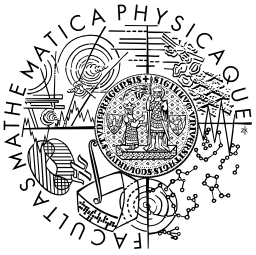


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Adéla Jalovcová

Prostorová epidemiologie

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jiří Dvořák, Ph.D.

Studijní program: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2022

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Tímto bych ráda poděkovala vedoucímu práce RNDr. Jiřímu Dvořákovi, Ph.D. jak za jeho podnětné rady a odbornou pomoc při psaní práce, tak za přátelský a lidský přístup ke studentům. Dále bych chtěla poděkovat své rodině za psychickou i finanční podporu po celou dobu mého studia a také svému partnerovi Honzovi za to, že mi během psaní práce vždy pomáhal a byl mi oporou.

Název práce: Prostorová epidemiologie

Autor: Adéla Jalovcová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Jiří Dvořák, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá představením nástrojů prostorové statistiky vhodných ke zkoumání prostorových epidemiologických dat. Práce představuje testy významnosti prostorové závislosti dat a aplikuje je na data o počtu nakažených Covidem 19. Jádrem práce je bayesovské modelování epidemiologických dat pomocí Integrated Nested Laplace Approximations. Práce shrnuje základní principy této metody a popisuje vybraný model pro představená data. Kromě prostorového aspektu dat práce ukazuje, jak model rozšířit o další proměnné působící na počet případů onemocnění nebo také o časovou složku. Mimo odhadu parametrů modelu se práce také zabývá testováním vhodnosti modelu pomocí obálkových testů.

Klíčová slova: prostorová epidemiologie, bodový proces, časoprostorový proces, náhodné pole, shlukování

Title: Spatial epidemiology

Author: Adéla Jalovcová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jiří Dvořák, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This work deals with spatial statistics methods that are suitable for analysing spatial epidemiological data. The work presents tests of spatial autocorrelation and applies them on data of the number of people infected by Covid 19. The main part of the work is Bayesian modelling of epidemiological data using Integrated Nested Laplace Approximations. We summarise the main principles of this method and present a chosen model for given data. Besides the spatial aspect of the data, the work shows how to incorporate other risk factors into the model and how to make the model spatio-temporal. Furthermore the work applies the model on the data and tests the suitability of the model with a global envelope test.

Keywords: spatial epidemiology, point process, space-time process, random field, clustering

Obsah

Úvod **2**

1 Data **3**
1.1 Příprava dat . . . 3
1.2 Prostorová struktura dat . . . 4
1.3 Značení . . . 7

2 Testování prostorové závislosti dat **8**
2.1 Teorie . . . 8
2.2 Aplikace na data . . . 9
2.3 Shrnutí . . . 14

3 Prostorový model **15**
3.1 Latentní Gaussovský model . . . 15
3.2 Integrated nested Laplace approximations . . . 16
3.2.1 Laplaceova aproximace . . . 16
3.2.2 Algoritmus . . . 17
3.3 Model . . . 18
3.4 Ekologická regrese . . . 19
3.5 Goodness of fit test . . . 19
3.5.1 Obálkový test . . . 20
3.6 Aplikace na data . . . 22
3.6.1 Prostorový model . . . 22
3.6.2 Ekologická regrese . . . 24
3.6.3 Goodness of fit test . . . 25
3.7 Implementace . . . 30

4 Časoprostorový model **31**
4.1 Teorie . . . 31
4.2 Aplikace na data . . . 31

Závěr **35**

Seznam použité literatury **36**

Seznam obrázků **39**

A Indexy okresů a týdnů **40**

Úvod

Epidemiologie, jak ji zná široká veřejnost především z posledních let, se zabývá vývojem a predikcí počtu nakažených danou nemocí v čase. Kromě časového vývoje mají ale epidemiologická data obvykle další stránku, a tou je stránka prostorová. Aktuální téma pandemie Covidu 19 nás tak motivovalo tuto často opomíjenou stránku dat zkoumat.
Prostorová epidemiologie používá několik různých nástrojů prostorové statistiky, jejichž volba obvykle závisí nejen na tom, co přesně chceme zkoumat, ale i na povaze získaných dat.
Z pohledu prostorové statistiky můžeme jednotlivé případy a jejich pozice v prostoru modelovat pomocí náhodného bodového procesu. Problémem tohoto přístupu je často nedostupnost takto podrobných údajů, proto se nejčastěji používá například pro zkoumání vzácných onemocnění. Ukázka takovéto analýzy je k nalezení například v Diggle (2013, Kapitola 9). V této práci se zaměříme naopak na agregovaná data. Téma typu zkoumaných dat a jejich sběru rozvádí například Illian (2008, kapitola 1.4.3).
V oboru epidemiologie se k modelování prostorových dat tradičně používá bayesovský přístup. Jednou z možností, jak data modelovat, jsou metody Markov Chain Monte Carlo, v poslední době se ale začíná prosazovat modelování pomocí Integrated Nested Laplace Approximations, zkráceně INLA, poprvé publikované v Rue a kol. (2009), a to především pro nižší časovou náročnost výpočtu, ale i jednodušší ladění řetězce, jelikož není nutné ověřovat konvergenci ke stacionárnímu rozdělení či volit návrhová rozdělení. Popularizaci této metody také pomohla implementace v R, konkrétně knihovna R-INLA veřejně dostupná z www.r-inla.org.
V první kapitole představíme data mapující vývoj počtu nakažených Covidem 19 v jednotlivých okresech České republiky mezi březnem 2020, kdy se podařilo zachytit první případy nákazy v ČR, a listopadem 2021, kdy jsme data začali zpracovávat. Zároveň ukážeme některé důležité charakteristiky dat, které použijeme v následujících kapitolách.
Druhá kapitola je věnována testování významnosti prostorové autokorelace dat pomocí testových statistik založených na Moranově a Gearyho indexu. Prozkoumáme možné předpoklady o datech a porovnáme výsledky těchto testů aplikovaných na data z předchozí kapitoly.
Třetí kapitola se věnuje modelování epidemiologických dat pomocí INLA. Představíme základní předpoklady pro použití metody a její algoritmus. Dále se budeme zabývat jedním z nejčastěji používaných modelů v epidemiologii a aplikujeme ho na data počtu případů Covidu 19. V druhé části kapitoly je věnován prostor kontrole modelu pomocí prediktivních aposteriorních charakteristik a obálkového testu. Na závěr představíme způsob, jakým do modelu zakomponovat další faktory ovlivňující výskyt onemocnění.
Závěrečná kapitola je věnována rozšíření čistě prostorového modelu o časovou složku. Představíme základní tvar modelu a jeho aplikaci na data.

1. Data

V následujících kapitolách budeme pracovat s daty o počtu nakažených v České republice v období od 1. března 2020, kdy se na území České republiky objevil první případ onemocnění Covidem 19 do 13. listopadu 2021. Data s názvem COVID-19: Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu, která jsou dostupná z Komenda M. (2021) ve formátu csv. Obsahují 48595 záznamů o sedmi údajích, kterými jsou následující

- *id*: unikátní identifikátor záznamu,
- *datum*: datum, kdy byl zaznamenán daný počet případů,
- *kraj_nuts_kod*: NUTS kód kraje, do kterého spadá okres, kde byl zaznamenán daný počet případů,
- *okres_lau_kod*: kód okresu, kde byl zaznamenán daný počet případů, tato proměnná má 77 možných hodnot, 76 okresů mimo Prahu, jednotlivé pražské části pak sjednotíme do jednoho celku, který budeme dále nepřesně nazývat okres, abychom byli konzistentní se zbytkem územních celků
- *kumulativni_pocet_nakazenych*: celkový (kumulativní) počet nakažených Covidem 19 k danému datu v daném okrese,
- *kumulativni_pocet_vylecenych*: celkový (kumulativní) počet vyléčených z onemocnění Covid 19 k danému datu v daném okrese,
- *kumulativni_pocet_umrti*: celkový (kumulativní) počet úmrtí na Covidem 19 k danému datu v daném okrese.

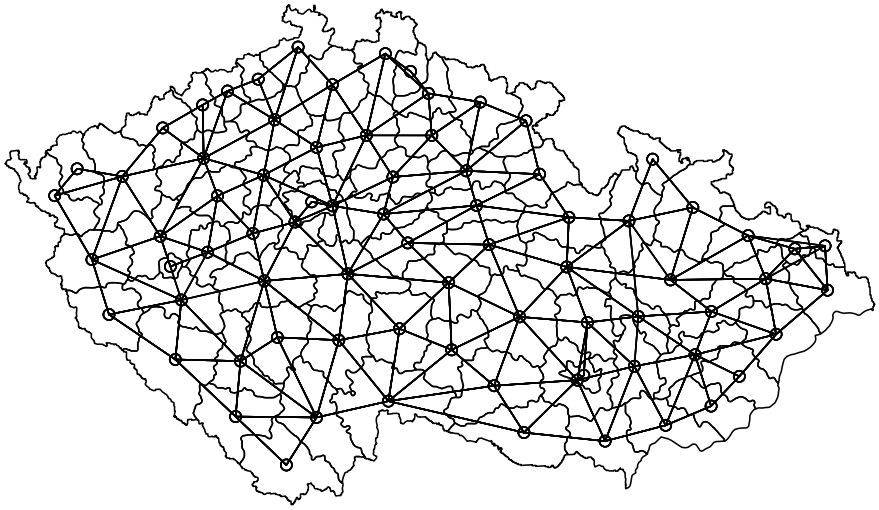
Kromě dat o Covidu-19 budeme používat další data o okresech ČR, konkrétně údaje o věku a počtu obyvatel dostupná z ČSÚ (2021) a rozloze okresů dostupných z ČSÚ (2018). Hlavní město Praha opět evidujeme jako jeden okres.

1.1 Příprava dat

Pro další práci s daty setřídíme záznamy o počtu nakažených tak, že první sloupec reprezentuje názvy okresů, další sloupce pak značí kumulativní počet případů onemocnění zaznamenaných v daný den. Následně spočteme z kumulativního počtu případů denní přírůstky případů a z nich následně součet přírůstků vždy za období jednoho týdne, kdy jednotlivé týdny jsou disjunktní a bezprostředně na sebe navazují. Získáváme tím 89 záznamů o týdenních přírůstcích pro všech 77 okresů. První týden začíná dnem 1.3.2020 a poslední týden začíná dnem 7.11.2021. Počtem vyléčených a počtem úmrtí se nebudeme dále zabývat.
Kromě výše uvedeného si dále pomocí datasetu o počtu obyvatel v okresech přepočteme počet případů onemocnění na 100 tisíc obyvatel. Dále si z údajů o počtu obyvatel a rozlohy okresů spočteme hustotu zalidnění jednotlivých okresů.

1.2 Prostorová struktura dat

Prostorová struktura dat vychází z geografického rozložení okresů v České republice, a to takovým způsobem, že dva okresy spolu sousedí, pokud sdílejí alespoň jeden bod hranice. Tím pádem je relace sousedství symetrická. Vzniklou strukturu sousedství můžeme znázornit jako graf z obrázku 1.1 nebo matici sousednosti z obrázku 1.2. U matice sice nevidíme, který řádek a sloupec reprezentuje který okres, ale je z něj patrné, že matice sousednosti je řídká, což pro nás bude významné při výpočtech v dalších kapitolách. Data pro vykreslení mapy jsou volně dostupná z GADM (2021).

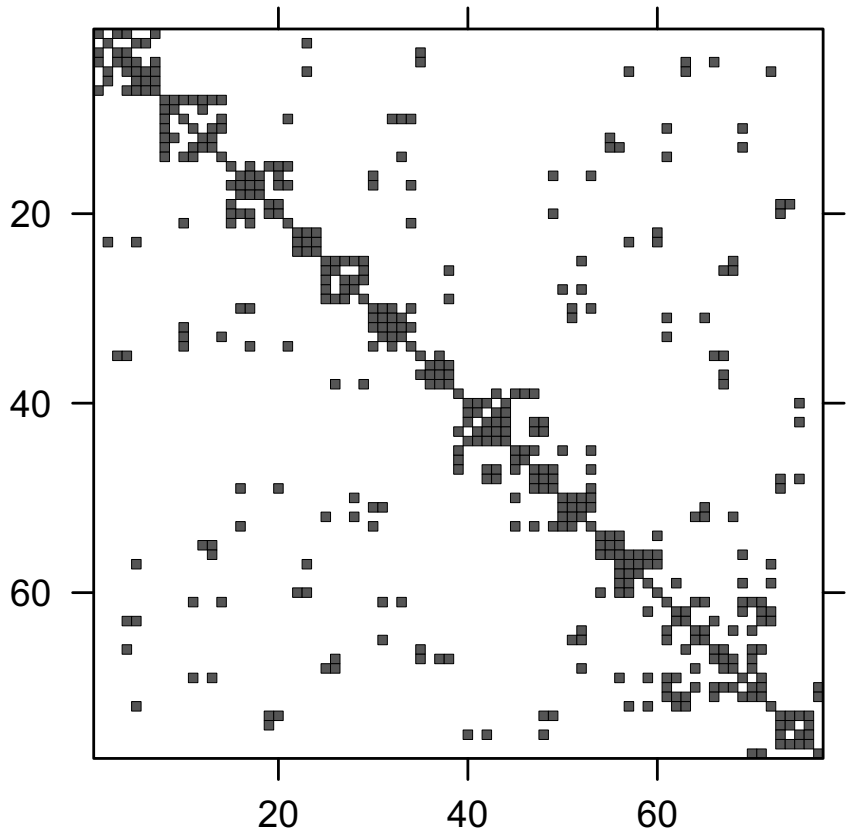


Obrázek 1.1: Znázornění relace sousedství mezi okresy

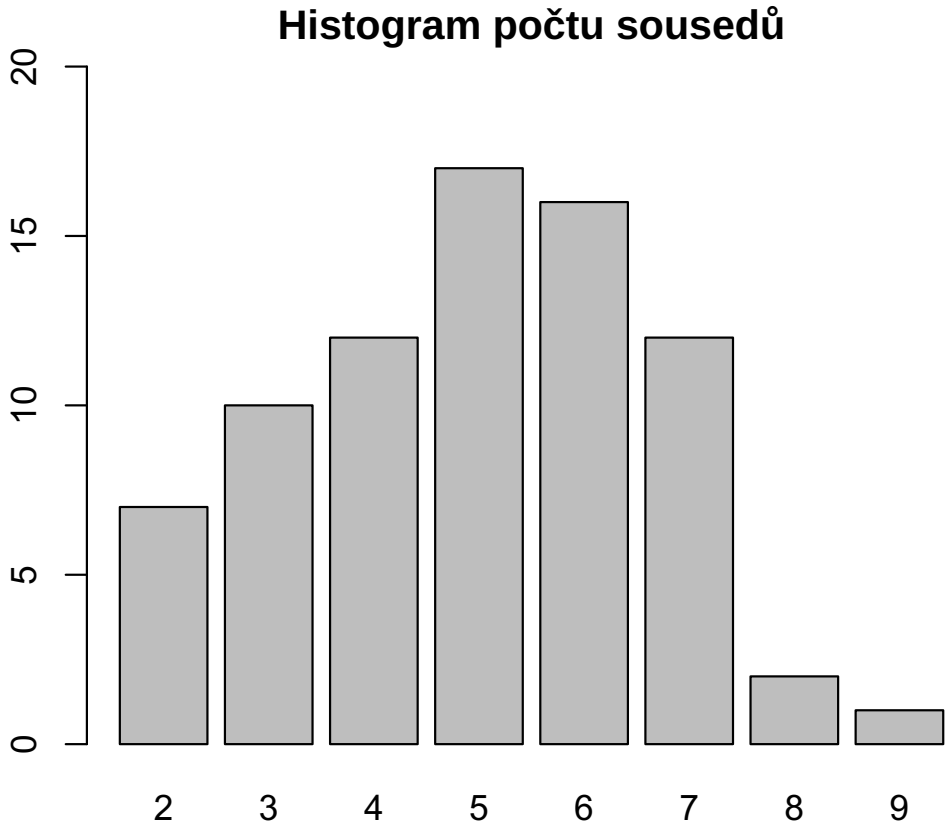
Pro lepší představu o datech se můžeme podívat na základní údaje o počtu sousedních okresů a histogram 1.3.

Minimum	Medián	Průměr	Maximum
2,00	5,00	4,96	9,00

Tabulka 1.1: Tabulka shrnující základní popisné charakteristiky počtu sousedů okresů



Obrázek 1.2: Matice sousednosti okresů, černá pole značí, že okresy jsou v sousedské relaci, bílá, že nejsou, pořadí okresů je dostupné v tabulce A.1.



Obrázek 1.3: Graf znázorňující četnost počtu sousedů okresů v České republice

1.3 Značení

V následujícím textu budeme jednotlivé okresy značit s_i pro $i = 1,...77$. Okresy mají dané pořadí, tak jak uvádí tabulka A.1 v přílohách práce. Pro zjednodušení značení budeme často zkracovat s_i pouze na i, především v situacích, kdy chceme ukázat, že nějaká charakteristika patří okresu s_i. Počet sousedů okresu s_i budeme značit $N(i)$.
Náhodný vektor udávající počet nakažených v jednotlivých okresech budeme značit y a náhodný vektor udávající počet nakažených na sto tisíc obyvatel budeme značit Z. Napozorované hodnoty náhodných veličin y a Z budeme značit y_{obs}, respektive Z_{obs}. Značení volíme na základě zvyklostí v těchto oborech.
Jednotlivé týdny budeme značit pomocí indexu $t = 1,...89$, kde $t = 1$ značí týden začínající dnem, kdy jsme poprvé pozorovali pozitivní test na Covid 19 v některém z okresů, tedy týden od 1.3. 2020 do 7.3. 2020. Naopak $t = 89$ značí poslední týden, který sledujeme, tedy dny od 7.11. 2021 do 13.11. 2021. Týdny jsou indexované chronologicky. Detailní seznam indexování týdnů je k nalezení v přílohách v tabulce A.2.

2. Testování prostorové závislosti dat

Prvním aspektem prostorových dat, na který se zaměříme, je zkoumání prostorové závislosti. Nejdříve představíme testy, kterými lze prostorovou autokorelaci testovat, a poté tyto metody aplikujeme na data představená v kapitole 1. Budeme pracovat s údaji o počtu případů na 100000 obyvatel, které jsou realizacemi náhodných veličin Z_1, Z_2,...,Z_{77}. Při konstrukci testů budeme vycházet z Cliff a Ord (1970) a Oyana (2020).

2.1 Teorie

V následující kapitole popíšeme vybrané metody, kterými můžeme testovat prostorovou autokorelaci, tedy hypotézu

$$H_0 : \text{Náhodné veličiny } Z_1, ..., Z_{77} \text{ jsou nezávislé.}$$

Abychom mohli tuto hypotézu testovat, musíme provést dva kroky, těmi jsou volba předpokladů a výběr testové statistiky.
Nejdříve se podíváme na možné předpoklady. První, co je třeba stanovit, je matice vah. Vyjdeme z matice sousednosti z obrázku 1.2.Označme $\omega_{i,j}$, i,$j = 1$,...,n prvek matice vah W na pozici [i,j]. Představíme si dvě možnosti, jak volit matici W.

- *binární váhy*: funkci, která přiřazuje sousedům x_i, x_j binární váhu $\omega_{i,j}$, můžeme vyjádřit jako

$$f(s_i, s_j) = \begin{cases} 1 & \text{pokud } s_i \text{ a } s_j \text{ sdílejí hranici,} \\ 0 & \text{jinak} \end{cases}$$

- *normalizované váhy*: Pro normalizované váhy můžeme funkci napsat jako

$$f(s_i, s_j) = \begin{cases} \frac{1}{N(i)} & \text{pokud } s_i \text{ a } s_j \text{ sdílejí hranici,} \\ 0 & \text{jinak,} \end{cases}$$

  kde $N(i)$ značí počet sousedů okresu s_i.

Jedná se o obvyklé volby vah. Vhodnost záleží na variabilitě rozdělení počtu sousedů, pro malou variabilitu počtu sousedů jsou vhodné binární váhy, pro větší variabilitu obvykle volíme normalizované váhy.
Pro náhodné pole Z dále předpokládáme, že platí buď předpoklad randomizace, nebo předpoklad normality. Předpoklad randomizace říká, že každá z $n!$ permutací pozorovaných hodnot je stejně pravděpodobná. Předpoklad normality popisuje za nulové hypotézy Z jako náhodné pole získané z nezávislých náhodných veličin s normálním rozdělením $N(\mu, \sigma^2)$. Oba předpoklady popisují situaci bez prostorové autokorelace.
Oba výše zmíněné předpoklady musíme stanovit na základě znalosti situace a vzorku dat, obvykle po konzultaci s odborníky na danou problematiku.

Jako testovou statistiku můžeme použít Moranův nebo Gearyho Index.

Definice 1. *Nechť $Z_i : i \in L$ je náhodné pole s konstantní střední hodnotou $\mathbb{E}Z_i = \mu$ a konstantním konečným rozptylem $varZ_i = \sigma^2, \sigma^2 > 0$. Pak definujeme Moranův index jako*

$$I = \frac{n}{\omega} \frac{\sum_{i \in L} \sum_{j \in L} \omega_{i,j}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i \in L}(Z_i - \bar{Z})^2},$$

kde $\bar{Z} = \frac{1}{n} \sum_{i \in L} Z_i$ a n počet prvků indexové množiny L. Gearyho index je dán vzorcem

$$c = \frac{n-1}{2\omega} \frac{\sum_{i \in L} \sum_{j \in L} \omega_{i,j}(Z_i - Z_j)^2}{\sum_{i \in L}(Z_i - \bar{Z})^2}.$$

Pro předpoklad normality se zpravidla používá asymptotický test, jelikož platí, že

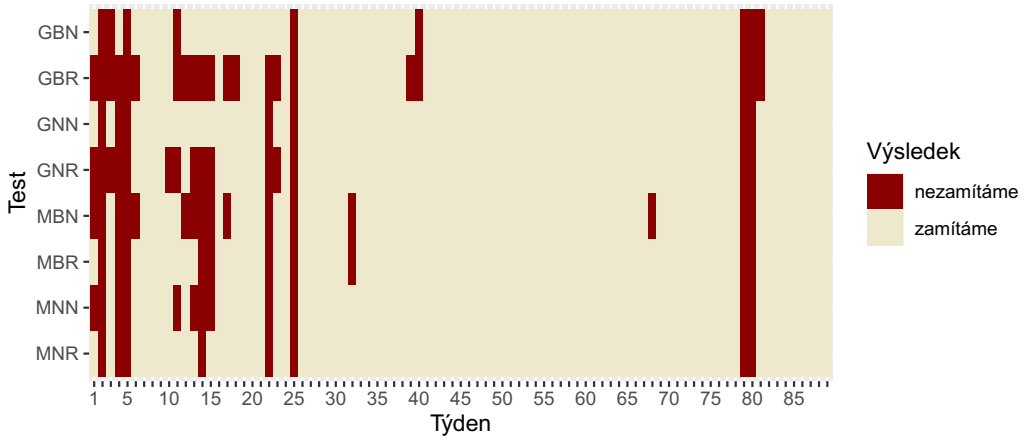
$$\frac{M_{obs} - E_g M}{\sqrt{var_g M}},$$

má asymptoticky normální rozdělení $N(0,1)$. Jako M značíme jednu z testových statistik z definice 1, střední hodnotu a rozptyl M za předpokladu normality značíme $E_g M$ a $var_g M$, M_{obs} je pak napozorovaná hodnota testové statistiky.
Pokud platí předpoklad randomizace, je více možností, jak test provést. První možností je asymptotický test obdobný tomu pro předpoklad normality, druhou možností je pak permutační test nebo pro zrychlení výpočtu Monte-Carlo test. Výpočet momentů a asymptotické rozdělení těchto statistik dále rozvádí a dokazují například v Bivand a Wong (2018) a Sen (2010).

2.2 Aplikace na data

Pokud v praxi sestavujeme test prostorové autokorelace, máme s využitím výše zmíněných metod osm možností, jelikož můžeme libovolně nakombinovat výpočet vah v modelu, testovou statistiku pomocí Gearyho nebo Moranova indexu a předpoklad randomizace nebo normality. Volbu testu z těchto možností zakládáme na odborné znalosti situace či oboru, z kterého pochází zkoumaná data.
Pokud bychom chtěli testovat autokorelaci dat představených v kapitole 1, zvolili bychom si jeden test, na základě kterého bychom udělali závěr. Níže ale provedeme testování autokorelace všemi výše popsanými testy a porovnáme jejich výsledky.
Testy budeme značit zkratkou o třech písmenech, kde první reprezentuje index, od kterého je odvozena testová statistika (M = Moranův, G = Gearyho), druhé reprezentuje tvar matice vah (B = binární, N = normalizované) a třetí písmeno značí předpoklad (R = randomizace, N = normality).
Výsledky všech testů můžeme znázornit pomocí grafu 2.1.

Z grafického znázornění výsledků můžeme vidět, že testy nejsou ekvivalentní, ale ve velkém množství případů se výsledky většiny z nich shodují. Ze znalosti toho, jak se infekční nemoci jako Covid 19 šíří, výsledek, kdy velká většina testů zamítá nulovou hypotézu, dává smysl. Zároveň je dobré si všimnout, že testy nejčastěji nezamítají nulovou hypotézu v prvních týdnech pandemie, kdy bylo na území České republiky malé množství pozitivních testů.



Obrázek 2.1: Výsledky testů autokorelace na hladině $\alpha = 0{,}05$

K lepšímu náhledu na výsledky testů nám mohou pomoci histogramy v obrázku 2.2.

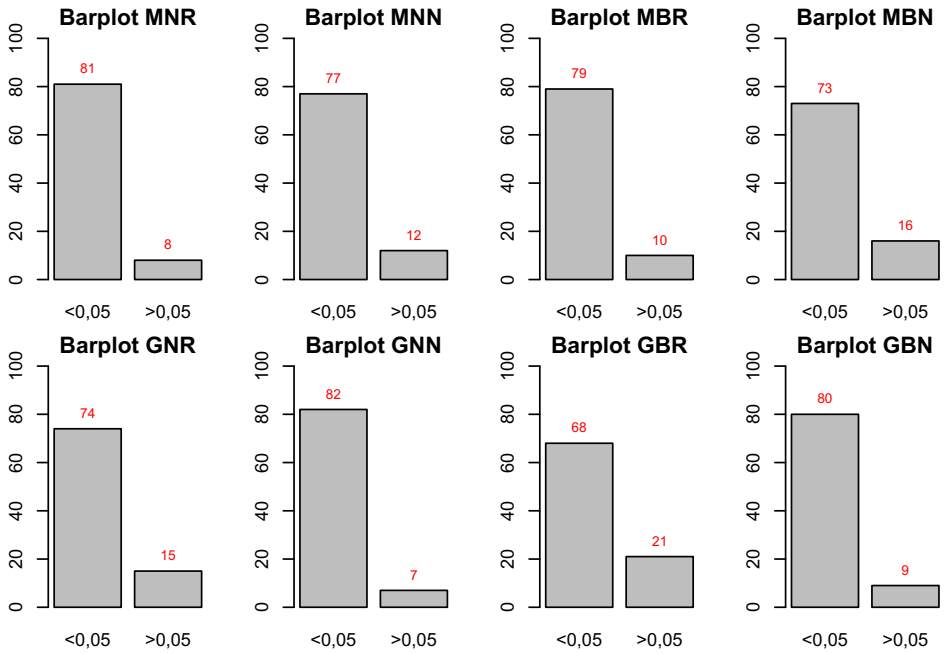
Na obrázku 2.2 můžeme vidět, že všechny testy zamítají nulovou hypotézu v 68 až 81 případech z 89 a naopak nezamítají v 8 až 21 případech. Nejméně často pak zamítá test pomocí Gearyho indexu s předpokladem binárních vah a randomizace, naopak nejčastěji zamítá test sestavený pomocí Moranova indexu s předpokladem normalizovaných vah a randomizace.
Graf 2.3 pak znázorňuje četnosti toho, že v daném týdnu zamítlo dané množství testů nulovou hypotézu.

Jak můžeme vidět, žádný test nezamítne nulovou hypotézu v pěti případech. Konkrétně jde o týdny 2, 5, 25, 79 a 80. Počet případů na sto tisíc obyvatel znázorňuje obrázek 2.4.

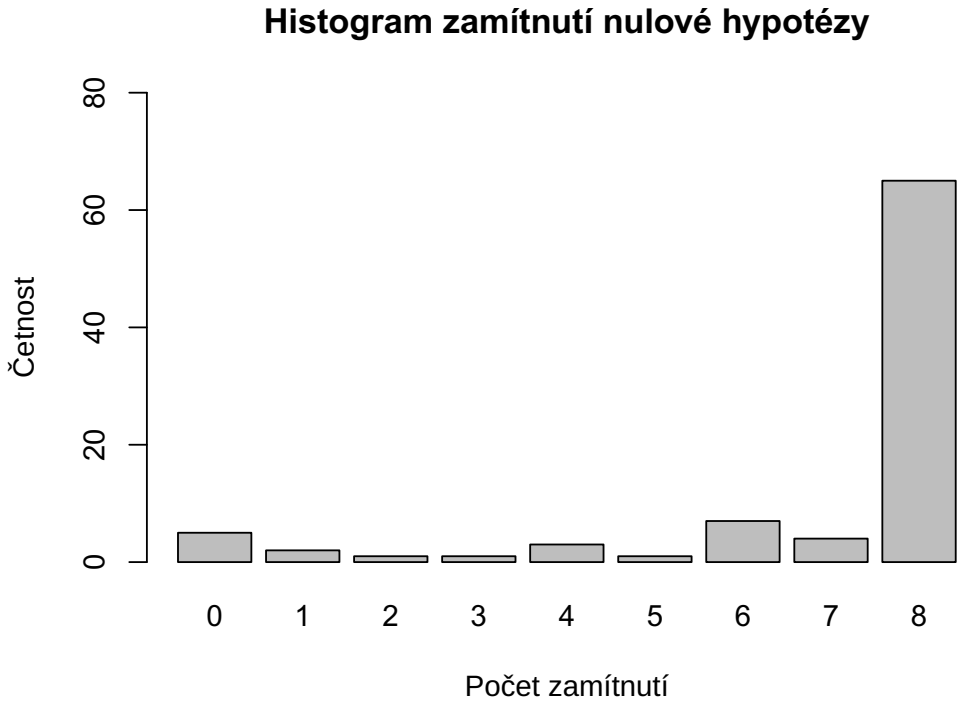
Mohlo by se zdát, že některé ze zobrazených týdnů, kdy žádný test nezamítl nulovou hypotézu, mají kladnou prostorovou autokorelaci, ale ukázalo se, že není statisticky významná.
Naopak všechny testy zamítají nulovou hypotézu například pro týdny 8 a 55. Jejich grafické znázornění můžeme vidět na obrázku 2.5. Zvolili jsme právě tyto dva týdny, abychom ukázali data z dvou různých situací, osmý týden je ukázkou ze začátku epidemie, zatímco během týdne 55 doznívala jedna z nejsilnějších vln epidemie. Je důležité vědět, že v obrázcích 2.4 a 2.5 jednotlivé grafy mají jiné zobrazení přiřazující počtu případů na počet obyvatel barvu, a to z důvodu velkého nárůstu počtu případů v čase. Konkrétní čísla uvádí tabulka 2.1.

Týden	2	5	8	55	79	80
Počet nakažených	162	1795	679	65516	1666	2574

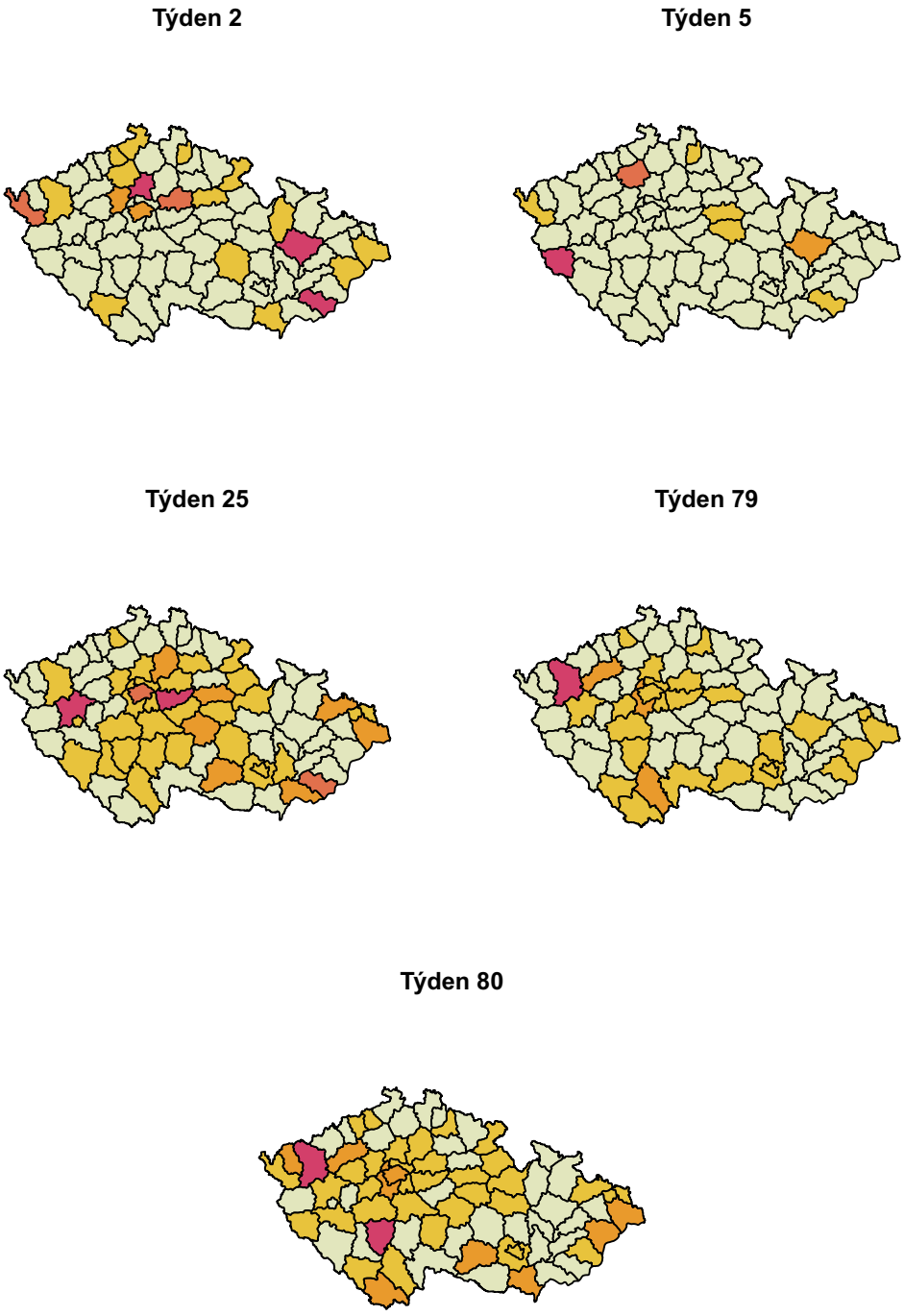
Tabulka 2.1: Tabulka počtu nakažených v týdnech z obrázků 2.4 a 2.5



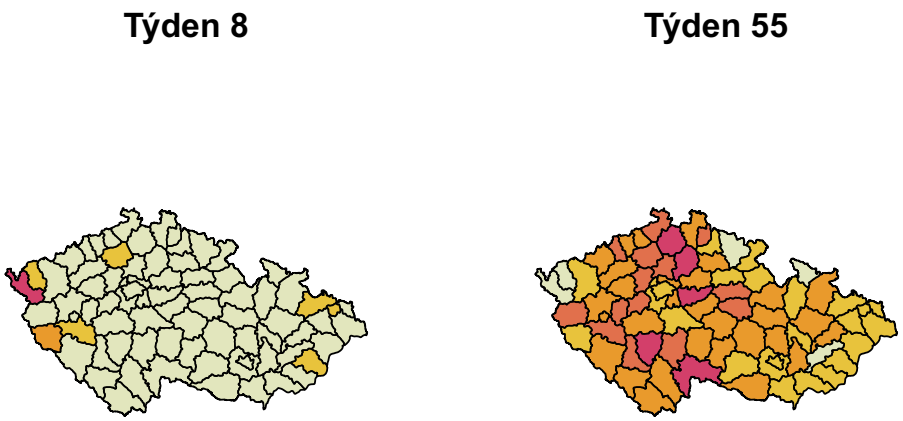
Obrázek 2.2: Graf četnosti zamítnutí nulové hypotézy pro jednotlivé testy. Popisy osy x u grafů značí spočtenou p-hodnotu testů.



Obrázek 2.3: Histogram počtu případů reprezentujících, kolik testů zamítlo nulovou hypotézu v jednotlivých týdnech



Obrázek 2.4: Znázornění počtu případů Covidu 19 na počet obyvatel v jednotlivých okresech pro týdny, kdy žádný z testů nezamítl nulovou hypotézu. Barevná škála značí nejsvětlejší barvou okresy s nízkým počtem nakažených na sto tisíc obyvatel, naopak tmavá červená značí nejvyšší počty případů. Barevná škála je pro každý z týdnů jiná.



Obrázek 2.5: Znázornění počtu případů Covid 19 na počet obyvatel v jednotlivých okresech pro vybrané týdny, kdy všechny testy zamítly nulovou hypotézu. Barevná škála značí nejsvětlejší barvou okresy s nízkým počtem nakažených na sto tisíc obyvatel, naopak tmavá červená značí nejvyšší počty případů. Barevná škála je pro každý z týdnů jiná.

2.3 Shrnutí

Z provedených testů můžeme vidět, že pro většinu týdnů data obsahují významné prostorové autokorelace. Nejčastěji testy nezamítají nulovou hypotézu pro týdny na začátku pandemie, zejména to lze pozorovat u týdnů 2 a 5, kde nezamítl hypotézu žádný test. To je pravděpodobně zapříčiněno několika faktory ze začátku pandemie, kdy byly počty případů ve většině okresů nízké a skoro či úplně navzájem nezávislé, až postupem času se vytvořila výraznější lokální ohniska. Zároveň musíme vzít v potaz, že teprve začínalo testování tohoto onemocnění a nemuseli jsme v každém z okresů zachytit stejný podíl nakažených. Z obrázků 2.4 a 2.5 můžeme vidět, že je těžké jen na základě vizuálního znázornění rozhodnout o tom, zda je prostorová autokorelace dat statisticky významná.

3. Prostorový model

V následující kapitole se zaměříme na prostorový model počtu nakažených v jednom daném týdnu pomocí bayesovských metod. Představíme vhodné modely pro epidemiologická data a aplikujeme je na data počtu nakažených Covidem 19 z kapitoly 1.
Naším hlavním cílem bude na základě modelu odhadnout relativní riziko nákazy v jednotlivých oblastech. Jak uvádí Beaglehole R. (2002, kapitola 2) relativní riziko definujeme jako podíl pravděpodobnosti výskytu onemocnění u skupiny obyvatel vystavené danému faktoru oproti té u lidí, kteří nebyli vystaveni danému faktoru. Prvním krokem k získání odhadu relativního rizika je odvození prostorového modelu.
K prostorovým epidemiologickým modelům se dá přistoupit několika způsoby. My se zaměříme na modely pro data na mříži (lattice). I tam bychom mohli zvolit několik možných postupů. Jelikož ale epidemie jako například Covid 19 se dynamicky mění a v praxi je nutné modely vytvářet poměrně rychle, aby odpovídaly aktuální situaci, musí být v tomto oboru brán ohled nejen na výpočetní náročnost, ale i na časovou náročnost a množství práce investované do odvozování modelu. I kvůli tomu se v prostorové epidemiologii prosazuje bayesovský přístup k odhadování modelů za pomoci INLA v porovnání s metodami Markov Chain Monte Carlo. Porovnání těchto metod je dohledatelné v Khan a kol. (2021). Rozšíření této metody napomáhá i to, že je implemetována jako knihovna například v R, což usnadňuje získání praktických výsledků.

3.1 Latentní Gaussovský model

Základem pro bayesovské modelování pomocí Integrated Nested Laplace Approximations je latentní Gaussovský model, popsaný například v Opitz (2017) a Martino a Riebler (2019). Skládá se ze tří vrstev: hyperparametry, latentní Gaussovské pole a věrohodnostní model. Pro data y předpokládáme, že jsou podmíněně při daných hodnotách latentního gausovského pole θ nezávislá. Veličina y_i závisí pomocí link funkce na prediktoru η_i, jehož obecný tvar je

$$\eta_i = b_0 + \sum_j \beta_j z_{ij} + \sum_k w_k f^k(x_{ik}), \tag{3.1}$$

kde b_0 značí absolutní člen, z jsou známé kovariáty s lineárním efektem β, w je známý vektor vah a f je vektor neznámých funkcí kovariát x. Gausovské pole je pak vektor $\theta = (b_0, \beta, f)$ Vrstvy hiearchického modelu potom vypadají následovně:

$$\psi \sim \pi(\psi), \tag{3.2}$$

$$\theta|\psi \sim N(0, Q(\psi)^{-1}), \tag{3.3}$$

$$y|\theta, \psi \sim \Pi_i \pi(y_i|\eta_i,\psi). \tag{3.4}$$

Předpis (3.2) reprezentuje vektor hyperparmaterů ψ a jeho rozdělení $\pi(\psi)$, které nemusí být nutně gaussovské. V předpisu 3.3 můžeme vidět, že podmíněné rozdělení vektoru parametrů θ při daném ψ je vícerozměrné normální se střední

hodnotou nula a rozptylovou maticí $Q(\psi)^{-1}$, která může být závislá na hyperparametrech ψ. Nakonec předpis 3.4 už znázorňuje věrohodnostní model pro data y za podmínky parametrů a hyperparametrů. Rozdělení π zde značí obecné pravděpodobnostní rozdělení, které nemusí reprezentovat v každé rovnici to samé.

3.2 Integrated nested Laplace approximations

K výpočtu budeme v praxi používat R s balíkem R-INLA, jehož samotný název je zkratkou pro Integrated Nested Laplace approximations, které k odhadu parametrů používá. Metoda byla představena v článku Rue a kol. (2009).
Abychom mohli použít INLA, musí daný latentní gaussovský model splňovat následující požadavky:

1. Každé z pozorování y_i závisí pouze na jednom prvku θ_i latentního gaussovského pole θ
2. Počet prvků vektoru hyperparametrů není příliš velký. Obvykle se v literatuře uvádí menší než 15 (Martino a Riebler, 2019).
3. Latentní Gaussovské pole θ může být velké, ale jeho matice přesnosti Q musí být řídká.
4. Funkce f z předpisu (3.1) jsou hladké.

Kromě toho musí platit:

5. Výsledkem, který chceme získat, jsou jednorozměrné aposteriorní marginály $\pi(\theta_i|y)$ a $\pi(\psi_j|y)$, nikoliv vícerozměrné aposteriorní rozdělení $\pi(\theta, \psi|y)$.

I pokud jsou tyto požadavky splněny, v ojedinělých případech mohou nastat při výpočtu problémy, zejména pokud máme málo pozorování. Více o tom lze nalézt například v Martino a Riebler (2019).

3.2.1 Laplaceova aproximace

Jádrem samotného výpočtu je Laplaceova aproximace popsaná také například v Blangiardo a Cameletti (2015, kapitola 4). Na rozdíl od Markov Chain Monte Carlo metody není založena na simulaci, ale na analytickém odvození. Mějme integrál

$$\int f(x)\,\mathrm{d}x = \int \exp(\log f(x))\,\mathrm{d}x,$$

jehož aproximaci chceme získat. Pomocí $f(x)$ zde značíme hustotu náhodné veličiny. Nyní provedeme aproximaci funkce $\log f(x)$ pomocí Taylorova polynomu v bodě x_0:

$$\log f(x) \approx \log f(x_0) + (x - x_0)\frac{\partial \log f(x)}{\partial x}\Big|_{x=x_0} + \frac{(x-x_0)^2}{2}\frac{\partial^2 \log(f(x))}{\partial x^2}\Big|_{x=x_0}$$

Nyní volme $x_0 = \operatorname{argmax}_x \log f(x)$, pak platí $\frac{\partial \log f(x)}{\partial x}\Big|_{x=x_0} = 0$ a aproximace se zjednoduší na

$$\log f(x) \approx \log f(x_0) + \frac{(x-x_0)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x_0}.$$

Dosazením do integrálu získáváme

$$\begin{aligned}\int f(x)\,\mathrm{d}x &\approx \int \exp\left(\log f(x_0) + \frac{(x-x_0)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x_0}\right)\,\mathrm{d}x \\ &= \exp(\log f(x_0)) \int \exp\left(\frac{(x-x_0)^2}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x_0}\right)\,\mathrm{d}x.\end{aligned} \tag{3.5}$$

Můžeme si všimnout, že integrál 3.5 připomíná distribuční funkci normálního rozdělení. Využijeme toho, že $\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x_0} \leq 0$ a pro $\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x_0} \neq 0$ definujme

$$\sigma^2 = \frac{-1}{\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x_0}}.$$

Po dosazení získáme

$$\int f(x)\,\mathrm{d}x \approx \exp(\log f(x_0)) \int \exp\left(\frac{-(x-x_0)^2}{2\sigma^2}\right)\,\mathrm{d}x,$$

kde integrand je hustotou normálního rozdělení se střední hodnotou x_0 a rozptylem σ^2 přenásobený konstantou. Díky tomu, můžeme odhadnout určitý integrál na intervalu (α, β) jako

$$\int_\alpha^\beta f(x)\,\mathrm{d}x \approx f(x_0)\sqrt{2\pi\sigma^2}(\Phi(\beta) - \Phi(\alpha)),$$

kde $\Phi(.)$ značí distribuční funkci normálního rozdělení $N(x_0, \sigma^2)$.

3.2.2 Algoritmus

Celý algoritmus Integrated nested Laplace approximations se skládá ze čtyř hlavních kroků.

1. K výpočtu budeme nejdříve potřebovat odhad $\tilde{\pi}(\psi|y)$. Tento odhad je založený na vztahu
$$\tilde{\pi}(\psi|y) \propto \frac{\pi(\theta, \psi, y)}{\tilde{\pi}_G(\theta|\psi, y)}\Big|_{\theta=\theta^*(\psi)},$$
kde $\tilde{\pi}_G(\theta|\psi, y)$ je gaussovská aproximace podmíněného rozdělení θ a $\theta^*(\psi)$ je modus $\pi(\theta|\psi)$. Modus $\tilde{\pi}(\psi|y)$ nalezneme pomocí optimalizace $\log(\tilde{\pi}(\psi|y))$ vzhledem k ψ. V praxi se při optimalizaci používá například quasi-Newtonova metoda. Tento odhad použijeme k stanovení K-tice bodů $\{\psi_1,...,\psi_K\}$ v oblasti, kde je vysoký odhad hustoty $\tilde{\pi}(\psi|y)$.

2. Pro body vybrané v prvním kroku spočteme $\tilde{\pi}(\psi_1|y),...,\tilde{\pi}(\psi_K|y)$.

3. Pro každý bod ψ_k odhadneme hustotu $\pi(\theta_i|\psi,y)$ jako $\tilde{\pi}(\theta_i|\psi_i,y)$ pro každé $k = 1,...K$ pomocí Laplaceovy aproximace z podkapitoly 3.2.1, případně zjednodušené Laplaceovy aproximace, jejíž algoritmus je popsán ve Wood (2019).

4. Provedeme numerickou integraci integrálu na pravé straně výrazu

$$\pi(\theta_i|y) = \int\int \pi(\theta, \psi|\theta) d\theta_{-i} d\psi = \int \pi(\theta_i|\psi, y)\pi(\psi|y) d\psi.$$

   K numerické integraci použijeme body ψ_1,...,ψ_K z bodu 1 a 2 stejně jako odhady hustot $\tilde{\pi}(\psi|y)$ a $\tilde{\pi}(\theta|\psi_i,y)$ z bodů 1 a 3.

Kroky 2 a 3 dávají tomuto postupu přívlastek nested, zatímco numerická integrace v kroku 4 přidává do názvu integrated. Laplaceova transformace obvykle dosahuje nejpřesnějších výsledků, ale pro zjednodušení, a tím pádem i zrychlení výpočtu se používá pro případy, kde je to potřeba, zjednodušená Laplaceova aproximace. Podrobnější popis této metody je k nalezení například v Rue a kol. (2009).

3.3 Model

K modelování epidemiologických dat existuje i v rámci bayesovských metod několik přístupů. My se zaměříme především na BYM (Besag-York-Mollié) model, neboli konvoluční model představený v Besag a kol. (1991). Tento model jsme zvolili, protože je v oboru matematické epidemiologie často používaný, a to především díky tomu, že zahrnuje jak prostorové, tak lokální vlivy na počet nakažených a zároveň díky metodě INLA není příliš výpočetně náročný. Porovnání s dalšími bayesovskými prostorovými modely je k nalezení například v Best a kol. (2005) nebo Lee (2011). Předpokládáme:

$$y_i|u_i \sim Poisson(E_i \cdot \exp(\beta_0 + u_i + v_i)), \tag{3.6}$$

kde y_i značí počet nakažených v okresu s_i, E_i značí očekávaný počet nakažených v regionech $i = 1,...,n$, koeficient u_i značí prostorově závislý efekt příslušný okresu s_i a koeficient v_i značí nezávislý prostorový efekt pro daný okres. Koeficient β_0 zde reprezentuje absoluntí člen a předpokládáme pro něj obvykle, jak uvádí (Gómez-Rubio, 2020, kapitola 3.2), apriorní rozdělení $N(0,1000)$. Zároveň předpokládáme , že

$$u_i|u_{-i} \sim N\left(\frac{1}{N(i)}\sum_{j=1}^{n} w_{ij}u_j, \sigma_{ui}^2\right),$$

kde $\sigma_{ui}^2 = \sigma_u^2/N(i)$ a w_{ij} jsou prvky matice sousednosti. Prvky na diagonále definujeme jako nulové, tedy $W_{ii} = a_{ii} = 0$. Matice S je zde definována jako $S = \text{diag}(\sigma_{u1},...\sigma_{un})$ Odtud pak platí, že náhodný vektor u má vícerozměrné normální rozdělení s parametry

$$u \sim N(0, (I - \alpha W)S^2),$$

kde α je konstanta zajišťující, aby rozdělení celého vektoru u bylo správně definováno, konkrétně variační matice $(I - \alpha W)S^2$ musí být pozitivně definitní. Pro v_i pak předpokládáme

$$v_i \sim N(0, \sigma_v^2).$$

Označme $\psi = (\sigma_u^2, \sigma_v^2)$, pak tento vektor nazýváme vektorem hyperparametrů. Správná volba apriorních rozdělení hyperparmetrů není jednoduchá, protože bez

důkladné znalosti chování těchto parametrů a situace, kterou modelují, může špatná volba významně ovlivnit výsledný model. Základní volbou pro BYM jsou tak zvané minimálně informativní apriorní rozdělení, ty definujeme pomocí

$$log(\tau_u) \sim logGamma(1; 0{,}0005),$$

$$log(\tau_v) \sim logGamma(1; 0{,}0005),$$

kde platí vztah $\sigma_u = \frac{1}{\tau_u}$ a $\sigma_v = \frac{1}{\tau_v}$. Srovnání s dalšími volbami rozdělení hyperparametrů nabízí například Ugarte a kol. (2014, kapitola 4.1).
Očekávaný počet nakažených můžeme získat více způsoby, nejčastější je v praxi poměr celkového počtu nakažených a počtu obyvatel bez stratifikace nebo se stratifikací podle věku a pohlaví. Mějme I územních celků a J skupin podle věku a pohlaví. Nechť $y_{ij,obs}$ značí pozorovaný počet nakažených v i-tém územním celku v j-té skupině podle pohlaví a věku a Pop_{ij} značí počet obyvatel v i-tém územním celku a j-té skupině podle pohlaví a věku. Pak platí

$$E_i = \sum_{j=1}^{J} Pop_{ij} \cdot r_j,$$

kde

$$r_j = \frac{\sum_{i=1}^{I} y_{ij,obs}}{\sum_{i=1}^{I} Pop_{ij}}.$$

Obvyklým důvodem, proč volit odhad bez standardizace, je nedostatek dat o tom, do jaké věkové skupiny a skupiny podle pohlaví nakažení přísluší. V takovém případě postupujeme stejně, jen dosazujeme $J = 1$ a vzorec se tím zjednoduší.

3.4 Ekologická regrese

Model z kapitoly 3.3 můžeme rozšířit o faktory ovlivňující dané onemocnění, můžeme pak kromě prostorového efektu pozorovat i vliv těchto faktorů na počet případů či počet úmrtí v populaci. Tyto proměnné začleníme do modelu jednoduchým rozšířením předpisu 3.6. Získáváme tak rozdělení pozorování za podmínky parametrů

$$y_i|u_i \sim Poisson(E_i \cdot \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + u_i + v_i)).$$

Stejně jako u modelu 3.6 jsou u_i prostorově strukturované efekty, v_i jsou nezávislé efekty specifické pro jednotlivé územní celky, β_0 je absolutní člen a členy $\beta_j x_{ji}$ reprezentují vliv dalších faktorů. Údaj x_{ji} je pozorování j-té charakteristiky pro i-tý územní celek, β_j je pak prediktor, který chceme odhadnout. Pro vektor $\beta = (\beta_0, \beta_1,...,\beta_n)$ předpokládáme apriorní rozdělení $N(0, 1000I)$, kde I je jednotková matice (Gómez-Rubio, 2020, kapitol 3.2).

3.5 Goodness of fit test

Nedílnou součástí modelování prostorových dat bayesovskými metodami je kontrola, jak dobře model odpovídá našim datům. Existuje mnoho možností jak model otestovat, ať už jde o prediktivní charakteristiky modelu, obálkové testy

nebo významnost členů zahrnutých do modelu.
Při výběru metody musíme vzít v úvahu nejen, co chceme zjistit, ale i jaká data máme k dispozici. Z toho důvodu například nebudeme rozvádět metodu krosvalidace, kdy se data rozdělují na modelovací a kontrolní skupinu a pro náš případ by nebylo jednoduché tyto skupiny podle požadavků metody vytvořit.
Jednou ze dvou metod, které představíme, je kontrola pomocí aposteriorní prediktivní p-hodnoty a aposteriorního prediktivního rozdělení. Ty jsou definovány, jak uvádí Blangiardo a Cameletti (2015, kapitola 5.6), následovně.

Definice 2. *Nechť y jsou pozorovaná data, $y_i^\star$ je replikované pozorování, neboli náhodná veličina se stejným rozdělením jako veličina y_i a zároveň na ní podmíněně nezávislá za daného θ, θ_i je parametr modelu, pak aposteriorní prediktivní rozdělení definujeme jako*

$$p(y_i^\star|y) = \int p(y_i^\star|\theta_i)p(\theta_i|y)d\theta_i$$

a aposteriorní prediktivní p-hodnotu jako

$$p = p(y_i^\star \leq y_i|y).$$

Z definice vyplývá, že pokud pro mnoho indexů i vychází aposteriorní prediktivní p-hodnota blízko 0 nebo 1, model příliš dobře data neaproximuje. Obdobné platí, pokud získáváme nezvykle nízké hodnoty aposteriorního prediktivního rozdělení vztaženo k jeho rozptylu.

3.5.1 Obálkový test

Druhou možností, jak otestovat model, je porovnání dat se simulacemi z odhadnutého modelu. Jak navrhuje (Blangiardo a Cameletti, 2015, Kapitola 5.6), porovnání můžeme provést pomocí souhrných charakteristik MSE (mean square error) nebo R^2. My v této kapitole nahradíme souhrné charakteristiky globálním obálkovým testem představeným v Myllymäki a kol. (2016), který je citlivější na lokální porušení nulové hypotézy. Konkrétně představíme obálkový test ve formě extreme length rank.
Hlavní myšlenkou testu je zjistit, zda jsou naše data v kontextu modelu v nějakém smyslu extrémní, jelikož je ale rozdělení některých popisných charakteristik nebo veličin těžké odvodit analyticky, je vhodnou volbou porovnávat naše pozorování oproti simulacím z modelu. Formálně řečeno nulová hypotéza našeho testu je

$$H_0: \text{ Data } y_{obs} \text{ jsou realizací odhadnutého modelu.}$$

Obálková metoda obvykle sleduje nějakou funkcionální nebo vektorovou charakteristiku procesu. Jelikož se v našem případě jedná o proces na mřížce, bude pro nás testovanou charakteristikou přímo počet nakažených v jednotlivých územních celcích y_i, kde i je index okresu. Ač zde nemá pořadí okresů žádné logicky dané pořadí, nijak to nebrání ho použít k testování.
Prvním krokem je vygenerování $m-1$ simulací z odhadnutého modelu. Je potřeba volit m dostatečně velké, například v Myllymäki a kol. (2016, kapitola 4.4) doporučují volit $m \geq 2500$, jelikož my ale budeme používat modifikovanou verzi testu, lze testovat i s menším počtem simulací. Vygenerovaná data označíme jako

vektor $y_{sim,k} = (y_{k1},...,y_{kn})$ pro $k = 1,...m$, kde n značí počet územních celků. Za $y_{sim,1}$ dosazujeme data, která chceme testovat.
Nejdříve definujeme extrémní pořadí R_k vektoru $y_{sim,k}$ jako

$$R_k = \min_{i=1,...,n} R_{ki},$$

kde bodové pořadí R_{ki} je definováno předpisem (3.7). Určení bodových pořadí záleží na tom, zda testujeme jednostranným nebo oboustranným testem, proto pro každé $i = 1,...,n$ definujeme hrubé pořadí $r_{1i},...,r_{mi}$ prvků $y_{1i},...,y_{mi}$ tak, že nejmenší prvek y_{ki} má hrubé pořadí 1. Pokud dojde k rovnostem, pak hrubá pořadí zprůměrujeme. Pak definujeme R_{ki} jako

$$R_{ki} = \begin{cases} r_{ki}, & \text{pro jednostranný test, kde malá hodnota je považována za extrémní} \\ m + 1 - r_{ki}, & \text{pro jednostranný test, kde velká hodnota je považována za extrémní} \\ \min(r_{ki}, m + 1 - r_{ki}), & \text{pro oboustranný test} \end{cases} \quad (3.7)$$

Jelikož při výpočtu bodových pořadí dochází často k rovnostem, modifikujeme tuto metodu seřazením podle extreme length measure, tu dobře popisuje například článek Mrkvička a kol. (2020, sekce 2.5). Mějme $\mathbf{R}_k = (R_{k[1]}, R_{k[2]},...,R_{k[d]})$ vektor bodových pořadí seřazený od nejmenšího po nejvyšší. Extreme length measure pak definujeme jako

$$R_k^{\text{erl}} = \frac{1}{m} \sum_{k'=1}^{m} \mathbb{1}(\mathbf{R}_{k'} \prec \mathbf{R}_k),$$

kde

$$\mathbf{R}_{k'} \prec \mathbf{R}_k \leftrightarrow \exists l \leq n : R_{k'[i]} = R_{k[i]} \forall i < l, R_{k'[l]} < R_{k[l]}.$$

Na základě extreme length measure definujeme p-hodnotu jako

$$p^{\text{erl}} = 1 - \sum_{k=1}^{m} \mathbb{1}(R_1 \prec R_k)/m.$$

I v tomto případě může dojít k rovnostem mezi R_1 a R_i pro $i = 2,...,m$, ale je to výrazně méně pravděpodobné než pro metodu bez modifikace. Navíc můžeme odvodit globální obálku $\{y_{\text{low}}^{(\alpha)}, y_{\text{upp}}^{(\alpha)}\}$ tak, že platí

$$P(\exists i \in \{1,...,n\} : y_i \notin [y_{\text{low i}}^{(\alpha)}, y_{\text{upp i}}^{(\alpha)}]) \leq \alpha.$$

Definujme R_α^{erl} jako největší z hodnot $\{R_1^{\text{erl}}, R_2^{\text{erl}},...,R_m^{\text{erl}}\}$ splňující

$$\sum_{i=1}^{m} \mathbb{1}(R_i^{\text{erl}} < R_{(\alpha)}^{\text{erl}}) \leq \alpha m.$$

Zároveň definujeme množinu indexů $I_\alpha = \{i \in 1,...,m : R_i^{\text{erl}} \geq R_\alpha^{\text{erl}}\}$, pak pro oboustranný test definujeme

$$y_{\text{low i}}^{(\alpha)} = \min_{k \in I_\alpha} y_{ki},$$

$$y_{\text{upp i}}^{(\alpha)} = \max_{k \in I_\alpha} y_{ki}.$$

Výhodou globálních obálek je názorné grafické znázornění.

3.6 Aplikace na data

V této kapitole aplikujeme představenou teorii na data o počtu nakažených onemocněním Covid 19. Na rozdíl od kapitoly 2 budeme v této kapitole pracovat s absolutními počty případů a ne s jejich přepočtem na sto tisíc obyvatel.

3.6.1 Prostorový model

Jak bylo popsáno v podkapitole 3.3, budeme modelovat počty nakažených y_i v jednotlivých okresech jako

$$y_i|u_i \sim Poisson(E_i \cdot \exp(\beta_0 + u_i + v_i))$$

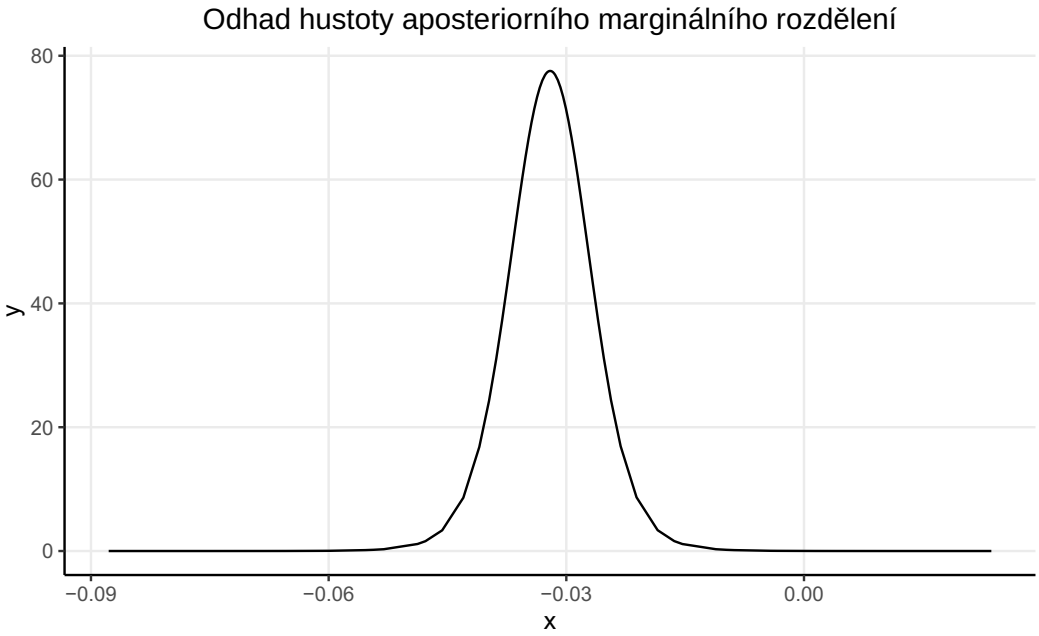
K použití tohoto modelu musíme předpokládat, že počet případů má poissonovské rozdělení. Takový předpoklad běžně využívají články mapující prostorový vývoj epidemie Covidu 19, například DiMaggio a kol. (2020) nebo Ngwira a kol. (2021). Pomocí metody INLA, získáváme odhad β_0 a odhady 154 koeficientů. Následující výsledky jsou odhady pro 45. mapovaný týden pandemie. Ten jsme vybrali jako ukázku, jelikož necelý rok po začátku pandemie už bylo dostatečně rozšířené testování na Covid 19 a tudíž z tohoto období máme data, o kterých se domníváme, že dobře mapují opravdový výskyt onemocnění.

```
> i.out3[["summary.fixed"]]

                mean          sd    0.025quant    0.5quant  0.975quant
(Intercept) -0.0320546 0.006236794 -0.04448508 -0.03205094 -0.01966396
```

```
> i.out3[["summary.random"]][["data.suicides|S|names"]]

ID        mean         sd   0.025quant    0.5quant    0.975quant
1  0.03063736 0.03138030 -0.031479160  0.03080897  0.0917430813
2 -0.52071267 0.03977228 -0.599639968 -0.52042543 -0.4434529224
3  0.13015375 0.02883375  0.073120836  0.13029631  0.1863418426
4 -0.23631047 0.03488994 -0.305469526 -0.23608655 -0.1684595918
5 -0.23748972 0.04004447 -0.316881794 -0.23722810 -0.1596237571
6 -0.26579335 0.03717849 -0.339505377 -0.26554863 -0.1935082856
7 -0.26455897 0.03456064 -0.333036274 -0.26434708 -0.1973222124
.
.
.
77 -0.15392904 0.01202610 -0.177617321 -0.15391595 -0.1303377984
78  0.02362663 0.05010702 -0.080693868  0.02496433  0.1193506344
79 -0.51021373 0.05620355 -0.616452517 -0.51185828 -0.3930830178
80  0.12019725 0.04920437  0.015121704  0.12218586  0.2121269167
81 -0.22058742 0.05118443 -0.315820201 -0.22267728 -0.1118514469
82 -0.23892054 0.05172506 -0.341458011 -0.23881862 -0.1370035966
83 -0.26901287 0.05321887 -0.376273052 -0.26848050 -0.1649975730
84 -0.25942931 0.05113592 -0.358426531 -0.26022401 -0.1550027822
.
.
.
154 -0.1505232 0.04196364 -0.2332365 -0.1515795 -0.06079557
```



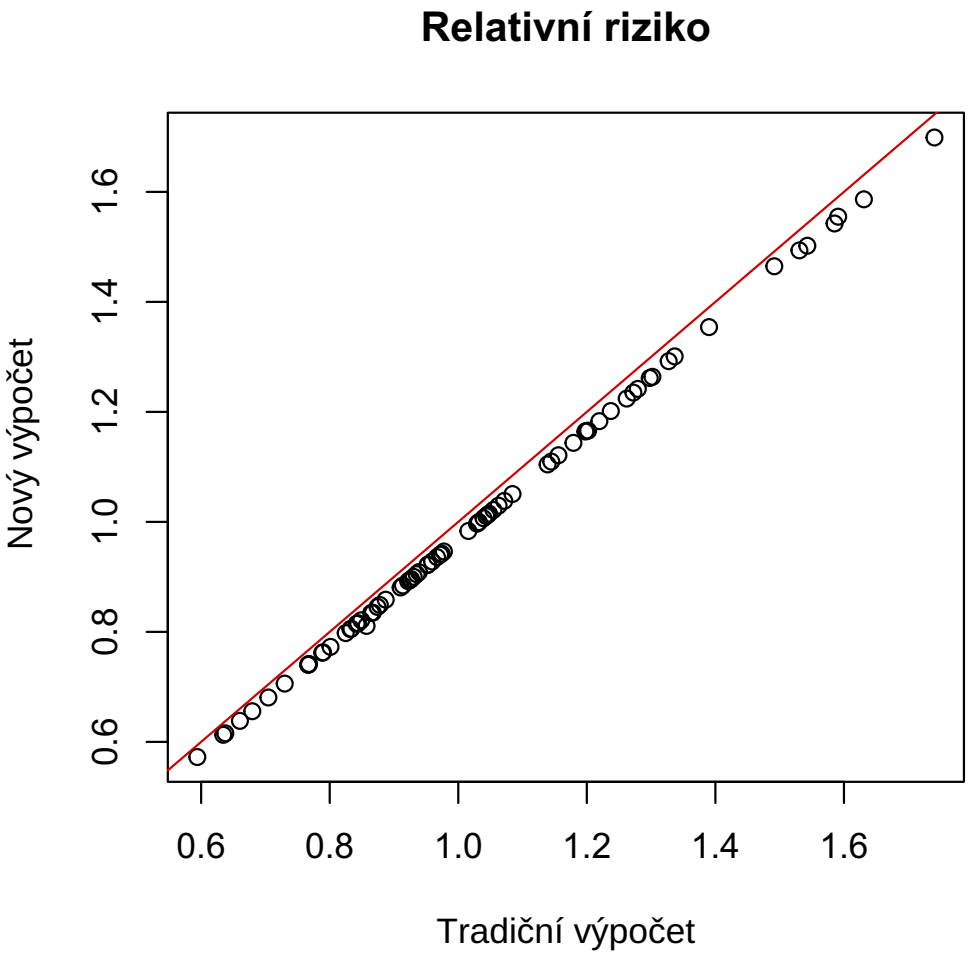
Obrázek 3.1: Odhad hustoty aposteriorního marginálního rozdělení β_0 pomocí INLA

Intercept zde reprezentuje odhad $\hat{\beta}_0$, zatímco prvních 77 koeficientů jsou odhady $\widehat{u_i + v_i} = \hat{\xi}_i$ po řadě pro i = 1,...,77 a odhady pod čísly 78 až 154 reprezentují odhady $\hat{u_i}$ po řadě pro i = 1,...,77.
Kromě těchto souhrnných informací nám také model dává informaci o tom, jak podle odhadů vypadají hustoty marginálních aposteriorních rozdělení odhadovaných parametrů. Tyto výsledky model dává ve formě matice o dvou sloupcích udávající v jakých bodech má hustota rozdělení jakou hodnotu. Výsledek můžeme snadno graficky znázornit, například pomocí obrázku 3.1

Podle Lawson (2006) je základní zkoumanou charakteristikou v prostorové epidemiologii obvykle relativní riziko, k jehož výpočtu spočtené odhady použijeme.
Jak jsme zmínili v úvodu kapitoly 3, relativní riziko je podílem pravděpodobnosti výskytu onemocnění u skupiny obyvatel vystavené danému faktoru oproti pravděpodobnosti výskytu u skupiny obyvatel, která danému faktoru vystavena nebyla. V našem případě je tímto faktorem příslušnost k danému okresu. Vzhledem k povaze zjišťování výskytu nemoci Covid-19 znamená příslušnost k okresu, že daný nemocný byl pozitivně testován v daném okrese.
Z této definice můžeme snadno odvodit výpočet odhadu relativního rizika.

$$\begin{aligned}
\zeta_i &= \frac{\frac{\hat{y_i}}{n_i}}{\frac{\sum_{j=1,j\neq i}^{m} \hat{y_j}}{\sum_{j=1,j\neq i}^{m} n_j}} = \frac{E_i \cdot \exp(\hat{\beta}_0 + \hat{u}_i + \hat{v}_i)}{n_i} \cdot \frac{\sum_{j=1,j\neq i}^{m} n_j}{\sum_{j=1,j\neq i}^{m} E_j \cdot \exp(\hat{\beta}_0 + \hat{u}_j + \hat{v}_j)} \\
&= \frac{y \cdot \frac{n_i}{n} \cdot \exp(\hat{\beta}_0 + \hat{u}_i + \hat{v}_i)}{n_i} \cdot \frac{\sum_{j=1,j\neq i}^{m} n_j}{\sum_{j=1,j\neq i}^{m} y \frac{n_j}{n} \cdot \exp(\hat{\beta}_0 + \hat{u}_j + \hat{v}_j)} \\
&= \exp(\hat{u}_i + \hat{v}_i) \cdot \frac{\sum_{j=1,j\neq i}^{m} n_j}{\sum_{j=1,j\neq i}^{m} n_j \cdot \exp(\hat{u}_j + \hat{v}_j)}
\end{aligned}$$

V epidemiologické literatuře se v tomto kontextu jako relativní riziko udává pouze



Obrázek 3.2: Znázornění relativního rizika v 45. týdnu nákazy

člen $\exp(\hat{u}_i + \hat{v}_i)$, což ukazuje jistou nekonzistentnost s udávanou definicí relativního rizika. Porovnání výsledků pomocí těchto dvou metod znázorňuje obrázek 3.2. Pro účely grafického znázornění můžeme převést tuto spojitou veličinu na kategorickou, jelikož jsou rozdíly mezi metodami poměrně malé, výsledný obrázek 3.3(nahoře) pak vypadá pro obě metody stejně.

3.6.2 Ekologická regrese

Jelikož na výskyt a šíření nemoci mají často vliv další faktory, ukážeme nyní na datech o počtu pozitivně testovaných na Covid 19, jak tyto proměnné do modelu zařadit. Konkrétně budeme pro ilustraci uvažovat hustotu osídlení a průměrný věk obyvatel v jednotlivých okresech. Budeme uvažovat následující model.

$$y_i|u_i \sim Poisson(E_i \cdot \exp(\beta_0 + \beta_1 \cdot \text{hustota osídlení} + \beta_2 \cdot \text{průměrný věk} + u_i + v_i))$$

Pro 45. týden nákazy pak dostáváme následující výsledky.

```
>round(i.out[["summary.fixed"]],digits=6)
```

```
                 mean       sd   0.025quant  0.5quant   0.975quant
(Intercept)  0.537565 1.006761  -1.443499    0.537356   2.517877
x1           0.000009 0.000079  -0.000146    0.000009   0.000164
x2          -0.013415 0.023704  -0.060084   -0.013411   0.033188
```

Proměnná `x1` ve výstupu výše reprezentuje hustotu osídlení a `x2` preprezentuje průměrný věk obyvatel. Obdobně jako u modelu, který nezahrnoval tyto proměnné, získáváme odhady prostorových proměnných

```
> i.out[["summary.random"]][["data.suicides|S|names"]]

ID         mean         sd   0.025quant     0.5quant    0.975quant
1       0.023313724 0.03470589 -0.045206835  0.023439418  0.0910728102
2      -0.530639768 0.04358125 -0.616956348 -0.530386544 -0.4458079092
3       0.130706862 0.02918166  0.073000936  0.130845905  0.1875866714
4      -0.238791585 0.03568092 -0.309496208 -0.238570515 -0.1693806486
5      -0.242723659 0.04208200 -0.326039643 -0.242490722 -0.1607784730
6      -0.270119250 0.03809608 -0.345594262 -0.269888367 -0.1959959901
7      -0.269292802 0.03596656 -0.340488017 -0.269096270 -0.1992551288
.
.
.
152     0.217029721 0.04729512  0.12108546  0.217500871  0.30994392
153     0.285124242 0.04987936  0.17915040  0.286951886  0.37889775
154    -0.182368472 0.19778959 -0.57271671 -0.182070590  0.20598156
```

Obdobně jako pro model bez přidaných kovariát můžeme spočíst relativní riziko jako $\zeta_i = \exp(u_i + v_i)$ a znázornit ho graficky pomocí obrázku 3.3 (dole). Koeficienty β_1,...,β_n zde nejsou zahrnuty do výpočtu, protože rizikový faktor, který chceme obrázkem znázornit, je příslušnost k danému okresu. Zahrnutím dalších kovariát do modelu se ovšem mění výsledný odhad $\hat{u}_i + \hat{v}_i$ a můžeme tak získat odlišný odhad relativního rizika než v kapitole 3.6.1. Z obrázku 3.3 (dole) můžeme ale vidět, že v našem případě se výsledky příliš neliší.

3.6.3 Goodness of fit test

Součástí modelování dat by měla být také kontrola toho, jak dobře model odpovídá datům. K tomu použijeme prediktivní charakteristiky a obálkový test z kapitoly 3.5.

Výpočet prediktivních charakteristik, konkrétně prediktivního rozdělení a prediktivní p-hodnoty, je v R součástí knihovny INLA. Spočtené výsledky můžeme ilustrovat například pomocí histogramu aposteriorních prediktivních p-hodnot 3.5 nebo grafem 3.4.

Z obrázků 3.4 a 3.5 můžeme vidět, že model dobře aproximuje data, jelikož v grafu 3.4 můžeme vidět, že odhadnuté hodnoty velmi dobře odpovídají naměřeným hodnotám a z 3.5 vidíme, že aposteriorní prediktivní p-hodnoty nabývají ve většině případů hodnoty mezi 0,4 a 0,6 a téměř žádné se nevyskytují blízko nuly nebo jedné.

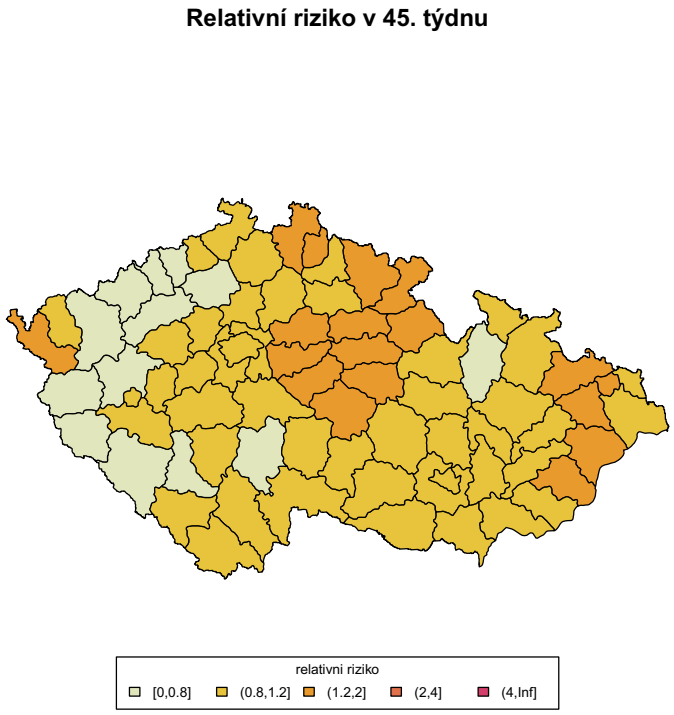
Kromě těchto charakteristik můžeme vhodnost modelu otestovat obálkovým testem z kapitoly 3.5.1, kde pozorované hodnoty porovnáváme se sadou simulací

z odhadnutého modelu. Při testování jsme použili 9999 simulací z odhadnutého modelu. Jelikož z kapitoly 3.1 víme, že veličiny y_i pro $i = 1,...,77$ jsou podmíněně nezávislé za daných hodnot latentního gausovského pole, můžeme pro každý okres simulovat počty případů nezávisle na ostatních okresech z rozdělení

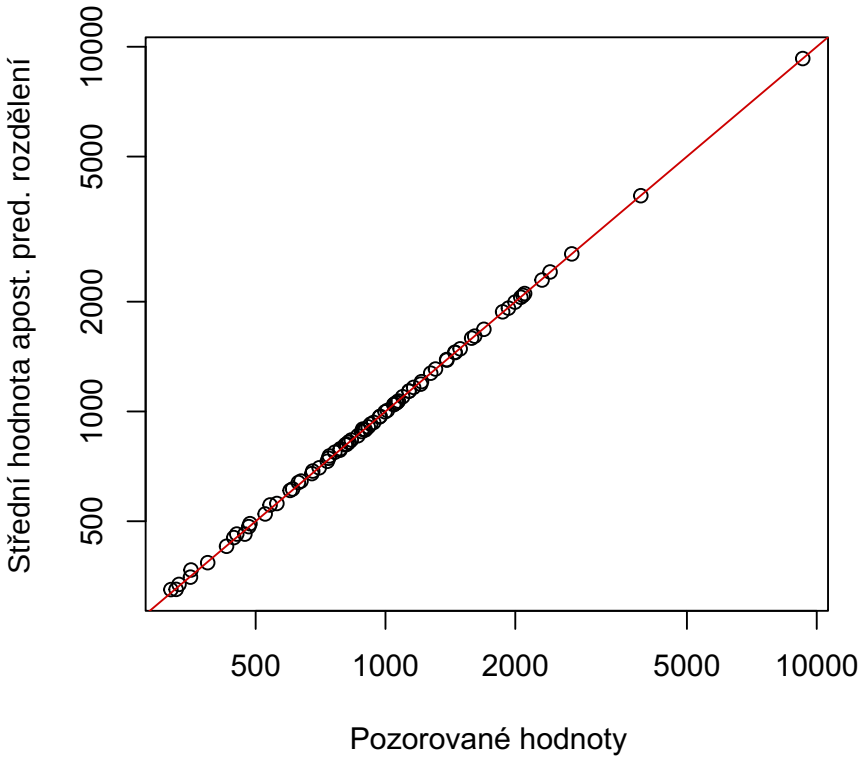
$$y_i \sim Poisson(E_i \cdot \exp(\hat{\beta}_0 + \hat{u}_i + \hat{v}_i)).$$

Výsledek pak díky R knihovně GET získáváme i s grafickým znázorněním 3.6.

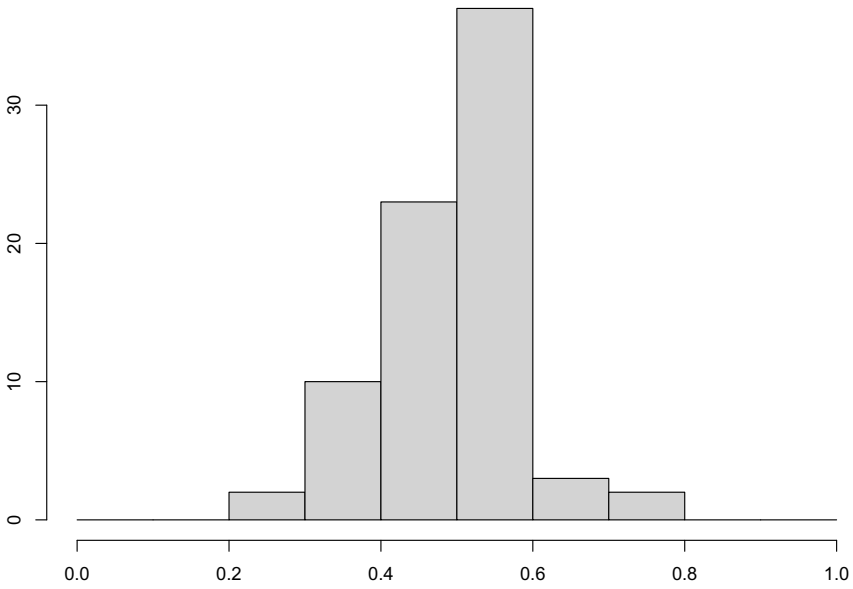
Z p-hodnoty $p = 0{,}144$ můžeme udělat závěr, že na hladině $\alpha = 0{,}05$ nezamítáme nulovou hypotézu, a tudíž nemůžeme vyloučit, že data jsou realizací našeho modelu. Tomu odpovídá i grafické znázornění, kdy datová křivka je celá obsažena v obálce.



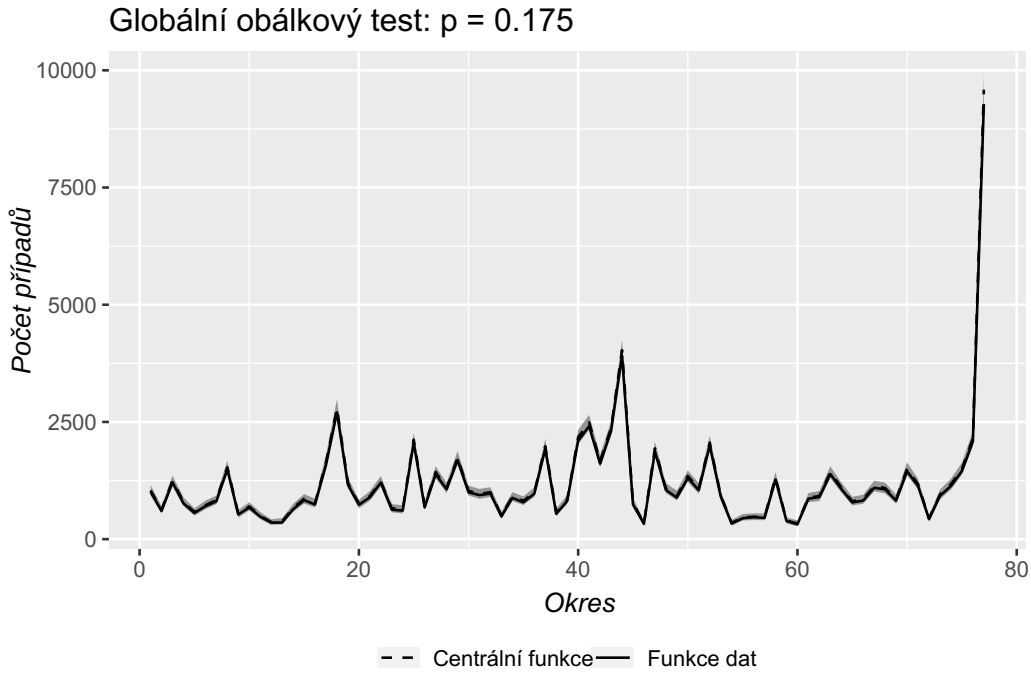
Obrázek 3.3: Znázornění relativního rizika v 45. týdnu nákazy pomocí čistě prostorového modelu (nahoře) a pomocí ekologické regrese (dole)



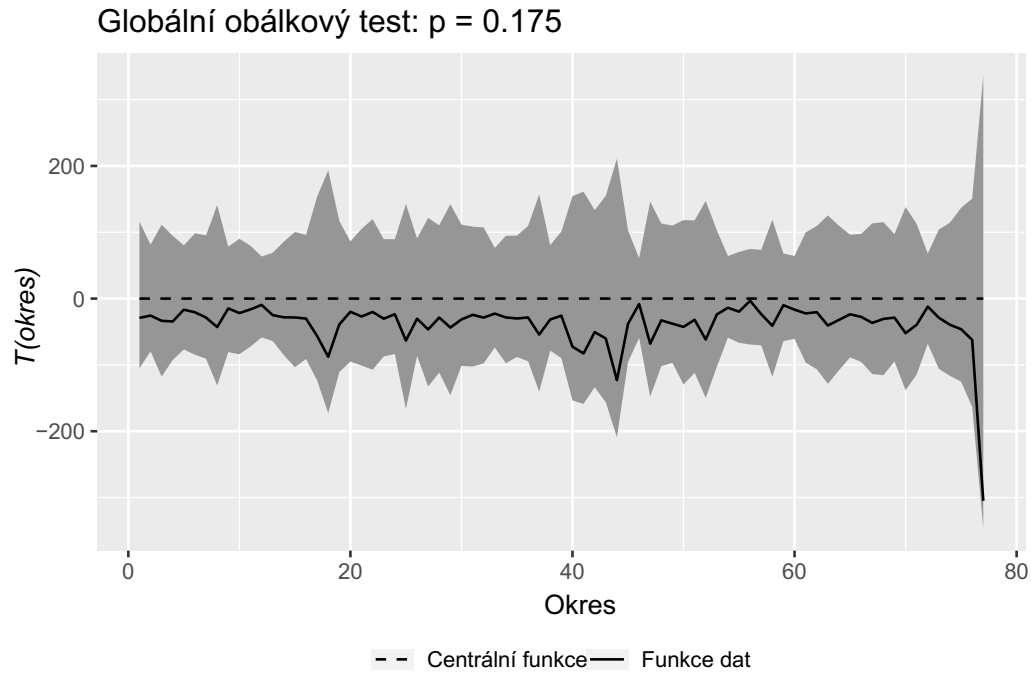
Obrázek 3.4: Graf aposteriorní střední hodnoty proti pozorovaným hodnotám, osy x a y mají logaritmické měřítko.



Obrázek 3.5: Histogram aposteriorních prediktivních p-hodnot



Obrázek 3.6: Výsledek obálkového testu pro model nákazy Covidem-19 v 45. týdnu sledování výskytu nemoci



Obrázek 3.7: Výsledek obálkového testu pro model nákazy Covidem-19 v 45. týdnu sledování výskytu nemoci. Horní a dolní obálka i data jsou posunuty tak, aby centrální linie obálky byla všude nulová. Toto posunutí slouží k přehlednějšímu grafickému znázornění.

3.7 Implementace

K výsledkům z kapitoly 3.6 jsme došli pomocí implementace v R verze 4.1.3 s využitím R Studia 2022.02.0 Build 443 a knihovny INLA verze 21.11.22. Jelikož práce s touto knihovnou není pro většinu uživatelů bez problémů, následující část věnujeme představení hlavních funkcí použitých při implementaci.
K získání výsledku musíme mít definovanou relaci sousedství územních celků. Ať už je tato definice součástí dat nebo ji vytvoříme ručně či pomocí funkce `poly2nb` z balíku `spdep`, musíme nakonec získat formát `list` se stejným počtem položek jako je počet územních celků v datech. Každá položka seznamu je pak výčet indexů, které reprezentují územní celky, s kterými daný okres sousedí.
Nyní už můžeme definovat tvar modelu.

```
formula1 <- y ~ 1 + f(data.spat\$names,
                      model="bym",
                      graph=LDN.adj,
                      scale.model=TRUE)
```

Složky vektoru y zde reprezentují počty nakažených v jednotlivých územních celcích, `model="bym"` pak říká, že používáme Besag-York-Mollié model a jako `graph` dosazujeme výše zmíněný list sousednosti. Kromě těchto povinných proměnných zde můžeme nastavit také apriorní rozdělení, s kterými chceme pracovat, my jsme zvolili přednastavené minimálně informativní apriorní rozdělení hyperparametrů. Dalším důležitým prvkem je argument `scale.model=TRUE`. Toto nastavení se vztahuje k zobecněnému rozptylu vektoru parametrů θ a zajistí porovnatelnost odhadů v různých modelech. Více tuto myšlenku rozvání Blangiardo a Cameletti (2015). Jádrem implementace je funkce inla.

```
model.output <- inla(formula1,
                     family="poisson",
                     data=data.spat,
                     offset = log(E),
                     control.predictor=list(link=1, compute=TRUE),
                     control.compute=list(return.marginals.predictor=TRUE),
                     verbose=TRUE)
```

Proměnnou formula jsme vysvětlili výše, `family="poisson"` říká, pomocí jakého rozdělení modelujeme počet nakažených, a E je očekávaný počet nakažených. Význam nastavení `control.predictor` a `control.compute` je nastavení výpočtu částí modelu, které se v základní podobě nepočítají, ale my je potřebujeme k výpočtu aposteriorních prediktivních charakteristik. Data, která dosazujeme do modelu, musí být ve specifickém formátu, měl by obsahovat obvykle tři sloupce: jména okresů, které ale musí být očíslované, nikoliv jejich názvy slovy, druhou proměnnou je obvykle počet nakažených y a třetí očekávaný počet nakažených E.
Celá implementace je k nalezení v kódu, který je přiložen k této práci a vychází z kódů doprovázejících knížku Blangiardo a Cameletti (2015), které jsou k dispozici online na `https://sites.google.com/a/r-inla.org/stbook/`.

4. Časoprostorový model

Pandemie mají kromě prostorového rozložení případů další zajímavý aspekt, a tím je vývoj v čase. V následující kapitole proto ukážeme, jak zobecnit prostorový model na časoprostorový.

4.1 Teorie

Časoprostorový model obdobně jako prostorový popisuje náhodnou veličinu počtu případů v jednotlivých okresech pomocí Poissonova rozdělení.

$$y_{it} \sim Poisson(\lambda_{it}),$$

kde i značí identifikátor územního celku a t značí čas. V tomto kontextu pracujeme s časem jako diskrétní veličinou. Obvykle se jedná o dny, týdny nebo měsíce. Parametr λ_{it} závisí na dalších parametrech.

$$\lambda_{it} = E_{it} \cdot \exp(\beta_0 + u_i + v_i + Temp_t), \tag{4.1}$$

kde β_0, u_i, v_i zastávají stejnou roli jako u prostorového modelu. Člen $Temp_i$ značí obecný tvar členu závislého na čase, jednoduchou volbou pro tento člen může být například

$$Temp_t = (\gamma + \delta_i) \cdot t \tag{4.2}$$

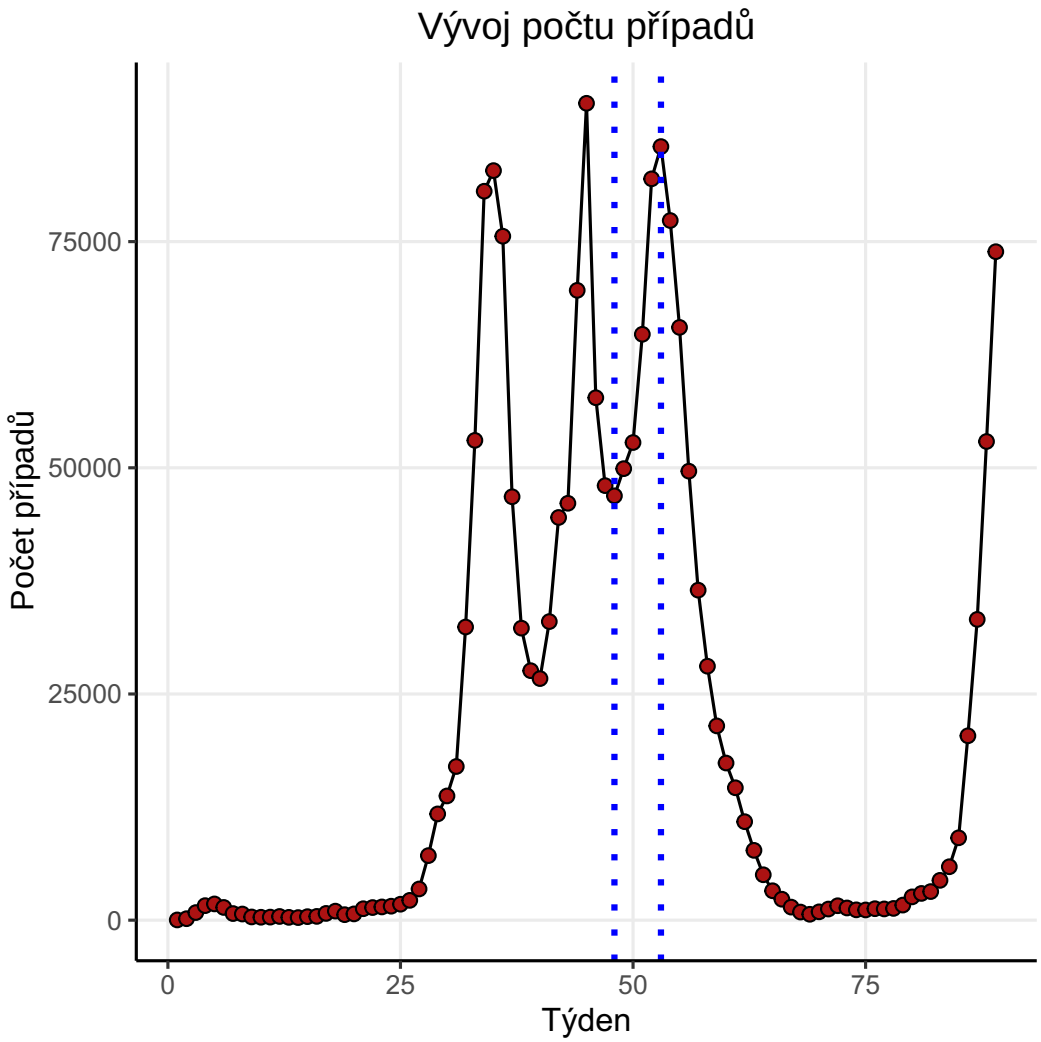
Člen γ zde udává globální trend v čase, pro který obvykle uvažujeme minimálně informativní apriorní rozdělení $N(0{,}1000)$, zatímco člen δ_i značí interakci mezi prostorovými a časovými vlivy. Předpokládáme

$$\delta_i \sim N\Big(0, \frac{1}{\tau_\delta}\Big),$$

kde τ_δ značí hyperparametr. Pro u_i a v_i platí stejné předpoklady jako v kapitole 3.3.
Jak uvádí (Blangiardo a Cameletti, 2015, kapitola 7), nevýhodou reprezentace času jako v předpisu (4.2) je, že předpokládáme, že vývoj případů v čase je v jednotlivých okresech exponenciální. Takový model je pak problém aplikovat mimo jiné na epidemická data, kde se trend v čase rychle mění, tam je pak možné tento model použít jen na krátký časový úsek. Další možné volby časových členů stejně jako rozšíření ekologické regrese o časový aspekt jsou k nalezení například v Schrödle a Held (2010).

4.2 Aplikace na data

Jak jsme uvedli, aplikovat časoprostorový model z rovnic (4.1) a (4.2) na celý průběh epidemie Covid 19 je problematické, jelikož trend v čase je velmi proměnlivý. Proto aplikujeme tento model na data o počtu případů v jednotlivých okresech v týdnech $t = 48, 49, 50, 51, 52, 53$. Z obrázku 4.1 můžeme vidět, že se jedná o relativně krátký časový úsek, kdy máme dostatek dat v porovnání například s počátkem sledování počtu nakažených.
Z výpočtu pomocí R-INLA získáváme odhady koeficientů.



Obrázek 4.1: Průběh týdenních přírůtků nakažených Covidem 19 mezi týdny 1 a 89

```
>model.par[["summary.fixed"]]

                    mean          sd  0.025quant    0.5quant  0.975quant
(Intercept) -0.03102924 0.006152323 -0.04319294 -0.03102478 -0.01891079
week        -0.02264417 0.001182145 -0.02496462 -0.02264439 -0.02032461
```

Proměnná week zde značí koeficient γ z vyjádření (4.2). Dále získáváme odhady prostorových koeficientů. Prvních 77 značí odhady $\hat{u}_i + \hat{v}_i$ a druhých 77 značí odhady $\hat{u}_i$ z rovnice (4.1).

```
>model.par[["summary.random"]][["SpatTempData|S|NAME"]]

     ID         mean          sd   0.025quant     0.5quant     0.975quant
1     1 -0.406696141 0.04453516 -0.494699107 -0.406504064 -0.3198444991
2     2 -0.594296157 0.04682317 -0.686857884 -0.594081161 -0.5030168758
3     3 -0.117546961 0.03788181 -0.192325247 -0.117410350 -0.0435997776
4     4 -0.425493032 0.04359140 -0.511617202 -0.425309887 -0.3404688307
```

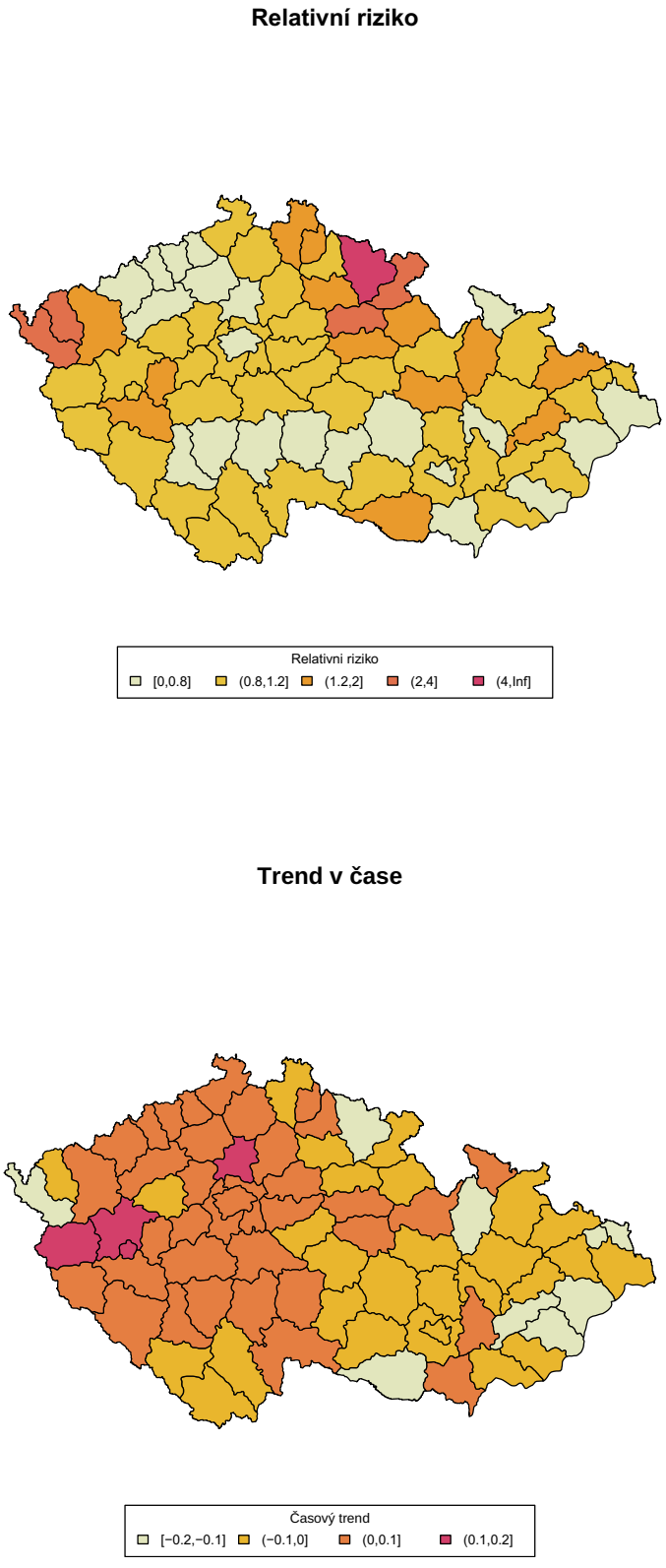
```
5      5 -0.289902419 0.04854435 -0.385871300 -0.289678220 -0.1952706350
6      6 -0.507991254 0.04770038 -0.602292929 -0.507770236 -0.4150070489
7      7 -0.524472713 0.04438661 -0.612178425 -0.524282537 -0.4379072899
.
.
.
77    77 -0.243881017 0.01412472 -0.271665736 -0.243868932 -0.2161919935
78    78 -0.406348765 0.05399568 -0.513109495 -0.406281121 -0.3000550754
79    79 -0.590628266 0.05610186 -0.699692157 -0.591117303 -0.4783092488
80    80 -0.119048648 0.04878310 -0.216819059 -0.118641801 -0.0239032330
81    81 -0.422147870 0.05277793 -0.524818647 -0.422602330 -0.3164453953
82    82 -0.288130403 0.05649699 -0.399179071 -0.288231193 -0.1765275263
83    83 -0.507726610 0.05659142 -0.619658188 -0.507620552 -0.3964864146
84    84 -0.523227870 0.05373603 -0.628936080 -0.523328524 -0.4169307810
.
.
.
154 154 -0.24295814 0.03422164 -0.3119034920 -0.243305949 -0.171094419
```

A nakonec ještě získáváme odhady koeficientů $\hat{\delta}_i$ značících interakci časového a prostorového vlivu.

```
>model.par[["summary.random"]][["SpatTempData|S|names"]]

 ID          mean           sd    0.025quant     0.5quant    0.975quant
1   1  0.070292592 0.010184840  0.0503348501  0.070278949  9.030769e-02
2   2  0.086854637 0.010618965  0.0660524574  0.086838280  1.077284e-01
3   3  0.056293454 0.008717836  0.0392033896  0.056284172  7.341937e-02
4   4  0.087972254 0.009891689  0.0685899370  0.087958655  1.074122e-01
5   5  0.064276472 0.011170729  0.0423850167  0.064262087  8.622763e-02
6   6  0.080533706 0.010854539  0.0592685430  0.080517503  1.018692e-01
7   7  0.088532016 0.010059095  0.0688232860  0.088517655  1.083023e-01
.
.
.
77 77  0.048583586 0.003179734  0.0423431966  0.048582597  5.482380e-02
```

Je důležité si uvědomit, že v tomto kontextu záporné hodnoty některých odhadů $\hat{\delta}_i$ nemusí znamenat, že absolutní počet případů klesá. Jde pouze o porovnání s ostatními okresy, pokud celkový počet případů roste jako v našem případě, znamená záporná hodnota odhadu, že počet případů v daném okresu buď roste pomaleji v porovnání se zbytkem okresů, nebo v extrémním případě může počet případů v daném okrese i klesat. Obdobně neplatí, že by kladná hodnota odhadu indikovala vždy růst počtu případů v daném okrese.
Obdobně jako u prostorového modelu můžeme znázornit relativní riziko v jednotlivých okresech, navíc můžeme pro jednotlivé okresy znázornit i trend v čase.



Obrázek 4.2: Relativní riziko nákazy v jednotlivých okresech v týdnech (nahoře) a trend nákazy v čase (dole) v týdnech 48-53

Závěr

V této práci jsme zmapovali základní metody pro zkoumání agregovaných epidemiologických dat.
V první části jsme se zaměřili na testování významnosti prostorové autokorelace pomocí statistik založených na Moranově a Gearyho indexu. Při aplikaci na data jsme ukázali, že záleží na volbě předpokladů testů a v závislosti na jejich volbě můžeme získat různé výsledky. Souhrně ale můžeme říci, že všechny testy ve velké většině případů zamítly nulovou hypotézu, která předpokládala, že počet nakažených na sto tisíc obyvatel je prostorově nezávislý, což vzhledem k infekční povaze zkoumaného onemocnění dává smysl.
Druhá část práce byla věnována prostorovému modelu odhadovanému pomocí Integrated Nested Laplace Approximations. Představili jsme základní algoritmus metody a jeden z v praxi nejčastěji používaných modelů, tedy Beság-York-Mollié model. Dále jsme ukázali, jak model rozšířit o další faktory, jejichž vliv na onemocnění se v praxi často zkoumá. Poté jsme model rozšířili na časoprostorový, který nám nabízí zajímavý pohled na časový vývoj počtu případů v jednotlivých územních celcích.
Dále jsme ukázali možnosti, jak zkontrolovat, zda odhadnutý model dobře odpovídá datům, a to jak pomocí prediktivní aposteriorních charakteristik, tak obálkových testů.
Velká část práce byla věnována aplikaci na reálná data o počtu nakažených Covidem 19 v České republice. Ukázali jsme, jak teorii implementovat pomocí balíku R-INLA. Jelikož práce s touto knihovnou může být poměrně náročná, může přiložený kód sloužit jako ukázka, jak s touto knihovnou pracovat.

Seznam použité literatury

Beaglehole R., Bonita R., K. T. (2002). *Basic epidemiology*. World Health Organisation. ISBN 9241544465.

Besag, J., York, J. a Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20. doi: 10.1007/bf00116466.

Best, N., Richardson, S. a Thomson, A. (2005). A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**(1), 35–59. doi: 10.1191/0962280205sm388oa.

Bivand, R. S. a Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, **27**(3), 716–748. doi: 10.1007/s11749-018-0599-x.

Blangiardo, M. a Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, Ltd. doi: 10.1002/9781118950203.

Cliff, A. D. a Ord, K. (1970). Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography*, **46**, 269. doi: 10.2307/143144.

ČSÚ (2018). Základní charakteristika okresů. [online], [cit. 20.03.2022]. URL `https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&katalog=31737&pvo=RSO07D`.

ČSÚ (2021). Věkové složení obyvatelstva - 2020. [online], [cit. 20.03.2022]. URL `https://www.czso.cz/csu/czso/vekove-slozeni-obyvatelstva-2020`.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Taylor & Francis Ltd. ISBN 978-1-4665-6024-6.

DiMaggio, C., Klein, M., Berry, C. a Frangos, S. (2020). Black/african american communities are at highest risk of COVID-19: spatial modeling of new york city ZIP code–level testing results. *Annals of Epidemiology*, **51**, 7–13. doi: 10.1016/j.annepidem.2020.08.012.

GADM (2021). Gadm database of global administrative areas, version 4.0.4 [cit 1.10.2021]. URL `https://gadm.org/`.

Gómez-Rubio, V. (2020). *Bayesian Inference with INLA*. Chapman & Hall/CRC Press. ISBN 9781032174532. Boca Raton, FL.

Illian, Antti Penttinen, H. S. D. S. J. I. (2008). *Statistical Analysis and Modelling*. John Wiley & Sons. ISBN 978-0-470-01491-2.

Khan, K., Luo, H. a Xi, W. (2021). Computing with r-inla: Accuracy and reproducibility with implications for the analysis of covid-19 data. doi: 10.48550/ARXIV.2111.01285.

Komenda M., Panoska P., B. V. (2021). Covid-19: Přehled aktuální situace v Čr. onemocnění aktuálně [online]. [cit. 14.11.2021]. URL `https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19`. Vývoj: společné pracoviště ÚZIS ČR a IBA LF MU. ISSN 2694-9423.

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology*. John Wiley & Sons. ISBN 978-0-470-01484-4.

Lee, D. (2011). A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, **2**(2), 79–89. doi: 10.1016/j.sste.2011.03.001.

Martino, S. a Riebler, A. (2019). Integrated nested laplace approximations (inla). doi: 10.48550/ARXIV.1907.01248.

Mrkvička, T., Myllymäki, M., Jílek, M. a Hahn, U. (2020). A one-way ANOVA test for functional data with graphical interpretation. *Kybernetika*, pages 432–458. doi: 10.14736/kyb-2020-3-0432.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H. a Hahn, U. (2016). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(2), 381–404. doi: 10.1111/rssb.12172.

Ngwira, A., Kumwenda, F., Munthali, E. C. a Nkolokosa, D. (2021). Spatial temporal distribution of COVID-19 risk during the early phase of the pandemic in malawi. *PeerJ*, **9**, e11003. doi: 10.7717/peerj.11003.

Opitz, T. (2017). Latent gaussian modeling and inla: A review with focus on space-time applications. doi: 10.48550/ARXIV.1708.02723.

Oyana, T. J. (2020). *Spatial Analysis with R*. Taylor & Francis Ltd. ISBN 0367860856. URL `https://www.ebook.de/de/product/39501684/tonny_j_oyana_spatial_analysis_with_r.html`.

Rue, H., Martino, S. a Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71** (2), 319–392. doi: 10.1111/j.1467-9868.2008.00700.x.

Schrödle, B. a Held, L. (2010). A primer on disease mapping and ecological regression using $${\texttt{INLA}}$$. *Computational Statistics*, **26**(2), 241–258. doi: 10.1007/s00180-010-0208-2.

Sen, A. (2010). Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical Analysis*, **8**(2), 175–184. doi: 10.1111/j.1538-4632.1976.tb01066.x.

Ugarte, M. D., Adin, A., Goicoa, T. a Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate bayesian inference. *Statistical Methods in Medical Research*, **23**(6), 507–530. doi: 10.1177/0962280214527528.

Wood, S. N. (2019). Simplified integrated nested laplace approximation. *Biometrika*. doi: 10.1093/biomet/asz044.

Seznam obrázků

1.1 Znázornění relace sousedství mezi okresy 4
1.2 Matice sousednosti okresů, černá pole značí, že okresy jsou v sousedské relaci, bílá, že nejsou, pořadí okresů je dostupné v tabulce A.1. 5
1.3 Graf znázorňující četnost počtu sousedů okresů v České republice 6

2.1 Výsledky testů autokorelace na hladině $\alpha = 0{,}05$ 10
2.2 Graf četnosti zamítnutí nulové hypotézy pro jednotlivé testy. Popisy osy x u grafů značí spočtenou p-hodnotu testů. 11
2.3 Histogram počtu případů reprezentujících, kolik testů zamítlo nulovou hypotézu v jednotlivých týdnech 11
2.4 Znázornění počtu případů Covidu 19 na počet obyvatel v jednotlivých okresech pro týdny, kdy žádný z testů nezamítl nulovou hypotézu. Barevná škála značí nejsvětlejší barvou okresy s nízkým počtem nakažených na sto tisíc obyvatel, naopak tmavá červená značí nejvyšší počty případů. Barevná škála je pro každý z týdnů jiná. 12
2.5 Znázornění počtu případů Covid 19 na počet obyvatel v jednotlivých okresech pro vybrané týdny, kdy všechny testy zamítly nulovou hypotézu. Barevná škála značí nejsvětlejší barvou okresy s nízkým počtem nakažených na sto tisíc obyvatel, naopak tmavá červená značí nejvyšší počty případů. Barevná škála je pro každý z týdnů jiná. 13

3.1 Odhad hustoty aposteriorního marginálního rozdělení β_0 pomocí INLA 23
3.2 Znázornění relativního rizika v 45. týdnu nákazy 24
3.3 Znázornění relativního rizika v 45. týdnu nákazy pomocí čistě prostorového modelu (nahoře) a pomocí ekologické regrese (dole) . . 27
3.4 Graf aposteriorní střední hodnoty proti pozorovaným hodnotám, osy x a y mají logaritmické měřítko. 28
3.5 Histogram aposteriorních prediktivních p-hodnot 28
3.6 Výsledek obálkového testu pro model nákazy Covidem-19 v 45. týdnu sledování výskytu nemoci 29
3.7 Výsledek obálkového testu pro model nákazy Covidem-19 v 45. týdnu sledování výskytu nemoci. Horní a dolní obálka i data jsou posunuty tak, aby centrální linie obálky byla všude nulová. Toto posunutí slouží k přehlednějšímu grafickému znázornění. 29

4.1 Průběh týdenních přírůtků nakažených Covidem 19 mezi týdny 1 a 89 32
4.2 Relativní riziko nákazy v jednotlivých okresech v týdnech (nahoře) a trend nákazy v čase (dole) v týdnech 48-53 34

A. Indexy okresů a týdnů

V následující příloze jsou k nahládnutí dvě tabulky. Tabulka A.1 ukazuje, jak jsou přiřazeny okresům indexy, které jsme používali při analýze dat. Tabulka A.2 obdobně ukazuje přiřazení indexů týdnům.

Index	Nazev.okresu	Index	Nazev okresu
1	Ústí nad Labem	40	Frýdek-Místek
2	Chomutov	41	Karviná
3	Děčín	42	Nový Jičín
4	Litoměřice	43	Opava
5	Louny	44	Ostrava-město
6	Most	45	Šumperk
7	Teplice	46	Jeseník
8	České Budějovice	47	Olomouc
9	Český Krumlov	48	Přerov
10	Jindřichův Hradec	49	Prostějov
11	Písek	50	Ústí nad Orlicí
12	Prachatice	51	Chrudim
13	Strakonice	52	Pardubice
14	Tábor	53	Svitavy
15	Břeclav	54	Domažlice
16	Blansko	55	Klatovy
17	Brno-venkov	56	Plzeň-jih
18	Brno-město	57	Plzeň-sever
19	Hodonín	58	Plzeň-město
20	Vyškov	59	Rokycany
21	Znojmo	60	Tachov
22	Cheb	61	Benešov
23	Karlovy Vary	62	Beroun
24	Sokolov	63	Kladno
25	Hradec Králové	64	Kolín
26	Jičín	65	Kutná Hora
27	Náchod	66	Mělník
28	Rychnov nad Kněžnou	67	Mladá Boleslav
29	Trutnov	68	Nymburk
30	Žďár nad Sázavou	69	Příbram
31	Havlíčkův brod	70	Praha-východ
32	Jihlava	71	Praha-západ
33	Pelhřimov	72	Rakovník
34	Třebíč	73	Kroměříž
35	Česká Lípa	74	Uherské Hradiště
36	Jablonec nad Nisou	75	Vsetín
37	Liberec	76	Zlín
38	Semily	77	Hlavní město Praha
39	Bruntál		

Tabulka A.1: Tabulka okresů s přiřazenými indexy, jak jsou používány během celé práce

Index	Prvni den tydne	Index	Prvni den tydne
1	01.03.2020	46	10.01.2021
2	08.03.2020	47	17.01.2021
3	15.03.2020	48	24.01.2021
4	22.03.2020	49	31.01.2021
5	29.03.2020	50	07.02.2021
6	05.04.2020	51	14.02.2021
7	12.04.2020	52	21.02.2021
8	19.04.2020	53	28.02.2021
9	26.04.2020	54	07.03.2021
10	03.05.2020	55	14.03.2021
11	10.05.2020	56	21.03.2021
12	17.05.2020	57	28.03.2021
13	24.05.2020	58	04.04.2021
14	31.05.2020	59	11.04.2021
15	07.06.2020	60	18.04.2021
16	14.06.2020	61	25.04.2021
17	21.06.2020	62	02.05.2021
18	28.06.2020	63	09.05.2021
19	05.07.2020	64	16.05.2021
20	12.07.2020	65	23.05.2021
21	19.07.2020	66	30.05.2021
22	26.07.2020	67	06.06.2021
23	02.08.2020	68	13.06.2021
24	09.08.2020	69	20.06.2021
25	16.08.2020	70	27.06.2021
26	23.08.2020	71	04.07.2021
27	30.08.2020	72	11.07.2021
28	06.09.2020	73	18.07.2021
29	13.09.2020	74	25.07.2021
30	20.09.2020	75	01.08.2021
31	27.09.2020	76	08.08.2021
32	04.10.2020	77	15.08.2021
33	11.10.2020	78	22.08.2021
34	18.10.2020	79	29.08.2021
35	25.10.2020	80	05.09.2021
36	01.11.2020	81	12.09.2021
37	08.11.2020	82	19.09.2021
38	15.11.2020	83	26.09.2021
39	22.11.2020	84	03.10.2021
40	29.11.2020	85	10.10.2021
41	06.12.2020	86	17.10.2021
42	13.12.2020	87	24.10.2021
43	20.12.2020	88	31.10.2021
44	27.12.2020	89	07.11.2021
45	3.1.2021		

Tabulka A.2: Tabulka týdnů s přiřazenými indexy, jak jsou používány během celé práce