UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Na COVID-19 s log-podíly

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **prof. RNDr. Karel Hron, Ph.D.**
Vypracoval(a): **Adéla Czolková**
Studijní program: B0541A170010 Matematika a její aplikace
Studijní obor Matematika a její aplikace
Forma studia: prezenční
Rok odevzdání: 2022

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Adéla Czolková

Název práce: Na COVID-19 s log-podíly

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: prof. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2022

Abstrakt: V souvislosti s pandemickou situací kolem onemocnění COVID-19 jsme zahlceni množstvím dat, která ze své povahy vyžadují vedle zaměření se na absolutní informaci v nich obsaženou, např. počty osob v kategoriích nakažení/hospitalizovaní/zemřelí, také analýzu (log-)podílů mezi jednotlivými kategoriemi. Cílem bakalářské práce je nalézt zajímavé souvislosti právě užitím relativní informace obsažené v tomto typu dat, které nebyla věnována taková pozornost jako absolutní informaci.

Klíčová slova: COVID-19, kompoziční data, log-podíly

Počet stran: 50

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Adéla Czolková

Title: Analyzing COVID-19 using log-ratios

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: prof. RNDr. Karel Hron, Ph.D.

The year of presentation: 2022

Abstract: In connection with the pandemic situation around COVID-19, we are overwhelmed by the amount of data that require in addition to focusing on the absolute information contained in them (numbers of people in categories infected/hospitalized/dead) also analyzing log-ratios between categories. The aim of the thesis is to find interesting relationships using the relative information contained in this type of data, to which was not given such attention as to the absolute information.

Key words: COVID-19, compositional data, log-ratios

Number of pages: 50

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením pana prof. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Poděkování

Ráda bych poděkovala prof. RNDr. Karlu Hronovi, Ph.D. za odborné vedení a pomoc při zpracování této bakalářské práce. Děkuji také doc. RNDr. Jitce Machalové, Ph.D. za cenné rady ohledně splajnů.

Úvod

Jak již název napovídá, tato bakalářská práce se zabývá analýzou dat souvisejících s onemocněním COVID-19 a zaměřuje se na relativní informaci obsaženou v těchto datech.

COVID-19 asi není potřeba nějak blíže představovat, všichni o něm mnoho slyšeli a mnozí ho i prodělali. Každý z nás pandemii nějak prožíval a na základě svých zkušeností si vytvářel vlastní názory na celou situaci spojenou s tímto onemocněním. Je tedy jasné, že se v některých otázkách týkajících se pandemie všichni neshodneme.

Když jsem přemýšlela nad tématem své bakalářské práce, napadlo mě, že bych mohla využít situace a zabývat se daty získanými během pandemie, proto jsem si nakonec vybrala právě toto téma. Cílem mé práce bylo zaměřit se na data jiným způsobem, než jsme dosud měli možnost vidět, a pracovat s relativní informací, kterou tato data obsahují. Mým záměrem nebylo přijít s nějakým novým objevem nebo názorem, ale jen ukázat, jak jinak se na data můžeme dívat a jaké to přináší výsledky.

Kapitola 1

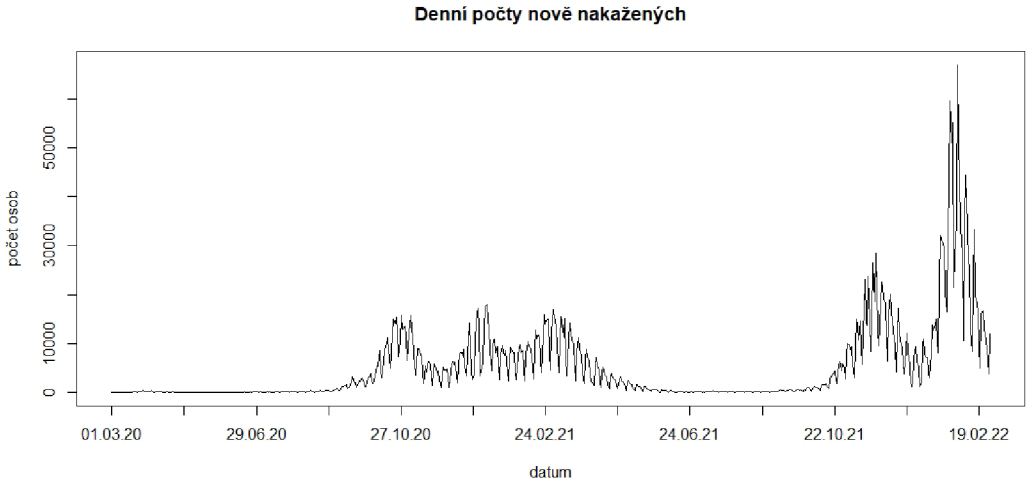
COVID-19

V prvních měsících roku 2020 začala celosvětová pandemie koronaviru SARS-CoV-2, který způsobuje onemocnění COVID-19. Jedná se o infekční respirační onemocnění, jehož typickými příznaky jsou dráždivý kašel, ztráta chuti a čichu, horečka, bolesti hlavy a svalů nebo únava, u novějších variant viru je častým projevem onemocnění i rýma. Může také dojít k postižení dalších orgánů, jako jsou například plíce nebo srdce. Virus se přenáší vzdušnou cestou kapénkami, a to hlavně při kýchání nebo kašli.

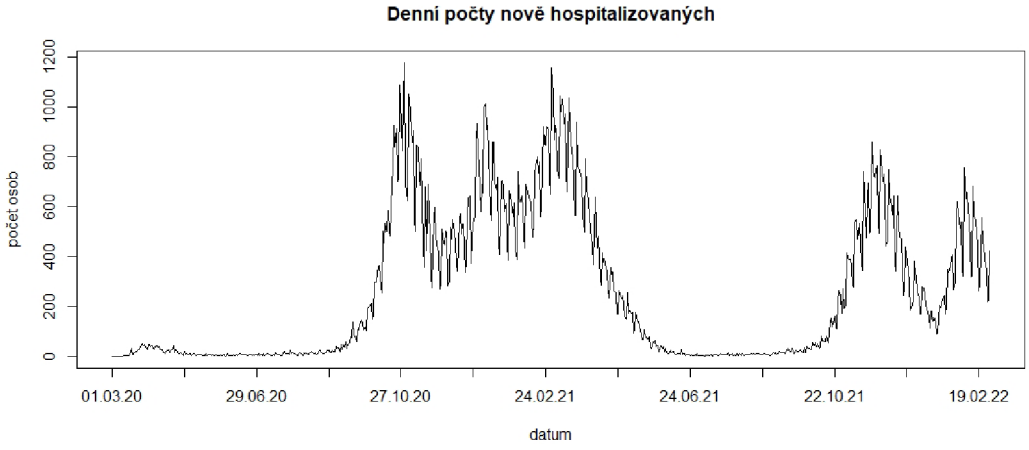
Od počátku pandemie jsme zahlcováni velkým množstvím nejrůznějších dat týkajících se šíření tohoto koronaviru. Česká data jsou denně zveřejňována na webových stránkách Ministerstva zdravotnictví ČR [2], jejich sběrem a zpracováním se zabývá Ústav zdravotnických informací a statistiky ČR.

Ve většině případů jsou prezentována data, která nesou absolutní informaci. Jsou to hlavně počty nově nakažených, hospitalizovaných, vyléčených či zemřelých. Z těchto čísel si rychle uděláme představu o tom, jak se pandemie vyvíjí, jak rychle se nákaza šíří, zda přibývá hospitalizovaných pacientů a podobně. Můžeme se dívat jak na celkové počty, tj. počty od začátku pandemie, tak na denní počty. Najdeme také data týkající se jednotlivých krajů, okresů nebo obcí. U hospitalizovaných pacientů jsou uváděny rovněž i počty osob na jednotkách intenzivní péče nebo počty potřebných dýchacích přístrojů. Zjistíme také, kolik bylo každý den provedeno testů a jakých. V průběhu pandemie přibyla i data o vakcinaci, tedy kolik bylo využito vakcín, kolik je očkovaných a kolika dávkami. V posledních

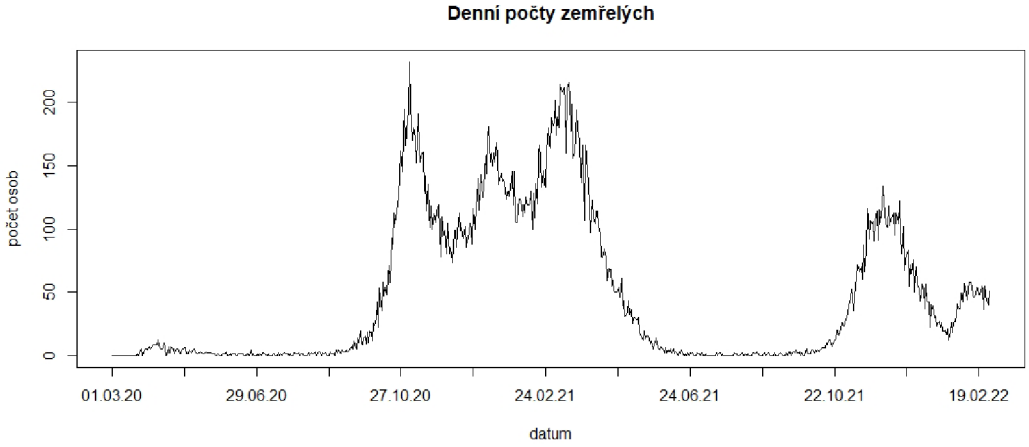
měsících se objevují také informace o počtu reinfekcí, kterých začalo postupně přibývat.



Obrázek 1.1: Vývoj denních počtů nově nakažených v období od 1. 3. 2020 do 28. 2. 2022.



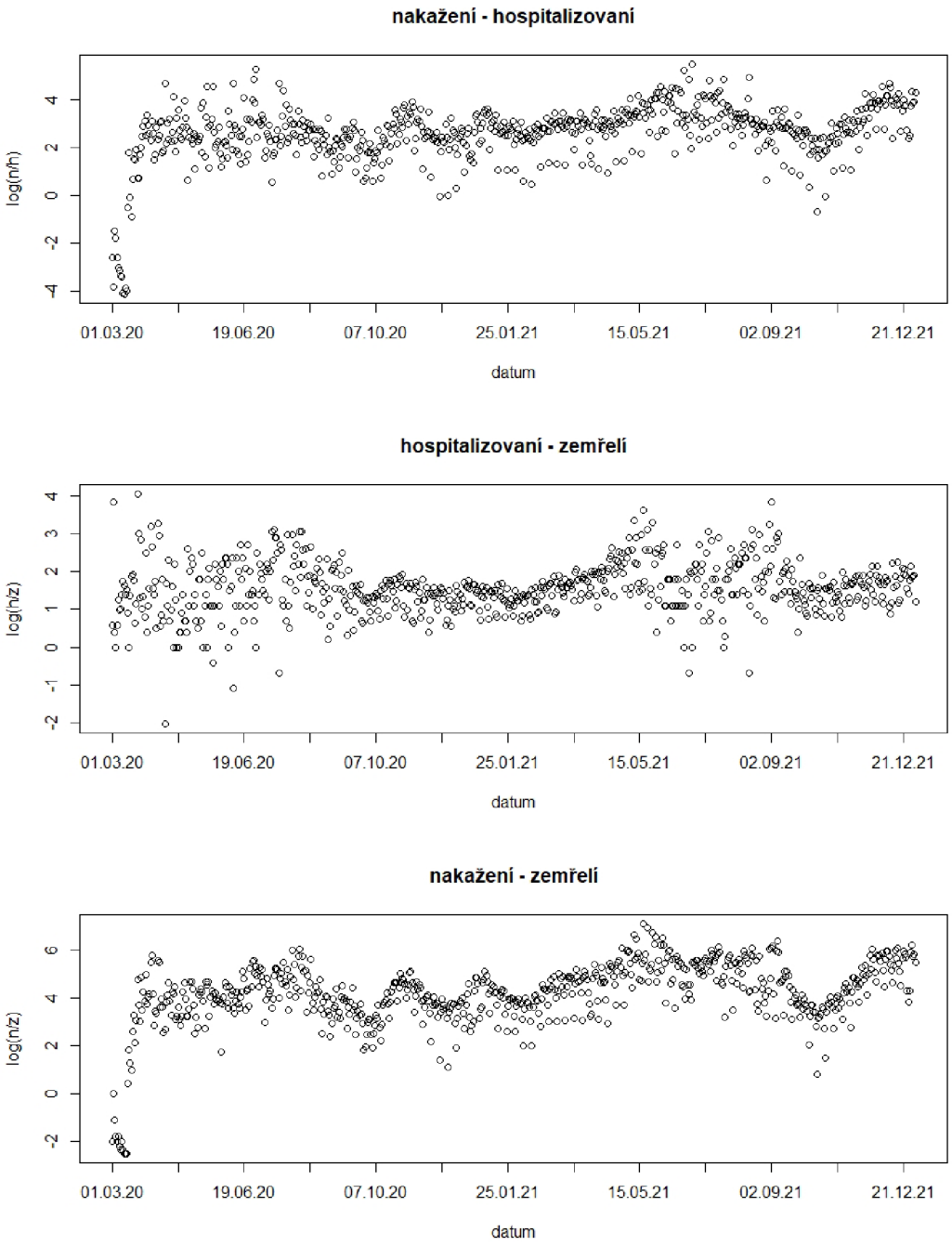
Obrázek 1.2: Vývoj denních počtů nově hospitalizovaných v období od 1. 3. 2020 do 28. 2. 2022.



Obrázek 1.3: Vývoj denních počtů zemřelých v období od 1. 3. 2020 do 28. 2. 2022.

Zajímavé výsledky nám však poskytuje také relativní informace obsažená v datech, i když jí nebyl věnován takový prostor jako absolutní informaci. Může nás například zajímat, jaká část pozitivně testovaných osob byla hospitalizována nebo zemřela, kolik hospitalizovaných pacientů je v těžkém stavu, kolik procent testů vyšlo s pozitivním výsledkem, jaký je podíl reinfekcí mezi nově nakaženými. Zde však musíme počítat s určitou proměnlivostí dat. Třeba v případě, kdy nás zajímá, jaká část nakažených, kteří byli pozitivně testováni v určitém období (například v konkrétním týdnu), byla hospitalizována, nelze jednoznačně určit, po jaké době od nakažení (nebo pozitivního testu) jsou nemocní hospitalizováni. Dále také nevíme, jak dlouho jsou hospitalizováni a za jak dlouho případně umírají. Musí nám proto stačit průměrné doby, a i ty se v průběhu pandemie měnily [5]. Musíme se tedy spokojit s tím, že budeme pracovat pouze s přibližnými hodnotami. Této problematice je věnována i část třetí kapitoly této práce.

V datech se také projevuje vývoj koronaviru, který má vliv i na výše popsanou proměnlivost dat. V průběhu pandemie bylo zaznamenáno několik jeho mutací, které měly vzájemně různé některé vlastnosti. Každá varianta viru se šířila jinou rychlostí a měla vliv i na průběh nemoci. Zpočátku se virus nešířil nějak rychle, ale v mnoha případech způsoboval poměrně těžký průběh onemocnění, a to hlavně



Obrázek 1.4: Relativní informaci můžeme získat z dat například pomocí log-podílů, kterými se zabývají další části této práce. Výpočet těchto konkrétních log-podílů bude popsán v druhé části třetí kapitoly tohoto textu.

u starších a dlouhodobě nemocných osob. Naopak jedna z nejnovějších variant – omikron – se šíří poměrně rychle, ale průběh nemoci je u většiny nakažených lehký.

Na vývoj pandemie může mít vliv také očkování, které sice před nákazou úplně neochrání, ale ve většině případů brání těžkému průběhu onemocnění. Častou otázkou je i to, jak fungují protilátky vytvořené po prodělání nemoci a zda ovlivňují průběh onemocnění při případné reinfekci.

Dalším faktorem ovlivňujícím pandemii jsou vládní protiepidemická opatření, která byla velmi často měněna a upravována a je i otázkou, jak velký byl skutečně jejich vliv, tímto se ale tato práce zabývat nebude.

Za zmínku ale stojí například vývoj testovací strategie. Zpočátku se provádělo poměrně málo testů, a to hlavně v případě podezření na nákazu. Postupně se však začalo i s preventivním testováním, a kromě PCR testů se začaly používat i testy antigenní, u kterých ale byla uváděna menší spolehlivost. Později se ve způsobu testování projevila i očkovací strategie, když se přestalo s preventivním testováním očkovaných osob. Testováni byli tedy hlavně neočkovaní, kteří se na rozdíl od očkovaných museli prokazovat negativním testem například v restauracích a na hromadných akcích. Nakonec se však od antigenních testů začalo upouštět a byly opět preferovány PCR testy.

Kromě testování můžeme zmínit třeba základní pravidla platící téměř po celou dobu pandemie – nošení roušek nebo respirátorů, dezinfekce rukou, dodržování dvoumetrových rozestupů, minimalizace kontaktů. Dalšími opatřeními byla například omezení ve školách nebo službách, a i jejich úplné uzavření, omezení kapacit na hromadných akcích nebo v obchodech. Asi nejpřísnějším opatřením bylo uzavření okresů na jaře 2021.

Můžeme se zamyslet rovněž nad tím, jaký vliv má na šíření viru třeba roční období nebo teplota vzduchu. Mohli bychom se domnívat, že COVID-19 je sezónním onemocněním podobně jako například chřipka, ale těžko říct, zda tato závislost nemoci na ročním období či počasí není pouze zdánlivá. Této závislosti se budeme věnovat na konci třetí kapitoly.

Kapitola 2

Kompoziční data

Data nesoucí relativní informaci se nazývají **kompoziční data** [3, 6, 9]. Ta tak obvykle popisují části nějakého celku, většinou jako vektory proporcí nebo procent. ***D*-složkovou kompozicí** nazýváme vektor $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, kde $x_1, x_2, \ldots, x_D$ jsou reálná kladná čísla. Relativní informace je obsažena v poměrech mezi složkami dané kompozice a součty hodnot jednotlivých složek pro nás nemají význam.

Většina kompozic má konstantní součet $\kappa > 0$, často $\kappa = 1$ pro proporce nebo $\kappa = 100$ pro data vyjádřená v procentech. Jednotky, ve kterých jsou data vyjádřena, můžeme libovolně měnit, protože vynásobení kompozice kladnou konstantou nezmění poměry mezi složkami. Každou kompozici $\mathbf{x} = [x_1, x_2, \ldots, x_D] \in \mathbb{R}_+^D$ tedy můžeme přeškálovat, aby měla požadovaný součet složek κ (například chceme-li kompozici vyjádřit v proporcích s konstantním součtem $\kappa = 1$), a to tak, že

$$\mathcal{C}(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right],$$

kde $\mathcal{C}(\mathbf{x})$ se nazývá **uzávěr**. Množinou všech kompozic s konstantním součtem κ (libovolně zvoleným, ale pevným) je **simplex**

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \mid x_i > 0,\ i = 1, 2, \ldots, D;\ \sum_{i=1}^{D} x_i = \kappa \right\},$$

což je vlastně $(D-1)$-dimenzionální množina v $\mathbb{R}^D$.

Někdy nás mohou zajímat pouze některé složky kompozice. V tomto případě můžeme pracovat jen se **subkompozicí**, tedy vektorem $\mathbf{x}_S = [x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$, kde $S = \{i_1, i_2, \ldots, i_s\}$ je množina indexů vybraných složek.

Je také možné sečíst vybrané složky do jedné. Pro množinu vybraných indexů $A = \{i_1, \ldots, i_a\}$, $D - a \geq 1$ vypočítáme hodnotu

$$x_A = \sum_{j \in A} x_{i_j},$$

nová kompozice je pak ve tvaru $\mathbf{x}' = [\mathbf{x}_{\bar{A}}, x_A]$, kde $\bar{A}$ je množina zbývajících indexů. Jedná se o kompozici v $\mathcal{S}^{D-a+1}$ a tato operace se nazývá **amalgamace**.

Nejčastěji se pracuje s trojsložkovou kompozicí, kterou lze dobře zobrazit. Odpovídající simplex je reprezentován trojúhelníkem v $\mathbb{R}^3$ s vrcholy $[\kappa, 0, 0]$, $[0, \kappa, 0]$, $[0, 0, \kappa]$. Ekvivalentní reprezentací je také **ternární diagram**, což je rovnostranný trojúhelník, kde pro kompozici $\mathbf{x} = [x_1, x_2, x_3]$ je x_i vzdálenost od protější strany i-tého vrcholu pro $i = 1, 2, 3$. Tento diagram je vhodný pro interpretaci výsledků.

2.1 Log-podíly

Pro analýzu kompozičních dat jsou vhodné například **log-podíly**, které mají několik dobrých vlastností. Pro libovolnou kompozici $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ totiž platí:

$$\ln \frac{x_i}{x_j} = \ln \frac{\lambda \cdot x_i}{\lambda \cdot x_j}, \; \forall i, j = 1, 2, \ldots, D, \; i \neq j, \; 0 < \lambda \in \mathbb{R},$$

$$\ln \frac{x_i}{x_j} = -\ln \frac{x_j}{x_i}, \; \forall i, j = 1, 2, \ldots, D, \; i \neq j.$$

Z první rovnosti plyne, že hodnota log-podílu se nezmění při vynásobení kompozice reálnou kladnou konstantou, a proto nezávisí na jednotkách, ve kterých je kompozice dána. Z druhé rovnosti pak vidíme, že při záměně pořadí složek kompozice se změní pouze znaménko log-podílu, tedy můžeme říct, že jeho hodnota nezávisí až na znaménko na pořadí složek kompozice.

Při používání log-podílů však nesmíme zapomínat na to, že nemůžeme počítat s nulami, protože $\ln 0$ není definován. Předpokládá se sice, že všechny složky kom-

pozice jsou kladné, ale v reálných datech se často mohou objevovat i nuly. Někdy se také může stát, že některé hodnoty jsou velmi malé, a tak jsou považovány za nuly. To se jedná o hodnoty pod detekčním limitem měřícího přístroje, ovšem nuly mohou také vznikat (jako u některých pozorování v našem případě) jako důsledek celkově malých četností v kompozičním vektoru. Ani v jedné z obou situací nelze tyto nuly považovat za strukturní, tedy např. v druhém případě nelze vyloučit, že při větším rozsahu souboru by byly nulové četnosti eliminovány. Abychom mohli počítat log-podíly, musíme tyto hodnoty vhodně upravit. Nejjednodušší úpravou je (mimo případné amalgamace složek) jejich nahrazení malou hodnotou, která nebude považována za nulu, například 2/3, nebo 1/2.

2.2 Souřadnice kompozice

Protože jsou kompoziční data reprezentována pomocí vektorů, mohlo by nás napadnout pracovat s nimi jako vektory v reálném prostoru tak, jak jsme zvyklí, tedy jako s prvky euklidovského vektorového prostoru. Ale ukazuje se, že euklidovská geometrie není pro analýzu kompozičních dat vhodná, protože dobře nepopisuje relativní rozdíly mezi kompozicemi. Například rozdíl mezi $[10, 60, 30]$ a $[20, 50, 30]$ není stejný jako mezi $[20, 50, 30]$ a $[30, 40, 30]$. V prvním případě se první složka zdvojnásobila, zatímco v druhém případě se zvýšila pouze o polovinu. Euklidovská vzdálenost mezi dvěma kompozicemi je však v obou případech stejná, neboť rozdíl mezi první a druhou složkou je 10 jednotek u obou dvojic.

Potřebujeme tedy jinou geometrii, která bude vhodná pro práci s kompozičními daty. Budeme proto pracovat v simplexu, kde je definována **Aitchisonova geometrie**, jejíž struktura odpovídá struktuře euklidovského vektorového prostoru. Základními operacemi jsou **perturbace**

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_D y_D] \in \mathcal{S}^D \text{ pro } \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$$

a **mocnění**

$$\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha] \in \mathcal{S}^D \text{ pro } \mathbf{x} \in \mathcal{S}^D, \alpha \in \mathbb{R}.$$

Simplex s těmito operacemi $\left(\mathcal{S}^D, \oplus, \odot\right)$ tvoří vektorový prostor.

Další operací, kterou budeme při práci s kompozičními daty potřebovat, je **perturbační rozdíl**

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1},$$

kde $\mathbf{y}^{-1} = \mathcal{C}\left[y_1^{-1}, y_2^{-1}, \ldots, y_D^{-1}\right]$ je inverzní kompozice k $\mathbf{y}$.

Abychom získali strukturu euklidovského vektorového prostoru, musíme definovat ještě další tři operace:

- Aitchisonův skalární součin

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \text{ pro } \mathbf{x}, \mathbf{y} \in \mathcal{S}^D,$$

- Aitchisonovu normu

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j}\right)^2} \text{ pro } \mathbf{x} \in \mathcal{S}^D,$$

- Aitchisonovu vzdálenost

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2}.$$

Chceme-li analyzovat kompoziční data pomocí standardních mnohorozměrných statistických metod, potřebujeme je vyjádřit ve vhodných reálných souřadnicích, protože tyto metody jsou definovány ve standardním reálném euklidovském prostoru.

Existuje několik souřadnicových reprezentací, které převádějí kompozice do reálného prostoru. Jsou to například:

- alr souřadnice (additive log-ratio coordinates, aditivní log-podílové souřadnice)

$$\mathrm{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right),$$

- clr koeficienty (centred log-ratio coefficients, centrované log-podílové koeficienty)

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right).$$

Nejlepší je však vyjádřit kompozici pomocí $D-1$ reálných souřadnic v ortonormální bázi $\mathbb{R}^{D-1}$, a to pomocí ilr souřadnic (isometric log-ratio coordinates, izometrické log-podílové souřadnice). Jejím speciálním případem jsou **bilance**, které jsou výsledkem postupného sekvenčního binárního dělení kompozice. Dělení probíhá tak, že se nejprve složky celé kompozice rozdělí do dvou skupin, každá skupina se pak dále dělí na dvě a celý proces takto pokračuje, dokud není každá skupina tvořena pouze jednou složkou kompozice. V i-tém kroku dělení vždy spočítáme jednu souřadnici a po dokončení dělení tak máme $D-1$ souřadnic v ortonormální bázi vzhledem k Aitchisonově geometrii, kde i-tá souřadnice je ve tvaru

$$z_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln \frac{\sqrt[r_i]{x_{i_1} \cdots x_{i_{r_i}}}}{\sqrt[s_i]{x_{j_1} \cdots x_{j_{s_i}}}} = \frac{1}{\sqrt{r_i s_i (r_i + s_i)}} \sum_{k=1}^{r_i} \sum_{l=1}^{s_i} \ln \frac{x_{i_k}}{x_{j_l}},$$

$i = 1, \ldots, D-1$ a r_i, s_i jsou počty složek ve dvou vytvořených skupinách v i-tém kroku dělení. Každá souřadnice navíc obsahuje (agreguje) všechny párové log-podíly mezi složkami z obou skupin. Známe-li souřadnice kompozice, můžeme s ní pracovat jako s vektorem v $\mathbb{R}^{D-1}$.

Zatímco k analýze kompozičních dat je vhodnější reálný prostor, pro interpretaci výsledků je lepší simplex. Data tedy nejprve vyjádříme ve vhodných reálných souřadnicích, provedeme potřebné výpočty a výsledky poté převedeme pomocí odpovídajícího inverzního zobrazení zpět do simplexu, abychom je mohli lépe interpretovat.

2.3 Centrování v ternárním diagramu

Jak už bylo řečeno, nejčastěji se pracuje s kompozicemi, které mají tři složky. Tyto kompozice se dají dobře zobrazit v ternárním diagramu. V některých přípa-

dech se však zobrazená data mohou nacházet blízko hranice ternárního diagramu, a to může způsobovat problémy s interpretací, protože obraz u hranice je pokřivený vlivem variability relativní informace v obsažené kompozici . Abychom se těmto problémům vyhnuli a data lépe zobrazili, provedeme centrování, což znamená, že zobrazená kompoziční data vhodným způsobem přesuneme k těžišti diagramu tak, aby se kompoziční centrum, které vypočítáme pomocí geometrických průměrů pro každou složku, nacházelo ve středu diagramu jako neutrální prvek vzhledem k Aitchisonově geometrii.

Centrování tedy provedeme pomocí perturbačního rozdílu s vhodně zvolenou kompozicí $[p_1, p_2, p_3]$, v tomto případě kompozičním centrem. Nová kompozice je pak určena vektorem

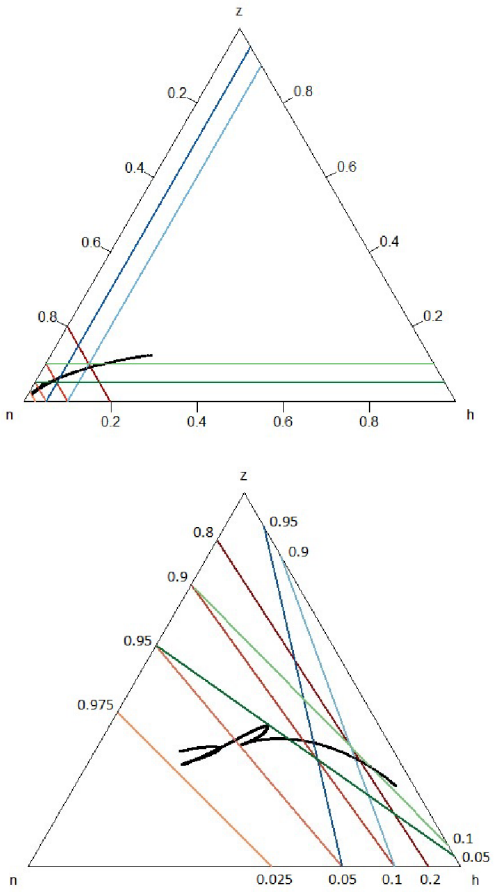
$$[y_1, y_2, y_3] = \mathcal{C}\left[\frac{x_1}{p_1}, \frac{x_2}{p_2}, \frac{x_3}{p_3}\right] \in \mathcal{S}^3,$$

kde $[x_1, x_2, x_3] \in \mathcal{S}^3$ je původní kompozice.

Můžeme perturbovat také souřadnicovou síť, aby byly zachovány původní hodnoty v diagramu. Je dobré si uvědomit, že tato síť je tvořena úsečkami spojujícími dvě strany trojúhelníku a po perturbaci se změní pouze jejich poloha v diagramu – zůstanou úsečkami.

Z obrázku 2.1 je jasné, jak velký vliv má centrování. Z nevycentrovaného ternárního diagramu (nahoře) téměř nic nevyčteme – vidíme, že zobrazená data se nachází v jednom rohu trojúhelníku, ale nejsme schopni je nějak podrobněji popsat. Po vycentrování dat (dole) už je tvar křivky mnohem zřetelnější a můžeme lépe vidět a popsat její chování. Z horního diagramu bychom mohli usoudit, že křivka je čarou spojující dva body, která se nijak nekroutí, ve skutečnosti se však jedná o mnohem komplikovanější křivku.

Ve zobrazených diagramech vidíme také část souřadnicové sítě. Jednotlivé úsečky odpovídají stejným souřadnicím v obou diagramech a jsou barevně rozlišeny, aby bylo dobře vidět, jak se změnila jejich poloha po centrování. Odstíny červené odpovídají hodnotám 0,975, 0,95, 0,9 a 0,8, jsou to souřadnice pro první složku kompozice zde označené jako n. Odstíny modré a zelené reprezentují hod-



Obrázek 2.1: Horní ternární diagram zobrazuje původní kompoziční data (tvořící černou křivku), pod ním je pak diagram znázorňující stejná data po vycentrování. Úsečky odpovídají souřadnicové síti. Tomuto konkrétnímu diagramu a jeho vytváření je věnována druhá část třetí kapitoly této práce.

noty 0,05 a 0,1. Modré úsečky jsou souřadnice pro druhou složku kompozice (h) a zelené pro třetí složku (z). Je zřejmé, že čím světlejší je barva, tím vyšší hodnotě složky úsečka odpovídá.

Kapitola 3

Log-podíly a COVID-19

V této kapitole se podíváme na dva případy, ve kterých lze použít log-podíly k analýze covidových dat. Nejprve se zaměříme pouze na celkové počty nakažených v několika zemích a budeme zkoumat, jak se vyvíjely. Tento přístup je vhodný spíše k analýze kratšího časového úseku (například jednoho nebo dvou měsíců). Dále se budeme zabývat téměř celým dosavadním obdobím pandemie, zaměříme se nejen na počty nakažených, ale i na počty hospitalizovaných nebo zemřelých a na vyhlazování časové řady kompozic. Kromě využití log-podílů si tak v této kapitole také ukážeme různé přístupy k vyhlazování dat – lokální regresi a splajny.

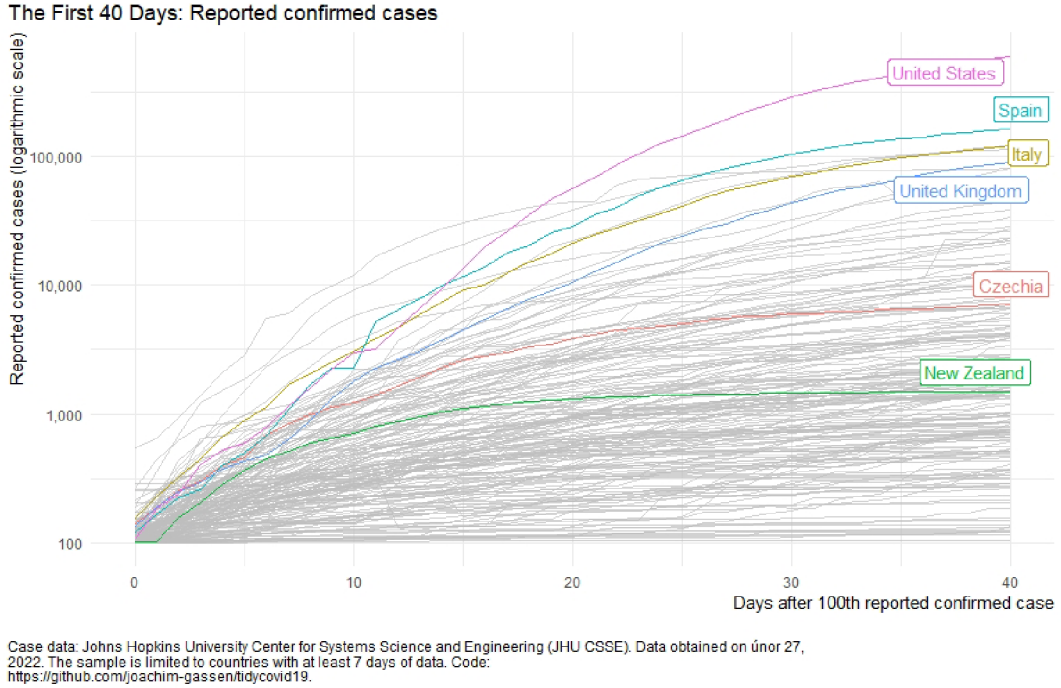
3.1 Časová řada log-podílů

Myšlenka analýzy vývoje celkového počtu nakažených pomocí log-podílů vychází z článku „Why log ratios are useful for tracking COVID-19“[7], který se zaměřuje na porovnávání šíření koronaviru ve vybraných státech na počátku pandemie. Budeme pracovat se stejnými zdroji dat a kódy, jaké jsou uvedeny v tomto článku.

Nejprve je asi důležité zmínit význam logaritmického měřítka grafu, které umožňuje zobrazovat hodnoty v rozpětí mnoha řádů – můžeme tak dobře pozorovat průběh růstu nejen velkých, ale i malých hodnot, jejichž vývoj by byl v grafu s klasickým (nezlogaritmovaným) měřítkem špatně viditelný. V tomto případě se

zaměříme na graf celkového počtu nakažených s logaritmickým měřítkem.

Protože chceme srovnávat počty nakažených v různých zemích, nebudeme se dívat pouze na česká data, ale vybereme si i několik dalších států (Itálii, Španělsko, Velkou Británii, Spojené státy americké a Nový Zéland). Hodnoty z těchto zemí si můžeme zobrazit v grafu s logaritmickým měřítkem.



Obrázek 3.1: Graf zobrazuje vývoj celkového počtu nakažených 40 dní od 100. potvrzeného případu nákazy ve vybraných zemích – České republice, Itálii, Španělsku, Velké Británii, Spojených státech a Novém Zélandu. Byl vytvořen na základě kódu převzatého z [7].

Nyní se nabízí otázka, jak vlastně vývoj pandemie v jednotlivých zemích porovnávat. Mohlo by nás například napadnout, zda by nebylo lepší uvažovat hodnoty na počet obyvatel, protože třeba 100 000 potvrzených případů nákazy bude pro každou zemi znamenat něco jiného (představme si například porovnání České republiky a Spojených států, jejichž počet obyvatel se významně liší). Tato úprava může být vhodná, pokud chceme porovnávat jednotlivé hodnoty mezi sebou, ale pro nějakou podrobnější analýzu vývoje celkového počtu nakažených tato úprava nemá význam. V grafu dojde pouze ke změně polohy křivek (posunutí nahoru,

nebo dolů), ale jejich tvar se nezmění.

Na polohu křivky má také vliv testování, tj. procento odhalených případů. Záleží proto na tom, jak probíhá testování v dané zemi – kolik se provádí testů, kdo je testován a jakými testy. Tvar křivky se však může změnit pouze v případě, že dojde ke změně testovacího režimu v daném státě (to z dat ale nevyčteme). Ukazuje se tedy, že porovnávání poloh křivek není příliš vhodné, protože musíme předpokládat, že se přístup k testování v jednotlivých zemích liší.

Při porovnávání situací v jednotlivých státech, bychom se proto měli zaměřit spíše na tvar křivek. Zajímá-li nás pouze krátký časový úsek pandemie (zvolme si třeba jeden měsíc), můžeme předpokládat, že se testovací režimy v jednotlivých státech nezmění, tudíž tvar křivky bude ovlivněn pouze vývojem počtu nových případů nákazy, ale ne způsobem testování.

Tvar křivky lze popsat pomocí toho, jaký má křivka sklon. Proto se v tomto případě můžeme dívat na log-podíly, neboť sklon křivky v logaritmickém měřítku je totéž co logaritmus podílu po sobě jdoucích dvou hodnot:

$$\ln Y_t - \ln Y_{t-1} = \ln \frac{Y_t}{Y_{t-1}},$$

kde Y_t, Y_{t-1} jsou počty nově nakažených ve dnech t, $t-1$.

Při počítání log-podílů však nesmíme zapomínat na to, že se v datech mohou objevit problematické hodnoty – nuly, a s těmi se musíme nějakým způsobem vypořádat. Jednou z možností je proložení vypočtených hodnot spojitou křivkou, která odstraní případné nesrovnalosti způsobené výskytem nul.

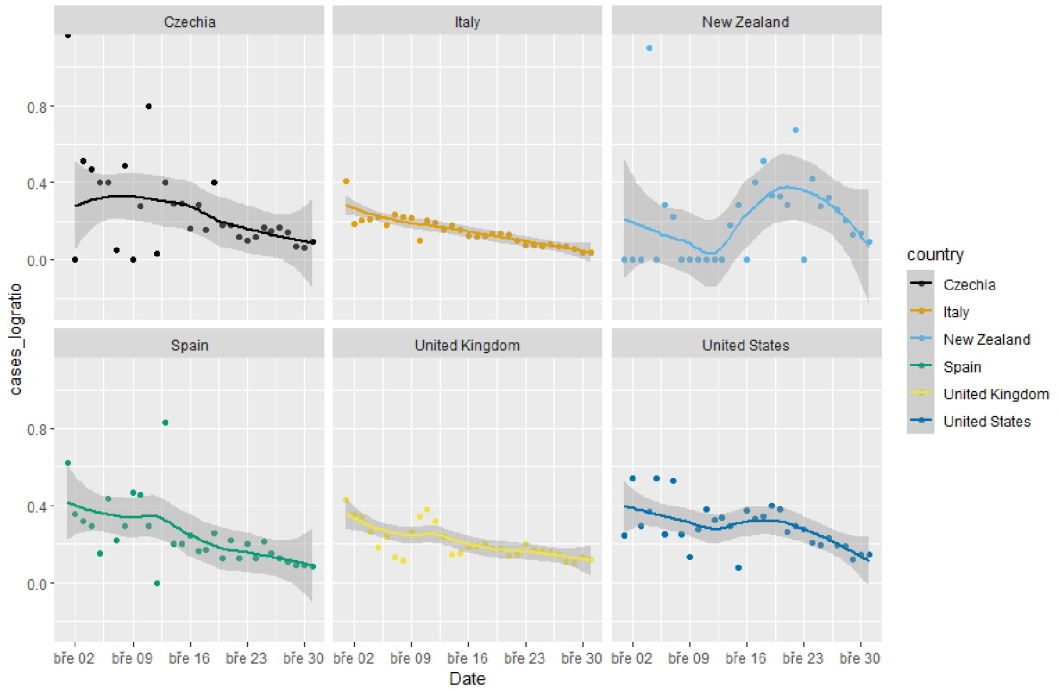
V tomto případě můžeme k vyhlazení dat použít například lokální regresi, která je vhodná pro menší datové soubory a byla použita i ve zmíněném článku [7]. Lokální regrese vychází z postupného počítání odhadů hodnot regresní funkce – ve zvoleném bodě x_0 spočítáme odhadovanou hodnotu pomocí několika okolních bodů, kterým přiřadíme váhy $K_{i0} = K(x_i, x_0)$ tak, že nejbližší body k x_0 mají nejvyšší váhu, zatímco nejvzdálenější body mají váhu nulovou. V bodě x_0 pak spočítáme odhadovanou hodnotu lokální regresní funkce pomocí vážené metody

nejmenších čtverců, tj. hledáme koeficienty $\hat{\beta}_0$ a $\hat{\beta}_1$, které minimalizují výraz

$$\sum_{i=1}^{n} K_{i0} \left(y_i - \beta_0 - \beta_1 x_i\right)^2,$$

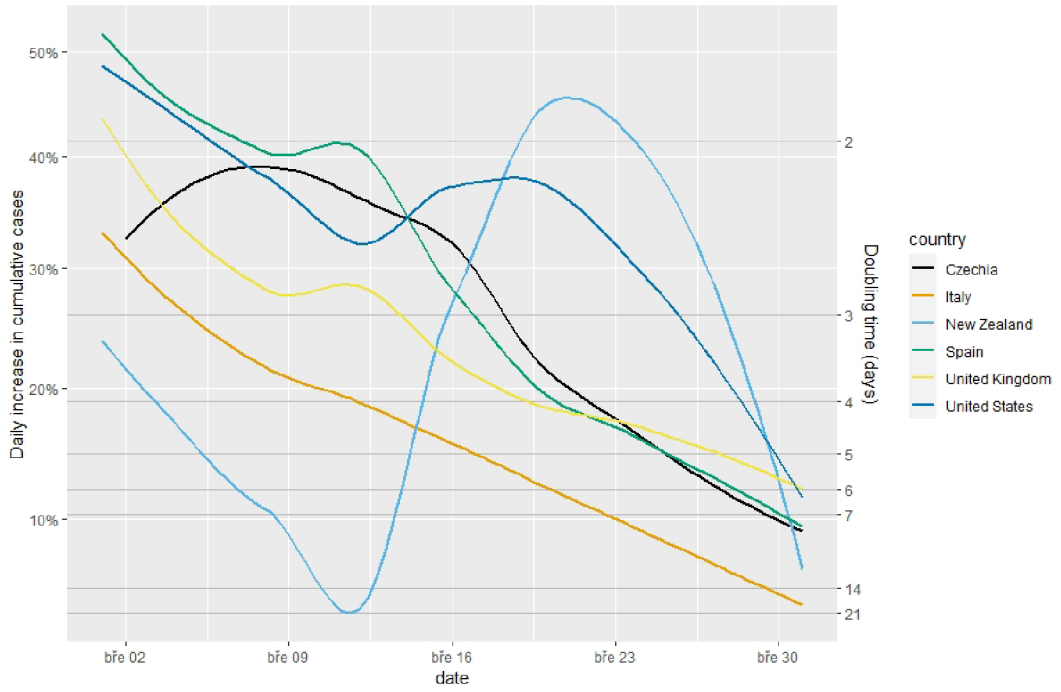
který odpovídá lineární regresi, můžeme však použít i konstantní nebo kvadratickou regresi. Důležitý je také výběr počtu okolních bodů, které pro odhad využijeme – pokud použijeme příliš málo bodů (z malého okolí), pak se výsledná funkce může velmi kroutit a náš odhad bude hodně lokální, naopak pokud vezmeme příliš mnoho bodů (z velkého okolí), dostaneme hodně obecný odhad.

K vytváření následujících grafů byla použita funkce `geom_smooth()` s nastavením `method = "loess"`, která vytvoří nejen výslednou křivku, ale i její bodový interval spolehlivosti, tj. pás spolehlivosti kolem regresní funkce.



Obrázek 3.2: Jednotlivé grafy zobrazují vývoj v březnu 2020 ve vybraných zemích. Kromě hodnot log-podílů jsou zobrazeny také pásy spolehlivosti kolem regresních funkcí. Grafy byly vytvořeny na základě kódu z [7].

Nyní můžeme porovnávat výsledné křivky a za tímto účelem si je zobrazíme do jednoho grafu.



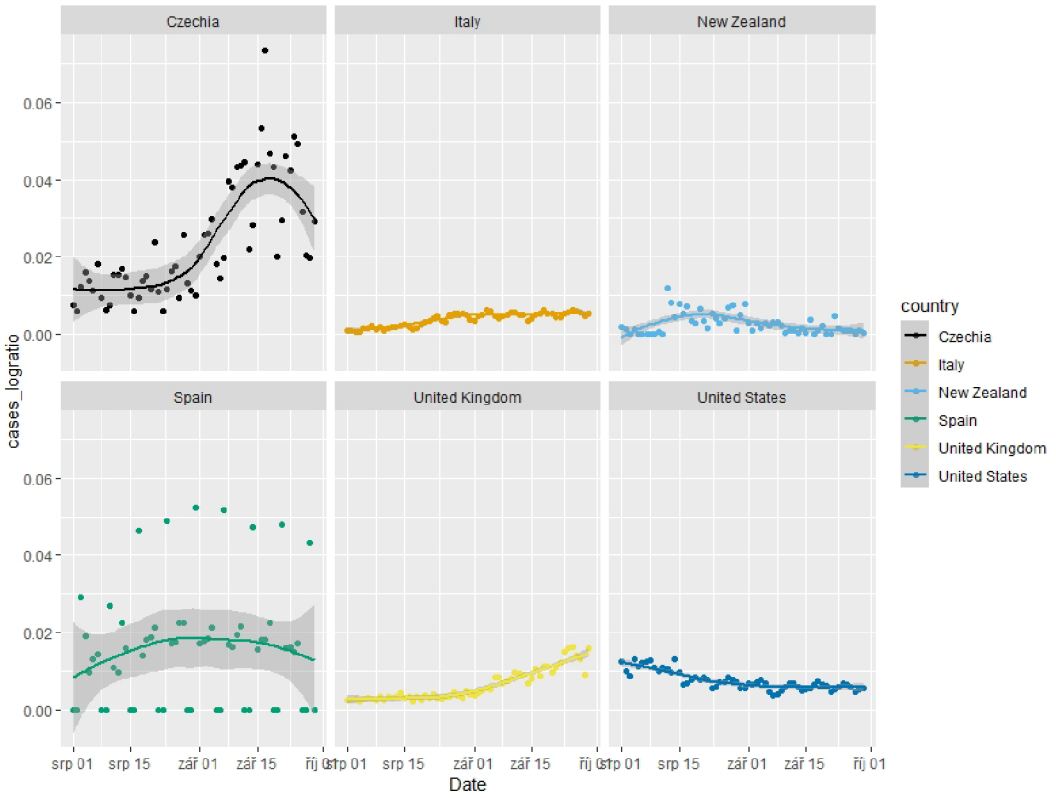
Obrázek 3.3: Zobrazení křivek pro vybrané státy v jednom grafu s upravenými osami pro lepší interpretaci. Graf je opět vytvořen na základě kódu z [7].

V grafu můžeme uvažovat alternativní osy, abychom mohli výsledky pozorování lépe interpretovat – hodnoty log-podílů nejsou pro interpretaci příliš vhodné. První možností je například osa denních přírůstků celkového počtu případů v procentech: $100\,(e^r - 1)$, kde r je hodnota log-podílu. Další možností může být třeba osa s počtem dní, za které se celkový počet případů zdvojnásobí: $\ln(2)/r$.

Je určitě nutné, zamyslet se nad tím, co nám vlastně křivky říkají. Vidíme, že na konci března 2020 již mají všechny klesající tendenci, což znamená, že v té době měly už všechny zkoumané země prvotní nárůst nových případů nákazy za sebou. V případě České republiky víme, že první případy se objevily na počátku března, kdy můžeme pozorovat rostoucí trend křivky. Naopak třeba v Itálii se nákaza začala šířit již v únoru, takže březnové přírůstky už nebyly z tohoto pohledu tak velké, to však ale neznamená, že situace v zemi byla nějak dobrá. Stačí si jen vzpomenout, jak rychle se virus rozšířil, a došlo tak k přetížení nemocnic. Je tedy jasné, že z grafů nevyčteme nic o tom, jak daná země zvládá nárůst nových

případů nákazy, ale jen to, jakou rychlostí se v ní virus šíří. Pokud se šíření viru zrychluje, křivka je rostoucí, naopak v případě zpomalování je křivka klesající.

Nemusíme se samozřejmě zaměřovat pouze na začátek pandemie jako v [7], ale stejný postup můžeme aplikovat i na jiná období v průběhu pandemie. Podívejme se například, jak vypadala situace na přelomu léta a podzimu 2020, kdy začala druhá vlna koronaviru.

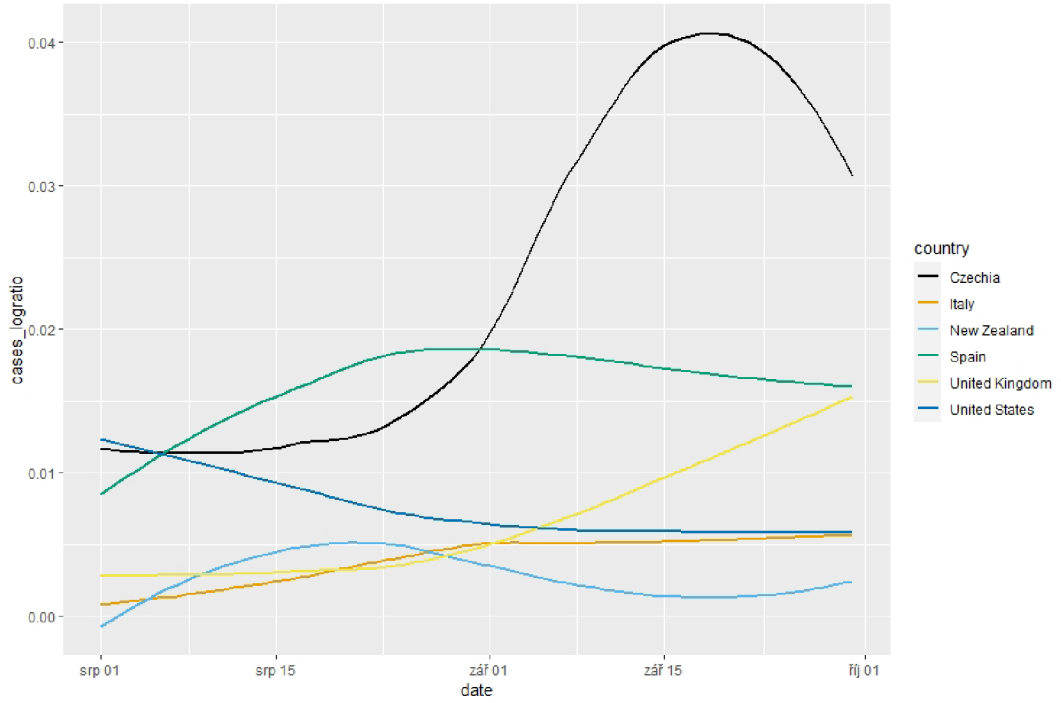


Obrázek 3.4: Grafy zobrazují vyhlazené log-podíly v období od 1. 8. 2020 do 30. 9. 2020. Byly vytvořeny na základě kódu z [7].

V grafech můžeme pozorovat, že se vývoj v jednotlivých zemích výrazně lišil. Je zajímavé, že v Itálii, Velké Británii, Spojených státech a na Novém Zélandu nedocházelo k výraznému nárustu nebo poklesu počtu nakažených ani k výrazným výkyvům hodnot jednotlivých log-podílů. Zatímco ve Španělsku můžeme vidět určitou periodicitu v hodnotách log-podílů, ale po vyhlazení hodnot zjistíme, že také nedocházelo k výrazným změnám v počtu nakažených. Podíváme-li se

blíže na odpovídající data, můžeme si všimnout, že periodicita je pravděpodobně způsobena testovacím režimem – zjistíme například, že nulové hodnoty byly zaznamenány o víkendech, což nemusí nutně znamenat, že neprobíhalo testování, protože se zdá, že víkendové hodnoty byly zapisovány až v pondělí. Naopak situace v České republice byla naprosto odlišná od ostatních zkoumaných zemí. Počet nových případů nákazy začal na konci srpna prudce růst a zpomalil se na konci září, což odpovídá nástupu druhé vlny koronaviru, která přišla právě v září 2020.

Pro lepší srovnání si můžeme křivky opět zobrazit v jednom grafu.



Obrázek 3.5: Zobrazení křivek v jednom grafu nám umožní lépe je porovnat. Můžeme dobře vidět, jak výrazně se liší vývoj v České republice. Opět byl použit kód z [7].

3.2 Kompoziční časová řada

V této části se budeme věnovat práci s kompoziční časovou řadou týkající se covidových dat. Jednotlivé kompozice vytvoříme tak, že rozdělíme denní počet (původně) pozitivně testovaných na tři části, tedy budeme pracovat s trojsložkovými kompozicemi. První složku bude tvořit počet nakažených (včetně reinfekcí) za zvolený den, kteří nebyli ani hospitalizováni, ani nezemřeli – tuto část budeme nazývat nakažení (budou-li v této části textu zmíněni nakažení, bude se jednat právě o tuto složku kompozice, nikoliv o všechny nakažené). Druhou složkou bude počet hospitalizovaných, kteří se vyléčili a nezemřeli – budeme ji nazývat hospitalizovaní. Třetí a poslední složkou pak bude počet zemřelých – zemřelí.

Nyní se musíme zamyslet nad tím, jak vlastně zjistit, kolik osob, které byly pozitivně otestovány daný den, bylo později hospitalizováno, nebo zemřelo. Je asi jasné, že každý z hospitalizovaných pacientů mohl být hospitalizovaný po jinak dlouhé době, která uplynula od pozitivního testu, a totéž platí pro zemřelé. Jak už bylo řečeno v první kapitole, nelze jednoznačně určit, kolik uběhne času od nakažení (nebo pozitivního testu) do hospitalizace a za jak dlouho pak případně pacient umírá. Z tohoto důvodu nám musí stačit pouze přibližné rozdělení denního počtu nakažených na tři složky, které provedeme tak, že použijeme průměrné doby (zpoždění) do vývoje onemocnění s tím, že zanedbáme fakt, že se mohly v průběhu pandemie měnit. Za průměrnou dobu od pozitivního testu do hospitalizace budeme považovat 15 dní a 9 dní jako dobu od začátku hospitalizace po úmrtí pacienta [1]. Dále budeme ještě předpokládat, že ke všem úmrtím došlo v nemocnicích, tj. že všichni zemřelí byli před smrtí hospitalizovaní. Nebudeme také rozlišovat možné situace, kdy byli pacienti hospitalizováni kvůli potížím, které s COVIDem-19 nesouvisely, a byli pozitivně testováni až v nemocnici, nebo příčinou jejich úmrtí nebyl COVID-19, ale zjistilo se, že byli nakažení. Z těchto úvah, je zřejmé, že hodnoty složek kompozice budou pouze přibližné a výsledky analýzy tak nebudou úplně přesné, ale pro účely této práce (a věříme, že pro hrubou představu i obecně) to budeme považovat za dostačující.

Budeme se zabývat obdobím od 1. 3. 2020 do 31. 12. 2021, tedy začneme

dnem, kdy byly v České republice odhaleny první 3 případy nákazy, a skončíme posledním dnem roku 2021. Pro každý den v tomto období vytvoříme kompozici, jak bylo naznačeno již výše. Od počtu pozitivně testovaných (nové případy + reinfekce, tyto počty jsou uváděny zvlášť) za zvolený den z tohoto období odečteme počet nově hospitalizovaných zaznamenaný o 15 dní později a počet zemřelých po 24 dnech, tímto způsobem dostaneme první složku kompozice. Dále od počtu nově hospitalizovaných, který jsme odčítali v předchozím kroku, odečteme také již použitý počet zemřelých, získáme tak druhou složku. Poslední složku už počítat nemusíme, protože ji tvoří přímo počet zemřelých, který jsme doposud odčítali při vytváření přechozích dvou složek.

Abychom se nevěnovali pouze teoretickému popisu tohoto problému, můžeme si ukázat, jak bychom mohli popsanou kompoziční časovou řadu vytvořit pomocí softwaru R. Podívejme se tedy na následující ukázku kódu.

```
# nacteni dat
data1 = read.csv("hospitalizace.csv", header = TRUE, sep = ",")
data2 = read.csv("nakazeni-vyleceni-umrti-testy.csv", header = TRUE, sep = ",")
data3 = read.csv("prehled-reinfekce.csv", header = TRUE, sep = ",")

# uprava dat
data1[is.na(data1)] = 0

pocet = which(data1[,"datum"] == data3[1,1])
reinfekce = c(rep(0,pocet-1),data3[,2])

# oznaceni konce roku 2021
end21 = which(data1[,"datum"]=="2021-12-31")

# vytvoreni vektoru potrebnych hodnot
nove_nakazeni = data2[35:(34+end21),"prirustkovy_pocet_nakazenych"] + reinfekce[1:end21]
nove_hospit = data1[16:(end21+15),"pacient_prvni_zaznam"]
nove_zemreli = data1[25:(end21+24),"umrti"]

# vytvoreni kompozice
kompozice = data.frame(nove_nakazeni - nove_hospit - nove_zemreli,nove_hospit - nove_zemreli,nove_zemreli)
```

```
colnames(kompozice) = c("nakazeni","hospitalizovani","zemreli")
```

Nejprve načteme data ze souborů, které jsou volně ke stažení na webových stránkách Ministerstva zdravotnictví ČR [2]. První soubor obsahuje data týkající se počtů hospitalizovaných, vážnosti jejich stavu, počtů potřebných přístrojů a počtů zemřelých. Druhý soubor se týká denních i celkových počtů nakažených, vyléčených, zemřelých a počtů provedených testů. V třetím souboru jsou pak uvedeny denní počty reinfekcí. Načtená data si můžeme zobrazit a dle potřeby upravit – doplníme chybějící nuly, aby se nám v datech neobjevovaly hodnoty `NA`, a vytvoříme si vektor reinfekcí. Protože se první reinfekce vyskytla až 8. 7. 2020, hodnoty pro předchozí dny doplníme v tomto vektoru nulami. V dalším kroku si označíme konec roku 2021, abychom snadno mohli vybrat odpovídající data. Dále si pak vytvoříme tři vektory, které budou obsahovat potřebné počty všech nově nakažených, hospitalizovaných a zemřelých. Všimněme si, že můžeme vektory vytvořit vhodně tak, abychom je od sebe mohli snadno odečítat, což provedeme hned v následujícím kroku. Vytvoříme si tabulku, jejíž tři sloupce budou odpovídat složkám výše popsané kompozice, v řádcích tak budou kompozice odpovídající jednotlivým dnům sledovaného období. Na závěr ještě může upravit názvy sloupců dle potřeby – pojmenujeme je podle jednotlivých složek.

Jakmile máme kompozice vytvořené, přejdeme k jejich analýze. Jak již bylo řečeno v předchozí kapitole, při práci s kompozičními daty se často používají log-podíly. Právě z tohoto důvodu musíme zajistit, aby všechny složky kompozic byly kladné. Snadno si všimneme, že ve vytvořené tabulce se nacházejí nuly, a dokonce i záporné hodnoty. Nuly jsou hlavně ve třetím sloupci (zemřelí), protože v některých dnech nikdo nemoci nepodlehl. Záporné hodnoty se pak nachází v prvních dvou sloupcích (nakažení, hospitalizovaní) a jsou důsledkem toho, že pracujeme s průměrnými dobami od pozitivního testu do hospitalizace a úmrtí. Tyto nulové a záporné hodnoty nahradíme hodnotou 2/3, což je úprava zmíněná v předchozí kapitole. Tuto úpravu provedeme snadno i v `R`.

```
# nahrazeni zapornych a nulovych hodnot hodnotou 2/3
for (i in 1:dim(kompozice)[1]){
        if (kompozice$nakazeni[i] <= 0){
                kompozice$nakazeni[i] = 2/3
        }
        if (kompozice$hospitalizovani[i] <= 0){
                kompozice$hospitalizovani[i] = 2/3
        }
        if (kompozice$zemreli[i] <= 0){
                kompozice$zemreli[i] = 2/3
        }
}
```

Po úpravě hodnot už nám nic nebrání v použití log-podílů, protože jsou hodnoty všech složek kladné. Spočítáme tedy log-podíly pro jednotlivé dvojice složek, tj. pro dvojice nakažení a hospitalizovaní, nakažení a zemřelí, hospitalizovaní a zemřelí. V R si můžeme vytvořit tabulku s výslednými hodnotami, kterou pak budeme dále používat.

```
# vypocet log-podilu
logpodily = data.frame(log(kompozice$nakazeni/kompozice$
    hospitalizovani),log(kompozice$hospitalizovani/kompozice$
    zemreli),log(kompozice$nakazeni/kompozice$zemreli))
colnames(logpodily) = c("nh","hz","nz")
```

Výslednou časovou řadu log-podílů můžeme vyhladit pomocí splajnů. **Splajny** jsou křivky, které slouží k vyhlazování dat $[x_i, y_i], i = 1, 2, \ldots, n$ [8]. Křivku konstruujeme tak, že si interval $[a, b]$, na kterém ji chceme zkonstruovat, rozdělíme na podintervaly $[\xi_j, \xi_{j+1}], j = 0, 1, \ldots, K$ a křivku konstruujeme po částech. Dělící body $a = \xi_0 < \xi_1 < \cdots < \xi_{K+1} = b$ se nazývají uzly. Je také možné zvolit uzly jako $\xi_i = x_i,\ i = 1, 2, \ldots, n$. Existuje několik možností, v jakém tvaru můžeme splajn $s_k(x)$ k-tého stupně vyjádřit. Nejznámější jsou **polynomické splajny** (na každém podintervalu aproximujeme data polynomem nejvýše stupně k) nebo **B-splajny** (konstruujeme s využitím bázových polynomických funkcí, typicky pomocí několika uzlů). Další z možností je například

$$s^k(x) = \sum_{i=0}^{k} b_i x^i + \sum_{j=1}^{K} c_j (x - \xi_j)_+^k,$$

kde $b_i, i = 0, 1, \ldots, k,\ c_j, j = 1, 2, \ldots, K$ jsou koeficienty a platí

$$(x - \xi_j)_+^k = \begin{cases} (x - \xi_j)^k, & \text{pokud } x > \xi_j, \\ 0, & \text{jinak.} \end{cases}$$

Polynomy, které se vyskytují u jednotlivých sčítanců, tvoří v tomto případě speciální splajnovou bázi.

Při konstrukci **vyhlazovacího splajnu** požadujeme, aby $s_k(x)$ minimalizoval výraz

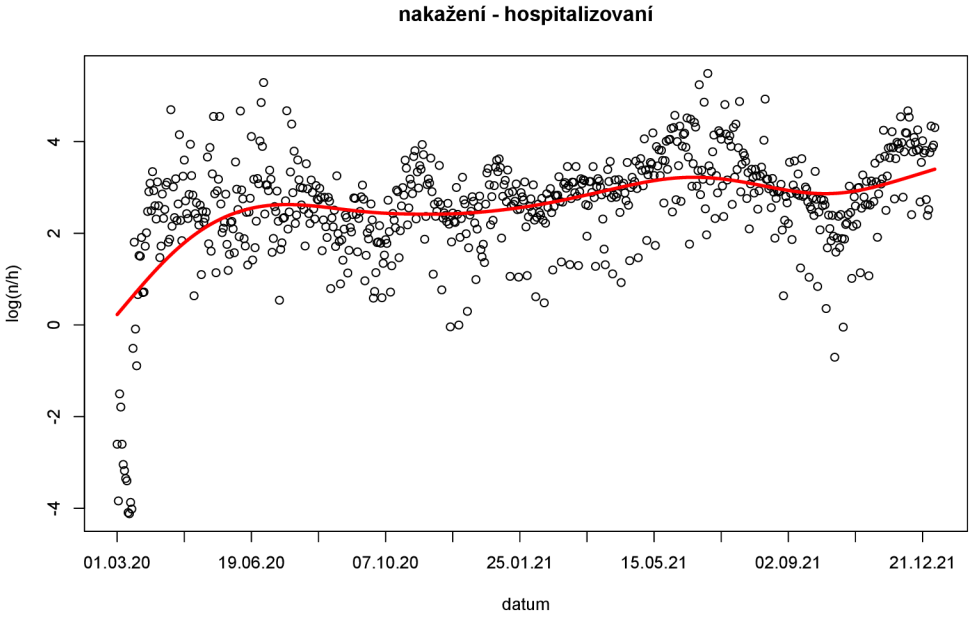
$$\sum_{i=1}^{n} [y_i - s_k(x_i)]^2 + \lambda \int_a^b [s_k''(x)]^2 dx,$$

kde $\lambda > 0$ je vyhlazovací parametr, který určuje míru hladkosti křivky. Čím je λ menší, tím více se bude křivka kroutit, a naopak, čím je λ větší, tím hladší bude křivka. Výraz $\sum_{i=1}^{n} [y_i - s_k(x_i)]^2$ určuje, jak dobře křivka aproximuje zadaná data, a výraz $\int [s_k''(x)]^2 dx$ zajišťuje hladkost křivky. Hodnoty splajnu v bodech $x_1, x_2, \ldots, x_n$ můžeme zapsat pomocí vektoru $\hat{\mathbf{s}} = \mathbf{S}_\lambda \mathbf{y}$, kde $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ a $\mathbf{S}_\lambda$ je matice odpovídající konkrétní volbě parametru λ.

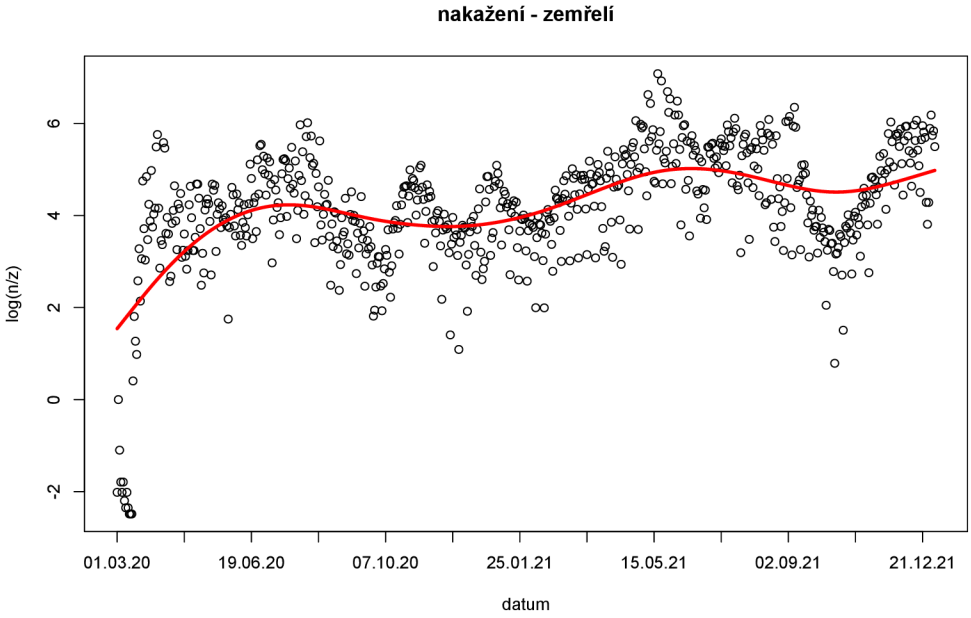
V `R` použijeme k vyhlazení dat pomocí splajnů funkci `smooth.spline()`, ve které můžeme určit míru vyhlazení pomocí parametru `df`, což jsou efektivní stupně volnosti, pro které platí $df = \sum_{i=1}^{n} (\mathbf{S}_\lambda)_{ii}$ a $df \in \langle 2, n \rangle \subset \mathbb{R}$. Čím menší zvolíme df, tím více vyhladíme naše data. Existují různé postupy, jak získat optimální hodnotu df (nebo λ), tím se však zabývat nebudeme, podrobnosti nalezneme například v [8].

Nyní se opět podívejme na kód, který je velmi jednoduchý. Zvolíme si $df = 6$, abychom dosáhli co nejlepšího vyhlazení časové řady log-podílů (nebudeme se tedy zabývat tím, jaká by byla optimální volba). Později si ukážeme, jak by se projevila volba vyšší hodnoty df.

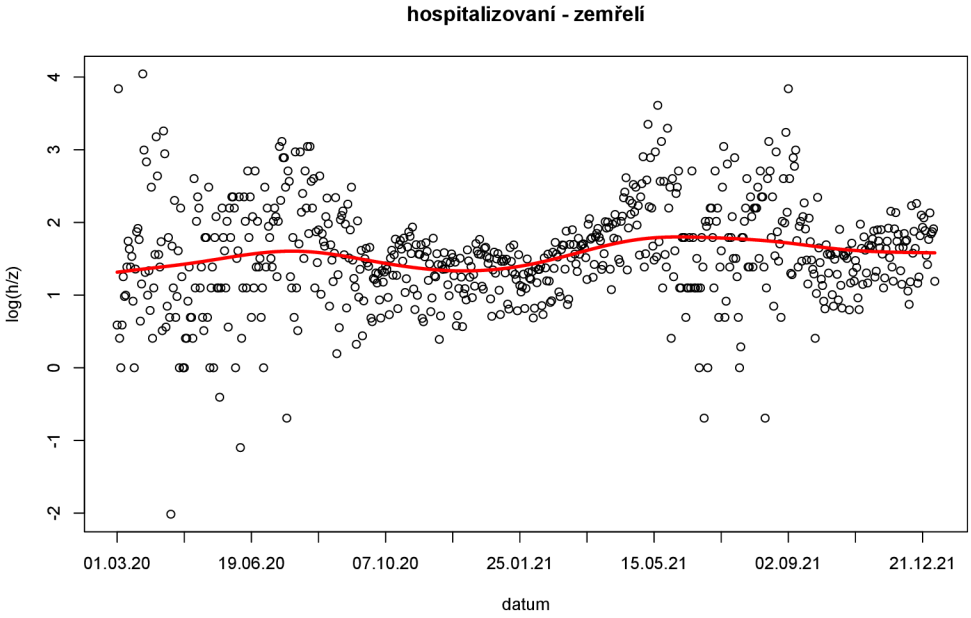
```
# vyhlazeni log-podilu pomoci splajnu
fit_nh = smooth.spline(logpodily$nh, df = 6)
fit_hz = smooth.spline(logpodily$hz, df = 6)
fit_nz = smooth.spline(logpodily$nz, df = 6)
```



Obrázek 3.6: Log-podíly nakažených a hospitalizovaných. Čím větší je hodnota log-podílu, tím více převažuje počet nakažených nad počtem hospitalizovaných.



Obrázek 3.7: Log-podíly nakažených a zemřelých. Čím větší je hodnota log-podílu, tím více převažuje počet nakažených nad počtem zemřelých.



Obrázek 3.8: Log-podíly hospitalizovaných a zemřelých. Čím větší je hodnota log-podílu, tím více převažuje počet hospitalizovaných nad počtem zemřelých.

Grafy (obrázky 3.6, 3.7, 3.8) zobrazují vypočítané log-podíly a jejich vyhlazení pomocí splajnů (červené křivky). Grafy samotných log-podílů jsou právě grafy, které jsme viděli již v první kapitole. V grafech s nakaženými můžeme vidět, že všechny hodnoty log-podílů na začátku března 2020 jsou záporné, je to způsobeno malým počtem pozitivních testů. Podívejme se například na kompozici odpovídající 1. 3. 2020, kdy byli pozitivně testováni tři lidé. Za 15 dní (16. 3.) bylo nově hospitalizováno 14 osob a za dalších 9 dní (25. 3.) zemřelo 5 pacientů. Po odečtení hodnot pak získáme zápornou hodnotu složky nakažení a tu nahradíme hodnotou 2/3. Je tedy jasné, jak vznikají odlehlé záporné hodnoty log-podílů. Ve všech třech následujících grafech vidíme pokles hodnot v podzimních a zimních měsících, kdy přicházely jednotlivé vlny koronaviru, a může to naznačovat příklon k těžšímu průběhu onemocnění. Naopak v létě pozorujeme o něco vyšší hodnoty log-podílů. V grafu pro dvojici hospitalizovaní-zemřelí (obrázek 3.8) je také dobře vidět změny rozptylu hodnot. Velký rozptyl mají hodnoty odpovídající obdobím, kdy byly počty hospitalizovaných i zemřelých poměrně malé,

takže se ve výsledných log-podílech projevily i malé změny poměru mezi hospitalizovanými a zemřelými, tj. v této situaci má například vliv, jestli v daný den někdo zemřel, nebo ne. Obecně ale můžeme říct, že se poměr mezi těmito dvěma složkami v průběhu sledovaného období výrazně neměnil, což vidíme z chování vyhlazovacího splajnu.

Podívejme se také na to, jak by se výše uvedené grafy vytvářely v R. Kromě toho můžeme vidět, jak je možné upravit formát, ve kterém je datum.

```
# uprava formatu pro datum
datum = format(as.Date(data1[1:end21,"datum"]),"%d.%m.%y")

# grafy
plot(logpodily$nh, main = "nakazeni - hospitalizovani", xlab =
    "datum", ylab = "log(n/h)", xaxt = "none")
axis(side = 1, at = seq(1,end21,55), labels = datum[seq(1,
   end21,55)])
lines(fit_nh, col = "red", lwd = 3)

plot(logpodily$hz, main = "hospitalizovani - zemreli", xlab =
   "datum", ylab = "log(h/z)", xaxt = "none")
axis(side = 1, at = seq(1,end21,55), labels = datum[seq(1,
   end21,55)])
lines(fit_hz, col = "red", lwd = 3)

plot(logpodily$nz, main = "nakazeni - zemreli", xlab = "datum"
   , ylab = "log(n/z)", xaxt = "none")
axis(side = 1, at = seq(1,end21,55), labels = datum[seq(1,
   end21,55)])
lines(fit_nz, col = "red", lwd = 3)
```

Dále už nebudeme pracovat s původními log-podíly, ale s jejich vyhlazenými hodnotami. Ty si v R můžeme uložit do nově vytvořených proměnných.

```
# ulozeni hodnot splajnu (vyhlazenych log-podilu)
log12 = fit_nh$y
log13 = fit_nz$y
log23 = fit_hz$y
```

Vyhlazené hodnoty můžeme vyjádřit v ortonormálních souřadnicích v $\mathbb{R}^2$ a poté určíme jejich výsledné souřadnice v ternárním diagramu – simplexu $\mathcal{S}^3$. Zde je dobré poznamenat, že je také možné nejprve vyjádřit původní nevyhlazené

hodnoty log-podílů v ortonormálních souřadnicích v $\mathbb{R}^2$, pak provést vyhlazování pomocí splajnů (vytváříme pouze dva splajny) a nakonec určit souřadnice vyhlazených hodnot v ternárním diagramu. Nezáleží na tom, jaký postup zvolíme, neboť v obou případech získáme stejné souřadnice v $\mathcal{S}^3$, pro vyhlazené hodnoty totiž platí

$$\mathbf{S}_\lambda \ln \frac{x_1}{x_2} + \mathbf{S}_\lambda \ln \frac{x_1}{x_3} = \mathbf{S}_\lambda \left(\ln \frac{x_1}{x_2} + \ln \frac{x_1}{x_3} \right),$$

kde $\mathbf{S}_\lambda$ je matice zmíněná při popisu vyhlazovacích splajnů a x_1, x_2, x_3 jsou složky kompozice.

Za tímto účelem vezmeme konkrétní volbu bilancí

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}} = \frac{\sqrt{2}}{\sqrt{3}} \cdot \frac{1}{2} \left(\ln \frac{x_1}{x_2} + \ln \frac{x_1}{x_3} \right),$$

$$z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3},$$

a potom využijeme jejich inverzní zobrazení zpět na simplex

$$x_1 = \exp\left(\frac{1}{\sqrt{6}} \left(\sqrt{3} z_1 + z_2 \right) \right),$$

$$x_2 = \exp\left(-\frac{1}{\sqrt{6}} \left(\sqrt{3} z_1 - z_2 \right) \right),$$

$$x_3 = \exp\left(-\frac{2}{\sqrt{6}} z_2 \right),$$

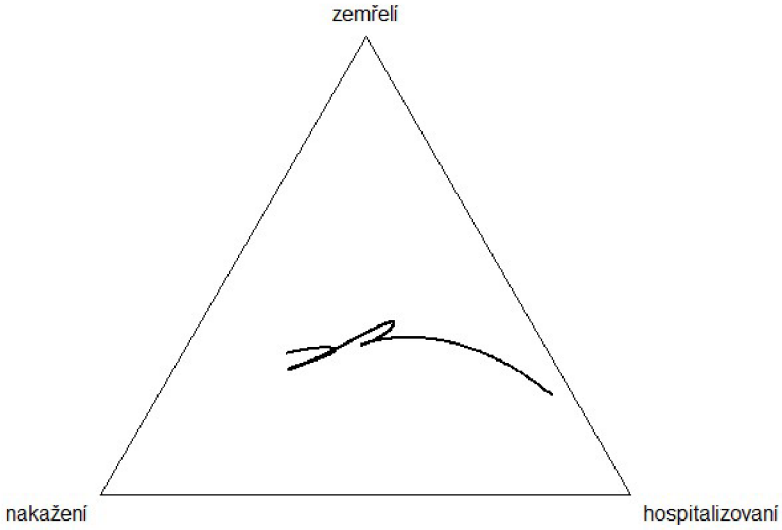
kde x_1, x_2, x_3 jsou složky kompozice – nakažení, hospitalizovaní, zemřelí, a z_1, z_2 jsou reálné souřadnice v $\mathbb{R}^2$. Vidíme, že do prvních dvou vztahů můžeme dosazovat přímo hodnoty log-podílů dvojic složek. V našem případě dosazujeme už vyhlazené hodnoty, které chceme vyjádřit v reálných ilr souřadnicích.

Opět se podívejme, jak bychom postupovali v `R`.

```
# transformace vyhlazenych hodnot
z1 = sqrt(2/3)*(1/2)*(log12+log13)
z2 = sqrt(1/2)*log23

x1 = exp(sqrt(1/6)*(sqrt(3)*z1 + z2))
x2 = exp(-sqrt(1/6)*(sqrt(3)*z1 - z2))
x3 = exp(-sqrt(4/6)*z2)
```

Po transformaci získáme trojice hodnot, které opět tvoří kompozice, a ty si můžeme zobrazit v ternárním diagramu. Nesmíme zapomenout na centrování, abychom mohli výsledný průběh vyhlazené kompoziční časové řady dobře popsat a interpretovat. Jak by vypadal diagram bez centrování bylo ukázáno v poslední části druhé kapitoly. Je třeba také ještě dodat, že se ve skutečnost nejedná o křivku, ale o body reprezentující jednotlivé kompozice, které jsou velmi blízko u sebe, tudíž výsledný obraz v malém měřítku vypadá jako spojitá křivka. Že je tomu opravdu tak, uvidíme později při použití barev.



Obrázek 3.9: Centrovaný ternární diagram zobrazující kompoziční časovou řadu tvořenou kompozicemi se složkami nakažení, hospitalizovaní a zemřelí. Řada začíná vpravo, kde je zobrazena kompozice odpovídající datu 1. 3. 2020.

Na tomto místě je důležité také vysvětlit, co nám vlastně říká poloha kompozice v ternárním diagramu. Jednotlivé vrcholy diagramu odpovídají složkám kompozice. Čím blíže je kompozice k danému vrcholu, tím větší je hodnota této složky

v poměru k ostatním složkám. V našem případě to znamená, že čím blíže jsou body k vrcholu označenému jako nakažení, tím více převažuje složka nakažených nad hospitalizovanými a zemřelými, což znamená, že počty hospitalizovaných a zemřelých jsou velmi malé vzhledem k počtu nakažených, pandemie se tedy vyvíjí dobrým směrem. Opačně to pak platí pro zbývající dva vrcholy – hospitalizovaní, zemřelí, čím blíže k nim jsou kompozice, tím horší je situace, tedy tím větší část lidí, kteří se virem nakazí, musí být hospitalizována, nebo umírá. Podíváme-li se například, jaký vývoj naznačují kompozice odpovídající závěru roku 2021 (levý konec křivky), můžeme říct, že se situace zlepšuje, protože body se přibližují k vrcholu nakažených (viz obrázek 3.9).

Pro zajímavost lze ještě doplnit, že pokud bychom pracovali i s daty ze začátku roku 2022, kdy se rozšířila varianta omikron, doplněné kompozice by se dále přibližovaly k vrcholu nakažených, protože denní počty nově nakažených byly rekordně vysoké, ale na zatížení nemocnic to již příliš vliv nemělo.

Ani v tomto případě nevynecháme ukázku kódu. Vytvoříme si tabulku s kompozicemi a pomocí knihovny `compositions` vytvoříme centrovaný ternární diagram.

```
komp = data.frame(x1,x2,x3)
colnames(komp) = c("nakazeni","hospitalizovani","zemreli")

library(compositions)
plot.acomp(acomp(komp), center = TRUE, pch = ".", cex = 2)
```

Dodejme ještě, že k vytváření ternárního diagramu v `R` lze použít i jiné knihovny, například knihovnu `zCompositions`.

Jak již bylo zmíněno dříve, je možné kompozicím zobrazeným v ternárním diagramu přiřadit barvy v závislosti na nějaké proměnné. Nejprve se zaměříme na to, zda a jaký vliv měla epidemická situace v Česku na poměry mezi jednotlivými složkami. Zajímá nás tedy, jestli je možné říci, že pokud rychle rostl počet pozitivních testů, tak se zvyšoval i poměr pacientů, kteří museli být hospitalizováni, a rostla i úmrtnost. To by znamenalo, že v obdobích, kdy byla situace nejhorší (rychle rostly počty nově odhalených případů), byla pravděpodobnost hospitali-

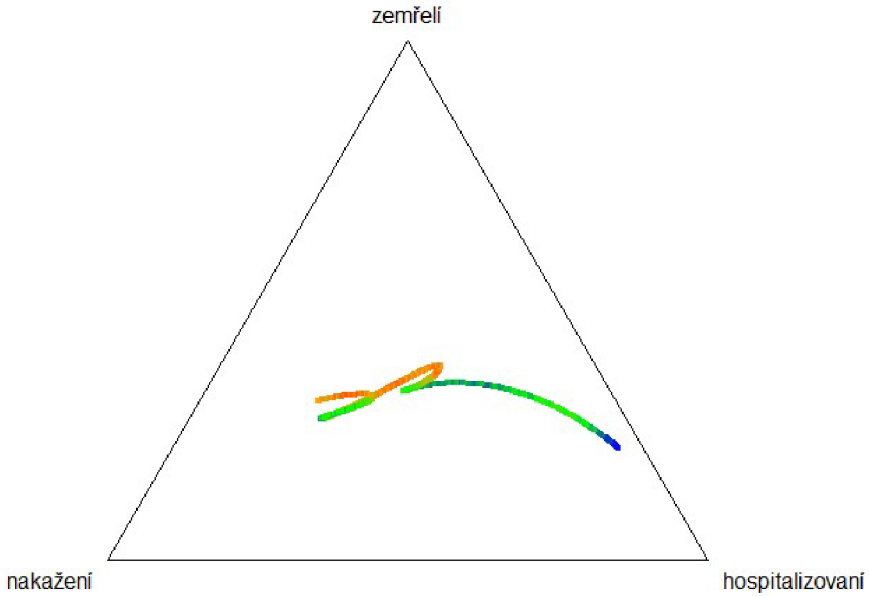
zace nebo úmrtí vyšší než v době, kdy byly denní počty pozitivních testů nízké. Za tímto účelem si vytvoříme barevnou škálu (zvolíme si, že bude obsahovat například 50 barev) pomocí čtyř barev – modrá, zelená, oranžová, červená, kde modrá odpovídá malým přírůstkům nových případů nákazy, a naopak červená odpovídá těm vysokým. Nyní přiřadíme denním počtům všech nově nakažených jednu barvu z vytvořené škály. Abychom dosáhli lepšího rozdělení barev, můžeme si počty nově nakažených logaritmovat. Nyní už jen vytvoříme ternární diagram (stejný jako výše) a nastavíme použití barev.

Opět následuje odbočka ke kódu. Při vytváření barevné škály můžeme použít některou z doplňkových knihoven v `R`, v tomto případě byla využita knihovna `paletteer`. Ternární diagram vytváříme zase pomocí knihovny `compositions` (nemusíme ji načítat znovu, pokud jsme s ní už pracovali).

```
# vytvoreni barevne skaly
library(paletteer)
paleta = colorRampPalette(c("blue","green","orange","red"),
    space = "rgb")
barvy = paleta(50)

# barvy podle poctu vsech nove nakazenych (zlogaritmovanych
    hodnot)
log_nak = ifelse(log(nove_nakazeni) != (-Inf), log(nove_
    nakazeni), 0)
n_rank = as.factor(as.numeric(cut(log_nak, 50)))
plot.acomp(acomp(komp), center = TRUE, pch = ".", cex = 4, col
     = barvy[n_rank])
```

Ve výsledném ternárním diagramu (obrázek 3.10) vidíme, že poměr hospitalizovaných mezi ostatními složkami byl největší na začátku pandemie, kdy bylo prováděno málo testů, a ještě jsme neměli dostatek informací o průběhu nemoci, takže se dá předpokládat, že pro jistotu byli hospitalizováni i lidé s lehčím průběhem. Dále pak rostl poměr nakažených, a jakmile začaly počty všech nově nakažených růst, zvýšil se i poměr hospitalizovaných a zemřelých. Vidíme ale, že když se počty všech nakažených začaly blížit nejvyšším hodnotám zaznamenaným během dané koronavirové vlny, poměr nakažených začal opět růst, což nás může překvapit, protože bychom mohli očekávat, že s rostoucím počtem

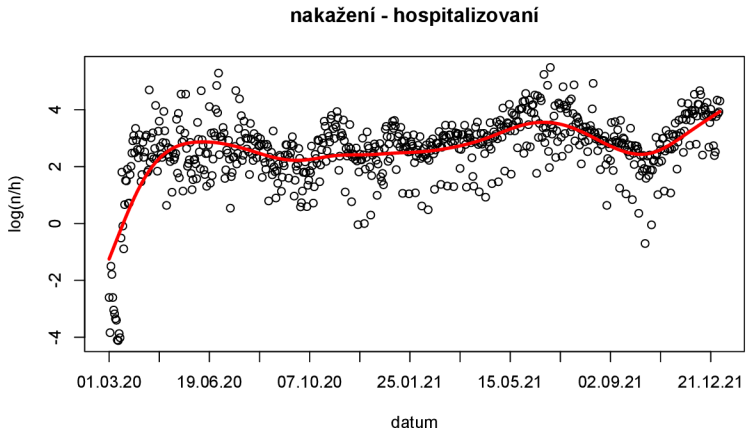


Obrázek 3.10: Ternární diagram, v němž byly jednotlivým kompozicím přiřazeny barvy podle počtu všech nově nakažených za daný den. Barevná škála je poměrně intuitivní – modrá odpovídá nejnižším hodnotám, červená nejvyšším hodnotám.

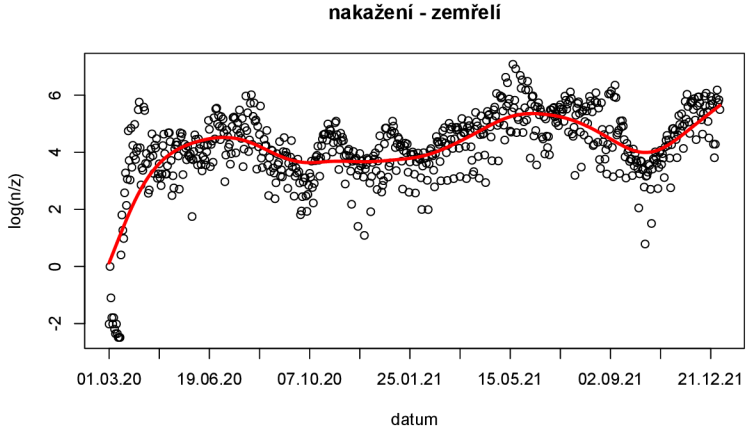
všech nakažených bude růst i poměr hospitalizovaných a zemřelých. To může být způsobeno například testovacím režimem, neboť v obdobích, kdy byly přírůstky malé, se méně testovalo a počty provedených testů se začaly zvyšovat až poté, co se zjistilo, že začala další vlna. Tudíž je možné, že někteří nakažení, kteří neměli příznaky nemoci, nebyli odhaleni, a na to, že se nemoc někde rozšířila se přišlo, až když začalo přibývat nakažených s příznaky nebo těžším průběhem. Všimněme si také, že obecně se poměr nakažených stále zvyšoval a poměr hospitalizovaných a zemřelých v nové vlně nebyl nikdy tak velký jako v té předchozí.

Nyní se podívejme, jak se projeví volba jiné hodnoty df (efektivní stupně volnosti) při konstrukci splajnů. Zvolme si nyní $df = 10$. Vidíme (obrázky 3.11, 3.12, 3.13), že na první pohled se tvar splajnů příliš nezmění, ale ve výsledném ternárním diagramu (obrázek 3.14) vidíme již významnější změny. Můžeme pozorovat stále stejný trend, který jsme viděli i při volbě $df = 6$, ale nyní se projeví i některé lokální efekty v datech, a to je v našem případě spíše nežádoucí, protože

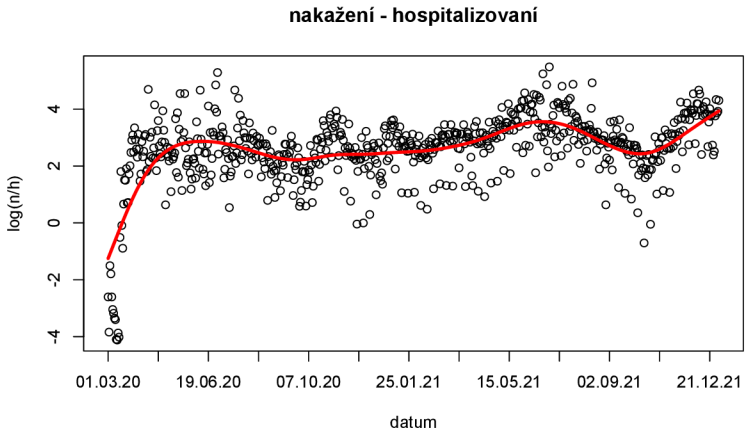
nás zajímá zejména hlavní (obecný) trend ve vytvořené kompoziční časové řadě, a ten není nyní tak dobře viditelný jako v předchozím případě. Budeme proto nadále pracovat se splajny s $df = 6$.



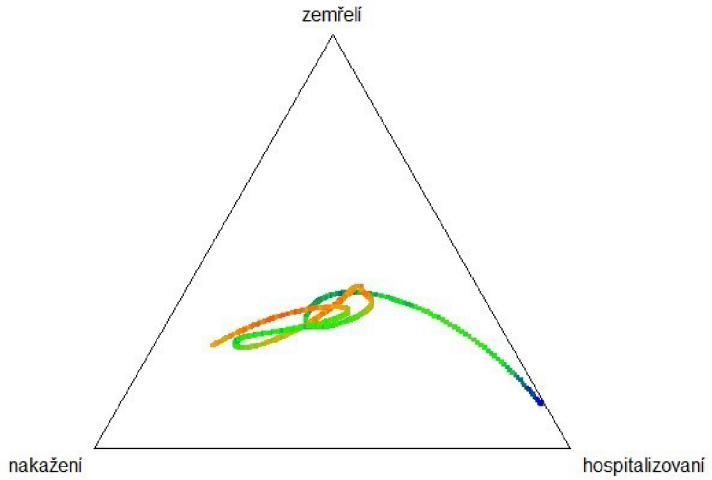
Obrázek 3.11: Vyhlazené log-podíly nakažených a hospitalizovaných pomocí splajnů s volbou $df = 10$.



Obrázek 3.12: Vyhlazené log-podíly nakažených a zemřelých pomocí splajnů s volbou $df = 10$.



Obrázek 3.13: Vyhlazené log-podíly hospitalizovaných a zemřelých pomocí splajnů s volbou $df = 10$.



Obrázek 3.14: Ternární diagram vytvořený pomocí splajnů s volbou $df = 10$ a barvami podle celkových počtů nakažených.

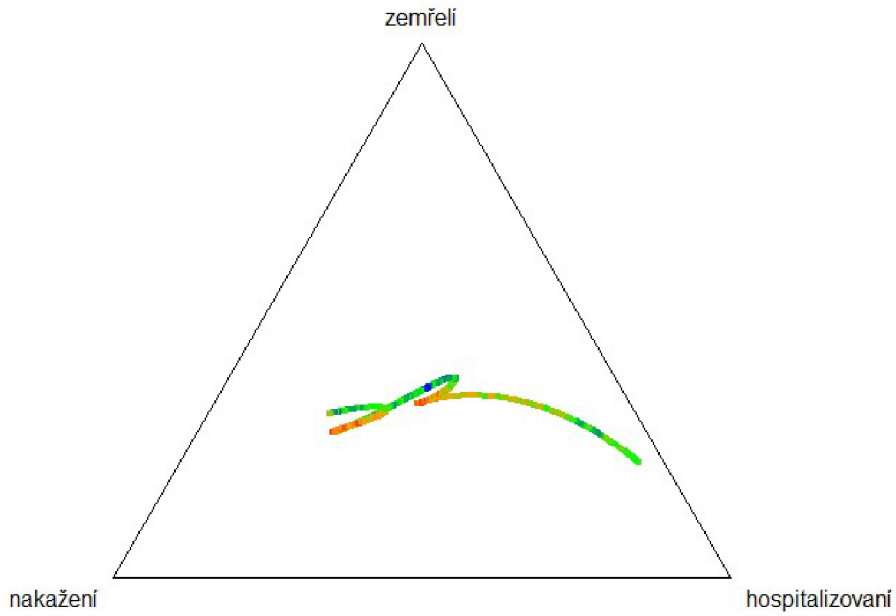
Dále nás také může napadnout, zda existuje nějaká závislost poměrů mezi jednotlivými složkami kompozice na ročním období, respektive na teplotě vzduchu (jak již bylo naznačeno v závěru první kapitoly). Z dat můžeme vyčíst, že nejvíce nových případů nákazy bylo odhaleno většinou v zimních měsících, na jaře se pak situace zlepšovala, v letních měsících byly přírůstky nově nakažených velmi malé a na podzim čísla zase začala rychle stoupat. Můžeme si také všimnout, že v zimě roste relativní počet hospitalizovaných a úmrtí, naopak v létě nebylo zatížení nemocnic tak velké. Podívejme se tedy blíže na to, jak závisí poměry složek na teplotě vzduchu, která byla naměřena v odpovídající den. Poznamenejme ještě, že pracujeme s teplotami ze dne, kdy byli lidé pozitivně testováni, a ne ze dne, kdy se nakazili, ale předpokládáme, že za pár dní se teplota nějak výrazně nezměnila. Také musíme myslet na to, že na různých místech v Česku se teploty často liší. V tomto případě použijeme průměrné denní teploty naměřené v pražském Klementinu, tyto hodnoty jsou k dispozici na webových stránkách Českého hydrometeorologického ústavu [10], můžeme si je tedy snadno stáhnout a nahrát do R. Použijeme stejné barvy jako v předchozím případě, bude se jen lišit jejich přiřazení k jednotlivým kompozicím. Tedy modrá barva odpovídá nejnižším teplotám a červená těm nejvyšším.

```
# barvy podle prumerne teploty v Klementinu
teploty = read.csv2("teploty.csv", header = TRUE, sep = ";")
    [,4]

t_rank = as.factor(as.numeric(cut(teploty,50)))

plot.acomp(acomp(komp), center = TRUE, pch = ".", cex = 4, col
    = barvy[t_rank])
```

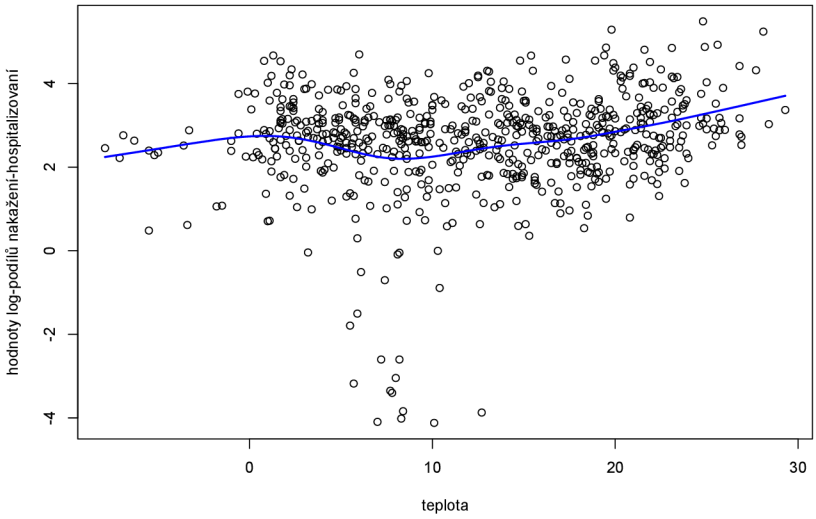
Z výsledného ternárního diagramu (obrázek 3.15) se může zdát, že by mohl existovat nějaký vliv teploty na poměry mezi složkami. Vidíme také, jak se situace vyvíjela v jednotlivých ročních obdobích (nízké teploty znamenají zimu a vysoké léto). Je však otázkou, jak velký je ve skutečnosti vliv teploty – museli bychom zjistit, jestli neexistuje ještě nějaký jiný činitel, která má podobný vliv jako teplota.



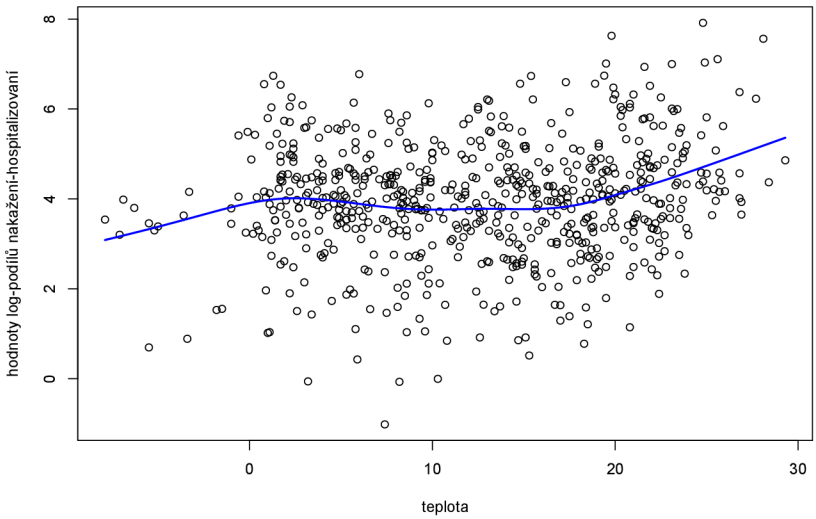
Obrázek 3.15: Ternární diagram, v němž byly jednotlivým kompozicím přiřazeny barvy podle denní průměrné teploty naměřené v Klementinu v Praze. Barevná škála opět odpovídá naší intuici, tedy nejnižší teploty jsou označeny modře a nejvyšší červeně.

Kromě vlivu teploty na celé kompozice můžeme zkoumat například i její vliv na poměry mezi jednotlivými dvojicemi složek – je možné se zaměřit na to, jak teplota ovlivňuje hodnoty log-podílů.

Nejprve se zaměříme pouze na závislost log-podílů nakažených a hospitalizovaných na teplotě. Příslušné log-podíly jsme již spočítali při konstrukci ternárního diagramu, tudíž se nabízí jejich využití. Zobrazíme si je tedy v grafu v závislosti na teplotě a zobrazené hodnoty můžeme opět vyhladit pomocí splajnu – zvolíme si $df = 6$, abychom dobře viděli trend v datech, v tomto případě bychom však mohli zvolit i vyšší počet efektivních stupňů volnosti, protože nebudeme vytvářet ternární diagram, a zachycení některých lokálních efektů by nám proto příliš nevadilo. Na obrázku 3.16 vidíme odlehlé záporné hodnoty, které odpovídají prvním dnům pandemie, a ty způsobují pokles splajnu v okolí teploty 10°C. Abychom odstranili vliv těchto hodnot, vynecháme prvních 20 pozorování a vytvoříme si nový graf (obrázek 3.17), ve kterém můžeme pozorovat změnu chování splajnu

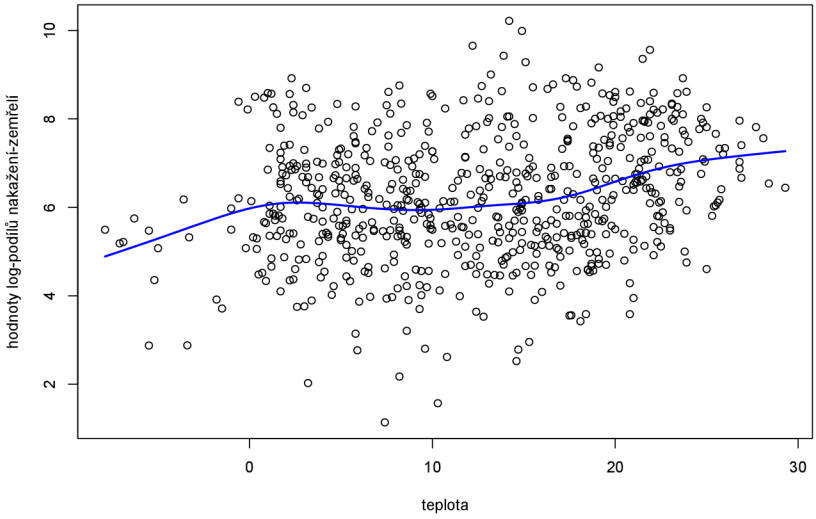


Obrázek 3.16: Graf znázorňuje závislost log-podílů nakažených a hospitalizovaných na teplotě vzduchu v období od 1. 3. 2020 do 31. 12. 2021.



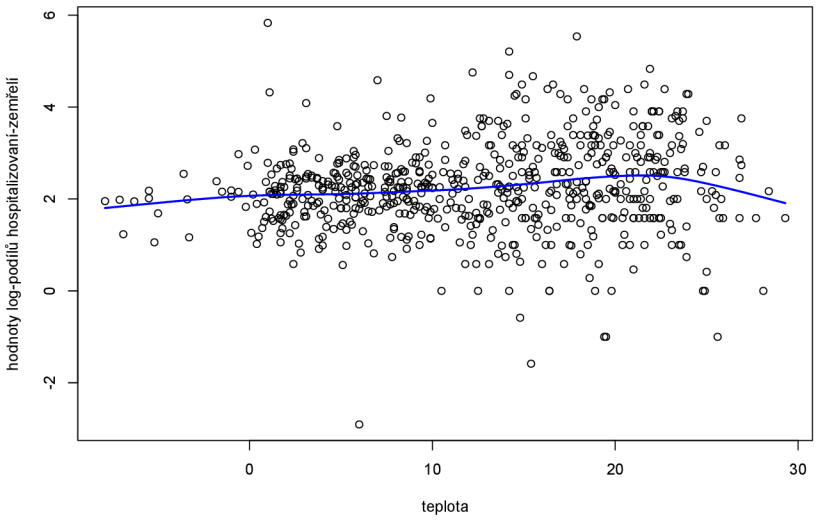
Obrázek 3.17: Graf znázorňuje závislost log-podílů nakažených a hospitalizovaných na teplotě vzduchu v období od 21. 3. 2020 do 31. 12. 2021 po vynechání prvních 20 pozorování.

– nevidíme již takový pokles v okolí teploty 10°C. Pro lepší interpretaci je také možné použít při výpočtu log-podílů jiný typ logaritmu. Dosud jsme používali přirozený logaritmus, nyní zkusíme použít binární logaritmus (se základem 2). Změna základu logaritmu nijak neovlivní chování výsledného splajnu (nový splajn je $(1/\ln 2)$-násobkem toho původního), dojde pouze ke změně funkčních hodnot a jejich možné jednodušší interpretaci (hodnota log-podílu k znamená, že hodnota složky v čitateli je 2^k-krát vyšší než hodnota složky ve jmenovateli). Z tohoto důvodu použijeme binární logaritmus i pro výpočet log-podílů mezi zbylými dvojicemi složek.



Obrázek 3.18: Graf znázorňuje závislost log-podílů nakažených a zemřelých na teplotě vzduchu v období od 21. 3. 2020 do 31. 12. 2021 po vynechání prvních 20 pozorování.

V grafech (obrázky 3.17, 3.18, 3.19) vidíme, že závislost jednotlivých log-podílů na teplotě vzduchu není nějak významná. Nemůžeme tedy například s jistotou říct, že by v létě mělo více pozitivně testovaných lehký průběh než v zimě. Ale v tomto případě by se mohl také projevit vliv testování, protože v létě se méně testovalo, neprobíhalo například preventivní testování zaměstnanců, které často odhalilo i nakažené, kteří neměli žádné příznaky, a v létě se tak možná



Obrázek 3.19: Graf znázorňuje závislost log-podílů hospitalizovaných a zemřelých na teplotě vzduchu v období od 21. 3. 2020 do 31. 12. 2021 po vynechání prvních 20 pozorování.

vůbec nepřišlo na to, že jsou nakažení. Jediné log-podíly, které nezávisejí na testovacím režimu, jsou ty pro hospitalizované a zemřelé. V odpovídajícím grafu (obrázek 3.19) vidíme, že teplota nemá téměř žádný vliv na poměr mezi těmito dvěma složkami. Můžeme si také všimnout změny rozptylu hodnot log-podílů, která byla popsána výše. Zde se taky můžeme zaměřit na interpretaci - vidíme, že se hodnoty log-podílů hospitalizovaných a zemřelých pohybují kolem 2, a tedy počty hospitalizovaných jsou 4krát vyšší než počty zemřelých, z čehož plyne, že přibližně 20 % osob, které byly hospitalizovány, zemřelo.

Také musíme ještě vzít v úvahu, že pro nejnižší a nejvyšší teploty máme poměrně málo pozorování, a proto nemůžeme příliš dobře posoudit vliv těchto teplot.

Na závěr ještě dodejme, že bychom mohli zkoumat i jiné závislosti. Za zmínku stojí třeba vliv očkování a každého z nás by určitě napadly i jiné faktory, které by mohly mít nějaký vliv na průběh nemoci nebo na šíření koronaviru.

Závěr

Viděli jsme jiný pohled na data týkající se COVIDu-19 zaměřený na relativní informaci obsaženou v těchto datech. Ukázali jsme si několik možností, jak lze využít log-podíly – můžeme je použít při analýze křivky v logaritmickém měřítku nebo při práci s kompozičními daty. Log-podíly lze také proložit hladkou křivkou, můžeme využít například lokální regresi a splajny.

Díky této práci jsem se mohla podívat na data získaná během pandemie jinak, než je prezentováno třeba na internetu, a překvapilo mě, kolik je možností, jak s daty pracovat. Líbilo se mi, že má práce nebyla pouze teoretická, ale měla jsem možnost vyzkoušet si i práci s reálnými daty. Ocenila jsem také, že jsem si mohla procvičit práci se softwarem `R`.

Myslím si, že tato práce je přínosná právě tím, že přináší jiný pohled na covidová data, než jsme zvyklí, a byly v ní použity metody, které se zaměřují na relativní informaci obsaženou v datech a jsou z mého pohledu méně známé. Na druhou stranu mohou být velmi přínosné v tom smyslu, že tento jiný pohled nás může přivést k dalším nápadům, jak s daty pracovat a jaké informace v nich hledat.

Netvrdila bych, že jsem ve své práci přišla s nějakým zázračným objevem, ale přesto se domnívám, že některé popsané výsledky jsou poměrně zajímavé i pro širší veřejnost.

Literatura

[1] Burýšek, J.: Kdy člověk začíná být infekční? Průvodce nákazou covidem-19. Seznam Zprávy. [online]. 2020, [cit. 2022-04-06]. Dostupné z: https://www.seznamzpravy.cz/clanek/kdy-clovek-zacina-byt-infekcni-pruvodce-nakazou-covidem-19-127187

[2] COVID-19 v ČR: Otevřené datové sady a sady ke stažení | Onemocnění Aktuálně MZČR [online]. [cit. 2022-03-13]. Dostupné z: https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19

[3] Eynatten, H. von, Pawlowsky-Glahn, V., Egozcue, J. J.: Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams. Mathematical Geology, Vol. 34, No. 3 (2002), s. 249–257.

[4] Fišerová, E., Hron, K.: Statistical inference in orthogonal regression for three-part compositional data using a linear model with type-II constraints. Communications in Statistics - Theory and Methods, 41(13–14) (2012), s. 2367–2385.

[5] Gámiz, M. L., Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P.: Missing link survival analysis with applications to available pandemic data. Computational Statistics & Data Analysis, Vol. 169, 107405 (2022).

[6] Hron, K.: Advances in compositional data analysis. Wiley StatsRef: Statistics Reference Online, 2018, s. 1–5.

[7] Hyndman, R. J.: Why log ratios are useful for tracking COVID-19. [online]. 2020, [cit. 2022-03-13]. Dostupné z: https://robjhyndman.com/hyndsight/logratios-covid19/

[8] James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) (2. vydání)*. Springer, 2021. [cit. 2022-03-13]. Dostupné z: https://www.statlearning.com/

[9] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R.: *Modeling and Analysis of Compositional Data.* Wiley, 2015.

[10] Portál ČHMÚ : Historická data : Počasí : Praha Klementinum. Český hydrometeorologický ústav [online]. [cit. 2022-03-13]. Dostupné z: https://www.chmi.cz/historicka-data/pocasi/praha-klementinum

[11] Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2009.