UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Funkcionální ANOVA: analýza pobaltských a finských časových řad teplot vzduchu

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **Mgr. Ondřej Vencálek Ph.D.**
Vypracoval(a): **Bc. Hanuš Hanečka**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2018

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Hanuš Hanečka

Název práce: Funkcionální ANOVA: analýza pobaltských a finských časových řad teplot vzduchu

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek Ph.D.

Rok obhajoby práce: 2018

Abstrakt: Cílem práce je představit základy funkcionální analýzy dat s důrazem na metodu *funkcionální analýzy rozptylu*, kterou aplikujeme na čtyři skupiny časových řad teplot vzduchu naměřených v oblasti *Pobaltí* a *Finska*. Čtenáři se v práci seznámí s přístupy k vyhlazení časových řad teplot vzduchu a tyto znalosti využijí v kapitole zaměřené na *funkcionální analýzu rozptylu*. Součástí práce je i podkapitola s popisem kódu v jazyku `R`.

Klíčová slova: funkcionální analýza dat, funkcionální analýza rozptylu, programovací jazyk R, časové řady, Pobaltí

Počet stran: 88

Počet příloh: 1 CD

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Hanuš Hanečka

Title: Functional ANOVA: the analysis of Baltic and Finnish air temperature time series

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek Ph.D.

The year of presentation: 2018

Abstract: The main aim of this work is to introduce the basics of functional data analysis. The emphasis is put on *functional analysis of variance*, which is applied on four groups of *Baltic* and *Finnish* air temperature time series. Different approaches to smooting air temperature time series are presented. The knowledge of approaches is necessary to understant *functional analysis of variance*. A part of this thesis is a subchapter with the description of the code written in the programming language `R`.

Key words: functional data analysis, functional analysis of variance, programming language R, time series, Baltic

Number of pages: 88

Number of appendices: 1 CD

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně pod vedením pana Mgr. Ondřeje Vencálka Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Úvod **7**

1 Hrátky s meteorologickými stanicemi **8**
1.1 Prezentace datové sady 8
1.2 Příprava datové sady k analýze 18

2 Z funkcionálních dat k odhadu hladké funkce **20**
2.1 Základní myšlenka 20
2.2 Bázové systémy 22
2.3 Odhad vektoru koeficientů bázového rozvoje funkce průměrné denní teploty 26
2.4 Volba počtu bázových funkcí 32

3 Funkcionální ANOVA a časové řady průměrných denních teplot **41**
3.1 Náš záměr 41
3.2 Model funkcionální analýzy rozptylu 45
3.3 Bodová minimalizace 48
3.4 Popis skriptu pro odhad regresních funkcí pomocí bodové minimalizace 51
3.5 Odhad regresních funkcí pomocí bázového rozvoje 58
3.6 Popis skriptu pro odhad regresních funkcí pomocí bázového rozvoje 62
3.7 Testování 69
3.8 Experiment: vyšetření vlastností FP testu 74
3.9 Dodatek: výpočet p-hodnoty *FP* testu 79

Závěr **81**

Literatura **86**

Poděkování

Mé velké poděkování patří Mgr. Ondřeji Vencálkovi Ph.D. za odborné vedení, trpělivost a ochotu, kterou mi v průběhu zpracování diplomové práce věnoval. Děkuji také mé rodině a blízkému okolí za podporu v průběhu studia.

Úvod

V dnešní době jsme obklopeni velkým množstvím informací, které mohou být důležitým zdrojem nových poznatků, jenž přispějí ke zkvalitnění nejen lidského života. Velkým pomocníkem při snaze najít ve spleti čísel a textu nápovědu pro důležitá rozhodnutí, mající dopad na nás všechny, je *statistika*, neustále se rozvíjející matematická disciplína. S růstem výpočetní techniky se *statistice* začínají otevírat nové možnosti a směry. Jedním z ních je oblast funkcionální analýzy dat.

Naším cílem je představit čtenářům základní myšlenku tohoto směru s důrazem na metodu *funkcionální analýzy rozptylu*, kterou využíváme při analýze časových řad průměrných denních teplot naměřených v oblasti *Pobaltí* a *Finska*. Zajímá nás například, zda a jak se od sebe liší chování časových řad naměřených ve čtyřech regionech *Pobaltí*.

V první kapitole se seznámíme se strukturou analyzovaných dat. Ve druhé kapitole se naučíme vyhlazovat, tj. odhadovat průběh funkce z původních pozorování časové řady. Na znalosti získané z této kapitoly navážeme v poslední části věnující se *funkcionální analýze rozptylu*.

Naším záměrem není jen získat teoretické vědomosti, ale i praktickou dovednost tak, abychom byli schopni použít získané znalosti i na jiný typ dat. Součástí poslední kapitoly je podkapitola popisující náš kód v jazyku `R`. Po celou dobu analýzy pracujeme v prostředí softwaru `R` ve verzi 3.3.1.

Kapitola 1

Hrátky s meteorologickými stanicemi

Cílem kapitoly je představit čtenáři analyzovanou datovou sadu a popsat námi vyzkoušené postupy, jak s těmito daty v našich podmínkách co nejefektivněji pracovat v prostředí statisického programu `R`.

1.1. Prezentace datové sady

Datová sada časových řad průměrných denních teplot z meteorologických stanic, se kterou pracujeme, je přístupná na stránkách projektu *European Climate Assessment & Dataset project (ECA&D)*. Najdeme zde nejen časové řady denních průměrných teplot, ale i další datové sady, které mají souvislost s klimatickými extrémy a změnou klimatu. Jedná se například o časové řady denního množství srážek, denní vlhkosti a denní hloubky sněhu [1].

Cílem projektu *ECA&D*[1] je poskytnout vědecké komunitě a širší veřejnosti přístup ke kvalitním datovým sadám z oblasti klimatologie v rámci regionu Evropy a Blízkého východu [2]. Data jsou poskytována *ECA&D* národními meteorologickými službami a univerzitami. Aktuálně na projektu participuje 68 účastníků z 63 zemí, přičemž počet meteorologických stanic se neustále zvyšuje[2].

[1] *ECA&D* je projekt založený v roce 1998 celoevropským sdružením národních meteorologických úřadů *EUMETNET* (aktuální počet zainteresovaných úřadů ve skupině *EUMETNET* je 31)[4]. Projekt je financován ze zdrojů Evropské komise a od roku 2010 je páteří Regionálního datového střediska Světové meteorologické stanice pro Evropu a Blízký východ.[5]

V současnosti se celkový počet meteorologických stanic v analyzované datové sadě pohybuje kolem hodnoty 3 900[3]. Data jsou v pravidelných intervalech aktualizována a doplňována o informace z nových meteorologických stanic.

Je důležité zmínit, že datové sady z meteorologických stanic zveřejněné na webových stránkách projektu jsou k dispozici zdarma výhradně k nekomerčnímu výzkumu a vzdělávacím účelům.[6]

Z vlastní zkušenosti víme, že není jednoduché se k takovému souboru informací dostat jinou cestou. Pokud se někdo z Vás čtenářů snažil někdy nějakým způsobem získat data z metorologické stanice, například z pražské stanice *Klementinum*, tak mohl být nemile zaskočen. Klementinská časová řada, která je ve střední Evropě nejdelší souvisle měřenou časovou řadou denních teplot a tlaku vzduchu, nebyla pro obyčejného uživatele dostupná[7][8]. V současnosti je situace odlišná. Klementinská řada je součástí datového balíčku projektu *ECA&D* a již je k dispozici přímo na webové stránce *Českého hydrometeorologického úřadu* (zkráceně *ČHMÚ*), což před zhruba rokem a půl nebylo možné i přes fakt, že řada byla stažitelná na stránkách projektu[2][9]. Jedná se o jedinou českou meteorologickou stanici zastoupenou v projektu členů *Světové meteorologické organizace*. O významu projektu pro běžného nekomerčního uživatele či studenta může svědčit následující prohlášení převzaté ze stránek *ČHMÚ*: „*Dovolujeme si upozornit, že většinu informací a produktů (data, zpracování dat, posudky apod.) ČHMÚ poskytuje jako placenou službu.*“[9]

Pojďme se blíže podívat na námi analyzovanou datovou sadu. Aktuálně máme k datu 7. 1. 2018 k dispozici 4 289 časových řad průměrných denních teplot měřených v 50 státech napříč Evropou a Blízkým Východem.

V datovém balíčku najdeme tři druhy dat, dvě datové tabulky a sadu časových řad průměrných denních teplot. První datová tabulka popisuje meteorologické stanice kde byla data získávána, druhá dodává bližší informaci o časových řadách sady. V tabulce 1.1 vidíme názornou ukázku prvního typu dat. Řádky této datové

[2]Pro stažení klementinské rady přejdeme na portál *ČHMÚ*→ Historická data→ Počasí→ *Praha Klementinum*→ Denní data ze stanice *Praha Klementinum* [10].

matice specifikují jednotlivé meteorologické stanice. První proměnou je *identifikační číslo stanice*, podle kterého můžeme dohledat časovou řadu průměrných denních teplot naměřených na meteorologické stanici s tímto identifikačním údajem. Podotýkáme, že sloupec *nadmořská výška* je měřen v metrech nad mořem. *Zeměpisná šířka a délka* jsou proměnné, které charakterizují polohu meteorologické stanice. Tyto souřadnice dále využijeme pro vizualizaci polohy meteorologických stanic na mapě.

ID číslo	**Jméno Stanice**	**Země**	**Zeměpisná šířka**	**Zeměpisná délka**	**Nadmořská výška**
27	Klementinum	CZ	+50:05:26	+014:25:09	191
28	Helsinki	FI	+60:10:00	+024:57:00	4
29	Jyvaskyla	FI	+62:24:00	+025:40:59	137
30	Sodankyla	FI	+67:22:00	+026:39:00	179
31	Marseille	FR	+43:18:18	+005:23:48	75
32	Bourges	FR	+47:04:00	+002:22:00	161

Tabulka 1.1: Ukázka datové tabulky 1.

Nejprve převedeme hodnoty sloupečků *Zeměpisná šířka a Zeměpisná délka* do takového formátu, jež je vhodný pro vyobrazení v prostředí statistického softwaru `R`. V současnosti máme tyto proměnné ve formátu *stupně:minuty:sekundy* a naším záměrem je převést hodnoty proměnných *Zeměpisné šířky a délky* na desetinné stupně. Vzorec použitý pro přepočet na stupně je uvedený níže (1.1).

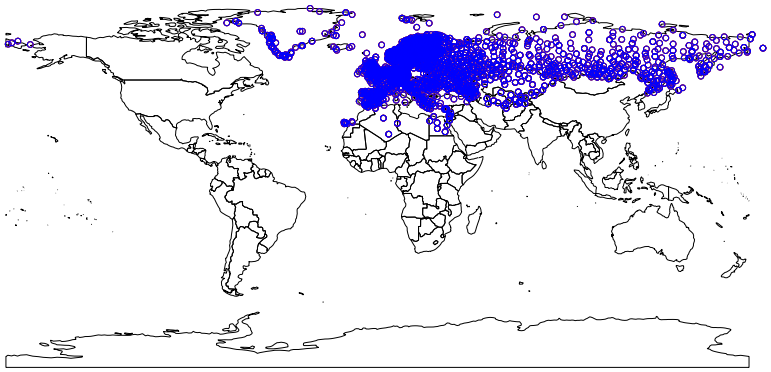
$$Desetinné\ stupně = stupně + \frac{minuty}{60} + \frac{sekundy}{3600} \tag{1.1}$$

Podobu datové matice po transformaci vystihuje následující tabulka 1.2.

ID číslo	**Jméno Stanice**	**Země**	**Zeměpisná šířka**	**Zeměpisná délka**	**Nadmořská výška**
27	Klementinum	CZ	50,09	14,42	191
28	Helsinki	FI	60,17	24,95	4
29	Jyvaskyla	FI	62,40	25,68	137

Tabulka 1.2: Ukázka upravené datové tabulky 1.

Upravenou datovou množinu 1.2 využijeme pro vizualizaci polohy meteorologických stanic ve statistickém programu `R`. Nejprve použijeme[3] knihovnu `rworldmap`[4]. Na obrázku 1.1 vidíme výslednou vizualizaci polohy meteorologických stanic.



Obrázek 1.1: Vizualizace polohy meteorologickych stanic v prostředí `R` s využitím balíčků `rworldmap`.

Podstatná část meteorologických stanic je lokalizovaná na evropském kontinentu. Zbylé meteorologické stanice najdeme v kavkazké oblasi, asijské části Ruska, Grónsku, Africe a na Blízkém východě. Dokonce máme přístup k datům ze stanice v *Jeruzalému*, která je položena ve výšce 815 metrů. Teplota je na jeruzalémské stanici systematicky měřena od roku 1950.

Stát	Německo	Švédsko	Rusko	Finsko	Španělsko
Počet	1031	799	579	392	203

Tabulka 1.3: Tabulka pěti států s nejvyšším počtem meteorologických stanic v námi analyzované datové sadě

[3]Na straně 16 popisujeme obdobnou vizualizaci s využitím balíčku `plotGoogleMaps`.

[4]Původně jsme polohu meteorologických stanic vizualizovali pomocí balíčku `Ggplot2`, avšak tiskárna odmítala vytisknout stránku s touto vizualizací.

Z tabulky 1.3 lze vyčíst, že 3022 stanic je dislokováno v pěti evropských státech. Jedná se zhruba o 78% z celkového počtu 3888 stanic[5].

Nyní představíme druhou datovou tabulku 1.4 popisující sadu časových řad průměrných denních teplot.

ID číslo časové řady	Jméno Stanice	Země	Nadmořská výška	Způsob měření	Začátek měření	Konec měření
100030	Stockholm	SE	44	TG6	1756:01:01	2003:12:31
100032	Kremsmuenster	AT	383	TG7	1876:01:01	2010:12:31
100034	Graz	AT	366	TG7	1901:01:01	2010:12:31
100036	Innsbruck	AT	577	TG7	1901:01:01	2010:12:31
100038	Salzburg	AT	437	TG7	1901:01:01	2010:12:31
100080	Praha-Klementinum	CZ	191	TG14	1775:01:01	2017:10:31
106031	Praha-Klementinum	CZ	191	TG5	1775:01:01	2004:12:31

Tabulka 1.4: Ukázka datové tabulky 2, popisující sadu časových řad průměrných denních teplot

Tabulka 1.4 představuje seznam časových řad datové sady. První proměnou je *identifikační číslo časové řady*, podle kterého můžeme dohledat časovou řadu průměrných denních teplot. Hodnoty proměnných pro příslušné meteorologické stanice *jméno stanice, země, nadmořská výška* se shodují s hodnotami stejně značených proměnných v datové tabulce 1.1. Hodnoty sloupce *Způsob měření*[6] vystihují postup, jakým byl spočítán teplotní průměr pro konkrétní den. Charakteristiky *Začátek měření, Konec měření* přesně specifikují datum, kdy došlo k zahájení měření a ukončení měření teploty v meteorologické stanici[7]. Podotýkáme, že *identifikační číslo stanice* z datové sady představené v tabulce 1.1 se neshoduje s *identifikačním číslem časové řady*[8].

Rozdíl mezi počtem stanic a počtem časových řad[9] je způsoben tím, že teplotní průměry měřené v meteorologicé stanici mohou být počítány různými způsoby. Můžeme tedy pracovat s vícero časovými řadami teplotních průměrů pocházejících

[5]Ke dni 7. 1. 2018.

[6]Podrobný popis jednotlivých postupů výpočtu průměrných denních teplot najdeme na [12].

[7]Je možné, že měření jako takové nebylo v meteorologické stanici ukončeno akorát nedochází k přenosu dat z meteorologické stanice do databáze. Hodnoty proměnných *Zahájení* a *Ukončení měření* jsou seřazeny v pořadí Rok:Měsíc:Den.

[8]V datové sadě popisující jednotlivé časové řady 1.4 máme k dispozici také proměnné *Zeměpisná šířka* a *délka*, mohou sloužit jako charakteristiky, přes které bychom v případě potřeby eventuálně mohli spárovat datovou sadu 1.4 s datovou sadou 1.2

[9]3888 stanic a 4289 časových řad

z jedné meteorologické stanice, avšak lišící se způsobem výpočtu průměrné denní teploty. Názorný příklad vidíme v posledním a předposledním řádku ukázky datové tabulky 1.4.

Datovou tabulku 1.4 využijeme pro základní průzkumovou analýzu meteorologických stanic. Zajímalo nás, ze kterých meteorologických stanic pochází nejstarší záznamy průměrné denní teploty. Tabulka 1.5 poskytuje kýženou informaci.

Jméno Stanice	**Země**	**Nadmořská výška**	**Začátek měření**	**Konec měření**
Stockholm	SE	44	1756:01:01	2003:12:31
Milan	IT	150	1763:01:01	2008:11:30
Praha-Klementinum	CZ	191	1775:01:01	2017:10:31
Hohenpeissenberg	DE	977	1781:01:01	2017:10:30
Bologna	IT	53	1814:01:01	2003:12:31

Tabulka 1.5: Tabulka meteorologických stanic s nejstarším záznamem průmerné denní teploty v rámci námi analyzované datové sady

Nejstarší záznam v analyzované datové sadě časových řad byl naměřen roku 1756 ve *Stockholmu*. Na třetím místě je pražské *Klementinum*. Meteorologická měření na klementinské hvězdárně byla zahájena roku 1752, avšak kvůli neúplnosti záznamů považujeme až rok 1775 za počátek klementinské časové řady.[7]

Jméno Stanice	**Země**	**Nadmořská výška**	**Začátek měření**	**Konec měření**
Sedom	IL	-388	1959:03:01	2016:12:31
Gilgal	IL	-255	1989:01:01	1999:12:31
Massada	IL	-200	1974:07:01	2016:12:31
Hazeva	IL	-135	1988:01:01	2016:12:31
Qeqertarsuatsiaat	GL	-100	2003:09:21	2003:09:21

Tabulka 1.6: Tabulka vzestupně seřazených meteorologických stanic s nejnižší nadmořskou výškou v rámci námi analyzované datové sady

Dále nás u meteorologických stanic zajímala nadmořská výška. V tabluce 1.6 jsou vzestupně seřazené meteorologické stanice s nejnižší nadmořskou výškou z datové sady projektu *ECA*&D.

Čtyři z pěti meteorologických stanic s nejnižší nadmořskou výškou jsou lokalizovány v Izraeli a jedna je z Grónska. Naší pozornost upoutala grónská meteorologická stanice *Qeqertarsuatsiaat* ze které je k dispozici pouze jeden záznam průměrné denní teploty. Očekávali bychom, že grónská stanice bude z datové sady odstraněna pro nedostatečný počet pozorování.

Podívejme se na meteorologické stanice z pohledu nejvyšší nadmořské výšky.

Jméno Stanice	**Země**	**Nadmořská výška**	**Začátek měření**	**Konec měření**
Summit-1	GL	3250	1991:01:09	1994:06:14
Summit-2	GL	3202	1997:11:05	2016:12:31
Sonnblick	AT	3106	1901:01:01	2010:12:31
Zugspitze	DE	2964	1900:08:01	2017:10:30
Sulak	RU	2927	1930:08:15	2013:12:31

Tabulka 1.7: Tabulka sestupně seřazených meteorologických stanic s nejvyšší nadmořskou výškou v rámci námi analyzované datové sady

První dvě meteorologické stanice v pořadí z tabulky 1.7 jsou dislokované v Grónsku a pozičně leží na stejném vrcholu.

ID číslo stanice	**ID číslo časové řady**	**Rok**	**Měsíc**	**Den**	**Průměrná teplota**	**Validace měření**
132	100851	1950	1	1	8,4	0
132	100851	1950	1	2	8,6	0
132	100851	1950	1	3	11,4	0
132	100851	1950	1	4	14,8	0
132	100851	1950	1	5	12,3	0
132	100851	1950	1	6	7,7	0

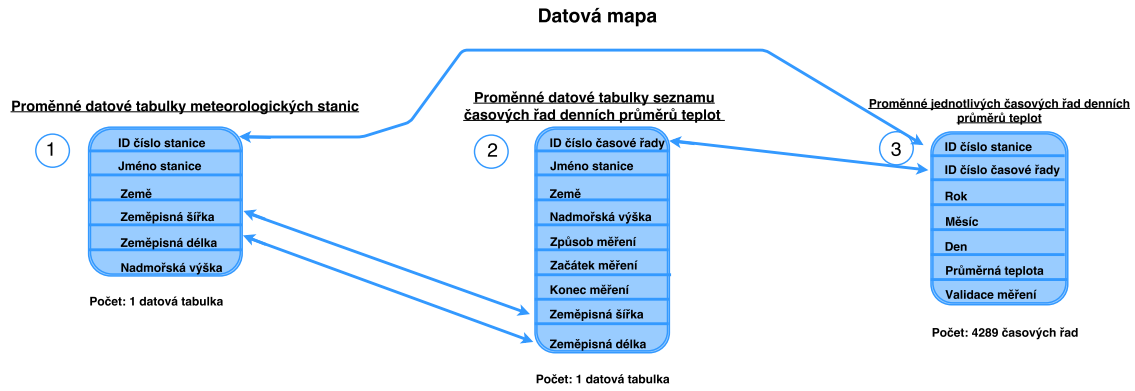
Tabulka 1.8: Ukázka časové řady denních průměrných teplot naměřených v jeruzalémské meteorologické stanici

Nyní představíme třetí druh dat dominující v datovém balíčku. Tabulka 1.8 je ukázkou formátu časové řady průměrných denních teplot naměřených v jeruzalémské meteorologické stanici. Jedná se o jednu z 4289 časových řad datové sady projektu *ECA*&D.

První proměnou je *identifikátor stanice*. Jde o proměnou, která je součástí datové tabulky meteorologických stanic 1.1. Druhá proměnná představuje *iden-*

tifikátor časové řady. Tato proměnná je zastoupena v datové tabulce seznamu časových řad 1.4 a předesíláme, že je pro nás klíčová pro nahrávání časových řad do programu `R`. Dalšími proměnnými jsou *Rok*, *Měsíc* a *Den* jejíž hodnoty upřesňují datum pořízení záznamu průměrné denní teploty zaznamenané ve sloupci *Průměrná teplota*. Proměnná *Validace měření* nabývá jedné ze třech hodnot: nuly, jedničky a devítky. Nula nám dodává informaci o tom, že pozorování bylo ověřeno tedy validováno. Hodnota jedna svědčí o pochybnosti zda napozorovaná hodnota je správná. Devítka indikuje chybějící hodnotu průměrné denní teploty pro daný den.

Nyní čtenáři na obrázku 1.2 popíšeme a zvizualizujeme vztahy mezi datovými soubory jejíž úkázky jsme představili výše.



Obrázek 1.2: Datová mapa, vizualizace struktůry dat a vztahů mezi datovými tabulkam využitými při analýze meteorologických stanic.

Jednotlivé řádky schémat proměnných s čísly 1, 2, 3 představují proměnné datových tabulek a časových řad. První datovou tabulku jsme představili v ukázce 1.1. Jedná se o seznam meteorologických stanic, který jsme využili za účelem získání absolutního počtu stanic a vizualizace polohy meteorologickych stanic 1.1. Druhá datová tabulka je podrobně popsána v ukázce 1.4. Tato tabulka nám poskytuje detailní informaci o jednotlivých časových řadách. Pokud by bylo potřeba, tak lze jednotlivé odpovídající objekty(řádky tabulek) mezi první datovou a druhou datovou tabulkou spárovat pomocí hodnot proměnných *Zeměpisná šířka* a

Zeměpisná délka[10]. Struktůra proměnných časových řad prezentována schématem 3 je pro všech 4289 časových řad shodná. První datovou tabulku můžeme opět propojit[11] s vybranou časovou řadou s využitím proměnné *Id číslo stanice*.

Stěžejní je pro nás vztah mezi druhou datovou tabulkou a časovými řadami. Napadlo nás využívat datovou tabulku, jako seznam časových řad přes který nahráváme požadované časové řady do statistického softwaru `R` skrze proměnnou *Id číslo časové řady*[12]. Hodnoty *Proměnné Id číslo časové řady* využíváme v námi naprogramované funkci pro nahrání dat do `R`.

Na závěr ukážeme výsledek vizualizace polohy meteorologických stanic s pomocí balíčku `plotGoogleMaps` v prostředí `R`, jako alternativu k výkreslení polohy 1.1. Knihovna `plotGoogleMaps` usnadňuje komunikaci se službou *Google mapa*. Předpokladem pro využití této služby je funkční Google učet a API klíč[13], opravňující nás k používání *Google mapy*. Služba je dostupná v neplacené a placené verzi. Pro naše účely postačí neplacená verze.

Opět používáme transformovaná data z tabulky 1.2. K dispozici máme polohu meteorologických stanic v úhlovém systému (*zeměpisná délka*, *zeměpisná šířka*). Souřadnice vhodně převedeme pomocí projekce na kartézský systém. Jedná se o transformaci zobrazující souřadnice z trojrozměrného úhlového systému na dvojrozměrný systém, který je snáze využitelný pro tvorbu mapy[11].

Vše je připravené a můžeme přejít k samotné vizualizaci. V prohlížeči se nám otevře nová záložka mapy světa s vyznačenými meteorologickými stanicemi. Ukázku mapy vidíme na obrázku 1.3 a 1.4.

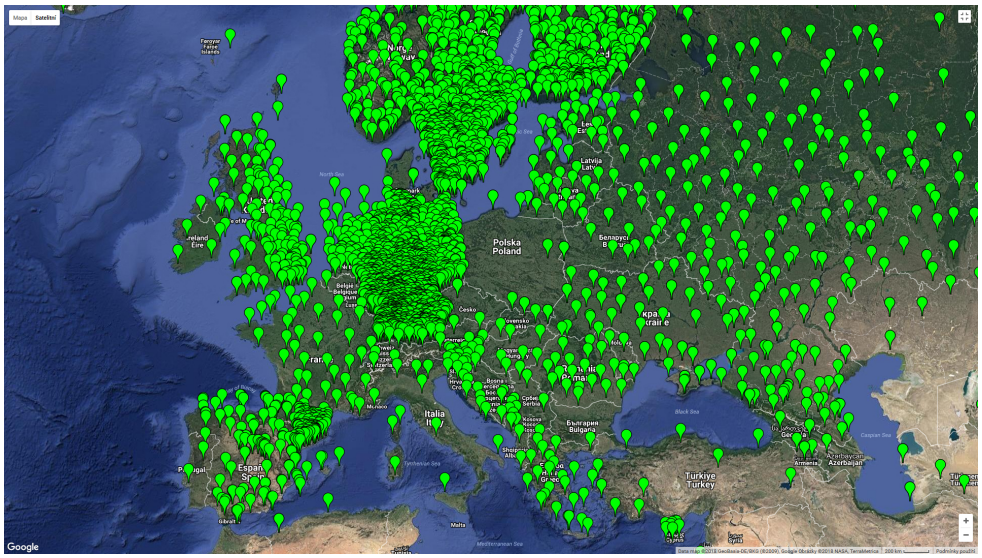
Výhodou google mapy je, že se nejedná o statický obrázek ale o dynamickou

[10] V naší situaci by to byl zcela zbytečný krok vzhledem k podobnosti datové tabulky 1.1 a 1.4

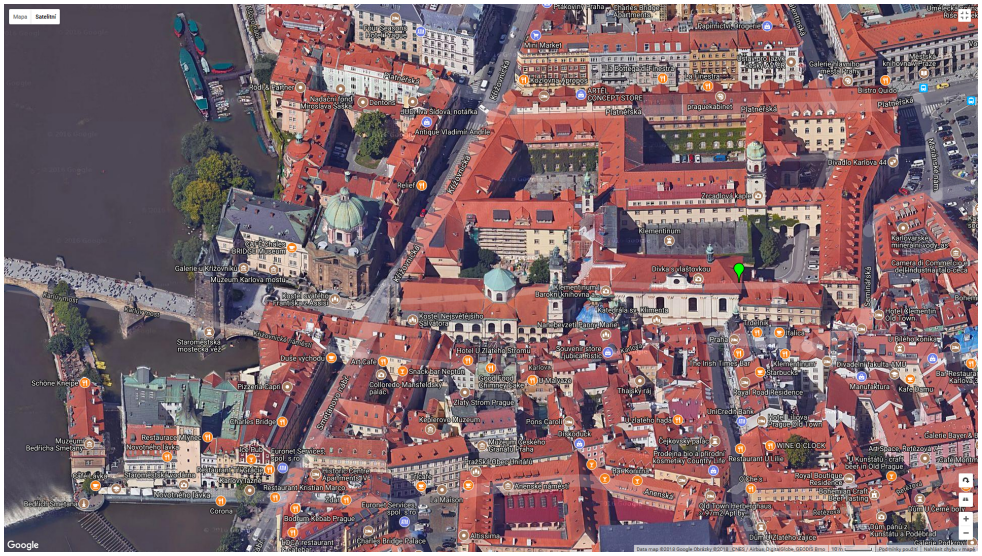
[11] Propojit ve smyslu vyhledání časových řad z datové sady jejíž hodnoty proměnné *Id číslo stanice* odpovídají hodnotě stejné proměnné v řádku tabulky 1.2(řádek reprezentuje meteorologickou stanici). Mohu tak dohledat všechny časové řady z datové sady, které byly naměřeny na určité meteorologické stanici. Pro tento účel budeme používat druhou datovou tabulku 1.4.

[12] Postup zpracování a nahrávání dat podrobněji popisujeme v následující podkapitole.

[13] Jedná se o protokol umožňující komunikaci mezi `R` a *Google mapou*. Podrobnější informace jak získat API klíč najdeme na stránce [13].



Obrázek 1.3: Ukázka vizualizace polohy meteorologickych stanic v prostředí `R` s využitím balíčků `plotGoogleMaps`.



Obrázek 1.4: Ukázka vizualizace polohy meteorologické stanice *Praha-Klementinum* v prostředí `R` s využitím balíčků `plotGoogleMaps`.

mapu se kterou můžeme dále manipulovat (přibližovat, oddalovat) a udělat si tak lepší představu o poloze meteorologických stanic zahrnutých v datové sadě projektu *ECA*&D.

1.2. Příprava datové sady k analýze

Než přejdeme k samotné analýze, je potřeba upravit data do vhodné podoby. Z počátku jsme nevěděli, které časové řady budeme analyzovat a které ne. Napadlo nás, že nejlepší variantou je celý proces čištění a nahrávání dat do `Rka` automatizovat.

Stažené datové tabulky časových řad jsou v textovém formátu (`txt`). Na prvních 18 řádcích je text specifikující datový soubor a proměnné v tabulce. Při nahrávání datových tabulek pomocí funkce `read.table` docházelo k opakovaným chybám, `Rku` vadilo oněch 18 řádků. Naštěstí jsme našli skript napsaný v jazyku *Python*[14][15], který problém s řádky vyřešil. Skript postupně prošel všech 4289 datových tabulek a odstranil u každé z nich řádky s textem. Celá operace trvala okolo tří minut. Problém s nahráváním dat do `Rka` byl vyřešen.

Pro snadnější manipulaci byla v `Rku` každá datová tabulka upravena a převedena na formát `csv`. Úprava spočívala v rozdělení sloupce *Datum* na dílčí sloupce *Rok*, *Měsíc*, *Den*. Napsali jsme v `Rku` jednoduchý `for cyklus` využívající funkce `separate` z balíčku `tidyr` pro rozdělení hodnot z proměnné *Date* do dílčích sloupců. Vykonání úkonů trvalo `Rku` 14 minut. Výsledkem úpravy je datová tabulka ve formátu 1.8.

V dalším kroku jsme datové tabulky přejmenovali, abychom zjednodušili jejich nahrávání. Náš `for cyklus` projde každou datovou tabulku, uloží hodnotu *identifikačního čísla stanice* do paměti softwaru a touto hodnotou přejmenuje datovou tabulku. To v praxi znamená, že časová řadu se jménem *TG_SOUID176097* nese po úspěšném dokončení `for cyklu` název *176097*.

Nové pojmenování datových tabulek odpovídá příslušným hodnotám proměnné *Identifikační číslo časové řady* ze seznamu časových řad (1.4). Posledním krokem bylo napsání funkce `nahravani`, která využívá této shody. Princip fungování funkce je následující: ze seznamu vyfiltrujeme identifikační čísla požadovaných

[14] Všem zájemcům o začátečnické lekce *Pythonu* doporučujeme navštívit stránku [14] respektive [15]. Na stránkách jsou pěkně a detailně vysvětleny základy jazyka *Python*.

[15] Pro správu a úpravu skriptů v jazyku *Python* používáme textový editor `Atom`[16].

časových řad, uložíme je do vektoru, který vložíme do funkce `nahravani`. Funkce z číselných hodnot přidáním koncovky `.csv` vytvoří názvy a nahraje soubory z adresáře do `Rka`.

Na závěr zmiňme, že naší první myšlenkou bylo využít pro správu datové sady databázový systém `MySQL`. Naučili jsme se základy práce s tímto systémem, navázali spojení s `Rkem` pomocí balíčku `RMySQL` a zdárně překonali praktické problémy při nahrávání dat do systému.

Nahrát data ze souborů ve formátu `csv` do databáze `MySQL` může být v některých případech problém. V příloze je `php` skript použitelný pro import velkých `csv` souborů do `MySQL`. Standardním způsobem do databáze nešlo nahrát jeden `csv` soubor se zhruba 88 milóny řádků. V soubory byla uložena data ze všech 4289 datových tabulek. Námi použitý skript operaci zvládl do šesti minut.

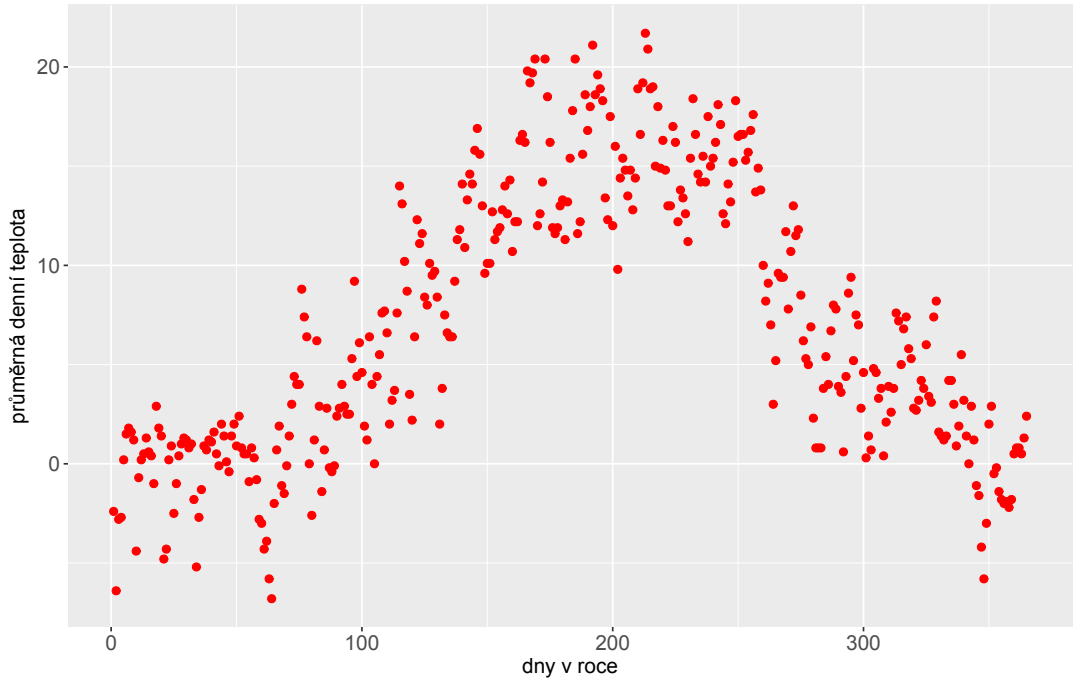
Kapitola 2

Z funkcionálních dat k odhadu hladké funkce

Cílem kapitoly je na praktických příkladech demonstrovat postup, jak z diskrétních pozorování průměrné denní teploty odhadnout hladké funkce průměrné denní teploty. Teorie k této kapitole byla čerpána z [17][18][19][20].

2.1. Základní myšlenka

Na úvod se přenesme do bavorského městečka *Waldassen* ležícího nadohled českých hranic a města *Chebu*.



Obrázek 2.1: Časová řada průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen*

Na obrázku 2.1 vidíme pozorování průměrné denní teploty roku 1951 z meteorologické stanice *Waldassen*. Jedná se o roční výseč z původní časové řady s pravidelnými zaznamy průměrné denní teploty v období 1949 až 1957.

Záznamy průměrné denní teploty na obrázku 2.1 chápeme jako diskrétní pozorování hladké funkce průměrné denní teploty, která je našemu zraku skryta. Naším cílem je s pomocí naměřených dat průměrné denní teploty odhadnout průběh této funkce. Hodnoty průměrné denní teploty, jakožto diskrétní pozorování hladké funkce, představují jeden z příkladů tzv. *funkcionálních dat*.

K dispozici je 365 pozorování průměrné denní teploty y_j s indexem $j{=}1, \ldots, 365$ měřených v časech t_j, kde $t_j{=}1, \ldots, 365$. Časové okamžiky t_j reprezentují dny v roce. Měření y_j považujeme za *funkcionální data* a řídíme se modelem 2.1:

$$y_j = x(t_j) + \epsilon_j. \tag{2.1}$$

Diskrétní pozorování průměrné denní teploty y_j je podle modelu 2.1 rovno součtu funkční hodnoty hladké funkce průměrné denní teploty $x(t_j)$ měřené v časovém okamžiku t_j a náhodné odchylky ϵ_j. Vyhlazením původních pozorování y_j, tj. odhadem funkčních hodnot $x(t_j)$, se snažíme odstranit z dat náhodné odchylky způsobující nehladkost spojnice původních (měřených) hodnot.

Funkční hodnotu funkce průměrné denní teploty $x(t)$ pozorovanou v časovém okamžiku t, kde $t \in [0, 365]$ reprezentujeme rozvojem bázových funkcí $\phi_k(t)$ podle vztahu 2.2.

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}(t) \tag{2.2}$$

Bázové funkce $\phi_k(t)$ jsou lineárně nezávislé funkce, jejichž lineární kombinace vyjadřuje funkci průměrné denní teploty $x(t)$. Tyto bázové funkce volíme podle povahy dat a jejich funkční hodnoty jsou nám známe. K odhadu funkčních hodnot funkce $x(t)$ nám však chybí hodnoty koeficientů c_k. Vektor kooficientů $\boldsymbol{c}$ délky K odhadneme z pozorování průměrné denní teploty y_j měřených v časových okamžicích t_j představujících dny roku. Je také potřeba zvolit vhodnou délku

vektoru koeficientů $\boldsymbol{c}$, tj. vhodný počet bázových funkcí K.

V další části textu se zaměříme na popis stěžejních bázových systémů a jejich možné případy použití v praxi.

2.2. Bázové systémy

Bázový systém můžeme chápat jako nekonečný systém funkcí, kde funkce z tohoto systému jsou mezi sebou lineárně nezávislé a zároveň můžeme pomocí lineární kombinace K bázových funkcí libovolně přesně aproximovat funkci průměrné denní teploty. Pro tento účel se často používají bázové systémy *Fourierovy báze* a *B-splinové báze*[1]. Jedním z příkladů bázového systému jsou bázové funkce v modelu regresní úlohy s jednou proměnnou:

$$1, t, t^2, t^3, \dots t^K \dots \tag{2.3}$$

Je dobré poznamenat, že nikde není napsáno, který bázový systém je vhodný pro aproximaci funkce. Záleží na aplikaci a povaze dat. *Fourierova báze* najde své úplatnění u periodických dat a *B-splinová báze* u neperiodicých. Vraťmě se k obrázku 2.1 z úvodu kapitoly. Bylo již zmíněno, že na této vizualizaci jsou vyobrazeny průměrné denní teploty za rok 1951. Jedná se pouze o část časové řady průměrných denních teplot měřených na meteorologické stanici *Waldassen*. Tato časová řada nezačíná a ani nekončí denními záznamy průměrné denní teploty z roku 1951, ale pokračuje do dalšího roku. V našem případě jsou analyzované datové tabulky 1.8 reprezentující časové řady průměrné denní teploty periodickými daty.

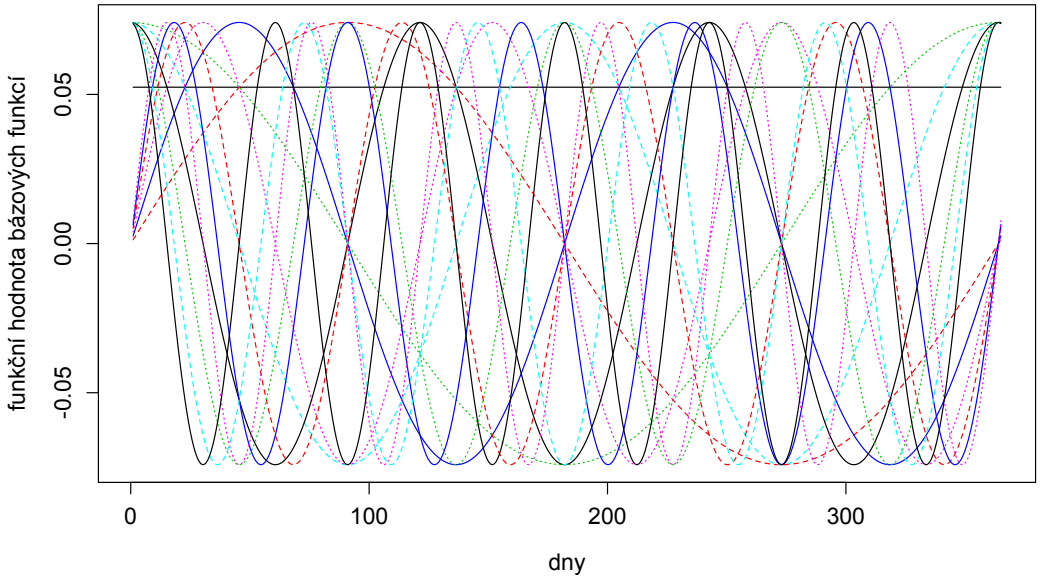
V dalším textu se omezíme na popis *Fourierovy báze* a *B-splinovové báze*. Začneme u *Fourierovy báze*. Tento bázový systém lze popsat jako nekonečnonou posloupnost funkcí *sinus* a *cosinus* s klesající periodou. Součástí *Fourierovy báze* je také konstanta.

[1] Dalším příkladem bázového systému je *Polynomiální bázový systém*, *Polygonální báze* a *Exponenciální báze*

Ve vztahu 2.4 představujeme vzorce pro jednotlivé bázové funkce $\phi_k(t)$.

$$\phi_0(t) = 1, \phi_{2r-1}(t) = \sin(r\omega t), \phi_{2r}(t) = \cos(r\omega t), r = 1, 2 \ldots \tag{2.4}$$

Funkce *sinus* a *cosinus* jsou periodické, což vysvětluje důvod, proč je systém vhodný zejména pro periodická data. Parametr úhlové rychlosti ω určuje periodu $2\pi/r\omega$. S rostoucí hodnotou úhlové rychlosti ω se snižuje perioda funkce. Podívejme se na vizualizaci *Fourierových bázových funkcí* 2.4.



Obrázek 2.2: *Fourierova báze* pro 13 bázových funkcí na intervalu $[0, 365]$

Na obrázku 2.2 pozorujeme 13 bázových funkcí, jejichž tvar vychází ze vztahu 2.4. Všimněme si horizontální linie ve funkční hodnotě 0,05. Jedná se o funkční hodnoty první bázové funkce $\phi_0(t)$, která nabývá konstantní funkční hodnoty pro $t \in [0, 365]$[2].

Důvody, jež nás vedly k volbě 13 bázových funkcí, budou představeny v další podkapitole. Dosazením prvních 13 bázových funkcí z *Fourierovy báze* do 2.2

[2]Všimněme si, že amplituda bázových funkcí na obrázku 2.2 je nižší než jedna, což je způsobeno přenásobením *bázových funkcí* konstantou nevyjímaje první *bázové funkce* $\phi_0(t)$.

získáme odhad funkce průměrné denní teploty.

$$\hat{x}(t) = c_0 + c_1 sin(\omega t) + c_2 cos(\omega t) + \ldots + c_{11} sin(6\omega t) + c_{12} cos(6\omega t). \tag{2.5}$$

Pro výpočet funkčních hodnot odhadu funkce průměrné denní teploty $\hat{x}(t)$ nám prozatím chybí odhad vektoru koeficientů $\mathbf{c}$. Zmíněnou problematikou se budeme zabývat v další podkapitole.

Výběr bázového systému také ovlivňuje derivaci odhadu funkce průměrné denní teploty $x(t)$. Derivaci odhadu této funkce vyjádříme pomocí rozvoje derivace bázových funkcí ϕ_k:

$$Dx(t) = \sum_{k=1}^{K} c_k D\phi_k(t) = c' D\phi. \tag{2.6}$$

Derivace bázových funkcí *Fourierovy báze* $D\phi_k(t)$ mají následující tvar:

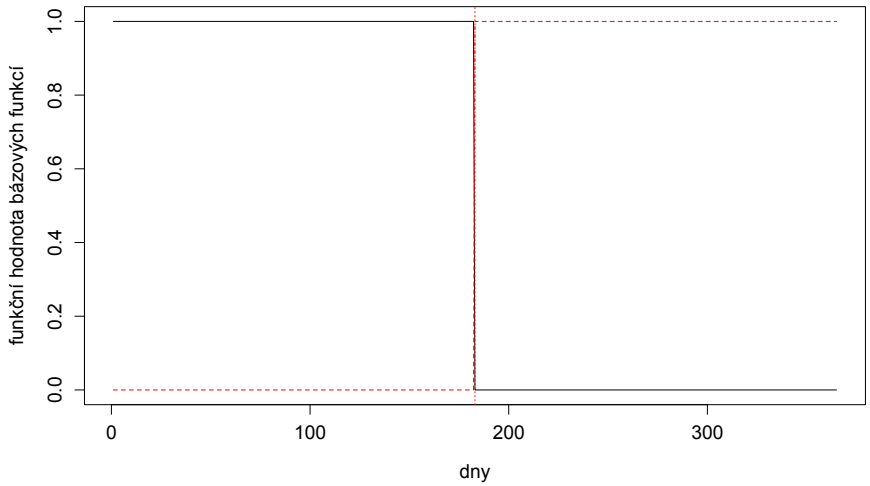
$$Dsin(r\omega t) = r\omega cos(r\omega t),$$
$$Dcos(r\omega t) = -r\omega sin(r\omega t).$$

Přejděme na popis bázového systému *B-splinovové báze*. Nyní předpokládáme, že odhad hladké funkce průměrné denní teploty můžeme chápat jako splinovou funkci, což je po částech spojitá polynomiální funkce. To znamená, že celý definiční obor hladké funkce průměrné denní teploty $[0, 365]$ máme rozdělený na posobě navazující podintervaly $[t_i, t_{i+1})$, na nichž máme odhadnuté polynomické funkce řádu k, které na sebe hladce navazují. Pro každou splinovou funkci platí, že ji lze vyjádřit jako lineární kombinací bázových funkcí 2.2 ze systému *bázových splinů*. Bázové funkce $\phi_k(t)$ tohoto systému generujeme podle rekurentního vztahu [20]:

$$B_{i,1}(t) = 1, t \in [t_i, t_{i+1}), B_{i,1}(t) = 0, t \notin [t_i, t_{i+1}), i = 1, \ldots, n \tag{2.7}$$

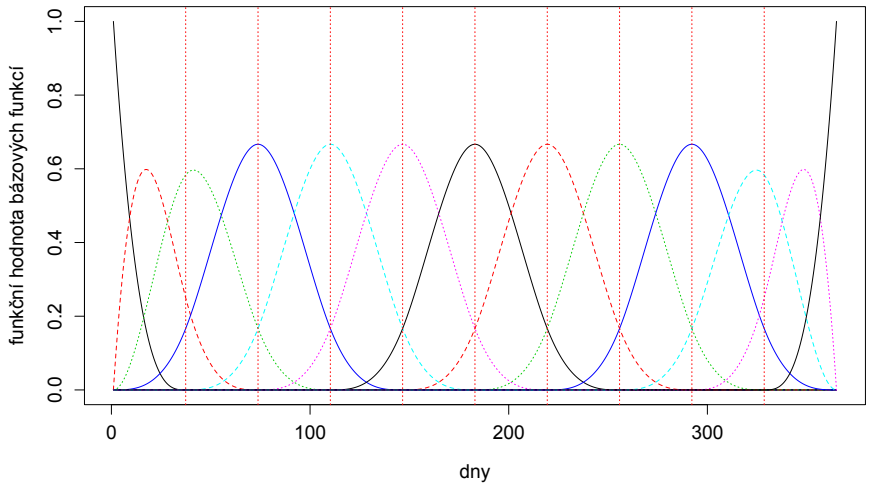
$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t) \tag{2.8}$$

Index k značí *řád bázové spline funkce* $B_{i,k}(t)$.



Obrázek 2.3: *Bázové spline funkce* $B_{1,1}(t)$ a $B_{2,1}(t)$ řádu 1 na intervalu $[0, 365]$

V grafu 2.3 pozorujeme dvě bázové funkce řádu 1 vygenerované podle vztahu 2.7. Interval [0,365] máme rozdělený na dva podintervaly [0,183) a [183,365]. První bázová funkce $B_{1,1}(t)$ nabývá funkční hodnoty jedna na intervalu [0,183) a funkční hodnoty nula na intervalu [183,365]. U bázové funkce $B_{2,1}(t)$ je tomu přesně naopak. Její funkční hodnota je nulová na intervalu [0,183) a rovna jedné na intervalu [183,365].



Obrázek 2.4: 13 *bázových spline funkcí* řádu 4 na intervalu $[0, 365]$

Bázové spline funkce vyššího řádu vznikají rekurentně podle vzorece 2.8. Nyní máme připravené bázové funkce, známe jejich funkční hodnoty na intervalu $[0, 365]$.

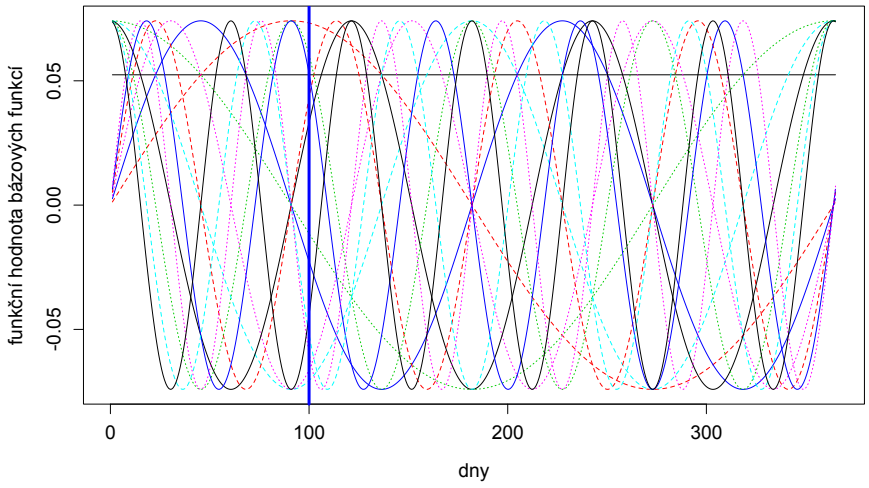
Pro odhad hladké funkce průměrné denní teploty potřebujeme odhadnout vektor koeficientů $\mathbf{c}$ *bázového rozvoje*. Jakmile odhadneme tyto koeficienty, tak nám již nic nebrání ve výpočtu funkční hodnoty odhadu hladké funkce průměrní denní teploty v libovolném časovém okamžiku $t \in [0, 365]$. Jinými slovy, získáme odhad průběhu funkce průměrné denní teploty.

2.3. Odhad vektoru koeficientů bázového rozvoje funkce průměrné denní teploty

Pro odhad koeficientů c_k využijeme standartní verze *metody nejmenších čtverců*. Hledáme odhad koeficientů c_k minimalizující kritérium:

$$SMSSE(\mathbf{y}|\mathbf{c}) = \sum_{j=1}^{365}[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2 \tag{2.9}$$

Hodnoty bázových funkcí $\phi_k(t)$ jsou nám známé. Podívejme se znovu na vizualizaci 13 bázových funkcí z *Fourierovy báze*.



Obrázek 2.5: *Fourierova báze* pro 13 bázových funkcí na intervalu $[0, 365]$ s vyznačenými funkčními hodnotami bázových funkcí v časovém okamžiku 100.

Funkční hodnoty bázových funkcí $\phi_k(t)$ pro časový okamžik $t = 100$ ze vzorce 2.9 zjistíme tak, že se podíváme na funkční hodnoty 13 bázových funkcí v čase t=100. Sledujeme modrou vertikální linii a postupně se díváme na funkční hodnoty 13 bázových funkcí $\phi_k(100)$ a dosadíme je do vztahu 2.9.

Kritérium 2.9 můžeme zapsat v maticovém tvaru:

$$SMSSE(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \boldsymbol{\phi}\mathbf{c})'(\mathbf{y} - \boldsymbol{\phi}\mathbf{c}). \tag{2.10}$$

Kde $\mathbf{y}$ je vektor napozorovaných hodnot průměrné denní teploty v časových okamžicích $t = 1, 2, \ldots, 365$. Složkami vektoru $\mathbf{c}$ délky K jsou koeficienty c_k. *Designová matice* $\boldsymbol{\phi}$ obsahuje funkční hodnoty známých bázových funkcí $\phi_k(t_j)$.

Výraz 2.10 zderivujeme podle složek vektru $\mathbf{c}$ a derivace položíme rovny nule. Získáme soustavu rovnic:

$$2\boldsymbol{\phi}\boldsymbol{\phi}'\mathbf{c} - 2\boldsymbol{\phi}'\mathbf{y} = 0. \tag{2.11}$$

Řešením soustavy 2.11 je odhad $\hat{\mathbf{c}}$, minimalizující hodnotu kritéria 2.10:

$$\hat{\mathbf{c}} = \boldsymbol{\phi}'\boldsymbol{\phi}^{-1}\boldsymbol{\phi}'\mathbf{y}. \tag{2.12}$$

Vyrovnané hodnoty představující odhady funkčních hodnot funkce průměrné denní teploty v čase $t = 1, 2, \ldots, 365$ počítáme:

$$\hat{\mathbf{y}} = \boldsymbol{\phi}\hat{\mathbf{c}} = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{y}. \tag{2.13}$$

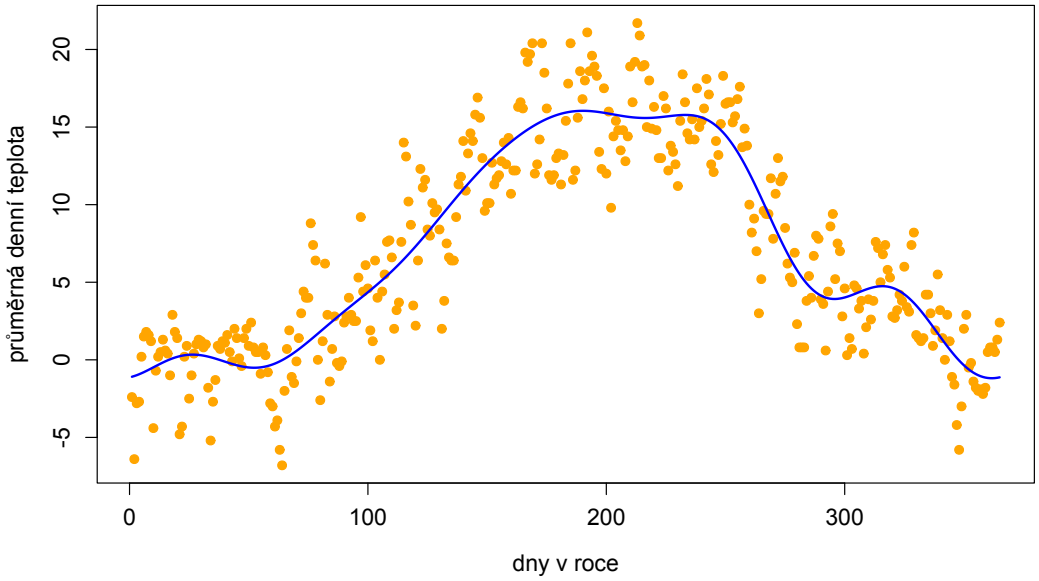
Matici $\mathbf{S} = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'$ nazýváme tzv. *Projekční matici*[3]. Označme vektor vyrovnaných hodnot v čase $t = 1, \ldots, 365$:

$$\hat{\mathbf{x}}(\mathrm{t}) = \hat{\mathbf{y}}. \tag{2.14}$$

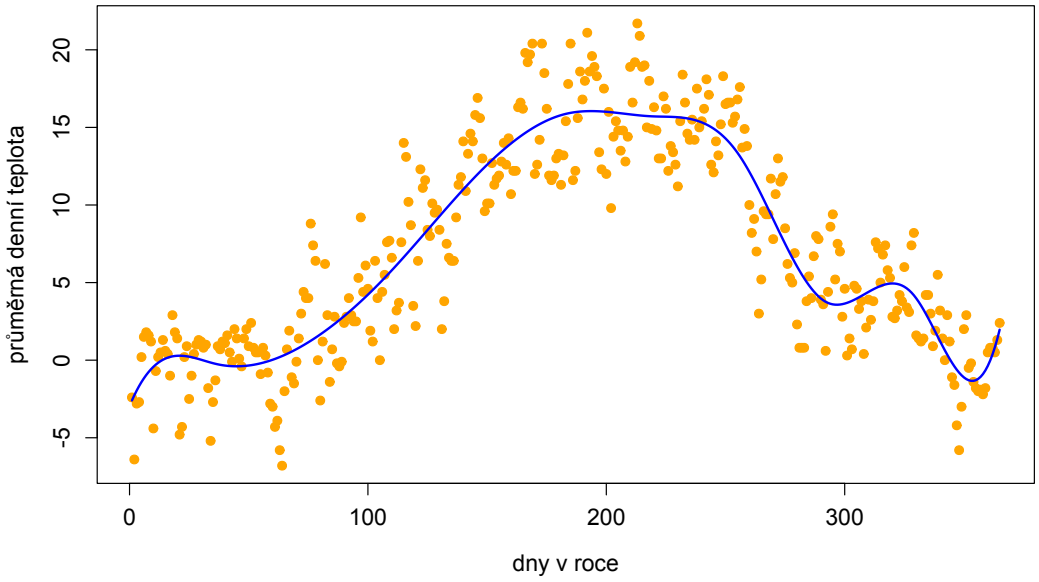
Matice $\mathbf{S}$ konvertuje vektor naměřených hodnot průměrné denní teploty $\mathbf{y}$ na vektor odhadů funkčních hodnot průměrné denní teploty $\hat{\mathbf{x}}(\mathbf{t})$.

Pro srovnání ukážeme odhad funkce průměrné denní teploty s využitím *Fourierovy báze* a *B-spline báze*. Pro odhad vektoru koeficientů $\mathbf{c}$ byla v grafu 2.6 použita *Metoda nejmenších čtverců* s 13 *Fourierovými bázovými funkcemi* $\phi_k(t)$. Při odhadu vektoru koeficientů $\mathbf{c}$, kterému odpovídají vyrovnané hodnoty na obrázku 2.7 používáme 13 bázových funkcí z *B-spline bázového systému*.

[3] V odborné literatuře se pro matici $\mathbf{S}$ používá také název *Hat matice*.



Obrázek 2.6: Odhad funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *Fourierovy báze*.

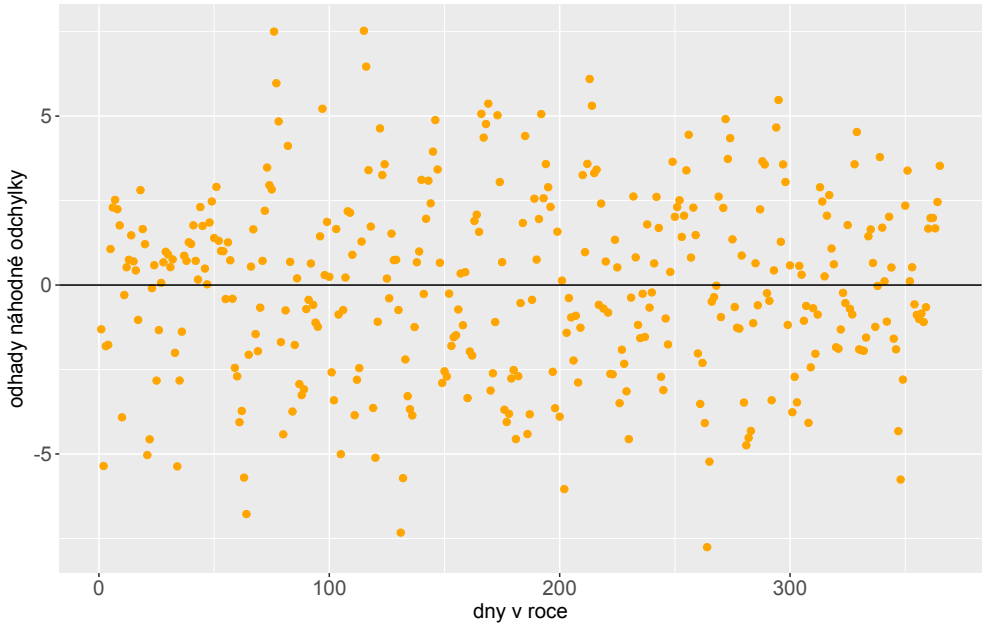


Obrázek 2.7: Odhad funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *B-spline báze*.

Odhady průběhu funkce průměrné denní teploty prezentované v grafech 2.6 a 2.7, vypadají na první pohled použitelně pro další analýzu. Podíváme-li se na zimní měsíce, tj. začátek a konec roku, pozorujeme odlišnost mezi grafy. Průběh odhadu funkce s využitím *Fourierových bázových funkcí* lépe zachycuje informaci z dat v zimních měsících než odhad průběhu funkce pomocí bázových funkcí *B-splinového systému*. Časové řady průměrných denních teplot představují periodická data. V našem příkladu odhadujeme průběh funkce pouze pro rok 1951. Je logické požadovat, aby odhad průběhu funkce průměrné denní teploty v roce 1951 hladce navazoval na odhad průběhu funkce v roce 1952, což je požadavek, jenž v grafu 2.7 není splněn.

Klasickou *Metodu nejměnších čtverců* využijeme v situacích, předpokládáme-li, že náhodná odchylka ϵ_j z modelu 2.1 je bílý šum, tj. posloupnost nekorelovaných náhodných veličin s nulovou střední hodnotou a stejným rozptylem.

V grafu 2.8 vizualizujeme odhady náhodných odchylek vyjádřené jako rozdíly původních pozorování průměrných denních teplot y_j a odhadů funkčních hodnot funkce průměrné denní teploty $\hat{x}(t_j)$, kde $j = 1, \ldots, 365$.



Obrázek 2.8: Odhad náhodných odchylkek z dat průměrných denních teplot měřených v roce 1951 na meteorologické stanici *Waldassen*.

Podívejme se na obrázek náhodných odchylek 2.8 podrobneji. Celkem máme 365 odhadů, z toho 178 odhadů náhodné odchylky leží pod a 187 nad horizontální linkou urcující nulovou strední hodnotu v case $t \in [0, 365]$. Průměr náhodných odchylek průmerné denní teploty vychází řádově 10^{-4}, je tedy prubližne nulový [4]. Odhad směrodatné odchylky je roven $2,71$.

Podle našeho mínění není v příkladu porušen předpoklad bílého šumu, pokud by porušen byl, tak i z grafu 2.6 je patrné, že v tomto případě nejde a dramatický prohřešek.

Zavedeme ještě klíčový pojem pro další kapitoly, tzv. *stupeň volnosti vyhlazení*:

$$df = stopa\mathbf{S} = stopa(\mathbf{S}\mathbf{S}') \tag{2.15}$$

Stupeň volnosti vyhlazení vyjádříme, jako stopu *Projekční matice*, tzn. součet diagonálních prvků matice $\mathbf{S}$ nebo matice $\mathbf{S}\mathbf{S}'$. Rovnost mezi *stopami* matic vychází z *idempotence* matice $\mathbf{S}$:

$$\mathbf{S} = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'$$

$$\mathbf{S}' = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'$$

$$\mathbf{S}\mathbf{S}' = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}(\boldsymbol{\phi}'\boldsymbol{\phi})(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}' = \boldsymbol{\phi}(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}' = \mathbf{S}.$$

[4] Což vychází z podstaty odhadu náhodných odchylek

V případech, kde je výše zmíněný předpoklad o náhodné odchylce v podobě bílého šumu výrazně porušen, je vhodné využít pro odhad vektoru koeficientů $\mathbf{c}$ *váženou metodu nejměnších čtverců*:

$$SMSSE(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \boldsymbol{\phi}\mathbf{c})'\mathbf{W}(\mathbf{y} - \boldsymbol{\phi}\mathbf{c}), \tag{2.16}$$

kde $\mathbf{W}$ je symetrická pozitivně definitní matice, jenž je inverzí varianční maticí náhodných odchylek $\boldsymbol{\Sigma_e}$:

$$\mathbf{W} = \boldsymbol{\Sigma_e}^{-1}. \tag{2.17}$$

Odhad vektoru koeficientů $\hat{\mathbf{c}}$ je ve tvaru:

$$\hat{\mathbf{c}} = (\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{W}\mathbf{y}. \tag{2.18}$$

Hat matici počítáme následovně:

$$\mathbf{S} = \boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{W}. \tag{2.19}$$

Dosadíme-li za matici vah $\mathbf{W}$ jednotkovou matici $\mathbf{I}$, dostaneme se na původní model 2.10. V klasickém modelu *Metody nejmenších čtverců* je předpokládaný tvar varianční matice $\boldsymbol{\Sigma_e} = \sigma^2\mathbf{I}$.

Parametr σ^2 odhadneme, jako součet čtverců odhadů náhodných odchylek podeleny *stupněm volnosti* $n - K$, kde K je počet *bázových funkcí*:

$$\mathrm{s}^2 = \frac{1}{365 - K}\sum_{j=1}^{365}(y_j - \hat{y}_j)^2. \tag{2.20}$$

Odhad varianční matice náhodných odchylek pro model *Vážené metody nejmenších čtverců* počítáme ze vztahu:

$$\hat{\boldsymbol{\Sigma}}_\mathbf{e} = (N - 1)^{-1}\mathbf{E}'\mathbf{E}. \tag{2.21}$$

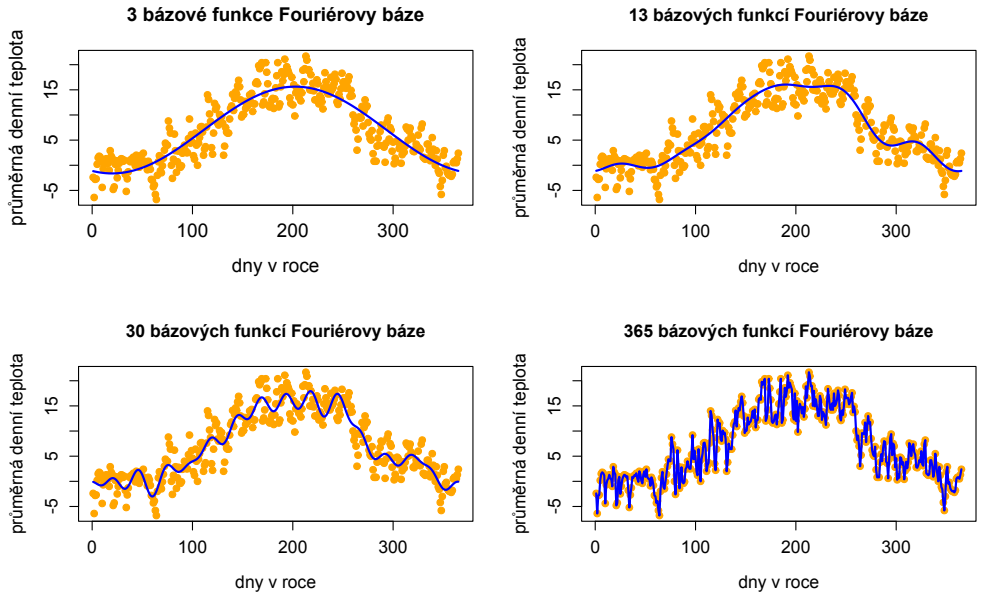
N je počet replikací odhadů funkce průměrné denní teploty a matice $\mathbf{E}$ představuje matici odhadů náhodných odchylek s N řádky a 365 sloupci. Řádek této matice je vektor odhadů náhodných odchylek jehož složkami jsou pozorování $\hat{\epsilon}_j$ z obrázku 2.8, bereme-li jako jednu z replikací odhad funkce průměrné denní teploty na meteorologické stanici *Waldassen*.

Pracujeme s časovými řadami a nereplikujeme měření konkrétní časové řady. Můžeme vzít N časových řad měřených na meteorologických stanicích, které jsou

si svojí polohou blízké. Z těchto dat odhadnout příslušné funkce průměrných denních teplot, náhodné odchylky a napočítat odhad varianční matice $\hat{\mathbf{\Sigma}}_{\mathbf{e}}$ reflektující varianční strukturu náhodných odchylek na meteorologických stanicích, z jejichž dat byly získány odhady náhodných odchylek. Matice bude regulární v případě, že N bude alespoň rovno 365.

2.4. Volba počtu bázových funkcí

Jak zvolit počet bázových funkcí K? Čím větší K, tím lépe přiléhá odhad funkce průměrné denní teploty k napozorovaným hodnotám, zároveň však ztrácíme míru obecnosti. Příliš malé K naopak vede k odhadu velmi obecné hladké funkce bez potřebné interpretační hodnoty. Na obrázku 2.9 vidíme názornou ukázku, jak počet *bázových funkcí* ovlivňuje vyhlazení původních pozorování.



Obrázek 2.9: Odhady funkcí průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* při různém nastavení počtu bázových funkcí *Fourierovy báze*.

Hledáme počet *bázových funkcí* zajišťující odhad hladké funkce průměrné denní teploty, jejíž průběh zároveň kopíruje charakter dat. Jedná se o kompromis. Na konflikt mezi vyhlazením a ztrátou informace můžeme pohlížet z jiného úhlu pohledu. Pro hodně velké K jsou funkční hodnoty funkce *vychýlení* odhadu

v časových okamžicích t malé, tj. nezáporné a blízké nule:

$$Bias[\hat{x}(t)] = x(t) - E[\hat{x}(t)] \tag{2.22}$$

Předpokládáme-li nulovou střední hodnotu náhodných odchylek ϵ_j v modelu 2.1, je *vychýlení* pro $K = n$ rovno nule, ale současně je neúměrně vysoká variabilita odhadu:

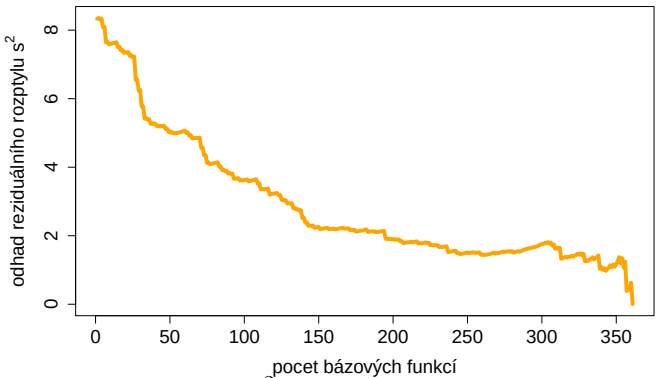
$$Var[\hat{x}(t)] = E[\hat{x}(t) - E[\hat{x}(t)]]^2. \tag{2.23}$$

Tyto úvahy vedou ke snížení počtu bázových funkcí, který redukuje variabilitu odhadu. Musíme dávat pozor, aby při zmenšení počtu K nedošlo zároveň k neúměrnému navýšení *vychýlení* (2.22). Naší snahou je najít kompromis mezi *vychýlením* a *variabilitou* odhadu. Tato idea nás směřuje k minimalizaci *Střední čtvercové chyby*:

$$MSE[\hat{x}(t)] = Bias^2[\hat{x}(t)] + Var[\hat{x}(t)]. \tag{2.24}$$

Ve většině aplikací nemůžeme pro neznalost $x(t)$ a z toho plynoucí neschopnost výpočtu *vychýlení* (2.22), kritérium (2.24) minimalizovat. Důležitější je myšlenka stojící za touto formulí: jsme ochotni tolerovat lehké *vychýlení*, pokud povede k výraznému snížení variability odhadu a celkovému snížení hodnoty *Střední čtvercové chyby*. Toho v praxi můžeme dosáhnout *penalizací funkce za nehladkost*. To sice povede k *vychýlení* odhadu, ale my jsme schopni ovlivnit míru *vychýlení* nastavením váhy penalizačního členu. O *penalizaci nehladkosti* ještě bude psáno.

Jedna z možných strategií výběru počtu bázových funkcí je odhadnout parametr σ^2 (2.20) při různém nastavení počtu bázových funkcí a zvolit nastavení dostatečně minimalizující odhad parametru σ^2.



Obrázek 2.10: Odhad parametru σ^2 při různém nastavení počtu bázových funkcí z *Fourierovy báze*.

Na obrázku 2.10 pozorujeme vývoj odhadu parametru σ^2 v závislosti na počtu *bázových funkcích*. Odhady počítáme z pozorování průměrných denních teplot měřených na stanici *Waldassen*. Od počtu 153 do zhruba 300 *bázových funkcích* se odhad parametru stabilně pohybuje kolem hodnoty 2. Vezmeme-li jako optimální počet 153 *bázových funkcích*, nebude odhad funkce použitelný pro analýzu[5].

Idea využít pro volbu vhodného počtu *bázových funkcích* odhad parametru σ^2 souvisí se sofistikovanějším přístupem tzv. *zobecněnou cross-validací*. K této metodě se vrátíme později.

Pomocí *penalizace za nehladkost* můžeme v kombinaci s *váženou metodou nejmenších čtverců* odhadnout hladký průběh funkce i přes větší počet *bázových funkcí*, tak aby měl výsledný odhad interpretační hodnotu. Nejprve je potřeba kvantifikovat *nehladkost funkce*:

$$PEN_2(x) = \int_0^{365} [D^2 x(t)]^2 \, \mathrm{d}t. \tag{2.25}$$

Čtverec druhé *derivace* $[D^2 x(t)]^2$ funkce $x(t)$ v časovém okamžiku t nazýváme *zakřivením* funkce v časovém okamžiku t. Pro *přímky* například platí, že jejich *druhá derivace* je nulová, tj. nemají žádné *zakřivení*. *Nehladkost funkce* vyjadřujeme ve formě určitého integrálu, integrační mezí je časový interval $[0, 365]$.

Pracujeme-li s funkcí s vysokou variabilitou, jejíž hodnoty druhé derivace jsou neúměrně vysoké, použijeme derivaci vyššího řadu. *Nehladkost funkce* vyjádříme obecně ve tvaru:

$$PEN_m(x) = \int_0^{365} [D^m x(t)]^2 \, \mathrm{d}t. \tag{2.26}$$

D^m představuje *derivaci* řádu m.

Dále je potřeba modifikovat *váženou metodu nejmenších čtverců*, tj. zakomponovat do ní *penalizaci za nehladkost* $\lambda PEN_2(x)$. Minimalizované kritérium *penalizované vážené metody nejmenších čtverců* je tvaru:

$$PENSSE_\lambda(x|\mathbf{y}) = [\mathbf{y} - \mathbf{x}(\mathrm{t})]' \mathbf{W} [\mathbf{y} - \mathbf{x}(\mathrm{t})] + \lambda PEN_2(x), \tag{2.27}$$

kde $\mathbf{x}(\mathrm{t})$ je vektor funkčních hodnot funkce x, pozorovaných v časových okamžicích t, $t = 1, \ldots, 365$. Hledáme odhad průběhu takové funkce, která minimalizuje

[5] Již při počtu 30 *bázových funkcích* je odhad průběhu funkce komplikovaný, viz obrázek 2.9.

hodnotu kritéria (2.27) přes prostor všech funkcí x, pro které je definována nehladkost $PEN_2(x)$. Parametr λ určující míru vyhlazení původních hodnot nazýváme *parametrem vyhlazení*. Pro malé hodnoty parametru λ je funkce velmi nehladká. Jde-li $\lambda \to 0$, při dostatečném počtu bázových funkcí napozorované hodnoty inerpolujeme. V opačném případě větší hodnoty λ přinášejí znatelnější vyhlazení pozorovaných hodnot.

Penalizační člen za nehladkost $PEN_m(x)$ můžeme s využitím vzorce 2.2 přepsat do tvaru:

$$PEN_m(x) = \int_0^{365} [D^m x(t)]^2 \, \mathrm{d}t = \int_0^{365} [D^m \mathbf{c}'\boldsymbol{\phi}(t)]^2 \, \mathrm{d}t$$

$$= \int_0^{365} [\mathbf{c}' D^m \boldsymbol{\phi}(t) D^m \boldsymbol{\phi}(t)' \mathbf{c}] \, \mathrm{d}t = \mathbf{c}' \int_0^{365} [D^m \boldsymbol{\phi}(t) D^m \boldsymbol{\phi}(t)' \, \mathrm{d}t] \mathbf{c}$$

$$= \mathbf{c}'\mathbf{R}\mathbf{c},$$

kde

$$\mathtt{R} = \int_0^{365} [D^m \boldsymbol{\phi}(t) D^m \boldsymbol{\phi}(t)'] \, \mathrm{d}t$$

je matice s K řádky a K sloupci, v praxi počítaná analyticky. Takto vyjádřený *penalizační člen* dosaďme do minimalizovaného kritéria (2.27):

$$PENSSE_\lambda(x|\mathbf{c}) = (\mathbf{y} - \boldsymbol{\phi}\mathbf{c})'\mathbf{W}(\mathbf{y} - \boldsymbol{\phi}\mathbf{c}) + \lambda \mathbf{c}'\mathbf{R}\mathbf{c}. \tag{2.28}$$

Zderivováním kritéria (2.28) podle složek vektoru $\mathbf{c}$ získáme soustavu:

$$2\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi}\mathbf{c} - 2\boldsymbol{\phi}'\mathbf{W}\mathbf{y} + 2\lambda\mathbf{R}\mathbf{c} = 0. \tag{2.29}$$

Odhad $\hat{\mathbf{c}}$ vektoru koeficientů $\mathbf{c}$ počítáme:

$$\hat{\mathbf{c}} = (\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi} + \lambda\mathbf{R})^{-1}\boldsymbol{\phi}'\mathbf{W}\mathbf{y}. \tag{2.30}$$

Vektor vyhlazených hodnot $\hat{y}_j$ měřených v časových okamžicích t, $t = 1, \ldots, 365$ vyjádříme:

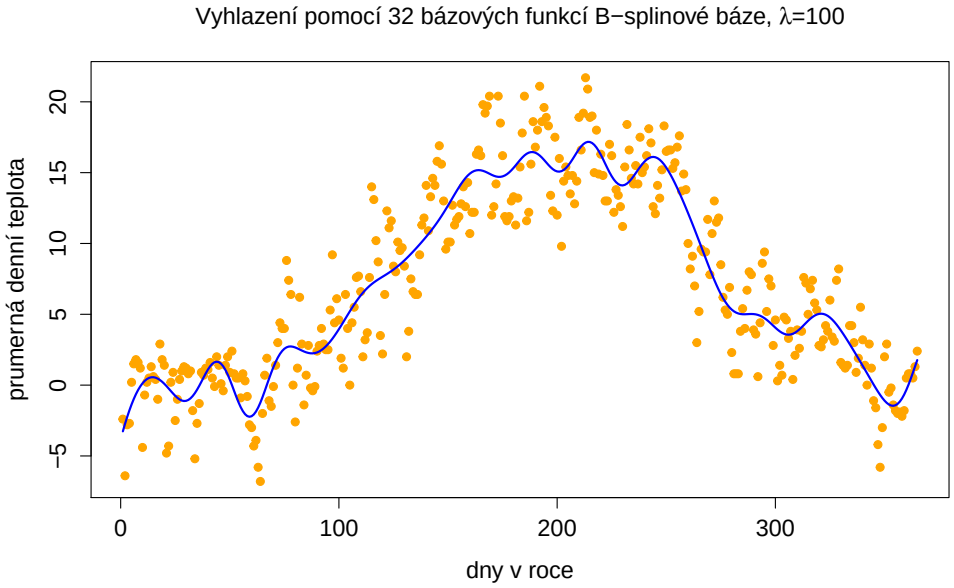
$$\hat{\mathbf{y}} = \boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi} + \lambda\mathbf{R})^{-1}\boldsymbol{\phi}'\mathbf{W}\mathbf{y} = \mathbf{S}_{\phi,\lambda}\mathbf{y}. \tag{2.31}$$

Projekční matici počítáme ve tvaru:

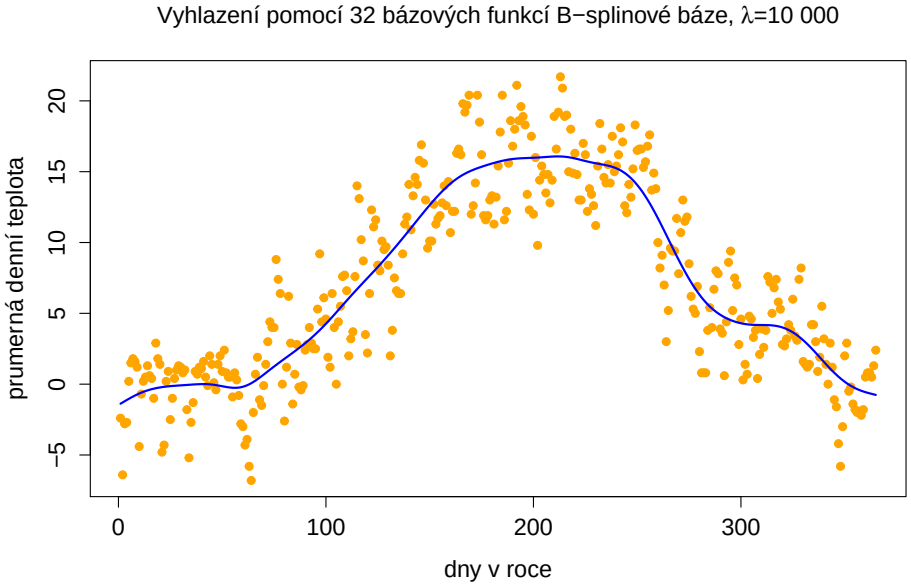
$$\mathbf{S}_{\phi,\lambda} = \boldsymbol{\phi}(\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi} + \lambda\mathbf{R})^{-1}\boldsymbol{\phi}'\mathbf{W}. \tag{2.32}$$

Tato matice není *idempotentní*, tj. neplatí rovnice $\mathbf{S}_{\phi,\lambda}\mathbf{S}_{\phi,\lambda} \neq \mathbf{S}_{\phi,\lambda}$. Srovnáme-li *projekční matici* (3.4) s *projekční matici vážené metody nejmenších čtverců*, vidíme jedinou odlišnost v přidání $\lambda\mathbf{R}$ do *inverzní* matice $(\boldsymbol{\phi}'\mathbf{W}\boldsymbol{\phi})^{-1}$. Matice budou identické, nastavíme-li hodnotu parametru $\lambda = 0$.

Podívejme se na vizualizaci odhadu průběhu funkce průměrné denní teploty při malé a velké hodnotě parametru λ.



Obrázek 2.11: Odhad průběhu funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *B-spline bázových funkcí*.



Obrázek 2.12: Odhad průběhu funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *B-spline bázových funkcí*.

Na obrázcích 2.11, 2.12 vidíme, jak nastavení *vyhlazovacího parametru* λ ovlivňuje vyhlazení dat. Čím větší je hodnota parametru λ, tím jsou pozorování průměrných denních teplot vyhlazenější. Hlavně si všimněme, že odhad funkce průměrné denní

teploty na obrázku 2.12 ve srovnání s odhadem funkce v grafu 2.7, obstojně zachycuje informaci v zimních měsících.

Zbývá nám určit optimální hodnotu *vyhlazovacího parametru* λ při daném počtu bázových funkcí K. Pro tento úkol využijeme ideu tzv. *cross-validace*: pozorování průměrných denních teplot rozdělíme do *tréninkové datové sady* a *validační datové sady*. Na datech *tréninkové sady* natrénujeme model, tedy odhadneme vektor koeficientů $\hat{\mathbf{c}}$ a vypočítáme vyrovnané hodnoty pro pozorování z *validační sady*, což jsou pozorování nevyužitá pro odhadu koeficientů $\hat{\mathbf{c}}$. Následně podle vybraného kritéria porovnáme tyto původní pozorování z *validační sady* s jejich vyrovnanými hodnotami a získáme srovnaní napříč modely.

V našem případě použijeme extrémnější variantu *cross-validace*. V *tréninkové sadě* je zahrnuto obecně $n - 1$ pozorovní, tj. 364 pozorování průměrné denní teploty, *validační sada* pojímá zbylé jedno pozorování. Opět na pozorováních z *tréninkové sady* odhadneme vektor koeficientů $\hat{\mathbf{c}}$ a použijeme jej na výpočet vyrovnané hodnoty pro jedno pozorování z *validační sady*. Tento postup výpočtu vyrovnané hodnoty zopakujeme pro každé ze zbylých 364 pozorování průměrné denní teploty. Získáme 365 vyrovnaných hodnot a spočítáme *residuální součet čtverců*. Celý proces této extrémnější *cross-validace* zopakujeme pro různé hodnoty parametru λ a vybereme hodnotu *vyhlazovacího parametru* minimalizující *residuální součet čtverců*.

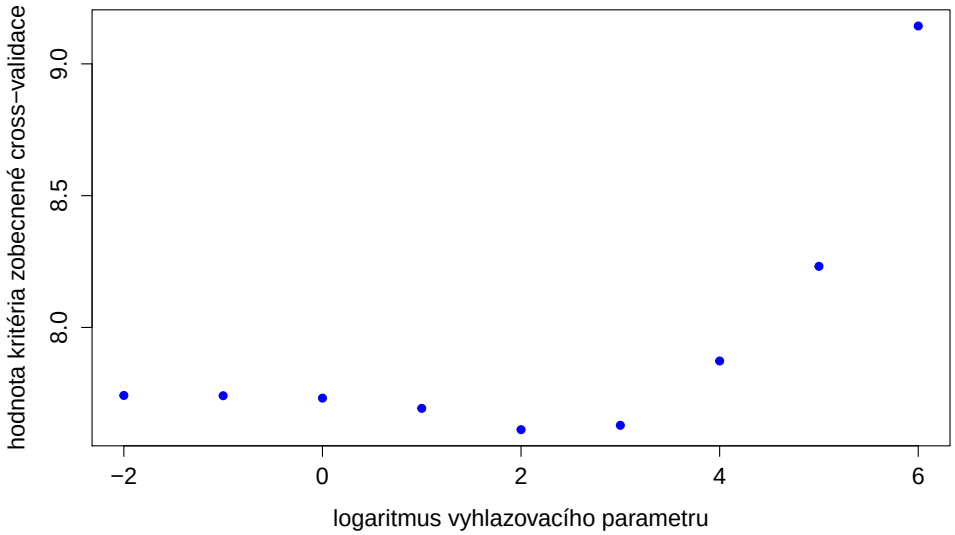
Tato metoda má dvě nevýhody. První je početní náročnost. Pro tisíce pozorování není tato forma extrémnější *cross-validace* nejvhodnějším přístupem. Druhou nevýhodou je, že při minimalizaci *cross-validačního* kritéria, v podobě *residuálního součtu čtverců*, má metoda tendenci vybrat hodnotu parametru λ nedostatečně vyhlazující data.

Dopad výše zmíněných nedostatků mírní *zobecněná cross-validace*, jednoduší verze *cross-validace* nevyžadující n krát odhad vektoru koeficientů $\hat{\mathbf{c}}$:

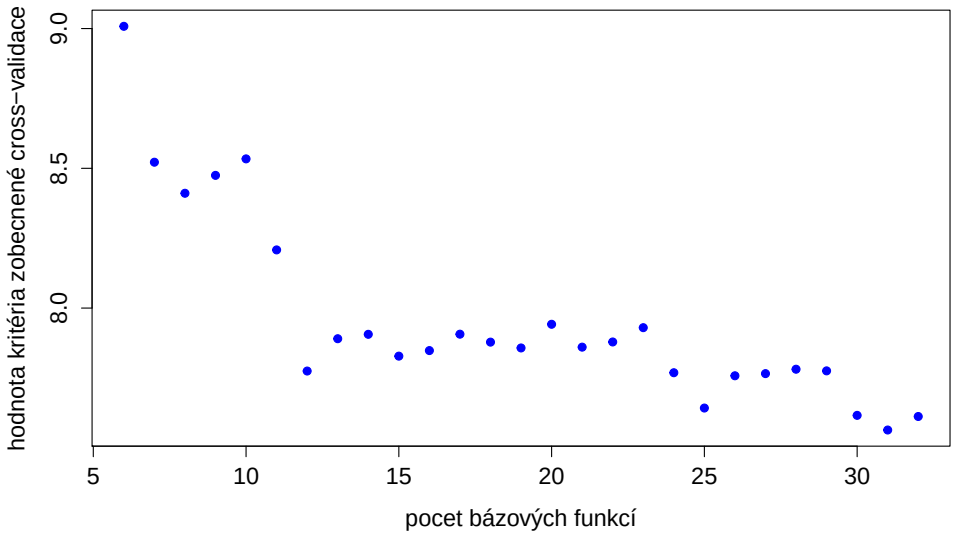
$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right), \tag{2.33}$$

kde pravý faktor představuje odhad parametru σ^2 (2.20) a *stupeň volnosi vyhla-*

zení $df(\lambda)$ počítáme dle vzorce 2.15. Nyní napočítáme hodnoty kritéria *zobecněné cross-validace* 2.33 pro různá nastavení parametru λ při počtu 32 *bázových funkcích* a vybereme hodnotu parametru λ minimalizující hodnotu kritéria 2.33.



Obrázek 2.13: Hodnoty kritéria *zobecněné cross-validace*(2.33) počítané z pozorování průměrných denních teplot měřených v roce 1951 na meteorologické stanici *Waldassen* při počtu 32 *bázových funkcích řádu* 5 *z B-splinové báze*.

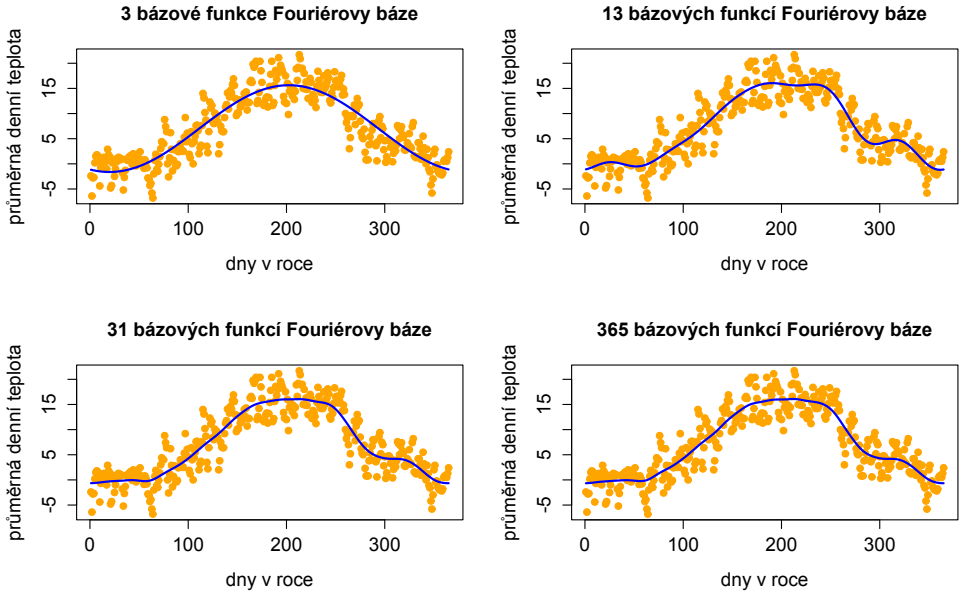


Obrázek 2.14: Hodnoty kritéria *obecné cross-validace*(2.33) při různém nastavení počtu *bázových funkcí řádu 5 z B-splinové báze*

Pro lepší orientaci v grafu 2.13 je zlogaritmována osa x, tj. osa zlogaritmovaných hodnot parametru λ. Nejvhodnějším nastavením parametru λ dle kritéria *obecné cross-validace* je hodnota 100, minimalizující kritérium 2.33 na 7.612, viz 2.11. Takové nastavení váhy *penalizačního členu* není pro náš účel příliš žádoucí.

Tento postup výběru *parametru vyhlazení* zopakujeme pro odlišný počet *bázových funkcí B-splinové báze* a výsledky zaneseme do grafu 2.14.

Na obrázku 2.14 pozorujeme nízkou hodnotu kritéria *obecné cross-validace* pro počty *bázových funkcí* 13 až 32. Odhadujeme-li průběh funkce s využitím *penalizace nehladkosti* v kombinaci s *metodou nejmenších čtverců* namísto jen klasické *metody nejmenších čtverců*, nastavujeme vyhlazení odhadu funkce primárně vyhlazovacím parametrem λ. Počet *bázových funkcí* již nehraje, tak významnou roli.



Obrázek 2.15: Odhady funkcí průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* při různém nastavení počtu bázových funkcí *Fourierovy báze*. Pro odhad průběhu funkcí byla použita *standardní metoda nejmenších čtverců* s *penalizací nehladkosti*.

Obrázek 2.15 je podobný grafu 2.9 s tím rozdílem, že při odhadu průběhu funkcí byla použita *metoda nejmenších čtverců* s *penalizací za nehladkost*. Je vidět, že při vhodném nastavení vyhlazovacího parametru dostaneme přijatelný odhad

průběhu funkce i přes různé nastavení počtu *bázových funkcí*.

Zaujalo nás, že při odhadu průběhu funkce vyjádřené lineární kombinací 32 *bázových funkcí* z *Fourierovy báze* (2.15), nám *zobecněná cross-validace* doporučila velmi malou a nevhodnou hodnotu *vyhlazovacího parametru*, která by vedla k příliš málo zjednodušujícímu odhadu funkce.

Museli jsme parametr λ nastavit manuálně tak, aby byl odhad dostatečně hladký, tudíž interpretovatelný. *Zobecněná cross-validace* je náš rádce, ale není všemocná. Vždy záleží na konkrétní situaci a účelu využití vyhlazených dat.

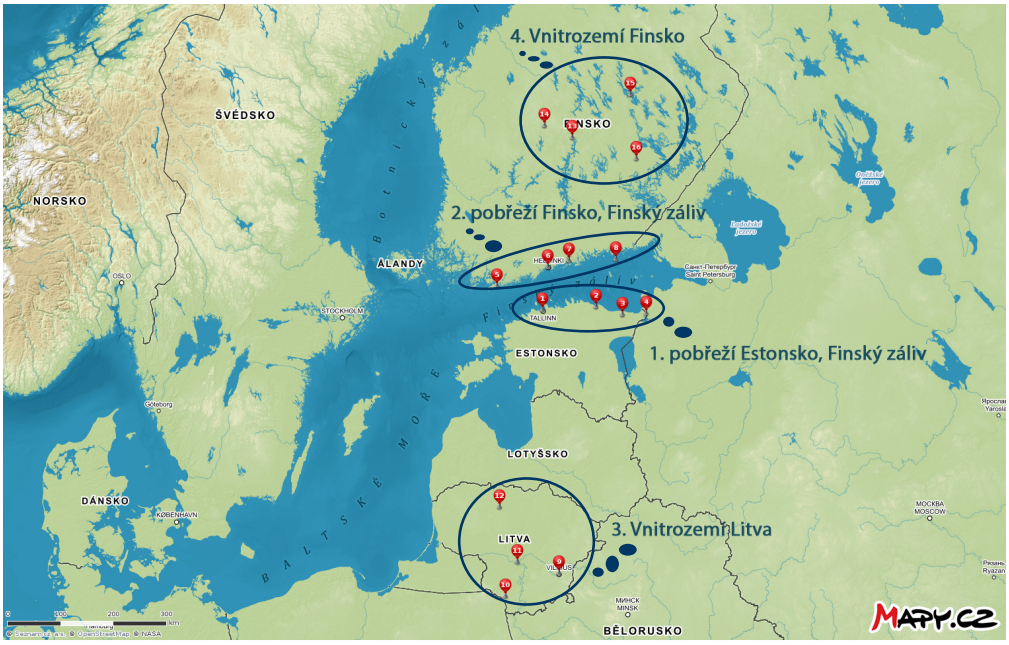
Kapitola 3

Funkcionální ANOVA a časové řady průměrných denních teplot

Cílem kapitoly je představit čtenáři model *funkcionální analýzy rozptylu* a ukázat její praktické využití na vybraných časových řadách průměrných denních teplot z projektu **ECA&D**. Teorie k této kapitole byla čerpána z [21][22][23].

3.1. Náš záměr

Podívejme se nejprve na mapu 3.1 zobrazující čtyři skupiny meteorologických stanic rozdělených do skupin dle jejich polohy, ze kterých pochází záznamy časových řad průměrných denních teplot.



Obrázek 3.1: Vizualizace skupin *baltských* a *finských* meteorologických stanic podle jejich polohy. Podklad pro obrázek pochází z webové stránky *mapy.cz*.

Na obrázku 3.1 vidíme *pobaltské* a *finské* meteorologické stanice, rozdělené podle blízkosti jejich vzájemné polohy do čtyř skupin se stejným počtem stanic:

- 1. skupina: pobřeží *Estonsko*, *Finský záliv*
- 2. skupina: pobřeží *Finsko*, *Finský záliv*
- 3. skupina: vnitrozemí *Litva*
- 4. skupina: vnitrozemí *Finsko*.

Region *Pobaltí* a *Finska* jsme zvolili záměrně pro naši sympatii k této oblasti. V každé skupině jsou čtyři meteorologické stanice, z nichž máme k dispozici kvalitní záznamy, tj. časové řady průměrných denních teplot bez chybějících pozorování ve formátu 1.8.

V naší analýze nepracujeme přímo se skupinami meteorologických stanic, nýbrž se stejně uspořádanými skupinami časových řad průměrných denních teplot měřených na těchto stanicích. cc

ID číslo časové řady	**Jméno Stanice**	**Země**	**Způsob výpočtu**	**Začátek měření**	**Konec měření**	**Skupina**
175002	Johvi	Estonsko	TG7	1959:01:01	2015:12:31	1.
175018	Malla	Estonsko	TG7	1945:02:01	2015:12:31	1.
175054	Tallinn	Estonsko	TG7	1883:01:01	2015:12:31	1.
175030	Narva	Estonsko	TG7	1920:04:01	2015:12:31	1.
142219	Helsinki	Finsko	TG15	2003:07:17	2017:11:30	2.
142023	Hanko	Finsko	TG15	1963:01:01	2017:11:30	2.
142304	Porvoo	Finsko	TG15	1984:01:13	2017:11:30	2.
142399	Kotka	Finsko	TG15	1986:04:12	2017:11:30	2.
100629	Vilnius	Litva	TG1	1900:01:01	2009:12:31	3.
100621	Kaunas	Litva	TG1	1901:01:01	2009:12:31	3.
105278	Lazdijai	Litva	TG1	1936:01:01	2009:12:31	3.
105282	Siauliai	Litva	TG1	1937:01:01	2009:12:31	3.
100088	Jyvaskyla	Finsko	TG1	1951:01:01	2017:11:30	4.
144780	Kuipio	Finsko	TG15	2005:06:24	2017:11:30	4.
144550	Multia	Finsko	TG15	2008:09:25	2017:11:30	4.
144040	Juva	Finsko	TG15	1992:10:16	2017:11:30	4.

Tabulka 3.1: Tabulka časových řad průměrných denních teplot zahrnutých do *funkcionální analýzy rozptylu*. Data pocházejí z datové sady projektu *ECA&D*.

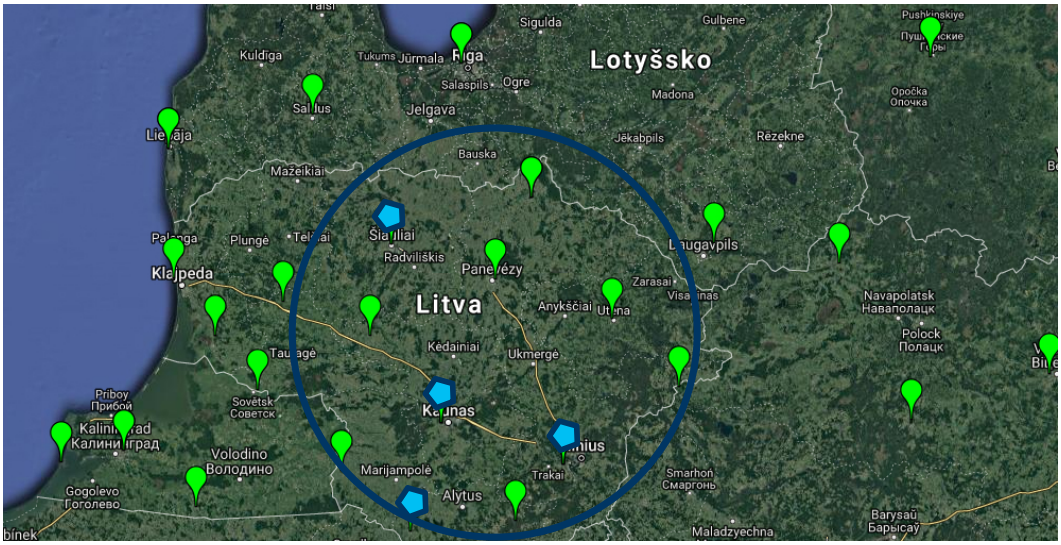
Nebudeme pracovat s celými časovými řadami. Z každé časové řady uvedené v tabulce 3.1 vybereme pozorování pro rok 2009. Napříč všemi skupinami máme nyní

16 ročních časových řad s 365 napozorovanými hodnotami průměrných denních teplot. K otázce proč zrovna rok 2009 se ještě vrátíme.

Naším záměrem je zjistit, zda a případně porovnat, jak se od sebe liší takto uspořádané skupiny časových řad průměrných denních teplot pro rok 2009. Využijeme získané poznatky z kapitoly 2 a model *Funkcionální analýzy rozptylu*.

V tabulce časových řad 3.1 si ještě všimněme rozdílného *způsobu výpočtu* průměrných denních teplot napříč časovými řadami. O způsobu *TG1* v podstatě nic nevíme, tento způsob reprezentuje průměr denních teplot z neznámého intervalu. V našem případě to nejspíše znamená, že *Litevci* nedodávají projektu *ECA&D* informace o *způsobu výpočtu* průměrných denních teplot. Druhý způsob *TG7* představuje výpočet průměrné denní teploty jako průměr tří a více pozorování denních teplot. Nevíme ani přibližně, v jaké hodině byly záznamy denních teplot pozorovány. *TG15* je průměr osmi záznamů denních teplot. Víme jen, že pro každý den bylo k dispozici 8 hodnot použitých pro výpočet průměru.

Snažili jsme se alespoň o to, aby v rámci skupiny byly časové řady se stejným *způsobem měření*, což v kombinaci se špatnou kvalitou časových řad, tj. velkým množstvím chybějících hodnot v časových řadách, nebyl jednoduchý úkol. To je také odpověď na případnou otázku, proč máme v každé skupině čtyři časové řady.



Obrázek 3.2: Vizualizace *pobaltských* meteorologických stanic. V modrém kruhu jsou vyobrazeny potencionální kandidáti pro skupinu č.3 *Vnitrozemí Litva*.

Podívejme se na obrázek 3.2 zobrazující meteorologické stanice *Pobaltí* a *Finska*, ze kterých máme k dispozici data. Zaměřme se na modrý kruh představující vnitrozemí *Litvy*, potažmo *Lotyšska*, kde máme potencionální kandidáty časových řad pro skupinu číslo 3, tj. vnitrozemí *Litva*. Z 12 časových řad byly přípustné vzhledem k velkému množství chybějících dat jen čtyři, zvýrazněné modrou značkou.

Po naší zkušenosti můžeme konstatovat, že kvalita dat z oblasti *Pobaltí*, zejména z *Litvy* a *Lotyšska*, v datové sadě *ECA&D* není valná. Často se nám stávalo, že celková doba měření, tedy *začátek* a *konec měření* uvedené v seznamu časových řad (1.4), neodpovídala realitě. Některé časové řady měly poskytnout informaci o průměrných denních teplotách například z roku 2011, namísto napočítaných hodnot jsme mohli v hojném počtu případů pracovat leda tak s chybějícími hodnotami.

Snažili jsme se pro naši analýzu najít vhodný rok z nedaleké minulosti, ve kterém nemáme napříč všemi časovými řadami žádné chybějící pozorování. Tyto kritéria splnil již zmiňovaný rok 2009.

Dalším příkladem využití *funkcionální analýzy rozptylu* je analýza chování jedné časové řady napříč desetiletími. V našem případě pracujeme s časovou řadou průměrných denních teplot měřených na meteorologické stanici *Tallinn*. Bližší informace o této časové řadě najdeme v tabulce 3.1.

Náš pustup předpřípravy *tallinnské* řady pro účel analýzy je následující: Z celé časové řady vybereme pozorování průměrných denních teplot za období 1976 až 2015. Tento čtyřicetiletý časový úsek rozdělíme na čtyři desetiletí: $1976 - 1985$, $1986 - 1995$, $1996 - 2005$ a $2006 - 2015$. V rámci každé skupiny, tj. desetiletí, máme k dispozici 10 ročních časových řad průměrných denních teplot měřených na stejné meteorologické stanici, tedy v *Tallinnu*.

Opět nás zajímá, zda a jak se od sebe liší takto definované skupiny ročních časových řad.

Než člověk začne používat statistický aparát, jenž mu ukáže příběh skrývající se za čísly, měl by mít nejprve vlastní představu o analyzované oblasti zájmu.

Vraťme se k prvnímu příkladu s časovými řadami z *Pobaltí* a *Finska*, viz obrázek 2.15. Intuitivně očekáváme, že pozorování průměrných denních teplot ve skupině *vnitrozemí Litva* budou nabývat vyšších hodnot než ve zbylých třech skupinách. Naopak rozdíly mezi skupinami časových řad *pobřeží Estonsko* a *pobřeží Finsko* by neměly být výrazné vzhledem k jejich vzájemné poloze. Ve skupině časových řad *vnitrozemí Finsko* předpokládáme vůbec nejnižší hodnoty průměrných denních teplot ve srovnání s ostatnímy skupinami. Pracujeme s následující úvahou: čím více půjdeme na sever, tím chladnější počasí lze očekávat, tj. naměříme nižší hodnoty průměrných denních teplot než v jižnějších polohách. Mezi skupinami časových řad by měly být patrné rozdíly. Největší rozdíl bychom podle naší úvahy měli pozorovat mezi skupinami *vnitrozemí Litva* a *vnitrozemí Finsko*.

U časové řady průměrných denních teplot měřených v *Tallinnu* také očekáváme rozdíly mezi skupinami ročních časových řad zařazených do skupin podle příslušnosti roku měření časové řady k jednotlivým desetiletím. Největší rozdíl bychom mohli vzhledem k oteplování planety pozorovat mezi prvním $(1976 - 1985)$ a posledním $(2006 - 2015)$ desetiletím.

Nyní představíme model stojící za *funkcionální analýzou rozptylu*.

3.2. Model funkcionální analýzy rozptylu

V předchozí sekci jsme čtenáři představili příklady využití *funkcionální analýzy rozptylu* na datové sadě projektu *ECA&D*. Pro názornost při vysvětlování a popisu modelu *funkcionální analýzy rozptylu* zkráceně *FANOVY*, využijeme prvního příkladu, tj. čtyř skupin *baltských* a *finských* časových řad průměrných denních teplot, viz tabulka 3.1. Naším cílem je tyto čtyři skupiny časových řad porovnat a prozkoumat, zda se od sebe výrazně liší.

První krok, který musíme před samotnou analýzou udělat, je odhadnout z dat jednotlivých časových řad hladké průběhy funkcí, tj. vyhladit původní pozorování. To znamená, že časovou řadu průměrných denních teplot zastupuje odhad funkce průměrné denní teploty. Pro přesnost v teorii *funkcionální analýzy roz-*

ptylu nepracujeme s napozorovanými hodnotamy časových řad, ale nahrazujeme je funkčními hodnotami příslušné funkce této časové řady. Při odhadu průběhu funkce vycházíme z poznatků kapitoly 2.

Stejně jako u jiných statistických modelů a metod i pro *FANOVU* zavádíme nulovou hypotézu a to ve tvaru:

$$u_1(t) = u_2(t) = u_3(t) = u_4(t) \quad \forall t \in [0, 365],$$

kde $u_g(t)$ představuje funkci střední hodnoty pro skupinu g. Za platnosti *nulové hypotézy* nabývají funkce skupinových středních hodnot stejných funkčních hodnot v každém časovém okamžiku t, kde $t \in [0, 365]$. To znamená, že průběhy funkcí $u_g(t)$, $g = 1, \ldots, 4$ jsou identické.

Uvažujeme čtyři skupiny funkcí průměrných denních teplot. V každé skupině jsou čtyři funkce. Pojďme si představit teoretický model *funkcionální analýzy rozptylu* pro vyjádření m-té funkce ve skupině g:

$$Temp_{mg}(t) = u(t) + \alpha_g(t) + \epsilon_{mg}(t), \tag{3.1}$$

kde m je indikátor funkce uvnitř skupiny a g specifikuje skupinu funkcí. Například $Temp_{21}(t)$ je v našem příkladu označení funkce průměrné denní teploty meteorologické stanice *Malla*, patřící do první skupiny *pobřeží Estonsko, Finský záliv*. Podle modelu (3.1) vyjadřujeme funkční hodnotu této funkce v čase t součtem funkčních hodnot funkcí celkové střední hodnoty $u(t)$, skupinového efektu $\alpha_g(t)$ a náhodné odchylky $\epsilon_{mg}(t)$ v časovém okamžiku t. Funkci $u(t)$ chápeme jako funkci střední hodnoty v celé populaci (stanic v oblasti *Pobaltí* a *Finska*) bez rozlišení skupiny. Funkční hodnoty funkce efektu $\alpha_g(t)$ vytvářejí rozdíl mezi funkčními hodnotami funkcí celkové střední hodnoty $u(t)$ a skupinové střední hodnoty $u_g(t)$. Funkce střední hodnoty pro skupinu g je ve tvaru:

$$u_g(t) = u(t) + \alpha_g(t). \tag{3.2}$$

Za předpokladu platnosti nulové hypotézy nabývají funkce efektů nulových funkčních hodnot ve všech časových okamžicích $t \in [0, 365]$. Jinými slovy řečeno, průběhy funkcí skupinových středních hodnot $u_g(t)$ se od sebe neliší a zároveň se neliší od průběhu funkce celkové střední hodnoty $u(t)$.

Funkce $u(t)$, $\alpha_g(t)$ chápejme v modelu 3.1, jako regresní funkce. Pro jednoznačné určení funkcí efektů $\alpha_g(t)$ požadujeme splnění podmínky:

$$\sum_{g=1}^{4} \alpha_g(t) = 0 \quad \forall t \in [0, 365]. \tag{3.3}$$

Dle této restrikce platí, že součet funkčních hodnot funkcí efektů α_g je v každém časovém okamžiku t roven nule.

Funkci $\epsilon_{mg}(t)$ v modelu 3.1 považujeme za funkci náhodné odchylky m-té meteorologické stanice ve skupině funkcí g.

Definujme *designovou matici* $\mathbf{Z}$ s 16 řádky a pěti sloupci. Řádek této matice z_{mg} se vztahuje k funkci $Temp_{mg}(t)$ z modelu 3.1. Označení z_{mg} odpovídá řádku matice pro m-tou funkci ve skupině g. Tento řádek má jedničku na první a $(g+1)$-ní pozici. Na ostatních pozicích má nulu. Například řádek z_{21} souvisí s funkcí $Temp_{21}(t)$ a má na první a druhé pozici jedničku. Zbylé položky tohoto řádku jsou nulové. Složku na pozici j v řádku z_{mg} označme $z_{(mg)j}$. Celou *designovou matici* ukážeme na příkladu v další sekci.

Přeznačme regresní funkce z modelu 3.1 do formy:

$$\beta_1(t) = u(t), \beta_2(t) = \alpha_1(t), \beta_3(t) = \alpha_2(t), \beta_4(t) = \alpha_3(t), \beta_5(t) = \alpha_4(t)$$
$$\boldsymbol{\beta}(\mathrm{t}) = (u(t), \alpha_1(t), \alpha_2(t), \alpha_3(t), \alpha_4(t)).$$

Nyní můžeme model 3.1 přeznačit do podoby:

$$Temp_{mg}(t) = \beta_1(t) + \beta_{g+1}(t) + \epsilon_{mg}(t). \tag{3.4}$$

Maticový zápis modelu 3.4:

$$\mathbf{Temp}(\mathrm{t}) = \mathbf{Z}\boldsymbol{\beta}(\mathrm{t}) + \boldsymbol{\epsilon}(\mathrm{t}) \quad \forall t \in [0, 365], \tag{3.5}$$

kde $\mathbf{Temp}(\mathrm{t})$ je vektor funkčních hodnot funkcí průměrné denní teploty v čase t, v našem případě má délku 16. $\boldsymbol{\beta}(\mathrm{t})$ považujeme za vektor funkčních hodnot regresních funkcí $\beta_j(t)$, $j = 1, \ldots, 5$ v čase t. Vektor $\boldsymbol{\epsilon}(\mathrm{t})$ je vektor funkčních hodnot funkcí náhodných odchylek v čase t délky 16.

Pro odhad vektoru regresních funkcí $\boldsymbol{\beta}(\mathrm{t})$ použijeme modifikovanou *metodu nejmenších čtverců* aplikovatelnou na funkce. Hledáme odhady regresních funkcí $\beta_j(t)$ minimalizující kritérium:

$$LMSSE(\boldsymbol{\beta}(\mathrm{t})) = \sum_{g=1}^{4}\sum_{m=1}^{4}\int_0^{365}[Temp_{mg}(t) - \sum_{j=1}^{5} z_{(mg)j}\beta_j(t)]^2\,\mathrm{d}t \tag{3.6}$$

za podmínky $\sum_{j=2}^{5}\beta_j(t) = 0 \quad \forall t \in [0, 365]$. Podotkněme, že kritérium je ve formě určitého integrálu, integrujeme přes časový interval [0 365]. Ještě kritérium zapíšme v maticové podobě:

$$LMSSE(\boldsymbol{\beta}(\mathrm{t})) = \int_0^{365}[\mathbf{Temp}(\mathrm{t}) - \mathbf{Z}\boldsymbol{\beta}(\mathrm{t})]'[\mathbf{Temp}(\mathrm{t}) - \mathbf{Z}\boldsymbol{\beta}(\mathrm{t})]\,\mathrm{d}t. \tag{3.7}$$

Minimalizací tohoto kritéria získáme odhady regresních funkcí $\hat{u}(t)$, $\hat{\alpha_g}(t)$.

V následující sekci čtenáři představíme první přístup vedoucí k odhadu regresních funkcí a minimalizaci kritéria (3.7).

3.3. Bodová minimalizace

Nemáme-li žádnou podmínku na tvar regresních funkcí $\beta_j(t)$,$j = 1, \ldots, 5$, můžeme minimalizovat hodnotu kritéria 3.6 minimalizací $||\mathbf{Temp}(\mathrm{t}) - \mathbf{Z}\boldsymbol{\beta}(\mathrm{t})||^2$ euklidovské normy v jednotlivých časových okamžicích t. To znamená, že nebereme v potaz celý časový interval $[0, 365]$, ale jen diskrétní časové okamžiky, ve kterých byly zaznamenány pozorování časových řad. Je důležité, aby tato síť diskrétních časových okamžiků byla pro všechny analyzované časové řady shodná.

Postup je následující: vezmeme odhady funkčních hodnot funkcí napříč všemi skupinami pro první časový okamžik 0,5, tj. $\mathbf{Temp}(0, 5)$, a použitím modelu *vícenásobné lineární regrese* odhadneme funkční hodnoty regresních funkcí $u(0, 5)$, $\alpha_g(0, 5)$ pro $g = 1, \ldots, 4$. Odhad funkční hodnoty $\hat{u}(0, 5)$ představuje v našem příkladu výběrový průměr 16 odhadů funkčních hodnot uložených ve vektoru $\mathbf{Temp}(0, 5)$. Odhad funkční hodnoty $\hat{\alpha_1}(0, 5)$ je rozdílem výběrového průměru vyhlazených hodnot z první skupiny a celkového průměru $\hat{u}(0, 5)$ v čase $0, 5$. Tento postup zopakujeme v našem případě 365 krát, tedy pro celou síť diskrétních

časových okamžiků, a získáme odhady funkčních hodnot jednotlivých regresních funkcí v časech $t = 0,5; 1,5; 2,5; \ldots; 364,5$ utvářejících síť časových okamžiků. Interpolací takto odhadnutých funkčních hodnot získáme odhad průběhu regresních funkcí na intervalu [0,365]. Velkou výhodou je, že při odhadu funkčních hodnot regresních funkcí je v regresním modelu využita stejná designová matice pro každý diskrétní časový okamžik z časové sítě.

Představme pro příklad *pobaltských* a *finských* časových řad průměrných denních teplot *designovou matici* $\mathbf{Z}$, o níž již byla řeč:

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Matice $\mathbf{Z}$ má 17 řádků a 5 sloupců. Poslední řádek má své opodstatnění. Je do matice přidán proto, aby byla zajištěna podmínka jednoznačnosti určení funkcí efektů. Podívejme se na tuto záležitost podrobněji. Pro časové okamžiky $t = 0,5; 1,5; 2,5; \ldots; 364,5$ hledáme odhady funkčních hodnot regresních funkcí minimalizujících hodnotu kritéria:

$$\sum_{i=1}^{17}(Temp_i(t) - z_i'\boldsymbol{\beta}(\mathrm{t}))^2 = \tag{3.8}$$

$$\sum_{i=1}^{4}(Temp_i(t) - (\beta_1(t) + \beta_2(t)))^2 + \sum_{i=5}^{8}(Temp_i(t) - (\beta_1(t) + \beta_3(t)))^2 +$$

$$\sum_{i=9}^{12}(Temp_i(t) - (\beta_1(t) + \beta_4(t)))^2] + \sum_{i=13}^{16}(Temp_i(t) - (\beta_1(t) + \beta_5(t)))^2 + (\beta_2(t) + \beta_3(t) + \beta_4(t) + \beta_5(t))^2. \quad (3.9)$$

V rovnosti (3.8) jsme akorát rozepsali minimalizační kritérium metody nejmenších čtverců. Tento součet náhodných odchylek v čase t bez posledního členu 3.9 představuje rozepsané minimalizační kritérium *metody nejmenších čtverců* bez využití 17. řádku matice **Z**. Všimněme si, že poslední člen (3.9) je jen součtem funkčních hodnot efektů v časovém okamžiku t. Posledním řádkem, tj. 17. řádkem matice **Z**, vyjadřujeme levou stranu podmínky $\sum_{g=1}^{4} \alpha_g(t) = 0$. Nulu vyskytující se na pravé straně zapíšeme na sedmnáctou pozici vektoru **Temp**(t) pro časové okamžiky $t = 0,5; 1,5; 2,5; \ldots; 364,5$. Označme:

$$\beta_0^*(t) = \beta_1(t) + \beta_2(t)$$
$$\beta_2^*(t) = (\beta_1(t) + \beta_3(t)) - (\beta_1(t) + \beta_2(t)) = \beta_3(t) - \beta_2(t)$$
$$\beta_3^*(t) = \beta_4(t) - \beta_2(t)$$
$$\beta_4^*(t) = \beta_5(t) - \beta_2(t)$$

kde $\beta_0^*(t)$ je funkční hodnotou funkce střední hodnoty pro první skupinu, $\beta_2^*(t)$ vyjadřuje rozdíl mezi funkčními hodnotamy funkcí střední hodnoty druhé skupiny a střední hodnoty první skupiny v čase t. Stejný význam vzhledem k $\beta_0^*(t)$ mají i funkční hodnoty funkcí $\beta_3^*(t)$ a $\beta_4^*(t)$.

Po tomhle přeznačení můžeme najít jednoznačně určenou čtveřici $\beta_0^*(t)$, $\beta_2^*(t)$, $\beta_3^*(t)$ a $\beta_4^*(t)$, která minimalizuje:

$$\sum_{i=1}^{4}(Temp_i(t) - (\beta_0^*(t))^2 + \sum_{i=5}^{8}(Temp_i(t) - (\beta_0^*(t) + \beta_2^*(t)))^2 + \sum_{i=9}^{12}(Temp_i(t) - (\beta_0^*(t) + \beta_3^*(t)))^2] + \sum_{i=13}^{16}(Temp_i(t) - (\beta_0^*(t) + \beta_4^*(t)))^2$$

Této čtveřici *regresních funkcí* odpovídá nekonečmě mnoho pětic $\beta_1(t)$,...,$\beta_5(t)$. Zdůrazněme, že součet výše uvedených členů, odpovídá součtu prvních šesnácti členů 3.8 bez uvážení sedmnáctého členu (3.9). My vybereme takovou pětici,

která minimalizuje poslední člen součtu 3.8, tj. právě sedmnáctý člen $(\beta_2(t) + \beta_3(t) + \beta_4(t) + \beta_5(t))^2$. Ten je vždy nezáporný, nejmenší hodnota kterou může mít je nula, a to právě když $\beta_2(t) + \beta_3(t) + \beta_4(t) + \beta_5(t) = 0$. Tím je splněna podmínka $\sum_{g=1}^{4} \alpha_g(t) = 0 \quad \forall t \in [0, 365]$.

3.4. Popis skriptu pro odhad regresních funkcí pomocí bodové minimalizace

Postup pro *bodovou minimalizaci* a odhad průběhu funkcí $u(t)$, $\alpha_g(t)$, kde $g = 1, 2, 3, 4$, ukážeme ve výpočetním prostředí statistického softwaru **R**. Pomohou nám funkce z balíčku `fda`.

Nejprve se podíváme na prvních 8 hodnot sítě diskrétních časových okamžiků, uložených ve vektoru délky 365 s názvem `dny`.

```
[1]   0.5   1.5   2.5   3.5   4.5   5.5   6.5   7.5
```

Místo klasického indexování jednotlivých dnů 1,2,3...,365, nám přišlo vhodnější použít toto časové vyjádření a to s ohledem k výpočtu průměrných denních teplot z pozorování denních teplot měřených v průběhu celého dne. Časovému okamžiku 0, 5 odpovídá 16 vyhlazených pozorování průměrných denních teplot naměřených na *pobatských* a *finských* meteorologických stanicích 1. 1. 2009.

Nyní ukážeme, jak vyhladit pozorování jedné časové řady. Druhý a elegantnější způsob použitelný na větší množství časových řad představíme v další sekci.

```
basis = create.fourier.basis(c(0,365), 13)
basismat=eval.basis(dny, basis)
```

Příkazem `create.fourier.basis` vygenerujeme na intervalu [0,365] 13 *bázových funkcí fourierovy báze*. Funkce `eval.basis` nám napočítá *designovou matici* funkčních hodnot bázových funkcí s 365 řádky a 13 sloupci. Do funkce je potřeba dosadit síť časových okamžiků, ve kterých chceme zjistit funkční hodnoty pro 13 bázových funkcí z *fourierovy báze*, a také onu bázi.

```
yfit=basismat %*% lsfit(basismat,y,intercept=FALSE)$coef
```

`Lsfit` je funkcí *metody nejmenších čtverců*. Používáme ji pro odhad koeficientů bázového rozvoje. Do argumentu funkce `lsfit` patří *designová matice* bázových funkcí, původní pozorování časové řady průměrných denních teplot, tedy 365 nevyhlazených hodnot, a informace, zda budeme v modelu uvažovat absolutní člen. V modelu bázového rozvoje je již absolutní člen zahrnut v podobě konstanty c_0, nastavením parametru `intercept=TRUE` by byl v modelu uvažován další absolutní člen. Přidáním koncovky `$coef` za funkci, posíláme `Rku` žádost o vypsání odhadu koeficientů bázového rozvoje funkce průměrné denní teploty. Nyní program vezme první řádek designové matice, ve kterém jsou uložené funkční hodnoty bázových funkcí v časovém okamžiku $0,5$, a pronásobí je s odhadem vektoru koeficientů bázového rozvoje. Tento krok opakuje `Rko` pro každý řádek designové matice `basismat`. Výsledkem je vektor vyhlazených hodnot `yfit`.

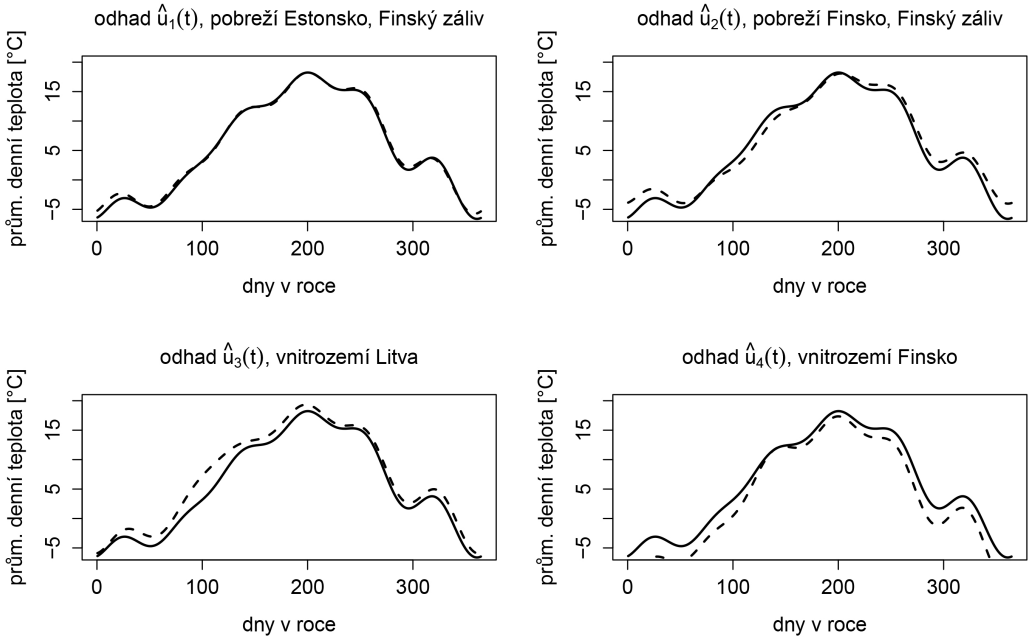
Celý postup zopakujeme pro zbývající časové řady. Pro usnadnění využijeme například `for` cyklus a vyhlazené hodnoty postupně uložíme do sloupců `matice`, která má 365 řádků a 16 sloupců. Do této `matice` přidáme sedmnáctý sloupec plný nul délky 365, tj. v každém řádku na sedmnácté pozici je nula korespondující se sedmnáctým řádkem *designové matice*. Tím máme zaručeno splnění podmínky 3.3. Vše potřebné pro odhad funkčních hodnot regresních funkcí v časových okamžicích reprezentovaných vektorem `dny` je připraveno. Pro odhad funkčních hodnot v časovém okamžiku $t = 0,5$ použijeme příkaz:

```
lm(formula = matice[1,]~zmat-1)
```

Funkce `lm` je standadní funkcí v prostředí `R` pro práci s lineárními modely. Matice `zmat` je již představenou designovou maticí $\mathbf{Z}$ s rozměry 17×5. Doplněním -1 do argument funkce za `zmat` nařídíme `Rku` neuvažovat absolutní člen, který je v našem modelu obsažen. Přidáním koncovky `$coef` za funkci `lm` necháme vypsat odhady funkčních hodnot regresních funkcí v čase 0,5. Tuto operaci s pomocí `for` cyklu zopakujeme pro zbylé časové okamžiky vektoru `dny`, tj. 364 krát a odhady funkčních hodnot regresních funkcí uložíme do matice `maticeodhadu` s 365 řádky a 5 sloupci. Pro vizualizaci odhadu průběhu funkce celkové střední hodnoty stačí vzít a vykreslit první sloupec. Druhý sloupec nám dává informaci

o odhadu průběhu funkce efektu první skupiny, pro zbylé sloupce platí analogie.

Na obrázku 3.3 vidíme vizualizaci zmíněného prvního sloupce matice `matice-odhadu`. Samotný odhad průběhu celkové střední hodnoty není až tak výjimečný, proto jsme k němu zobrazili jednotlivé odhady průběhu funkcí skupinových středních hodnot. Nicméně nás zaujal teplotní výkyv, tj. pokles funkčních hodnot celkového průměru vyhlazených pozorování průměrných denních teplot v časovém intervalu [30, 50]. Obdobný výkyv celkového průměru průměrných denních teplot, v tomto případě růstu funkčních hodnot celkového průměru, pozorujeme v časových okamžicích [295, 315]. Jde možná o střednědobé odchylky teploty v daném roce oproti dlouhodobému průměru. Možná však jde i o jev typický pro dané období v roce.



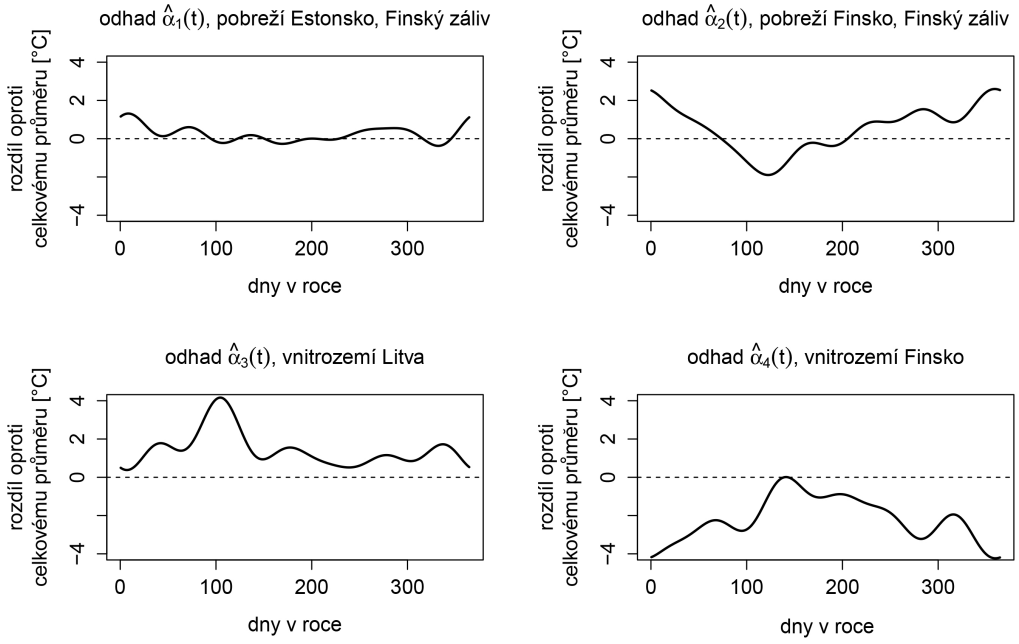
Obrázek 3.3: Odhady průběhu funkcí skupinových středních hodnot $u_g(t)$ ve srovnání s odhadem průběhu funkce celkové střední hodnoty, tj. funkce celkového průměru na intervalu [0, 365] pro rok 2009. Tlustou čarou je značen průběh celkového průměru a čárkovanou průběh funkce skupinových průměrů.

Z obrázku 3.3 můžeme na první pohled vyčíst, že průběh funkce průměru skupiny *pobřeží Estonsko* $\hat{u}_1(t)$ se na intervalu $[0, 365]$ výrazně neliší od průběhu funkce celkového průměru $\hat{u}(t)$. Totéž můžeme tvrdit také pro skupinu *pobřeží*

Finsko $\hat{u}_2(t)$. Průměrné chování průměrných denních teplot bylo na pobřeží *Finska* a *Estonska* velmi podobné, což není vzhledem k blízké poloze meteorologických stanic v rámci těchto dvou skupin překvapující informace.

Funkce skupinového průměru *vnitrozemí Litva* $\hat{u}_3(t)$ nabývá vyšších funkčních hodnot na celém časovém intervalu $[0, 365]$, zejména však v jarním období. V této oblasti napočítáme vyšší průměrné denní teploty, než v ostatních skupinách. Patrný je rozdíl mezi funkcemi skupinového průměru *vnitrozemí Litva* $\hat{u}_3(t)$ a *vnitrozemí Finsko* $\hat{u}_4(t)$, kde po celý rok pozorujeme podprůměrné záznamy průměrných denních teplot.

Ještě se podívejme podrobněji na rozdíly funkčních hodnot jednotlivých skupinových průměrů a celkového průměru.



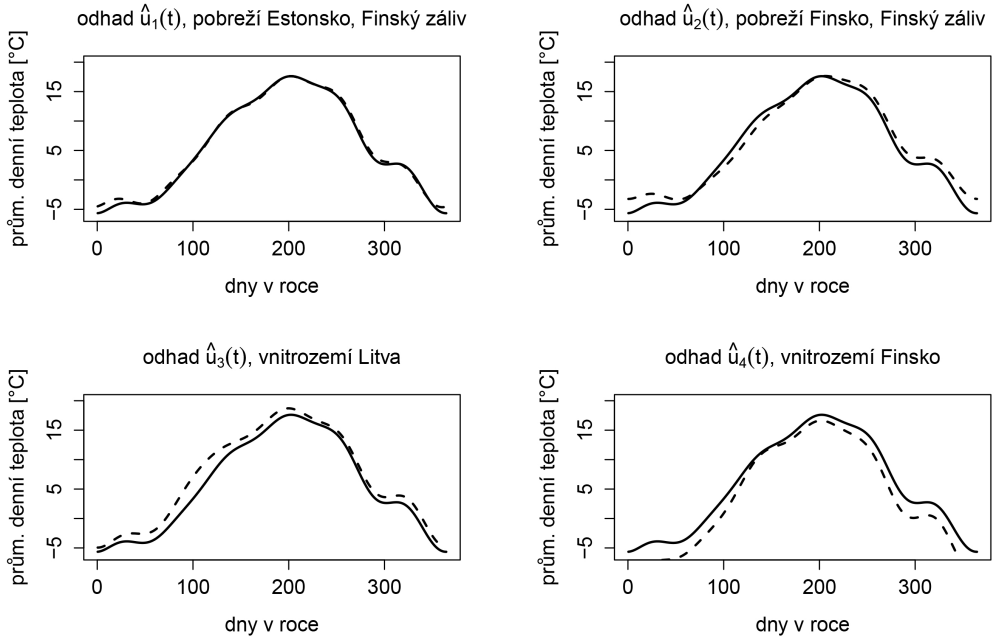
Obrázek 3.4: Odhady průběhu funkcí efektů $\alpha_g(t)$ na intervalu $[0, 365]$ pro rok 2009. Značení je následující: *pobřeží Estonsko* $\hat{\alpha}_1(t)$, *pobřeží Finsko* $\hat{\alpha}_2(t)$, *vnitrozemí Litva* $\hat{\alpha}_3(t)$, *vnitrozemí Finsko* $\hat{\alpha}_4(t)$.

Na obrázku 3.4 vidíme, že odhad funkce efektů $\hat{\alpha}_2(t)$ *pobřeží Finsko* je variabilnější než $\hat{\alpha}_1(t)$ *pobřeží Estonsko*. Z toho lze usoudit, že i když se skupinové funkce průměrů $\hat{u}_1(t)$ a $\hat{u}_2(t)$ chovají podobně, tak funkce $\hat{u}_1(t)$ je lehce stabilnější než finský protějšek $\hat{u}_2(t)$.

Ještě si všimněme, že největší rozdíly mezi funkčními hodnotami $\hat{u}_2(t)$, $\hat{u}_4(t)$ a $\hat{u}(t)$ pozorujeme v zimních měsících. Ve *vnitrozemí Finska* napočítáme v zimních měsících v průměru o 6° C nižší průměrnou denní teplotu než na *pobřeží Finska*.

Komentáře k obrázkum 3.3 a 3.4 jsou v souladu s naší úvahou z úvodu této kapitoly: čím více půjdeme na sever, tím nižší teploty lze očekávat.

Je potřeba zmínit, že při vyhlazení původních pozorování časových řad průměrných denních teplot byla použita metoda nejmenších čtverců bez *penalizace za nehladkost*. Podívejme se, jak bude vypadat obdoba grafu 3.3, zopakujeme-li celý proces a využijeme-li při vyhlazení dat právě *penalizace za nehladkost* (2.25) s nastavením penalizačního parametru λ na 10 000.

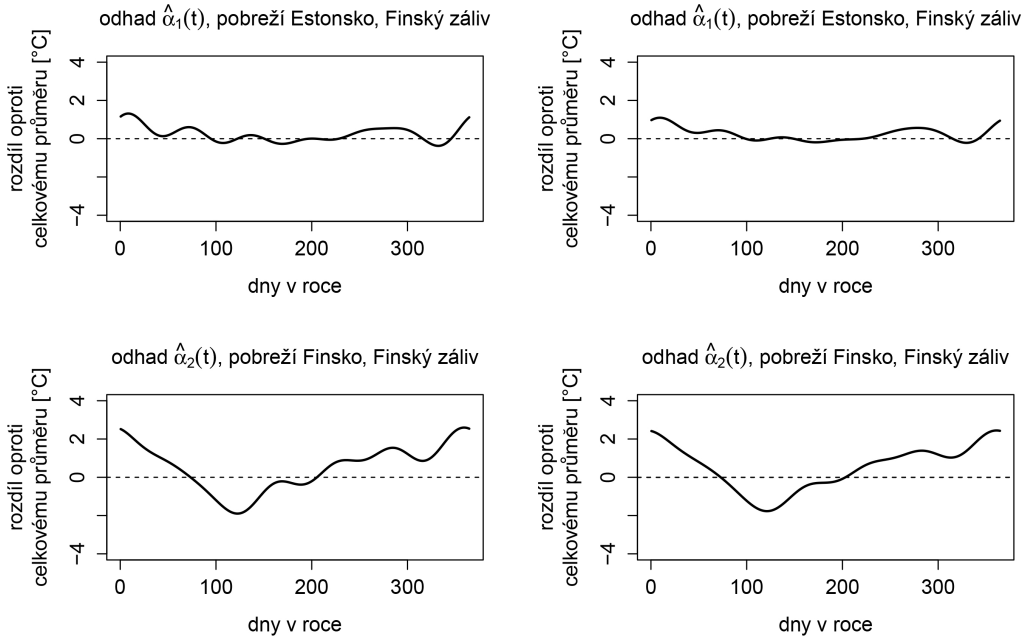


Obrázek 3.5: Odhady průběhu funkcí skupinových středních hodnot $u_g(t)$ ve srovnání s odhadem průběhu funkce celkové střední hodnoty, tj. funkce celkového průměru na intervalu [0, 365] pro rok 2009. Tlustou čarou je značen průběh celkového průměru a čárkovanou průběh funkce skupinových průměrů. Jedná se obdobu grafu 3.3 s tím rozdílem, že při vyhlazení původních pozorování byla použita *penalizace za nehladkost*.

Odhady průběhu regresních funkcí, tj $\hat{u}(t)$ a $\hat{u}_g(t)$ jsou lehce vyhlazenější, avšak interpretace obrázku 3.5 je totožná s interpretací grafu 3.3. Původně nás napadlo, že pokud více vyhladíme původní pozorování průměrných denních teplot, tedy využijeme *penalizace za nehladkost*, tak ve větší míře ovlivníme vyhlazení odhadu

průběhu regresních funkcí. Podívejme se na graf 3.6 srovnávající odhady efektů skupiny jedna $\hat{\alpha}_1(t)$ a dva $\hat{\alpha}_2(t)$ při různém způsobu vyhlazení původních pozorování.

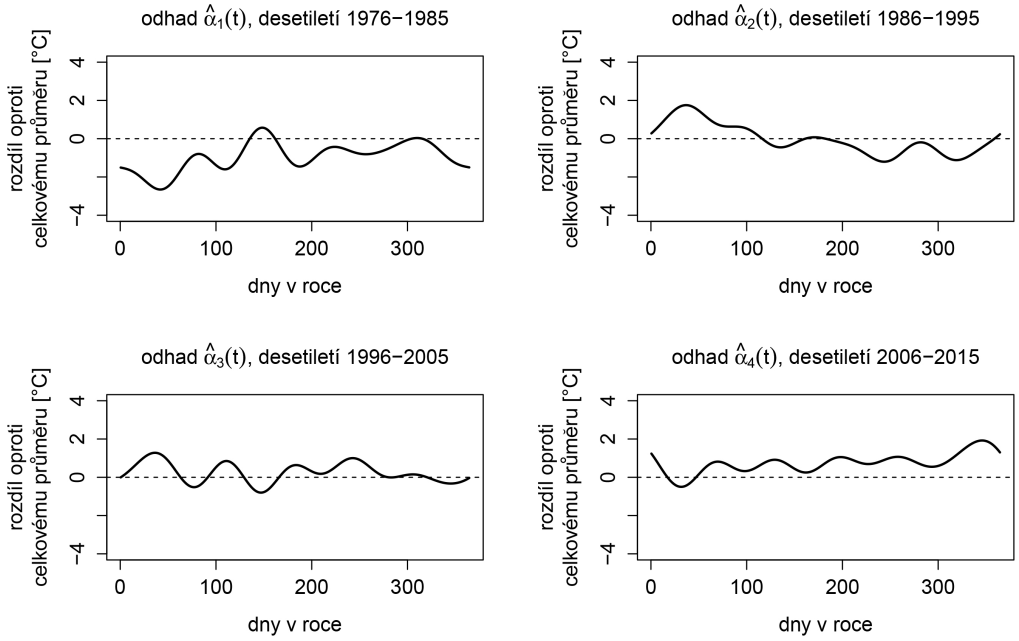
Na obrázku 3.6 vidíme, že odhady regresních funkcí na levé straně se příliš neliší od odhadů na straně pravé. Pokud bychom chtěli ve větší míře ovlivnit vyhlazení odhadu průběhu regresních funkcí, tak určitě není dobrý nápad vydat se cestou navýšení hodnoty vyhlazovacího parametru λ. Došlo by k velké ztrátě informace z původních pozorování a odhady funkcí průměrných denních teplot by nereflektovaly dostatečně původní pozorování, což by se přeneslo na odhady regresních funkcí. Pokud chceme více ovlivnit vyhlazení odhadů regresních funkcí, tak použití metody *bodové minimalizace* není nejvhodnějším nápadem. Vhodný způsob řešení představíme v následující sekci.



Obrázek 3.6: Odhady průběhu funkcí efektů $\alpha_1(t)$ a $\alpha_2(t)$ na intervalu $[0, 365]$ pro rok 2009 při různém způsobu vyhlazení původních pozorování průměrných denních teplot. Na levé straně jsou odhadnuté průběhy funkcí efektů z dat vyhlazených bez využití *penalizace za nehladkost*. Na pravé straně jsou vyobrazeny odhady funkcí efektů napočítané ze záznamů průměrných denních teplot vyhlazených s použitím *penalizace za nehladkost* (2.25), s váhou $\lambda = 10\,000$.

Na závěr kapitoly se přesuneme ke druhému příkladu časové řady průměrných denních teplot měřených na meteorologické stanici *Tallinn*. Máme opět čtyři skupiny ročních časových řad průměrných denních teplot rozdělených do skupin podle příslušnosti roku měření průměrných denních teplot do desetiletí.

Provedeme stejné operace jako v předchozím příkladu a vizualizujeme odhady průběhu funkcí efektů desetiletí.



Obrázek 3.7: Odhady průběhu funkcí efektů desetiletí na intervalu [0, 365].

V grafu 3.7 pozorujeme, že funkční hodnoty odhadu efektu desetiletí 1976 − 1985 jsou podprůměrné, tj. naměřili bychom v tomto desetiletí v průměru nižší hodnoty průměrných denních teplot ve srovnání s celkovým průměrem za celé období. Naopak funkční hodnoty odhadu efektu pro desetiletí 2006 − 2015 jsou lehce nadprůměrné. Největší rozdíly mezi odhady funkcí efektů $\hat{\alpha}_g(t)$ pro $g = 1, \ldots, 4$, pozorujeme v zimních měsících. V poslední sekci se dozvíme, zda jsou tyto rozdíly mezi funkčními hodnotami odhadů funkcí $\hat{\alpha}_g(t)$ významné.

3.5. Odhad regresních funkcí pomocí bázového rozvoje

Cílem této sekce je čtenáři ukázat druhý přístup k odhadu průběhu regresních funkcí v modelu *funkcionální analýzy rozptylu*, který vede k minimalizaci kritéria:

$$LMSSE(\boldsymbol{\beta}(\mathrm{t})) = \int_0^{365} [\mathbf{Temp}(\mathrm{t}) - \mathbf{Z}\boldsymbol{\beta}(\mathrm{t})]'[\mathbf{Temp}(\mathrm{t}) - \mathbf{Z}\boldsymbol{\beta}(\mathrm{t})]\,\mathrm{d}t.$$

a zároveň k lepší kontrole nad vyhlazením odhadů funkčních hodnot těchto funkcí.

Jak již název sekce napovídá, využíváme nabytých vědomostí z kapitoly č. 2. Stěžejní je pro nás princip odhadu průběhu funkcí pomocí *bázového rozvoje* v kombinaci s *penalizací za nehladkost*, ovlivňující vyhlazení napozorovaných hodnot.

V předchozí části jsme se snažili větším vyhlazením původních záznamů průměrných denních teplot časových řad *Pobaltí* a *Finska* ovlivnit hladkost odhadu průběhu regresních funkcí a narazili jsme na nepřekročitelné limity.

Teď je však naše filozofie zcela opačná: z původních pozorování sice opět odhadneme příslušné průběhy funkcí průměrných denních teplot, ovšem bez použití *penalizace za nehladkost*. Nechceme totiž přijít o žádnou informaci z původních dat, jejíž ztráta by mohla negativně ovlivnit odhad regresních funkcí. Hlavním záměrem je mít dobře interpretovatelné odhady regresních funkcí, čehož dosáhneme, použijeme-li *penalizaci za nehladkost* až k vyhlazení odhadu funkčních hodnot regresních funkcí.

Opět pro vysvětlení modelu využijeme příkladu časových řad průměrných denních teplot měřených na *pobaltských* a *finských* meteorologických stanicích.

Jak funkce průměrných denních teplot $Temp_i(t)$, tak i regresní funkce $u(t)$, $u_g(t)$, $g = 1, \ldots, 4$, vyjádříme bázovým rozvojem. Můžeme použít například *bázové funkce fourierovy báze* nebo *B-splinové báze*. My používáme vzhledem k periodicitě dat *bázové funkce fourierovy báze*.

Funkční hodnoty průměrných denních teplot vyjádříme:

$$\mathbf{Temp}(\mathrm{t}) = \mathbf{C}\boldsymbol{\phi}(t) \quad \forall t \in [0, 365],$$

kde $\mathbf{Temp}(\mathrm{t})$ je v našem příkladu vektor funkčních hodnot funkcí průměrných

denních teplot v časovém okamžiku t, délky 16. Matice $\mathbf{C}$ je matice koeficientů bázového rozvoje s 16 řádky a K_y sloupci, tj. počet bázových funkcí při bázovém rozvoji funkcí $Temp_i(t)$ pro $i = 1, \ldots, 16$. Vektor $\boldsymbol{\phi}(t)$ je vektor funkčních hodnot bázových funkcí v čase t, uložených do vektoru délky K_y. V jednotlivých řádcích matice $\mathbf{C}$ jsou uloženy vektory koeficientů pro funkce průměrných denních teplot, tedy druhý řádek matice $\mathbf{C}$ odpovídá vektoru koeficientů bázového rozvoje funkce průměrné denní teploty $Temp_2(t)$ *Malli*.

Postup pro výjádření funkčních hodnot vektoru regresních funkcí $\boldsymbol{\beta}(t)$ v časovém okamžiku t je obdobný jako u funkcí průměrných denních teplot:

$$\boldsymbol{\beta}(\mathrm{t}) = \mathbf{B}\boldsymbol{\theta}(t),$$

kde vektor $\boldsymbol{\beta}(\mathrm{t})$ je v našem případě délky 5. Matice $\mathbf{B}$ je matice koeficientů bázového rozvoje regresních funkcí s 5 řádky a K_β sloupci. Počet sloupců K_β odpovídá počtu bázových funkcí použitých při bázovém rozvoji regresních funkcí $\beta_j(t)$, $j = 1, \ldots, 5$. Například druhý řádek matice $\mathbf{B}$ odpovídá vektoru koeficientů bázového rozvoje funkce $\beta_2(t)$, což je funkce střední hodnoty první skupiny. Vektor $\boldsymbol{\theta}(t)$ je vektor funkčních hodnot bázových funkcí v čase t, použitých pro bázový rozvoj regresních funkcí a je délky K_β.

V některých situacích je možno využít stejné báze pro funkce průměrných denních teplot a regresních funkcí, tj. $\boldsymbol{\phi}(t) = \boldsymbol{\theta}(t)$. Avšak v jiných analýzách je záhodno pracovat s odlišnými bázemi. V našem příkladu používáme stejný bázový systém, tj. *fourierovu bázi*, ale s odlišným počtem využitých bázových funkcí. Tedy $K_y \neq K_\beta$.

Pojďme si představit tvar *penalizačního členu penalizace za nehladkost* regresní funkce $\beta_j(t)$:

$$PEN_L(\beta_j(t)) = \int_0^{365} [L\beta_j(t)]^2 \, \mathrm{d}t, \tag{3.10}$$

kde L představuje *diferenciální operátor*. V praxi můžeme místo *operátoru posunutí* použít *derivaci*, tj. penalizační člen ve tvaru 2.25. Obě dvě cesty povedou k vyhlazení odhadu průběhu regresních funkcí. Využíváme-li pro *bázový rozvoj* regresních funkcí *bázové funkce fourierovy báze*, je doporučeno použití *diferenciálního operátoru* ve formě [24]:

$$L\beta_j(t) = \omega^2 D\beta_j(t) + D^3\beta_j(t). \tag{3.11}$$

Jak bude patrno z následujícího textu, čím větší je hodnota *vyhlazovacího parametru* λ, tím více se vyhlazení průběhu regresních funkcí blíží linenární kombinaci prvních tří *bázových funkcí fouriérovy báze* [25]:

$$c_0 + a_1 \sin(\omega t) + b_1 \cos(\omega t),$$

a platí $L[c_0 + a_1 \sin(\omega t) + b_1 \cos(\omega t)] = 0$.

Při použití *diferenciálního operátoru* 3.11 na dvojice *bázových funkcí fouriérovy báze* využijeme vztahu:

$$L[a_k \sin(k\omega t) + b_k \cos(k\omega t)] = \omega^3 k(1-k^2)[a_k \cos(k\omega t) - b_k \sin(k\omega t)]. \tag{3.12}$$

Předpokládejme, že $\beta_j(t) = a_k \sin(k\omega t) + b_k \cos(k\omega t)$:

$$\begin{aligned}
D\beta_j(t) &= a_k k\omega \cos(k\omega t) - b_k k\omega \sin(k\omega t) \\
D^2\beta_j(t) &= -a_k k^2\omega^2 \sin(k\omega t) - b_k k^2\omega^2 \cos(k\omega t) \\
D^3\beta_j(t) &= -a_k k^3\omega^3 \cos(k\omega t) + b_k k^3\omega^3 \sin(k\omega t) \\
\omega^2 D\beta_j(t) + D^3\beta_j(t) &= \omega^3 k(1-k^2)[a_k \cos(k\omega t) - b_k \sin(k\omega t)].
\end{aligned}$$

Pokud je $k = 1$, výraz nabývá nulové hodnoty. Podívejme se na následující případ, kdy $k = 2$. Funkci $\beta_j(t)$ rozvineme s pomocí 5 *bázových funkcí*:

$$\beta_j(t) = c_0 + a_1 \sin(\omega t) + b_1 \cos(\omega t) + a_2 \sin(2\omega t) + b_2 \cos(2\omega t).$$

Aplikací *diferenciálního operátoru* dle 3.12 získáme:

$$L\beta_j(t) = \omega^3 2(1-2^2)[a_2 \cos(2\omega t) - b_2 \sin(2\omega t)]$$

a dosadíme do 3.10:

$$PEN_L(\beta_j(t)) = \int_0^{365} [\omega^3 2(1-2^2)[a_2 \cos(2\omega t) - b_2 \sin(2\omega t)]]^2 \, \mathrm{d}t.$$

Jedná se o určitý integrál s integrační mezí $[0, 365]$ a s hodnotou přibližně $[2^2(1-2^2)^2]$. Obecně platí, že při použití konečného počtu *bázových funkcí fouriérovy báze* je *penalizační člen za nehladkost* 3.10 přibližně roven hodnotě $[k^2(1-k^2)^2]$, kde k je nejvyšší hodnota indexu indentifikujícího dvojice funkcí $(a_k \sin(k\omega t) + b_k \cos(k\omega t))$ v rámci bázového rozvoje regresní funkce. Pro rostoucí k se hodnota

výrazu zvyšuje.

Přepneme zpět do maticového zápisu, *penalizační člen penalizace za nehladkost* vektoru regresních funkcí $\boldsymbol{\beta}$(t) je ve tvaru:

$$PEN_L(\boldsymbol{\beta}(\mathrm{t})) = \int_0^{365} [L\boldsymbol{\beta}(t)]'[L\boldsymbol{\beta}(t)]\,\mathrm{d}t.$$

Vektor $L\boldsymbol{\beta}(t) = (L\beta_1(t), L\beta_2(t), L\beta_3(t), L\beta_4(t), L\beta_5(t))'$ je délky pět.

A konečně, hledáme odhad matic koeficientů $\mathbf{B}$ a $\mathbf{C}$ minimalizujících hodnotu kritéria *penalizované metody nejmenších čtverců* ve formě:

$$LMSSE(\boldsymbol{\beta}) = \int_0^{365} [\mathbf{C}\boldsymbol{\phi}(t) - \mathbf{ZB}\boldsymbol{\theta}(t)]'[\mathbf{C}\boldsymbol{\phi}(t) - \mathbf{ZB}\boldsymbol{\theta}(t)]\mathrm{d}t + \lambda \int_0^{365} [L\boldsymbol{\beta}(t)]'[L\boldsymbol{\beta}(t)]\,\mathrm{d}t$$

V praxi máme při minimalizaci kritéria již odhad $\hat{\mathbf{C}}$ matice koeficientů $\mathbf{C}$ k dispozici a soustředíme se čistě na odhad matice $\mathbf{B}$.

Jen naznačíme, že při odhadu $\mathbf{B}$ využijeme vlastností *stopy matice* a *Kroneckerova součinu*. Detailní technický postup, jak se dostat k odhadu matice $\mathbf{B}$, je podrobně popsán v [22].

3.6. Popis skriptu pro odhad regresních funkcí pomocí bázového rozvoje

Vracíme se do prostředí statistického softwaru **R** a na příkladu časových řad průměrných denních teplot z oblasti *Pobaltí* a *Finska* názorně ukážeme postup pro odhad průběhu regresních funkcí v modelu *funkcionální analýzy rozptylu*.

V první řadě potřebujeme vyhladit původní pozorování 16 časových řad průměrných denních teplot. V porovnání s přístupem k vyhlazení dat prezentovaným v sekci *Bodová minimalizace* využijeme nyní efektivnějšího způsobu. Opět použijeme funkce balíčku `fda`.

Nejprve uložíme pozorování jednotlivých časových řad do sloupců matice `data` s 365 řádky a 16 sloupci. Např. ve třetím sloupci máme uložené záznamy časové řady průměrných denních teplot *Tallinn*.

```
dny
```

Ve vektoru `dny` jsou uloženy na jednotlivých pozicích časové okamžiky stejné časové sítě jako v předchozí sekci. Tento vektor je totožný se svým jmenovcem ze sekce *Bodová minimalizace*.

```
dayrange   = c(0,365)
daybasis65 = create.fourier.basis(dayrange, 65)
```

Opět s využitím funkce `create.fourier.basis`, která již byla představena, zadefinujeme na časovém intervalu [0, 365] 65 *bázových funkcí fourierovy báze*.

```
smoothList = smooth.basis(dny, data, daybasis65)
daytempfd  = smoothList$fd
```

Funkcí `smooth.basis` vytvoříme list skládající se z 8 částí, ve třech z nich je uložena informace o vyhlazení původních pozorování, *zobecněné cross-validaci*[1] a *residuálním součtu čtverců*. Argumenty funkce je sít časových okamžiků, matice původních pozorování časových řad a báze. My vybereme část `smoothList$fd`

[1] Ve vektoru `smoothList$gcv` jsou uskladněné hodnoty kritéria *zobecněné cross-validace* (2.33) pro vyhlazená pozorování jednotlivých časových řad, tento vektor je v našem příkladu délky 16.

obsahující informace o vyhlazení, tj. odhad matice koeficientů $\hat{\mathbf{C}}$ a vektor $\boldsymbol{\phi}(t)$ funkčních hodnot bázových funkcí v časových okamžicích t, kde $t \in$ `dny`. Část listu pojmenujeme `daytempfd`. Zde tedy máme uloženo vše potřebné pro výpočet vyhlazených pozorování časových řad.

```
tempy2cMap = smoothList$y2cMap
daytempfd$fdnames = list(NULL, nazev, NULL)
```

Z objektu `smooth.basis` vyfiltrujeme matici `tempy2cMap` s 65 řádky a 365 sloupci ve tvaru $(\boldsymbol{\phi}'\boldsymbol{\phi})^{-1}\boldsymbol{\phi}$. Matice konvertuje původní pozorování časové řady na odhad vektoru koeficientů $\hat{\mathbf{c}}$ bázového rozvoje příslušné funkce průměrných denních teplot, viz 2.12. Využitím funkce `list` přidáme do objektu `daytempfd` názvy meteorologických stanic, kde byly měřeny záznamy analyzovaných časových řad.

Ještě musíme zajistit splnění podmínky 3.3, $\sum_{g=1}^{4} \alpha_g(t) = 0$. Chceme aby na 17. pozici ve vektoru vyhlazených hodnot **Temp**(t) v čase $t \in$ `dny` byla nula korespondující s 17 řádkem matice $\mathbf{Z}$, která je totožná s *designovou maticí* použitou v příkladu při *Bodové minimalizaci*[2]. Myšlenka je stejná jako v předchozí sekci, viz strana 50.

```
coef = daytempfd$coefs
coef = cbind(coef,matrix(0,65,1))
```

Příkazem `daytempfd$coefs` necháme `Rko` vypsat matici koeficientů $\mathbf{C}'$ s rozměry 65×16. Vytvoříme sloupcový vektor délky 65 plný nul `matrix(0,65,1)` a spojíme pomocí funkce `cbind` nulový sloupeček s maticí `coef`. Výsledkem je matice s 65 řádky a 17 sloupci.

Převedeme *designovou matici* $\mathbf{Z}$ pojmenovanou `zmat` do vhodného formátu:

```
p = 5
xfdlist = vector("list",p)
for (j in 1:p) xfdlist[[j]] <- zmat[,j]
```

Vytvoříme list `xfdlist` obsahujících v řádcích sloupce matice $\mathbf{Z}$. Jednoduše řečeno přeskládáme sloupce *designové matice* do listu. První sloupec matice $\mathbf{Z}$

[2]Podobu matice $\mathbf{Z}$ najdeme na straně 49.

je teď první řádek v první části listu `xfdlist[[1]]`, druhý sloupec je prvním řádkem ve druhé části `xfdlist[[2]]` atd...

Na intervalu [0, 365] vytvoříme pro regresní funkce bázi z 13 *bázových funkcí fourierova systému*:

```
betabasis  <- create.fourier.basis(dayrange, 13)
```

Báze je pro všech 5 regresních funkcí shodná. Nyní je potřeba zadefinovat pro *penalizaci nehladkosti diferenciálního operátoru* $\omega^2 D\beta_j(t) + D^2\beta_j(t)$:

```
harmaccelLfd <- vec2Lfd(c(0,(2*pi/365)^2,0), dayrange)
```

Obecně je *diferenciální operátor* lineární kombinací derivací funkce $x(t)$ do řádu m, kde každá derivace funkce řádu $i = 0, \ldots, m-1$ může být pronásobena váhovou funkcí $b_i(t)$:

$$Lx(t) = b_0(t)x(t) + b_1(t)Dx(t) + ... + b_{m-1}(t)D^{m-1}x(t) + D^m x(t) \qquad (3.13)$$

Námi aplikovaný *diferenciální operátor* je lineární kombinací derivací regresní funkce do řádu 3. Funkční hodnoty váhových funkcí pro nultou, druhou derivaci funkcí $\beta_j(t)$, $j = 1, \ldots, 5$, jsou na intervalu $[0, 365]$ nulové a funkční hodnoty váhové funkce pro první derivaci nabývají na intervalu [0, 365] konstantní hodnoty $(2\pi/365)^2$, což odpovídá hodnotě parametru ω^2 v 3.11. Vektorem $c(0, (2 * pi/365)^2, 0)$ definujeme váhy derivací regresní funkce $D^i\beta_j$ pro i=0, ...,m-1 (kde m=3) a tím také specifikujeme samotný řád operátoru. Vektor `dayrange` upřesňuje interval, na kterém je *diferenciální operátor* definován. Funkce `vec2Lfd` vytvoří objekt *diferenciálního operátoru*, který aplikujeme na odhady funkčních hodnot regresních funkcí v modelu. Objekt `harmaccelLfd` operátoru využijeme v dalším kroku. Předtím však zadefinujeme objekt pro odhad regresní funkce.

```
betafd = fd(matrix(0,13,1), betabasis)
```

Objekt `betafd` zastupuje v `Rku` regresní funkci $\beta_j(t)$ a sdělujeme přes něj `Rku` informaci jak má vypadat regresní funkce, kterou odhadujeme, tj. specifikujeme počet bázových funkcí a bázový systém využitý pro odhad průběhu regresní funkce. Příkazem `matrix(0,13,1)` vytvoříme matici o 13 řádcích a 1 sloupci, tj.

počtu bázových funkcí využitých v rámci bázového rozvoje. Požadujeme odhad 13 koeficientů bázového rozvoje příslušné regresní funkce, tedy odhad příslušného řádku matice **B**.

Pomocí funkce `fdPar` vytvoříme objekt, kde upřesníme naše požadavky na odhad regresní funkce $\beta_j(t)$:

```
lambda    = 1e+8
betafdPar = fdPar(betafd, harmaccelLfd, lambda)
```

Argumentem funkce je objekt představující regresní funkci `betafd`, *diferenciální operátor* upřesňující podobu *penalizačního členu* a hodnota vyhlazovacího parametru λ. Alternativně místo *diferenciálního operátoru* můžeme použít druhou derivaci, tj. *penalizační člen* ve tvaru 2.25:

```
betafdPar <- fdPar(betafd, 2, lambda)
```

Vzhledem k tomu, že na regresní funkce v modelu *funkcionální analýzy rozptylu* máme stejné požadavky (báze, počet bázových funkcí, *penalizace za nehladkost*), je nejvhodnější zduplikovat objekt `betafdPar` a uložit duplikace do listu `betalist`:

```
betalist <- vector("list",p)
for (j in 1:p) betalist[[j]] <- betafdPar
```

Do první části `betalist[[1]]` ukládáme informaci o odhadu regresní funkce $\beta_1(t)$, druhá část `betalist[[2]]` obsahuje informaci o odhadu regresní funkce $\beta_2(t)$...

Přejdeme k finálnímu kroku, odhadneme matici koeficientů bázového rozvoje regresních funkcí **B**:

```
fRegressList = fRegress(daytempfd, xfdlist, betalist)
```

Funkce `fRegress` je funkcionální obdobou funkce `lm`. Do argumentu funkce dosadíme informaci o vyhlazení původních záznamů časových řad průměrných denních teplot. Dále *designovou matici* **Z** ve formě listu, kde jednotlivé části listu, tj. transponované řádky matice **Z**, korespondují s příslušnými částmi listu `betalist`, v nichž jsou uloženy informace pro odhady regresních funkcí. To znamená, že první část listu `xfdlist` představuje první transponovaný sloupec matice **Z** a

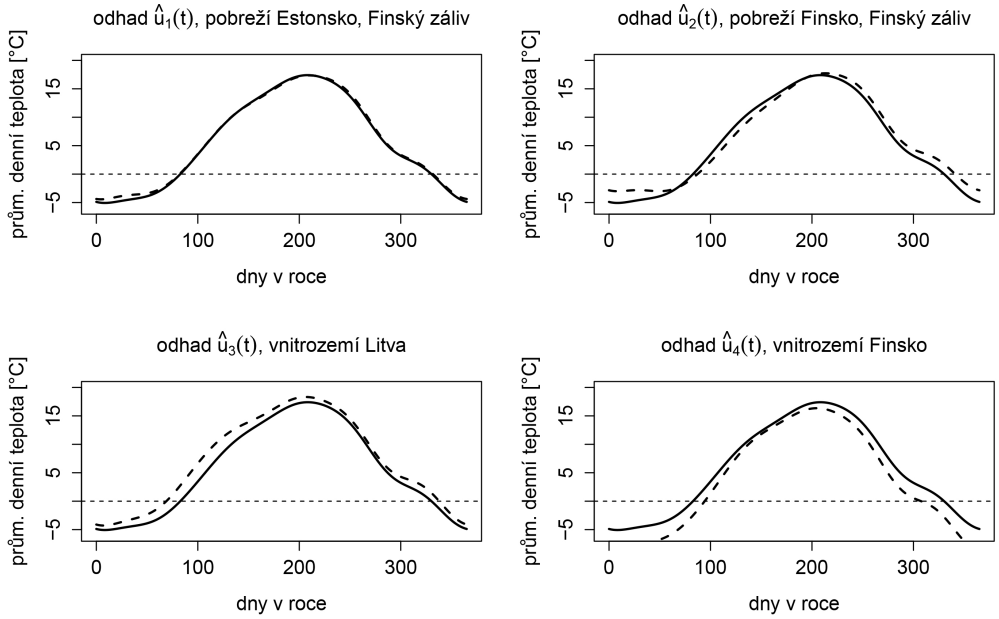
vztahuje se k první části listu `betalist`, která specifikuje požadavky na odhad funkce celkové střední hodnoty. Informaci o vyhlazení průběhu regresních funkcích uložíme do listu `betaestlist`:

```
  betaestlist = fRegressList$betaestlist
```

A nyní jednotlivě nebo pomocí `for cyklu` vizualizujeme průběhy regresních funkcí:

```
par(mfrow=c(3,2))
for (j in 1:p) {
 betaestParfdj <- betaestlist[[j]]
 plot(betaestParfdj$fd, xlab="dny v roce", ylab="denní prům. [°C]",)
}
```

Porovnejme odhady průběhu regresních funkcí získané metodami *bodové minimalizace* a *bázového rozvoje*. Vykreslíme totožné grafy jako v předchozí sekci.

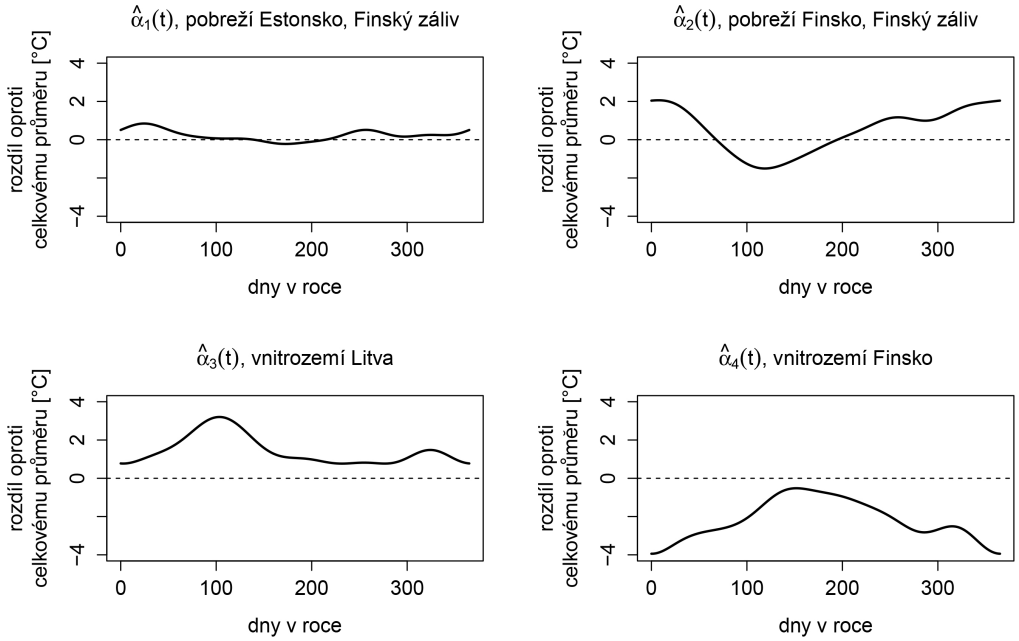


Obrázek 3.8: Odhady průběhu funkcí skupinových středních hodnot ve srovnání s odhadem průběhu funkce celkové střední hodnoty, tj. funkce celkového průměru na intervalu [0, 365] pro rok 2009. Tlustou čarou je značen průběh celkového průměru a čárkovanou průběh funkce skupinových průměrů. Pro odhad regresních funkcí byl použit *penalizační člen* s *diferenciálním operátorem* a $\lambda = 10^8$

Podívejme se na obdobu grafu 3.3 zobrazující celkovou funkci průměrných denních teplot a funkce skupinových průměrů. A dodejme, že při výpočtu odhadu průběhu regresních funkcí byla použita *metoda nejmenších čtverců s penalizací nehladkosti* tvaru 3.10, tj. penalizační člen s *diferenciálním operátorem* a váhou λ rovnou hodnotě 10^8.

Ve srovnání s grafem 3.3 jsou funkce celkového průměru a skupinových průměrů vyhlazenější. Interpretace obrázku se však s větším vyhlazením nezměnila a je stejná jako u 3.3.

Vizualizujme odhad průběhu funkcí skupinových efektů.

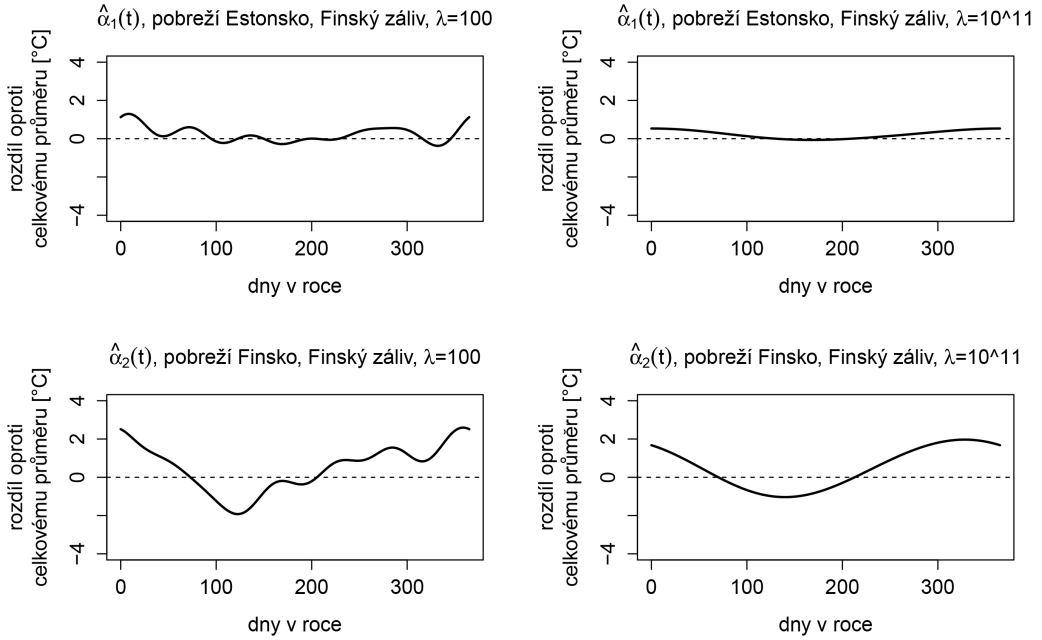


Obrázek 3.9: Odhady průběhu funkcí efektů $\alpha_g(t)$, $g = 1, \ldots, 4$, na intervalu [0, 365] roku 2009. Pro odhad regresních funkcí byl použit *penalizační člen* s *diferenciálním operátorem* a $\lambda = 10^8$

Porovnejme odhady průběhu funkcí efektů polohy s odhady v obrázku 3.4. Je evidentní, že odhady v grafu 3.9 jsou vyhlazenější, ale ne na úkor ztráty přínosné informace. Opět můžeme konstatovat, že komentář napsaný pro obrázek 3.4 je platný i zde.

Použijeme-li verzi *penalizačního členu* s derivací ve tvaru 2.25 a nastavíme-li vyhlazovací parametr λ na hodnotu 10^6, dostaneme velmi podobný obrázek jako 3.9.

Na závěr ukážeme odhady průběhu funkcí $\hat{\alpha_1}(t)$ a $\hat{\alpha_2}(t)$ při malé a velké hodnotě parametru λ. V levé polovině grafu vizualizujeme odhady průběhu funkcí skupinových efektů při malé hodnotě parametru λ, v pravé části máme odhady stejných funkcí při vysoké hodnotě λ.



Obrázek 3.10: Odhady průběhu funkcí skupinových efektů $\alpha_g(t)$ na intervalu [0, 365] pro rok 2009 při různém nastavení parametru λ. Pro odhad regresních funkcí byl použit *penalizační člen* s *lineárním operátorem posunu*.

Odhady průběhu funkcí na levé straně jsou v podstatě totožné s odhady funkcí efektů pomocí metody *bodové minimalizace*, viz levá polovina obrázku 3.6, což nás překvapilo. Čekali jsme, že při nastavení nízké hodnoty parametru λ dostaneme podobné odhady, ale nečekali jsme, že budou skoro totožné. Dále vidíme, že odhady funkcí efektů na straně pravé jsou příliš jednoduché, tj. příliš vyhlazené.

Zrovna v našem příkladu se intepretační hodnota při použití metod *bázového rozvoje* a *bodové minimalizace* pro odhad průběhu regresních funkcí příliš neliší. Avšak mohou nastat případy a často nastanou, kdy je intepretace odhadu regresních funkcí získaných přes bodovou minimalizaci vzhledem k nehladkosti značně složitá a je proto žádoucí použít metodu představenou v této sekci.

3.7. Testování

V předchozích sekcích této kapitoly byl popsán základní model *funkcionální analýzy rozptylu*. Naučili jsme se odhadovat a posléze vyhlazovat průběhy regresních funkcí $u(t)$, $\alpha_g(t)$, $g = 1, \ldots, 4$. Poslední část, která dokončí naší mozaiku zvanou *funkcionální analýza rozptylu*, se vztahuje k testování nulové hypotézy:

$$u_1(t) = u_2(t) = u_3(t) = u_4(t) \quad \forall t \in [0, 365], \tag{3.14}$$

jež byla detailně okomentována v úvodu sekce 3.2. Nulovou hypotézu zamítáme v případě, kdy existuje časový okamžik $t \in [0, 365]$, pro který platí, že funkční hodnota jedné nebo více funkcí skupinových středních hodnot $u_g(t)$, $g = 1, \ldots, 4$, nesplňují rovnost 3.14.

První možností je využít tzv. *bodovou F statistiku* ve tvaru:

$$F_n(t) = \frac{SSR_n(t)/(l-1)}{SSE_n(t)/(n-l)}, \tag{3.15}$$

kde

$$SSR_n(t) = \sum_{g=1}^{4} n_g(\hat{u}_g(t) - \hat{u}(t))^2,$$

$$SSE_n(t) = \sum_{g=1}^{4} \sum_{m=1}^{4} n_g(Temp_{mg}(t) - \hat{u}_g(t))^2.$$

Symbol l značí počet skupin ($l = 4$), n je celkový počet funkcí, v našem případě roven 16. Počet pozorování v rámci skupiny představuje symbol n_g ($n_g = 4$ pro $g = 1, \ldots, 4$). Průběh funkce $SSR_n(t)$ na intervalu [0, 365] nám dává informaci o meziskupinové variabilitě. Nabývá-li tato funkce v každém časovém okamžiku $t \in [0, 365]$ nízkých funkčních hodnot, tj. nezáporných a blízkých nule, je to signál, že mezi skupinami jsou malé nebo žádné rozdíly. Průběh funkce $SSE_n(t)$ na intervalu $[0, 365]$ popisuje variabilitu uvnitř jednotlivých skupin. K praktickému výpočtu *bodové F statistiky* se vrátíme později.

Velká část testů pro nulovou hypotézu 3.14 vychází z testové statistiky 3.15. Nyní čtenáři představíme ucelený přehled těchto testů [23].

- *Faraway* a *Zhang Chen* navrhli využít jako testovou statistiku určitý integrál $\int_0^{365} SSR_n(t)\,\mathrm{d}t$. Testy z ní vycházející nazýváme testy založenými na L^2-normě. Rozdělení pravděpodobnosti této testové statistiky může být aproximováno $\beta\chi^2_d$ rozdělením. Pro odhad parametrů β a d můžeme použít *naivní* nebo *vychýlení redukující* metodu. Test využívající pro odhad parametrů naivní přístup nazýváme L^2N a test používající *vychlýlení redukující* metodu L^2B testem. V případě malého rozsahu výběru je možnost použít bootstrap k odhadu rozdělení testové statistiky. Tuto variantu testu označujeme L^2b.

- *Cuevas* nabízí modifikaci testů založených na L^2-normě postavenou na statistice:

$$\sum_{1\le i<j\le 4} n_i \int_0^{365} (\hat{u}_i(t) - \hat{u}_j(t))^2\,\mathrm{d}t.$$

  Z *Cuevasovy* testové statistiky vycházejí dvě testové procedury. První pro homoskedastické případy, zkráceně *CH*, a druhá pro heteroskedastické (*CS*).

- Test používající upravenou testovou statistiku 3.15:

$$F_n(t) = \frac{\int_0^{365} SSR_n(t)\mathrm{d}t/(k-1)}{\int_0^{365} SSE_n(t)\mathrm{d}t/(n-k)} \tag{3.16}$$

  nazýváme testem typu F (*F-type*). Rozdělení pravděpodobnosti testové statistiky 3.16 je aproximováno $F_{(k-1)\kappa,(n-k)\kappa}$ rozdělením. Podle zvolené metody odhadu parametru κ rozlišujeme dva testy, *LN* (naivní metoda) a *LB* (vychýlení redukující metoda) test. Pro malý rozsah výběru funkcí existuje bootstrapová varianta zvaná *Fb* test.

- *Zhang* a *Liang* doporučují *globalizaci bodového F-testu* (*GPF*), založenou na testové statistice $\int_0^{365} F_n(t)\,\mathrm{d}t$. V některých případech může být lepší (ve smyslu větší síly), brát jako testovou statistiku $sup_{t\in[0,365]} F_n(t)$ a simulovanou kritickou hodnotou. Tento test označujeme za *Fmaxb* test. *Zhang* zjistil, že *Fmaxb* test je obecně silnější než *GPF* test.

- Testy založené na analýze náhodných projekcích pojmenováváme jako *TRP testy*.

Vraťme se opěk k našemu příkladu časových řad z oblasti *Pobaltí* a *Finska*. Na vyhlazená pozorování použijeme výše představené testy a vyjádříme se k zamítnutí či nezamítnutí nulové hypotézy 3.14. Testy jsou implementovány v `rkovském` balíčku `fdANOVA`. Donedávna neexistoval balíček, který by většinu ze zmíněných testů obsahoval. Využijeme funkci `fanova.tests`:

```
fanova.tests(x=matice[,-c(17)],
             group.label = as.character(group.label.temp))
```

V argumentu funkce je `matice` vyhlazených pozorování průměrných denních teplot s 365 řádky a 16 sloupci, poslední sloupec plný nul nebereme v potaz. Ještě do funkce dosadíme identifikační vektor `as.character(group.label.temp)`, který rozřadí jednotlivé odhady průběhu funkcí průměrných denních teplot, tj. sloupce `matice`, do čtyř skupin. Nastavení všech ostatních parametrů necháme na výchozích hodnotách. Podívejme se na výstup této funkce upravený do tabulky 3.2:

Název testu	**hodnota testové statistiky**	**p-hodnota**
$L2^N$	15535.12	0
$L2^B$	15535.12	0
$L2^b$	15535.12	0
CH	62140.47	0.0034
CS	62140.47	0
LN	60.99749	0
LB	60.99749	0
Lb	60.99749	0.0003
GPF	70.03239	0
$Fmaxb$	238.9033	0

Tabulka 3.2: Přehled výsledků testů hypotézy *funkcionální analýzy rozptylu*, tj. testových statistik a příslušných p-hodnot, pro vyhlazené pozorování průměrných denních teplot získaných z *baltských* a *finských* časových řad.

U všech 10 testů v tabulce 3.2 vyšla p-hodnota nulová nebo blízká nule. Zamítáme tedy nulovou hypotézu o shodném průběhu funkcí skupinových středních hodnot (3.14). Mezi skupinami časových řad průměrných denních teplot naměřených v oblasti *Pobaltí* a *Finska* jsou statisticky významné rozdíly, což je tvrzení jenž

není v rozporu se závěry učiněnými v sekci *Bodová minimalizace* v komentářích k obrázkům 3.3 a 3.4 na straně 53 a 54.

Nyní se podrobněji zaměříme na *FP* test, představený *Góreckým* a *Smagou* [23], používající testovou statistiku 3.16. Vycházíme z bázové reprezentace m-té funkce průměrné denní teploty v g-té skupině:

$$Temp_{mg}(t) = \sum_{i=1}^{K} c_{mgi}\phi_i(t) = \mathbf{c}'_{mg}\boldsymbol{\phi}(t) \quad \forall t \in [0, 365], \tag{3.17}$$

kde $\mathbf{c}_{mg}$ je vektor koeficientů bázového rozvoje m-té funkce ve skupině g délky K. Téma bázové reprezentace funkce je popsáno ve druhé kapitole. Vyjádřením funkcí zahrnutých v analýze pomocí bázového rozvoje 3.17 *Górecký* a *Smaga* dokázali, že testovou statistiku 3.16 lze vyjádřit ve tvaru:

$$\frac{(a-b)/(l-1)}{(c-a)/(n-l)}, \tag{3.18}$$

kde

$$a = \sum_{g=1}^{l} \frac{1}{n_g}\mathbf{1}'_{n_g}\mathbf{C}_g\mathbf{J}_\phi\mathbf{C}'_g\mathbf{1}_{n_g},\ b = \frac{1}{n}\sum_{g=1}^{l}\sum_{h=1}^{l}\mathbf{1}'_{n_g}\mathbf{C}_g\mathbf{J}_\phi\mathbf{C}'_h\mathbf{1}_{n_h},\ c = \sum_{g=1}^{l} stopa(\mathbf{C}_g\mathbf{J}_\phi\mathbf{C}'_g).$$

Vektor $\mathbf{1}_{n_g}$ je vektor jedniček délky n_g (je počet funkcí zahrnutých do g-té skupiny) a n je celkový počet funkcí. Matice $\mathbf{C}_g = (c_{mgi})$, $m = 1, \ldots, n_g$, $i = 1, \ldots, K$, je matice koeficientů bázového rozvoje funkcí patřících do skupiny g s rozměry $n_g \times K$, v našem případě 4×13. Matice $\mathbf{J}_\phi = \int_0^{365}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'\mathrm{d}t$ je matice s K řádky a K sloupci.

Výpočet testové statistiky 3.16 může být tedy založen pouze na znalosti odhadů koeficientů c_{mgi} a matice $\mathbf{J}_\phi$. Podle *Góreckého* a *Smagy* libovolné rozdělení funkcí do skupin nezmění hodnoty b, c. Proto *Górecki* a *Smaga* přišli s permutační variantou testu založenou na testové statistice 3.18. Poznamenejme, že rozdělení testové statistiky *FP* testu 3.18 není aproximováno $F_{(k-1)\kappa,(n-k)\kappa}$ rozdělením. Detailnější popis výpočtu p-hodnoty najdeme v dodatku. Test je také implementován v balíčku `fdANOVA`. Ukážeme aplikaci *FP* testu na našich datech.

Na intervalu [0, 365] vygenerujeme bázi 13 bázových funkcí z *Fouriérova systému*.

```
basis = create.fourier.basis(c(0,365), 13)
```

Spočítáme matici $\mathbf{J}_\phi$.

```
own.cross.prod.mat=inprod(basis,basis)
```

Dále odhadneme transponovanou matici koeficientů bázového rozvoje $\mathbf{C}'$ s rozměry 13×16.

```
own.basis=Data2fd(dny,data,basis)$coef
```

Do funkce `Data2fd` dosazujeme vektor časových okamžiků `dny`, matici nevyhlazených záznamů časových řad s názvem `data` s rozměry 365×16 a bázi. Teď máme vše potřebné napočítané a přejdeme k *FP* testu:

```
fanova.tests(x=NULL, as.character(group.label.temp),test = c("FP"),
params = list(paramFP = list(B.FP = 100, basis = "own",
own.basis = own.basis,own.cross.prod.mat =own.cross.prod.mat)))
```

FP test je jediným testem v balíčku `fdANOVA`, u kterého můžeme nastavit v argumentu funkce `fanova.tests` `x=NULL`, to znamená, že do argumentu funkce přímo nedosadíme matici vyhlazených pozorování průměrných denních teplot. Jediné, co potřebujeme do funkce dosadit, je vzhledem ke konstrukci testové statistiky 3.18 odhad matice koeficientů bázového rozvoje $\mathbf{C}$, která je uložena v objektu `own.basis`, a matici $\mathbf{J}_\phi$ uloženou pod názvem `own.cross.prod.mat`. V případě, že nenastavíme parametr `basis = "own"`, je potřeba dosadit do funkce vyhlazená pozorování, tj. například `x=matice[,-c(17)]`. Hodnotou parametru `B.FP` nastavíme počet permutací. Podívejme se na výstup této funkce:

```
P test - permutation test based on a basis function representation
Test statistic = 60.9975  p-value = 0
```

P-hodnota je nulová, výsledek je tedy stejný jako u testů v tabulce 3.2, nulovou hypotézu opět zamítáme. I podle výsledku tohoto testu jsou mezi skupinami

časových řad průměrných denních teplot z oblasti *Pobaltí* a *Finska* rozdíly.

Nejprve nás napadlo, zda změna nastavení parametru `B.FP` by mohla mít vliv na zamítnutí či nezamítnutí nulové hypotézy. Zkusili jsme přenastavit hodnotu parametru na 100 000, výsledek byl totožný s výstupem prezentovaným výše.

3.8. Experiment: vyšetření vlastností FP testu

Vzhledem k výsledku testu, tj. nulové p-hodnotě nás zajímalo, jak bude test reagovat, vytvoříme-li z původních skupin nové skupiny, kde v našem případě v každé z nich bude právě jeden unikátní zástupce z původní skupiny. V rámci nových skupin se jedná o náhodný výběr bez vracení. Získáme nové čtyři skupiny po čtyřech funkcích, takto rozdělené skupiny by měly být velmi podobné. Dopředu očekáváme, že nedojde k zamítnutí nulové hypotézy.

Nejprve vytvoříme pomocí příkazu `sample` náhodné permutace uvnitř skupin:

```
perm1=sample(1:4,4)
perm2=sample(5:8,4)
perm3=sample(9:12,4)
perm4=sample(13:16,4)
```

tj, například

```
perm1
  2 3 1 4
```

Nyní vezmeme z vektorů `perm1` až `perm4` hodnoty na prvních pozicích a uložíme je do vektoru `skup1`, tímto postupem získáme unikátní indikátory pro skupinu, která splňuje naše požadavky, tj. obsahuje po jedné funkci z každé původní skupiny. Pro indikátory funkcí druhé skupiny vezmeme hodnoty z vektorů permutací na druhém místě a opět uložíme do vektoru `skup2`. Analogický postup je pro třetí a čtvrtou skupinu.

```
skup1=c(perm1[1],perm2[1],perm3[1],perm4[1])
skup2=c(perm1[2],perm2[2],perm3[2],perm4[2])
```

Ve vektor `skup1` jsou pro ilustraci obsaženy indikátory funkcí:

```
2 6 10 15.
```

Vytvoříme prázdný finální vektor indikátorů a požadujeme, aby na pozicích 2, 6, 10, 15, byly jedničky. Tím `Rku` sdělujeme informaci, že tyto funkce patří do první skupiny. Stejný postup uplatníme pro druhou, třetí a čtvrtou skupinu.

```
group.label.temp=c()
group.label.temp[skup1]=1
group.label.temp[skup2]=2
```

Tímto způsobem jsme získali nový vektor indikátorů `group.label.temp`, dávající nám informaci o složení nových skupin:

```
2 1 4 3 3 1 4 2 3 1 2 4 3 2 1 4
```

Původně měl tento vektor uložené hodnoty:

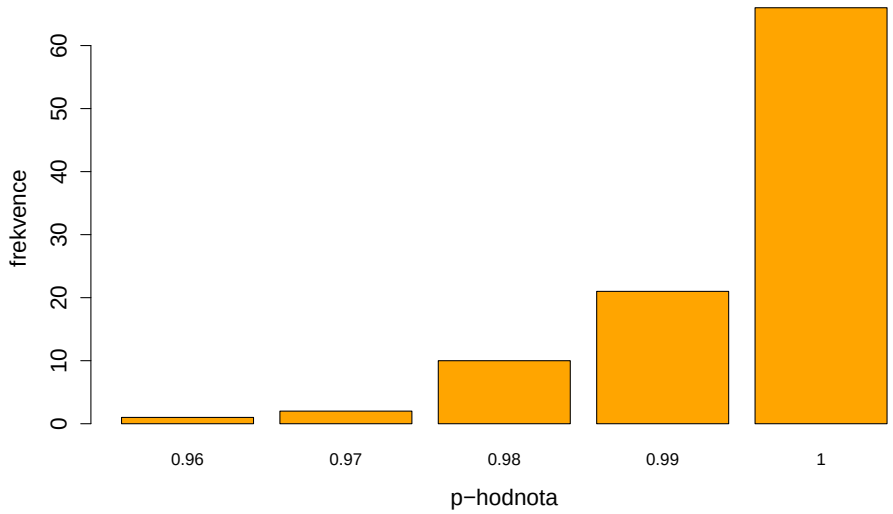
```
1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4.
```

Máme nové uspořádání skupin odhadů funkcí průměrných denních teplot a teď s takto uspořádanými skupinami otestujeme s využitím *FP* testu nulovou hypotézu o shodě průběhu funkcí skupinových středních hodnot na časovém intervalu $[0, 365]$. Celou výše představenou proceduru od tvorby náhodných permutací uvnitř skupin až po testování hypotézy zopakujeme pomocí `for cyklu` 100 krát a výsledné p-hodnoty uložíme do vektoru `phodnota` délky 100.

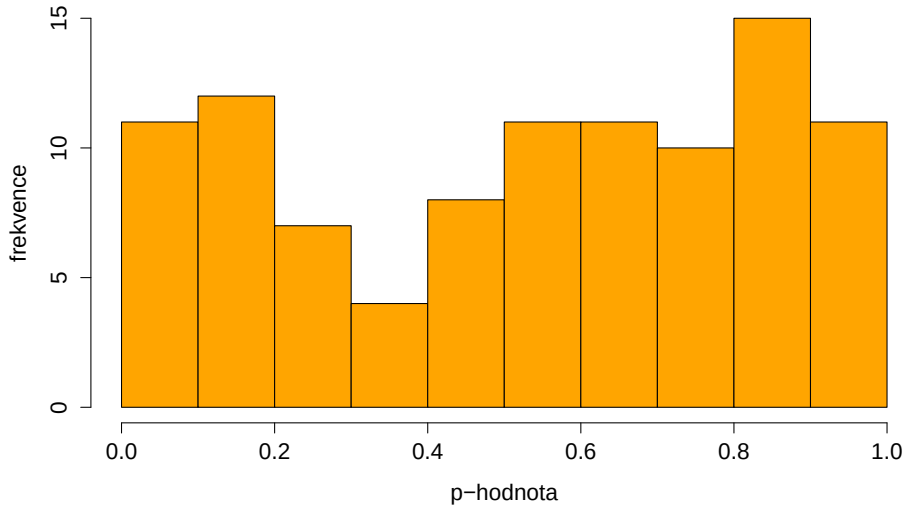
Podívejme se na výsledný histogram pro vektor `phodnota` 3.11. V 76 případech byla výsledná p-hodnota rovna 1. Nejnižší p-hodnota v rámci opakovaného testování byla 0,96. To je výsledek, který jsme vzhledem k velké podobnosti jednotlivých skupin odhadu funkcí čekali. Ve všech 100 případech nebyla nulová hypotéza zamítnuta.

Ještě nás zajímá, jak se bude *FP* test chovat v případě, že dojde k náhodnému zařazení funkcí do jednotlivých skupin. Může se stát, že během simulace v jedné nové skupině budou například dvě funkce ze skupiny původní, to je rozdíl mezi první a druhou simulací. Čekáme, že v některých případech dojde k zamítnutí nulové hypotézy a v některých nikoliv. Opět jako při první simulaci použijeme funkci `sample(1:16,16)`, která nám vrátí náhodně uspořádaných šestnáct čísel.

Vezmeme první čtveřici a uložíme ji do vektoru `perm1`, druhou čtveřici uložíme do vektoru `perm2`. Vektory nám dávají informaci o rozřazení funkcí do skupin. Dále už je postup stejný, jako v předchozí simulaci. Histogram 3.12 je obdobou histogramu 3.11.



Obrázek 3.11: Histogram záznamů p-hodnot *FP* testu během první simulace.



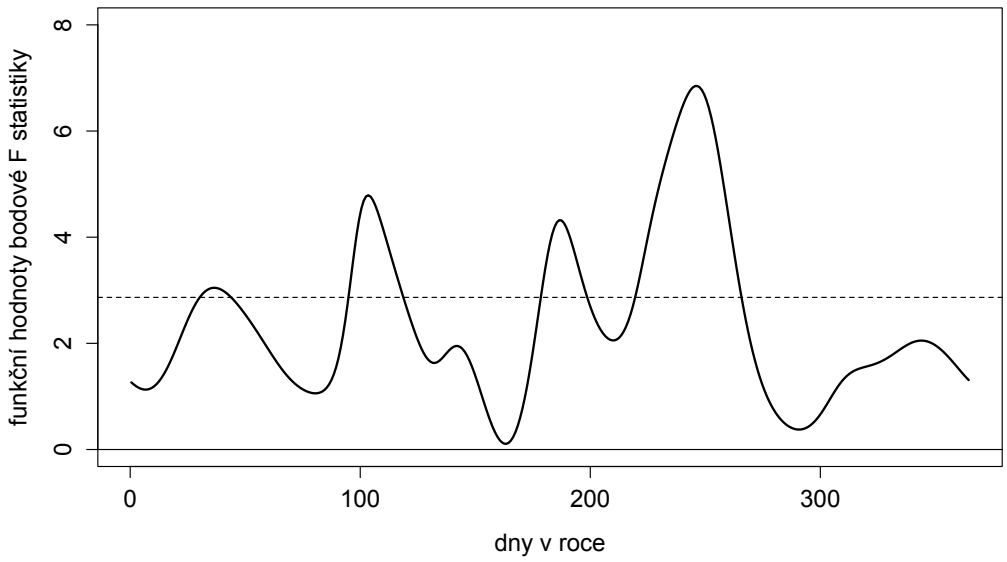
Obrázek 3.12: Histogram záznamů p-hodnot *FP* testu během druhé simulace s náhodným rozřazením odhadu funkcí do skupin.

V sedmi případech ze 100 došlo k zamítnutí nulové hypotézy, tj. p-hodnota

byla menší nebo rovna hodnotě 0.05. Možná bychom čekali, že vícekrát dojde k zamítnutí nulové hypotézy. Na druhou stranu pořád je to výsledek, který jsme očekávali. Odhady funkcí průměrných denních teplot jsou náhodně rozřazeny do jednotlivých skupin, které si byly v průběhu simulace někdy méně podobné a někdy více podobné, což se odráží v histogramu 3.12.

Na závěr kapitoly se vracíme k *tallinnské časové řadě* průměrných denních teplot. Zajímá nás, zda jsou mezi skupinami ročních časových řad představujících jednotlivá desetiletí rozdíly. Vypočítáme průběh funkce testové statistiky $F_n(t)$ (3.15).

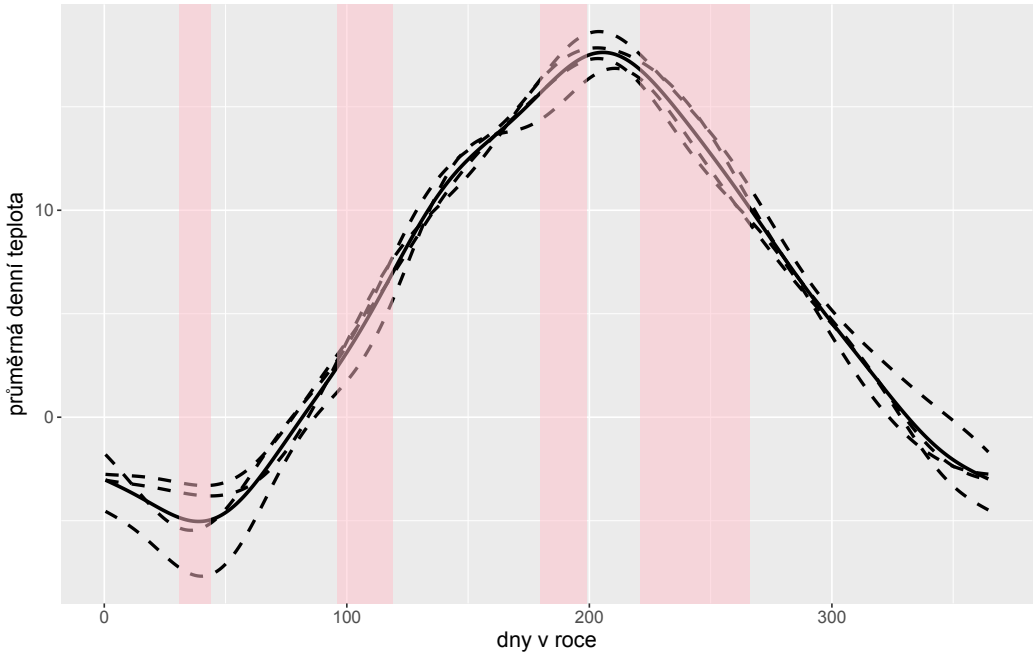
Postup je následující: ve vybraných časových okamžicích t=0,5;1,5;...;364,5, kde každému takovému časovému okamžiku je přiřazeno 40 vyhlazených záznamů průměrných denních teplot, napočítáme hodnotu testové statistiky jednofaktorové analýzu rozptylu. Získáme vektor, ve kterém jsou uloženy funkční hodnoty funkce testové statistiky $F_n(t)$ a je délky 365. Interpolací těchto hodnot získáme průběh funkce $F_n(t)$, viz obrázek 3.13.



Obrázek 3.13: Průběh funkce *bodové F* statistiky $F_n(t)$ (3.15) pro vyhlazená pozorování průměrných denních teplot časové řady *Tallinn*.

Čárkovaná linie v grafu 3.13 představuje na intervalu $[0, 365]$ kritickou hodnotu $F_{3,36,0,95}$. Vidíme, že nulovou hypotézu o shodě průběhu funkcí skupinových středních hodnot nejrazantněji zamítáme v srpnových a zářijových dnech. V tomto období nejvíce překračujeme kritickou hodnotu.

Ještě se podívejme na vizualizaci 3.14 odhadu regresních funkcí v modelu *funkcionální analýzy rozptylu*, kde je červeně zvýrazněné období, ve kterém zamítáme nulovou hypotézu.



Obrázek 3.14: Odhady regresních funkcí v modelu *funkcionální analýzy rozptylu* pro vyhlazená pozorování průměrných denních teplot časové řady *Tallinn*. Červeně je zvýrazněné období, ve kterém zamítáme nulovou hypotézu.

Na první pohled nás překvapilo, že v zimním období nedošlo u většího počtu dnů k zamítnutí nulové hypotézy. Vysvětlení je následující: v zimním období pozorujeme zvýšenou variabilitu uvnitř jednotlivých skupin, tj. vyšší funkční hodnoty funkce $SSE_n(t)$, což i přes viditelné rozdíly mezi funkcemi skupinových průměrů nestačí k zamítnutí nulové hypotézy.

Náš závěr, tedy zamítnutí nulové hypotézy v tomto příkladu i v příkladu s časovými řadami průměrných denních teplot naměřenými v oblasti *Pobaltí* a *Finska*, odpovídá naší úvaze z úvodu této kapitoly.

3.9. Dodatek: výpočet p-hodnoty *FP* testu

FP test je permutační test. Myšlenka výpočtu p-hodnoty při 100 permutacích je následující: Nejprve napočítáme hodnotu testové statistiky při našem původním rozřazení funkcí do skupin a hodnotu uložíme. V dalším kroku 16 funkcí náhodně rozřadíme do čtyř skupin a opět napočítáme testovou statistiku. Tento poslední krok zopakujeme ještě 99 krát. Získáme 100 hodnot testové statistiky při různém rozřazení funkcí do skupin.

P-hodnotu chápeme, jako podmíněnou pravděpodobnost, že F statistika nabyde stejné nebo ještě větší hodnoty než její zjištěná hodnota za platnosti nulové hypotézy. Můžeme ji odhadnout vztahem $\frac{S}{N}$, kde S je počet případů, kdy hodnota testové statistiky počítané z přeuspořádaných skupin funkcí byla větší nebo rovna hodnotě testové statistiky při námi uvažovaném rozřazení funkcí do skupin. N je celkový počet permutací, tj. výpočtů testové statistiky při náhodném uspořádání skupin funkcí.

Například p-hodnota $0,97$ znamená, že v 97 případech ze 100 se testová statistika realizovala do stejné nebo kritičtější hodnoty ve srovnání s hodnotou testové statistiky při námi uvažovaném uspořádání skupin. [26]

Závěr

V úvodu práce jsme čtenáři představili analyzovanou datovou sadu časových řad průměrných denních teplot a náš postup při očištění a nahrávání dat do `R`. Během našich pokusů při čištění dat jsme zíkali základní povědomí o programovacím jazyku *Python* a databázovém systému `MySQL`. Dozvěděli jsme se, že v prostředí softwaru `R` existuje balíček `plotGooglemaps`, který využíváme pro propojení polohy meteorologických stanic s *google mapou*.

Ve druhé kapitole jsme se naučili odhadovat z původních pozorování na příkladu časové řady z meteorologické stanice *Waldassen* příslušnou funkci průměrné denní teploty. Seznámili jsme se s *penalizačním členem*, což je prostředek ovlivňující míru vyhlazení naměřených záznamů. V této kapitole je představen i způsob volby váhy penalizačního členu, tzv. *zobecněná cross-validace*. Zjistili jsme, že i při větším počtu využitých bázových funkcí během odhadu průběhu funkce může být odhad dostatečně hladký, nastavíme-li vhodně váhu *penalizačního členu*. Při volbě váhy nespoléháme pouze na výsledky zobecněné cross-validace. Může se stát, že tato metoda doporučí volbu váhy, která nemusí být pro naše další účely vhodná. Vždy záleží na konkrétní aplikaci.

Třetí kapitola je věnována samotné *funkcionální analýze rozptylu*. Metodu aplikujeme na skupiny časových řad průměrných denních teplot naměřených v oblasti *Pobaltí* a *Finska*. Ukázali jsme dva přístupy pro odhad regresních funkcí *funkcionální analýzy rozptylu* a to *bodovou minimalizaci* a *odhad regresních funkcí pomocí bázového rozvoje*. Chceme-li mít větší kontrolu nad vyhlazením odhadů regresních funkcí, je vhodné využít přístup pomocí bázového rozvoje. Součástí kapitoly je také popis skriptu z prostředí `R`. Získáváme tak nejen teoretické zna-

losti, ale i návod, jak postupovat, aplikujeme-li metodu na jiná data.

Poslední podkapitola se týká testování nulové hypotézy *funkcionální analýzy rozptylu*. Stručně jsme sepsali přehled testů, které jsou obsaženy v softwaru `R` v balíčku `fdANOVA` a zaměřeli se podrobněji na *FP* test, který je založen na bázové reprezentaci analyzovaných funkcí. Opět jsme sepsali i postup, jak implementovat test v prostředí `R`. *FP* test jsme aplikovali na vyhlazená pozorování skupin časových řad průměrných denních teplot z *Pobaltí* a *Finska*. Zajímalo nás, zda mezi těmito skupinami jsou statisticky významné rozdíly. Nulová hypotéza byla na 5% hladině významnosti zamítnuta. Tento závěr nebyl v rozporu s naší úvahou. Na severu naměříme nižší teploty než na jihu. Například na začátku dubna roku 2009 naměříme ve *vnitrozemí Finska* teploty v průměru o 5°C nižší než ve *vnitrozemí Litvy*.

Na závěr jsme uskutečnili experiment, jehož cílem bylo vyšetřit validitu *FP* testu. V první fázi jsme opakovaně sestavili skupiny odhadů funkcí, které by si měly být velmi podobné a provedli testování. Očekávali jsme, že test ve všech případech vyhodnotí rozdíly mezi skupinami jako nevýznamné a nezamítne nulovou hypotézu. Tato simulace naše očekávání potvrdila. Ve druhé fázi jsme provedli podobnou simulaci s tím rozdílem, že odhady funkcí byly náhodně rozřazeny do skupin. V sedmi případech ze 100 došlo k zamítnutí nulové hypotézy. Skupiny si byly v průběhu simulace někdy méně podobné a někdy více podobné.

Seznam tabulek

1.1 Ukázka datové tabulky 1. 10
1.2 Ukázka upravené datové tabulky 1. 10
1.3 Tabulka pěti států s nejvyšším počtem meteorologických stanic v námi analyzované datové sadě 11
1.4 Ukázka datové tabulky 2, popisující sadu časových řad průměrných denních teplot 12
1.5 Tabulka meteorologických stanic s nejstarším záznamem průmerné denní teploty v rámci námi analyzované datové sady 13
1.6 Tabulka vzestupně seřazených meteorologických stanic s nejnižší nadmořskou výškou v rámci námi analyzované datové sady 13
1.7 Tabulka sestupně seřazených meteorologických stanic s nejvyšší nadmořskou výškou v rámci námi analyzované datové sady 14
1.8 Ukázka časové řady denních průměrných teplot naměřených v jeruzalémské meteorologické stanici 14

3.1 Tabulka časových řad průměrných denních teplot zahrnutých do *funkcionální analýzy rozptylu*. Data pocházejí z datové sady projektu *ECA&D*. 42
3.2 Přehled výsledků testů hypotézy *funkcionální analýzy rozptylu*, tj. testových statistik a příslušných p-hodnot, pro vyhlazené pozorování průměrných denních teplot získaných z *baltských* a *finských* časových řad. 71

Seznam obrázků

1.1 Vizualizace polohy meteorologickych stanic v prostředí `R` s využitím balíčků `rworldmap`. 11
1.2 Datová mapa, vizualizace struktůry dat a vztahů mezi datovými tabulkam využitými při analýze meteorologických stanic. 15
1.3 Ukázka vizualizace polohy meteorologickych stanic v prostředí `R` s využitím balíčků `plotGoogleMaps`. 17
1.4 Ukázka vizualizace polohy meteorologické stanice *Praha-Klementinum* v prostředí `R` s využitím balíčků `plotGoogleMaps`. 17

2.1 Časová řada průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* 20
2.2 *Fourierova báze* pro 13 bázových funkcí na intervalu $[0, 365]$ . . . 23
2.3 *Bázové spline funkce* $B_{1,1}(t)$ a $B_{2,1}(t)$ řádu 1 na intervalu $[0, 365]$. 25
2.4 13 *bázových spline funkcí* řádu 4 na intervalu $[0, 365]$ 25
2.5 *Fourierova báze* pro 13 bázových funkcí na intervalu $[0, 365]$ s vyznačenými funkčními hodnotamy bázových funkcí v časovém okamžiku 100. 26
2.6 Odhad funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *Fourierovy báze*. 28
2.7 Odhad funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *B-spline báze*. 28
2.8 Odhad náhodných odchylkek z dat průměrných denních teplot měřených v roce 1951 na meteorologické stanici *Waldassen*. 29
2.9 Odhady funkcí průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* při různém nastavení počtu bázových funkcí *Fourierovy báze*. 32
2.10 Odhad parametru σ^2 při různém nastavení počtu bázových funkcí z *Fourierovy báze*. 33
2.11 Odhad průběhu funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *B-spline bázových funkcí*. 36

2.12 Odhad průběhu funkce průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* s využitím *B-spline bázových funkcí*. 36
2.13 Hodnoty kritéria *zobecněné cross-validace*(2.33) počítané z pozorování průměrných denních teplot měřených v roce 1951 na meteorologické stanici *Waldassen* při počtu 32 *bázových funkcích řádu* 5 *z B-splinové báze*. 38
2.14 Hodnoty kritéria *obecné cross-validace*(2.33) při různém nastavení počtu *bázových funkcí řádu 5 z B-splinové báze* 38
2.15 Odhady funkcí průměrné denní teploty měřené v roce 1951 na meteorologické stanici *Waldassen* při různém nastavení počtu bázových funkcí *Fourierovy báze*. Pro odhad průběhu funkcí byla použita *standardní metoda nejmenších čtverců* s *penalizací nehladkosti*. . . 39

3.1 Vizualizace skupin *baltských* a *finských* meteorologických stanic podle jejich polohy. Podklad pro obrázek pochází z webové stránky *mapy.cz*. 41
3.2 Vizualizace *pobaltských* meteorologických stanic. V modrém kruhu jsou vyobrazeny potencionální kandidáti pro skupinu č.3 *Vnitrozemí Litva*. 43
3.3 Odhady průběhu funkcí skupinových středních hodnot $u_g(t)$ ve srovnání s odhadem průběhu funkce celkové střední hodnoty, tj. funkce celkového průměru na intervalu [0, 365] pro rok 2009. Tlustou čarou je značen průběh celkového průměru a čárkovanou průběh funkce skupinových průměrů. 53
3.4 Odhady průběhu funkcí efektů $\alpha_g(t)$ na intervalu [0, 365] pro rok 2009. Značení je následující: *pobřeží Estonsko* $\hat{\alpha}_1(t)$, *pobřeží Finsko* $\hat{\alpha}_2(t)$, *vnitrozemí Litva* $\hat{\alpha}_3(t)$, *vnitrozemí Finsko* $\hat{\alpha}_4(t)$. 54
3.5 Odhady průběhu funkcí skupinových středních hodnot $u_g(t)$ ve srovnání s odhadem průběhu funkce celkové střední hodnoty, tj. funkce celkového průměru na intervalu [0, 365] pro rok 2009. Tlustou čarou je značen průběh celkového průměru a čárkovanou průběh funkce skupinových průměrů. Jedná se obdobu grafu 3.3 s tím rozdílem, že při vyhlazení původních pozorování byla použita *penalizace za nehladkost*. 55
3.6 Odhady průběhu funkcí efektů $\alpha_1(t)$ a $\alpha_2(t)$ na intervalu [0, 365] pro rok 2009 při různém způsobu vyhlazení původních pozorování průměrných denních teplot. Na levé straně jsou odhadnuté průběhy funkcí efektů z dat vyhlazených bez využití *penalizace za nehladkost*. Na pravé straně jsou vyobrazeny odhady funkcí efektů napočítané ze záznamů průměrných denních teplot vyhlazených s použitím *penalizace za nehladkost* (2.25), s váhou $\lambda = 10\,000$. 56
3.7 Odhady průběhu funkcí efektů desetiletí na intervalu [0, 365]. . . . 57

3.8 Odhady průběhu funkcí skupinových středních hodnot ve srovnání s odhadem průběhu funkce celkové střední hodnoty, tj. funkce celkového průměru na intervalu [0, 365] pro rok 2009. Tlustou čarou je značen průběh celkového průměru a čárkovanou průběh funkce skupinových průměrů. Pro odhad regresních funkcí byl použit *penalizační člen* s *diferenciálním operátorem* a $\lambda = 10^8$ 66
3.9 Odhady průběhu funkcí efektů $\alpha_g(t)$, $g = 1, \ldots, 4$, na intervalu [0, 365] roku 2009. Pro odhad regresních funkcí byl použit *penalizační člen* s *diferenciálním operátorem* a $\lambda = 10^8$ 67
3.10 Odhady průběhu funkcí skupinových efektů $\alpha_g(t)$ na intervalu [0, 365] pro rok 2009 při různém nastavení parametru λ. Pro odhad regresních funkcí byl použit *penalizační člen* s *lineárním operátorem posunu*. 68
3.11 Histogram záznamů p-hodnot *FP* testu během první simulace. . . 76
3.12 Histogram záznamů p-hodnot *FP* testu během druhé simulace s náhodným rozřazením odhadu funkcí do skupin. 76
3.13 Průběh funkce *bodové F* statistiky $F_n(t)$ (3.15) pro vyhlazená pozorování průměrných denních teplot časové řady *Tallinn*. 77
3.14 Odhady regresních funkcí v modelu *funkcionální analýzy rozptylu* pro vyhlazená pozorování průměrných denních teplot časové řady *Tallinn*. Červeně je zvýrazněné období, ve kterém zamítáme nulovou hypotézu. 78

Literatura

[1] Daily data, In: *European Climate Assessment & Dataset project* [online]. [cit. 2018-12-02]. Dostupné z: http://www.ecad.eu/dailydata/index.php.

[2] ECA&D flyer, In: *ECA&D* [online]. [cit. 2018-12-02]. Dostupné z: http://www.ecad.eu/documents/ECAD_flyer.pdf.

[3] Download predefined subsets in ASCII, In: *ECA&D* [online]. [cit. 2018-12-02]. Dostupné z: http://www.ecad.eu/download/ECA_blend_station_tg.txt.

[4] About us-Eumetnet, In: *Eumetnet* [online]. [cit. 2018-12-02]. Dostupné z: http://eumetnet.eu/about-us/.

[5] European Climate Assessment & Dataset project, In: *ECA&D* [online]. [cit. 2018-12-02]. Dostupné z: http://www.ecad.eu/http://www.ecad.eu/.

[6] Terms and conditions of use, In: *Data Policy for ECA&D and E-OBS* [online]. [cit. 2018-12-02]. Dostupné z: http://www.ecad.eu/documents/ECAD_datapolicy.pdf.

[7] Praha Klementinum, In: *Portál ČHMÚ* [online]. [cit. 2018-12-02]. Dostupné z: http://portal.chmi.cz/historicka-data/pocasi/praha-klementinum.

[8] Časová analýza klementinské teplotní a srážkové řady pomocí Palmerova indexu závažnosti sucha , In: *http://www.amet.cz* [online]. [cit. 2018-12-02]. Dostupné z: http://www.amet.cz/klemcz.htm.

[9] Historická data - meteorologie a klimatologie, In: *Portál ČHMÚ* [online]. [cit. 2018-12-02]. Dostupné z: http://portal.chmi.cz/historicka-data/pocasi/zakladni-informace.

[10] Praha Klementinum, In: *Portál ČHMÚ* [online]. [cit. 2018-12-02]. Dostupné z: http://portal.chmi.cz/historicka-data/pocasi/praha-klementinum.

[11] Coordinate Reference Systems, In: *Spatial Data Analysis and Modeling with R* [online]. [cit. 2018-12-02]. Dostupné z: http://rspatial.org/spatial/rst/6-crs.html.

[12] Elements, In: *ECA&D* [online]. [cit. 2018-12-02]. Dostupné z: https://www.ecad.eu/dailydata/datadictionaryelement.php.

[13] Get API Key, In: *Google Maps APIs* [online]. [cit. 2018-10-04]. Dostupné z: https://developers.google.com/maps/documentation/javascript/get-api-key.

[14] Učíme se Python, In: *python.cz* [online]. [cit. 2018-16-04]. Dostupné z: https://python.cz/zacatecnici/.

[15] Začátečnický kurz, In: *Nauč se Python!* [online]. [cit. 2018-16-04]. Dostupné z: http://naucse.python.cz/course/pyladies/.

[16] Atom Editor, In: *Atom* [online]. [cit. 2018-16-04]. Dostupné z: https://atom.io/.

[17] Ramsay, James, Silverman, B. W. From functional data to smooth functions. In: *Functional Data Analysis* (2. vydání). Springer-Verlag New York, 2005. Stránky kapitoly od 54-75. ISBN: 978-0-387-40080-8.

[18] Ramsay, James, Silverman, B. W. Smoothing functional data by least squares. In: *Functional Data Analysis* (2. vydání). Springer-Verlag New York, 2005. Stránky kapitoly od 76-96. ISBN: 978-0-387-40080-8.

[19] Ramsay, James, Silverman, B. W. Smoothing functional data with a roughness penalty. In: *Functional Data Analysis* (2. vydání). Springer-Verlag New York, 2005. Stránky kapitoly od 97-126. ISBN: 978-0-387-40080-8.

[20] Hladíková Hana, Aproximace funkcí. In: *Katedra statistiky a pravděpodobnosti Fakulty informatiky a statistiky, Vysoká škola ekonomická v Praze* [online]. [cit. 2018-16-04]. Dostupné z: http://nb.vse.cz/~stepkova/cf/k05_interpol.pdf.

[21] Ramsay, James, Silverman, B. W. Modelling functional responses with multivariate covariates. In: *Functional Data Analysis* (2. vydání). Springer-Verlag New York, 2005. Stránky kapitoly od 234-256. ISBN: 978-0-387-40080-8.

[22] Ramsay, James, Silverman, B. W. Using the Kronecker product to express B. In: *Functional Data Analysis* (2. vydání). Springer-Verlag New York, 2005. Stránky podkapitoly od 248-250. ISBN: 978-0-387-40080-8.

[23] Górecki, Tomasz, Smaga, Lukasz. Analysis of variance for functional data. In: *fdANOVA: An R Software Package for Analysis of Variance for Univariate and Multivariate Functional Data* [online]. [cit. 2018-16-04]. Dostupné z: https://cran.r-project.org/web/packages/fdANOVA/vignettes/fdANOVA.pdf.

[24] Ramsay, James O., Hooker, Giles, Graves, Spencer. Functional Data Analysis with R and MATLAB, In: *books.google* [online]. [cit. 2018-16-04]. Dostupné z: https://books.google.cz/books?id=fNKHa8eV7WYC&pg=PA63&lpg=PA63&dq=a+harmonic+acceleration+operator&source=bl&ots=yDT33ieLzo&sig=0Ex4mHZbqt9BXCKAgWV8cBrwp9g&hl=cs&sa=X&ved=0ahUKEwj7-N3M0oXaAhVBZ1AKHVegAfkQ6AEIJzAA#v=onepage&q=a%20harmonic%20acceleration%20operator&f=false.

[25] Define a Linear Differential Operator Object, In: *Rdocumentation* [online]. [cit. 2018-16-04]. Dostupné z: https://www.rdocumentation.org/packages/fda/versions/2.4.7/topics/Lfd.

[26] Šuláková, Monika. Permutační testy, In: *Neparametrické metody* [online]. [cit. 2018-16-04]. Dostupné z: https://dokupdf.com/download/7-neparametricke-metody-nonparametric-methods-_5a0344b7d64ab2b9bdf79507_pdf.