MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Ústav matematiky a statistiky

Diplomová práce

Brno 2020 **Stanislava Nesvadbová**

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Analýza funkcionálních dat v regionálních klimatických modelech

Diplomová práce

Stanislava Nesvadbová

Vedoucí práce: doc. Mgr. Jan Koláček, Ph.D. **Brno 2020**

Bibliografický záznam

Autor:	Stanislava Nesvadbová Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Analýza funkcionálních dat v regionálních klimatických modelech
Studijní program:	Matematika
Studijní obor:	Statistika a analýza dat
Vedoucí práce:	doc. Mgr. Jan Koláček, Ph.D.
Akademický rok:	2019/20
Počet stran:	xii + 48
Klíčová slova:	funkcionální analýza dat; vyhlazování; Fourierova báze; regularizace; klimatický model; funkcionální analýza rozptylu; regresní parametry; funkcionální lineární model

Bibliografický záznam

Autor:	Stanislava Nesvadbová Prírodovedecká fakulta, Masarykova univerzita Ústav matematiky a štatistiky
Názov práce:	Analýza funkcionálnych dát v regionálnych klimatických modeloch
Študijný program:	Matematika
Študijný odbor:	Štatistika a analýza dát
Vedúci práce:	doc. Mgr. Jan Koláček, Ph.D.
Akademický rok:	2019/20
Počet strán:	xii + 48
Kľúčové slová:	funkcionálna analýza dát; vyhladzovanie; Fourierova báza; regularizácia; klimatický model; funkcionálna analýza rozptylu; regresné parametre; funkcionálny lineárny model

Bibliographic Entry

Author:	Stanislava Nesvadbová Faculty of Science, Masaryk University Department of Mathematics and Statistics

Title of Thesis:	Functional data analysis in regional climate models
Degree Programme:	Mathematics
Field of Study:	Statistics and Data Analysis
Supervisor:	doc. Mgr. Jan Koláček, Ph.D.
Academic Year:	2019/20
Number of Pages:	xii + 48
Keywords:	functional data analysis; smoothing; Fourier basis; regularization; climate model; regression parameters; functional analysis of variance; functional linear model

Abstrakt

Tato diplomová práce se zabývá analýzou funkcionálních dat pro klimatické modely. Zaměřuje se především na popis teorie analýzy funkcionálních dat, konkrétně na reprezentaci a vyhlazování dat a popis funkcionálního lineárního modelu. Tato teorie se pak aplikuje na reálná data s využitím balíku "fda" jazyku R. Primárním cílem práce je nalezení statistického modelu, který popisuje vztahy mezi regionálními a globálními klimatickými modely. Tyto vztahy jsou modelovány pomocí funkcionální analýzy rozptylu. Získané odhady jsou nakonec porovnány s výsledky z metody pro doplňování dat, která již byla dříveji aplikována na uvedená data.

Abstrakt

Táto diplomová práca sa venuje analýze funkcionálnych dát pre klimatické modely. Zameriava sa predovšetkým na popis teórie analýzy funkcionálnych dát, konkrétne na reprezentáciu a vyhladzovanie dát a popis funkcionálneho lineárneho modelu. Táto teória sa potom aplikuje na reálne dáta s využitím balíka "fda"jazyka R. Primárnym cieľom práce je nájdenie štatistického modelu, ktorý popisuje vzťahy medzi regionálnymi a globálnymi klimatickými modelmi. Tieto vzťahy sú modelované pomocou funkcionálnej analýzy rozptylu. Získané odhady sa nakoniec porovnajú s výsledkami z metódy pre dopĺňanie dát, ktorá už bola aplikovaná na uvedené dáta v minulosti.

Abstract

This diploma thesis deals with the analysis of functional data for climate models. It focuses mainly on the description of the functional data analysis theory, specifically on the representation and smoothing of data and description of the functional linear model. This theory is then applied on real data using the ”fda” package of the R language. The primary aim of this work is to find a statistical model which describes relationships between regional and global climate models. These relationships are modeled using functional analysis of variance. The estimates obtained by this model are then compared with the results of method of filling in the missing data, which has already been applied to the data.

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2017/2018

Ústav: Ústav matematiky a statistiky

Studentka: Bc. Stanislava Nesvadbová

Program: Matematika

Obor: Statistika a analýza dat

Ředitel *Ústavu matematiky a statistiky* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

Název práce: Analýza funkcionálních dat v regionálních klimatických modelech

Název práce anglicky: Functional data analysis in regional climate models

Oficiální zadání:
Při zkoumání vztahů a závislostí mezi regionálními klimatickými modely se ukazuje užitečná analýza funkcionálních dat. Student v práci nejprve popíše teorii analýzy funkcionálních dat, která patří k aktuálním trendům matematické statistiky. Tuto teorii využije k analýze dat. Zaměří se na implementaci dat do balíku " fda" jazyka R, volbu vhodné báze, semimetriky a interpretaci výsledků. Cílem práce by měl být statistický model popisující vztahy mezi regionálními klimatickými modely.

Literatura:
RAMSAY, J. O. a B. W. SILVERMAN. *Applied functional data analysis : methods and case studies*. New York: Springer, 2002. x, 190. ISBN 0387954147.

RAMSAY, J. O. a B. W. SILVERMAN. *Functional data analysis*. New York: Springer-Verlag, 1997. xiv, 310. ISBN 0387949569.

FERRATY, Frédéric a Philippe VIEU. *Nonparametric functional data analysis : theory and practice*. New York: Springer, 2006. xx, 258. ISBN 0387303693.

The Oxford handbook of functional data analysis. Edited by Frédéric Ferraty - Yves Romain. New York: Oxford University Press, 2011. xvi, 494. ISBN 9780199568444.

Marie Strnádková, Analýza rozptylu a chybějící data v klimatických modelech, bakalářská práce.

Jazyk závěrečné práce: slovenština

Vedoucí práce: doc. Mgr. Jan Koláček, Ph.D.

Datum zadání práce: 31. 8. 2017

V Brně dne: 16. 11. 2017

Souhlasím se zadáním (podpis, datum):

.....	.....	.....
Bc. Stanislava Nesvadbová **studentka**	doc. Mgr. Jan Koláček, Ph.D. **vedoucí práce**	prof. RNDr. Jan Slovák, DrSc. **ředitel Ústavu matematiky a statistiky**

Poďakovanie

Na tomto mieste by som sa chcela poďakovať doc. Mgr. Janovi Koláčkovi, Ph.D. za vedenie diplomovej práce, poskytnuté študijné materiály a za cenné rady v priebehu jej písania. Poďakovanie patrí aj mojej rodine za podporu počas celého štúdia.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracovala samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Brno 31. května 2020

. . . . . .

Stanislava Nesvadbová

Obsah

Prehľad použitého značenia

Pre jednoduchšiu orientáciu v texte predkladáme prehľad základného značenia, ktoré sa v celej práci vyskytuje.

X	funkcionálna náhodná veličina
χ	pozorovaná funkcionálna náhodná veličina
$D^m x(t)$	m-tá derivácia funkcie x v bode t
λ	vyhladzovací parameter
$PEN(x)$	penalizačná funkcia
$\mathrm{tr}(A)$	stopa matice A
RCM	regionálny klimatický model
GCM	globálny klimatický model

Úvod

Hlavnou témou tejto diplomovej práce je analýza funkcionálnych dát, ktorá predstavuje jeden z aktuálnych smerov matematickej štatistiky. Medzi odvetviami, kde nachádza svoje uplatnenie, je nepochybne aj klimatológia. Práve v súvislosti s ňou sa budeme funkcionálnymi dátami zaoberať.

Cieľom práce je najskôr popísať teóriu analýzy funkcionálnych dát a následne aplikovať tieto poznatky na reálne dáta z oblasti klimatológie. V prvej kapitole sa preto pre zoznámenie s touto témou stručne popíše klimatický systém a klimatický model. Ďalšie štyri kapitoly sa venujú teórií analýzy funkcionálnych dát. Najskôr sa zadefinujú základné pojmy pre funkcionálne premenné. V ďalších častiach sú popísané postupy pre reprezentáciu dát pomocou systémov bázových funkcií spolu s dvomi vyhladzovacími metódami. Dôležitou časťou je kapitola popisujúca funkcionálny lineárny model - konkrétne model funkcionálnej analýzy rozptylu.

Poslednú kapitolu predstavuje už spomínaná aplikačná časť. Tu je realizovaný jeden z hlavných cieľov práce a to nájsť model, ktorý popisuje vzťahy medzi klimatickými modelmi. V závere tejto kapitoly a zároveň celej práce sa porovnajú výsledky získané daným modelom s tým, čo už bolo k danej téme vytvorené. Konkrétne ide o metódu dopĺňania chýbajúcich dát, ktorá bola zostrojená a realizovaná v bakalárkej práci [10].
V zadaní práce je okrem už spomenutých cieľov uvedené takisto využitie semimetrík. Avšak medzičasom už bol tento prístup využitý v [13], z toho dôvodu sa už takýmto spôsobom nepostupovalo a zvolila sa iná forma spracovania.

Kapitola 1

Klimatické modely

V úvodnej kapitole tejto práce popíšeme základné pojmy z oblasti klimatických modelov, s ktorými budeme neskôr v texte pracovať. Budeme pritom čerpať z [1],[2],[3] a [4].

1.1 Klimatický systém

Pre porozumenie klimatickému modelu je nutné najskôr zadefinovať pojmy klíma a klimatický systém.

Klíma v užšom slova zmysle je definovaná ako štatistický popis pomocou priemeru a variability relevantných kvantít za dané časové obdobie alebo veľmi zjednodušene povedané "priemerné počasie". Časovým obdobím pritom môžu byť mesiace, tisíce až milióny rokov. Za klasickú dĺžku časového obdobia je Svetovou meteorologickou organizáciou (WMO) považovaných 30 rokov. Spomínanými kvantitami sú najčastejšie povrchové premenné ako teplota, vlhkosť, atmosférický tlak, zrážky a ďalšie meteorologické ukazovatele. Klíma v širšom slova zmysle je stav klimatického systému, čo zahŕňa aj vyššie spomínaný štatistický popis.

Klimatický systém je komplexný systém pozostávajúci z piatich hlavných zložiek: atmosféra, hydrosféra, kryosféra, povrch pevniny a biosféra, a interakcie medzi nimi. Klimatický systém sa vyvíja v čase pod vplyvom svojej vlastnej vnútornej dynamiky a takisto v dôsledku vonkajších síl ako vulkanické erupcie, solárne výkyvy a ľudská aktivita, napr. meniace sa zloženie atmosféry v dôsledku znečistenia či zmena vo využití povrchu pevniny. Hoci zložky klimatického systému sa významne líšia vo svojej stavbe, fyzikálnych a chemických vlastnostiach, štruktúre a správaní, všetky sú previazané pohybom hmoty a energie, pričom všetky podsystémy sú otvorené a prepojené.

1.2 Klimatický model

Klimatický model definujeme ako matematickú reprezentáciu fyzikálnych a chemických dejov prebiehajúcich v klimatickom systéme. Klimatické modely poskytujú výsledky, ktoré sú diskrétne v čase a priestore, čo znamená, že reprezentujú priemery pre oblasti, ktorých veľkosti závisia na rozlíšení modelu. Modely sa líšia vo svojej komplexnosti, počnúc od najjednoduchších, bez dimenzií, až po zložité, komponované z viacerých podmodelov.

Využívajú sa na množstvo účelov od štúdia dynamiky klimatického systému po predpovede budúceho stavu klímy. Práve predpoveď patrí k najdôležitejším cieľom klimatológie.

1.2.1 Globálny klimatický model

Najsofistikovanejším súčasným typom modelov sú tzv. GCM (*z angl. Global Climate Models/General Circulation Models*), teda Globálne klimatické modely alebo Všeobecné modely cirkulácie. Zatiaľ čo prvý z pojmov je všeobecnejší, v praxi sú oba názvy často používané v rovnakom zmysle a zameniteľné.

Pod GCM, pôvodne sa zaoberajúci predovšetkým cirkuláciou atmosféry, klimatológovia postupom času zahrnuli viacero zložiek zemského systému. Zemský systém pozostáva z pevniny, oceánov, atmosféry a ľadu (pólov). Zahŕňa prirodzené cykly planéty (cyklus vody, dusíka, uhlíka, atď.) a takisto hlboké zemské procesy. Tieto boli kedysi simulované pomocou samostatných modelov. Globálny klimatický systém má teda niekoľko častí, štandardne ide o model atmosféry, model oceánu, model ľadu, model biosféry a sústavu chemických modelov. Najnovšie boli ako podmnožina do GCM zahrnuté aj biochemické cykly (prenos chemických látok medzi živými vecami a ich prostredím) a ich interakcia s klimatickým systémom. Kľúčovou výpočtovou časťou je však model cirkulácie atmosféry.

Cirkulácia atmosféry je spojitý proces, pri ktorom sa hodnoty jednotlivých veličín (rýchlosť prúdenia, tlak, teplota, atď.) v čase a priestore kontinuálne menia. Tieto procesy sú popísané pomocou diferenciálnych rovníc. Základom modelu cirkulácie atmosféry je sústava diferenciálnych rovníc, ktorá tvorí tzv. dynamické jadro modelu.

Problémom je skutočnosť, že v súčasnosti nevieme získať presné riešenie tejto sústavy, teda hodnoty skúmaných veličín v čase a priestore. Pre konkrétne zadanie preto odhadujeme hodnoty pomocou numerických metód. Numerické riešenie úzko súvisí s chybami klimatických modelov. Diferenciálne rovnice opisujú, ako sa skúmaná veličina vyvíja na nekonečne malom priestorovom alebo časovom úseku. Princípom numerického riešenia je zväčšenie tohto limitne malého úseku na konečný. V priestore sa vytvorí sieť vzájomne prepojených bodov vzdialených od seba konečnú dĺžku a približné riešenie hľadáme v uzloch tejto siete. V priebehu výpočtu každý uzol siete vygeneruje jednu lineárnu rovnicu, z ktorej získavame hodnotu veličiny v uzlovom bode siete.

GCM vyobrazujú stav klimatického systému pomocou trojrozmernej mriežky pokrývajúcej celú zemeguľu (Obr. 1.1), ktorá má typicky horizontálne rozlíšenie od 250 do 600 km (10 až 20 vertikálnych vrstiev v atmosfére, v oceánoch niekedy až 30).

1.2.2 Regionálny klimatický model

V našej práci sa budeme primárne venovať Regionálnym klimatickým modelom RCM (Regional Climate Model). Fungujú podobne ako GCM, ale sú limitované len na obmedzenú časť Zeme, región.

Lokálne klimatické zmeny sú významne ovplyvnené lokálnymi topografickými prvkami, napr. pohoriami. GCM nie sú schopné vziať do úvahy tieto prvky, keďže používajú relatívne nízke rozlíšenie. Regionálny klimatický model je klimatický model s vyšším priestorovým rozlíšením ako globálny. Vyššie rozlíšenie modelu v tomto prípade znamená

Obr. 1.1: Sieť uzlov okolo Zeme (Zdroj: [4])

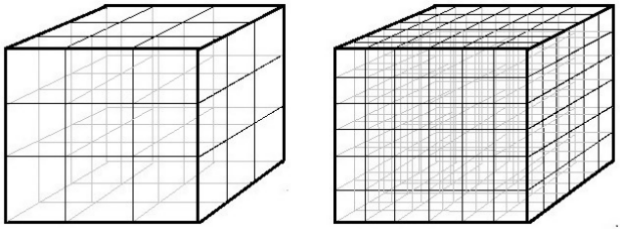
menšie vzdialenosti uzlov siete, a väčší vplyv topografických objektov. Toto môže byť veľmi užitočné napríklad v pobrežných oblastiach alebo pohoriach, kde môže vzdialenosť 50 km znamenať výraznú zmenu v počasí. S rastúcim počtom uzlov však rastie výpočtová náročnosť a práve preto sa RCM obmedzujú len na špecifickú oblasť, typicky približne 5000km x 5000km, teda napr. oblasť veľkosti východnej Európy alebo južnej Afriky.

Nevýhodou je, že keďže RCM pokrývajú iba obmedzenú oblasť, hodnoty na jej hraniciach musia byť explicitne dané. Na to aby mohli regionálne modely simulovať klímu vnútra daného úseku, preto závisia na informácií dostupnej v horizontálnych hraniciach, označovanej ako hraničné podmienky. Preto sa musia RCM v snahe o simuláciu počasia pre konkrétnu geografickú oblasť spoliehať na GCM, aby im poskytli hraničné podmienky v pravidelných intervaloch (napríklad šesťhodinových), aby regionálny model mohol vytvoriť detaily o regionálnej klímy vo svojej oblasti.

Je preto dôležité povedať, že regionálne modely nenahrádzajú globálne modely, ale môžu poskytnúť pridanú informáciu k simuláciám vykonaným s globálnymi modelmi.

Problémom pri zmenšovaní výpočetných krokov modelu je, že so zjemňovaní siete velmi rýchlo rastú nároky na výpočetný výkon. Pre predstavu o tejto vzrastajúcej náročnosti si vezmime z výpočetnej siete modelu kocku, ktorej hrana má dĺžku trojnásobku vzdialenosti uzlových bodov. Takáto sieť (Obr. 1.2 vľavo) obsahuje 64 uzlových bodov. Pre spresnenie výsledkov môžeme skrátiť vzdialenosť uzlov na polovicu. Dostávame tak sieť (Obr. 1.2 vpravo), ktorá už obsahuje 343 bodov. Dá sa ukázať všeobecný vzťah. Označme počet uzlov v pôvodnej sieti n. Zmenšením vzdialenosti medzi uzlovými bodmi k-krát, získame novú sieť obsahujúcu $((\sqrt[3]{n}-1)\,k+1)^3$.

Nárast súvisí s faktom, ze modely pracujú v trojrozmernom priestore a berúc do úvahy, že výpočetnou oblasťou je celá zemeguľa, je zrejmé, že pri snahe o čo najdetailnejšiu modelovú sieť môžeme rýchlo prísť na hranicu výpočetných možností súčasných počítačov.



Obr. 1.2: Zjemnenie výpočetnej siete. (Zdroj:[2])

1.2.3 Ďalšie druhy modelov

- **EBM** (Energy Balance Models) Najstaršími a najzákladnejšími klimatickými modelmi sú EBM - Modely rovnováhy energie. EBM nesimulujú klímu, ale miesto toho berú do úvahy rovnováhu medzi energiou vstupujúcou do atmosféry zo Slnka a energiou (teplom), ktoré sa uvoľňuje späť do vesmíru. V ich najjednoduchšej forme, pritom berú Zem ako jeden bod, poskytujúc iba globálne spriemerované hodnoty vypočítaných premenných.

- **RCM** (Radiative Convective Models) Tieto modely simulujú prenos energie vertikálne naprieč atmosférou, napríklad stúpanie teplého vzduchu smerom nahor. Dokážu vypočítať teplotu a vlhkosť rôznych vrstiev atmosféry. Tieto modely sú zvyčajne jednodimenzionálne - berúc do úvahy iba presun energie iba vo zvislom smere - ale môžu byť rozšírené.

- **IAM** (Integrated Assessement Model) Tieto modely zahŕňajú do klimatického modelu aj aspekty spoločnosti, simulujúc ako populácia, ekonomický rast a využitie energie ovplyvňujú klímu a interagujú s ňou.

Kapitola 2

Analýza funkcionálnych dát

V nasledujúcich štyroch kapitolách popíšeme vybrané časti z teórie analýzy funkcionálnych dát. Pri definovaní jednotlivých pojmov a opise postupov budeme čerpať predovšetkým z [5], [6], [8] a [7], a využijeme tiež [12], [10] a [11].

Zoznámme sa teraz s hlavnou témou práce. V štatistike bežne pracujeme s pozorovaniami, kde jednému elementu prislúcha jedna alebo viacero samostatných hodnôt. Avšak pokrok vo výpočtových nástrojoch v súčasnosti umožňuje zbierať, ale aj spracovávať čoraz viac údajov a pre jeden jav tak často pozorujeme veľké množstvo premenných. V mnohých odvetviach, vrátane klimatológie, tak vznikajú situácie, keď sa ukazuje užitočnejšie pri pozorovaniach uvažovať o všeobecnejších objektoch ako len o diskrétnych bodoch.

Majme napríklad premennú, ktorú pozorujeme v rôznych časových okamihoch na nejakom intervale (t_{min}, t_{max}). Tieto pozorovania môžeme vyjadriť aj formou zobrazenia $\{X(t_j)\}_{j=1,\ldots,J}$. Pri určitej hustote a blízkosti pozorovaní je potom rozumné začať o dátach uvažovať ako o spojitom zobrazení $\mathrm{X} = X(t)$; $t \in (t_{min}, t_{max})$, kedy hodnoty pozorujeme na nejakom kontinuu t. Tým býva zvyčajne čas, ale stretneme sa aj s vlnovou dĺžkou, pravdepodobnosťou a inými.

Rastúci počet takýchto situácií motivoval vývoj smeru, ktorý nazývame *analýza funkcionálnych dát*. Hlavnou myšlienkou je pozerať sa na dáta skôr ako na samostatné objekty - funkcie - než ako na postupnosti pozorovaní.

2.1 Funkcionálna náhodná veličina a funkcionálny náhodný výber

Chceme teda ako o náhodných veličinách uvažovať o funkciách. Známe pojmy náhodná veličina a náhodný výber môžeme rozšíriť na nekonečne rozmerné priestory a definovať tak funkcionálnu náhodnú veličinu a funkcionálny náhodný výber.

Definícia 1. *Náhodná veličina* X *sa nazýva funkcionálna náhodná veličina (f.n.v.), ak berie svoje hodnoty v nekonečne rozmernom priestore (alebo funkcionálnom priestore) E. Pozorovania χ_i sú realizáciami f.n.v* X *a nazývame ich funkcionálne dáta.*

Definícia 2. *Funkcionálnym náhodným výberom nazývame n funkcionálnych náhodných veličín* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *rovnako rozdelených ako f.n.v.* X. *Realizáciou funkcionálneho náhodného výberu je funkcionálny štatistický súbor, ktorý značíme* $\chi_1, \ldots, \chi_n$.

V prípade, že funkcionálnou náhodnou veličinou, resp. jej pozorovaním, je krivka, dostávame implicitne vzťah $\mathrm{X} = \{\mathrm{X}(t);\ t \in T\}$. Krivka je však iba špecifickým jednorozmerným prípadom, kde $T \subset \mathbb{R}$. Je dôležité spomenúť, že označenie funkcionálna premenná zahŕňa širšiu skupinu útvarov ako iba krivky. Môže ísť o rôzne zložité matematické objekty. Ďalej v tejto práci sa však budeme zaoberať iba krivkami.

2.2 Vybrané súhrnné štatistiky

Uvedieme tu základné popisné štatistiky, s ktorými sa v tejto práci viackrát stretneme. Definujeme si ich empirické verzie.

Nech $\{\chi_1, \ldots, \chi_n\}$ je výber n funkcionálnych pozorovaní. Výberový priemer potom definujeme vzťahom

$$\overline{\chi}(t) = \frac{1}{n}\sum_{i=1}^{n} \chi_i(t) \text{ pre } \forall t \in \mathbb{R}.$$

Výberový rozptyl je definovaný

$$\mathrm{var}(t) = \frac{1}{n-1}\sum_{i=1}^{n} [\chi_i(t) - \overline{\chi}(t)]^2 \text{ pre } \forall t \in \mathbb{R}.$$

A nakoniec výberová kovariancia

$$\mathrm{covar}(t) = \frac{1}{n-1}\sum_{i=1}^{n} [\chi_i(s) - \overline{\chi}(s)][\chi_i(t) - \overline{\chi}(t)] \text{ pre } \forall s,t \in \mathbb{R}.$$

Kapitola 3

Reprezentácia dát

Dáta, s ktorými pracujeme vo funkcionálnej analýze zvyčajne nepozorujeme priamo v podobe funkcií. Predpokladáme však, že existuje nejaká funkcia x, ktorá dáva vznik týmto hodnotám. Zároveň nedisponujeme pozorovaniami v každom bode t, ale iba odmedzeným počtom meraní. Pre jedno pozorovanie funkcie x_i máme zvyčajne záznam pozostávajúci z n párov (t_{ij}, y_{ij}), $j = 1, \ldots, n$. Jedným z prvých krokov vo funkcionálnej analýze je nájsť pre pozorované body ich funkcionálne vyjadrenie.

Funkciu x_i zvyčajne zostrojujeme osobitne pre každú sériu pozorovaní (t_{ij}, y_{ij}), a to nezávisle na i. Pre zjednodušenie budeme teda v tejto časti predpokladať, že odhadujeme jednu samostatnú funkciu a používať zjednodušené značenie x.

Je dôležité podotknúť, že ešte pred prevedením do funkcionálnej podoby predpokladáme určitú úroveň hladkosti dát. V prípade, že by za sebou idúce pozorované hodnoty boli navzájom príliš vzdialené, by nemalo veľký zmysel pristupovať k premenným ako k funkcionálnym, ale postačoval by prístup ako v prípade viacerých nezávislých premenných.

3.1 Systém bázových funkcií

Pre vyjadrenie hľadaných funkcií použijeme reprezentáciu pomocou *systému bázových funkcií*. Ide o užitočnú metódu, ktorá umožňuje uchovať informácie o dátach a zároveň poskytuje potrebnú flexibilitu. Je to nástroj s dostatočnou výpočtovou silou pre prácu s veľkým množstvom údajov.

Pri stavbe funkcie x používame množinu tzv. *bázových funkcií*. Ide o množinu známych funkcií $\{\phi_k\}_{k \in \mathbf{N}}$ s vlastnosťou, že ak vezmeme lineárnu kombináciu dostatočne veľkého počtu K týchto funkcií, každú funkciu vieme ľubovoľne dobre aproximovať. S jej pomocou potom môžeme funkciu $x(t)$ vyjadriť v tvare lineárneho rozvoja ako

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t),$$

ktorý nazývame *rozvoj bázových funkcií* v bode t. Parametre $c_1, c_2, \ldots, c_K$ sú koeficientami rozvoja.

Situáciu, kedy máme súbor N funkcií a ich vyjadrení $x_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t)$, $i = 1, \ldots, N$, môžeme zapísať v maticovom tvare ako

$$\mathbf{x}(t) = \mathbf{C}\boldsymbol{\phi}(t),$$

kde $\boldsymbol{x}(t)$ je vektor dĺžky N, $\boldsymbol{C}$ matica koeficientov s rozmermi $N \times K$ a $\boldsymbol{\phi}(t)$ vektor bázových funkcií dĺžky K. V tejto časti však nadalej uvažujeme iba samostatnú funkciu x.

Pre prehľadnosť a jednotnosť s ďalšími operáciami si ešte zavedieme operátor D pre m-tú deriváciu

$$D^m x(t) = \frac{d^m x(t)}{dt^m} = x^{(m)}(t), \quad m \geq 1$$

a platí $D^0 x(t) = x(t)$ a $D^{-1} f(t) = \int x(s)ds$. Spomeňme, že využitie derivácií je jednou z charakteristických čŕt analýzy funkcionálnych dát.

3.2 Typy báz

Výber bázy predstavuje podstatný krok pri analyzovaní funkcionálnych dát. Závisí na konkrétnom type dát, s ktorými pracujeme. V tejto práci sa bližšie pozrieme na dva v súčasnosti najpoužívanejšie systémy bázových funkcií: *Fourierovu bázu* a *B-splajnovú bázu*.

3.2.1 Fourierova báza

Definujeme teda najskôr Fourierovu bázu zloženú z konštantnej funkcie a zo sínov a kosínov:

$$\begin{aligned} \phi_0(t) &= 1 \\ \phi_{2r-1}(t) &= \sin(r\omega t), \qquad \text{pre } r = 1,2,\ldots \\ \phi_{2r}(t) &= \cos(r\omega t). \end{aligned}$$

Ako názov napovedá, báza súvisí s Fourierovým radom, a to tak, že $x(t)$ môžeme vyjadriť pomocou konečného počtu členov Fourierovho radu ako

$$x(t) = c_0 + \sum_{i=1}^{M} c_{2i} \sin(i\omega t) + \sum_{i=1}^{M} c_{2i+1} \cos(i\omega t).$$

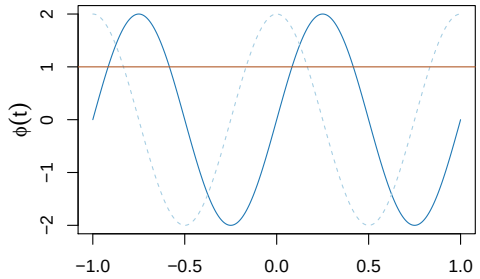
Celkový počet bázových funkcií je teda rovný $K = 2M + 1$, pri určovaní počtu funkcií totiž zvykneme uvažovať o celých pároch. V tomto prípade ide o periodickú bázu, kde je perióda T určená parametrom $\omega = 2\pi/T$. Keďže reprezentácia má periodický tvar, ide o prirodzenú voľbu pre tento typ dát.

Voľba Fourierovej bázy má niekoľko výhod súvisiacich najmä s jej výpočtovou jednoduchosťou. Uveďme napríklad jednoduchosť výpočtu derivácií:

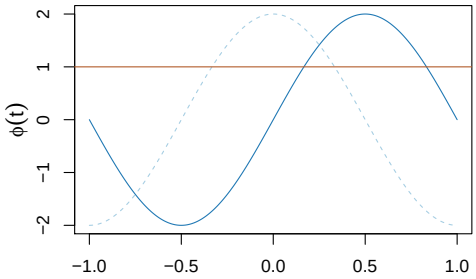
$$\begin{aligned} D\sin r\omega t &= \quad r\omega\cos(r\omega t), \\ D\cos r\omega t &= -r\omega\sin(r\omega t), \end{aligned} \tag{3.1}$$

pričom je vidieť, že aj derivácie vyšších rádov by boli pomerne ľahko spočítateľné.

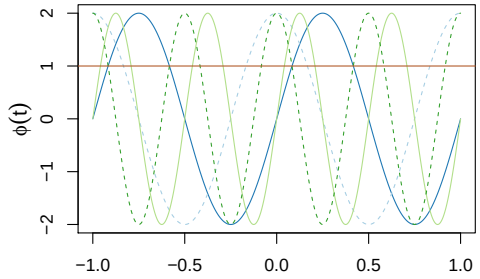
Séria grafov 3.1 až 3.4 vyobrazuje funkcie Fourierovej bázy na intervale $t \in [-1,1]$ pre rôzne hodnoty r a periódy T.



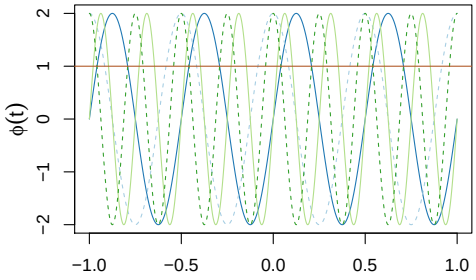
Obr. 3.1: Prvé 3 funkcie bázy; $T = 1$



Obr. 3.2: Prvé 3 funkcie bázy; $T = 2$



Obr. 3.3: Prvých 5 funkcií bázy; $T = 1$



Obr. 3.4: Prvých 5 funkcií bázy; $T = \frac{1}{2}$

3.2.2 B-splajnová báza

Najčastejšou voľbou pre neperiodické funkcionálne dáta sú splajnové bázy. Ide o bázy splajnových funkcií, ktoré sú zložené po častiach z polynómov. Vďaka tomu ponúkajú výpočtovú jednoduchosť polynómov, avšak s podstatne väčšou flexibilitou, často s využitím iba neveľkého počtu bázových funkcií.

V tejto časti najskôr popíšeme štruktúru splajnovej funkcie a potom bázový systém, ktorý sa zvyčajne používa na jej výstavbu - *B-splajn systém.*

Prvým krokom je rozdelenie intervalu, na ktorom chceme funkciu aproximovať, na L podintervalov oddelených hodnotami τ_l, $l = \{1, \ldots, L-1\}$. Deliace hodnoty nazývame *zlomové body*. Okrajové body intervalu označíme τ_0 a τ_L. Kľúčovým pojmom pri splajnových bázach sú však *uzly.* Platí, že každý uzol má hodnotu rovnú nejakému zlomovému bodu, ale naopak, v určitom bode zlomu môže byť niekoľko uzlov. Východzou situáciou vo väčšine prípadov je však jeden vnútorný uzol na jeden zlomový bod. Bez prítomnosti vnútorných uzlov sa splajnová báza redukuje na klasický polynóm.

Na každom intervale je splajnová funkcia tvorená polynómom pevného *rádu m*. Rád polynómu je počet konštánt potrebných na jeho definovanie a je o jedno vyšší než *stupeň polynómu m*. S prechodom do ďalšieho intervalu sa mení charakter polynómu, susediace polynómy sú ale spojené hladko, teda v bodoch spoja sú funkcie obmedzené tak, aby mali

rovnaké hodnoty.

Pre tvar splajnovej funkcie je rozhodujúci:

- rád polynómových segmentov m,
- postupnosťou uzlov τ_l.

Počet K bázových funkcií v splajnovom bázovom systéme je v situácií jedného uzlu na jeden zlomový bod rovný súčtu rádu polynómov a počtu vnútorných uzlov, tj.

$$K = m + L - 1.$$

Existuje viacero spôsobov, ako konštruovať systém bázových funkcií pre splajnovú bázu, jedným z najpoužívanejších je *B-splajnový bázový systém*.

Pre postupnosť uzlov τ_l definujme i-ty B-splajn rádu j funkcie pomocou rekurzívneho vzťahu ako

$$B_{i,0}(t) = \begin{cases} 1 \text{ ak } \tau_i \leq t < \tau_{i+1} \text{ a } t < \tau_{i+1}, \\ 0 \text{ inak .} \end{cases}$$

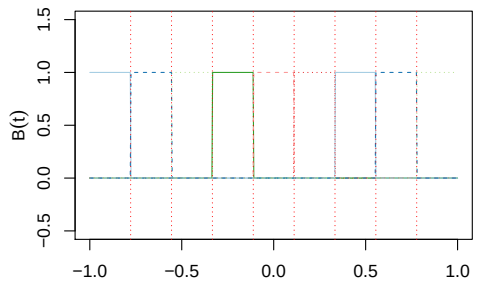
$$B_{i,j}(t) = \frac{t - \tau_i}{\tau_{i+j} - \tau_i} B_{i,j-1}(t) + \frac{\tau_{i+j+1} - t}{\tau_{i+j+1} - \tau_{i+1}} B_{i+1,j-1}(t).$$

Potom splajnová funkcia $S(t)$ tvorená B-splajnovými funkciami je definovaná ako

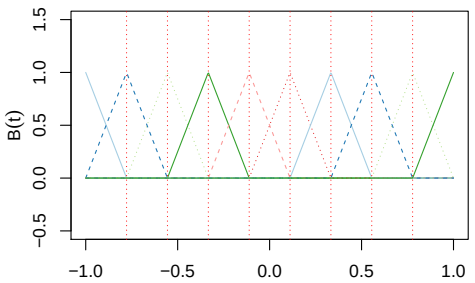
$$S(t) = x(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau),$$

kde $B_k(t, \tau)$ predstavuje hodnotu B-splajnovej funkcie v bode t definovanú postupnosťou τ_l.

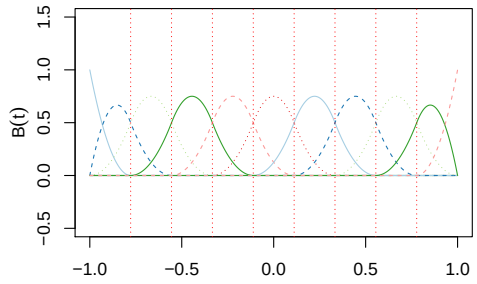
Na grafoch 3.5 až 3.8 sú bázy tvorené B-splajnovými funkciami na intervale $[-1, 1]$. Podľa vyššie uvedeného pravidla je počet funkcií K vždy rovný súčtu počtu vnútorných uzlov, ktorý bol v tomto prípade pevne zvolený 8, a rádu polynómov postupne $m = \{1, 2, 3, 4\}$.



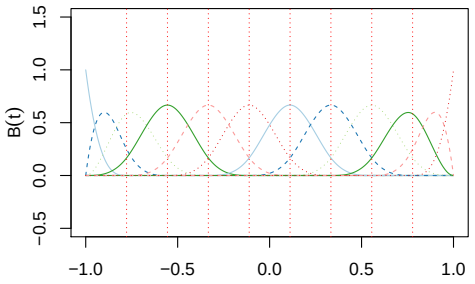
Obr. 3.5: $K = 9$ bázových funkcií; $m = 1$



Obr. 3.6: $K = 10$ bázových funkcií; $m = 2$



Obr. 3.7: $K = 11$ bázových funkcií; $m = 3$



Obr. 3.8: $K = 12$ bázových funkcií; $m = 4$

Existuje veľké množstvo ďalších báz, ktoré už nie sú tak často využívané, ale nachádzajú svoje uplatnenie v špecifických situáciách. Spomeňme napríklad najjednoduchšiu z báz - *konštantnú bázu*, tvorenú jedinou konštantnou funkciou $\boldsymbol{\phi}(t) = \{1\}$, či *monomiálnu bázu* $\boldsymbol{\phi}(t) = \{t^1, t^2, t^3, \dots\}$, ktorá pred príchodom B-splajn bázy patrila k veľmi obľúbeným bázam.

Kapitola 4

Vyhladzovanie dát

V predchádzajúcej časti sme objasnili, že zvyčajne nepozorujeme priamo funkciu $x(t)$, ale diskrétne body, pomocou ktorých chceme získať jej odhad. Ďalej treba povedať, že funkciu $x(t)$ navyše zvykneme pozorovať s nejakou chybou merania ε. Pozorovaná funkcia má tak tvar

$$y(t) = x(t) + \varepsilon(t).$$

Ďalej sa teda zameriame na metódy, ktoré pri hľadaní vyjadrenia $x(t)$ zohľadňujú aj chybu merania. Bližšie si predstavíme dva prístupy.

4.1 Lineárna regresia

Jednoduchší z týchto prístupov využíva metódu odhadu najmenších štvorcov. Pripomeňme, že vychádzame z n dvojíc pozorovaní $(y(t_j), t_j)$, pre ktoré máme vzťah $y(t_j) = x(t_j) + \varepsilon(t_j)$. Využitím týchto pozorovaní chceme získať vyrovnané (*angl. fitted*) hodnoty pozorovanej funkcie y pomocou odhadu x. Pripomeňme, že ten získame lineárnou kombináciou konečného počtu K bázových funkcií

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \boldsymbol{c}' \boldsymbol{\phi},$$

kde $\boldsymbol{c}$ je vektor koeficientov c_k a $\boldsymbol{\phi}$ vektor bázových funkcií ϕ_k. Problém hľadania vhodných koeficientov c_k môžeme previesť na hľadanie koeficientov lineárneho regresného modelu, kde je odozvou $y(t_j)$ a K premennými sú hodnoty bázových funkcií $\phi_1(t_j), \phi_2(t_j), \ldots, \phi_K(t_j)$.

Na základe popísaných vzťahov tak dostávame regresný model v tvare:

$$\begin{aligned}
y(t_1) &= x(t_1) + \varepsilon(t_1) = c_1\phi_1(t_1) + \cdots + c_k\phi_k(t_1) + \varepsilon(t_1), \\
y(t_2) &= x(t_2) + \varepsilon(t_2) = c_1\phi_1(t_2) + \cdots + c_k\phi_k(t_2) + \varepsilon(t_2), \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
y(t_n) &= x(t_n) + \varepsilon(t_n) = c_1\phi_1(t_n) + \cdots + c_k\phi_k(t_n) + \varepsilon(t_n).
\end{aligned}$$

Prevedením do maticového tvaru máme

$$\boldsymbol{y} = x(\boldsymbol{t}) + \boldsymbol{\varepsilon} = \boldsymbol{\Phi c} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\Phi}$ je matica o rozmeroch $n \times K$ obsahujúcu hodnoty $\phi_k(t_j)$,kde $x(\boldsymbol{t})$ je vektor hodnôt funkcie x pre vektor $\boldsymbol{t}$ a $\boldsymbol{\varepsilon}$ je vektor chýb. Predpokladáme, že chyby $\boldsymbol{\varepsilon}$ sú nezávislé, rovnako rozdelené náhodné veličiny s nulovou strednou hodnotou a konštantným rozptylom σ^2.

Odhad bázových koeficientov $\hat{\boldsymbol{c}}$ potom získame minimalizáciou *súčtu vyhladených štvorcov chýb* (*angl. smoothed sum of squared errors)*:

$$SMSSE(\boldsymbol{y}|\boldsymbol{c}) = \sum_{j=1}^{n} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2 \tag{4.1}$$

V maticovej forme má tento výraz tvar

$$\begin{aligned} SMSSE(\boldsymbol{y}|\boldsymbol{c}) &= (\boldsymbol{y} - \boldsymbol{\Phi c})'(\boldsymbol{y} - \boldsymbol{\Phi c}), \\ &= ||\boldsymbol{y} - \boldsymbol{\Phi c}||^2. \end{aligned}$$

Výsledný odhad koeficientov, ktorý minimalizuje (4.1), je potom

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{y},$$

a odtiaľ získaný vektor vyrovnaných hodnôt

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}\hat{\boldsymbol{c}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{y}.$$

Treba poznamenať, že nejde o veľmi komplexnú metódu a jej využitie je limitované. Nevýhodou je, že vyhladzovanie vieme regulovať len v obmedzenej miere; tento prístup funguje len v prípade, že zvolíme výrazne menší počet K bázových funkcií, než počet dostupných bodov n.

4.2 Regularizácia

O niečo silnejším nástrojom pre aproximáciu diskrétnych dát funkciou je *regularizácia* alebo *vyhladzovanie pomocou penalizačnej funkcie*. V praxi je využívanejší, než predchádzajúci spôsob, avšak ten sme si uviedli ako metódu, z ktorej budeme teraz vychádzať. Pri vyhladzovaní dát pomocou regularizácie potrebujeme najskôr určiť mieru hladkosti, resp. "hrubosti" odhadovanej funkcie pomocou tzv. *penalizačnej funkcie* $PEN(x)$. Následne minimalizujeme súčet štvorcov, ktorý ňou modifikujeme.

Na výber je viacero penalizačných funkcií, uvedieme si dve, ktoré patria k najpoužívanejším.

- **Operátor D**

  Prvý spôsob vychádza z vyjadrenia hrubosti funkcie pomocou derivácie. Za dokonale hladkú funkciu určite považujeme rovnú priamku, ktorá má deriváciu rovnú nule. Využitím operátora D, ktorý sme zaviedli v časti 3.1, môžeme definovať *zakrivenie*

krivky v bode t ako druhú mocninu jej druhej derivácie $[D^2x(t)]^2$. Za vhodnú mieru hrubosti potom považujeme

$$PEN_2(x) = \int [D^2x(s)]^2 ds,$$

keďže všade tam, kde je funkcia veľmi premenlivá, má výraz $[D^2x(t)]^2$ veľkú hodnotu, ktorá bude penalizovaná.

Zovšeobecnením tohto pojmu pre m-tú deriváciu dostávame

$$PEN_m(x) = \int [D^m x(s)]^2 ds.$$

- **Harmonický akcelerátor**

  Hladkosť však nemusí vždy súvisieť iba s deriváciou funkcie. O niečo všeobecnejšie môžeme tvrdiť, že ide o odklon pozorovanej funkcie od nejakej "základnej verzie" krivky. Napríklad pri periodických dátach môže byť takouto krivkou sínusoida. Zaveďme *harmonický akcelerátor* v tvare $L = \omega^2 D + D^3$, pre ktorý platí $Lsin(\omega x) = 0 = Lcos(\omega x)$. Príslušná penalizačná funkcia je potom

$$PEN_L(x) = \int [Lx(s)]^2 ds.$$

4.2.1 Penalizované vyhladzovanie

Pri vyhladzovaní pomocou penalizačnej funkcie chceme modifikovať súčet najmenších štvorcov tak, aby táto funkcia mala vplyv na odhad $x(t)$. Musíme sa pritom vysporiadať s dvomi protichodnými cieľmi: získať čo možno najhladšiu funkciu a zároveň dostatočne kopírovať pôvodné dáta tak, aby sme uchovali čo najviac informácií. Nasledujúci výraz predstavuje kompromis medzi týmito požiadavkami. Definujeme tak *penalizovaný súčet štvorcov chýb* (*angl. penalized sum of squared errors)* ako

$$\begin{aligned} PENSSE_\lambda(x|\boldsymbol{y}) &= \sum_{j=1}^{n} [y_j - \sum_{k}^{K} c_k \phi_k(t_j)]^2 + \lambda PEN(x) \\ &= (\boldsymbol{y} - x(\boldsymbol{t}))'(\boldsymbol{y} - x(\boldsymbol{t})) + \lambda PEN(x) \\ &= ||\boldsymbol{y} - x(\boldsymbol{t})||^2 + \lambda PEN(x), \end{aligned} \tag{4.2}$$

kde $x(\boldsymbol{t})$ je vektor hodnôt funkcie x pre vektor $\boldsymbol{t}$. Parameter λ nazývame *vyhladzovcí parameter*. Určuje pomer medzi variabilitou funkcie x a blízkosťou odhadu k pozorovaným hodnotám. S rastúcou hodnotou parametra λ sú nehladké funkcie vo väčšej miere penalizované a *PENSEE* preto musí klásť väčší dôraz na vyhladenie, než na presný odhad dát. Preto pre $\lambda \to \infty$ sa odhadovaná krivka blíži k štandardnej krivke lineárnej regresie. Naopak, pre $\lambda \to 0$ je v odhade čoraz viac variability a blížime sa k interpolácií dát.

Ukážme si teraz postup minimalizácie pre konkrétnu penalizačnú funkciu. Zvoľme $PEN_m(x)$. Vieme, že $x(t) = \boldsymbol{c}'\boldsymbol{\phi}(t)$. Penalizačnú funkciu potom vyjadríme v maticovom

tvare nasledovne:

$$\begin{aligned} PEN_m(x) &= \int [D^m x(s)]^2 ds \\ &= \int [D^m \boldsymbol{c}'\boldsymbol{\phi}(s)]^2 ds \\ &= \int \boldsymbol{c}' D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s)\, \boldsymbol{c}\, ds \\ &= \boldsymbol{c}' [\int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) ds] \boldsymbol{c} \\ &= \boldsymbol{c}'\boldsymbol{R}\boldsymbol{c}, \end{aligned}$$

Maticu $\boldsymbol{R}$ nazývame *penalizačná matica* a z predchádzajúceho vzťahu vidíme, že

$$\boldsymbol{R} = \int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) ds.$$

Súčtom sumy štvorcov chýb $SSE(\boldsymbol{y}|\boldsymbol{c})$ a λ-násobku penalizačnej funkcie $PEN_m(x)$ tak dostávame vyjadrenie pre (4.2) v tvare

$$\begin{aligned} PENSSE_m(\boldsymbol{y}|\boldsymbol{c}) &= (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c})'(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c}) + \lambda \boldsymbol{c}'\boldsymbol{R}\boldsymbol{c} \\ &= ||\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c}||^2 + \lambda \boldsymbol{c}'\boldsymbol{R}\boldsymbol{c} \end{aligned}$$

Odhad koeficientov $\hat{\boldsymbol{c}}$ pomocou penalizovaných najmenších štvorcov je

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda \boldsymbol{R})^{-1} \boldsymbol{\Phi}'\boldsymbol{y},$$

s vektorom odhadnutých hodnôt

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda \boldsymbol{R})^{-1} \boldsymbol{\Phi}'\boldsymbol{y}.$$

Vektor $\hat{\boldsymbol{y}}$ potom môžeme vyjadriť aj v tvare

$$\hat{\boldsymbol{y}} = \boldsymbol{S}_{\phi,\lambda}\boldsymbol{y}$$

kde

$$\boldsymbol{S}_{\phi,\lambda} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda \boldsymbol{R})^{-1} \boldsymbol{\Phi}'. \tag{4.3}$$

Maticu $\boldsymbol{S}_{\phi,\lambda}$ nazývame aj *vyhladzovací operátor.* Pre danú voľbu vyhladzovacieho parametra λ a bázových funkcií $\boldsymbol{\phi}$.

4.2.2 Výber vyhladzovacieho parametra λ

Účinným nástrojom na voľbu vyhladzovacieho parametra λ je *zovšeobecnená krížová validácia*. Ide o upravenú *krížovú validáciu*, ktorej nevýhodou je, že zvykne *podhladzovať*, teda nedostatočne vyhladzovať dáta. Výraz pre zovšeobecnenú krížovú validáciu definujeme nasledovne:

$$GCV(\lambda) = \frac{n^{-1} SSE}{[n^{-1} \mathrm{tr}(\boldsymbol{I} - \boldsymbol{S}_{\phi,\lambda})]},$$

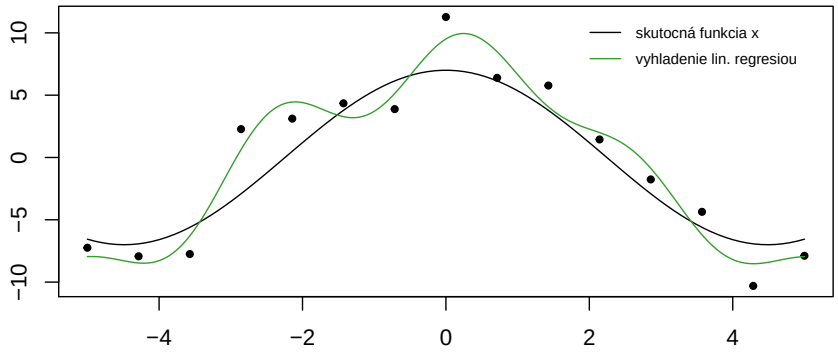
čo vieme prehľadnejšie zapísať aj ako

$$GCV(\lambda) = \left(\frac{n}{n - \mathrm{df}(\lambda)}\right)\left(\frac{SSE}{n - \mathrm{df}(\lambda)}\right)$$

využijúc fakt, že počet stupňov voľnosti pre vyhladzovací parameter získame ako stopu matice vyhladzovacieho operátora $\mathrm{df}(\lambda) = \mathrm{tr}(\boldsymbol{S}_{\phi,\lambda})$, ktorý je definovaný v (4.3). Metóda volí hodnotu λ, ktorá minimalizuje predchádzajúce výrazy.

Vplyv vyhladzovacieho parametra

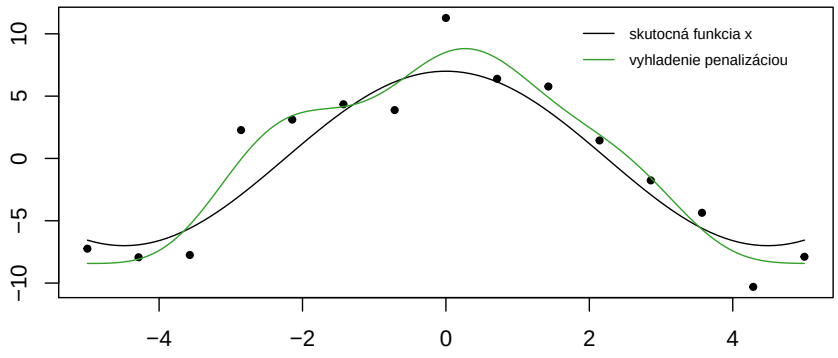
Ukážme si teraz vplyv vyhladzovacieho parametra na simulovaných pozorovaniach. Pre našu prácu má význam predovšetkým Fourierova báza, preto si príklad ilustrujeme práve na nej. Majme funkciu x definovanú na intervale $[-5,5]$, ktorá dáva vznik pozorovaniam $y(t_1),\ldots,y(t_{15})$. Pomocou týchto pozorovaní chceme funkciu x odhadnúť. Zvolíme pevný počet $K = 9$ bázových funkcií. Chceme ilustrovať, ako sa mení odhad funkcie pri meniacej sa hodnote λ, resp. inej metóde vyhladenia. Na Obr. 4.1 je pre vyhladenie použitá metóda lineárnej regresie. Penalizačná funkcia teda nie je prítomná. Vidieť, že odhadnutá krivka je v porovnaní so skutočnou príliš "zvlnená", priveľmi kopíruje pozorovania s chybou. Hovoríme teda, že dáta sú podhladené. Zavedieme preto do minimalizačného vzťahu penalizačnú funkciu a uvidíme, ako sa zmení aproximácia.



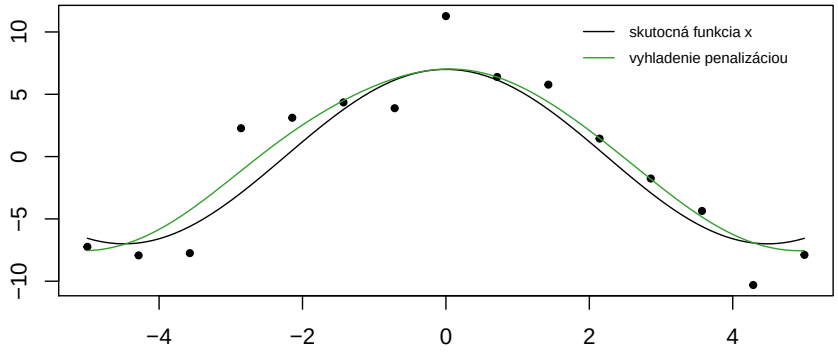
Obr. 4.1: Vyhladenie pomocou lineárnej regresie

Zvolili sme najskôr pomerne nízku hodnotu parametra $\lambda = 0,05$ (Obr. 4.2). Už tu pozorujeme, ako sa odhad v porovnaní s lineárnou regresiou priblížil skutočnej hodnote. Stále je však v odhade nežiadúco veľké množstvo variability.

Pri voľbe $\lambda = 1,5$ vidíme výrazné zlepšenie. Odhadovaná krivka (Obr. 4.3) je veľmi blízko svojej skutočnej hodnote a miera vyhladenosti je tomto prípade uspokojivá.

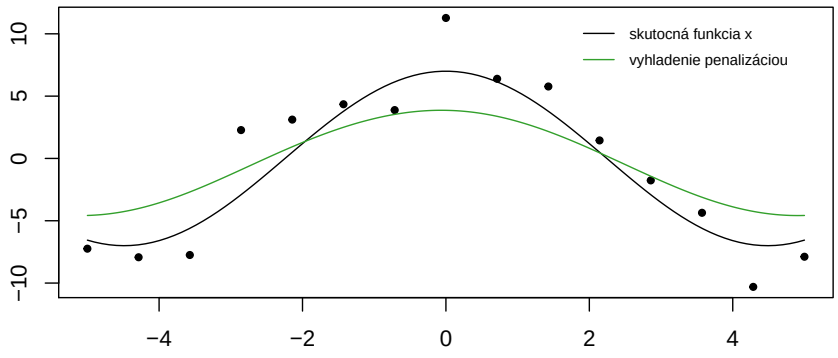


Obr. 4.2: Penalizované vyhladenie pre $\lambda = 0,05$



Obr. 4.3: Penalizované vyhladenie pre $\lambda = 1,5$

Úkažku, čo by sa stalo, keby sme pokračovali vo zvyšovaní hodnoty λ, vidíme na Obr. 4.4. Zvolili sme hodnotu $\lambda = 10$. To, čo vidíme, je v súlade s vyššie popísanou teóriou. Pri zväčšujúcej sa hodnote vyhladzovacieho parametra je stále väčší dôraz kladený na vyhladenie krivky a blížime sa k výsledkom ako v štandardnej lineárnej regresii. V tomto prípade hovoríme o *prehladených* dátach, resp. krivke.



Obr. 4.4: Penalizované vyhladenie pre $\lambda = 10$

Kapitola 5

Funkcionálny lineárny model

V predchádzajúcich kapitolách sme popísali, ako diskrétne dáta prevedieme na funkcionálne. Ďalej sa budeme zaoberať tým, ako pre takéto pozorovania zostrojiť lineárny model, presnejšie funkcionálny lineárny model. Funkcionálny lineárny model predstavuje lineárny vzťah medzi dvoma alebo viacerými premennými, pričom aspoň jedna z týchto premenných má funkcionálny charakter. Funkcionálne modely potom ďalej rozlišujeme podľa toho, či je takouto premennou závislá premenná, nezávislá premenná, prípadne obe.

5.1 Funkcionálna analýza rozptylu

V tejto práci sa snažíme nájsť a popísať vzťahy medzi jednotlivými regionálnymi, resp. globálnymi klimatickými modelmi, v závislosti na teplote. Teplotné pozorovania prevedieme do funkcionálnej podoby a budeme o nich uvažovať ako o závislej premennej. Ku každej z týchto teplotných kriviek potom vieme priradiť jednak globálny a jednak regionálny model, čo vyjadríme pomocou kategoriálnych premenných. Z toho dôvodu sa budeme bližšie zaoberať iba funkcionálnymi lineárnymi modelmi pre prípad funkcionálnej odozvy a skalárnych kovariátov.

Špecifickým prípadom takéhoto modelu je situácia, kedy koviariáty nadobúdajú iba hodnoty 0 a 1. Hovoríme vtedy o *funkcionálnej analýze rozptylu*, skrátene z *angl. functional analysis of variance - fANOVA*.

Zadefinujme si teraz tento model. Majme J skupín nezávislých funkcionálnych náhodných veličín X_{ij}, kde $i = 1,\ldots,n_j$, $j = 1,\ldots,J$, pričom platí $N = n_1 + \cdots + n_J$. Majme potom *jednofaktorový model funkcionálnej analýzy rozptylu* definovaný ako

$$\mathrm{X}_{ij}(t) = \mu(t) + \alpha_j(t) + \varepsilon_{ij}(t), \tag{5.1}$$

kde $\mu(t)$ je spoločná stredná hodnota pre všetky X_{ij}, $\alpha_j(t)$ je j-ty efekt faktora a predstavuje odchýlku jeho j-tej skupiny od celkovej strednej hodnoty a $\varepsilon_{ij}(t)$ je chyba merania. Pre chyby v modeli musí platiť $\mathrm{E}\varepsilon_{ij}(t) = 0$.

Všimnime si, že na rozdiel od klasickej analýzy rozptylu, všetky parametre v tomto modeli majú funkcionálny charakter a závisia na premennej t. Ich odhady získame ako

$$\hat{\mu}(t) = \overline{\mathrm{X}}(t) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathrm{X}_{ij}(t),$$

$$\hat{\alpha}_j(t) = \frac{1}{n_j}\sum_{i=1}^{n_j}(\mathrm{X}_{ij}(t) - \overline{\mathrm{X}}(t)).$$

Pre odhady chýb modelu ε_{ij} dostávame vzťah

$$\hat{\varepsilon}_{ij}(t) = \mathrm{X}_{ij}(t) - \hat{\mu}(t) - \hat{\alpha}_j(t).$$

Predstavujú variabilitu, ktorá nie je vysvetlená príslušnosťou pozorovania k nejakej kategórií j. Potrebujeme ešte zabezpečiť, aby boli jednotlivé efekty jednoznačne identifikovateľné. Je preto potrebné pridať podmienku nulového súčtu parametrov:

$$\sum_{j=1}^{J}\alpha_j(t) = 0.$$

Dvojfaktorová funkcionálna analýza rozptylu

V prípade, kedy máme prítomné dva faktory, hovoríme o *dvojfaktorovom modeli funkcionálnej analýzy rozptylu*. Rozšírime tak vzťah (5.1) o nový faktor s K skupinami. Majme teraz N nezávislých f.n.v X_{ijk} rozdelených do J skupín faktoru A a zároveň K skupín faktoru B, kde $i = 1,\ldots,n_{jk}; j = 1,\ldots,J$ a $k = 1,\ldots,K$ a platí $N = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}; n_j = \sum_{k=1}^{K} n_{jk}$ a $n_k = \sum_{j=1}^{J} n_{jk}$.

Dostávame potom model v tvare

$$X_{ijk}(t) = \mu(t) + \alpha_j(t) + \gamma_k(t) + \varepsilon_{ijk}(t), \tag{5.2}$$

kde $\mu(t)$ je celková stredná hodnota pre všetky X_{ijk}, $\alpha_j(t)$ je j-ty efekt faktoru A a predstavuje odchýlku strednej hodnoty j-tej skupiny od celkovej strednej hodnoty $\mu(t)$, $\gamma_k(t)$ je k-ty efekt faktoru B a predstavuje odchýlku strednej hodnoty k-tej skupiny od celkovej strednej hodnoty $\mu(t)$ a $\varepsilon_{ijk}(t)$ je chyba. Pre chyby opäť platí $\mathrm{E}\varepsilon_{ijk}(t) = 0$.

Pre odhady parametrov potom dostávame

$$\hat{\mu}(t) = \overline{\mathrm{X}}(t) = \frac{1}{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}\mathrm{X}_{ijk}(t)$$

$$\hat{\alpha}_j(t) = \frac{1}{n_j}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\mathrm{X}_{ijk}(t) - \overline{\mathrm{X}}(t))$$

$$\hat{\gamma}_k(t) = \frac{1}{n_k}\sum_{j=1}^{J}\sum_{i=1}^{n_{jk}}(\mathrm{X}_{ijk}(t) - \overline{\mathrm{X}}(t))$$

a pre odhady chýb ε_{ijk} dostávame vzťah

$$\hat{\varepsilon}_{ijk}(t) = \mathrm{X}_{ijk}(t) - \hat{\mu}(t) - \hat{\alpha}_j(t) - \hat{\gamma}_k(t).$$

Opäť platí, že parametre musia byť jednoznačne identifikovateľné, teda musí platiť

$$\sum_{j=1}^{J} \alpha_j(t) = \sum_{k=1}^{K} \gamma_k(t) = 0.$$

V tejto práci využijeme práve model dvojfaktorovej fANOVY. Model pre jeden faktor bol popísaný ako východzí model pre zavedenie a zadefinovanie základných vzťahov. Spomeňme tu, že pre prípad dvoch faktorov existujú podobne ako v klasickej analýze rozptylu modifikácie modelu, napr. pridanie interakčného členu. Ako sa však neskôr ukáže, z povahy dát vyplýva, že sa budeme zaoberať iba modelom bez interakcie.

5.1.1 Lineárny model

Preveďme teraz vzťah (5.2) do podoby všeobecnejšieho zápisu funkcionálneho lineárneho modelu. Majme súbor premenných z_{Aj} a z_{Bk}, kde $j = 1,\dots J$, $k = 1,\dots K$, reprezentujúcich jednotlivé skupiny postupne faktorov A a B. Definujeme ich nasledovne:

$$z_{Aji} = \begin{cases} 1 & \mathrm{X}_{ijk} \text{ patrí do } j\text{-tej skupiny} \\ 0 & \text{inak} \end{cases}$$

$$z_{Bki} = \begin{cases} 1 & \mathrm{X}_{ijk} \text{ patrí do } k\text{-tej skupiny} \\ 0 & \text{inak} \end{cases}$$

Pomocou týchto premenných potom môžeme definovať aj maticu plánu $\boldsymbol{Z}$ pozostávajúcu z núl a jednotiek. Každý riadok matice predstavuje jedno pozorovanie, každý stĺpec prináleží postupne premenným z_{Aj} a z_{Bk}, pre $j = 1,\dots J$ a $k = 1,\dots K$. Ďalej definujme príslušnú množinu parametrov. V tomto prípade sú regresnými parametrami funkcie. Definujeme teda množinu funkcií $\boldsymbol{\beta}$ ako vektor dĺžky $q = J + K + 1$ pomocou vzťahov

$$\begin{aligned} \beta_0 &= \mu, \\ \beta_1 = \alpha_1, \dots, \beta_J &= \alpha_J, \\ \beta_{J+1} = \gamma_1, \dots, \beta_{J+K} &= \gamma_K. \end{aligned}$$

Máme teda funkcionálny lineárny model v tvare

$$y_{ijk}(t) = \beta_0(t) + \sum_{j=1}^{J} z_{Aji}\beta_j(t) + \sum_{k=1}^{K} z_{Bki}\beta_{J+k}(t) + \varepsilon_{ijk}(t), \tag{5.3}$$

kde $y(t)$ je funkcionálna odozva, $\beta_0, \dots, \beta_{K+J}$ sú regresné koeficienty, pre ktoré musí platiť

$$\sum_{j=1}^{J} \beta_j = \sum_{k=1}^{K} \beta_{J+k} = 0$$

a pre chybu $\mathrm{E}\varepsilon_{ijk}(t) = 0$.

Vzťah (5.3) môžeme v maticovom tvare zapísať ako

$$\boldsymbol{y}(t) = \boldsymbol{Z}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t),$$

kde $\boldsymbol{y}(t), \boldsymbol{\varepsilon}(t)$ sú vektory dĺžky N, $\boldsymbol{Z}$ je matica $N \times q$ a $\boldsymbol{\beta}$ vektor dĺžky q. Všimnime si, že matica $\boldsymbol{Z}$ má rovnakú stavbu ako v prípade klasickej analýzy rozptylu. Takto zostavený model sa tak od bežného lineárneho modelu líši iba v tom, že vektory $\boldsymbol{y}$, $\boldsymbol{\beta}$ nie sú vektormi čísel, ale vektormi funkcií.

Uveďme tu ešte, prečo neuvádzame predpoklady pre model tak, ako ich poznáme z analýzy rozptylu. Dôvodom je, že vzhľadom na to, že sa nachádzame vo funkcionálnom priestore, predpoklady normality a zhodnosti rozptylov nevieme definovať a overiť tak, ako sme zvyknutí.

5.1.2 Odhad regresných parametrov

Ďalším krokom je odhad funkcií $\boldsymbol{\beta}$. V prípade štandardnej regresnej analýzy by sme tak urobili pomocou minimalizácie štvorcov rezíduií. Rovnako môžeme postupovať aj v prípade funkcionáleho modelu, musíme však upraviť minimalizačné kritérium berúc ohľad na skutočnosť, že výraz $\boldsymbol{y}(t) - \boldsymbol{Z\beta}(t)$ je teraz funkcia.

Chceme teda minimalizovať *štvorce chýb v lineárnom modeli* v tomto tvare:

$$LMSSE(\boldsymbol{\beta}) = \int [\boldsymbol{y}(t) - \boldsymbol{Z\beta}(t)]'[\boldsymbol{y}(t) - \boldsymbol{Z\beta}(t)]dt.$$

Avšak rovnako ako pri vyhladzovaní funkcií odozvy $\boldsymbol{y}$, aj v tomto prípade môžeme pri odhadovaní $\hat{\boldsymbol{\beta}}$ využiť k vyhladeniu týchto funkcií regularizáciu. Uvedieme si rovno postup pre tento prípad. Odhad pre situáciu bez pridanej penalizačnej funkcie potom získame jednoducho položením $\lambda = 0$.

Predpokladajme, že pozorované funkcie $\boldsymbol{y}$ majú svoje bázové vyjadrenia. To znamená, že ich vieme vyjadriť v tvare

$$\boldsymbol{y}(t) = \boldsymbol{C\phi}(t),$$

kde $\boldsymbol{y}$ je vektor s N pozorovaniami závislej premennej, $\boldsymbol{C}$ je matica $N \times K$ koeficientov rozvoja a $\boldsymbol{\phi}$ je vektor bázových funkcií dĺžky K_y, kde dolný index označuje, že v tomto prípade ide o bázu vzťahujúcu sa na vyjadrenie odozvy. Pomocou bázového rozvoja chceme teraz vyjadriť aj hľadaný odhad regresných parametrov ako $\hat{\boldsymbol{\beta}} = \boldsymbol{B\theta}$, kde $\boldsymbol{B}$ je matica $q \times K_\beta$ a K_β je počet bázových funkcií použitých na vyjadrenie parametrov $\boldsymbol{\beta}$. Uveďme tu, že predpokladáme, že bázy zvolené pre funkcie odozvy a tie pre regresné parametre sa nemusia zhodovať. Naopak, pre všetky regresné koeficienty budeme predpokladať použitie rovnakej bázy. Pre účely tejto práce je takéto určenie vyhovujúce.

Predpokladajúc nejakú voľbu operátora L, definujme ďalej penalizačnú funkciu pre $\boldsymbol{\beta}$ ako

$$PEN_L(\beta) = \int [L\boldsymbol{\beta}(s)]'[L\boldsymbol{\beta}(s)]ds. \tag{5.4}$$

Pre zjednodušenie zápisu už ďalej argumenty (s) a ds vynecháme. Pre *penalizované najmenšie štvorce* potom dostávame nasledujúci výraz:

$$PENSEE(y|\boldsymbol{\beta}) = \int (\boldsymbol{C\phi} - \boldsymbol{ZB\theta})'(\boldsymbol{C\phi} - \boldsymbol{ZB\theta}) + \lambda \int (L\boldsymbol{B\theta})'(L\boldsymbol{B\theta}). \tag{5.5}$$

Pomocou zadefinovania matíc

$$\boldsymbol{J}_{\phi\phi} = \int \boldsymbol{\phi}\boldsymbol{\phi}', \ \boldsymbol{J}_{\theta\theta} = \int \boldsymbol{\theta}\boldsymbol{\theta}', \ \boldsymbol{J}_{\phi\theta} = \int \boldsymbol{\phi}\boldsymbol{\theta}', \ \boldsymbol{R} = \int (L\boldsymbol{\theta})(L\boldsymbol{\theta})'$$

a vychádzajúc zo skutočnosti, že operácie integrovania a súčtu v (5.5) sú navzájom zameniteľné, môžeme (5.5) prepísať na tvar

$$PENSEE(y|\boldsymbol{\beta}) = \mathrm{tr}(\boldsymbol{C}'\boldsymbol{C}\boldsymbol{J}_{\phi\phi}) + \mathrm{tr}(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{B}\boldsymbol{J}_{\theta\theta}\boldsymbol{B}') - 2\mathrm{tr}(\boldsymbol{B}\boldsymbol{J}_{\theta\phi}\boldsymbol{C}'\boldsymbol{Z}) + \lambda\,\mathrm{tr}(\boldsymbol{B}\boldsymbol{R}\boldsymbol{B}'). \quad (5.6)$$

Keďže chceme minimalizovať tento výraz, potrebujeme spočítať jeho deriváciu vzhľadom na maticu $\boldsymbol{B}$ a to položiť rovné nule. Týmto postupom dostaneme, že $\boldsymbol{B}$ spĺňa

$$(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{B}\boldsymbol{J}_{\theta\theta} + \lambda\boldsymbol{B}\boldsymbol{R}) = \boldsymbol{Z}'\boldsymbol{C}\boldsymbol{J}_{\phi\theta}.$$

K ďalšej úprave budeme potrebovať Kroneckerov súčin. Detaily a vlastnosti tejto operácie, ktoré sme využili, je možno nájsť v [5]. Z predchádzajúceho vzťahu teda dostávame

$$[\boldsymbol{J}_{\theta\theta} \otimes (\boldsymbol{Z}'\boldsymbol{Z}) + \boldsymbol{R} \otimes \lambda\boldsymbol{I}]\mathrm{vec}(\boldsymbol{B}) = \mathrm{vec}(\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{J}_{\phi\theta}),$$

kde maticová operácia vec($\boldsymbol{A}$) predstavuje preusporiadanie prvkov matice $\boldsymbol{A}$ do stĺpcového vektora. Odhad $\hat{\boldsymbol{B}}$, ktorý hľadáme, potom dostávame ako riešenie systému lineárnych rovníc

$$\mathrm{vec}(\boldsymbol{B}) = [\boldsymbol{J}_{\theta\theta} \otimes (\boldsymbol{Z}'\boldsymbol{Z}) + \boldsymbol{R} \otimes \lambda\boldsymbol{I}]^{-1}\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{C}\boldsymbol{J}_{\phi\theta}). \quad (5.7)$$

Následným prevedením vec($\hat{\boldsymbol{B}}$) späť na maticu $q \times K_\beta$ a dosadením do vzťahu pre bázové vyjadrenie, ktorý sme si definovali na začiatku, dostávame odhad koeficientov $\boldsymbol{\beta}$ v tvare

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{B}}\boldsymbol{\theta}.$$

5.1.3 Rozptyl odhadu regresných parametrov

Chceli by sme teraz ešte vyjadriť rozptyl pre odhady, ktoré sme si odvodili. Budeme k tomu potrebovať maticu, ktorá zobrazuje pozorovania $\boldsymbol{y}$ na maticu koecifientov $\boldsymbol{c}$. Tú sme si už zadefinovali v časti (4.2). Pripomeňme, že ide o maticu zo vzťahu

$$\boldsymbol{c} = \boldsymbol{S}_{\phi,\lambda}\boldsymbol{y}.$$

Tento vzťah teraz potrebujeme rozšíriť pre N pozorovaní (každé získané z n diskrétnych bodov). Pôvodné dáta teda tvoria maticu $N \times n$, ktorú chceme zobraziť na maticu koeficientov $\boldsymbol{C}$ s rozmermi $N \times K_y$. Maticu tohto zobrazenia budeme značiť Y2C. Máme teda vzťah

$$\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{S}_{\phi,\lambda},$$

ktorý opäť pomocou Kroneckerovho súčinu môžeme previesť na tvar

$$\mathrm{vec}(\boldsymbol{C}) = (\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})\mathrm{vec}(\boldsymbol{Y}),$$

vďaka čomu potom môžeme vyjadriť hľadanú maticu zobrazenia Y2C ako

$$\text{Y2C} = \boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I}. \tag{5.8}$$

Z predošlej kapitoly už vieme, že $\boldsymbol{S}_{\phi,\lambda}$ má tvar

$$\boldsymbol{S}_{\phi,\lambda} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda_y \boldsymbol{R}_y)^{-1}\boldsymbol{\Phi}', \tag{5.9}$$

kde λ_y je vyhladzovací parameter použitý na vyhladenie dát a $\boldsymbol{R}_y$ je príslušná penalizačná matica. Ak ňou nahradíme maticu vo vzťahu (5.8), dostávame plné vyjadrenie pre zobrazenie Y2C.

Druhá matica, ktorú potrebujeme definovať, je tá, ktorá zobrazuje maticu koeficientov $\boldsymbol{C}$ pre funkcie odozvy na maticu koeficientov $\boldsymbol{B}$ pre regresné koeficienty $\boldsymbol{\beta}$. Označme toto zobrazenie C2B.
Miernou úpravou výrazu (5.7) dostávame

$$\text{vec}(\boldsymbol{B}) = [\boldsymbol{J}_{\theta\theta} \otimes \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{R}_\beta \otimes \lambda_\beta \boldsymbol{I}]^{-1}(\boldsymbol{J}'_{\phi\theta} \otimes \boldsymbol{Z}')\text{vec}(\boldsymbol{C}'),$$

kde λ_β a $\boldsymbol{R}_\beta$ sú vyhladzovací parameter a penalizačná matica tentokrát spojené s vyhladzovaním parametrov $\boldsymbol{\beta}$. Odtiaľ už potom dostávame samotné zobrazenie C2B v tvare

$$\text{C2B} = \boldsymbol{S}_\beta = [\boldsymbol{J}_{\theta\theta} \otimes \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{R}_\beta \otimes \lambda_\beta \boldsymbol{I}]^{-1}(\boldsymbol{J}'_{\phi\theta}\theta\boldsymbol{Z}').$$

Teraz môžeme definovať zobrazenie Y2B spojením zobrazení Y2C a C2B:

$$\text{Y2B} = \boldsymbol{S}_\beta(\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})$$

a príslušný vzťah

$$\text{vec}(\boldsymbol{B}) = \boldsymbol{S}_\beta(\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})\text{vec}(\boldsymbol{Y}').$$

Rozptyl pôvodných dát je daný

$$\text{var}[\text{vec}(\boldsymbol{Y}')] = \boldsymbol{\Sigma}_e \otimes \boldsymbol{I},$$

kde $\boldsymbol{\Sigma}_e$ je kovariačná matica rezíduí modelu a $\boldsymbol{I}$ je matica $N \times N$.

Spojením predchádzajúcich vzťahov teda konečne dostávame pre bázové koeficienty regresných parametrov $\boldsymbol{\beta}$ rozptyl v tvare

$$\text{var}[\text{vec}(\hat{\boldsymbol{B}})] = \boldsymbol{S}_\beta(\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})(\boldsymbol{\Sigma}_e \otimes \boldsymbol{I})(\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})\boldsymbol{S}'_\beta$$

a pre rozptyl samotných parametrov

$$\text{var}[\text{vec}(\hat{\boldsymbol{\beta}})] = (\Theta \otimes \boldsymbol{I})\boldsymbol{S}_\beta(\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})(\boldsymbol{\Sigma}_e \otimes \boldsymbol{I})(\boldsymbol{S}_{\phi,\lambda} \otimes \boldsymbol{I})\boldsymbol{S}'_\beta(\Theta \otimes \boldsymbol{I})',$$

kde $\boldsymbol{\Theta}$ je $N \times K_\beta$ matica hodnôt bázových funkcií pre $\boldsymbol{\beta}$ vyhodnotená v pôvodných bodoch pre pozorované dáta odozvy.
Za predpokladu, že matice $\boldsymbol{Z}'\boldsymbol{Z}$ a $\boldsymbol{J}_{\theta\theta}$ sú invertibilné, môžeme výraz zjednodušiť na

$$\text{var}[\text{vec}(\hat{\boldsymbol{\beta}})] = [\boldsymbol{J}^{-1}_{\theta\theta}\boldsymbol{J}'_{\phi\theta}\boldsymbol{S}_{\theta,\lambda_y}\boldsymbol{\Sigma}_e\boldsymbol{S}_{\theta,\lambda_y}\boldsymbol{J}_{\phi\theta}\boldsymbol{J}^{-1}_{\theta\theta}] \otimes (\boldsymbol{Z}'\boldsymbol{Z})^{-1}.$$

5.2 Posúdenie kvality odhadu

Rovnako ako v bežnom lineárnom modeli, aj v tomto prípade sa pri posudzovaní kvality odhadu pozrieme na hodnoty koeficientu R^2 a F-test. Rozdiel spočíva v tom, že sledujeme aj vývoj týchto veličín v závislosti na čase t.

Permutovaný F-test

Chceme zostrojiť test, ktorý by nám umožnil overiť, či má model, tak ako sme ho zostavili, významný vplyv na získané výsledky.

Typickým nástrojom na posúdenie celkovej kvality modelu je pri bežnej analýze rozptylu F-test. Ten je však založený na znalosti rozdelenia, z ktorého pochádzajú hodnoty štatistiky. V prípade funkcionálnej analýzy je práve tento bod problematický. Využíva sa preto modifikovaná verzia testu, tzv. *permutovaný F-test*. Rozdelenie pre prípad platnosti nulovej hypotézy sa v tomto prípade konštruuje priamo z pozorovaní. Zakladá sa na myšlienke, že v prípade, že medzi výstupnou premennou a kovariátmi nie je nami skúmaný vzťah, náhodné preusporiadanie párov odozva-kovariáty by nemalo mať na výsledok vplyv. Chceme teda testovať nulovú hypotézu, že príslušná kombinácia kovariátov nemá vplyv na závislú premennú.

Najskôr definujeme funkcionálnu verziu F-štatistiky

$$F(t) = \frac{\mathrm{var}[\hat{\boldsymbol{y}}(t)]}{\frac{1}{n}\sum(y_i(t) - \hat{y}_i(t))^2},$$

ktorá vyjadruje pomer rozptylu predikovaných kriviek a sumy štvorcov rezíduí.

Spočítame túto štatistiku pre náš model. Jej funkcionálnu podobu potom zredukujeme do podoby samostatnej hodnoty určením maxima tejto funkcie

$$F^* = \max_{t \in [t_{min}, t_{max}]} F(t),$$

kde $[t_{min}, t_{max}]$ je interval, na ktorom definujeme pozorovania $\boldsymbol{y}(t)$.

Hodnotu F^* potom vezmeme za pozorovanú testovú štatistiku. Ďalším krokom je zostrojenie empirického rozdelenia pomocou permutácií. Sformalizujme teraz tento postup.

Nasledujúce kroky opakujeme D-krát:

1. Permutujeme množinu indexov pozorovaní $\{1, \ldots, n\}$ a dostávame nové indexy $\{i_1, \ldots, i_n\}$.

2. Preusporiadame hodnoty závislej premennej podľa nového indexovania, pričom matica plánu ostáva nezmenená.

3. Spočítame F-štatistiku pre takéto usporiadanie a určíme príslušnú hodnotu F_d^*.

4. Porovnáme F_d^* s pôvodnou hodnotou F^*. Definujeme premennú

$$I_d = \begin{cases} 1 & \text{ak } F_d^* > F^*, \\ 0 & \text{ak } F_d^* \leq F^*. \end{cases} \tag{5.10}$$

P-hodnotu testu potom určíme ako pomer všetkých spočítaných *F*-štatistík a tých, ktoré sú väčšie ako *F*-štatistika pre pôvodné usporiadanie:

$$p = \frac{1}{D}\sum_{d=1}^{D} I_d. \tag{5.11}$$

Nevýhodou tohto testu je, že vieme posúdiť iba vplyv celého lineárneho prediktoru, nie jednotlivých kovariátov samostatne. Naopak, výhoda spočíva v tom, že nie sme viazaní predpokladmi pre rozdelenie. Pre lepšie porozumenie konštrukcii a priebehu testu si ho detailnejšie ilustrujeme na reálnych dátach v praktickej časti tejto práce.

R^2 - štvorcový koeficient násobnej korelácie

Ďalším nástrojom na posúdenie kvality modelu je *štvorcový koeficient násobnej korelácie* R^2. Potrebujeme najskôr definovať dva výrazy, a to sumu štvorcov rezíduí:

$$\begin{aligned} SSE(t) &= \sum_{i}^{N} [y_i(t) - \hat{y}_i(t)]^2 \\ &= [\boldsymbol{y}(t) - \boldsymbol{Z}\hat{\boldsymbol{\beta}}(t)]^2, \end{aligned}$$

a sumu štvorcov chýb v prípade modelu pozostávajúceho iba z priemeru $\hat{\mu}$:

$$\begin{aligned} SSY(t) &= \sum_{i}^{N} [y_i(t) - \hat{\mu}(t)]^2 \\ &= \sum_{i}^{N} [y_i(t) - \overline{y}(t)]^2. \end{aligned}$$

Teraz môžeme definovať koeficient R^2 ako

$$R^2(t) = 1 - \frac{SSE(t)}{SSY(t)}. \tag{5.12}$$

Tak, ako sme zvyknutí, R^2 udáva mieru toho, ako dobre vieme danú závislú premennú v modeli predikovať pomocou lineárnej kombinácie zvyšných premenných.

Kapitola 6

Aplikácia na reálne dáta

V tejto kapitole aplikujeme doteraz popísanú teóriu funkcionálnych dát na reálne dáta. Skúmame vzťahy a závislosti medzi jednotlivými klimatickými modelmi. Spracovanie je realizované pomocou programu R a predovšetkým balíka `fda`, špeciálne určeného pre prácu s funkcionálnymi dátami.

6.1 Popis dátového súboru

Dátový súbor, ktorý v tejto práci spracovávame, obsahuje teplotné merania zhotovené počas 30 ročného časového intervalu. Máme tak k dispozícií priemerné mesačné teploty pre roky 1971-2000. Merania boli realizované vždy pomocou dvojice klimatických modelov - globálny model a regionálny model. V tomto prípade pracujeme s 8 rôznymi globálnymi modelmi (alebo len "GCM"): CNRM, CanESM, EC-EARTH, HadGEM2, IPSL, MICRO, MPI, NorESM; a 5 regionálnymi modelmi (alebo len "RCM"): RCA, RACMO, WRF, CCLM, RegCM.

Pre každú takúto dvojicu máme tabuľku 360 hodnôt (12 mesiacov $\times$ 30 rokov). Avšak dátový súbor považujeme za nekompletný, pretože neobsahuje pozorovania pre všetky kombinácie regionálneho a globálneho modelu. Z plného počtu možných 40 dvojíc tak máme k dispozícií záznamy iba pre 12. Náhľad na nekompletný súbor pozorovaní vidíme v tabuľke 6.1, kde sú uvedené dostupné hodnoty pre január 1971. Všetky hodnoty sú uvádzané v Kelvinoch.[1]

Vidíme odtiaľ, že zastúpenie pozorovaní je pomerne nerovnomerné, obzvlášť v prípade regionálnych modelov. Pre RCA model máme k dispozícií až 7 pozorovaní, zatiaľ čo pre zvyšné modely jedno, prípadne dve pozorovania.

Uveďme si na tomto mieste, ako budeme ďalej uvažovať nad dátami v rámci našej analýzy. Cieľom analýzy je popísať vzťahy medzi jednotlivými modelmi, k čomu využijeme namerané hodnoty teplôt. Keďže využívame metódy analýzy funkcionálnych dát, nebudeme ďalej uvažovať o teplotných meraniach ako o individuálnych pozorovaniach, ale za pozorovania vezmeme celé krivky. V prípade, že by sme uvažovali teplotnú krivku pre celých 30 rokov ako jedno pozorovanie pre nejakú kombináciu RCM-GCM, mali by sme 12 pozorovaní. Takýto počet pozorovaní je sám o sebe nízky. Keď naviac vezmeme

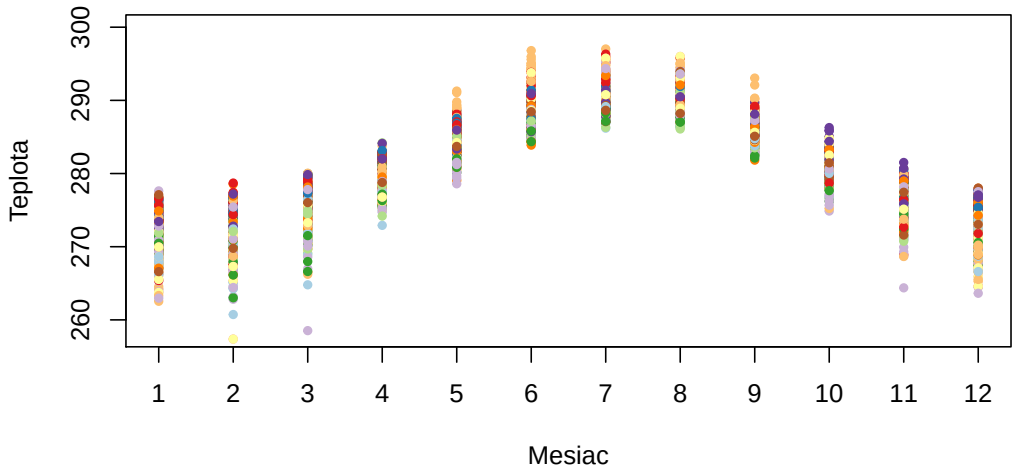
[1] 1 K = -272,15 °C

Tabuľka 6.1: Priemerné hodnoty pre január 1971

	CNRM	CanESM	EC-Earth	HDGEM2	IPSL	MICRO	MPI	NorESM
RACMO			268,912					
RCA	270,709	272,331		269,176	269,380	267,507	274,552	273,754
WRF			269,874		268,980			
CCLM				267,507				
RegCM				264,541				

do úvahy počet modelov, s ktorými pracujeme, zdá sa byť nedostatočný pre popis vzájomných vzťahov. Počas analýzy sa ukázalo ako najužitočnejšie uvažovať nad pozorovaniami zvlášť pre jednotlivé roky. Dostávame tak 30 funkcionálnych pozorovaní pre každú dvojicu RCM-GCM, čím máme k dispozícii celkovo 360 pozorovaných kriviek. Pripomeňme, že zatiaľ stále len vo forme diskrétnych bodov.

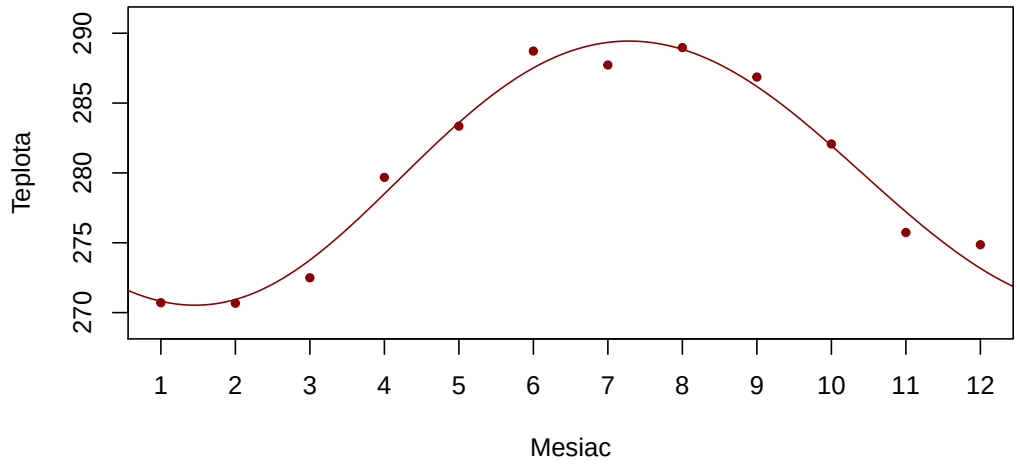
Na Obr. 6.1 sú vykreslené všetky dostupné hodnoty. Ako sa dalo očakávať, keďže hovoríme o teplotných meraniach, vidíme istý periodický charakter pozorovaní. Z tohto prvotného náhľadu môžeme ďalej pozorovať mierne zväčšený rozptyl pre zimnú sezónu oproti zvyšku roka. Vidíme tu tiež niekoľko odľahlejších hodnôt.



Obr. 6.1: Priemerné mesačné hodnoty pre všetky modely

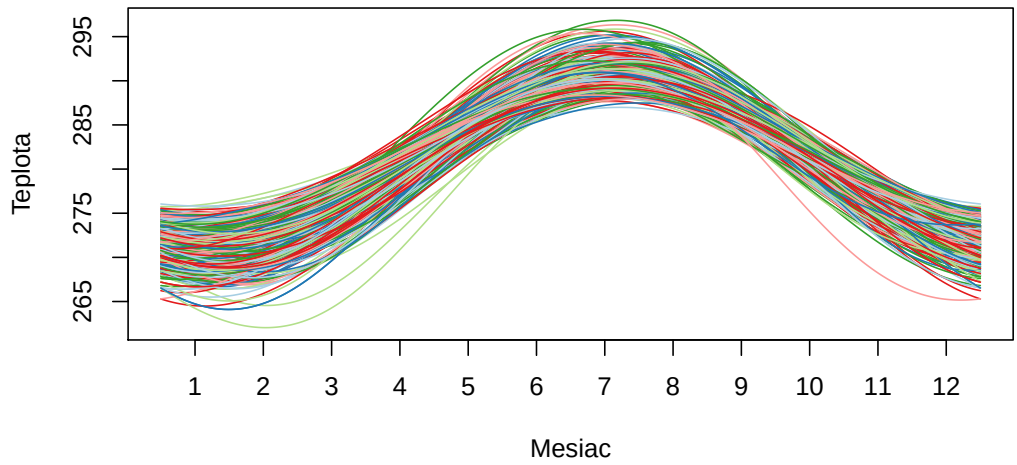
6.2 Úprava a vyhladenie dát

Prvým krokom analýzy je reprezentácia a vyhladenie dát. Ani pri vyššom počte zvolených bázových funkcií neviedla táto voľba k ujme na vyhladenosti kriviek. Vďaka tomu a pre uchovanie čo najväčšieho množstva informácií sme zvolili počet bázových funkcií 13. Ako najlepšia voľba spôsobu vyhladenia sa ukázala metóda regularizácie pomocou Fourierovej bázy za použitia harmonického akcelerátora ako penalizačnej funkcie. Z hľadiska bázy aj



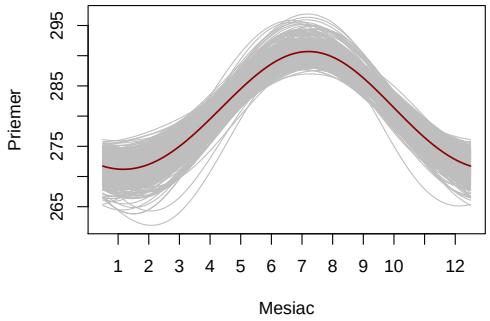
Obr. 6.2: Pôvodné a vyhladené hodnoty pre rok 1971 - model CNRM-RCA

penalizačnej funkcie ide zároveň o typické voľby pre prípad periodických dát. Optimálnu hodnotu parametra λ sme určili pomocou zovšeobecnenej krížovej validácie, táto hodnota vyšla $\lambda = 1,02$. Na Obr. 6.2 je ukážka pôvodných meraní spolu s vyhladenou krivku pre dvojicu modelov CNRM-RCA a pre rok 1971.
Vyššie popísaná voľba je aplikovaná na všetky pozorovania. Ich prevedením do funkcionálnej podoby dostávame 360 vyhladených kriviek znázornených na Obr. 6.3.

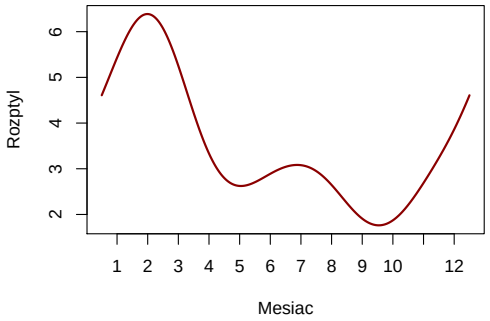


Obr. 6.3: Všetky vyhladené krivky

Bližšie sa teraz pozrieme na rozptyl dát. Už z Obr. 6.1 a 6.3 bolo možné vidieť, že väčšia variabilita medzi pozorovaniami je prítomná v zimných a letných mesiacoch. Rovnaký



Obr. 6.4: Priemer pre všetky krivky

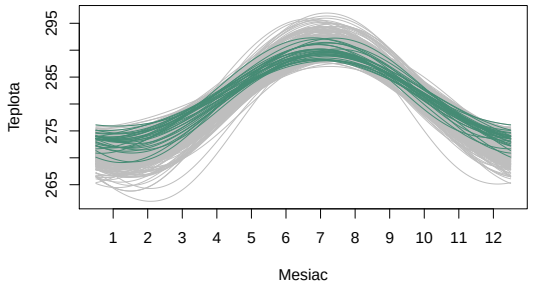


Obr. 6.5: Rozptyl pre všetky krivky

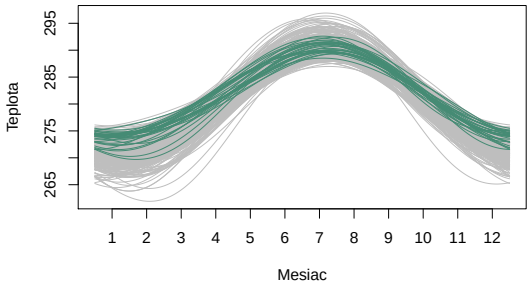
záver môžeme vyvodiť aj z Obr. 6.5. Rozptyl je v tomto období dvoj- až trojnásobný oproti zvyšku pozorovaného intervalu. Na Obr. 6.4 uvádzame aj priemernú teplotnú krivku pre pozorovania. Priemerné teploty sa pohybujú približne od 272 K po 290 K. Pre lepšiu predstavu, v európskych jednotkách ide o interval zhruba od 0 °C po 18 °C.

6.2.1 Skupiny klimatických modelov

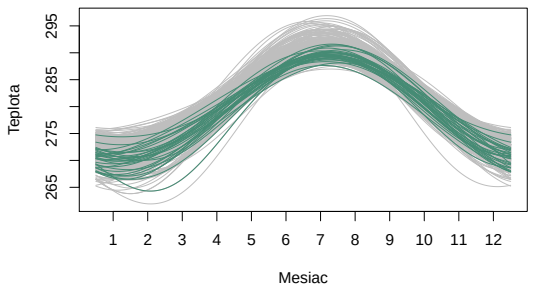
Prvým krokom pri skúmaní vzťahov medzi klimatickými modelmi bude graficky posúdiť vzájomnú blízkosť a podobnosť jednotlivých pozorovaných kriviek v rámci skupín naprieč modelmi. Pripomeňme, že skupinu predstavuje jedna z 12 kombinácií regionálneho modelu a globálneho modelu. Vykreslili sme si teda spoločne všetky krivky, pričom zvýraznené sú vždy tie, ktoré prislúchajú danej dvojici modelov.
Správanie, ktoré pozorujeme, v tomto prípade napovedá o ďalšom smerovaní v analýze. Celkovo môžeme konštatovať, že krivky v rámci jednej skupiny sú si vzájomne pomerne blízko, a zároveň vidíme, že naprieč rôznymi skupinami sa líšia úrovňou aj tvarom krivky. V rámci všetkých skupín pozorujeme to, čo aj pri plnom súbore kriviek a to že väčší rozptyl je predovšetkým v zimných mesiacoch a čiastočne v letných. Ďalšie výraznejšie vzorce správania nesledujeme. Ako problematické možeme ale napríklad vidieť zimné mesiace pre dvojicu WRF-IPSL (Obr. 6.17).



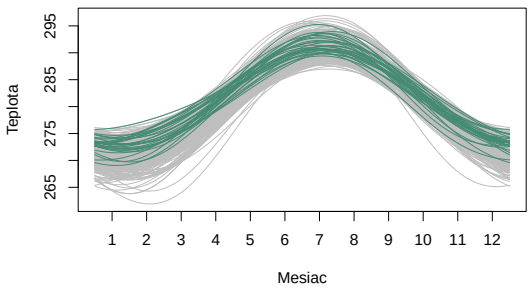
Obr. 6.6: Modely RCA-NorESM



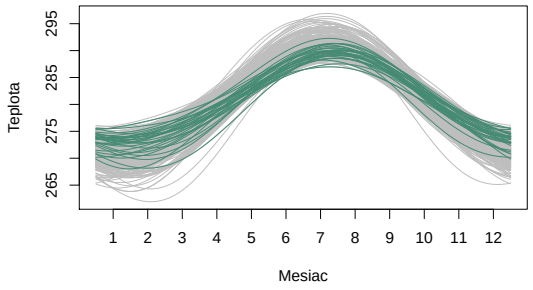
Obr. 6.7: Modely RCA-CNRM



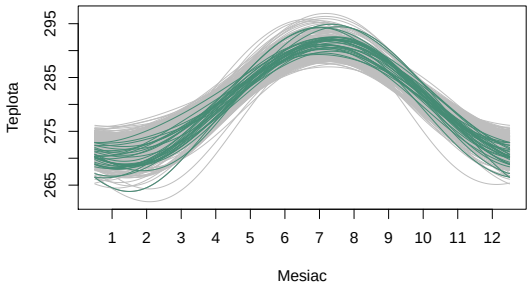
Obr. 6.8: Modely RCA-CanESM



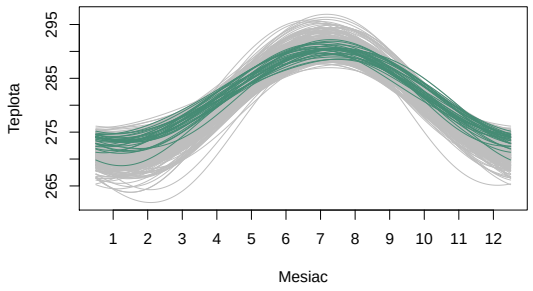
Obr. 6.9: Modely RCA-HadGEM2



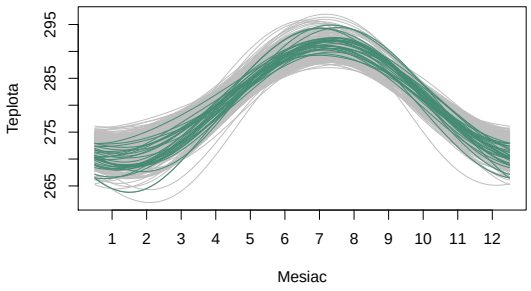
Obr. 6.10: Modely RCA-IPSL



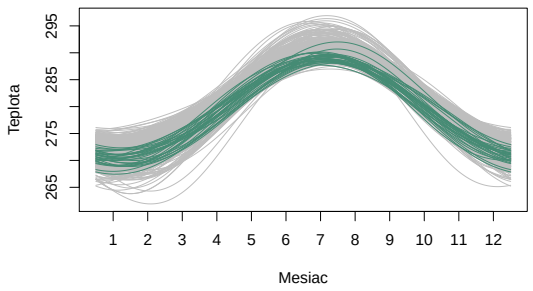
Obr. 6.11: Modely RCA-MPI



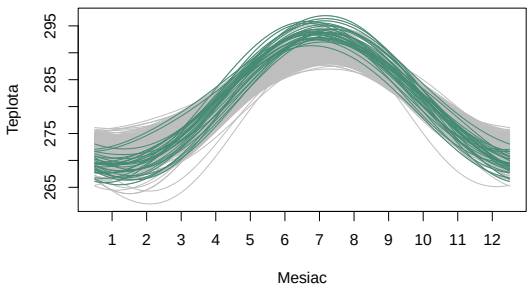
Obr. 6.12: Modely RCA-MICRO



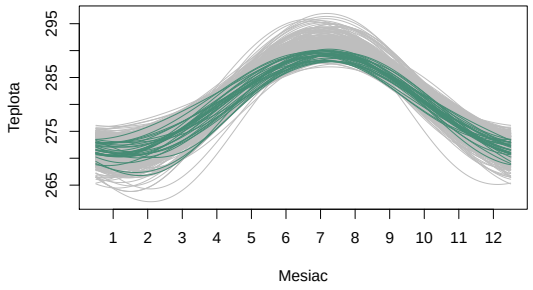
Obr. 6.13: Modely RCA-HadGEM2



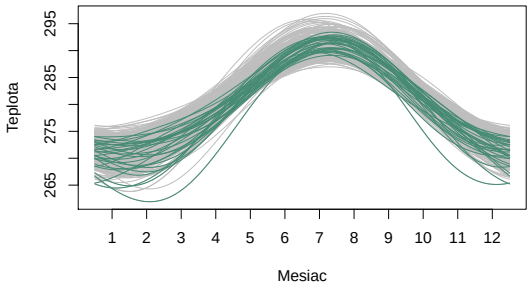
Obr. 6.14: Modely RCA-EC-Earth



Obr. 6.15: Modely RegCM-HadGEM2



Obr. 6.16: Modely WRF-EC-Earth



Obr. 6.17: Modely WRF-IPSL

Úvodný náhľad na pozorovania teda napovedá, že klimatické modely, pomocou ktorých boli uskutočňované merania teploty, majú vplyv na hodnoty daných meraní. Úlohou je teraz nájsť štatistický model, ktorý čo najlepšie popíše tieto vplyvy. O to sa pokúsime v nasledujúcej časti.

6.3 Model

Chceme nájsť model, ktorý popíše vzťahy medzi nameranou teplotou a klimatickými modelmi. Ukázalo sa vhodné použiť model funkcionálnej analýzy rozptylu. Keďže ide o dva rozličné súbory klimatických modelov, presnejšie pôjde o model *dvojfaktorovej funkcionálnej analýzy rozptylu*. Vzhľadom na to, že nemáme k dispozícii pozorovania pre všetky možné kombinácie týchto faktorov, zvolili sme model bez interakcie. Máme teda model v tvare, kde pozorované teplotné krivky sú funkcionálnou premennou a sú prerozdelené do 5 skupín faktoru *rcm* a 8 skupín faktoru *gcm*.
Daný model fANOVA potom definujeme ako

$$y_{ijk}(t) = \mu(t) + \alpha_{rcm,j}(t) + \gamma_{gcm,k}(t) + \varepsilon_{ijk}(t), \tag{6.1}$$

kde $\alpha_{rcm,j}(t)$ je j-ty efekt faktoru *rcm* a $\gamma_{gcm,k}(t)$ je k-ty efekt faktoru *gcm*

pre $j = 1,\ldots,5; k = 1,\ldots,8$. Podmienka identifikovateľnosti má potom tvar

$$\sum_{j=1}^{5} \alpha_{rcm,j}(t) = \sum_{k=1}^{8} \gamma_{gcm,k}(t) = 0.$$

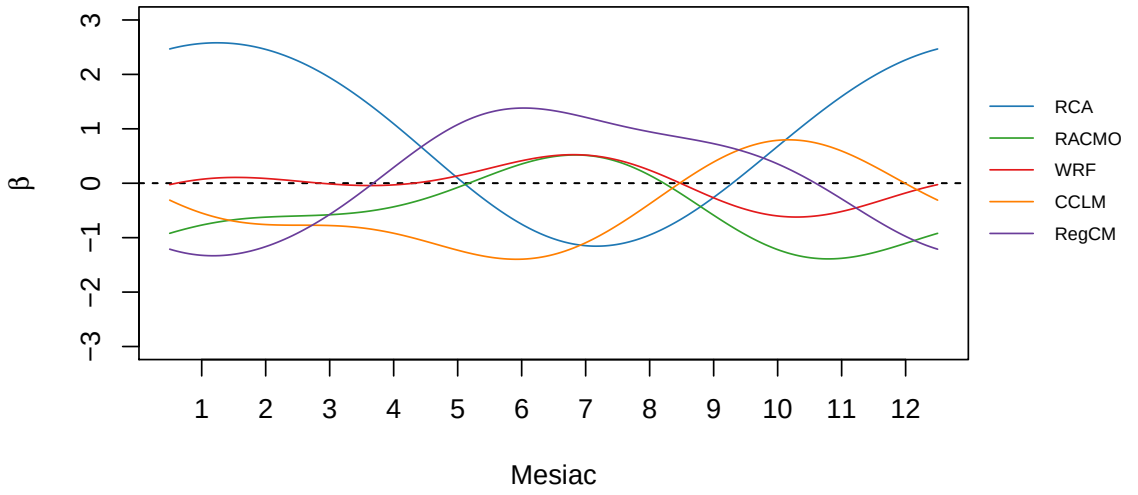
Definujme ďalej pre regionálny a globálny model indikátorové premenné $z_{rcm,j}$ a $z_{gcm,k}$, ktorých pozorovania majú hodnoty 0 alebo 1.
Model (6.1) má potom ekvivalentnú formuláciu

$$y_{ijk}(t) = \beta_0(t) + \sum_{j=1}^{5} z_{rcm,ji}\beta_j(t) + \sum_{k=1}^{8} z_{gcm,ki}\beta_{k+5}(t) + \varepsilon_{ijk}(t) \tag{6.2}$$

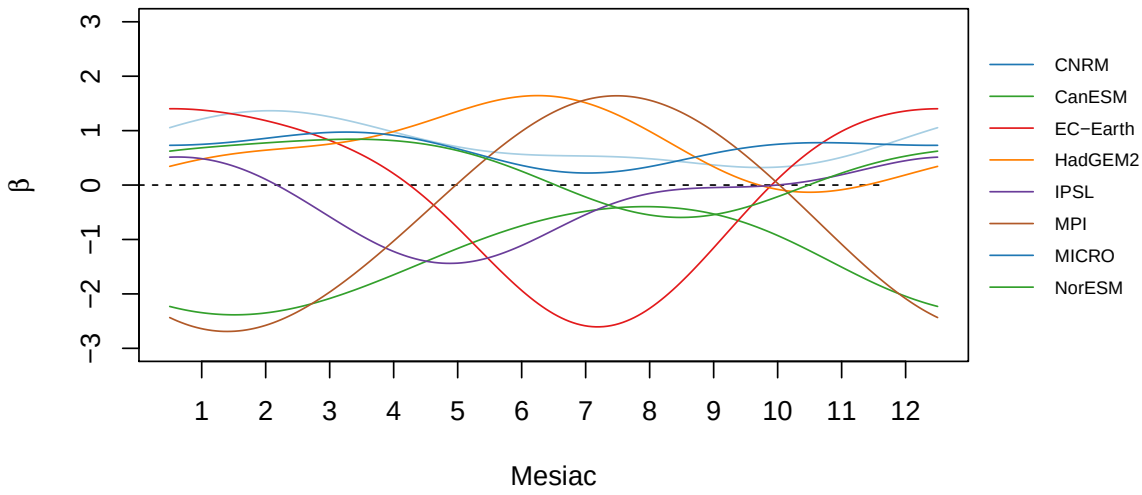
a musí opäť platiť

$$\sum_{j=1}^{5} \beta_j(t) = \sum_{k=1}^{8} \beta_{k+5}(t) = 0.$$

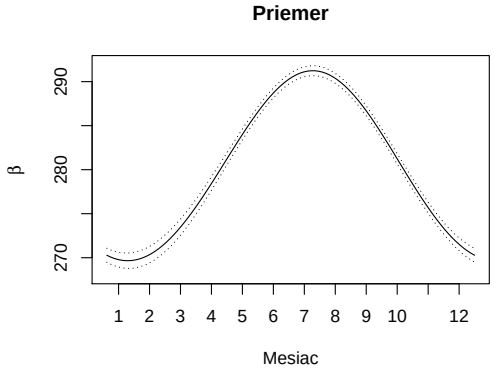
Ďalším krokom je konštrukcia matice plánu $\boldsymbol{Z}$. Konštruujeme ju rovnako ako v prípade klasickej analýzy rozptylu. Majme teda maticu s rozmermi 360×14, kde každý riadok reprezentuje jedno pozorovanie. Stĺpce potom predstavujú postupne celkovú strednú hodnotu a jednotlivé efekty faktorov. V prvom stĺpci pre strednú hodnotu má matica samé jednotky, v ďalších stĺpcoch sú jednotky tam, kde príslušné pozorovanie patrí ku konkrétnej skupine daného faktora, zvyšok sú nuly. Potrebujeme ešte ošetriť otázku indentifikovateľnosti efektov. Existuje na to viacero postupov, my použijeme ten, ktorý využíva pridanie "nulových pozorovaní", čo zabezpečíme voľbou nulových koeficientov $\boldsymbol{c}$ pre tieto pozorovania. Pridáme teda dve nulové pozorovania a zároveň do matice plánu pridáme korešpondujúce riadky. Pre každý faktor vytvoríme riadok, ktorý má jednotky na miestach (príslušných stĺpcoch matice) všetkých svojich skupín, všade inde nuly.
Ďalším krokom je samotný odhad parametrov $\boldsymbol{\beta}$. Ako bolo uvedené v časti (5.1.1), ide o funkcionálne objekty, volíme preto opäť bázové funkcie, pomocou ktorých ich vyjadríme. V (5.1.2) sme predpokladali, že táto báza sa nemusí zhodovať s bázou, ktorú sme použili na vyhladenie pozorovaní. Doplníme, že nemusí ísť ani o rovnaký typ bázy. Opäť sme však zvolili Fourierovu bázu s 13 bázovými funkciami. Voľba regularizačnej metódy pre vyhladzovanie parametrov v tomto prípade nepriniesla žiadne zlepšenie odhadov modelu, vystačili sme si preto s metódou lineárnej regresie.
Na obrázku 6.18 je už spoločne päť odhadnutých regresných koeficientov pre regionálne modely, na 6.19 sú koeficienty pre globálne modely. Vidíme, že najvýraznejší vplyv na teplotu má spomedzi regionálnych modelov RCA model. V zimných mesiacoch vidieť až takmer trojstupňovú odchýlku od priemeru. Naopak odhad pre WRF model sa od priemeru neodlišuje o viac ako pol stupňa.
V prípade globálnych modelov na Obr. 6.19 pozorujeme viacero modelov s výraznejším vplyvom, u EC-Earth, MPI a CanESM modelov vidíme v určitých mesiacoch opäť až trojstupňový odklon od priemeru.
Séria grafov 6.20-6.33 zobrazuje parametre samostatne, spolu s ich intervalmi spoľahlivosti. Uveďme, že je treba byť opatrný pri prípadnej interpretácií týchto intervalov spoľahlivosti. Ide iba o bodové odhady intervalu v pevnom čase. Nejde teda o globálne intervaly spoľahlivosti pre celé krivky parametrov.



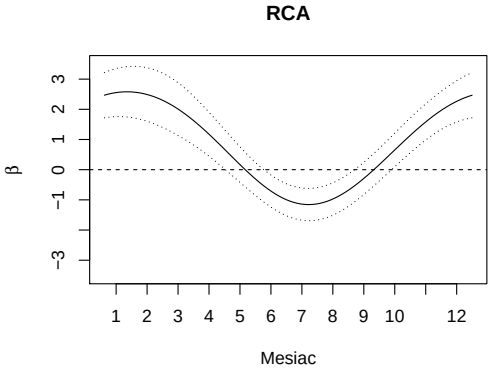
Obr. 6.18: Odhady pre RCM modely



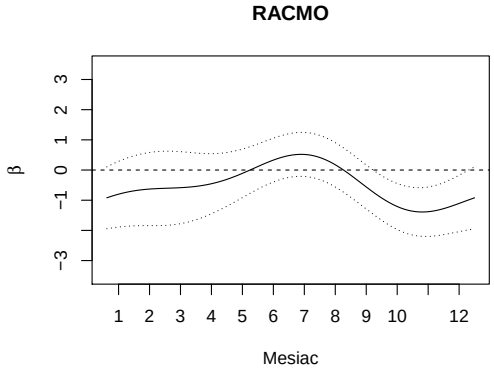
Obr. 6.19: Odhady pre GCM modely



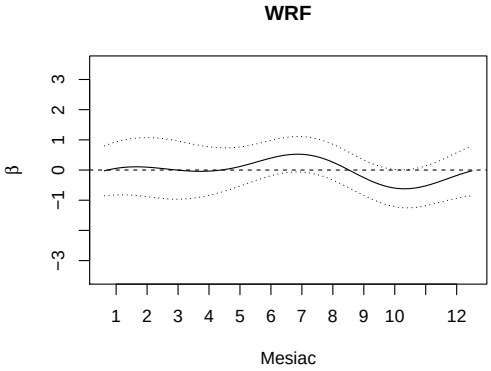
Obr. 6.20: Odhad spoločného priemeru



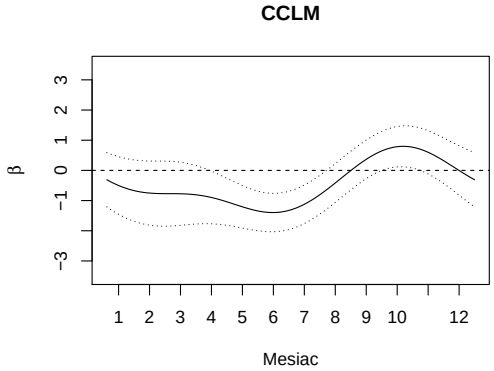
Obr. 6.21: Odhad β pre RCA



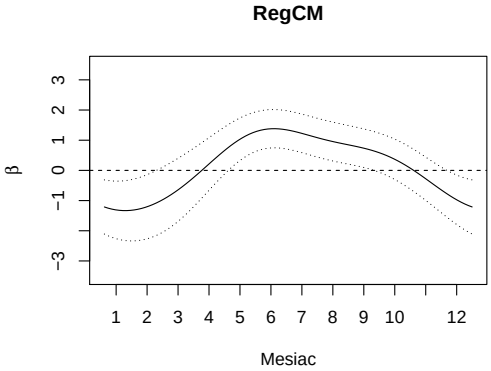
Obr. 6.22: Odhad β pre RACMO



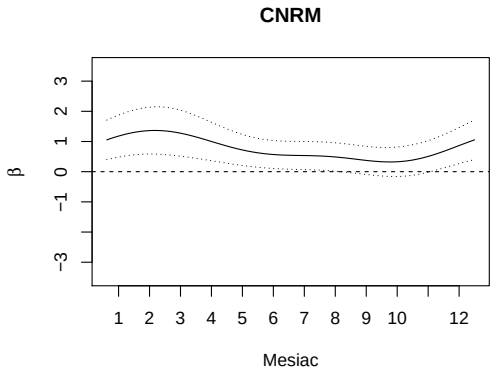
Obr. 6.23: Odhad β pre WRF



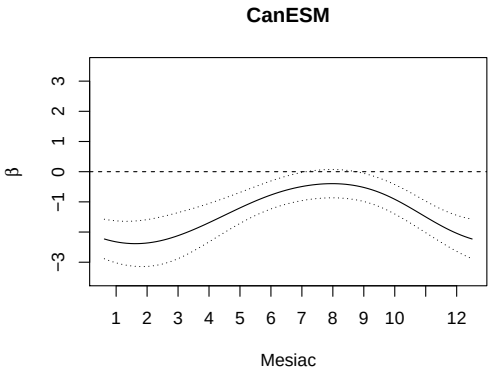
Obr. 6.24: Odhad β pre CCLM



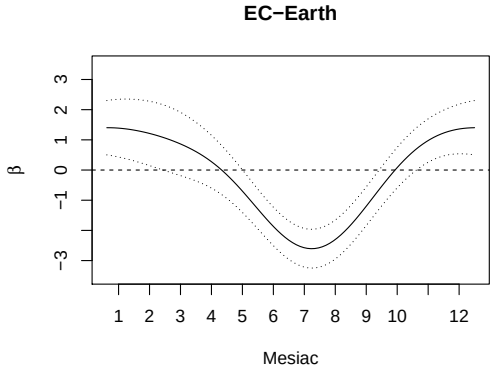
Obr. 6.25: Odhad β pre RegCM



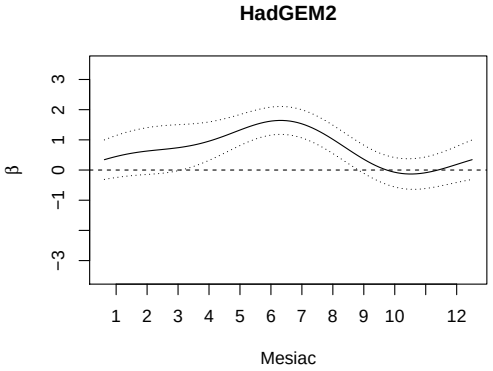
Obr. 6.26: Odhad β pre CNRM



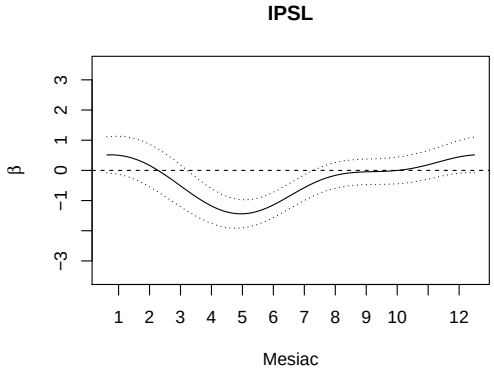
Obr. 6.27: Odhad β pre CanESM



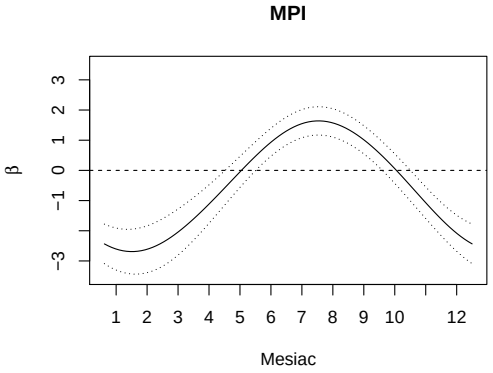
Obr. 6.28: Odhad β pre EC-Earth



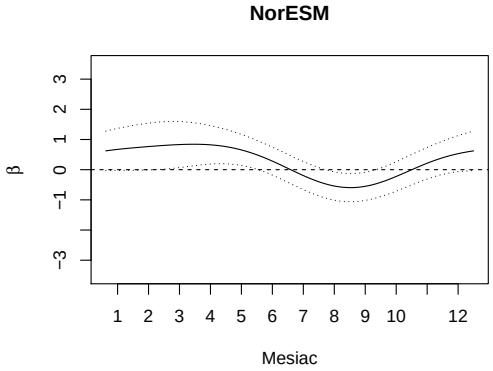
Obr. 6.29: Odhad β pre HadGEM2



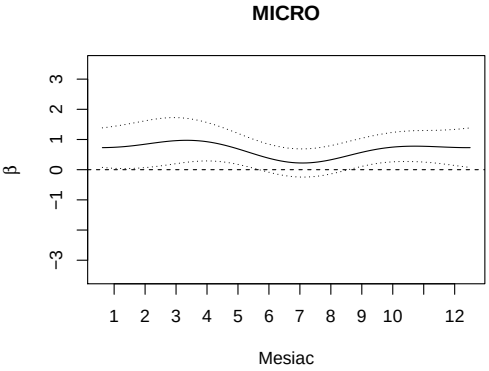
Obr. 6.30: Odhad β pre IPSL



Obr. 6.31: Odhad β pre MPI



Obr. 6.32: Odhad β pre NorESM



Obr. 6.33: Odhad β pre MICRO

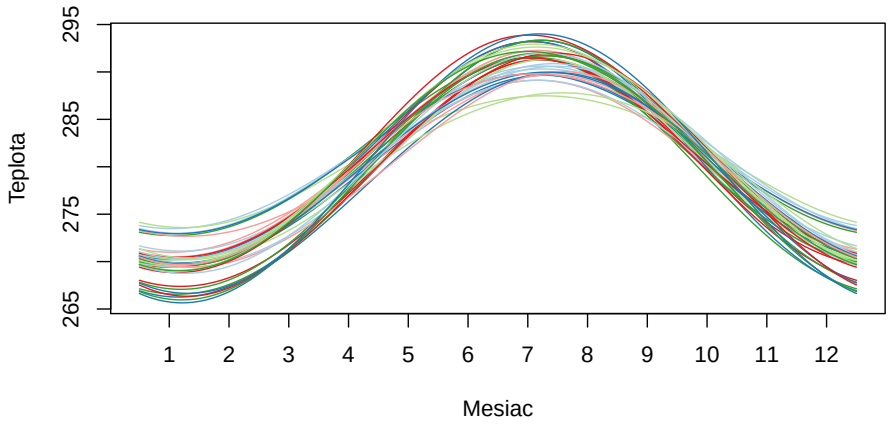
6.3.1 Posúdenie modelu

Ukážme si teraz, ako vyzerajú predikované hodnoty z modelu. Krivky sa spočítajú pomocou vzťahu:

$$\hat{y}_{ijk}(t) = \hat{\beta}_0(t) + \hat{\beta}_j(t) + \hat{\beta}_{k+5}(t), \tag{6.3}$$

kde $\hat{\beta}_0$ je odhadnutá spoločná stredná hodnota, $\hat{\beta}_j$ sú odhady efektov faktoru *rcm* a $\hat{\beta}_{k+5}$ odhady efektov faktora *gcm* pre $j = 1,\ldots,5; k = 1,\ldots,8$. Pre každú kombináciu klimatických modelov dostávame jednu odhadnutú krivku, máme tak spoločne 40 kriviek. Tie sú zobrazené na grafe 6.34.
Vychádzajúc ďalej zo vzťahov (6.2) a (6.3), pre rezíduá modelu máme vyjadrenie:

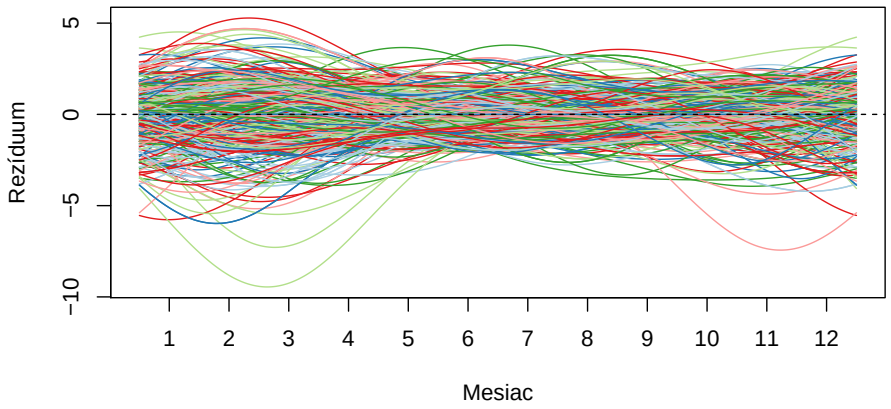
$$\hat{\varepsilon}_{ijk}(t) = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \hat{\beta}_0(t) - \hat{\beta}_j(t) - \hat{\beta}_{k+5}(t).$$



Obr. 6.34: Všetky predikované krivky

Graf 6.35 zobrazuje rezíduá modelu. Tak ako bolo na začiatku vidieť aj u pôvodných bodov a neskôr vyhladených kriviek, vyskytuje sa tu niekoľko pozorovaní, ktoré sa javia

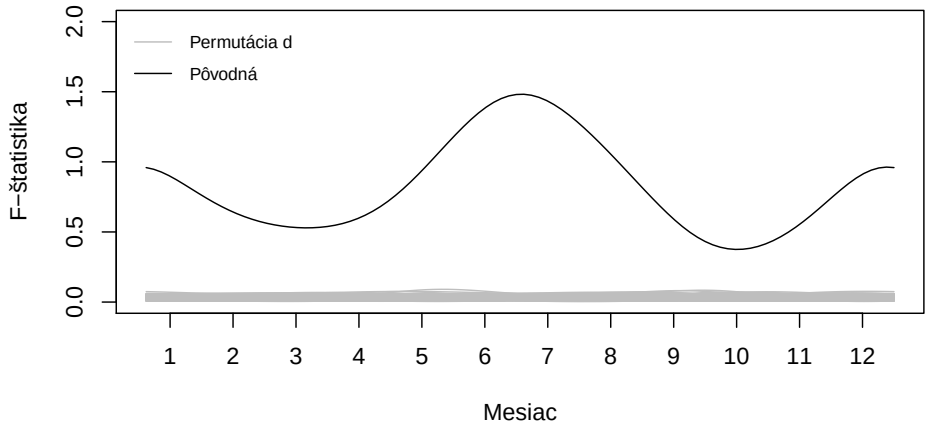
ako odľahlé. Táto skutočnosť sa prejavila aj v rezíduách. A aj tu opäť vidíme väčší rozptyl v zimnej sezóne. V tomto období sa hodnoty väčšiny rezíduí pohybujú do 4 teplotných stupňov, vo zvyšných mesiacoch je to do približne $2,5 - 3$ stupňov.



Obr. 6.35: Rezíduá modelu

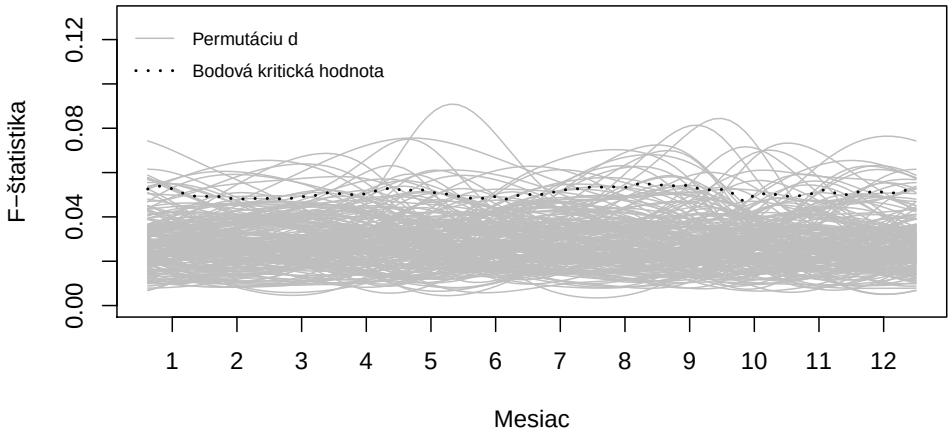
F-test a R^2

Chceli by sme sa teraz pozrieť, či sú vzťahy, ktoré sme odhadli, významné. Ako už bolo avizované v teoretickej časti, použijeme k tomu permutovaný F-test a koeficient R^2.
V prostredí `R` však funkcia, ktorá počíta tento test nie je úplne vhodná pre situáciu, kedy zavádzame podmienku nulového súčtu. Vytvorili sme preto funkciu, ktorá sa s týmto problémom vysporiada. Na obrázku 6.36 je vykreslená krivka príslušnej F-štatistiky spočítanej pre náš model.



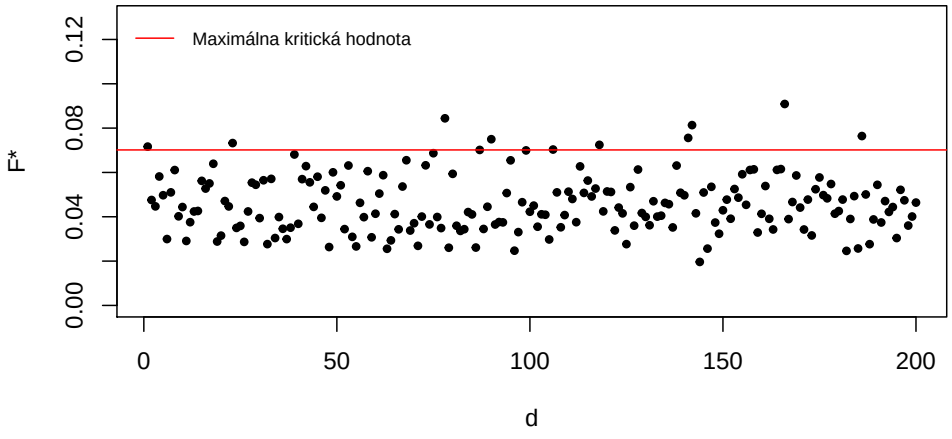
Obr. 6.36: Pozorovaná F-štatistika

Ako bolo uvedené v časti 5.2, túto štatistiku, resp. jej maximum, porovnávame s F-štatistikami spočítanými pre modely s permutovanými pozorovaniami. V tomto prípade bol zvolený počet permutácií $D = 200$. V spodnej časti grafu 6.36 sú vykreslené všetky tieto štatistiky. Sú tu znázornené pre porovnanie so štatistikou z pôvodného modelu. Vidieť, že vo všetkých prípadoch sú ich hodnoty na celom intervale veľmi blízko nule, niekoľnonásobne menšie ako pôvodná štatistika. Presný priebeh kriviek je potom lepšie vidieť na grafe 6.37. Na tomto grafe je zároveň vykreslená bodová 5%-ná kritická hodnota.



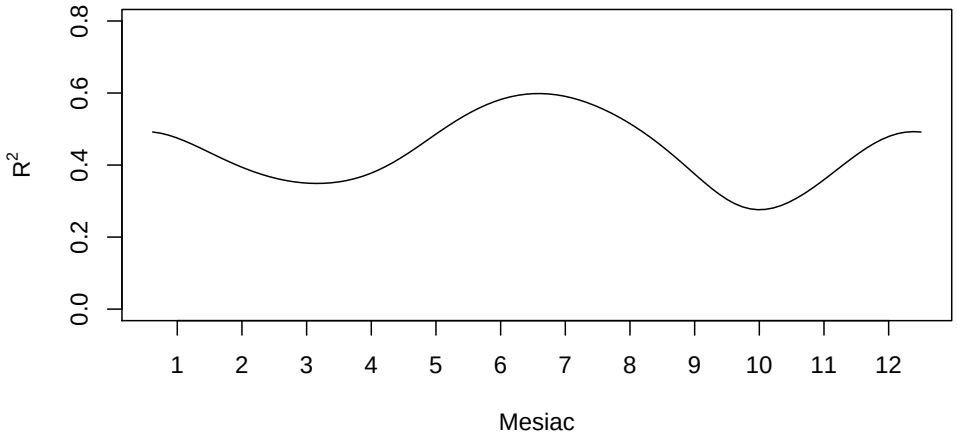
Obr. 6.37: F-štatistiky permutácií, $D = 200$

Určením maxima pre všetky spočítané krivky a následne určením 95%-ného kvantilu týchto hodnôt dostávame maximálnu úroveň kritickej hodnoty. Tá je v tomto prípade rovná $F^*_{0,95} = 0,07$. Tento postup je znázornený na 6.38.



Obr. 6.38: Maximálne hodnoty F-štatistík

Zhrňme teda výsledky testu. Pozorovaná hodnota F-štatistiky vyšla $F^* = 1,48$, určená ako maximum funkcie znázornenej na grafe 6.36, zatiaľ čo 95%-ný kvantil má hodnotu $0,07$, čo výrazne svedčí v prospech prítomnosti efektu skupín na výstupné hodnoty modelu. Naviac, vychádzajúc zo vzťahu (5.11), hodnoty F^* pre náš model a grafu 6.38 rovno vidíme, že p-hodnota testu je rovná nule. Zamietame teda nulovú hypotézu o neprítomnosti efektov klimatických modelov, tak ako sme ich stanovili v modeli.
Ďalej uvádzame hodnoty koeficientu R^2, ktorý udáva aký podiel variabilty medzi pozorovaniami je vysvetlený pomocou modelu. V tomto prípade sa hodnoty pohybujú okolo 50%, čo môžeme považovať za celkom uspokojivý výsledok.



Obr. 6.39: Koeficient R^2

Pri oboch krivkách vidíme veľmi podobný trend. Najsilnejší test a najvyššia hodnota R^2 vyšla pre letné a zimné mesiace. Naopak v prechodných obdobiach jary a jesene sú výsledky horšie. Môžeme si to vysvetliť tým, že v týchto obdobiach, kde je medzi skupinami väčšia variabilita , boli jednotlivé efekty ľahšie identifikovateľné. Naopak, v jarných a jesenných mesiacoch si boli pozorovania aj naprieč skupinami pomerne blízke.

6.4 Porovnanie výsledkov

V predchádzajúcej časti sme posudzovali celkovú kvalitu modelu. V tejto časti sa pokúsime porovnať výsledky, ku ktorým sme sa dopracovali, s tým, čo už bolo k danej téme urobené. Dátový súbor, ktorý sme spracovávali, už bol predmetom analýzy v bakalárskej práci [10]. K problému sa ale pristupovalo odlišne. Podrobný postup je popísaný v [10], na tomto mieste si iba načrtneme metódu, ktorá tu bola vytvorená. K dátam sa opäť pristupovalo ako k funkcionálnym, avšak vzťahy sa skúmali skôr lokálne, než globálne. Bola teda skonštruovaná metóda, pomocou ktorej sa bodovo dopĺňali jednotlivé hodnoty v nekompletných tabuľkách pre diskrétne pozorovania. I keď dáta boli súčasne vyhladené na celom 30-ročnom intervale, algoritmus počítal hodnoty pre každý mesiac a rok samostatne. Cieľom v tejto časti je porovnať výsledky, ktoré poskytuje nami nájdený model s výsledkami

pochádzajúcimi z popísaného algoritmu. Pri voľbe hodnôt, ktoré porovnávame, budeme vychádzať z tých, ktoré si zvolili v uvedenej práci. V tomto prípade boli vybrané dve časové obdobia. Mesiac január v roku 1971 a mesiac júl v roku 1986. Na základe limitácií, ktoré vyplývali z povahy posudzovania hodnôt v danej práci, sa nakoniec použili pre porovnanie iba štyri dvojice modelov z možných 12. Celkovo tak porovnávame 8 hodnôt. Cieľom tohto porovnania je zistiť, ktorý z uvedených postupov poskytuje lepší odhad, teda generuje menšiu chybu. Pre dané body preto porovnáme odhady so skutočnou hodnotou z dátového súboru a určíme chybu.

Pôvodné merania

Uvedieme najskôr pôvodné hodnoty, z ktorých sa pri analýze vychádzalo. Tie sú v tabuľkách 6.2 a 6.3.

Tabuľka 6.2: Pôvodné hodnoty pre január 1971

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				269,176	269,380			
WRF			269,874		268,980			
CCLM								
RegCM								

Tabuľka 6.3: Pôvodné hodnoty pre júl 1986

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				295,986	287,757			
WRF			289,393		289,017			
CCLM								
RegCM								

6.4.1 Výsledky získané pomocou modelu

Odhady - model fANOVA

Pre získanie odhadov v konkrétnych časových bodoch pre náš model sme príslušné predikované krivky vyčíslili v týchto bodoch. Hodnoty sú obsiahnuté v tabuľkách 6.4 a 6.5.

Tabuľka 6.4: Odhadnuté hodnoty pomocou modelu - január 1971

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				272,749	272,757			
WRF			271,154		270,262			
CCLM								
RegCM								

Tabuľka 6.5: Odhadnuté hodnoty pomocou modelu - júl 1986

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				295,986	287,757			
WRF			289,393		289,017			
CCLM								
RegCM								

Chyby - model fANOVA

Ako už bolo povedané, zaujímajú nás chyby, ktorých sa dopúšťame pri odhadoch pomocou zvoleného modelu fANOVA. Spočítame teda rozdiel rozdiel odhadnutých a pôvodných hodnôt. Tabuľky 6.6 a 6.7 uvádzajú tieto hodnoty.

Tabuľka 6.6: Chyby - model fANOVA - január 1971

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				-3,573	-3,378			
WRF			-1,280		-1,283			
CCLM								
RegCM								

Tabuľka 6.7: Chyby - model fANOVA - júl 1986

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				4,458	-1,715			
WRF			0,300		-2,121			
CCLM								
RegCM								

Vidíme, že zrejme najlepšie odhady model poskytuje pre dvojicu WRF a EC-Earth, zatiaľ čo hodnoty pre modely HadGEM2 a RCA sa javia ako najproblematickejšie.

6.4.2 Výsledky získané pomocou algoritmu

V [10] bola uvedená doplňovacia metóda aplikovaná v dvoch modifikáciách. Rozdiel spočíval v spôsobe reprezentácie dát, resp. vo voľbe bázy. Pripomeňme skutočnosť, že k dátam sa tu pristupovalo ako k pozorovaniam na celom intervale 30 rokov. V prvom prípade išlo o Fourierovu bázu s periódou dĺžky jedného roku. V druhom prípade sa použila B-splajnová báza. Výsledky z nášho modelu porovnáme s oboma týmito postupmi.

Chyby - metóda dopĺňania - Fourierova báza

Uvedieme najskôr chyby, ktoré generovala metóda pri použití Fourierovej bázy s 13 bázovými funkciami a ročnou periódou. Tabuľky 6.8 a 6.9 obsahujú tieto hodnoty.

Tabuľka 6.8: Metóda dopĺňania - Fourierova báza - január 1971

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				-2,240	-2,605			
WRF			-0,776		-3,389			
CCLM								
RegCM								

Tabuľka 6.9: Metóda dopĺňania - Fourierova báza - júl 1986

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				5,141	-3,052			
WRF			-0,226		-0,156			
CCLM								
RegCM								

Výsledky, ktoré boli obdržané v tomto prípade, sú podobné tým, ktoré sme spočítali pomocou modelu. Model HadGEM2 sa opäť zdá byť problematickým, obzvlášť pre mesiac júl.

Chyby - metóda dopĺňania - B-splajn báza

V prípade voľby B-splajnovej bázy bol zvolený počet bázových funkcií tak, aby bol uzol v každom bode merania, čiže výrazne vyšší ako v prípade Fourierovej bázy. Chyby, ktorých sa dopustil algoritmus pri použití tejto bázy sú uložené v tabuľkách 6.10 a 6.11.

Tabuľka 6.10: Metóda dopĺňania - B-splajn báza - január 1971

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				-0,786	-1,406			
WRF			0,882		-0,717			
CCLM								
RegCM								

Tabuľka 6.11: Metóda dopĺňania - B-splajn báza - júl 1986

	CNRM	CanESM	EC-Earth	HGEM2	IPSL	MICRO	MPI	NorESM
RACMO								
RCA				6,283	-1,699			
WRF			1,879		1,491			
CCLM								
RegCM								

Porovnanie uvedených hodnôt s chybami nášho modelu v tomto prípade nie je na prvý pohľad také jednoznačné. Zdá sa, že v mesiaci január poskytuje algoritmus značne lepšie odhady, ale pre júl vidíme možné zhoršenie. V záverečnej časti sa podrobnejšie pozrieme na tieto výsledky.

Zhrnutie

Zhrnieme teraz výstup zo všetkých uvedených postupov. Tabuľka 6.12 obsahuje priemerné hodnoty chýb, ktoré vznikli pri počítaní odhadov tromi vyššie popísanými spôsobmi. Úvadzame priemery zvlášť pre obe časové obdobia a takisto celkovú priemernú chybu pre danú metódu.

Tabuľka 6.12: Priemerné chyby počítané tromi spôsobmi

	Model fANOVA	Algoritmus - Fourier	Algoritmus - B-splajn
Január 1971	2,378	2,253	0,948
Júl 1986	2,148	2,144	2,838
Spoločne	2,263	2,198	1,893

Tu už vidíme kvantifikované to, čo sme doteraz pozorovali. Ako najlepšia metóda, teda generujúca v priemere najmenšie chyby, sa ukázal byť dopĺňací algoritmus s použitím B-splajn bázy. Zvyšné dva prístupy, vrátane toho, ktorý je reprezentovaný v tejto práci, poskytujú o niečo horšie, ale vzájomne veľmi podobné výsledky.
Dané porovnanie má však svoje limitácie. Zvolený počet bodov je bohužiaľ pomerne malý. Vo viacerých tabuľkách sme si mohli všimnúť, že jedno číslo je výrazne vyššie než všetky ostatné. Zvyčajne išlo o dvojicu modelov HadGEM2-RCA. Aj spočítané priemery môžu mať preto obmedzenú výpovednú hodnotu.
Rozdiel v dvoch metódach, ktoré boli použité v bakalárskej práci, spočíval vo voľbe bázy. Kľúčovým rozdielom však nebol samotný typ bázy, ale počet bázových funkcií. Ako sa v práci aj uvádza, rozdiel vo výsledkoch sa pripisuje práve tomuto dôvodu. Je to naviac v súlade aj s našimi výsledkami. V prípade Fourierovej bázy s periódou rovnou jednému roku sa predpokladalo veľmi podobné správanie kriviek naprieč jednotlivými periódami a individuálne vplyvy rokov sa tu viac-menej zotierajú. Postup, ktorý sme zvolili v tejto diplomovej práci má však práve ten istý nedostatok. Množstvo pozorovaní z rôznych rokov sme využili na určenie jednej priemernej hodnoty reprezentujúcej skupinu kriviek. Práve z tejto skutočnosti môže plynúť podobnosť výsledkov pre spomenuté dva postupy.
Nie je preto až takým prekvapením, že pri voľbe veľmi vysokého počtu bázových funkcií a individuálnych bodových odhadoch sboli dosiahnuté lepšie výsledky. Aj keď pre krivky v našom modeli bol takisto použitý relatívne vysoký počet bázových funkcií (13 bázových funkcií pre merania v 12 časových bodoch), pri modelovaní priemerných efektov sa časť tejto variability nevyhnutne potlačí.

Záver

V tejto práci sme sa venovali analýze funkcionálnych dát ako jednému zo súčasných trendov matematickej štatistiky. Prvá časť práce sa zameriavala na popis príslušnej teórie. Boli popísané postupy pre prevedenie diskrétnych pozorovaní na funkcionálne dáta. Konkrétne sa definovali najpoužívanejšie systémy bázových funkcií, Fourierov a B-splajnový. Následne sa popísala metóda lineárnej regresie a regularizácie ako vyhladzovacie metódy. Rozsiahlu časť tvoril popis funkcionálneho lineárneho modelu a odhadu jeho parametrov. V ďalšej časti sme tieto poznatky aplikovali na reálne dáta, ktorými boli teplotné merania získané pomocou série klimatických modelov.
Cieľom práce bolo nájsť model, ktorý popisuje vzťahy a závislosti medzi jednotlivými klimatickými modelmi. Vhodným modelom sa ukázal byť dvojfaktorový model funkcionálnej analýzy rozptylu. Pomocou tohto modelu sme odhadli, aký vplyv majú jednotlivé globálne a regionálne modely na merania teploty. Tieto výsledky sme overili pomocou vhodných testov a ukázali sa ako významné.
V záverečnej časti sme potom hodnoty získané pomocou modelu ešte porovnali s výsledkami, ktoré vychádzajú z metódy uvedenej v [10]. Z porovnávaných hodnôt vyšlo, že táto metóda poskytuje o niečo lepšie alebo podobné výsledky, v závislosti na voľbe bázy, ako nami zvolený model fANOVA. Avšak nezanedbateľnou prednosťou, ktorú má postup uvedený v tejto práci je, že pomocou modelu, ktorý sme zostrojili, vieme popísať vzťahy medzi klimatickými modelmi globálne a potenciál analýzy funkcionálnych dát je v tomto prípade využitý vo väčšej miere.

Zoznam použitej literatúry

[1] *Q&A: How do climate models work?* [online]. [cit. 2018-11-12]. Dostupné z: https://www.carbonbrief.org/qa-how-do-climate-models-work.

[2] HNILICA, Jan. *Klimatické modely a jejich systematické chyby* [online]. Praha: Ústav pro hydrodynamiku AV ČR, [cit. 2018-11-12]. Dostupné z: https://kfa.mff.cuni.cz/?p=57.

[3] *Klima, klimatický systém, klimatické modely* [online]. Praha: Katedra fyziky atmosféry MFF UK [cit. 2018-11-12]. Dostupné z: https://kfa.mff.cuni.cz/?p=57

[4] *The Atmosphere around climate models* [online]. Lawrence Livermore National Laboratory, 2017, [cit. 2018-11-12]. Dostupné z: https://str.llnl.gov/december-2017/bader.

[5] RAMSAY, J. O. a B. W. SILVERMAN. *Functional data analysis.* 2nd ed. New York:Springer, c2005. ISBN 978-0387-40080-8.

[6] RAMSAY, J. O., Giles HOOKER a Spencer GRAVES. *Functional data analysis with R and MATLAB.* New York: Springer, c2009. Use R!. ISBN 978-0-387-98184-0.

[7] KOLÁČEK, Jan. *Text prezentácií k predmetu M7777 Aplikovaná analýza funkcionálních dat*. Brno, 2019. Masarykova univerzita.

[8] FERRATY, Frédéric a Philippe VIEU. *Nonparametric functional data analysis: theory and practice.* New York: Springer, c2006. ISBN 978-0387-30369-7.

[9] STRNÁDKOVÁ, Marie. *Analýza rozptylu a chybějící data v klimatických modelech.* Brno, 2017. Bakalárska práca. Masarykova univerzita. Vedúca práce Marie Budíková.

[10] BRÁBLÍK, Jan. *Metody doplňování dat v klimatických modelech.* Brno, 2019. Bakalárska práca. Masarykova univerzita. Vedúci práce Jan Koláček.

[11] ŽILKOVÁ, Alexandra. *Analýza funkcionálních dat v modelu tinnitu.* Brno, 2018. Bakalárska práca. Masarykova univerzita. Vedúci práce Jan Koláček.

[12] RAMSAY, James, Hadley WICKHAM, Spencer GRAVES a Giles HOOKER. *Functional Data Analysis: Package 'fda'. Version 2.4.8.* 2018. Dostupné z: https://cran.r-project.org/web/packages/fda/fda.pdf.

[13] HOLTANOVÁ, E., MENDLIK, T., KOLÁČEK, J., HOROVÁ, I., MIKŠOVSKÝ, J. *Similarities within a multi-model ensemble: functional data analysis framework*, Geosci. Model Dev., 12, 735–747, 2019. Dostupné z: https://doi.org/10.5194/gmd-12-735-2019.