	0,79; 0,65	0,87; 0,82	0,67; 0,58	0,62; 0,53
s7	0,52; 0,50	0,73; 0,72	0,28; 0,26	0,03; 0,29	0,78; 0,64	0,87; 0,82	0,67; 0,58	0,62; 0,54
s8	0,72; 0,67	0,83; 0,80	0,54; 0,47	0,50; 0,43	0,79; 0,63	0,87; 0,82	0,68; 0,58	0,62; 0,54
s9	0,67; 0,60	0,79; 0,75	0,46; 0,35	0,43; 0,35	0,78; 0,64	0,87; 0,82	0,67; 0,58	0,61; 0,54
s10	**0,95; 0,94**	**0,95; 0,94**	**0,87; 0,84**	**0,83; 0,79**	0,80; 0,66	0,87; 0,83	0,68; 0,58	0,62; 0,54
s11	0,81; 0,74	0,87; 0,83	0,66; 0,56	0,61; 0,51	0,77; 0,62	0,86; 0,82	0,66; 0,56	0,61; 0,52
Scénář 7	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	0,63; 0,57	0,91; 0,90	0,63; 0,58	0,51; 0,46	0,77; 0,69	0,94; 0,93	0,74; 0,69	0,64; 0,57
s7	**0,84; 0,80**	**0,95; 0,94**	**0,76; 0,72**	**0,65; 0,60**	0,73; 0,67	0,93; 0,93	0,71; 0,68	0,60; 0,56
s8	0,48; 0,46	0,87; 0,87	0,38; 0,35	0,29; 0,27	0,77; 0,69	0,94; 0,93	0,75; 0,69	0,64; 0,57
s9	0,80; 0,76	0,94; 0,93	0,73; 0,69	0,62; 0,57	0,77; 0,69	0,94; 0,93	0,74; 0,69	0,63; 0,58
s10	0,51; 0,49	0,89; 0,88	0,49; 0,45	0,38; 0,35	0,76; 0,69	0,94; 0,93	0,73; 0,69	0,62; 0,57
s11	0,66; 0,64	0,92; 0,92	0,65; 0,63	0,53; 0,52	0,74; 0,68	0,94; 0,93	0,72; 0,69	0,61; 0,57
Scénář 8	aglom., Ward, Pearson (a6)				*k*-průměrů (a7)			
	CCR	*RI*	*RIadj*	*Jac*	*CCR*	*RI*	*RIadj*	*Jac*
s6	**0,42**; 0,40	**0,84; 0,84**	**0,30; 0,29**	**0,24; 0,23**	**0,42**; 0,40	**0,84; 0,84**	0,26; 0,25	0,21; 0,21
s7	0,34; 0,32	0,83; 0,83	0,17; 0,17	0,16; 0,15	**0,42**; 0,40	**0,84; 0,84**	0,26; 0,25	0,21; 0,20
s8	0,39; 0,37	**0,84**; 0,83	0,25; 0,23	0,20; 0,19	0,41; 0,40	**0,84; 0,84**	0,26; 0,25	0,21; 0,21
s9	0,33; 0,32	0,83; 0,82	0,18; 0,18	0,16; 0,16	**0,42**; 0,40	**0,84; 0,84**	0,26; 0,25	0,21; 0,21
s10	0,39; 0,38	**0,84**; 0,83	0,27; 0,25	0,22; 0,21	0,41; 0,40	**0,84; 0,84**	0,26; 0,25	0,21; 0,20
s11	0,35; 0,33	0,83; 0,82	0,21; 0,19	0,18; 0,17	**0,42; 0,41**	**0,84; 0,84**	0,26; 0,25	0,21; 0,21

Legenda: v tabulce jsou zobrazeny hodnoty: průměr; dolní kvartil. Tučně jsou zvýrazněny maximální hodnoty (průměr; dolní kvartil) pro danou simulaci

Na základě hodnot hodnoticích kritérií v tabulce 7.12 lze porovnat metody a6 a a7 z hlediska jejich robustnosti na použitou kombinaci shlukovacích proměnných. Metoda *C*-průměrů (a6) dosahuje v každém scénáři podobných hodnot hodnoticích kritérií bez ohledu na použitou kombinaci shlukovacích proměnných (s6–s11). Oproti tomu hodnoty hodnoticích kritérií se u Wardovy metody liší v závislosti na použité kombinaci proměnných. Pro některé kombinace proměnných je úspěšnost lepší než u metody *C*-průměrů. Pro jiné kombinace je však horší. Z toho lze usuzovat, že metoda *C*-průměrů je robustnější na použití shlukovacích proměnných, avšak v případě vhodně zvolených shlukovacích proměnných je vhodnější použít Wardovu metodu (a6).

Pořadí shlukovacích proměnných dle úspěšnosti v jednotlivých scénářích se liší, a proto z nich nelze jednoznačně vybrat a doporučit nejvhodnější kombinaci proměnných na základě výsledků v tabulce 7.12. Z tohoto důvodu přistupuji k výběru nejlepší kombinace shlukovacích proměnných dle skórovacího systému zavedeného v oddílu 7.1.2, který ohodnocuje shlukovací

postupy napříč všechmi simulovanými scénáři. Počet získaných bodů 1 a 2 je uveden v tabulce 7.13. Na základě zvoleného hodnoticího systému nejlépe dopadla kombinace s6+a6 (celkem bodů 50,74). Kombinace proměnných s6 ($\max \triangleleft$, $\overline{k_i}$ a %pos), kterou jsem navrhl, při použití Wardovy metody s mírou nepodobnosti založené na Pearsonově korelačním koeficientu překonala v této simulační studii doporučované postupy s7+a7, viz D'Urso (2004) a s8+a7, viz Genolini a Falissard (2011).

Tabulka 7.13 Počet získaných bodů za všechny scénáře

Shlukovací postup	Součet bodů 1 (*pr*)	Součet bodů 2 (*q*25)	Součet bodů celkem
s6+a6	**25,76**	**24,98**	**50,74**
s7+a6	11,85	13,05	24,89
s8+a6	11,72	11,53	23,25
s9+a6	17,49	15,68	33,17
s10+a6	22,09	19,08	41,17
s11+a6	21,97	19,90	41,87
s6+a7	20,01	12,92	32,94
s7+a7	19,24	12,75	31,99
s8+a7	20,27	13,03	33,30
s9+a7	20,91	15,01	35,91
s10+a7	21,86	17,00	38,87
s11+a7	19,82	12,85	32,68

8. Vyhodnocení zpřesnění odhadů regresních modelů shlukovou analýzou

Různé mikropanelové datové soubory, nasimulované v kapitole 7, dokázala nejlépe klasifikovat do správných shluků kombinace proměnných $\max \triangleleft$, $\overline{k_i}$ a %pos, kterou jsem navrhl při použití Wardovy metody s mírou nepodobnosti založenou na Pearsonově korelačním koeficientu. V této kapitole v souvislosti s jedním z cílů práce ověřuji, zda navržený shlukovací postup zpřesní odhady smíšených lineárních modelů (LMM). Míra využitelnosti shlukování pro zpřesnění regresních modelů je zjišťována pomocí simulovaných dat. Nejprve se na generovaná data aplikuje navržený shlukovací postup. Dále se odhadnutá náležitost do shluků začlení do odhadovaného modelu jako kategoriální pomocná vysvětlující proměnná. Zároveň je odhadnut populační regresní model pro všechny objekty bez využití náležitosti do shluku. Oba modely (s náležitostí do shluku a bez ní) jsou porovnány pomocí odhadu průměrné chyby odhadu, AIC a BIC.

Postup navržený v této kapitole plánuji ve spolupráci s neurology použít na reálná rezonanční data, viz kapitola 1.2. Velkou skupinu (v řádech tisíců) pacientů s roztroušenou sklerózou chceme rozřadit do shluků dle vývoje objemu celkového mozku (v cm^3) a vybraných regionálních struktur. Změna objemu v čase bude odhadována pomocí LMM za předpokladu normálně rozdělené vysvětlované proměnné (např. celkový objem mozku). Motivací pro shlukování je vysoká míra interindividuální heterogenity vývoje v čase, která vede k nepřesným a zkresleným odhadům populačního vývoje objemu mozku v čase. Vyskytují se různé skupiny pacientů, které mají buď lineární, nebo nelineární klesající trend v čase. Vývoj individuálních trajektorií uvnitř těchto skupin (lineární a nelineární vývoj) je značně variabilní. U pacientů s lineárním vývojem se různí anualizovaná míra atrofie (směrnice lineárního vývoje). U nelineárního vývoje lze očekávat různé typy funkční závislosti (logaritmická, polynomy) objemu mozku na čase.

8.1 Data a metodologie

Data jsem simuloval v souladu s očekávaným průběhem lineární a nelineární atrofie (úbytku) mozku. Využil jsem rozdělení tří shluků ze scénáře 8 v oddílu 7.1 (dva nelineární vývoje pro shluk 1 a 2 a jeden lineární vývoj ve shluku 3). K těmto shlukům jsem přidal nový, čtvrtý shluk s lineárním vývojem s menším sklonem ve srovnání se shlukem 3. Trajektorie byly vygenerovány pomocí následujících funkcí:

shluk 1 (nelineární):

$y_{it} = \beta_0 + \beta_1 \log_{\beta_2}(t) + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(6{,}0; 6{,}5), \beta_2 \sim \mathrm{U}(0{,}04; 0{,}08),$
$\varepsilon_{it} \sim \mathrm{N}(0;0{,}4^2),$

shluk 2 (nelineární):

$y_{it} = \beta_0 + \beta_1 D_{it}^2 + \varepsilon_{it}, \quad \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(-0{,}25; -0{,}20), \varepsilon_{it} \sim \mathrm{N}(0; 0{,}4^2),$

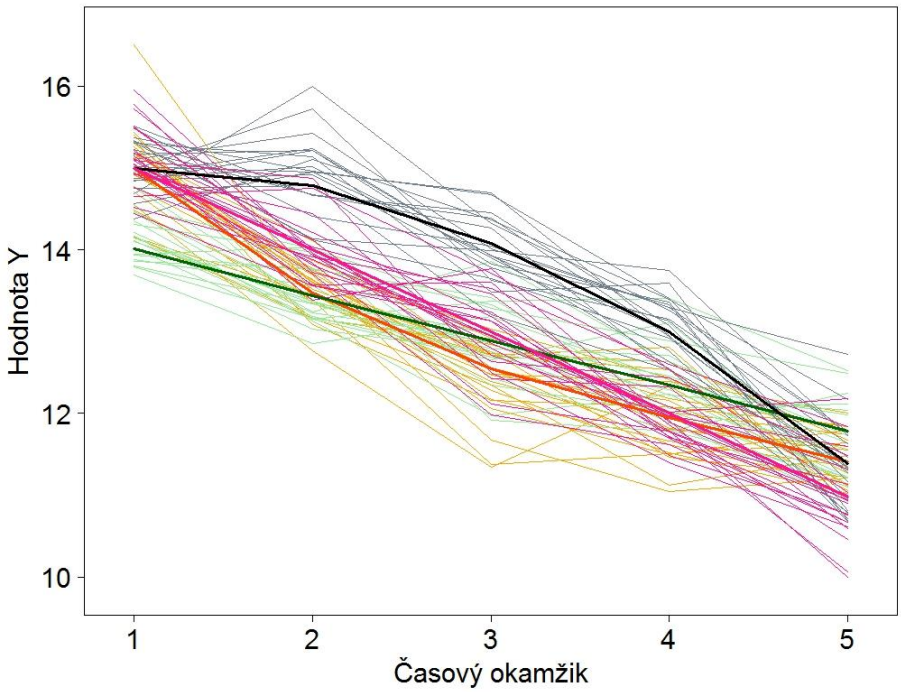
shluk 3 (lineární):

$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \ \beta_0 \sim \mathrm{N}(15; 0{,}1^2), \beta_1 \sim \mathrm{U}(-1{,}05; -0{,}95), \ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}4^2),$

shluk 4 (lineární):

$y_{it} = \beta_0 + \beta_1 D_{it} + \varepsilon_{it}, \ \beta_0 \sim \mathrm{N}(14; 0{,}1^2), \beta_1 \sim \mathrm{U}(-0{,}6; -0{,}5), \ \varepsilon_{it} \sim \mathrm{N}(0; 0{,}4^2).$

V každém shluku je 1000 trajektorií. Každý objekt je měřen pětkrát v pravidelných časových intervalech délky 1, stejně jako v oddílu 7.1.2, tj. $t = 1, 2, \ldots, T$. D_{it} značí časovou vzdálenost časového okamžiku t od počátku ($t = 1$) pro i-tý objekt (v letech), tj. v tomto generujícím procesu je $D_{it} = 0, 1, 2, 3, 4$. Vysvětlovaná proměnná je normálně rozdělená v každém časovém okamžiku, je tedy splněn předpoklad normality pro následné použití LMM. Obrázek 8.1 zobrazuje náhodný výběr 20 (z 1000) trajektorií pro každý vygenerovaný shluk. Trajektorie náležející do shluku 1 jsou zakresleny oranžově (nelineární konvexní pokles), trajektorie ve shluku 2 (nelineární konkávní pokles) mají barvu šedou. Třetí shluk je růžový (lineární pokles, prudší sklon) a zelené trajektorie ve shluku 4 se vyznačují pomalejším lineárním poklesem. Tučné čáry zobrazují reprezentativní trajektorii daného shluku. Jedná se o spojnici průměrných hodnot pro jednotlivá t. Kromě simulovaných „empirických“ hodnot y_{it} jsem uložil do datové matice skutečné střední hodnoty, kde skutečná střední hodnota pro i-tý objekt a t-tý časový okamžik je rovna systematické složce (bez ε_{it}) vygenerovaných funkcí pro shluky 1 až 4, tedy $y_{it} - \varepsilon_{it}$.



Obrázek 8.1 Ilustrace generovaných dat (náhodný výběr 20 trajektorií pro každý shluk).

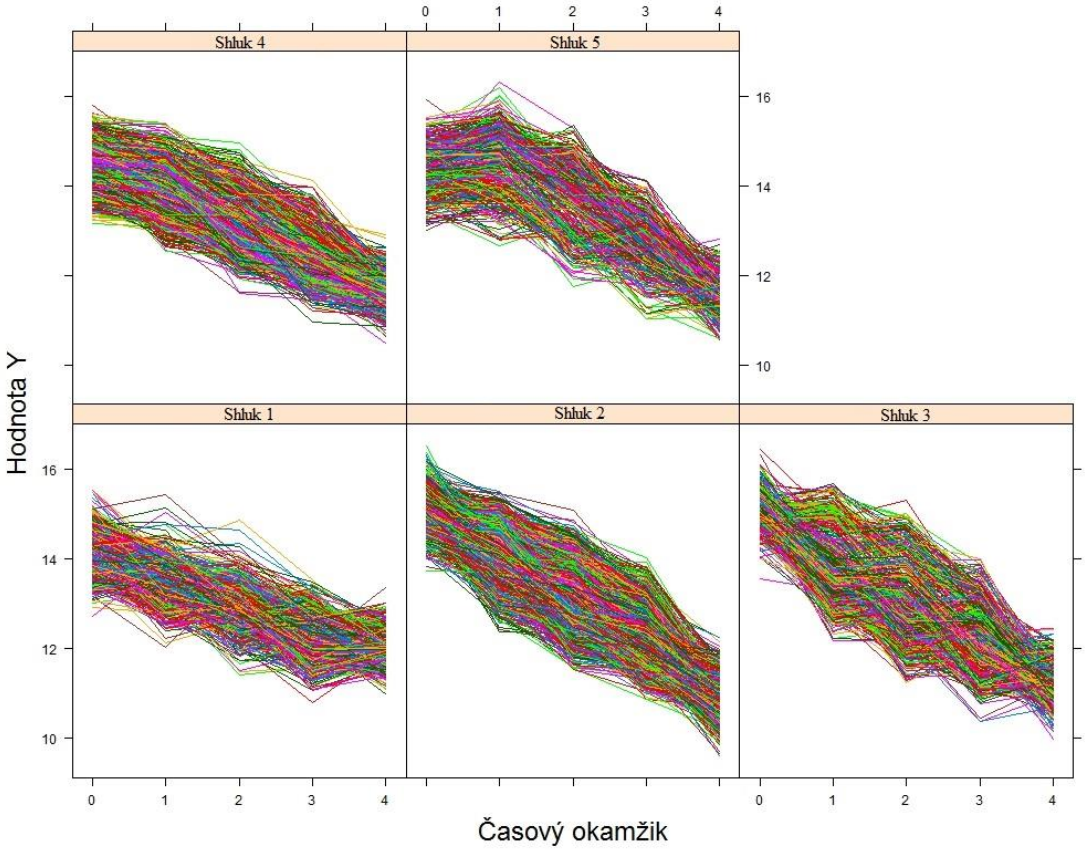
Na všechna data (4000 trajektorií) jsem aplikoval shlukovací postup, který jsem určil jako nejlepší (proměnné $\max \triangleleft$, $\overline{k_i}$ a %pos; Wardova metoda s Pearsonovou mírou nepodobnosti). Optimální počet shluků jsem stanovil pomocí tří kritérií: obrysový (silhouette) koeficient, viz Rousseeuw (1987), Calinského-Harabaszův index, viz Calinski a Harabasz (1974) a Dunnův koeficient, viz Dunn (1974). K jejich výpočtu jsem použil funkci *NbClust()* v balíčku *NbClust*, Charrad a kol. (2014). Výsledkem shlukování je diskrétní proměnná, jejíž hodnoty určují disjunktní náležitost do jednoho z *C* shluků.

V následné regresní analýze jsou odhadovány dvě skupiny LMM. V obou skupinách modelů je vysvětlovaná proměnná *Y*. Jedna skupina modelů obsahuje jako vysvětlující proměnnou čas (časovou proměnnou *t*) a je uvažována lineární a nelineární (kvadratická, kubická a logaritmická) závislost *Y* na *t*. Ve druhé skupině modelů jsou navíc odhadovány interakce časové proměnné *t* (t, t^2, t^3, ln t) s kategoriální proměnnou určující náležitost do jednoho ze shluků. Modely jsou popsány blíže v oddílu 8.3. Kvalita modelů je srovnána pomocí průměrných kvadratických chyb (*Mean Squared Error of Estimate*, **MSEE**)). Tato průměrná chyba je pro každý model počítána jako průměrný čtvercový rozdíl odhadnuté hodnoty Y_{it} a skutečné střední hodnoty $y_{it} - \varepsilon_{it}$, tj. $\mathrm{MSEE} = \frac{\sum_{i=1}^{4000}\sum_{t=1}^{5}\left(Y_{it} - \left(y_{it} - \varepsilon_{it}\right)\right)^2}{20000}$.

MSEE porovnává přesnost odhadů parametrů pevných efektů časové proměnné *t* modelu s náležitostí do shluku a bez ní. Dalšími srovnávacími kritérii odhadu modelů je AIC a BIC.

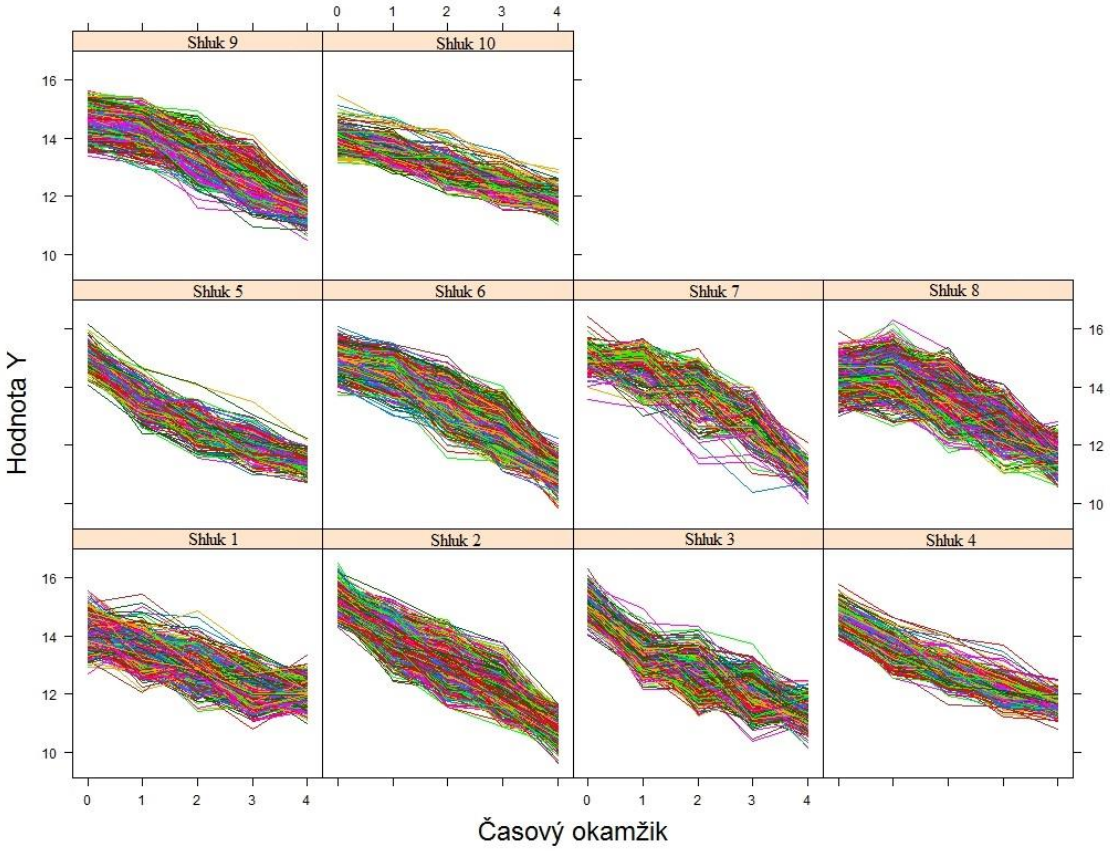
8.2 Výsledky shlukování a klasifikace

Na všechna data jsem aplikoval vybraný shlukovací postup 20krát. V každé iteraci jsem měnil počet shluků od 2 do 21. Trajektorie jsem rozdělil do deseti shluků na základě kritérií pomáhajících zvolit optimální počet, viz oddíl 8.1. Hodnoty těchto kritérií pro 2 až 21 shluků jsou zobrazeny na obrázku P1 (v příloze). Pro zvolená kritéria platí, že čím je hodnota tohoto kritéria větší, tím je výsledek shlukování lepší. Pokud bych vybral počet shluků striktně dle maximálních hodnot těchto kritérií, zvolil bych dva shluky. Tento počet se jeví nedostačující při pohledu na průběhy trajektorií na obrázku 8.1. Na obrázku lze totiž rozlišit, že se v datech vyskytují alespoň tři tvary poklesu (lineární, konvexní, konkávní) hodnot *Y* v čase. Na základě kritérií se dále nabízí vybrat jako optimální počet pět nebo deset shluků. Pět shluků se nejeví jako ideální počet, jelikož při tomto počtu je na základě inspekce profilu shluků zjevná nepříznivá vnitroshluková heterogenita. Profily pěti shluků zobrazuje obrázek 8.2.



Obrázek 8.2 Ilustrace trajektorií při rozdělení trajektorií do pěti shluků.

Deset shluků jsem zvolil, jelikož hodnota obrysového koeficientu a Calinského-Harabaszova indexu je stále relativně vysoká a dochází u nich ke zlomu a následnému (pro větší počet shluků) výraznému poklesu hodnot. Obrázek 8.3 zobrazuje všechny trajektorie rozdělené do deseti shluků. Obrázek ukazuje, že identifikovaných deset shluků vykazuje zjevnou homogenitu tvaru trajektorií uvnitř shluku a heterogenitu mezi shluky. Shluky se mezi sebou liší buď tvarem, nebo horizontální úrovní trajektorií v nich obsažených. Trajektorie ve shlucích 1, 2 a 10 mají lineární průběh. Shluk 2 obsahuje trajektorie s větším sklonem ve srovnání s trajektoriemi ve shluku 1 a 10. Směrnice trajektorií ve shluku 1 a 10 se příliš neliší. Shluk 10 vykazuje nižší náhodné kolísání (vnitroobjektovou variabilitu) okolo lineárního trendu než shluk 1. Ve zbylých sedmi shlucích převládají trajektorie s nelineárním tvarem. Tyto shluky se mezi sebou liší v míře nelinearity (zakřivení) a ve vnitroobjektové variabilitě.



Obrázek 8.3 Ilustrace trajektorií při rozdělení trajektorií do deseti shluků.

Profily deseti shluků doplňuje tabulka 8.1. Shluky jsou popsány pomocí průměrných hodnot a mediánů shlukovacích proměnných.

Tabulka 8.1 Profily odhadnutých shluků

shluk	počet trajektorií	max◁	$\overline{k}$	%pos
shluk 1	542	-0,03; -0,23	1,36; 1,30	1,02; 0,86
shluk 2	775	-0,05; -0,11	-1,15; -1,08	-0,69; -0,69
shluk 3	342	-0,30; -0,29	-0,37; -0,32	0,99; 0,86
shluk 4	282	-0,23; -0,21	0,38; 0,35	-0,69; -0,69
shluk 5	283	-0,25; -0,24	-0,27; -0,29	-0,69; -0,69
shluk 6	492	0,38; 0,37	-0,64; -0,58	-0,69; -0,69
shluk 7	196	0,68; 0,65	-0,81; -0,75	0,88; 0,86
shluk 8	474	0,86; 0,86	0,48; 0,35	1,32; 0,86
shluk 9	307	0,37; 0,37	0,26; 0,24	-0,69; -0,69
shluk 10	307	0,05; 0,11	1,35; 1,26	-0,69; -0,69

Legenda: První z dvojice čísel uvnitř tabulky je vždy průměr, druhé pak medián.

8.3 Výsledky regresní analýzy

V každé iteraci jsou odhadovány dvě skupiny LMM s náhodným průsečíkem a náhodným sklonem. Čas (t, t^2, t^3 a ln t) je vysvětlující proměnná s pevným efektem. Všechny modely jsou odhadnuty metodou REML. V první skupině jsou odhadovány pevné efekty času na celém souboru bez ohledu na náležitost do shluku. Níže jsou popsány odhadované čtyři modely formou zadání do funkce *lmer()*, která byla k odhadu použita:

- lineární: $Y \sim t + (t \mid \mathrm{id})$,
- kvadratický: $Y \sim t + t^2 + (t \mid \mathrm{id})$,
- kubický: $Y \sim t + t^2 + t^3 + (t \mid \mathrm{id})$,
- logaritmický: $Y \sim \ln t + (t \mid \mathrm{id})$,

kde Y značí vysvětlovanou proměnnou, t je pomocná časová proměnná s pevným efektem a $(t \mid \mathrm{id})$ značí, že jsou v modelu odhadovány náhodné efekty (sklon a průsečík).

Druhá skupina modelů odhaduje pevné efekty času pro každý shluk zvlášť. Za tímto účelem je do modelu přidána interakce náležitosti do jednoho z deseti shluků (kategoriální proměnná, která nabývá 10 hodnot) a času. V nelineárních modelech (kvadratický, kubický a logaritmický) je pro referenční skupinu (shluk) uvažován lineární trend. Na základě předchozí grafické analýzy byl vybrán shluk 10 jako referenční shluk z důvodu lineárního vývoje s nízkou mírou vnitroobjektové variability. Ve druhé skupině jsou odhadovány tyto čtyři modely:

- lineární: $Y \sim shluk + t + t{:}shluk + (t \mid \mathrm{id})$,
- kvadratický: $Y \sim shluk + t + t{:}shluk + t^2{:}shluk + (t \mid \mathrm{id})$,
- kubický: $Y \sim shluk + t + t{:}shluk + t^2{:}shluk + t^3{:}shluk + (t \mid \mathrm{id})$,

- logaritmický: $Y \sim shluk + t + \ln t{:}shluk + (t \mid \mathrm{id})$,

kde *shluk* značí kategoriální proměnnou „náležitost do shluku“ a *t*:*shluk* interakci *t* s proměnnou *shluk*.

Kvalita modelů je srovnána pomocí celkové MSEE, AIC a BIC. Hodnoty těchto kritérií jsou uvedeny v tabulce 8.2. Na základě hodnot těchto kritérií je nejvhodnějším modelem popisujícím vývoj *Y* v čase kubický nebo kvadratický model s náležitostí do shluku, tj. druhá skupina modelů. Nejnižší hodnota MSEE (rovná 0,092) byla spočítána u kvadratického modelu a nejnižší hodnota AIC (rovná 24674,5) a BIC (rovná 25022,2) u kubického modelu. Všechny nelineární modely z druhé skupiny (s náležitostí do shluku) odhadly výrazně lépe pevné efekty (hodnoceno kritériem MSEE) a nasimulované hodnoty *Y* (AIC, BIC) než první skupina nelineárních modelů bez informace o náležitosti do shluku. V první skupině modelů (bez náležitosti do shluku) se rovněž jeví jako nejvhodnější kubický model.

Tabulka 8.2 Kritéria pro výběr LMM

Model	$MSEE^1$	$MSEE^2$	AIC^1	AIC^2	BIC^1	BIC^2
Lineární	0,110	0,114	36125,1	30421,4	36172,6	30611,1
Kvadratický	0,109	**0,092**	36075,1	25294,4	36130,4	25563,1
Kubický	0,109	**0,095**	36049,4	**24674,5**	36112,6	**25022,2**
Logaritmický	0,215	0,116	41346,0	30563,2	41393,4	30760,8

Legenda: v tabulce jsou zobrazeny hodnoty: MSE, AIC a BIC. Horní indexy 1, 2 u těchto kritérií (v záhlaví tabulky) značí, o jakou skupinu se jedná. Hodnota 1 značí první skupinu modelů (bez náležitosti do shluků) a hodnota 2 značí druhou skupinu modelů (s náležitostí do shluků).

V závěru této kapitoly se nachází v tabulkách 8.3 a 8.4 pro srovnání odhady paremetrů (pevných a náhodných efektů) kubického modelu z první a druhé skupiny. V tabulkách je kromě bodových odhadů pevných efektů a jejich středních chyb hodnota testového kritéria *t* a P-hodnota individuálního *t*-testu o parametrech se Satterthwaiteovu aproximací počtu stupňů volnosti. Individuální t-testy byly provedeny pomocí balíčku *lmerTest*.

Tabulka 8.3 Odhad pevných a náhodných efektů pro kubický model bez náležitosti do shluku

Pevné efekty	Odhad	SE	testové kritérium *t*	P-hodnota
Průsečík	15,630	0,027	571,227	< 0,0001
t	-0,877	0,020	-43,293	< 0,0001
t^2	0,061	0,013	4,778	< 0,0001
t^3	-0,013	0,002	-6,187	< 0,0001

Náhodné efekty	Odhad rozptylu	Odhad směrodatné odchylky
Náhodný průsečík	0,366	0,605
t (náhodný sklon)	0,020	0,143
reziduum (odhad náhodné chyby)	0,254	0,504

V tabulce 8.4 jsou nejprve průsečíky, tj. průměrné hodnoty *Y* k počátku, průměrných trajektorií (prvních 10 odhadů) pro jednotlivé shluky formou odchylky od odhadu průsečíku 14,46 pro referenční skupinu (shluk č. 10). Oproti referenční skupině se signifikantně liší průsečík všech devíti shluků 1 až 9. Odhad pevného efektu pro *t* (11. odhadovaný parametr) je odhad lineárního efektu času na *Y* pro referenční skupinu. Další odhadované parametry jsou interakce času a náležitosti do shluku. Pomocí statistické signifikance (P-hodnot) pevných efektů interakcí náležitosti do skupiny a času v tabulce 8.4 lze validovat subjektivní úsudky o shlucích provedené na základě diagnostiky grafů v oddílu 8.2. Na základě této diagnostiky jsem usoudil, že referenční skupinu charakterizuje lineární klesající vývoj v čase. Tento úsudek se potvrdil, jelikož interakce „shluk č.10:t^2“ a „shluk č.10:t^3“ nejsou statisticky signifikantní (P-hodnota > 0,01). Na jednoprocentní hladině významnosti nezamítáme hypotézu o lineárním průběhu trajektorií ve shluku 10. Na základě subjektivního hodnocení profilů shluků jsem dále předpokládal, že i shluky 1 a 2 mají lineární průběh. Tento úsudek se jasně nepotvrdil, jelikož interakce „shluk č.1:t^3“, „shluk č.2:t^2“ a „shluk č.2:t^3“ jsou statisticky signifikantní a zdá se být vhodnější tyto shluky trajektorií charakterizovat kubickým nebo kvadratickým modelem. Na základě signifikance interakcí náležitosti s (nelineárním) časem (shluk:t^2 nebo shluk:t^3) se zdá být vhodné charakterizovat i zbývající shluky 3, 4, 5, 6, 7, 8 a 9 nelineárním poklesem.

V neposlední řadě je zajímavé porovnat odhad směrodatné odchylky náhodné chyby. V modelu s náležitostí do shluků (tabulka 8.4) je hodnota nevysvětlené variability (0,391) menší než v modelu bez náležitosti do shluku (tabulka 8.3; 0,504). Z toho lze rovněž usuzovat na preferenci modelu s náležitostí do shluku v této studii.

Tabulka 8.4 Odhad pevných a náhodných efektů pro kubický model s náležitostí do shluku

Pevné efekty	Odhad	SE	testové kritérium *t*	P-hodnota
Průsečík	14,460	0,074	194,412	< 0,0001
shluk č.1	0,333	0,093	3,573	0,0004
shluk č.2	2,469	0,088	28,096	< 0,0001
shluk č.3	2,911	0,103	28,411	< 0,0001
shluk č.4	1,598	0,108	14,864	< 0,0001
shluk č.5	2,416	0,107	22,493	< 0,0001
shluk č.6	0,922	0,095	9,724	< 0,0001
shluk č.7	0,988	0,119	8,294	< 0,0001
shluk č.8	-0,655	0,095	-6,858	< 0,0001
shluk č.9	0,340	0,105	3,232	0,0012
t	-0,463	0,056	-8,218	< 0,0001
shluk č.1:t	-0,185	0,071	-2,617	0,0089
shluk č.2:t	-1,221	0,067	-18,327	< 0,0001
shluk č.3:t	-1,796	0,078	-23,127	< 0,0001
shluk č.4:t	-0,923	0,081	-11,330	< 0,0001
shluk č.5:t	-1,373	0,081	-16,866	< 0,0001
shluk č.6:t	0,066	0,072	0,924	0,3555
shluk č.7:t	0,014	0,090	0,158	0,8745
shluk č.8:t	1,105	0,072	15,266	< 0,0001
shluk č.9:t	0,310	0,080	3,889	0,0001
shluk č.10:t^2	-0,052	0,036	-1,456	0,1454
shluk č.1:t^2	-0,038	0,027	-1,403	0,1607
shluk č.2:t^2	0,374	0,023	16,583	< 0,0001
shluk č.3:t^2	0,640	0,034	18,886	< 0,0001
shluk č.4:t^2	0,263	0,037	7,034	< 0,0001
shluk č.5:t^2	0,398	0,037	10,686	< 0,0001
shluk č.6:t^2	-0,140	0,028	-4,942	< 0,0001
shluk č.7:t^2	0,048	0,045	1,075	0,2825
shluk č.8:t^2	-0,597	0,029	-20,725	< 0,0001
shluk č.9:t^2	-0,235	0,036	-6,574	< 0,0001
shluk č.10:t^3	0,009	0,006	1,479	0,1392
shluk č.1:t^3	0,016	0,004	3,704	0,0002
shluk č.2:t^3	-0,057	0,004	-15,325	< 0,0001
shluk č.3:t^3	-0,077	0,006	-13,884	< 0,0001
shluk č.4:t^3	-0,026	0,006	-4,302	< 0,0001
shluk č.5:t^3	-0,042	0,006	-6,820	< 0,0001
shluk č.6:t^3	-0,002	0,005	-0,336	0,7368
shluk č.7:t^3	-0,047	0,007	-6,440	0,0000
shluk č.8:t^3	0,064	0,005	13,463	< 0,0001
shluk č.9:t^3	0,020	0,006	3,324	0,0009

Tabulka 8.4. Pokračování

Náhodné efekty	**Odhad rozptylu**	**Odhad směrodatné odchylky**
Náhodný průsečík	0,118	0,343
t (náhodný sklon)	0,000	0,021
reziduum (odhad náhodné chyby)	0,153	0,391

Závěr

Součástí této práce je ucelený přehled metod regresní a shlukové analýzy mikropanelových dat. V české literatuře neexistuje monografie, která by sdružovala popis tradičních a nejnovějších regresních metod. Jako tradiční metody z historického pohledu a četnosti jejich užívání vnímám MMM a LMM. Popis tradičních metod a postup, jak je aplikovat, nabízejí Pekár a Brabec (2012). Z nich vychází nejnovější metody, jimiž se rovněž zabývám. Nejnovější metody, mezi něž lze řadit GMMM, GLMM, GLMM, ZIGLMM, CLMM, GAMM a bayesovský přístup k regresnímu modelování, nejsou v domácí literatuře popsány v jednom zdroji. Jedná se o velmi pokročilé statistické metody. Z vlastní zkušenosti vím, že analytik musí investovat hodně času a projít mnoho zdrojů, pokud chce do oblasti regresní analýzy mikropanelových dat proniknout a tyto metody správně aplikovat na reálná data.

Všechny popisované metody (kapitola 2 a 3) jsou implementovány v dnes celosvětově rozšířeném a obecně užívaném systému R. Pro aplikaci některých metod lze vybírat dokonce z více funkcí. Rozdíly mezi funkcemi jsou často zásadní, začátečníkovi však nemusí být zřejmé. Z tohoto důvodu popisuji v kapitole 4 možnosti analýzy v systému R s důrazem na rozdíly mezi funkcemi. Stručný přehled metod a jejich dostupných aplikací v systému R pomohou uživateli zorientovat se a správně si zvolit metodu vhodnou pro konkrétní studii.

Zásadní problém používaných regresních metod spatřuji v nejednoznačnosti iterativních odhadů parametrů kovarianční matice náhodných chyb. Odhady parametrů lze zpřesnit v případě známé kovarianční struktury. Pokud struktura není známá, nebo se v datech vyskytuje směs rozdělení, doporučuji prozkoumat před regresní analýzou vztahy mezi náhodnými chybami.

V práci se kromě regresní analýzy zabývám shlukovou analýzou mikropanelových dat. Shluková analýza je možnost, jak v datech nalézt podskupiny, které v sobě zahrnují objekty homogenní z hlediska tvaru dynamiky vývoje sledovaných proměnných. Díky nalezeným shlukům lze lépe popsat a modelovat různé vzory chování náhodných chyb. Pro shlukování trajektorií mikropanelových dat je nabídka metod výrazně omezena. Shlukování trajektorií mikropanelových dat je relativně mladá oblast (2. pol. 90. let minulého století) v oboru statistiky. Existují však již zahraniční články a kapitoly v monografiích, které popisují stávající přístupy, viz Liao (2005); Wang a kol. (2006). Z důvodu omezeného výběru metod shlukové analýzy je cílem této disertační práce navrhnout vlastní postup shlukování mikropanelových dat a otestovat, zda tento postup shlukování zpřesní odhady regresních parametrů LMM. Další motivací pro

vývoj nových přístupů ke shlukování tohoto typu dat je fakt, že se shlukování trajektorií stává v poslední době často i primárním cílem analýzy dat.

Navrhl jsem několik kombinací shlukovacích proměnných, které reprezentují tvar individuálních časových trajektorií, a otestoval pomocí simulační studie jejich využitelnost pro shlukování trajektorií. Na základě provedené simulační studie lze konstatovat, že nejlépe shlukuje trajektorie kombinace shlukovacích proměnných (s6) $\max \triangleleft$, $\overline{k_i}$ a %pos při aplikaci aglomerativního shlukování s Wardovým algoritmem a mírou nepodobnosti založené na Pearsonově korelačním koeficientu (a6).

Ze simulací vyplývá, že je vhodnější použít metodu aglomerativního shlukování v případě, že individuální trajektorie mají tvar jako ve scénáři 1 a 5. Pro tyto scénáře dosáhla kombinace proměnných s6 při aplikaci metody a6 maximálně spolehlivých výsledků na rozdíl od metody *C*-průměrů. Metoda *C*-průměrů dopadla hůře v kombinaci s proměnnými, které doporučují jiní autoři (první diference $\mathbf{v}_{it}$, druhé diference $\mathbf{a}_{it}$ a původní empirické hodnoty $\mathbf{y}_i$). Kombinace proměnných, kterou navrhuji, s6 při aplikaci a6 dosáhla samostatně lepších výsledků než zavedené kombinace (s7 a s8). Zároveň kombinace s6 s a6 dopadla samostatně lépe než v kombinaci s již používanými shlukovacími proměnnými (s7 a s8). Lze učinit závěr, že v případě, kdy má každý objekt stejný počet měření, je vhodné použít shlukovací postup navržený v této práci, tj. aplikovat Wardovu metodu s mírou nepodobnosti založené na Pearsonovu korelačním koeficientu na shlukovací proměnné $\max \triangleleft$, $\overline{k_i}$ a %pos. Dalším závěrem, který vyplývá z provedené studie, je ten, že metoda *C*-průměrů je robustnější na použití shlukovacích proměnných, avšak v případě vhodně zvolených shlukovacích proměnných je vhodnější použít Wardovu metodu s mírou nepodobnosti založené na Pearsonově korelačním koeficientu.

Výhodou shlukování na základě navržených charakteristik dynamiky je možnost shlukovat data s výskytem chybějících hodnot vysvětlované proměnné $\mathbf{y}_i$ v některých časových okamžicích *t*. Jiné shlukovací postupy, viz D'Urso (2004) a Genolini a Falissard (2011), vyžadují pro shlukování data bez chybějících hodnot. V případě výskytu chybějících hodnot se tyto hodnoty musí dodatečně dopočítat některou z technik k tomu určených (například lineární interpolace, dosazení hodnoty z předchozího nebo následujícího časového okamžiku, neparametrický odhad). Doplnění chybějících hodnot $\mathbf{y}_i$ a aplikace metody *C*-průměrů na kompletní matici $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N)'$, kde $\mathbf{y}_i = (y_{i1}, \ldots, y_{it}, \ldots, x_{iT})'$ je v balíčku *kml* systému R, který je určený pro shlukování proměnných. Doplňování chybějících hodnot však může vnést do dat zkreslení. Mnou navržené charakteristiky dynamiky jsou rezistentní vůči chybějícím hodnotám.

Dokáží si poradit s typem návrhu běžným v praxi, kdy objekty jsou měřeny v různém počtu časových okamžiků s nestejnou délkou časové vzdálenosti mezi dvěma sousedními měřeními. Dalším přínosem této práce pak je vytvořený skript jazyka R přiložený na CD. Tento skript je možno použít pro analýzu vlastních mikropanelových dat.

Podle mého názoru by validitu závěrů o vhodnosti výběru metody shlukování bylo vhodné ověřit i na datech, kdy předpokládaná náležitost do shluků (počet objektů) není stejná. Je možné, že pro různě velké shluky metoda *C*-průměrů dosáhne lepších výsledků, jelikož Wardův algoritmus má tendence vytvářet stejně velké shluky. V provedené simulaci dále dopadla špatně euklidovská vzdálenost. Důvodem může být to, že navržené návrhy studie obsahovaly stejně velké shluky. Z toho vyplývá, že by bylo vhodné ověřit, zda je můj závěr obecně platný i pro jiný typ dat. Ve své budoucí vědecké práci se budu zabývat validací těchto závěrů na dalších v úvahu připadajících návrzích, ve kterých se například liší počet objektů ve shluku nebo počet opakovaných měření.

Námětem pro další výzkum je rovněž rozšíření mého přístupu na vícerozměrnou shlukovou analýzu, pokud chceme simultánně shlukovat více proměnných $\mathbf{Y} = (\mathbf{y}_{1t}, \ldots, \mathbf{y}_{it}, \ldots, \mathbf{y}_{Nt})'$, kde $\mathbf{y}_{it} = (y_{i1t}, \ldots, y_{ijt}, \ldots, y_{iJt})'$. V případě silné vzájemné závislosti uvažovaných proměnných v $\mathbf{Y}$ je možnost v prvním kroku najít latentní proměnnou, která reprezentuje tyto proměnné, některou z metod redukce dimenze, například metodou hlavních komponent. Další možností je spočítat shlukovací proměnné s2 a s6 pro každou proměnnou Y_j a v druhém kroku shlukovat na základě všech spočítaných shlukovacích proměnných. Tento postup dává všem proměnným Y_j stejnou váhu. Z tohoto důvodu nemusí být výsledek optimální. Navržené postupy je třeba otestovat. Inspiraci budu rovněž hledat v rozšíření algoritmu *kml* na vícerozměrnou shlukovou analýzu, která je implementovaná v balíčku *kml3d*, viz Genolini a kol. (2015).

Dalším výzkumným cílem je testování použitelnosti mého přístupu na trajektorie s velkým počtem opakovaných měření v čase (počet měření pro každý objekt je v řádu desítek). Pro shlukování takového typu dat se používají například metody založené na modelu (např. modely směsí rozdělení), se kterými lze můj přístup porovnat.

V závěrečné kapitole této práce ověřuji, zda navržený shlukovací postup zpřesní odhady regresních modelů LMM a hodnot vysvětlované proměnné na regresní analýze uměle generovaných dat. Na základě MSEE, AIC a BIC a odhadu variability náhodných chyb je preferován model, zahrnující jako pomocnou vysvětlující proměnnou náležitost do shluku před modelem bez této informace. Závěr provedené analýzy je ten, že má význam trajektorie nejprve shlukovat

a výsledky shlukové analýzy využít pro zpřesnění regresních modelů. Do budoucna plánuji provést důkladnou simulační studii, která potvrdí nebo vyvrátí závěr provedené analýzy.

V oddílu 6.1 této práce přicházím s myšlenkou regresního přístupu nazvaného regresní model prostorových bloků. Mohlo by se jednat o alternativu k LMM a GLMM. Zásadní problém těchto modelů tkví v numerickém přístupu odhadu parametrů modelu. Cílem myšlenky regresního modelu prostorových bloků bylo navrhnout regresní přístup, který neodhaduje parametry modelu iterativně, ale využívá OLS nebo ML pro data s obecnou kovarianční a korelační strukturou náhodných chyb. Uvažuji dvoukrokový přístup. V prvním kroku se shlukují individuální trajektorie, například pomocí navrženého přístupu ke shlukování. Ve druhém kroku se odhaduje vhodný model pomocí LM nebo GLM pro každý prostorový blok obsahující měření provedená v čase t pro objekty náležící do c-tého shluku. Je třeba matematicky prokázat správnost mé úvahy. Dále by bylo vhodné porovnat navržený blokový přístup s LMM a GLMM pomocí simulační studie.

Svůj návrh ke shlukování trajektorií plánuji aplikovat na reálná data, aktuálně v oblasti financí a medicíny. V oblasti medicíny hledám (v týmu s neurology z 1. LF UK) fenotypy roztroušené sklerózy. K dispozici máme longitudinální rezonanční snímky (MRI) celého mozku a jeho dílčích struktur. Aktuálně plánujeme shlukovat pacienty na základě individuálních trajektorií vyjadřujících vývoj objemu celého mozku a vybraných struktur s cílem identifikovat skupiny pacientů s podobným vývojem. Shluky budou vzájemně porovnány z hlediska klinických a kognitivních proměnných pacientů.

V oblasti financí se zabývám (v týmu s kolegy z University Mateje Bela v Banskej Bystrici) modelováním predikce finanční tísně podniku viz Kráľ a kol. (2014b); Stachová a kol. (2015). Sledovanou vysvětlovanou proměnnou je finanční tíseň podniku. Obecná definice finanční tísně podniku je situace, kdy není firma schopna dostát svým finančním závazkům.

V různých operačních definicích finanční tísně jsou sledovány absolutní hodnoty vybraných finančních ukazatelů v jednotlivých letech separátně. V plánu je navrhnout kompozitní skóre finanční tísně podniku, které zohledňuje dynamiku vývoje vybraných finančních ukazatelů a je lépe predikovatelné než dosud užívané definice.

Literatura

1. ABRAHANTES, J.C, AERTS, M. (2012). A solution to separation for clustered binary data. *Statistical Modelling*, roč. 12, č. 1, s. 3–27.
2. AGRESTI, A. (2010). Analysis of Ordinal Categorical Data. 2. vyd. Wiley.
3. ALTMAN, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, roč. 23, č. 4, s. 589–609.
4. AMIT, G., GAVRIELY, N., INTRATOR, N. (2009). Cluster analysis and classification of heart sounds. *Biomedical Signal Processing and Control*, roč. 4, 2009, s. 26–36.
5. ANKERST, M., BREUNIG, M., KRIEGEL, H.P., SANDER, J. (1999). OPTICS: ordering points to identify the clustering structure. In: *Proceedings of the 1999 ACM-SIGMOD International Conference on Management of Data*, Philadelphia, PA, June 1999, s. 49–60.
6. ASAR, O., ILK, O. (2013). mmm: an R package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine*, roč. 112, č.3, s. 649-654.
7. ASAR, O., ILK, O. (2014). *An R package for analyzing multivariate longitudinal data with multivariate marginal models*. Online manuál k balíčku verze 1.4. URL: https://cran.r-project.org/web/packages/mmm/mmm.pdf
8. BAAYEN, R.H., DAVIDSON, D.J., BATES, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, roč. 59, č. 4, s. 390-412.
9. BALCAEN, S., OOGHE, H. (2006). 35 Years of Studies on Business Failure: an Overview of the Classic Statistical Methodologies and Their Related Problems. *The British Accounting Review*, roč. 38, č. 1, s. 63–93.
10. BALDWIN, N., BORRELLI, S. (2008). Education and economic growth in the United States: cross-national applications for an intra-national path analysis. *Policy Sciences*, roč. 41, č. 3, 2008, s. 183–204.
11. BALTAGI, B.H. (2012). *Econometric Analysis of Panel Data*. 4. vyd. Chichester: John Wiley & Sons, Ltd.
12. BATES, D., KLIEGL, R., VASISHTH, S., BAAYEN, H. (2015b). *Parsimonious mixed models*. ArXiv e-print; submitted to Journal of Memory and Language. URL: http://arxiv.org/abs/1506.04967
13. BATES, D., MAECHLER, M., BOLKER, B., WALKER, S. (2015a). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, roč. 67, č. 1, s. 1–48.
14. BATES, D.M., MAECHLER, M., BOLKER, B.M., ET AL. (2016). *Linear Mixed-Effects Models using 'Eigen' and S4*. Online manuál k balíčku verze 1.1-11. URL: https://cran.r-project.org/web/packages/lme4/lme4.pdf
15. BECK, N., KATZ, J. (1995). What to do (and not to do) with time-series-cross-section data in comparative politics. *American Political Science Review*, roč. 89, s. 634–647.
16. BECKER, G.S. (1993). The Economic Way of Looking at Life. *Coase-Sandor Institute for Law & Economics* [online], Working Paper No. 12, URL: http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1509&context=law_and_economics
17. BELISLE, C.J.P. (1992). Convergence theorems for a class of simulated annealing algorithms on Rd. *Journal of Applied Probability*, roč. 29, s. 885–895.

18. BELL, A., FAIRBROTHER, M., JONES, K. (2016). *Fixed and random effects: making an informed choice*. Under review, URL: https://www.researchgate.net/publication/299604336
19. BERGER, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.*, roč. 1, č. 3, s. 385–402.
20. BEZDEK, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.
21. BEZDEK, J.C., KELLER, J., KRISNAPURAM, R., PAL, N.R. (1999). *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Dordrecht: Kluwer Academic Publishers.
22. BOLKER, B. (2014). *Dealing with quasi-models in R*. URL: https://cran.r-project.org/web/packages/bbmle/vignettes/quasi.pdf
23. BOLKER, B. (2016). *Wald errors of variances*. Online návod, poslední aktualizace z 2.1.2016. URL: https://rpubs.com/bbolker/waldvar
24. BOLKER, B.M., BROOKS, M.E., CLARK, C.J., GEANGE, S.W., POULSEN, J.R., STEVENS, M.H., WHITE, J.S. (2008). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, roč. 24, č. 3, s. 127–135.
25. BOLKER, B., BROOKS, M., GARDNER, B., LENNERT, C., MINAMI, M. (2012A). *Owls example: a zero-inflated, generalized linear mixed model for count data*. URL: https://groups.nceas.ucsb.edu/non-linear-modeling/projects/owls/WRITEUP/owls.pdf
26. BOLKER, B., SKAUG, H., MAGNUSSON, A., NIELSEN, A. (2012b). *Getting started with the glmmADMB package*. Online tutoriál k balíčku, URL: http://glmmadmb.r-forge.r-project.org/glmmADMB.pdf
27. BOX, G.E.P., JENKINS, G.M., REINSEL G.C. (1994). *Time Series Analysis: Forecasting and Control*. 3. vyd., Prentice Hall.
28. BRANDA, M. (2013). *Zobecněné lineární modely v pojišťovnictví*. Online, URL: http://www.nfvp.cz/res/data/000170.pdf
29. BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. (1984). *Classification and Regression Trees*. New York: Chapman and Hall, Wadsworth, Inc.
30. BRESLOW, N. (2003). WHITHER PQL? *UW Biostatistics Working Paper Series* [online], Working Paper 192, URL: http://biostats.bepress.com/uwbiostat/paper192/
31. BREZIGAR-MASTEN, A., MASTEN, I. (2012). CART-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications*, roč. 39, č. 11, s. 10153–10159.
32. BROOKS, S.P., GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, roč. 7, č. 4, s. 434–455.
33. BROWN, H., PRESCOTT, R. (2006). *Applied Mixed Models in Medicine*, 2. vyd. New York, NY: John Wiley and Sons.
34. BROYDEN, C.G. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, roč. 6, s. 76–90.
35. BRYK, A.S., RAUDENBUSH, S.W., SELTZER, M., CONGDON, R. (1988). *An Introduction to HLM: Computer Program and Users' Manual*. Chicago: University of Chicago Department of Education.

36. BRYK, A.S., RAUDENBUSH, S.W., CONGDON, R.T. (1996). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago: IL: Scientific Software International, Inc.
37. BUERKNER, P.CH. (2016). *Fit Bayesian generalized linear mixed models using Stan for full Bayesian inference*. Online manuál k balíčku verze 0.8.0, URL: https://cran.r-project.org/web/packages/brms/brms.pdf
38. BYRD, R. H., LU, P., NOCEDAL, J. AND ZHU, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, roč. 16, s. 1190–1208.
39. CARLIN, B.P. ET AL. (2006). Elements of hierarchical Bayesian inference. In: Clark, J.S. and Gelfand, A.E., (eds.) *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*, Oxford University Press, s. 3–24.
40. CENTER FOR ECONOMIC STUDIES. *Longitudinal Business Database*. URL: https://www.census.gov/ces/dataproducts/datasets/lbd.html
41. CONGDON, P. (2006). *Bayesian statistical modeling*. 2. vyd. Chichester: John Wiley & Sons, Ltd.
42. D'URSO, P. (2004). Fuzzy c-means clustering models for multivariate time-varying data: different approaches. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, roč. 12, č. 3, s. 287–326.
43. DAVIDIAN, M., GILTINAN, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.
44. DE LA CRUZ-MESÍA, R., QUINTANA, F.A., MARSHALL, G. (2008). Model-based clustering for longitudinal data. *Comput. Statist. Data Anal.*, roč. 52, s. 1441–1457.
45. DE STEFANO, N., STROMILLO, M.L., GIORGIO, A., BORTOLOZZI, M.L., BATTAGLINI, M., BALDINI, M., PORTACCIO, E., AMATO, M.P., SORMANI, M.P. (2015). Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. *J Neurol Neurosurg Psychiatry*, č. 0, s. 1–7, doi: 10.1136/jnnp-2014-309903.
46. DEADRICK, D., BENNETT, N., RUSSELL, C. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management*, roč. 23, č. 6, s. 745–757.
47. DIGGLE, P.J., HEAGERTY, P., LIANG, K.Y., ZEGER, S.L. (2002). *Analysis of Longitudinal Data*. 2 vyd. Oxford: Oxford University Press.
48. DITRICH, J. (2015). Použití logistické regrese pro male datové výběry. Informační bulletin České statistické společnosti, roč. 26, č. 4, s. 8–17.
49. DORIE, V. (2015). *Maximum a posteriori estimation for linear and generalized linear mixed-effects models in a Bayesian setting*. Online manuál k balíčku verze 1.0-4, URL: https://cran.r-project.org/web/packages/blme/blme.pdf
50. DRAFT r-sig-mixed-models FAQ (2013). *GLMM. Package comparison*. Online, URL: http://glmm.wikidot.com/pkg-comparison
51. DUNN, J.C. (1974). Well separated clusters and fuzzy partitions. *Journal on Cybernetics*, roč. 4, s. 95–104.
52. ELROD, P., KEANE, M.P. (1995). A Factor-Analytic Model for Representing the Market Structure in Panel Data. *Journal of Marketing Research*, roč. 32, č. February, s. 1–16.
53. ENGLAND, P., FARKAS, G., KILBOURNE, B.S., DOU, T. (1988). Explaining occupational sex segregation and wages: Findings from a model with fixed effects. *American Sociological Review*, roč. 53, s. 544–558.

54. ERDEM, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science*, roč. 15, s. 359–378.
55. EYRING, J.D., JOHNSON, D.S., FRANCIS, D.J. (1993). A cross-level units of analysis approach to individual differences in skill aquisition. *Journal of Applied Psychology*, roč. 78, č.5, s. 805–814.
56. FRALEY, CH., RAFTERY, A.E., MURPHY, T.B., SCRUCCA, L. (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597*. Department of Statistics, University of Washington. URL: https://www.stat.washington.edu/research/reports/2012/tr597.pdf
57. FERREIRA, L., HITCHCOCK, D.B. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics - Simulation and Computation*, roč. 38, s. 1925–1949.
58. FERRON, J (1997). Moving between hierarchical modeling notations. *Journal of Educational and Behavioral Statistics*, roč. 22, s. 119–123.
59. FITZMAURICE, G.M., LAIRD, N.M., WARE, J.H. (2011). *Applied Longitudinal Analysis*. 2. vyd. Hoboken: John Wiley & Sons.
60. FLETCHER, R. (1970). A New Approach to Variable Metric Algorithms. *Computer Journal*, roč. 13, č. 3, s. 317–322.
61. FLETCHER, R., REEVES, C.M. (1964). Function minimization by conjugate gradients. *Computer Journal*, roč. 7, č. 2, s. 149–154.
62. FIRTH, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, roč. 80, č. 1, s. 27–38.
63. FITZGERALD, J., GOTTSCHALK, P., MOFFITT, R. (1998). An analysis of sample attrition in panel data: the Michigan panel study of income dynamics. *The Journal of Human Resources*, roč. 33, s. 251–299.
64. FORBELSKÁ, M., HUDCOVÁ, H., BERNARDOVÁ, I., SVOBODOVÁ, J. (2015). Lineární smíšené regresní modely při sledování obsahu těžkých kovů v sedimentech řeky Moravy. Informační bulletin *České statistické společnosti*, roč. 26, č. 3, s. 13–24.
65. FOX, J. (2002). *Linear mixed models: An appendix to "An R and S-PLUS companion to applied regression"*. URL: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf
66. FOX, P.A., HALL ,A.D., SCHRYER, N.L. (1978). The PORT mathematical subroutine library. *ACM Trans. Math. Software*, roč. 4, s. 104–126.
67. GALECKI, A., BURZYKOWSKI, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. New York: Springer.
68. GELNAROVÁ, E. (2007). *Prediktivní odhady incidence a prevalence nádorových onemocnění na populační úrovní*. Rigorózní práce na MUNI (Přírodovědecká fakulta), Brno, 2007. URL: http://is.muni.cz/th/133275/prif_r/rigorozni_prace_Gelnarova.pdf
69. GENOLINI, CH., ALACOQUE, X., SENTENAC, M., ARNAUD, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, roč. 65, č. 4, s. 1–34. URL: http://www.jstatsoft.org/v65/i04/.
70. GENOLINI, CH. (2016). *kml: C-means for longitudinal data*. Online manuál k balíčku verze 2.4.1, URL: https://cran.r-project.org/web/packages/kml/kml.pdf
71. GELMAN, A. (2004). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall / CRC.
72. GELMAN, A. (2005). Analysis of variance: why it is more important than ever. *The Annals of Statistics*, roč. 33, č. 1, s. 1–52.

73. GELMAN, A., JAKULIN, A., PITTAU, M.G., SU, T.S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, roč. 2, č. 4, s. 1360–383.
74. GENOLINI, CH., FALISSARD, B. (2011). Kml: A package to cluster longitudinal data. *Computer methods and programs in biomedicine*, roč. 104, s. 112–121.
75. GRABUSTS, P. (2011). The choice of metrics for clustering algorithms. In: *Proceedings of the 8th International Scientific and Practical Conference, Environment. Technology. Resources*, roč. 11, s. 70–76.
76. GREENE, W.H. (1994). Some Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Working Paper EC-94-10: *Department of Economics, New York University*, URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1293115
77. GILKS, W.R., RICHARDSON, S., SPIEGELHALTER, D.J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
78. GIVON, M. (1984). Variety seeking through brand switching. *Marketing Science*, roč. 3, č. 1, s. 1–22.
79. GOLDFARB, D. (1970). A Family of Variable Metric Updates Derived by Variational Means. *Mathematics of Computation*, roč. 24, č. 109, s. 23–26.
80. GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, roč. 73, č. 1, s. 43–56.
81. GOLDSTEIN, H. (1987). *Multilevel Models in Educational and Social Research*. London: Oxford University Press.
82. GOSLEE, S.C., URBAN, D.L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, roč. 22, č. 7, 2007, s. 1–19.
83. GREEN, B.F., TUKEY, J.W. (1960). Complex analyses of variance: General problems. *Psychometrika*, roč. 25, č. 7, s. 127–152.
84. HADFIELD, J.D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, roč. 33, č. 2, s. 1–22.
85. HAGENAARS, J.A., MCCUTCHEON, A.L. (2002). *Applied Latent Class Analysis*. 1. vyd. Cambridge: Cambridge University Press.
86. HALL, D.B. (2000). Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, roč. 56, č. 4, s. 1030–1039.
87. HAN, J., KAMBER, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers.
88. HÄRDLE, W., SIMAR, L. (2007). *Applied Multivariate Statistical Analysis*. New York: Springer.
89. HARTIGAN, J.A., WONG, M.A. (1979). A C-means clustering algorithm. *Applied Statistics*, roč. 28, s. 100–108.
90. HARTLEY, H.O., RAO, J.N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, roč. 54, č. 1, s. 93–108.
91. HASTIE, T.J., TIBSHIRANI, R.J. (1990). Generalized Additive Models. Chapman and Hall/CRC.
92. HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J.H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag.
93. HEBÁK, P., HUSTOPECKÝ, J., PECÁKOVÁ, I., PRŮŠA, M., PLAŠIL, M., ŘEZANKOVÁ, H., SVOBODOVÁ, A., VLACH, P. (2007). *Vícerozměrné statistické metody [3]*. 2. vyd. Praha: Informatorium.

94. HEDEKER, D., GIBBONS, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, roč. 50, č. 4, s. 933–944.
95. HOFMANN, D.A, GRIFFIN, M.A., GAVIN, M.B. (2000). The Application of Hierarchical Linear Modeling to Organizational Research. In: Klein, K.J., Kozlowski, S.W.J. (eds.) *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*. San Francisco: Jossey Bass Inc., A Wiley Company, s. 467–511.
96. HOFMANN, D.A., STETZER, A. (1998). The role of safety climate and communication in accident interpretation: Implications for learning from negative events. *Academy of Management Journal*, roč. 41, č. 6, s. 644–657.
97. HOFFMAN, L., ROVINE, M.J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, roč. 39, č. 1, s. 101–117.
98. HØJSGAARD, S., HALEKOH, U., YAN J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, roč. 15, č. 2, s. 1–11.
99. HOTHORN, T., HORNIK, K., ZEILIS, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, roč. 15, č. 3, 2006, s. 651–674.
100. HOWARD, A.L. (2015). Leveraging Time-Varying Covariates to Test Within- and Between-Person Effects and Interactions in the Multilevel Linear Model. *Emerging Adulthood*, roč. 3, č. 6, s. 400–412.
101. HOX, J.J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.
102. HSIAO, C. (2003). *Analysis of Panel Data*. Cambridge: Cambridge University Press.
103. HUBERT, L., ARABIE, P. (1985). Comparing partitions. *Journal of Classification*, roč. 2, č. 1, s. 193–218.
104. CHAMBERLAIN, G. (1982). Multivariate Regression Models for Panel Data. *Journal of Econometrics*, roč. 19, č. 1, s. 5–46.
105. CHARRAD, M., GHAZZALI, N., BOITEAU, V., NIKNAFS, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, roč. 61, č. 6, s. 1–36. URL http://www.jstatsoft.org/v61/i06/.
106. CHILD, J. (1974). Managerial and organizational factors associated with company performance. Part1. *Journal of Management Studies*, roč. 11, s. 175–189.
107. CHILD, J. (1975). Managerial and organizational factors associated with company performance. Part2. *Journal of Management Studies*, roč. 12, s. 12–27.
108. CHINTAGUNTA, T. (1994). Heterogeneous Logit Implications for Brand Positioning. *Journal of Marketing Research*, roč. 31, č. May, s. 304–311.
109. CHOUAKRIA, A.D., NAGABHUSHAN, P.N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, roč. 1, č.1, s. 5–21.
110. CHRISTENSEN, R.H.B. (2015a). *ordinal - Regression Models for Ordinal Data*. Online manuál k balíčku verze 2015.6–28, URL: http://www.cran.r-project.org/package=ordinal/
111. CHRISTENSEN, R.H.B. (2015b): *A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package*. Online tutoriál k balíčku. URL: https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf
112. CHUA, A.S., EGOROVA, S., ANDERSON, M.C., POLGAR–TURCSANYI, CHITNIS, T., WEINER, H.L., GUTTMANN, CH.R.G., BAKSHI, R., HEALY, B.C.

(2015). Handling changes in MRI acquisition parameters in modeling whole brain lesion volume and atrophy data in multiple sclerosis subjects: Comparison of linear mixed-effect models. *NeuroImage: Clinical*, roč. 8, s. 606–610.

113. CHUNG, Y., RABE-HESKETH, S., DORIE, V., GELMAN, A., LIU, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, roč. 78, č. 4, s. 685–709.
114. JARMIN, R.S., MIRANDA, J. (2002). The Longitudinal Business Database. *Center for Economic Studies* [online], Discussion Paper CES-WP-02-17, URL: https://www.census.gov/ces/pdf/CES-WP-02-17.pdf
115. JENNRICH, R.I., SCHLUCHTER, M.D.: (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, roč. 42, č. 4, s. 805–820.
116. JIMMIESON, N., GRIFFIN, M.A. (1998). Linking staff and client perceptions of the organization: A field study of client satisfaction with health service. *Journal of Occupational and Organizational Psychology*, roč. 71, č. 1, s. 81–96.
117. JOHNSTON, A., BRUNO, A., WATANABE, J., QUANSAH, B., PATEL, N., DAKIN, S., ET. AL. (2008). Visually-based temporal distortion in dyslexia. *Vision Research*, roč. 48, č. 17, s. 1852–1858.
118. KAHN, B.E., KALWANI, M.U., MORRISON, D.G. (1986). Measuring Variety-Seeking and Reinforcement Behaviours Using Panel Data. *Journal of Marketing Research*, roč. 23, s. 89–100.
119. KANNAN, P.K., SANCHEZ, S.M. (1994). Competitive Market Structures: A Subset Selection Analysis. *Management Science*, roč. 40, č. 11 (November), s. 1484–1499.
120. KEANE, M.P. (1997). Modeling Heterogeneity and State Dependence in Consumer Choice Behaviour. *Journal of Business & Economic Statistics*, roč. 15, č. 3 (July), s. 310–327.
121. KOESTLER, D.C., MARSIT, C.J., CHRISTENSEN, B.C., KELSEY, K.T., HOUSEMAN, E.A. (2014). A recursively partitioned mixture model for clustering time-course gene expression data. *Translational Cancer Research*, roč. 3, č. 3, s. 217–232.
122. KOHONEN, T. (2001). *Self-organizing maps*. 3. vyd. Berlin: Springer.
123. KRÁĽ, P., SOBÍŠEK, L., STACHOVÁ, M. (2014a). A Distance Based Measure of Data Quality. *Metodološki zvezki – Advances in Methodology and Statistics*, roč. 11, č. 2, s. 107–120.
124. KRÁĽ, P., STACHOVÁ, M., SOBÍŠEK, L. (2014b). Utilization of Repeatedly Measured Financial Ratios in Corporate Financial Distress Prediction in: Slovakia. In: *Applications of mathematics and statistics in economy (AMSE)*. Jerzmanovice, 27.08.2014 – 31.08.2014. Wroclaw : Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, s. 156–163.
125. KREFT, I., DE LEEUW, J. (1998). *Introducing Multilevel Modeling*. London: Sage.
126. KUZNETSOVA, A., BROCKHOFF, P.B., CHRISTENSEN, R.H.B. *lmerTest: Tests in Linear Mixed Effects Models*. Online manuál k balíčku verze 2.0-32. URL: http://CRAN.R-project.org/package=lmerTest
127. LAIRD, N.M., WARE, J.H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, roč. 38, č. 4, s. 963–974.
128. LAMBERT, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, roč. 34, č. 1, s. 1–14.
129. LAMOTTE, L.R. (2014). Fixed-, random-, and mixed-effects models. *Wiley StatsRef: Statistics Reference Online* [online], DOI: 10.1002/9781118445112.stat03169.
130. LATOUR, D., LATOUR, K., WOLFINGER, R.D (1994). *Getting Started with PROC MIXED*. Cary: SAS Institute, Inc.

131. LECKIE, G., BELL. A. (2013). *Cross-Classified Multilevel Models – Stata Practical. LME VLE Module 12*. URL: http://www.bristol.ac.uk/cmm/learning/course.html

132. LEE, Y., NELDER, J.A. (2006). Double hierarchical generalized linear models (with discussion). *Applied Statistics*, roč. 55, s. 139–185.
133. LEE, Y., NELDER, J.A., PAWITAN, Y. (2006). *Generalised linear models with random effects: unified analysis via h-likelihood*. London: Chapman & Hall.
134. LEE, Y., NOH, M. (2015). *Double Hierarchical Generalized Linear Models*. Online manuál k balíčku verze 1.5. URL: https://cran.r-project.org/web/packages/dhglm/dhglm.pdf

135. LECHNEROVÁ, Z. (2009). Charakteristika panelového výzkumu a jeho vývoj. *Data a výzkum – SDA Info*, roč. 3, č. 1, s. 31–51.
136. LI, Q., RACINE, J.C. (2007). *Nonparametric Econometrics*. New Jersey: Princeton University Press.
137. LIANG, K.Y., ZEGER, S.L. (1986). Longitudinal data analysis using Generalized Linear Models. *Biometrika*, roč. 73, s. 13–22.
138. LIAO, T.W. (2005). Clustering of time series data – a survey. *Pattern Recognition*, roč. 38, č. 11, s. 1857–1874.
139. LIN, X., ZHANG, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, roč. 61, č. 2, s. 381–400.
140. LINDLEY, D.V., SMITH, A.F.M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society*. Series B (Methodology), roč. 34, č. 1, s. 1–41.
141. LINDSTROM, M.J., BATES, D.M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, roč. 46, č. 3, s. 673–687.
142. LITTELL, R.C., MILLIKEN, G.A., STROUP, W.W., WOLFINGER, R.D. (2006). *SAS System for Mixed Models*. 2. vyd. Cary: SAS Institute, Inc.
143. LITTLE, R.J.A., RUBIN, D.B. (2002). *Statistical Analysis with Missing Data*. 2. vyd. New York: Wiley-Interscience.
144. LOMBARDO, R., FALCONE, M. (2011). *Crime and Economic Performance. A cluster analysis of panel data on Italy's NUTS 3 regions*. Working Paper, University of Calabria, s.1–33, URL: http://www.ecostat.unical.it/RePEc/WorkingPapers/WP12_2011.pdf

145. LORCH, R.F., MYERS, J.L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory and Cognition*, roč. 16, č.1, s. 149–157.
146. LUCAS, A. (2015). *amap: Another Multidimensional Analysis Package*. Online manuál k balíčku verze 0.8-14, URL: https://cran.r-project.org/web/packages/amap/amap.pdf

147. MACKOVIČOVÁ, L., SOBÍŠEK, L., STACHOVÁ, M. (2011). Porovnanie metód DBSCAN a CLARA na reálných dátach. *Forum Statisticum Slovacum*, roč. 7, č. 7, s. 124–129.

148. MACQUEEN, J.B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. roč. 1, s. 281–297.

149. MARTINETZ, T.M., BERKOVICH, S.G., SCHULTEN, K.J. (1993). "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, roč. 4, č. 4, s. 558–569.

150. MCCULLOCH, C.E., SEARLE, S.R., NEUHAUS, J.M. (2008). Generalized, Linear, and Mixed Models. 2. vyd., Hoboken: John Wiley & Sons.
151. MENARD, S. (2002). *Longitudinal Research (Quantitative Applications in the Social Sciences)*. 2. vyd. London: Sage Publications.
152. MOLENBERGHS, G., VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. Berlin: Springer-Verlag.
153. MOULTON, B.R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, roč. 32, s. 385–397.
154. MOULTON, B.R. (1987). Diagnostics for group effects in regression analysis. *Journal of Business and Economic Statistics*, roč. 5, s. 275–282.
155. MULLIN, B.P. (2010). *A Quantitative Analysis of Human Capital as an Economic Development Tool for the South*. disertační práce. URL: http://gradworks.umi.com/34/38/3438492.html
156. MUNDLAK, Y. (1978). Pooling of Time-Series and Cross-Section Data. *Econometrica*, roč. 46, č. 1, s. 69–85.
157. NASH, J.C. (2014). On best practice optimization methods in R. *Journal of Statistical Software*, roč. 60, č. 2, s. 1–14.
158. NELDER, J.A., MEAD, R. (1965). A simplex method for function minimization. *Computer Journal*, roč. 7, s. 308–313.
159. NOCEDAL, J., WRIGHT, S.J. (2006). *Numerical Optimization*. 2. vyd. New York: Springer.
160. NOVÁK, P. (2007). Analýza panelových dat. *Acta Oeconomica Pragensia*, roč. 15, č. 1, s. 71–78.
161. O'HARA, R.B., KOTZE, D.J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, roč. 1, č. 2, s. 118–122.
162. PAAP, R. (2002). What are the advantages of MCMC based inference in latent variable models?. *Statistica Neerlandica*, roč. 56, č. 1, s. 2–22.
163. PATTERSON, H.D., THOMPSON, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, roč. 58, č. 3, s. 545–554.
164. PEKÁR, S., BRABEC, M. (2012). *Moderní analýza biologických dat. 2. Lineární modely s korelacemi v prostředí R*. 1. vyd. Brno: Masaryk University Press.
165. PEŠOUT, P. (2010). *Přístupy k shlukování funkčních dat*. Disertační práce na VŠE (FIS), Praha, 2010. URL: https://isis.vse.cz/zp/portal_zp.pl?podrobnosti_zp=14829
166. PINHEIRO, J.C., BATES., D.M. (1996). Unconstrained Parametrizations for Variance-Covariance Matrices. *Statistics and Computing*, roč. 6, č. 3, s. 289–296.
167. PINHEIRO, J.C., BATES, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
168. PINHEIRO, J.C., BATES D.M., DEBROY, S., SARKAR, D., AND R CORE TEAM (2016). *nlme: Linear and Nonlinear Mixed Effects Models*. Online manuál k balíčku verze 3.1-126. URL: https://cran.r-project.org/web/packages/nlme/nlme.pdf
169. PINHEIRO, J.C., CHAO, E.C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, roč. 15, č. 1, s. 58–81.
170. POWELL, M.J.D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. Report No. DAMTP 2009/NA06, Cambridge: Centre for Mathematical Sciences, University of Cambridge. URL: http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf

171. PRENTICE, R.L., ZHAO, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous response. *Biometrics*, roč. 47, č. 3, s. 825–839.
172. PROCHASKA, J.O., VELICER, W.F., GUADAGNOLI, E., ROSSI, J.S., DICLEMENTE, C.C. (1991). Patterns of Change: Dynamic Topology Applied to Smoking Cessation. *Multivariate Behavioral Research*, roč. 26, č. 1, s. 83–107.
173. PROSSER, R., RASBASH, J., GOLDSTEIN, H. (1996). *MLn User's Guide*. London: Institute of Education.
174. QUINLAN, J.R. (1993). *C4.5: Programs for Machine Learning*. Los Altos: Morgan Kaufmann.
175. R CORE TEAM (2013). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2012. URL: http://www.r-project.org/
176. RAMEY, J.A. (2015). *clusteval: Evaluation of Clustering Algorithms*. Online manuál k balíčku verze 0.1. URL: http://CRAN.R-project.org/package=clusteval
177. RAND, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, roč. 66, s. 846–850.
178. RASBASH, J. (2010). *What are mulitlevel models and when do we use them?*. University of Bristol, Center for Multilevel Modeling. URL: http://www.cmm.bristol.ac.uk/learning-training/multilevel-models/whatwhy
179. RAUDENBUSH, S.W., BRYK, A.S. (2002). *Hierarchical linear models: Applications and data Analysis Methods*. 2. vyd. SAGE Publications, Inc.
180. RAUDENBUSH, S.W., YANG, M.-L., YOSEF, M. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, roč. 9, č. 1, s. 141–157.
181. ROSENBERG, B. (1973). Linear Regression with Randomly Dispersed Parameters. *Biometrika*, roč. 60, s. 65–72.
182. ROSS, M.L. (2012). *The Oil Curse: How Petroleum Wealth Shapes the Development of Nations*. Princeton, NJ: Princeton University Press.
183. ROUSSEEUW, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, roč. 20, s. 53–65.
184. RUPPERT, D., CARROLL, R. (1988). *Transformation and Weightening in Regression*. London and New York: Chapman and Hall.
185. RUSPINI, E. (2002). *Introduction to Longitudinal Research*. London: Routledge.
186. ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V. (2009). *Shluková analýza dat*. 2. vyd. Praha: Professional Publishing.
187. ŘEZANKOVÁ, H., SOBÍŠEK, L. (2011). Interpretation and Combination of Results of Dimensionality Reduction Methods by Fuzzy Cluster Analysis. In: *ASMDA 2011*. Rome, 07.06.2011 – 10.06.2011. Roma : Sapieza Universita di Roma, s. 1164–1171.
188. SAS INSTITUTE (1992). *SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements*. Cary: NC: Author.
189. SAS INSTITUTE (1996). *SAS/STAT Software: Changes and Enhancements through Release 6.11*. Cary: SAS Institute.
190. SATO, M., SATO, Y., JAIN, L.C. (1997). *Fuzzy Clustering Models and Applications*. Tokio: Springer.

191. SEARLE, S.R., CASELLA, G. AND MCCULLOCH, C.E. (1992). *Variance Components*. New York: Wiley.

192. SELTMAN, H. (2014). *Experimental Design for Behavioral and Social Sciences*. URL: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

193. SHANNO, D.F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comput.*, roč. 24, č. 111, s. 647–656.

194. SCHALL, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, roč. 78, č. 4, s. 719–727.

195. SINGER, J.D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*, roč. 24, č. 4, s. 323–355.

196. SINGER, J., WILLETT, J. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

197. SNIJDERS, T.A.B., BOSKER, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2. vyd. London: Sage Publishers.

198. SOBÍŠEK, L., ŘEZANKOVÁ, H. (2011). Srovnání metod pro redukci dimenzionality aplikovaných na ordinální proměnné. *Acta Oeconomica Pragensia*, roč. 19, č. 1, s. 3–19.

199. SOBÍŠEK, L., STACHOVÁ, M. (2015). Can clustering based on predictors' dynamics improve the accuracy of financial distress classifier? In: *Applied statistics*. Ribno, 20.09.2015 – 23.09.2015. Ljubljana : Narodna in univerzitatetna knijižnica, s. 56.

200. SOBÍŠEK, L., STACHOVÁ, M., PECÁKOVÁ, I. (2015). Analýza longitudinálních dat pomocí smíšeného lineárního modelu v programu R. *Forum statisticum slovacum*, roč. 11, č. 6, s. 138–143.

201. SOBÍŠEK, L., VINTROVÁ, V., VINTR, T., PASTOREK, L., ŘEZANKOVÁ, H. (2011). Aplikace metod shlukové analýzy na data z pojišťoven. *Forum Statisticum Slovacum*, roč. 7, č. 5, s. 145–151.

202. STACHOVÁ, M., KRÁĽ, P., SOBÍŠEK, L., KAKAŠČÍK, M. (2015). Analysis of financial distress of Slovak companies using repeated measurements. In: *Applications of mathematics and statistics in economy (AMSE)*. Jindřichův Hradec, 02.09.2015 – 06.09.2015. Prague : University of Economics, Prague, Oeconomica Publishing House.

203. STACHOVÁ, M., SOBÍŠEK, L. (2013). Predikcia predčasného ukončenia poistnej zmluvy pomocou podmienených stromových štruktúr. *Forum statisticum slovacum*, roč. 9, č. 7, s. 201–206.

204. STACHOVÁ, M., SOBÍŠEK, L. (2016). Financial Distress Criteria Defined by Clustering of Longitudinal Data. In: *The 10th International Days of Statistics and Economics (MSED 2016)*. Praha, 8.09.2015 – 10.09.2015. Slaný : Melandrium, 2015, s. 1703–1712.

205. STRASSER, H., WEBER, C. (1999). On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*, roč. 8, s. 220–250.

206. STUKEL, T.A., DEMIDENKO, E. (1997). Two-Stage Method of Estimation for General Linear Growth Curve Models. *Biometrics*, roč. 53, č. 2, s. 720–728.

207. STUKEL, T.A., DEMIDENKO, E., DYKES, J., KARAGAS, M.R. (2001). Two-stage methods for the analysis of pooled data. *Statistics in Medicine*. roč. 20, č. 14, s. 2115–2130.

208. ŠMILAUER, P. (2007). *Moderní regresní metody*. URL: http://regent.jcu.cz/MRM.pdf

209. THAI, H.T., MENTRÉ, F., HOLFORD, N.H.G., VEYRAT-FOLLET, CH., COMETS, E. (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical Statistics*, roč. 12, č. 3, s. 129–140.
210. TINOCO, M. H., WILSON, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, roč. 30, s. 394–419.
211. TREMBLAY, A. (2015). *LMERConvenienceFunctions: Model Selection and Post-hoc Analysis for (G)LMER Models*. Online manuál k balíčku verze 2.10. URL: http://cran.r-project.org/web/packages/LMERConvenienceFunctions/index.html
212. UHER, T., HAVRDOVA, E., SOBISEK, L., KRASENSKY, J., VANECKOVA, M., SEIDL, Z., TYBLOVA, M., RAMASAMY, D., ZIVADINOV, R., HORAKOVA, D. (2016b). Is no evidence of desease activity an achievable goal in MS patients on intramuscular interferon beta-1a treatment over long-term follow-up?. *Multiple Sclerosis Journal*, doi: 10.1177/1352458516650525.
213. UHER, T., VANECKOVA, M., SOBISEK, L., TYBLOVA, M., SEIDL, Z., KRASENSKY, J., RAMASAMY, D., ZIVADINOV, R., HAVRDOVA, E., KALINCIK, T., HORAKOVA, D. (2016a). Combining clinical and MRI markers enhances prediction of 12-year disability in multiple sclerosis. *Multiple Sclerosis Journal*, doi: 10.1177/1352458516642314.
214. ULLAH, A., ROY, N. (1988). Nonparametric and semiparametric econometrics of panel data, Chapter 17. In Ullah, A., Giles, D.E.A. (eds.) *Handbook on Applied Economic Statistics*. New York: Marcel Dekker.
215. VENABLES, W.N., RIPLEY, B.D. (2002). *Modern Applied Statistics with S*. 4. vyd. New York: Springer.
216. VERBEKE, G., MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. New Yorc-Berlin-Heidelberg: Springer-Verlag.
217. VERBYLA, A.P. (1990). A conditional derivation of residual maximum likelihood. *The Australian Journal of Statistics*, roč. 32, č.2, s. 227–230.
218. VERMUNT, J.K., MAGIDSON, J. (2002). Latent class cluster analysis. In Hagenaars, J. A., McCutcheon, A.L. (eds.). *Applied Latent Class Analysis*. Cambridge: Cambridge University Press, s. 89–106.
219. WANG, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, roč. 90, s. 43–52.
220. WANG, N., CARROLL, R.J., LIN. X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, roč. 100, s. 147–157.
221. WANG, X., SMITH, K., HYNDMAN, R. (2006). Characteristic-Based Clustering for Time Series Data. *Data Mining and Knowledge Discovery*, roč. 13, s. 335–364.
222. WANG, W., YANG, J., MUNTZ, R.: STING (1997). A statistical information grid approach to spatial data mining. In: *Proceedings of the 1997 International Conference on Very Large Data Base (VLDB'97)*, Athens, Greek, 1997, s. 186–195.
223. WEST, B.T., WELCH, K.B., GALECKI, A. (2014). *Linear mixed models: A practical guide using statistical software*. 2. vyd. CRC Press.
224. WILLIAMS, J.D., BIRCH, J.B., ABDEL-SALAM, A.S. (2015). Outlier robust nonlinear mixed model estimation. *Statistics in Medicine*, roč. 34, č. 8, s. 1304–1316.
225. WINKELMANN, L., WINKELMANN, R. (1998). Why are the unemployed so unhappy? Evidence from panel data. *Economica*, roč. 65, s. 1–15.

226. WOOD, S.N. (2006a). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
227. WOOD, S.N. (2006b). Low rank scale invariant tensor product smooths for Generalized Additive Mixed Models. *Biometrics*, roč. 62, č. 4, s. 1025–1036.
228. WOOD, S.N. (2015). *gamm4 - Fit generalized additive mixed models via a version of mgcv's gamm function, using lme4 for estimation*. Online manuál k balíčku verze 0.2-3. URL: https://cran.r-project.org/web/packages/gamm4/gamm4.pdf
229. WOOD, S.N. (2016). *mgcv - Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. Online manuál k balíčku verze 1.8-12. URL: https://cran.r-project.org/web/packages/mgcv/mgcv.pdf
230. YEUNG, K.Y., RUZZO, W.L. (2001). *Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data" (to appear in Bioinformatics)*. Online doplňující material ke článku. URL: http://faculty.washington.edu/kayee/pca/supp.pdf
231. ZESHUI, X., MEIMEI, X. (2011). Distance and similarity meansures for hesitant fuzzy sets. *Information Sciences*, roč. 181, č. 11, s. 2128–2138.
232. ZUUR, A.F., IENO, E.N., WALKER, N.J., SAVELIEV, A.A., SMITH. G.M. (2009). *Mixed Effects Models and Extensions In Ecology with R*. New York: Springer.

Přílohy

Porovnání nejužívanějších funkcí (nejen R) shrnuje tabulka P1 pro LMM z hlediska použité metody pro odhad parametrů, možnosti provedení a výběru celkového testu o modelu a individuální testy o parametrech, možnosti a výběru intervalových odhadů, volby struktury náhodných efektů a struktury reziduí. V poznámkách k tabulce níže jsou objasněny některé pojmy vyskytující se v tabulce. Tabulka P2 porovnává funkce pro odhad GLMM. Tabulka P3 zahrnuje počet získaných bodů pro každý testovaný shlukovací scénář získaných v osmi scénářích. Obrázky P1 až P5 ilustrují hodnoty kritérií pro určení optimálního počtu shluků v 5ti iteracích v simulační studii v kapitole 8.

Tabulka P1. Porovnání softwarových funkcí pro odhad LMM

SW	Funkce	Metoda odhadu parametrů	Odhady (testy o signifikanci)[9]	Odhady intervalů spolehlivosti	Struktura matice náhodných efektů	Struktura reziduí
R nlme	lme	ML, REML	Waldův t test, test věrohodnostním poměrem (logLR), sekvenční a marginální F testy	Waldův interval pro parametry fixních a náhodných efektů	vícerozměrné sdružené (*nested*) náhodné efekty	prostorová a časová
					diagonální, bloková struktura	spojitá a diskrétní heteroskedasticita
					křížená (*crossed*) struktura možno modelovat, ale pomalé	
R lme4	lmer	ML, REML	F test, logLR, post-hoc MCMC	post-hoc MCMC, profily věrohodnosti, parametrický bootstrapping	sdružené a křížené	žádná
SAS	PROC MIXED	REML,ML, MIVQUE0, nebo Typ I–Typ III	Waldův t a F test		vícerozměrná,komplexní	
SAS	PROC GLIMMIX	pseudo-věrohodnost, Laplace, GHQ, REML, PQL	Waldův, logLR, test pro fixní efekty typu III	Waldův,LR	Vícerozměrné,sdružené nebo křížené	
SAS	HPMIXED	REML	Waldův t a F test, test typu III and chí-kvadrát test	Waldovi intervaly fixních a náhodných efektů	vícerozměrná, komplexní	
HLM	HLM	REML,FML			HLM,sdružené a/nebo křížené náhodné efekty	
MLWiN		ML, MCMC			HLM, sdružené/křížené náhodné efekty	
Stata	xtmixed//xtreg(random-intercept model)	REML,ML	Waldův, logLR	Waldův	HLM, sdružené/křížené,4 typy kovarianční struktury, heteroskedasticita náhodných efektů	Heteroskedasticita, obecná (*unstructured*), exponenciální, pruhovaná (*banded*)

[9] Statistické testy o celkovém modelu a odhadu jednotlivých parametrů.

Sloupec „Odhady (testy o signifikanci)[1]„ sdružují seznam možných individuálních testů o parametrech (Waldův *t*-test, test poměrem věrohodnosti (logLR)) a simultánních testů o modelu (Waldův chí-kvadrát test, logLR, podmíněný *F*-test, test využívající simulaci Markovova řetězce Monte Carlo (MCMC)).

Tabulka P2. Porovnání softwarových funkcí pro odhad GLMM

SW	Funkce	Metoda odhadu parametrů			Odhady (testy o signifikanci parametrů, intervalové odhady)		MCMC glmm	Další specifika modelu		
		PQL	Laplaceova aproximace	GHQ	Waldův χ^2 test nebo F test	Stupně volnosti		Křížový náhodný efekt	Spojitá prostorová /časová korelace	Overdispersion
SAS	PROC GLIMMIX	x	x	x	x	BW, S, KR		x	x	QL
	PROC NLMIXED			x	x	BW, S, KR				Dist
R	glmmPQL	x			x	BW			x	QL
	glmmML		x	x						
	glmer		x	x			x	x		QL
	glmmADMB		x							Dist
	GLMM	x			x			x	x	QL
GenStat/ASREML			x	x			x	x		Dist
AD Model Builder		x	x					x		x
HLM				x						
GLLAAMM (Stata)							x			Dist
WinBUGS							x	x		

Zkratky použité v Tabulce P2: BW, mezi-vnitro (between-within); dist, specifické rozdělení (např. negativně binomické); KR, Kenward-Roger; QL, kvazi-věrohodnost (quasilikelihood); S, Satterhwaite.

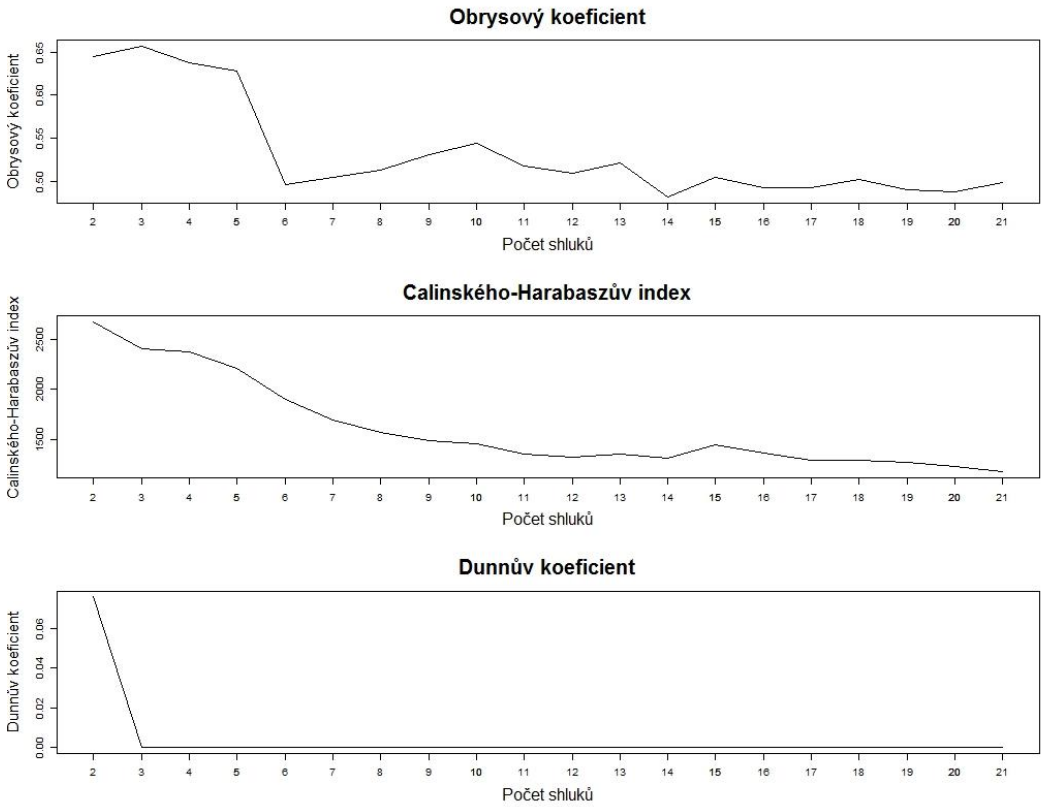
(zdroj: Bolker, a kol. (2008): Generalized linear mixed models: a practical guide for ecology and evolution. . Trends in Ecology and Evolution, roč. 24, č. 3, s. 127–135.)

Tabulka P3. Počet získaných bodů v jednotlivých scénářích pro každý testovaný shlukovací postup

Shlukovací postup	Scénář 1		Scénář 2		Scénář 3		Scénář 4	
	Součet bodů 1	Součet bodů 2	Součet bodů 1	Součet bodů 2	Součet bodů 1	Součet bodů 2	Součet bodů 1	Součet bodů 2
s1+a1	0,98	0,49	1,03	0,74	1,76	1,86	1,25	0,79
s2+a1	0,71	0,48	0,92	0,65	1,68	1,71	1,04	0,56
s3+a1	0,62	0,46	0,40	0,33	1,25	1,11	0,37	0,32
s4+a1	0,79	0,50	0,88	0,60	1,48	1,38	1,09	0,72
s5+a1	0,70	0,47	0,63	0,48	1,52	1,57	0,50	0,35
s6+a1	2,06	2,15	2,92	2,30	2,84	2,60	3,35	3,10
s1+a2	4,00	4,00	3,34	3,10	2,10	2,28	2,60	2,68
s2+a2	4,00	4,00	3,78	3,26	2,10	2,29	2,65	2,71
s3+a2	4,00	4,00	2,63	2,04	2,14	2,30	2,60	2,63
s4+a2	4,00	4,00	3,69	3,18	2,04	2,26	2,54	2,64
s5+a2	4,00	4,00	2,12	1,73	1,99	2,21	2,30	2,39
s6+a2	3,44	3,43	3,51	3,32	3,68	3,99	4,00	4,00
s1+a3	0,03	0,03	0,12	0,12	0,05	0,07	0,14	0,11
s2+a3	0,17	0,12	0,04	0,07	0,22	0,32	0,21	0,23
s3+a3	0,02	0,05	0,10	0,09	0,17	0,31	0,29	0,27
s4+a3	0,18	0,12	0,10	0,11	0,20	0,34	0,10	0,11
s5+a3	0,00	0,01	0,26	0,22	0,11	0,20	0,18	0,20
s6+a3	0,57	0,42	1,39	1,11	1,32	0,83	1,94	1,53
s1+a4	2,08	1,81	2,76	2,50	2,18	2,48	2,70	2,66
s2+a4	3,46	3,37	3,21	2,83	2,81	2,86	2,89	2,73
s3+a4	3,04	2,09	3,30	2,93	2,85	2,90	2,71	2,51
s4+a4	3,79	3,75	3,54	3,16	1,86	2,11	2,17	2,16
s5+a4	2,85	1,87	2,93	2,68	2,18	2,48	2,46	2,47
s6+a4	3,42	3,37	2,84	1,75	3,68	3,99	3,59	3,47
s1+a5	4,00	4,00	2,87	2,25	2,07	2,27	2,57	2,66
s2+a5	3,89	4,00	2,83	2,29	2,07	2,27	2,60	2,68
s3+a5	2,79	1,96	2,11	1,49	2,07	2,29	2,57	2,63
s4+a5	3,83	4,00	2,74	2,26	2,02	2,27	2,54	2,62
s5+a5	3,99	4,00	1,80	1,52	1,98	2,20	2,27	2,29
s6+a5	3,44	3,37	2,70	2,23	2,79	2,60	3,63	3,35
s1+a6	4,00	4,00	3,43	3,21	2,14	2,34	2,56	2,66
s2+a6	4,00	4,00	3,80	3,39	2,14	2,34	2,64	2,71
s3+a6	4,00	4,00	2,93	2,44	2,18	2,35	2,60	2,67
s4+a6	4,00	4,00	3,76	3,27	2,04	2,27	2,57	2,63
s5+a6	4,00	4,00	2,36	1,85	2,01	2,25	2,46	2,55
s6+a6	3,43	3,37	3,37	3,21	3,68	4,00	3,86	3,93
s1+a7	3,09	1,65	3,81	2,48	3,85	3,20	3,22	3,05
s2+a7	3,22	1,70	3,98	4,00	3,83	3,15	3,32	3,10
s3+a7	3,22	1,66	3,88	2,58	4,00	3,24	3,25	3,07
s4+a7	3,38	4,00	3,88	2,57	4,00	3,20	3,28	3,10
s5+a7	3,44	4,00	3,88	2,51	3,87	3,16	3,22	3,05
s6+a7	3,19	1,65	3,73	2,43	3,95	3,21	3,22	3,02

Tabulka P3. Pokračování

Shlukovací postup	**Scénář 5**		**Scénář 6**		**Scénář 7**		**Scénář 8**	
	Součet bodů 1	Součet bodů 2	Součet bodů 1	Součet bodů 2	Součet bodů 1	Součet bodů 2	Součet bodů 1	Součet bodů 2
s1+a1	2,56	2,55	0,68	0,61	2,75	2,76	1,69	1,67
s2+a1	2,55	2,54	1,20	1,07	2,80	2,86	1,62	1,62
s3+a1	2,28	2,11	0,38	0,41	2,61	2,54	0,42	0,37
s4+a1	2,40	2,52	1,13	1,10	2,78	2,85	1,40	1,35
s5+a1	2,32	2,30	0,31	0,40	2,71	2,77	0,72	0,81
s6+a1	3,86	3,81	2,14	2,10	2,74	2,74	3,42	3,36
s1+a2	4,00	4,00	1,78	1,71	3,10	3,01	2,62	2,57
s2+a2	4,00	4,00	2,88	2,67	3,48	3,42	2,72	2,78
s3+a2	4,00	4,00	2,46	2,39	3,46	3,37	2,38	2,48
s4+a2	2,49	2,53	1,95	2,01	2,95	3,02	2,57	2,57
s5+a2	2,36	2,40	0,79	0,74	2,76	2,82	2,21	2,25
s6+a2	3,88	3,84	2,72	2,72	3,73	3,71	4,00	4,00
s1+a3	0,03	0,06	0,08	0,11	0,00	0,00	0,12	0,13
s2+a3	0,32	0,29	0,10	0,14	0,58	0,55	0,06	0,07
s3+a3	0,57	0,42	0,42	0,45	1,40	1,36	0,72	0,63
s4+a3	0,23	0,18	0,00	0,00	0,96	1,00	0,31	0,26
s5+a3	0,07	0,06	0,24	0,25	0,46	0,50	0,84	0,81
s6+a3	3,34	2,66	2,09	2,14	2,62	2,67	3,14	3,13
s1+a4	3,83	3,81	2,67	2,70	3,00	2,94	3,17	3,10
s2+a4	3,98	4,00	3,69	3,57	3,72	3,63	3,17	3,13
s3+a4	4,00	4,00	3,11	3,01	3,95	4,00	3,13	3,02
s4+a4	2,44	2,52	2,15	2,06	2,78	2,84	2,79	2,76
s5+a4	2,43	2,54	1,25	1,27	2,46	2,60	2,57	2,60
s6+a4	3,86	3,81	2,74	2,83	3,02	2,92	3,70	3,68
s1+a5	3,78	4,00	2,28	2,11	2,78	2,83	2,60	2,70
s2+a5	3,98	4,00	2,42	2,18	2,84	2,88	2,61	2,68
s3+a5	3,98	4,00	2,11	1,74	2,68	2,67	2,50	2,48
s4+a5	2,47	2,49	1,53	1,56	2,68	2,77	2,55	2,64
s5+a5	2,45	2,48	1,34	1,38	2,69	2,73	2,23	2,26
s6+a5	4,00	4,00	3,20	2,82	2,79	2,81	3,29	3,28
s1+a6	4,00	4,00	2,86	2,84	3,19	3,17	2,72	2,79
s2+a6	4,00	4,00	3,00	2,87	3,54	3,57	2,90	2,94
s3+a6	4,00	4,00	2,59	2,52	3,58	3,63	2,78	2,81
s4+a6	2,47	2,52	1,71	1,78	2,82	2,89	2,72	2,83
s5+a6	2,44	2,47	1,49	1,53	2,80	2,89	2,49	2,55
s6+a6	4,00	4,00	4,00	4,00	3,10	2,94	3,64	3,64
s1+a7	2,98	2,26	3,59	3,15	3,86	3,62	3,23	3,26
s2+a7	3,04	2,27	3,59	3,16	3,79	3,53	3,23	3,25
s3+a7	3,26	2,29	3,57	3,10	4,00	3,76	3,23	3,30
s4+a7	3,20	2,27	3,56	3,02	3,84	3,61	3,23	3,22
s5+a7	2,94	2,27	3,64	3,23	3,83	3,57	3,23	3,28
s6+a7	3,06	2,27	3,56	3,11	3,88	3,60	3,23	3,28



Obrázek P1 Hodnoty kritérií pro určení optimálního počtu shluků.