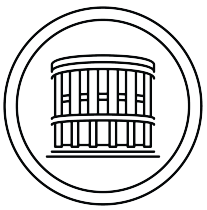
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



INFERENCIA LOGISTICKÉHO RASTU V DÁTACH S NÁHODNÝM ČASOVÝM POSUNOM

DIPLOMOVÁ PRÁCA

2026 Bc. Robert Valaška

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

INFERENCIA LOGISTICKÉHO RASTU V DÁTACH S NÁHODNÝM ČASOVÝM POSUNOM

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Katarína Boďová, PhD.

Bratislava 2026 **Bc. Robert Valaška**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Robert Valaška
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Inferencia logistického rastu v dátach s náhodným časovým posunom
Inference of logistic growth in randomly lagged data

Anotácia: Logistická obyčajná diferenciálna rovnica dobre aproximuje viacero biologických situácií. Jednou z aplikácií je evolučná dynamika vírusových variantov v populácii, v ktorej sú niektoré varianty lepšie adaptované (majú vyšší fitness) a sú preto dynamicky preferované. Zastúpenie lepšie adaptovaného variantu v čase v dostatočne veľkej populácii spĺňa logistickú dynamiku, ktorej parametrom je práve miera preferencie (fitness) sledovaného variantu. Dôležitou praktickou otázkou je odhadnúť fitness jednotlivých variantov, zastúpených v populácii z časových dát, ktoré prirodzene obsahujú rôzne formy náhodnosti. Aj keď existuje viacero prístupov na inferenciu fitness parametrov, táto praktická matematická otázka má fundamentálny problém. Tento problém súvisí s časovým posunom v identifikácii genetického variantu od momentu infikovania vírusom. Hlavné otázky práce sú: Ako časový posun dátových bodov vplýva na odhadnutú hodnotu fitness? Ako predošlá odpoveď závisí od náhodného rozdelenia posunov? Práca zahŕňa teoretickú časť, prácu s biologickými dátami, aj numerické výpočty.

Vedúci: Mgr. Katarína Boďová, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 09.01.2025

Dátum schválenia: 15.01.2025 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné vyhlásenie

Čestne vyhlasujem, že celú diplomovú prácu na tému „Inferencia logistického rastu v dátach s náhodným časovým posunom“, vrátane všetkých jej príloh a obrázkov, som vypracoval samostatne, a to s použitím literatúry uvedenej v priloženom zozname a nástrojov umelej inteligencie. Vyhlasujem, že nástroje umelej inteligencie som použil v súlade s príslušnými právnymi predpismi, akademickými právami a slobodami, etickými a morálnymi zásadami za súčasného dodržania akademickej integrity a že ich použitie je v práci vhodným spôsobom označené.

Bratislava, 7.5.2026

.....

Bc. Robert Valaška

Poďakovanie Touto cestou by som sa rád poďakoval svojej školiteľke Mgr. Kataríne Boďovej za jej pomoc, ochotu na konzultáciách, trpezlivosť a užitočné rady, ktoré mi veľmi pomohli pri písaní tejto práce. Taktiež ďakujem svojej rodine za morálnu podporu ku koncu môjho štúdia.

Abstrakt v štátnom jazyku

VALAŠKA, Robert: Inferencia logistického rastu v dátach s náhodným časovým posunom [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Katarína Boďová, PhD., Bratislava, 92 s.

V tejto práci sa zaoberáme inferenciou relatívnej fitness invadujúceho variantu vírusu. Pandémia COVID-19 v roku 2020 nastolila signifikantné výzvy po celom svete, medzi nimi aj potrebu epidemiologického výskumu rôznych variantov vírusu SARS-CoV-2 a ich vlastností. Budeme sa zameriavať hlavne na kvantifikáciu genetickej výhody invadujúceho variantu oproti udomácnenému variantu, ktorá sa nazýva už spomenutých pojmom fitness. Vychádzame zo základného exponenciálneho modelu, ktorý je najjednoduchším spôsobom modelovania šírenia vírusu a analyticky vypočítame fitness. Podiel invadujúceho variantu oproti udomácnenému variantu v čase sa dá aproximovať logistickou funkciou. Následne do modelu zapracujeme náhodné oneskorenie, zachytávajúce časový rozdiel medzi časom infikovania a časom zistenia prítomnosti vírusu. Analyticky zistíme, do akej miery táto náhodnosť ovplyvňuje odhad fitness, a výsledky podložíme aj numerickými experimentami a simuláciami. V ďalšej časti práce sa budeme venovať SIR modelu, ktorý lepšie odráža realistické podmienky. Budeme skúmať, či proporcia invadujúceho variantu v SIR modeli je stále konzistentná s logistickou funkciou a ako zapracované oneskorenie vplýva na presnosť odhadu fitness. Výsledky naznačujú, že jednoduchší exponenciálny model v simuláciách nevykazuje výrazné odchýlky odhadu fitness ani po aplikovaní náhodne rozdelených oneskorení. Realistickejší SIR model modelovaný pomocou diferenciálnych rovníc zväčša zachováva proporciu invadujúceho variantu konzistentnú s logistickou funkciou, ale existujú režimy, kde toto neplatí, napr. pri neúplnej imunite po prekonaní infekcie. Odhad fitness pomocou logistickej funkcie môže potom viesť k chybe na úrovni až 200%, pri režimoch konzistentných s logistickou funkciou približne do 80% v závislosti od oneskorenia. Taktiež existujú režimy, kde náhodnosť v oneskoreniach nehrá v SIR modeli zásadnú rolu, na rozdiel od strednej hodnoty oneskorenia.

Kľúčové slová: SARS-CoV-2, logistická funkcia, inferencia, fitness, časové oneskorenie, exponenciálny model, SIR model

Abstract

VALAŠKA, Robert: Inference of logistic growth in randomly lagged data [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Katarína Boďová, PhD., Bratislava, 92 p.

In this thesis we investigate the inference of relative fitness of invading virus variant. The COVID-19 pandemic in 2020 presented significant challenges around the world, including the need of epidemiological research of various SARS-CoV-2 variants and their traits. We will focus on quantifying the genetic advantage of an invading variant against an established variant, also called fitness. We will begin with the basic exponential model, which is the simplest way of modeling the virus' spread and allows us to analytically compute fitness. The proportion of the invading variant in time can be approximated by a logistic function. Then we will incorporate random delays into the model, representing a time delay between the time of infection and time of virus discovery. We will perform analytical calculation to help us understand how the random delays affect the fitness inference and we will support our findings with numerical experiments and simulations. The next part of the thesis will be dedicated to the SIR model, which conveys a more realistic approach. Our main goal is to find out if the proportion of invading variant is still consistent with the logistic function and how the applied delays affect fitness inference. The results suggest that the simpler exponential model does not display signs of significant fitness inference bias even after applying randomly distributed delays. The more realistic SIR model made up of differential equations mostly keeps the proportion of invading variant consistent with the logistic function, but there are certain regimes where this is not true, for example assuming an incomplete immunity after recovery. Inference of fitness using a logistic function in this scenario can yield errors of up to 200%, in regimes consistent with the logistic function, we can observe errors of up to 80% based on applied delays. There are also regimes in which the randomness in the delays does not play a big role in SIR model, contrary to the expected value of delay.

Keywords: SARS-CoV-2, logistic function, inference, fitness, reporting delay, exponential model, SIR model

Obsah

Úvod **13**

1 Epidemiológia šírenia vírusov **15**

1.1 Evolúcia a selekcia 15

1.2 Modelovanie šírenia vírusov 16

1.3 Epidemiológia SARS-CoV-2 16

2 Dátová analýza **19**

2.1 Základné charakteristiky 19

2.2 Experimentálne zistenia - logistická krivka 22

3 Exponenciálny model **29**

3.1 Deterministický model 29

3.1.1 Výpočet relatívnej fitness 30

3.1.2 Deterministický model s oneskorením 32

3.1.3 Simulácia v Pythone 33

3.2 Model s normálne rozdeleným náhodným oneskorením 35

3.3 Model s log-normálne rozdeleným náhodným oneskorením 46

4 SIR model **51**

4.1 Základný model 51

4.1.1 Analýza vstupných parametrov 54

4.1.2 Aplikácia oneskorenia 58

4.2 Model s dočasnou imunitou 68

4.3 Model bez krížovej imunity 72

4.4 Kombinovaný SIR model 78

Záver **83**

Zoznam použitej literatúry **87**

Príloha A - Skripty **91**

Úvod

Evolučná biológia sa v priebehu storočí vyvinula na výrazne relevantné odvetvie vedy. Po začiatku svetovej pandémie respiračného ochorenia COVID-19 sa vedci sústredili na porozumenie virálnej dynamiky vírusu SARS-CoV-2 a jeho variantov v snahe vyvinúť efektívne prostriedky proti jeho šíreniu. Globálne zdieľanie sekvenovaných dát cez platformy ako databáza GISAID sa stali veľmi dôležitými nástrojmi v boji proti šíreniu vírusu [22].

Vírus spôsobujúci chorobu COVID-19, SARS-CoV-2, prešiel za roky jeho sledovania rôznymi mutáciami, čím vznikli rôzne varianty vírusu s inými charakteristikami ovplyvňujúcimi infekčnosť a úmrtnosť. Najznámejšie a najviac rozšírené varianty niesli mená Alpha, Delta a Omicron. Vlny infekcií, ktoré boli pozorovateľné pri vzostupe každého z variantov, boli podrobne pozorované a dáta z nich uložené v databáze GISAID. Dáta s najlepšou kvalitou v rámci Európy pochádzali zo Spojeného kráľovstva, ktoré bolo ťažko zasiahnuté rýchlym vzostupom variantu Alpha na konci roka 2020 [4]. Neskôr v priebehu rokov 2021 a 2022 sa v databáze objavili aj podrobné informácie o novších variantoch Delta a Omicron, ktoré tiež predstavovali zvýšené riziko kvôli ich infekčnosti [10].

Aj najlepšie dáta však podliehajú istým nespresnostiam, spôsobeným reportovaním dát. Presnosť dát závisí od viacerých faktorov, napríklad od testovacej stratégie, rôzneho prístupu k testom, zásad reportovania prípadov, lokálnych ohnísk a podobne [3]. Jednou z neželaných nepresností v dátach je časové oneskorenie, ktoré je vyjadrené časovým rozdielom medzi skutočným časom infekcie a časom odhalenia infekcie. V reálnych dátach poznáme len druhý spomínaný čas. Toto oneskorenie informácií následne môže spôsobiť to, že niektoré varianty považujeme za menšiu alebo väčšiu hrozbu ako v skutočnosti sú. Súčasná literatúra objasňuje viacero štatistických metód používaných na zistenie reálnej incidencie vírusov, ale nie priamo na kvantifikáciu relatívneho rastu medzi viacerými variantmi cirkulujúcimi v jednej populácii, pričom rôzne varianty môžu mať rôzne rozdelenie oneskorenia [2]. Výsledky z článku [25] taktiež ukazujú, že jednotlivé varianty SARS-CoV-2 majú odlišnú inkubačnú dobu, čo sa v realite pretaví do rôzne dlhých oneskorení.

V tejto práci sa budeme venovať dynamike šírenia ochorenia COVID-19. Najprv stručne predstavíme šírenie vírusov z epidemiologického hľadiska. Následne ako motiváciu na hlavnú časť práce využijeme dáta z databázy GISAID, ktoré obsahujú veľmi podrobné informácie o každej sekvenovej vzorke vírusu. Budeme analyzovať dáta zo Spojeného kráľovstva za roky 2020 - 2024, pričom budeme klásť špecifický dôraz na interakciu dvoch variantov v populácii. Genetická výhoda invadujúceho variantu oproti udomácnenému variantu sa nazýva relatívna fitness, ktorú kvantifikujeme zo stiahnutých dát. Na základe dátovej analýzy prejdeme k modelom, pričom začneme jednoduchým exponenciálnym modelom. Analyticky vypočítame relatívnu fitness a zahrnieme náhodné oneskorenie s cieľom odhaliť režimy, kedy odhad fitness je najďalej od hodnoty fitness bez uvažovania oneskorenia. Následne naše zistenia overíme numerickými simuláciami vykonanými v Pythone. Potom prejdeme k SIR modelu, ktorý lepšie reprezentuje realitu ako exponenciálny model [24]. Bude nás zaujímať, za akých nastavení parametrov tento model poskytuje dobrú kvalitu výsledkov a zistíme, ako aplikované oneskorenie vplýva na odhad relatívnej fitness oproti exponenciálnemu modelu. Nakoniec sa pozrieme na viaceré modifikácie SIR modelu, napr. na model s neúplnou krížovou imunitou, model s dočasnou imunitou, ale aj ich kombinácie so zahrnutím asymetrickej krížovej imunity. Na záver zhrnieme výsledky zo všetkých modelov a porovnáme skúmané prístupy k modelovaniu šírenia vírusov.

1 Epidemiológia šírenia vírusov

V tejto kapitole sa oboznámime so šírením vírusov v populácii a ich populačnou dynamikou. Pozrieme sa na základné evolučné procesy vplývajúce na šírenie vírusu a na možnosti modelovania tohto fenoménu. Následne sa pozrieme na to, ako môžu na zozbierané dáta o ich šírení teoreticky vplývať ich nedokonalosti, pričom sa zameriame na nedávnu pandémiu vírusu SARS-CoV-2.

1.1 Evolúcia a selekcia

V roku 1859 vydal Charles Darwin prvé vydanie knihy Pôvod druhov (z angl. *The Origin of Species*), ktorá sa dnes považuje za základ evolučnej biológie. Darwin v tejto knihe prvýkrát definoval pojem prirodzenej selekcie (z angl. *natural selection*). Moderná formulácia sa dá kompaktne popísať v nasledovných troch bodoch [13]:

- Narodí sa viac jedincov, ako dokáže prežiť alebo sa rozmnožovať;
- Organizmy sa odlišujú vo svojej schopnosti prežiť a rozmnožovať sa - čiastočne kvôli rozdielom v genotype jedincov;
- Genotypy, ktoré vedú k prežitiu v súčasnom prostredí, sú početne v presile a preto prispievajú vo väčšej miere k potomkom v ďalšej generácii.

Procesom prirodzenej selekcie sa medzigeneračne zvyšuje prítomnosť genotypov, ktoré majú viac vyvinuté možnosti prežitia a reprodukcie. Postupne sa z populácie vytratia horšie adaptované genotypy a populácia sa teda progresívne stáva odolnejšou a dokáže lepšie prežiť v meniacom sa prostredí. Tento proces sa nazýva aj adaptácia.

Darwin určil jednotku selekcie ako individuálny organizmus, pretože schopnosť prežitia jednotlivca (resp. genotypu) je to, na čom záleží. Darwin tento fenomén (z angl. *„survival of the fittest“*) prirovnáva k prirodzenej selekcii. Ako sme už vyššie spomenuli, časom sa frekvencia jednotlivých genotypov pod vplyvom naturálnej selekcie mení, pričom prevažujú lepšie adaptované genotypy.

1.2 Modelovanie šírenia vírusov

Populačná genetika sa musí vysporiadať s faktormi ako veľkosť populácie, geografická distribúcia jedincov, mutácie a pod. Veľa z týchto faktorov je tak komplexných, že sa nedajú vyjadriť vo forme jednoduchých modelov a preto typicky pracujeme so zjednodušenými situáciami, v ktorých sa zaujímame len o niekoľko faktorov a ostatné zanedbáme. Účelové zjednodušenie komplexnej situácie zdroj [13] vysvetľuje pod pojmom model.

Matematické modely sú užitočnými prostriedkami na zjednodušenie, pochopenie a analýzu epidemiologických procesov. Konštrukcia modelu vyžaduje rovnováhu medzi realistickosťou a jednoduchosťou. Najjednoduchší model, ktorý sa okrem modelovania rastu populácie dá využiť aj na modelovanie šírenia vírusov, je exponenciálny model. Tento model predpokladá, že konkrétny genotyp sa v každom sledovanom časovom bode multiplikuje nejakým fixným podielom, teda podľa istého rastového koeficientu. Existuje aj spojitý exponenciálny model, ktorý sa od diskrétneho odlišuje tým, že vývoj šírenia vírusu sledujeme konštantne, resp. v infinitezimálne malých časových krokoch.

Okrem toho, že dnes máme k dispozícii veľa rôznych modelov, existujú aj rôzne úrovne detailov pri modelovaní šírenia vírusov. V práci sa najprv budeme venovať jednoduchým modelom a následne prejdeme na komplexnejšie a realistickejšie modely.

1.3 Epidemiológia SARS-CoV-2

Za posledné roky sa modelovanie šírenia vírusov dostalo do globálneho povedomia, a to kvôli celosvetovej pandémii vírusu SARS-CoV-2. Pandémia COVID-19 bola najväčšou globálnou epidémiou za posledné storočia. Okrem toho, že upozornila na nedostatky v zdravotníckych systémoch, poskytla aj veľa dát a vedci dokázali získať podrobný prehľad o šírení a typoch vírusov. Spomeňme si, že šírenie vírusu neprichádzalo v jednej veľkej vlne, ale vo viacerých vlnách, reprezentovaných jednotlivými variantmi vírusu, napr. Alpha, Delta, atď. Každá väčšia vlna bola spôsobená iným variantom vírusu, ktorý vďaka istej výhode oproti ďalším variantom dokázal v populácii prežiť najlepšie a preto sa najviac rozšíril. Túto relatívnu výhodu oproti iným variantom nazývame fitness, ktorá sa dá aj kvantifikovať, čo bude predmetom nasledujúcej kapitoly.

Po nájdení prvého variantu, ktorý sa začal šíriť v čínskom Wuchane v decembri 2019, sa počas rokov mutáciou vyvinuli nové varianty, ktoré boli v niektorých aspektoch lepšie ako ich predchodcovia. Primárne išlo o vyššiu infekčnosť, ktorá dovolila novému variantu rýchlejšie sa rozšíriť a efektívne z populácie vytlačiť starý variant. Prvá veľká svetová vlna bola spôsobená variantom Alpha, ktorý bol objavený v Spojenom kráľovstve na konci roku 2020. Následne bol nahradený variantom Delta, ktorý drží prvenstvo v počte zaznamenaných prípadov. Posledná veľká vlna bola spôsobená variantom Omicron, z ktorého neskôr vzniklo viacero nových variantov, ktoré už ale nemali globálny dopad.

Modelovanie šírenia vírusov, resp. ich variantov nie je vždy priamočiare. Ak chceme z modelov získať čo najpresnejšie výsledky, musíme mať k dispozícii čo najpresnejšie dáta. Aj najlepšie dáta ale obsahujú istú mieru nepresnosti, a to vo forme oneskorenia (z angl. *reporting delay*). Je vyjadrená ako rozdiel medzi časom, kedy bol človek infikovaný a časom, kedy sa infekcia odhalila. Táto odchýlka merania je dôležitá, ak istý čas po infekcii jedinec neprejavuje symptómy, ale už je schopný šíriť infekciu ďalej. Ďalšie nepresnosti v získaných dátach môžu zahrňovať rozdiely v testovacích kapacitách, vybavenosti laboratórií, spôsoboch reportovania prípadov a pod. V práci sa budeme venovať práve oneskoreniam, lebo na rozdiel od iných vyššie spomenutých faktorov nie je až tak preskúmaný ich dopad. Náš cieľ je vytvoriť obraz o tom, ako vplýva oneskorenie na vnímanie istého variantu vírusu relatívne k inému variantu vírusu.

2 Dátová analýza

V tejto kapitole predstavíme dáta, s ktorými budeme ďalej pracovať. Dáta budú slúžiť aj ako motivácia pre ďalšie kapitoly práce. Pozrieme sa na ich základné charakteristiky a vykreslíme ich priebeh. Následne vykonáme dátovú analýzu a výsledky spracujeme do prehľadných grafov, ktoré pomôžu získať jasný a podrobný prehľad o dátach.

2.1 Základné charakteristiky

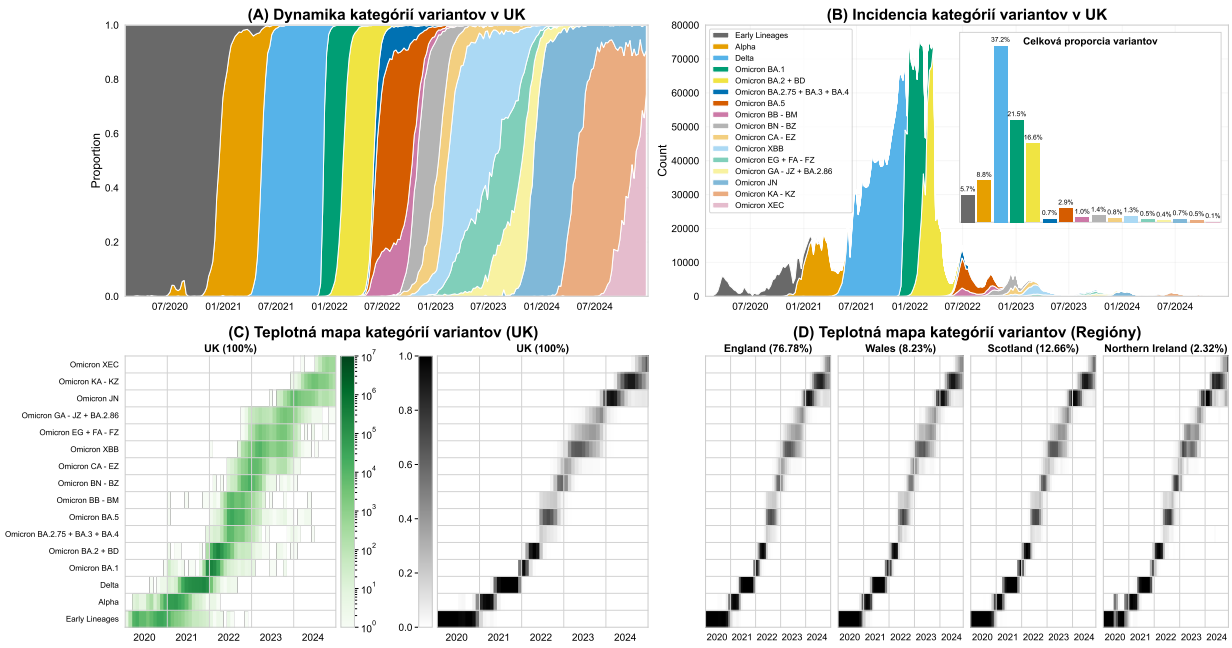
Dáta sme si stiahli z databázy GISAID [11]. GISAID obsahuje veľmi podrobné dáta o rôznych sledovaných vírusoch a ich variantoch, napr. SARS-CoV-2, influenza, mpox a pod. Pre účely našej práce sme si vybrali dáta pre vírus SARS-CoV-2 a budeme sledovať jeho dynamiku v Spojenom kráľovstve v období od januára 2020 do decembra 2024.

Dáta z databázy GISAID sú dostupné vo forme súborov formátu `.tsv`, pričom každý riadok v súbore predstavuje jeden prípad. Keďže každá vzorka bola samostatne sekvenovaná, vieme sa dozvedieť veľa podrobných informácií, od proteínovej sekvencie až po presný typ variantu. Pre účely dátovej analýzy nás budú zaujímať nasledujúce štyri charakteristiky dát:

- Accession ID: unikátny identifikátor merania;
- Collection date: dátum odobratia analyzovanej vzorky;
- Location: miesto odobratia analyzovanej vzorky;
- Lineage: presný variant vírusu.

Hlavný cieľ dátovej analýzy je priblížiť celkový priebeh ochorenia COVID-19 v celom Spojenom kráľovstve aj v každom z jeho štyroch regiónov. Začali sme skúmaním variantov vírusu SARS-CoV-2 a ich vývoja. Premennú Lineage sme zovšeobecnili na hlavnú subkategóriu (napr. Omicron EG.1, EG.2 a podobne boli označené ako EG). Taktiež sme prípadné podrobnejšie informácie z premennej Location zovšeobecnili na uvažované štyri regióny (Anglicko, Wales, Škótsko, Severné Írsko).

Z predošlej kapitoly vieme, že jednotlivé varianty už udomácnené v populácii časom môžu nahradiť nové varianty, ktoré majú nejakú výhodu oproti tým starým. Dátovú analýzu sme vykonali v Pythone pomocou knižníc `numpy`, `pandas` a `matplotlib` [5]. Výsledky analýzy s vizuálnou interpretáciou sú dostupné na Obr. 1 a v texte nižšie.



Obr. 1: Virálna dynamika vírusu SARS-CoV-2 a distribúcia kategórií variantov po regiónoch Spojeného kráľovstva. (A) Proporcia individuálnych kategórií variantov od januára 2020 do decembra 2024, týždenné dáta. **(B)** Incidencia individuálnych kategórií variantov spolu s celkovou proporciou prípadov pre každú kategóriu variantov, týždenné dáta. Vložený graf zobrazuje celkovú proporciu prípadov pre každú kategóriu variantov. Legenda je zdieľaná s (A). **(C)** Teplotná mapa proporcie a incidencie kategórií variantov pre Spojené kráľovstvo, mesačné dáta. **(D)** Teplotná mapa proporcie kategórií variantov pre Anglicko, Wales, Škótsko a Severné Írsko, mesačné dáta.

Na Obr. 1 môžeme vidieť celkovú dynamiku šírenia vírusu SARS-CoV-2 a incidenciu jeho variantov. Pre účely analýzy sme menej rozšírené varianty (primárne varianty objavené po roku 2022) spojili do kategórií variantov (napr. kategória Omicron KA - KZ obsahuje varianty KA, KB, KC, ..., KY, KZ). Kategórie boli vybrané tak, aby prechody medzi jednotlivými kategóriami variantov čo najviac pripomínali logistickú krivku, ale zároveň bola dodržaná chronológia nástupu variantov.

Graf (A) predstavuje proporciu každej kategórie variantov v čase. Všimnime si rozhranie medzi dvoma variantmi idúcimi po sebe, ktoré nápadne pripomína logistickú krivku. Tento fenomén bol skúmaný a podložený napríklad zdrojom [23]. Pri skorších variantoch (napr. Alpha, Delta, Omicron BA.1) máme vizuálne hladšie krivky, pretože v tom čase existovalo len málo variantov, ktoré ale mali vysokú incidenciu. Preto máme za toto obdobie aj vďaka vyšším testovacím kapacitám dostatok dát.

Na grafe (B) vidíme incidenciu jednotlivých kategórií variantov. Najviac prípadov sa vyskytlo práve u skorších variantov, ako napr. Alpha, Delta a skoré varianty Omicronu. V absolútnom počte prípadov vedie variant Delta, s 37.2% všetkých prípadov. Časom počty prípadov klesli kvôli vyššej imunite a zaočkovanosti. Preto okrem iného boli nové varianty viac obmedzené na lokálne epidémie a nešírili sa globálne ako skoršie varianty.

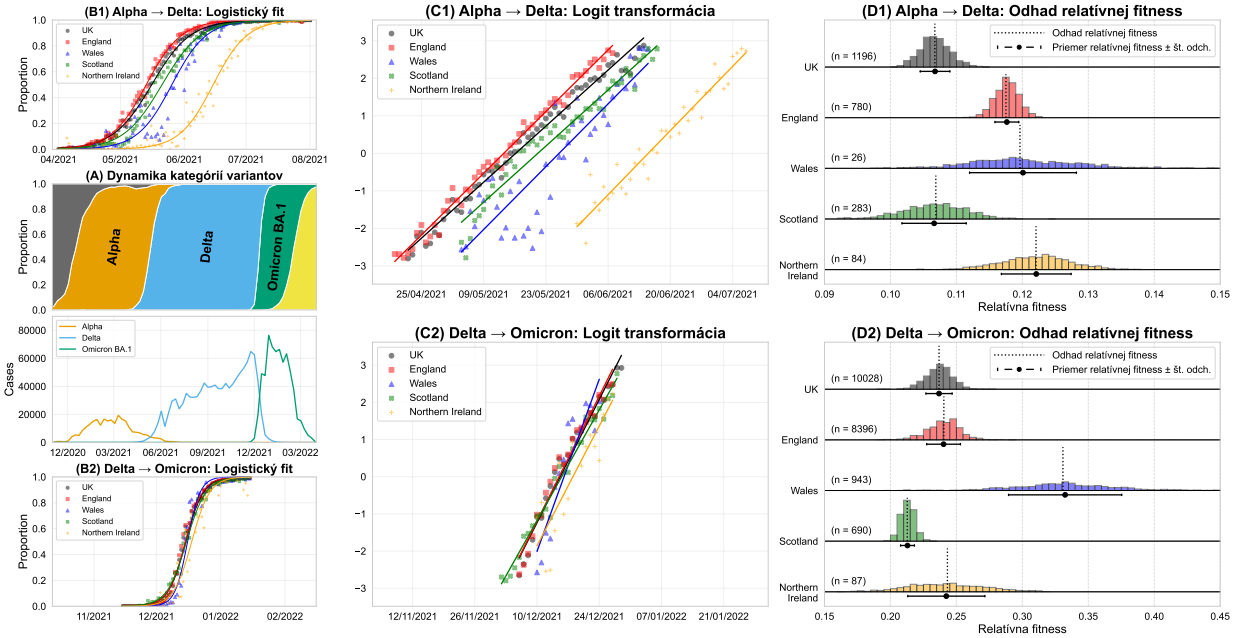
Graf (C) ponúka dva pohľady na šírenie SARS-CoV-2 v Spojenom kráľovstve - v absolútnych a relatívnych číslach. Na obrázku vľavo môžeme vidieť počet prípadov jednotlivých kategórií variantov za každý mesiac v logaritmickej škále a na obrázku vľavo podiel kategórií variantov za každý mesiac. Za zmienku stojí to, že aj rozdelenie variantov v čase na grafoch (C) a (D) má vizuálne logistický priebeh.

Nakoniec na grafe (D) pozorujeme podiel kategórií jednotlivých variantov podobne ako v grafe (C) vpravo, ale samostatne pre Anglicko, Wales, Škótsko a Severné Írsko. Vieme konštatovať, že priebeh ochorenia je pre každý región konzistentný s celou krajinou, teda vývoj SARS-CoV-2 vyzerá veľmi podobne pre každý región.

Keďže naše dáta obsahujú incidencie viacerých variantov idúcich po sebe, ako je možné vidieť na grafe (A), dáva zmysel vyčísliť relatívnu výhodu invadujúceho variantu oproti udomácnenému (predošlému) variantu. Táto výhoda sa nazýva fitness, pričom v relatívnej škále ju budeme volať relatívna fitness [23]. V nasledujúcej časti budeme pre rôzne vlny SARS-CoV-2 počítať relatívnu fitness a overovať jej robustnosť vzhľadom na charakter vlny.

2.2 Experimentálne zistenia - logistická krivka

V tejto časti sme sa bližšie pozreli na interakcie medzi variantmi. Vybrali sme dve najdôležitejšie vlny, a to prechod z variantu Alpha na Delta a prechod z variantu Delta na prvý Omicron variant (BA.1). Tieto vlny majú podľa grafu (A) z Obr. 1 logistický priebeh a zároveň obsahujú dostatočne veľa prípadov na to, aby boli naše výsledky robustné. Náš cieľ je kvantifikovať schopnosť invadujúceho variantu vytlačiť udomácnený variant z populácie a vykonať inferenciu (z angl. *inference*) relatívnej fitness z dostupných dát. Výsledky analýzy sú viditeľné na Obr. 2.



Obr. 2: Inferencia relatívnej fitness pre vlny Alpha → Delta a Delta → Omicron BA.1. **(A)** Prehľad virálnej dynamiky pre varianty Alpha, Delta a Omicron BA.1; vrchný obrázok analogicky ku grafu (A) z Obr. 1. Omicron BA.1 budeme pre účely tohto obrázku označovať iba ako Omicron. **(B1)** Proporcia variantu Delta pre Spojené kráľovstvo a jeho 4 regióny pre vlnu Alpha → Delta. Cez dátové body je nakreslená logistická krivka. **(B2)** Analogicky podľa (B1) pre vlnu Delta → Omicron. **(C1)** Logit transformácia vlny Alpha → Delta pre Spojené kráľovstvo a jeho 4 regióny. **(C2)** Analogicky podľa (C1) pre vlnu Delta → Omicron. **(D1)** Inferencia relatívnej fitness z (C1) a jej rozdelenie získané použitím neparametrického bootstrapu pre vlnu Alpha → Delta. Číslo „n“ reprezentuje priemerný počet denných prípadov okolo stredu logistickej krivky pre každý región. **(D2)** Analogicky podľa (D1) pre vlnu Delta → Omicron.

Pri analýze na Obr. 2 sme využili výsledky z analýzy na Obr. 1. Graf (A) nám približuje

časové okno, ktoré bude pri tejto analýze relevantné, spolu s vykreslením týždenných prípadov variantov Alpha, Delta a Omicron BA.1. Všimnime si incidenciu variantu Delta oproti variantu Alpha. Variant Delta začal získavať prevalenciu až vtedy, keď variant Alpha bol na ústupe. Naopak, variant Omicron BA.1 sa objavil v čase, kedy variant Delta bol na vrchole a vďaka väčšej genetickej výhode ho v Spojenom kráľovstve nahradil [19].

Následne sme pokračovali analyzovaním jednotlivých regiónov aj celého Spojeného kráľovstva na grafe (B1) zodpovedajúceho vlne Alpha → Delta a grafe (B2) zodpovedajúceho vlne Delta → Omicron BA.1. Môžeme vidieť, že všetky regióny majú proporciu invadujúceho variantu konzistentnú s krivkou logistickej funkcie. Taktiež môžeme pozorovať, že variant Delta sa v Severnom Írsku rozšíril neskôr ako v ostatných regiónoch. Pri variante Omicron BA.1 vidíme veľmi podobné dynamiky šírenia, ale mierne rýchlejší rast prevalencie vo Walesi.

Na získané logistické krivky pre každý región sme aplikovali logit (log-odds) transformáciu. Je to inverzná funkcia k logistickej funkcii, preto za predpokladu perfektnej logistickej krivky by sme mali dostať perfektnú lineárnu krivku. Na grafoch (C1) a (C2) vieme pozorovať transformované body. Na body sme následne aplikovali lineárnu regresiu. Pri lineárnej regresii uvažujeme dva argumenty krivky, všeobecne nazývané ako β_0 (intercept) a β_1 (smernica). Smernica predstavuje relatívnu fitness invadujúceho variantu, teda strmšia lineárne krivka predstavuje väčšiu výhodu invadujúceho variantu oproti udomácnenému. Na grafe (C1) voľným okom vidíme približne podobnú fitness pre každý región, na grafe (C2) má Wales trochu väčší sklon ako ostatné regióny. Pre kompletnosť treba uviesť, že do logit transformácie sme zahrnuli iba tie body, kde mal invadujúci variant podiel väčší ako 5% a menší ako 95%, teda sme vylúčili body na oboch okrajoch logistických kriviek z grafov (B1) a (B2) pre presnejší odhad fitness.

Nakoniec sme otestovali robustnosť našich dát a vykonali odhad relatívnej fitness na grafoch (D1) a (D2). Z grafov (C1) a (C2) sme vykonali inferenciu hodnoty relatívnej fitness invadujúceho variantu pre obe vlny a všetky regióny, čo môžeme vidieť v Tabuľke 1.

Vlna	UK	Anglicko	Wales	Škótsko	Severné Írsko
Alpha → Delta	0.106728	0.117508	0.119639	0.106881	0.122074
Delta → Omicron BA.1	0.236859	0.240492	0.330763	0.212754	0.243204

Tabuľka 1: Inferencia relatívnej fitness pre vlny Alpha → Delta a Delta → Omicron BA.1.

Hodnotu relatívnej fitness vieme vyjadriť ako rozdiel rastových koeficientov v spojitom exponenciálnom modeli (ukážeme v kapitole 3). Pozorujeme, že pre druhú vlnu vyšla relatívna fitness všeobecne vyššia ako pre prvú vlnu, čo indikuje väčšiu výhodu variantu Omicron BA.1 oproti variantu Delta ako variantu Delta oproti variantu Alpha. Pre porovnanie hodnôt s dostupnou literatúrou potrebujeme definovať reprodukčné číslo a generačný čas. Reprodukčné číslo je priemerný počet jedincov, ktorých jeden infekčný jedinec infikuje počas trvania choroby. [2]. Generačný čas alebo generačný interval je čas, ktorý ubehne medzi infekciou určitého jedinca a jeho následným infikovaním iného jedinca. Ak uvažujeme mediánový generačný čas T_g o dĺžke 4.7 dní [12], vieme spočítať relatívnu prenosnosť vírusu (z angl. *transmissibility*) [4] ako

$$\frac{R_{Delta}}{R_{Alpha}} = e^{sT_g} = 1.6514\,, \tag{1}$$

kde R_{Alpha} a R_{Delta} je reprodukčné číslo variantov Alpha a Delta a s je rozdiel rastových koeficientov. Dosadením nášho výsledku (0.106728) za s sme dostali výsledok rovný približne 1.65, čo znamená o 65% vyššiu prenosnosť variantu Delta oproti variantu Alpha. Toto je konzistentné s hodnotami spomínanými v literatúre [12] [20] [8]. Výpočet zopakujeme pre vlnu Delta → Omicron BA.1., uvažujúc $T_g = 3$:

$$\frac{R_{Delta}}{R_{Alpha}} = e^{sT_g} = 2.0352\,. \tag{2}$$

Môžeme vidieť, že Omicron BA.1 má až o približne 103% vyššiu prenosnosť. Toto číslo je tiež veľmi blízko k hodnotám v literatúre, podľa ktorej má Omicron BA.1 približne dvojnásobnú prenosnosť [17].

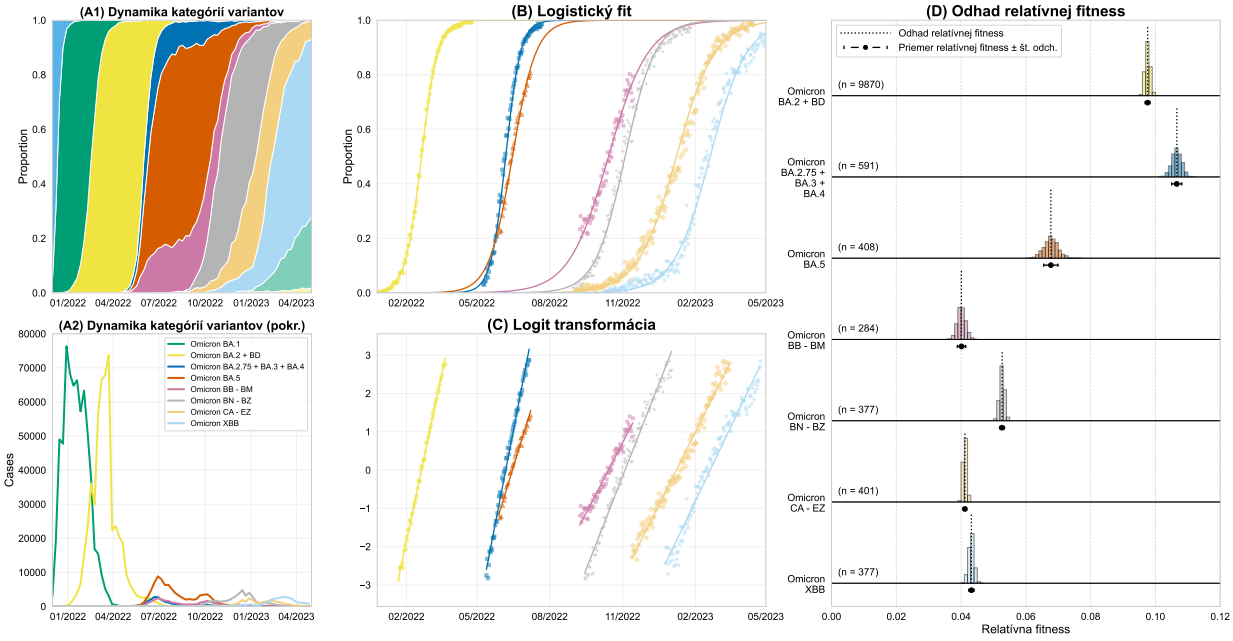
Vráťme sa k Obr. 2. Po inferencii hodnoty relatívnej fitness sme sa rozhodli otestovať jej robustnosť. Vyrobili sme hypotetické rozdelenie odhadu relatívnej fitness pomocou neparametrického bootstrapu [1]. Tento proces spočíva v tom, že si z jednej dátovej sady

(reprezentovanej bodmi z grafov (C1) a (C2)) vygenerujeme 1000 ďalších sád, do ktorých náhodne vyberieme rovnaký počet dátových bodov ako v originálnej dátovej sade, ale vyberáme s opakovaním. To znamená, že niektoré dátové body sa môžu v novej dátovej sade vyskytovať viackrát a niektoré ani raz. Následne po aplikácii lineárnej regresie a zistení smernice lineárnej krivky sme zostavili rozdelenie inferovanej relatívnej fitness a nakreslili histogramy do grafov (D1) a (D2). Tiež sme z rozdelenia vypočítali aj jeho strednú hodnotu (priemer) relatívnej fitness z vytvorených dátových sád a štandardnú odchýlku tohto odhadu a zaznačili ich do grafov. Môžeme vidieť, že odhady pre Spojené kráľovstvo a Anglicko majú malú štandardnú odchýlku vďaka dostatku presných dát z týchto regiónov. Wales, Škótsko a Severné Írsko majú kvôli menšiemu počtu dát väčšiu štandardnú odchýlku, ale stále nie je príliš veľká. Pozorujeme, že naše inferencie relatívnej fitness sú konzistentné nielen s priemernou hodnotou relatívnej fitness z nových dátových sád, ale aj s ostatnými regiónmi.

Všimnime si ešte číslo „n“ v ľavej časti grafov (D1) a (D2), ktoré označuje priemerný počet denných prípadov okolo stredu logistickej krivky. Pre väčšiu robustnosť sme nevybrali jeden dátový bod, ktorý bol najbližšie k stredu logistickej krivky (50% proporcia invadujúceho variantu), ale priemer denných prípadov približne zo strednej tretiny krivky (35% - 65% proporcia invadujúceho variantu). Pre prvú vlnu môžeme pozorovať výrazne menšie číslo pre Wales a Severné Írsko. Toto je spôsobené tým, že variant Alpha ustúpil skôr, ako sa naplno začal šíriť variant Delta (viditeľné v dolnej časti grafu (A)). Preto neprebiehala taká náhrada variantov ako napríklad v Škótsku, ale spoločne koexistovali oba varianty len veľmi krátku dobu a bolo zaregistrovaných spoločne málo prípadov. Pre druhú vlnu pozorujeme tento jav hlavne pre Severné Írsko. Za zmienku stojí aj veľký rozdiel medzi číslami „n“ v Spojenom kráľovstve v prvej a druhej vlne (1196 oproti 10028), pozorovateľné aj v Anglicku. Toto opäť pochádza z grafu (A), kde variant Omicron oveľa viac interaguje s variantom Delta ako variant Delta s variantom Alpha. Pri druhej vlne máme preto zaregistrovaných oveľa viac prípadov, pretože udomácnený variant ešte nebol na ústupe.

Na záver tejto časti sa vrátime na začiatok tejto kapitoly. Na Obr. 1 sme analyzovali všetky

kategórie variantov, z ktorých sme následne vybrali dve hlavné vlny a tie podrobne kvantifikovali na Obr. 2. Ostali ale ešte neskoršie vlny, ktoré by mohli byť predmetom ďalšej analýzy. Preto sme si vybrali niekoľko ďalších vĺn od Omicron BA.1 až po Omicron XBB a vykonali rovnakú analýzu ako predtým pre vlny Alpha → Delta a Delta → Omicron BA.1. Keďže sa jedná až o 7 vĺn, rozhodli sme sa skúmať dáta iba pre celé Spojené kráľovstvo. Neskoršie vlny navyše majú už pomerne malú incidenciu, takže by rozdeľovanie na regióny spôsobilo nestabilitu odhadu relatívnej fitness. Vykonaná analýza je dostupná na Obr. 3.



Obr. 3: Inferencia relatívnej fitness pre ďalšie Omicron vlny. (A1) Prehľad virálnej dynamiky pre varianty od Omicron BA.1 po Omicron XBB; analogicky ku grafu (A) z Obr. 1. **(A2)** Incidencia kategórií variantov, týždenné dáta. Legenda je zdieľaná s (A1), (B) a (C). **(B)** Proporcia kategórií variantov pre jednotlivé vlny. Na dáta je nafitovaná logistická krivka. Legenda popisuje krivky podľa invadujúceho variantu v danej vlne (napr. vlna Omicron BA.1 → Omicron BA.2 + BD je označená ako Omicron BA.2 + BD); toto označenie budeme používať aj na nasledujúcich grafoch. **(C)** Logit transformácia jednotlivých vĺn. **(D)** Inferencia relatívnej fitness z (C) a jej rozdelenie získané použitím neparametrického bootstrapu pre jednotlivé vlny. Číslo „n“ reprezentuje priemerný počet denných prípadov okolo stredu logistickej krivky pre každú vlnu.

Všimnime si na grafe (A2), že varianty Omicron BA.1 a BA.2 + BD majú veľmi vysokú incidenciu, pričom ostatné vlny už boli viac lokálne a nedosiahli globálne rozšírenie. Z

prvej vlny Omicron BA.1 → Omicron BA.2 + BD máme preto najviac dát, čo sa odráža aj na vysokej hodnote čísla „n“ na grafe (D). Keď sa pozrieme na graf (B), väčšina vĺn je vizuálne konzistentná s logistickou funkciou. Pre kategórie variantov Omicron BA.5 a Omicron BB - BM už ale nemáme perfektné dáta, preto aj napriek najlepšiemu možnému zostaveniu a zoradeniu kategórií variantov nebudú mať tieto dve vlny v základnej forme veľkú výpovednú hodnotu. Preto sme pre tieto dve vlny spravili výnimku a nepoužili okrajové časti kriviek. Zvolili sme taký rozsah, kde invadujúci variant má proporciu medzi 20% a 80%, čo nám dovolí presnejšie vypočítať sklon následnej lineárnej regresie. Na grafe (C) môžeme vidieť s-krivky po logit transformácii, pričom na väčšinu z nich sa dá už od začiatku pekne nafitovať lineárna krivka, okrem už spomínaných dvoch vĺn, pre ktoré sme spravili hore popísanú výnimku. Inferované hodnoty relatívnej fitness sme zapísali do Tabuľky 2.

Vlna	UK
Omicron BA.1 → Omicron BA.2 + BD	0.097591
Omicron BA.2 + BD → Omicron BA.2.75 + BA.3 + BA.4	0.106614
Omicron BA.2.75 + BA.3 + BA.4 → Omicron BA.5	0.067699
Omicron BA.5 → Omicron BB - BM	0.040085
Omicron BB - BM → Omicron BN - BZ	0.052684
Omicron BN - BZ → CA - EZ	0.041196
Omicron CA - EZ → Omicron XBB	0.043173

Tabuľka 2: Inferencia relatívnej fitness pre ďalšie Omicron vlny.

Graf (D) nám približuje robustnosť odhadu relatívnej fitness a rozdelenie jej odhadov pomocou neparametrického bootstrapu. Môžeme vidieť, že naše odhady nemajú veľkú štandardnú odchýlku, preto ich môžeme považovať za relatívne robustné. Výsledky pre vlny Omicron BA.5 a Omicron BB - BM treba potenciálne brať s rezervou, ale je namieste poznamenať, že sú konzistentné s odhadmi relatívnej fitness pre ostatné varianty a preto sú vierohodné. Taktiež konštatujeme, že pre každú vlnu máme dostatočný počet dát, keďže číslo „n“ je vždy väčšie ako 100.

Táto kapitola slúži ako motivácia pre nasledujúce kapitoly, pričom v prvej z nich sa budeme

venovať najjednoduchšiemu modelu, pomocou ktorého môžeme modelovať šírenie vírusu v populácii.

3 Exponenciálny model

V tejto kapitole sa pozrieme na jednoduchý exponenciálny model. Na základe dátovej analýzy v predchádzajúcej kapitole vieme konštatovať, že podiel variantu B oproti variantu A sa dá relatívne spoľahlivo modelovať logistickou krivkou. Ako ukážeme, tento výsledok je konzistentný s jednoduchým exponenciálnym modelom.

Najprv predstavíme deterministický model a jeho parametre. Následne otestujeme robustnosť modelu pridaním normálne a log-normálne rozdeleného náhodného oneskorenia, ktoré signalizuje časový rozdiel medzi infikovaním a odhalením infekcie. Nakoniec naše zistenia podložíme numerickými simuláciami v Pythone.

3.1 Deterministický model

Tento model sme si vybrali preto, lebo je to najjednoduchší zmysluplný model na modelovanie populačnej dynamiky [15]. Základný exponenciálny model predpokladá niekoľko vecí, ktoré môžu byť trochu nerealistické, ale ako sme ukázali v predošlej kapitole, výhodu invadujúceho variantu vieme modelovať pomocou logistickej funkcie [23]. Výhodou modelu je tiež možnosť explicitne vypočítať riešenie.

Nech $N(t)$ je veľkosť populácie v čase t a r_d je rastový koeficient populácie, ktorý vyjadruje zvýšenie podielu populácie za jednu časovú jednotku. Ak $r_d > 1$, hovoríme o zväčšovaní populácie. Ak $r_d < 1$, hovoríme o zmenšovaní populácie. Veľkosť populácie v čase t vieme modelovať ako súčet počtu jedincov v čase $t-1$ a prírastku jedincov za jednu časovú jednotku určenú parametrom r_d. Model vieme následne zapísať ako rovnicu (3).

$$N(t) = N(t-1) + r_d N(t-1) = (1+r_d)N(t-1) = (1+r_d)^t N(0)\,. \tag{3}$$

Tento model je diskrétny v čase. Za predpokladu, že N sledujeme v dostatočne jemných intervaloch, vieme použiť na modelovanie rastu populácie aj spojitý exponenciálny model. Pri spojitom modeli sa dajú ľahšie pozorovať trendy v dátach a sú lepšie interpretovateľné. Spojitý exponenciálny model definujeme ako rovnicu (4).

$$N(t) = N(0)e^{r_s t}\,. \tag{4}$$

Všimnime si, že spojitý model používa namiesto koeficientu r_d koeficient r_s. Spojitý model je konzistentný s diskrétnym modelom za predpokladu existencie vzťahu medzi r_d a r_s. Pre ukázanie vzťahu začneme porovnaním členov násobených $N(0)$ a následne vyjadríme r_s pomocou r_d a naopak.

$$\begin{aligned} (1 + r_d)^t &= e^{r_s t}\,, \\ r_s &= \ln(1 + r_d)\,, \\ r_d &= e^{r_s} - 1\,. \end{aligned} \tag{5}$$

Pre účely našej práce budeme ďalej pracovať so spojitým exponenciálnym modelom. Tento model vieme následne rozšíriť na dve rovnice, pričom každá reprezentuje jednu populáciu. Uvažujme samostatné populácie pre každý variant vírusu. Genetická výhoda jedného variantu voči druhému sa prejaví vo vyššej hodnote rastového koeficientu r_s, ktorý vyjadruje jeho infekčnosť. Definujeme dva varianty, A a B, pričom každý má svoj rastový koeficient r_A, resp. r_B. Model (4) následne vieme prepísať ako systém (6).

$$\begin{aligned} N_A(t) &= N_A(0)e^{r_A t}\,, \\ N_B(t) &= N_B(0)e^{r_B t}\,. \end{aligned} \tag{6}$$

Ak viaceré varianty koexistujú v nejakom časovom intervale, vieme na tomto intervale spočítať výhodu jedného variantu oproti druhému, nazývanú ako relatívna fitness. S týmto pojmom je úzko spätá štandardná logistická funkcia, ako sme ukázali v kapitole 2. V nasledujúcej časti vypočítame relatívnu fitness pre dva varianty vírusu A a B.

3.1.1 Výpočet relatívnej fitness

Nech variant A je udomácnený variant a variant B je invadujúci variant. Chceme analyticky vypočítať relatívnu fitness (výhoda variantu B oproti variantu A) pri uvažovaní dát bez oneskorenia. O dátach platí, že o každej infekcii vieme hneď, ako nastane. Nech $N_A(t)$, resp. $N_B(t)$ označujú počet infikovaných variantom A, resp. B v čase t. Nech r_A, resp. r_B vyjadrujú infekčnosť variantu A, resp. B. Nech $N_A(0)$, resp. $N_B(0)$ sú vopred určené konštanty vyjadrujúce počiatočný počet infikovaných variantom A, resp. B v čase 0.

Uvažujme model (6). Pre každé t budeme sledovať vývoj šírenia tak, že budeme počítať

podiel variantu B, ktorý označíme ako

$$X(t) = \frac{N_B(t)}{N_A(t) + N_B(t)}\,. \tag{7}$$

Vzťah (7) zodpovedá logistickej funkcii. Pomocou logit (log-odds) transformácie budeme vedieť logistickú funkciu transformovať na lineárnu funkciu a následne vypočítať smernicu ako deriváciu podľa t, čím dostaneme relatívnu fitness [18]. Najprv prepíšeme $X(t)$ do vhodného tvaru kompatibilného s logistickou funkciou, ktorý neobsahuje exponenciálne rastúce členy v čitateli aj menovateli. Pripomeňme si, že štandardná logistická funkcia má nasledovný predpis:

$$l(t) = \frac{1}{1 + e^{-t}}\,. \tag{8}$$

Výraz (7) predelíme členom $N_B(t)$ a následne upravíme:

$$X(t) = \frac{N_B(t)}{N_A(t) + N_B(t)} = \frac{1}{1 + \frac{N_A(t)}{N_B(t)}} = \frac{1}{1 + \exp\left(\ln\left(\frac{N_A(t)}{N_B(t)}\right)\right)} = l\left(-\ln\left(\frac{N_A(t)}{N_B(t)}\right)\right)\,. \tag{9}$$

Všimnime si, že upravený výraz vieme zapísať ako štandardnú logistickú funkciu s argumentom $-\ln\left(\frac{N_A(t)}{N_B(t)}\right)$. Otázka je, či tento argument je lineárny v t ako pôvodný argument t. Cez definíciu $N_A(t)$ a $N_B(t)$ vieme logaritmus prepísať do nasledovného tvaru:

$$\ln\left(\frac{N_A(t)}{N_B(t)}\right) = \ln\left(\frac{N_A(0)}{N_B(0)}\right) + \ln\left(\frac{e^{r_A t}}{e^{r_B t}}\right) = \ln\left(\frac{N_A(0)}{N_B(0)}\right) + (r_A - r_B)t\,. \tag{10}$$

Keďže $\ln\left(\frac{N_A(0)}{N_B(0)}\right)$ je konštanta, potvrdili sme, že výraz $-\ln\left(\frac{N_A(t)}{N_B(t)}\right)$ je lineárny v t a preto konzistentný s logistickou funkciou.

Pre porovnanie vypočítame relatívnu fitness pomocou aplikovania logit funkcie na výraz (7). Argument funkcie (8) vo výraze (10) dosadíme do (9) a upravíme na súčin:

$$X(t) = \frac{1}{1 + \exp\left(\ln\left(\frac{N_A(t)}{N_B(t)}\right)\right)} = \frac{1}{1 + \frac{N_A(0)}{N_B(0)} e^{(r_A - r_B)t}}\,. \tag{11}$$

Následne aplikujeme logit transformáciu v tvare napísanom v (12):

$$L(p) = \ln\left(\frac{p}{1-p}\right), p \in (0,1)\,. \tag{12}$$

Logit funkcia je inverzná k logistickej funkcii, preto aplikáciou logit funkcie na logistickú funkciu dostaneme pôvodnú funkciu, v našom prípade lineárnu. Predpoklady tejto transformácie spĺňame, lebo $X(t) \in (0,1)\,\forall t$. Aplikujeme na $X(t)$ a upravujeme na jednoduchší

tvar:

$$L(X(t)) = \ln\left(\frac{N_B(0)}{N_A(0)}e^{(r_B-r_A)t}\right) = \ln\left(\frac{N_B(0)}{N_A(0)}\right) + (r_B - r_A)t\,. \tag{13}$$

Opäť vidíme, že relatívna fitness závisí lineárne od člena t, ktorý ma koeficient $r_B - r_A$. Pre zistenie smernice krivky vypočítame deriváciu podľa t:

$$\frac{d}{dt}\left(\ln\left(\frac{N_B(0)}{N_A(0)}\right) + (r_B - r_A)t\right) = r_B - r_A\,. \tag{14}$$

Relatívna fitness nám vyšla rovná výrazu $r_B - r_A$, teda rozdielu medzi rastovými koeficientmi variantu B a variantu A. V našej práci pracujeme s predpokladom, že variant B ako invadujúci bude mať výhodu oproti variantu A, preto aj rozdiel $r_B - r_A$ bude kladný. V nasledujúcej časti sa pozrieme na prípad, ak nepoznáme skutočný čas infekcie a budeme mať tzv. oneskorené dáta.

3.1.2 Deterministický model s oneskorením

V časti 3.1.1 sme vypočítali relatívnu fitness ako $r_B - r_A$. Pozrieme sa na prípad, ak by sme pracovali s oneskorenými dátami, čo je v realite bežné. Náš model bude stále deterministický, len v ňom bude zahrnuté aj konštantné časové oneskorenie vyjadrené ako T_A pre variant A a T_B pre variant B. Každá infekcia variantom A bude teda posunutá dopredu v čase o T_A a každá infekcia variantom B o T_B. Cieľom bude zistiť, či sa hodnota relatívnej fitness bude odlišovať od deterministického modelu bez oneskorenia.

Do modelu (6) pridáme oneskorenia T_A, resp. T_B a prepíšeme ako model (15).

$$\begin{aligned} N_A(t) &= N_A(0)e^{r_A(t-T_A)}\,, \\ N_B(t) &= N_B(0)e^{r_B(t-T_B)}\,. \end{aligned} \tag{15}$$

Následne budeme pokračovať rovnakým spôsobom ako v predchádzajúcej časti bez časového oneskorenia. Najprv vypočítame podiel variantu B oproti variantu A ako

$$X(t) = \frac{N_B(t)}{N_A(t) + N_B(t)} = \frac{1}{1 + \exp\left(\ln\left(\frac{N_A(t)}{N_B(t)}\right)\right)} \tag{16}$$

a následne vyjadríme logaritmus v exponenciálnej funkcii ako

$$\begin{aligned} \ln\left(\frac{N_A(t)}{N_B(t)}\right) &= \ln\left(\frac{N_A(0)}{N_B(0)}\right) + \ln\left(\frac{e^{r_A(t-T_A)}}{e^{r_B(t-T_B)}}\right) = \\ &= \ln\left(\frac{N_A(0)}{N_B(0)}\right) + (r_A - r_B)t - (r_A T_A - r_B T_B)\,. \end{aligned} \tag{17}$$

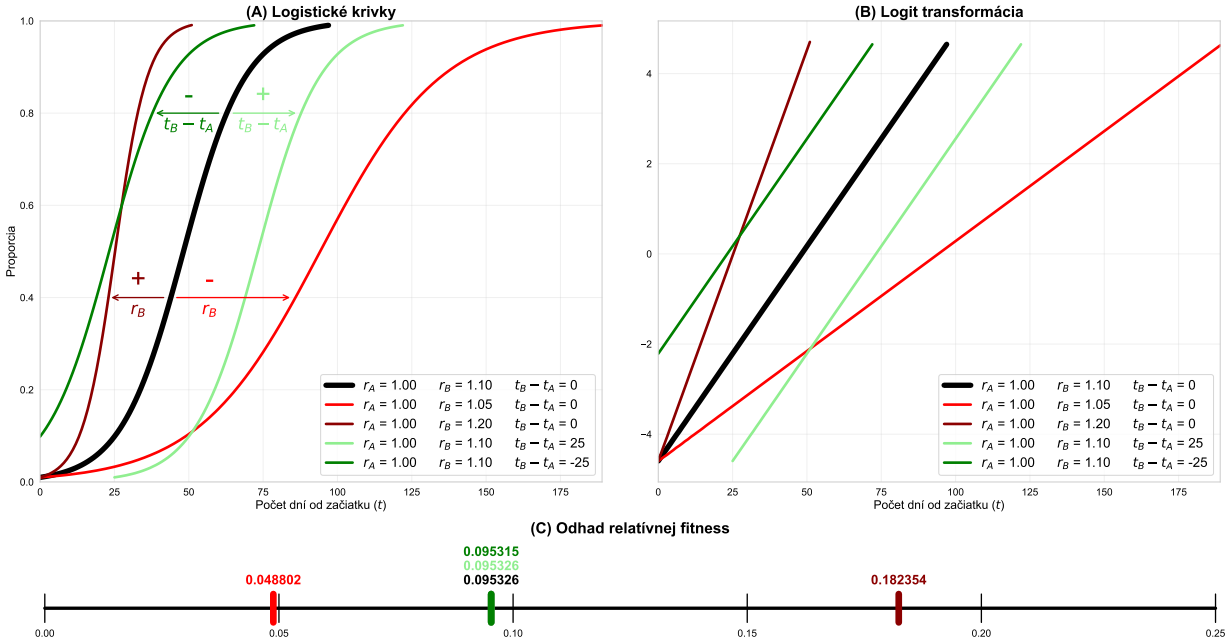
Všimnime si, že lineárny člen t opäť závisí iba od výrazu $r_A - r_B$. Fitness nám teda po aplikovaní logit transformácie $L(X(t))$ vyjde rovnaká ako pri deterministickom modeli bez konštantného časového oneskorenia, lebo pri derivácii nám konštantný výraz $r_A T_A - r_B T_B$ zanikne. V nasledujúcej časti overíme analytický výsledok na simuláciách v Pythone.

3.1.3 Simulácia v Pythone

V tejto časti overíme pozorovania z častí 3.1.1 a 3.1.2. Vybrali sme si niekoľko reprezentatívnych scenárov, ktoré sme následne nasimulovali v Pythone. Zamerali sme sa na dve hlavné oblasti: ako sa mení odhad relatívnej fitness s meniacimi sa hodnotami r_A, resp. r_B a či na jej hodnotu vplýva aj to, keď variant B začína neskôr alebo skôr ako variant A, čo môžeme brať ako formu konštantného oneskorenia variantu B. Označme čas t_A ako čas vstupu variantu A do populácie a t_B ako čas vstupu variantu B do populácie. Scenár, kedy variant B začne skôr ako variant A je nerealistický, ale v teoretickom prostredí nás môže zaujímať to, či po zaregistrovaní oboch variantov sa stále zachová logistika podielu B. Pre zjednodušenie budeme rozdiel medzi vstupom variantu B a vstupom variantu A do populácie označovať ako $t_B - t_A$.

Pre všetky simulácie sme uvažovali hodnotu $r_A = 1$, čo znamená, že variant A zostane na konštantnom počte infikovaných, ale predmetom bude skúmanie výhody variantu B oproti variantu A, teda pôjde iba o podiel $\frac{r_B}{r_A}$, nie o ich konkrétne hodnoty. Výsledky simulácií je možné vidieť na Obr. 4.

Postup spočíva v tom, že sme si zvolili štartovacie počty infikovaných $N_A(0) = 9900$ a $N_B(0) = 100$. Všimnime si, že podiel variantu B je v čase 0 rovný 99%. V Pythone sme implementovali diskrétny exponenciálny model pre dva varianty vychádzajúci z (3), ktorý ale vďaka dostatočnej granularite nebude prekážkou. Následne sme simulovali priebeh šírenia variantov A a B na grafe (A) dovtedy, pokiaľ variant B vďaka vyššej hodnote r_B nedosiahol 99% podiel. Posledná krivka začína technicky v čase $t = -25$, ale kvôli absencii variantu A sú všetky body do $t = 0$ rovné 1. Predposledná krivka začína od času $t = 25$, lebo až vtedy sa v populácii objaví variant B a teda všetky hodnoty do $t = 25$ sú rovné 0.



Obr. 4: Simulácie exponenciálneho modelu a inferencia relatívnej fitness. (A) Proporcia variantu B v čase pre vybrané kombinácie parametrov. **(B)** Logit transformácia kriviek z grafu (A). **(C)** Os s inferovanými hodnotami relatívnej fitness. Okolie bodu 0.095326 je zdieľané pre tri kombinácie parametrov.

Potom sme logistické krivky pomocou logit transformácie transformovali na lineárne krivky, ktoré je možné vidieť na grafe (B). Už teraz je možné pozorovať sklony jednotlivých lineárnych kriviek, pričom strmší sklon bude indikovať vyššiu relatívnu fitness. Následne sme pomocou nafitovania lineárnej regresie a vyčíslenia lineárneho parametra vypočítali relatívnu fitness, ktorú sme pre každú krivku zakreslili do grafu (C).

Hodnota relatívnej fitness nám pre základný scenár vyšla 0.095326. Uvedomme si, že toto je vlastne len rastový koeficient pre spojitý model r_S. Ak sa vrátime k výrazom v (5), tak výraz $\ln(1 + 0.1)$, kde 0.1 signalizuje rozdiel $r_B - r_A$, sa bude rovnať práve hodnote vypočítanej relatívnej fitness. Ak je rozdiel $r_B - r_A$ väčší, tak sa podľa očakávaní zväčšila aj hodnota relatívnej fitness, pričom pri nižšom $r_B - r_A$ sa naopak zmenšila. Posledné dva scenáre uvažovali len s posunom začiatku šírenia variantu B bez zmeny rastových koeficientov r_A a r_B, ale napriek tomu sme dostali rovnakú relatívnu fitness ako pri prvom scenári (až na zanedbateľné numerické odchýlky kvôli absencii časti poslednej s-krivky). Toto potvrdzuje náš analytický výsledok z časti 3.1.2, že konštantný časový posun nevp-

lýva na odhad hodnoty relatívnej fitness.

V nasledujúcich častiach sa budeme venovať náhodne rozdeleným oneskoreniam, pričom vyskúšame viacero rozdelení a budeme analyticky aj simuláciami zisťovať potenciálne odchýlky v odhade relatívnej fitness.

3.2 Model s normálne rozdeleným náhodným oneskorením

V tejto časti budeme zisťovať robustnosť exponenciálneho modelu po pridaní normálne rozdeleného náhodného časového oneskorenia pre variant A aj pre variant B. Na začiatok sme vybrali normálne rozdelenie, lebo pomocou jeho symetrie vieme analyticky vypočítať hodnotu relatívnej fitness, ktorá bude slúžiť ako odrazový bod pre numerické simulácie aj ďalšie rozdelenia. V rámci analytickej časti sa pozrieme na rôzne režimy náhodného oneskorenia a zistíme, pri ktorom režime registrujeme najväčšie odchýlky od deterministickej hodnoty relatívnej fitness. Následne naše zistenia overíme aj na numerických simuláciách v Pythone.

Odvodenie výpočtu relatívnej fitness

Pri výpočte relatívnej fitness budeme používať náhodné premenné τ_A a τ_B, ktoré pochádzajú z normálneho rozdelenia určeného parametrami μ_A a σ_A pre variant A, resp. μ_B a σ_B pre variant B. Nech $\widetilde{N}_A(t)$, resp. $\widetilde{N}_B(t)$ označuje počet jedincov infikovaných variantom A, resp. B v čase t. Nech $r_A, r_B, \widetilde{N}_A(0), \widetilde{N}_B(0)$ sú vopred určené konštanty. Definujeme nasledovný systém rovníc šírenia:

$$\begin{aligned} \widetilde{N}_A(t) &= \widetilde{N}_A(0)e^{r_A(t-\tau_A)}\,, \tau_A \sim \mathcal{N}(\mu_A, \sigma_A^2)\,, \\ \widetilde{N}_B(t) &= \widetilde{N}_B(0)e^{r_B(t-\tau_B)}\,, \tau_B \sim \mathcal{N}(\mu_B, \sigma_B^2)\,. \end{aligned} \tag{18}$$

Podiel variantu B v čase t budeme označovať ako $\widetilde{X}(t)$:

$$\widetilde{X}(t) = \frac{\widetilde{N}_B(t)}{\widetilde{N}_A(t) + \widetilde{N}_B(t)}\,. \tag{19}$$

Samotné $\widetilde{X}(t)$ môže mať kvôli náhodnosti veľmi veľa realizácií. Aby sme mohli vypočítať fitness, budeme pracovať so strednou hodnotou $\widetilde{X}(t)$, ktorú označíme $\mathbb{E}[\widetilde{X}(t)]$. Najprv si zlomok prepíšeme analogicky podľa časti 3.1 do tvaru štandardnej logistickej funkcie,

pričom upravíme aj argument logaritmu:

$$\begin{aligned}\mathbb{E}[\widetilde{X}(t)] &= \mathbb{E}\left[\frac{1}{1+\exp\left(\ln\left(\frac{\widetilde{N}_A(t)}{\widetilde{N}_B(t)}\right)\right)}\right] = \\ &= \mathbb{E}\left[\frac{1}{1+\exp\left(\ln\left(\frac{\widetilde{N}_A(0)}{\widetilde{N}_B(0)}\right)+(r_A-r_B)t-(r_A\tau_A-r_B\tau_B)\right)}\right].\end{aligned} \tag{20}$$

Všimnime si, výraz $\ln\left(\frac{\widetilde{N}_A(0)}{\widetilde{N}_B(0)}\right)+(r_A-r_B)t$ je deterministický, lebo nezávisí od žiadnej náhodnej premennej. Naopak, výraz $r_A\tau_A - r_B\tau_B$ je stochastický a ide o lineárnu kombináciu náhodných premenných τ_A, τ_B. Pre zjednodušenie budúcich výpočtov si každú časť označíme jedným písmenom:

$$\begin{aligned}D(t) &\coloneqq \ln\left(\frac{\widetilde{N}_A(0)}{\widetilde{N}_B(0)}\right)+(r_A-r_B)t\,,\\ S &\coloneqq r_A\tau_A - r_B\tau_B\,.\end{aligned} \tag{21}$$

Nová náhodná premenná S má normálne rozdelenie, lebo aj τ_A, τ_B sú normálne rozdelené a ich lineárna kombinácia tento atribút zachováva. Vypočítame si strednú hodnotu a varianciu a zapíšeme rozdelenie náhodnej premennej S:

$$\begin{aligned}m_S &\coloneqq \mathbb{E}[S] = r_A\mu_A - r_B\mu_B\,,\\ \sigma_S^2 &\coloneqq Var(S) = r_A^2\sigma_A^2 + r_B^2\sigma_B^2 - 2r_Ar_BCov(\tau_A,\tau_B) = r_A^2\sigma_A^2 + r_B^2\sigma_B^2 - 2r_Ar_B\sigma_A\sigma_B\rho\,,\\ S &\sim \mathcal{N}(m_S,\sigma_S^2)\,.\end{aligned} \tag{22}$$

Následne prepíšeme $\mathbb{E}(\widetilde{X}(t))$ do tvaru štandardnej logistickej funkcie (8):

$$\begin{aligned}\mathbb{E}[\widetilde{X}(t)] &= \mathbb{E}\left[\frac{1}{1+\exp\left(\ln\left(\frac{\widetilde{N}_A(t)}{\widetilde{N}_B(t)}\right)\right)}\right] = \mathbb{E}\left[\frac{1}{1+\exp\left(D(t)-S\right)}\right] = \\ &= \mathbb{E}[l(S-D(t))] =: I(t)\,.\end{aligned} \tag{23}$$

Výraz $I(t)$ ako strednú hodnotu vieme zapísať aj vo forme integrálu:

$$I(t) = \mathbb{E}[l(S-D(t))] = \int_{-\infty}^{\infty} l(s-D(t))\frac{1}{\sqrt{2\pi}\sigma_S}\exp\left(-\frac{(s-m_S)^2}{2\sigma_S^2}\right)ds\,. \tag{24}$$

Všimnime si, že v deterministickom modeli sme po logit transformácii mohli počítať deriváciu v ľubovoľnom bode logistickej funkcie, lebo v každom bode mala funkcia rovnaký sklon. Pri náhodných dátach toto ale nevieme úplne zaručiť, keďže funkcia podlieha náhodnosti a nemusí mať perfektný logistický rast. Našou stratégiou bude spočítať

deriváciu logit transformácie v bode, kde sa stredná hodnota rovná $\frac{1}{2}$, pretože vtedy má logistická funkcia najstrmší rast. Pripomeňme si, že štandardná logistická funkcia l podľa vzorca (8) má najstrmší rast (teda nadobúda hodnotu $\frac{1}{2}$) v bode 0. Definujme $\mathbb{E}[S - D(t)] = 0$. Označme t_{stred} ako čas, v ktorom bude táto definícia pravdivá. Neskôr dokážeme, v čase t_{stred} funkcia l nadobúda hodnotu $\frac{1}{2}$. Budeme teda zároveň predpokladať, že $\mathbb{E}(S) = D(t_{stred})$, teda $m_S = D(t_{stred})$. Čas t_{stred} sa dá následne odvodiť z definície $D(t)$:

$$t_{stred} = \frac{m_S - \ln\left(\frac{\widetilde{N}_A(0)}{\widetilde{N}_B(0)}\right)}{r_A - r_B}\,. \tag{25}$$

Existujú ale isté limitácie výrazu $m_S - D(t)$. Ak bude rozdiel medzi μ_A a μ_B veľký, tak člen m_S bude nadobúdať vysoké hodnoty. Následne by sa teoreticky mohlo stať, že by sme nemali také $D(t)$, pre ktoré by bola splnená podmienka $m_S = D(t_{stred})$, teda by sme nevedeli nájsť čas t_{stred}. Taktiež budeme pracovať s predpokladom, že variant B nezistíme skôr ako variant A. Potom dostaneme proporciu variantu B, ktorá začína od nuly.

Vypočítame si deriváciu logit transformácie. Derivácia funkcie $L(p)$ vyzerá nasledovne:

$$\frac{d}{dt}L(p) = \frac{d}{dt}\ln\left(\frac{p}{1-p}\right) = \frac{1-p}{p}\frac{p'(1-p)+pp'}{(1-p)^2} = \frac{p'}{p(1-p)}\,. \tag{26}$$

Pre našu funkciu $I(t)$ budeme mať nasledovný výraz:

$$\frac{d}{dt}L(I(t)) = \frac{\frac{d}{dt}I(t)}{I(t)(1-I(t))}\,. \tag{27}$$

Najprv vypočítame deriváciu výrazu $I(t)$ podľa t. Keďže sa integruje podľa hodnôt náhodnej premennej S, vieme deriváciu vsunúť dovnútra integrálu a spočítať ju:

$$\begin{aligned}\frac{d}{dt}I(t) &= \int_{-\infty}^{\infty}\frac{d}{dt}l(s-D(t))\frac{1}{\sqrt{2\pi}\sigma_S}\exp\left(-\frac{(s-m_S)^2}{2\sigma_S^2}\right)ds = \\ &= -\frac{dD}{dt}\int_{-\infty}^{\infty}l'(s-D(t))\frac{1}{\sqrt{2\pi}\sigma_S}\exp\left(-\frac{(s-m_S)^2}{2\sigma_S^2}\right)ds = \\ &= (r_B - r_A)\mathbb{E}(l'(S-D(t)))\,.\end{aligned} \tag{28}$$

Následne dosadíme:

$$\frac{d}{dt}L(I(t)) = \frac{(r_B - r_A)\mathbb{E}[l'(S-D(t))]}{I(t)(1-I(t))}\,. \tag{29}$$

Zostáva nám už iba určiť samotný sklon derivácie v čase t_{stred}. Pred výpočtom derivácie logit transformácie sme predpokladali, že vezmeme t_{stred} také, že $\mathbb{E}(S) = D(t_{stred})$, teda

$m_S = D(t_{stred})$. Definujme novú náhodnú premennú $U := S - D(t_{stred})$, kde $U \sim \mathcal{N}(0, \sigma_S^2)$, lebo $D(t_{stred})$ len zabezpečuje posun rozdelenia náhodnej premennej S o hodnotu m_S v čase t_{stred}. Zostáva ešte ukázať, že výber bodu t_{stred} bol správny. Predpokladali sme, že $\mathbb{E}(S - D(t_{stred})) = 0$. Funkcia $l(u)$ ako logistická funkcia má vlastnosť $l(u) + l(-u) = 1$. Toto platí, ak U je symetrické okolo 0, teda hustota $f_U(u)$ spĺňa $f_U(u) = f_U(-u)$, pričom $\mathbb{E}(U) = 0$. Potom bude platiť nasledovná rovnosť:

$$\begin{aligned} \mathbb{E}(l(u)) &= \int_{-\infty}^{\infty} l(u) f_U(u)\, du = \frac{1}{2} \int_{-\infty}^{\infty} (l(u) + l(-u)) f_U(u)\, du = \\ &= \frac{1}{2} \int_{-\infty}^{\infty} f_U(u)\, du = \frac{1}{2}\,. \end{aligned} \tag{30}$$

Ukázali sme, že za predpokladu $\mathbb{E}(S - D(t_{stred})) = 0$ funkcia $l(u)$ nadobúda hodnotu $\frac{1}{2}$, pričom $\mathbb{E}(U) = 0$. Preto bol bod t_{stred} vhodne zvolený a zachovali sme konzistenciu s deterministickým modelom. Následne dosadíme do (29) a počítame pre $t = t_{stred}$:

$$\begin{aligned} \frac{d}{dt} L(I(t_{stred})) &= \frac{(r_B - r_A)\mathbb{E}[l'(S - D(t_{stred}))]}{I(t_{stred})(1 - I(t_{stred}))} = \frac{(r_B - r_A)\mathbb{E}[l'(u)]}{\frac{1}{2}(1 - \frac{1}{2})} = \\ &= 4(r_B - r_A)\mathbb{E}[l'(u)]\,. \end{aligned} \tag{31}$$

Všimnime si deterministickú časť $4(r_B - r_A)$. Všetko ostatné je stochastická časť, ktorú sa pokúsime rozvinúť pomocou Taylorových radov a určiť, či sa dá izolovať istá korekcia výrazu $4(r_B - r_A)$, ktorá by významne ovplyvnila odhad fitness. Ak sa korekcia bude rovnať $\frac{1}{4}$, tak dostaneme ten istý odhad fitness ako pri deterministickom modeli. Keďže náhodnosť pochádza z parametrov σ_A, σ_B, budeme chcieť zistiť, pri akom výbere týchto parametrov budeme registrovať najväčšie odchýlky odhadu fitness oproti skutočnej fitness. Najprv si vyberieme režim $\sigma_B \ll \sigma_A$. Vráťme sa k výrazu (28), kde vystupuje člen $\mathbb{E}(l'(S - D(t)))$. V nájdenom čase t_{stred} bude platiť $\mathbb{E}[l'(S - D(t))] = \mathbb{E}[l'(u)]$:

$$\begin{aligned} \frac{d}{dt} I(t_{stred}) &= (r_B - r_A)\mathbb{E}[l'(u)] = \\ &= (r_B - r_A) \int_{-\infty}^{\infty} l'(u) \frac{1}{\sqrt{2\pi}\sigma_S} \exp\left(-\frac{u^2}{2\sigma_S^2}\right) du =: Y\,. \end{aligned} \tag{32}$$

Pri expanzii výrazu Y sa budeme venovať výrazu $l'(u)$, pretože ten môžeme následne presunúť z integrálu von a zvyšok integrálu sa zintegruje na 1. Vypočítame si preto prvé

dve derivácie $l'(u)$, pričom $l'(u)$ označíme ako $f(u)$ pre prehľadnosť:

$$\begin{aligned} f(u) \coloneqq l'(u) &= \frac{e^{-u}}{(1+e^{-u})^2}\,, \\ f'(u) \coloneqq l''(u) &= \frac{-e^{-u}+e^{-3u}}{(1+e^{-u})^4}\,, \\ f''(u) \coloneqq l'''(u) &= \frac{(e^{-u}+e^{-3u})(1+e^{-u})-4e^{-u}(-e^{-u}+e^{-3u})(1+e^{-u})^3}{(1+e^{-u})^8}\,. \end{aligned} \tag{33}$$

Taylorov rozvoj pre $E[f(u)]$ vyzerá nasledovne:

$$\begin{aligned} \mathbb{E}[f(u)] &\approx E\left[f(0)+f'(0)u+\frac{1}{2}f''(0)u^2++\frac{1}{6}f'''(0)u^3+O(u^4)\right] = \\ &= f(0)+f'(0)\mathbb{E}[u]+\frac{1}{2}f''(0)\mathbb{E}[u^2]+\frac{1}{6}f'''(0)\mathbb{E}[u^3]+O(u^4)\,. \end{aligned} \tag{34}$$

Na základe parity vieme povedať, že U má všetky nepárne momenty nulové, teda výrazy $\mathbb{E}[u^3]$, $\mathbb{E}[u^5]$, atď. sa budú rovnať nule. Navyše $U \sim \mathcal{N}(0,\sigma_S^2)$, preto $\mathbb{E}[u]=0$ a $\mathbb{E}[u^2]=\sigma_S^2$. Následne vypočítame $f(0)$, $f'(0)$ a $f''(0)$ a dosadíme:

$$\mathbb{E}(f(u)) \approx f(0)+\frac{\sigma_S^2}{2}f''(0)+O(\sigma_S^4) = \frac{1}{4}-\frac{\sigma_S^2}{16}+O(\sigma_S^4)\,. \tag{35}$$

Nakoniec dosadíme do výrazu (32):

$$\begin{aligned} Y &= (r_B-r_A)\left(\frac{1}{4}-\frac{\sigma_S^2}{16}+O(\sigma_S^4)\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_S}\exp\left(-\frac{(s-m_S)^2}{2\sigma_S^2}\right)ds = \\ &= (r_B-r_A)\left(\frac{1}{4}-\frac{\sigma_S^2}{16}+O(\sigma_S^4)\right)\,. \end{aligned} \tag{36}$$

Pripomeňme si, že podľa výrazu (32) vieme zapísať odhad fitness v čase t_{stred} ako $4(r_B-r_A)E(l'(u))$, teda $4(r_B-r_A)\left(\frac{1}{4}-\frac{\sigma_S^2}{16}+O(\sigma_S^4)\right)$. Ja zrejmé, že ak $\sigma_S^2=0$, tak sa korekcia bude rovnať $\frac{1}{4}$ a teda budeme mať čisto deterministický výsledok. Ak ale $\sigma_S^2>0$, tak odhad fitness bude menší ako ten deterministický. Horná hranica odhadu fitness bude $(r_B-r_A)\,\forall\sigma_S^2$. Pozrieme sa, od čoho závisí σ_S^2 pri režime $\sigma_B \ll \sigma_A$:

$$\sigma_S^2 = r_A^2\sigma_A^2+r_B^2\sigma_B^2-2r_Ar_B\sigma_A\sigma_B\rho = r_A^2\sigma_A^2\left(1+\frac{r_B^2}{r_A^2}\frac{\sigma_B^2}{\sigma_A^2}-2\frac{r_B}{r_A}\frac{\sigma_B}{\sigma_A}\rho\right) \approx r_A^2\sigma_A^2\,. \tag{37}$$

Zlomok $\frac{\sigma_B}{\sigma_A}$ sme pri vybranom režime považovali za veľmi malé číslo blízke nule, pričom vo výpočte sme dosadili nulu. Môžeme vidieť, že σ_S^2 závisí od σ_A. Čím je σ_A väčšie, tým viac sa vzdiali od malého σ_B a zväčší sa σ_S^2. Väčšie σ_S^2 potom zmenší korekciu, čím sa zmenší aj odhad fitness a viac sa vzdiali od deterministického odhadu fitness. Zmenšenie odhadu fitness oproti realite spôsobuje, že variant B zdanlivo predstavuje menšie riziko, než aké v

skutočnosti predstavuje. Čím je tento jav spôsobený a či sa prejaví v simuláciách uvidíme neskôr v nasledujúcej časti.

Existujú ale aj iné režimy, pri ktorých sa dá skúmať odchýlka odhadu relatívnej fitness. Keby sme mali napríklad režim $\sigma := \sigma_B \approx \sigma_A$, tak nám do výrazu σ_S^2 bude vstupovať aj korelácia medzi variantmi ρ. Pre rôzne hodnoty ρ budeme dostávať odlišné výrazy σ_S^2, preto sme do Tabuľky 3 nižšie spísali niekoľko pod-režimov s vypočítanou σ_S^2:

Pod-režim	σ_S^2
$\rho \approx 1$	$\sigma_S^2 \approx \sigma^2(r_A - r_B)^2$
$\rho \approx 0$	$\sigma_S^2 \approx \sigma^2(r_A^2 + r_B^2)$
$\rho \approx -1$	$\sigma_S^2 \approx \sigma^2(r_A + r_B)^2$

Tabuľka 3: Výpočet σ_S^2 na základe vybraného pod-režimu

Môžeme si všimnúť, že na rozdiel od režimu $\sigma_B \ll \sigma_A$ nebude odchýlka odhadu fitness závisieť od rozdielu medzi zašumenosťou dát, ale od škálovania rastových koeficientov r_A a r_B. Pri $\rho \approx -1$ budeme dostávať výrazne väčšie odchýlky ako pri $\rho \approx 0$. Súčasne konštatujeme, že ak máme režim $\sigma := \sigma_B \approx \sigma_A$, tak zrejme môžeme predpokladať aspoň čiastočne kladnú koreláciu medzi dvoma variantmi, preto výsledok pri $\rho \approx 1$ bude pre nás najviac relevantný. Zároveň môžeme pozorovať, že nám hodnota σ_S^2 bude priamo závisieť od rozdielu medzi r_A a r_B.

V nasledujúcej časti budeme pomocou simulácií v Pythone testovať rôzne režimy náhodného oneskorenia a overovať, či sa zistenia z tejto časti prejavia aj v realite.

Numerické experimenty

V tejto časti budeme simulovať náhodné oneskorenie pre variant A aj variant B a pozrieme sa na to, akým spôsobom vplýva na odhad relatívnej fitness. Nasledujúce simulácie sme vykonali v Pythone s použitím knižníc `numpy`, `pandas`, `matplotlib`, `sklearn` a `scipy` [7].

Podstatou simulácií bude zhodnotiť, akým spôsobom vplýva náhodný šum v dátach na presnosť modelu. Presnosť modelu budeme hodnotiť podľa odchýlky relatívnej fitness v

modeli bez náhodného šumu oproti relatívnej fitness v modeli s náhodným šumom. Opäť uvažujme dva varianty vírusu, kde A je udomácnený variant a B je invadujúci variant. Budeme pracovať v diskrétnom čase, ktorý má ale dostatočnú granularitu na to, aby sa model všeobecne dal považovať za spojitý (6).

V časti 3.1.3 sme vykonávali simulácie na exponenciálnom modeli bez náhodného oneskorenia. V tejto časti budeme uvažovať normálne rozdelené náhodné oneskorenie, pričom ho budeme modelovať pre variant A aj B samostatne. Budeme ho modelovať pomocou rozdelenia s určitou strednou hodnotou a disperziou:

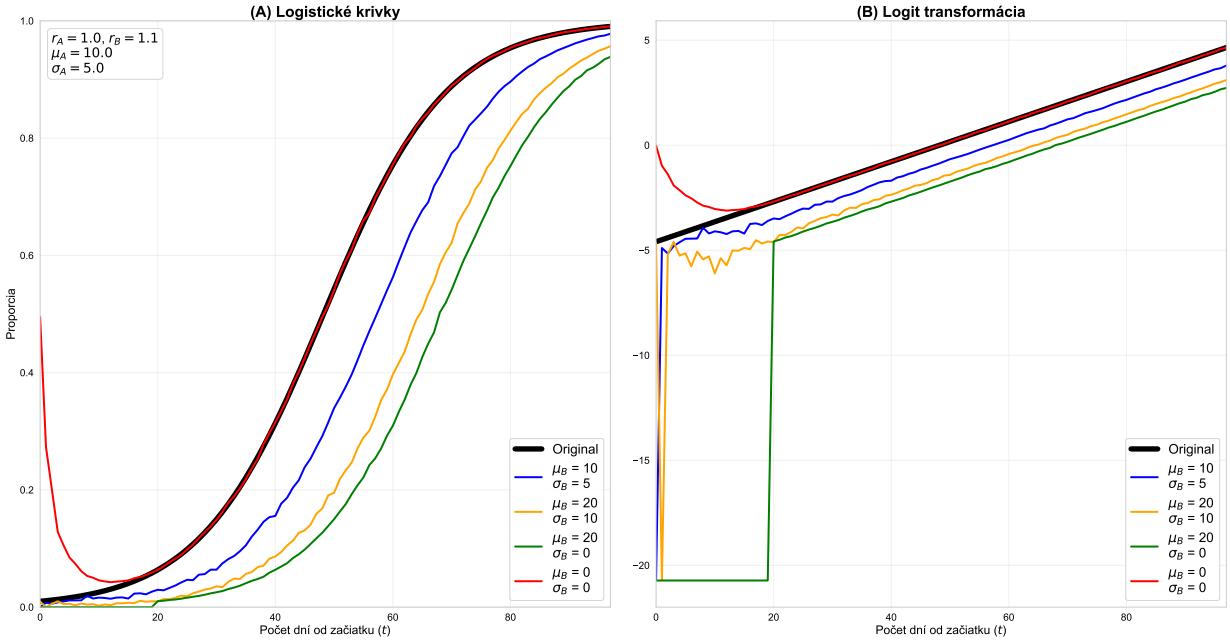
$$\begin{aligned} \tau_A &\sim \mathcal{N}(\mu_A, \sigma_A^2)\,, \\ \tau_B &\sim \mathcal{N}(\mu_B, \sigma_B^2)\,. \end{aligned} \tag{38}$$

Pripomeňme si, že pri základnom exponenciálnom modeli bez oneskorenia sme pracovali len s parametrami r_A, r_B a t_B, pričom t_B nemalo na odhad relatívnej fitness žiadny vplyv. Pri exponenciálnom modeli s náhodným oneskorením budeme pracovať so 7 parametrami: r_A, r_B, t_B, μ_A, σ_A, μ_B a σ_B.

Je dôležité si uvedomiť, že normálne rozdelenie je spojité, kým modelované oneskorenia v realite bývajú zachytené viac diskrétne (napr. raz za deň). Preto sme sa rozhodli diskretizovať realizácie náhodných premenných τ_A a τ_B do denných binov pomocou multinomického rozdelenia [21]. Pre odstránenie outlierov z chvostov rozdelenia a tiež elimináciu záporných hodnôt sme sa rozhodli brať do úvahy oneskorenia iba na intervale $(\mu_A - 2\sigma_A, \mu_A + 2\sigma_A)$, resp. $(\mu_B - 2\sigma_B, \mu_B + 2\sigma_B)$. Každé oneskorenie bude mať okolo seba bin s dĺžkou 1 (napr. náhodné oneskorenie na intervale $(4.5, 5.5)$ bude definované ako 5). Označme všeobecne dolnú hranicu binu i ako a_i a hornú hranicu binu i ako b_i. Pravdepodobnosť, že nejaká realizácia skončí v bine i sa dá vypočítať ako rozdiel hodnôt kumulatívnej distribučnej funkcie normálneho rozdelenia v bodoch b_i a a_i. Pre zabezpečenie jednotkového súčtu pravdepodobností sme vykonali renormalizáciu jednotlivých pravdepodobností.

Program v Pythone sme upravili tak, aby reflektoval zmeny popísané vyššie. Tento prístup je štatisticky ekvivalentný, ale pre program oveľa rýchlejší. Vygenerovali sme si pre každý deň toľko realizácií, koľko bolo v ten deň prípadov a následne sme ich priradili do správ-

neho binu, pričom každý bin predstavoval nejaký deň. Takto dostaneme nové hodnoty podielu variantu B v čase. Pre ukážku sme si vybrali viacero extrémnejších kombinácií μ_A, σ_A, μ_B a σ_B, ktoré sme následne aplikovali na jednu zvolenú kombináciu $r_A = 1$, $r_B = 1.1$ a $t_B = 0$. Tiež sme zafixovali dva zo štyroch ostatných parametrov, a to μ_A na hodnotu 10 a σ_A na hodnotu 5. Výsledky analýzy je možné vidieť na Obr. 5.



Obr. 5: Logistické krivky a ich logit transformácia pri aplikovaní normálne rozdeleného náhodného oneskorenia. (A) Logistické krivky pre jednotlivé režimy parametrov. **(B)** Logit transformácie logistických kriviek z grafu (A).

Zo časti 3.1.2 vieme, že deterministické oneskorenie neovplyvní hodnotu relatívnej fitness. Z toho je zrejmé, že za predpokladu $\sigma_A = 0$ a $\sigma_B = 0$ a pre ľubovoľné hodnoty μ_A a μ_B taktiež nebudeme pozorovať odchýlku v relatívnej fitness. Režimy parametrov na Obr. 5 sme zvolili tak, aby sme neporušili vyššie zavedený predpoklad $\mu \geq 2\sigma$. Pre všetky režimy môžeme pozorovať istý šum hlavne na začiatku kriviek, ale samotný stred krivky ostáva logistický, a preto na prvý pohľad nepredpokladáme veľkú odchýlku v relatívnej fitness. Pre výpočet relatívnej fitness sme odsekli konce krivky tak, že berieme do úvahy iba hodnoty, v ktorých je krivka medzi 10% a 90% proporcie variantu B. Toto môžeme spraviť, lebo inferencia relatívnej fitness pochádza práve z exponenciálneho rastu logistickej krivky v jej strede, zároveň však nechceme odseknúť z krivky viac, než je nutné. Taktiež sa týmto krokom zbavíme outlierov, ktorí mohli vzniknúť posunom prípadov variantu A alebo B.

Výsledky inferencie sme zapísali do Tabuľky 4.

Parametre	Pôvodná rel. fitness	Rel. fitness po pridaní oneskorenia	Rozdiel
$\mu_B = 10$, $\sigma_B = 5$	0.095319	0.095431	0.1181%
$\mu_B = 20$, $\sigma_B = 10$	0.095319	0.095433	0.1202%
$\mu_B = 20$, $\sigma_B = 0$	0.095319	0.09531	-0.0089%
$\mu_B = 0$, $\sigma_B = 0$	0.095319	0.072454	-23.9875%

Tabuľka 4: Hodnoty relatívnej fitness pre rôzne režimy parametrov a ich odchýlka od pôvodnej hodnoty relatívnej fitness.

Všimnime si, že tri zo štyroch režimov majú nízku odchýlku v relatívnej fitness. Posledný režim má ale veľmi veľkú odchýlku v relatívnej fitness, a to v neprospech variantu B. Ak sa vrátime naspäť k Obr. 5, tak logistická krivka režimu $\mu_B = 0$, $\sigma_B = 0$ (červená farba) vykazuje istú nekonzistentnosť na jej začiatku. Uvedomme si, že všetky režimy boli testované pri vstupe variantu B do populácie v čase $t = 0$, teda $t_B = 0$. Keďže je ale režim $\mu_B = 0$, $\sigma_B = 0$ založený na veľkom oneskorení prípadov variantu A, kým prípady variantu B zostávajú na mieste, tak nám umelo vzniká vysoký podiel variantu B v skorom štádiu pozorovania. Neporušíme predpoklad, že variant B nesmie byť objavený skôr ako variant A, ale napriek tomu spôsobíme viditeľne veľkú odchýlku relatívnej fitness, keďže body na začiatku krivky ovplyvnia smernicu nafitovanej lineárnej priamky tak, že bude menej strmá.

V tejto časti vyššie sme analyticky ukázali, že režim $\sigma_B \ll \sigma_A$ môže spôsobovať zápornú odchýlku relatívnej fitness od jej pôvodnej hodnoty. Berúc do úvahy už spomenutý predpoklad $\mu \geq 2\sigma$, presne toto naznačuje aj Obr. 5. Dôvod, kvôli ktorému ale táto odchýlka existuje, sa čiastočne vymyká realite, takže v záujme robustnejšieho zhodnotenia náhodných oneskorení sme pristúpili k modifikácii inferencie relatívnej fitness. Existujú dva hlavné spôsoby spôsoby, ktoré bez zmeny režimu parametrov majú potenciál zmeniť výsledok inferencie relatívnej fitness:

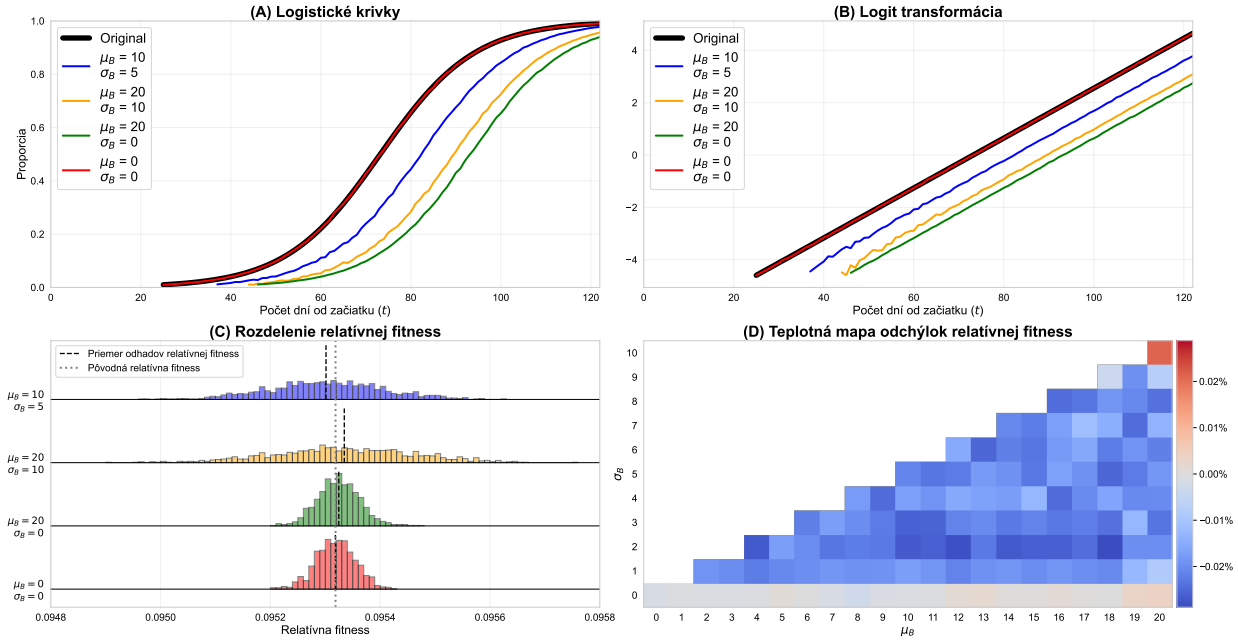
- Budeme brať do úvahy iba rastúcu časť s-krivky;
- Variant B bude objavený neskôr ako variant A, t.j. $t_B > 0$.

Jedno z možných riešení je brať do úvahy iba rastúcu časť s-krivky. Pri vyšších hodnotách μ_A a σ_A (napríklad aj takých, ako máme na Obr. 5) však stále v rastúcej časti s-krivky zachytíme aj časť nekonzistentnosti na začiatku. Ak by sme brali do úvahy menšiu časť s-krivky okolo stredu (napr. medzi 20% a 80% proporcie variantu B), tak stále existujú režimy, pri ktorých registrujeme odchýlku relatívnej fitness na základe tejto nekonzistentnosti. Tiež by sa patrilo poznamenať, že ani metodicky nedáva obrovský zmysel, aby variant B ako invadujúci variant bol na začiatku veľmi silný, potom výrazne spomalil a až následne začal postupne nahrádzať udomácnený variant A.

Preto bude lepšie použiť druhú možnosť, a to zabezpečiť, aby variant B vstúpil do populácie neskôr. Z časti 3.1.3 už vieme, že deterministická zmena vstupu variantu A alebo B neovplyvní odhad relatívnej fitness. Zvolili sme si $t_B = 25$ a opakovali simulácie. V rámci simulácií sme taktiež vykonali test robustnosti nasledovným spôsobom: vygenerovali sme náhodné oneskorenie 1000-krát a pre každú realizáciu sme vypočítali odhad relatívnej fitness. Následne sme zo všetkých realizácií vypočítali strednú hodnotu (priemer) odhadu relatívnej fitness a porovnali ju s pôvodnou hodnotou relatívnej fitness. Simulácie sme výrazne zrýchlili vďaka využitiu knižnice `multiprocessing` [6]. Výsledky simulácií je možné vidieť na Obr. 6 a v Tabuľke 5.

Parametre	Pôvodná rel. fitness	Rel. fitness po pridaní oneskorenia	Rozdiel
$\mu_B = 10$, $\sigma_B = 5$	0.095319	0.095301	-0.0182%
$\mu_B = 20$, $\sigma_B = 10$	0.095319	0.095335	0.0168%
$\mu_B = 20$, $\sigma_B = 0$	0.095319	0.095325	0.0062%
$\mu_B = 0$, $\sigma_B = 0$	0.095319	0.095318	-0.0002%

Tabuľka 5: Hodnoty relatívnej fitness pre rôzne režimy parametrov a ich odchýlka od pôvodnej hodnoty relatívnej fitness.



Obr. 6: Aplikácia normálne rozdeleného náhodného oneskorenia na inferenciu relatívnej fitness. (A) Logistické krivky pre jednotlivé režimy parametrov. **(B)** Logit transformácie logistických kriviek z grafu (A). **(C)** Histogram rozdelenia odhadov relatívnej fitness pre každý režim parametrov so zaznačenou strednou hodnotou (priemerom) odhadu. **(D)** Teplotná mapa pre všetky kombinácie μ_B a σ_B pri fixnom μ_A a σ_A. Zakreslené sú len tie kombinácie, ktoré spĺňajú predpoklad $\mu_B \geq 2\sigma_B$.

Všimnime si, že na grafe (A) aj (B) už nepozorujeme žiadne nekonzistentnosti. Keďže variant B začína teraz výrazne neskôr ako variant A, aj veľké oneskorenie prípadov variantu A za zachovania malého oneskorenia prípadov variantu B nemá veľký efekt na kvalitu kriviek. Na grafe (C) môžeme pozorovať histogram rozdelenia odhadov relatívnej fitness pre každý režim. Priemerné hodnoty odhadov fitness pre každú kombináciu μ_B, σ_B sme zaznačili do Tabuľky 5. Na grafe (D) sme za zachovania selekcie parametrov r_A, r_B, t_B, μ_A a σ_A vypočítali priemernú odchýlku fitness pre rôzne režimy parametrov μ_B a σ_B. Môžeme vidieť, že zobrazujeme iba selekcie, kde spĺňame predpoklad $\mu_B \geq 2\sigma_B$. Pre $\sigma_B = 0$ registrujeme najmenšie odchýlky relatívnej fitness, ale celkovo všade môžeme vidieť iba malé záporné odchýlky, okrem režimu so $\sigma_B = 10$, kde pozorujeme malé kladné odchýlky.

Kvôli upraveniu metodiky skúmania logistických kriviek a generovaniu normálne rozdelených náhodných oneskorení sme nenašli režim, ktorý by vykazoval signifikantné odchýlky

v odhadoch relatívnej fitness. Pre režim $\mu_B = 0$, $\sigma_B = 0$ sme prvotne videli veľkú zápornú odchýlku, ale po prehodnotení metodiky výpočtu sme ju prestali pozorovať. Keďže sme cielene otestovali extrémne a tým pádom trochu nerealistické režimy parametrov, môžeme si byť istí, že absencia vysokých odchýlok v takýchto režimoch bude viesť aj k absencii odchýlok v miernejších režimoch. Celkovo sme dostali výsledky podobné deterministickému modelu kvôli zanedbateľným odchýlkam v odhade relatívnej fitness.

V nasledujúcej časti sa budeme venovať inému rozdeleniu náhodných oneskorení, ktoré by sa malo ešte viac priblížiť realite: log-normálnemu rozdeleniu.

3.3 Model s log-normálne rozdeleným náhodným oneskorením

V tejto časti sa budeme venovať modifikácii modelu z časti 3.2. Normálne rozdelenie náhodného oneskorenia nahradíme log-normálnym rozdelením. Toto rozdelenie má ťažší pravý chvost a nenadobúda záporné hodnoty. Vo svete modelovania šírenia vírusov je realistickejšie ako normálne rozdelenie [2] a porovnaním s normálnym rozdelením dostaneme kompletnejší prehľad o robustnosti exponenciálneho modelu.

Porovnanie s normálnym rozdelením

V tejto časti sa budeme odvolávať na analytický výpočet relatívnej fitness z časti 3.2. Skúsime ho adaptovať na log-normálne rozdelenie a vysvetlíme, ktoré časti výpočtu sa už nebudú dať uskutočniť kvôli strate symetrie rozdelenia.

Log-normálne rozdelenie je rozdelenie náhodnej premennej, ktorej logaritmus je normálne rozdelený. Inými slovami, ak pre náhodnú premennú X platí $X \sim \text{Lognormal}(\mu, \sigma^2)$, potom $\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$. Toto rozdelenie nadobúda iba nezáporné hodnoty a má ťažký pravý chvost. Z toho vyplýva, že priemer a medián už nebudú mať rovnakú hodnotu, ale priemer bude väčší ako medián práve kvôli ťažkému pravému chvostu. V praxi je takéto rozdelenie viac realistické [2]. Hlavnú pozornosť budeme venovať simuláciám, kde zistíme prípadné rozdiely oproti normálnemu rozdeleniu.

Analytický výpočet relatívnej fitness začneme rovnako ako pri normálnom rozdelení.

Máme náhodné premenné τ_A a τ_B, ktoré pochádzajú z normálneho rozdelenia určeného parametrami μ_A a σ_A pre variant A, resp. μ_B a σ_B pre variant B. Nech $\widetilde{N}_A(t)$, resp. $\widetilde{N}_B(t)$ označuje počet jedincov infikovaných variantom A, resp. B v čase t. Nech $r_A, r_B, \widetilde{N}_A(0), \widetilde{N}_B(0)$ sú vopred určené konštanty. Pracovať budeme opäť s rovnakým systémom rovníc ako pri normálnom rozdelení, ale pri inak definovaných τ_A a τ_B:

$$\begin{aligned} \widetilde{N}_A(t) &= \widetilde{N}_A(0)e^{r_A(t-\tau_A)}\,, \tau_A \sim \text{Lognormal}(\mu_A, \sigma_A^2)\,, \\ \widetilde{N}_B(t) &= \widetilde{N}_B(0)e^{r_B(t-\tau_B)}\,, \tau_B \sim \text{Lognormal}(\mu_B, \sigma_B^2)\,. \end{aligned} \tag{39}$$

Podiel variantu B v čase t označíme ako $\widetilde{X}(t)$:

$$\widetilde{X}(t) = \frac{\widetilde{N}_B(t)}{\widetilde{N}_A(t) + \widetilde{N}_B(t)}\,. \tag{40}$$

Pre výpočet relatívnej fitness budeme pracovať so strednou hodnotou $\widetilde{X}(t)$, ktorú označíme $\mathbb{E}[\widetilde{X}(t)]$. Výraz $\mathbb{E}[\widetilde{X}(t)]$ vieme spočítať rovnakým spôsobom ako pri normálnom rozdelení:

$$\begin{aligned} \mathbb{E}[\widetilde{X}(t)] &= \mathbb{E}\left[\frac{1}{1+\exp\left(\ln\left(\frac{\widetilde{N}_A(t)}{\widetilde{N}_B(t)}\right)\right)}\right] = \mathbb{E}\left[\frac{1}{1+\exp\left(D(t)-S\right)}\right] = \\ &= \mathbb{E}[l(S-D(t))] =: I(t)\,. \end{aligned} \tag{41}$$

Pri normálnom rozdelení sme následne definovali výrazy $D(t)$ a S, pričom S bola náhodná premenná, ktorá vznikla lineárnou kombináciou náhodných premenných z normálneho rozdelenia a bola tiež normálne rozdelená. Keďže τ_A a τ_B sú log-normálne rozdelené, tak ich lineárna kombinácia nemá definované rozdelenie. Odvodenie v (41) stále platí bez ohľadu na rozdelenie náhodnej premennej S. Výraz $I(t)$ vieme tiež prepísať na integrál, ale iba vo všeobecnom tvare s použitím funkcie hustoty $f_S(s)$:

$$I(t) = \mathbb{E}[l(S-D(t))] = \int_{-\infty}^{\infty} l(s-D(t))f_S(s)\,ds\,. \tag{42}$$

Hustota $f_S(s)$ nemá explicitnú formuláciu, preto sa integrál nedá v tomto tvare explicitne vypočítať. Následne nám vznikne problém so symetriou, ktorý sa už nedá obísť. Definovali sme si novú náhodnú premennú $U := S - D(t_{stred})$, pričom bod t_{stred} označoval stred s-krivky, teda bod, ktorý spĺňa $\mathbb{E}(S - D(t_{stred})) = 0$, aby funkcia l nadobúdala hodnotu $\frac{1}{2}$. Kvôli absencii symetrie ale nebude tento predpoklad splnený, lebo výraz $\mathbb{E}(l(u)) = \frac{1}{2}$ nebude pravdivý. Logistická funkcia $l(u)$ má vlastnosť $l(u) + l(-u) = 1$, ktorá platí len

ak U je symetrické okolo 0, teda hustota $f_U(u)$ spĺňa $f_U(u) = f_U(-u)$, pričom $\mathbb{E}(U) = 0$. Inflexný bod s-krivky (jej stred) už nie je zosúladený s priemerom, čo pramení zo zvolenia asymetrického náhodného rozdelenia, teda napríklad log-normálneho.

Neskôr vo výpočte sme vykonali Taylorov rozvoj výrazu $\mathbb{E}(f(u))$, pri ktorom sme využili symetriu normálneho rozdelenia a mohli zanedbať všetky nepárne momenty. Toto nám dovolilo dospieť k dobre interpretovateľnému výsledku $\mathbb{E}(f(u)) \approx \frac{1}{4} - \frac{\sigma_S^2}{16} + O(\sigma_S^4)$. Pri log-normálnom rozdelení ale táto vlastnosť nie je a nepárne momenty nebudú nulové.

Ako sme pri normálnom rozdelení ukázali, analytické výsledky po určitých úpravách v simulácii neboli až tak významné a nenašli sme režim, pri ktorom by existovala výrazná odchýlka od pôvodnej hodnoty relatívnej fitness. V nasledujúcej časti sa vrátime k simuláciám, ale namiesto normálneho rozdelenia použijeme log-normálne rozdelenie. Následne ukážeme, či zmena rozdelenia odhalí režim, pri ktorom existuje nezanedbateľná odchýlka odhadu relatívnej fitness.

Numerické experimenty

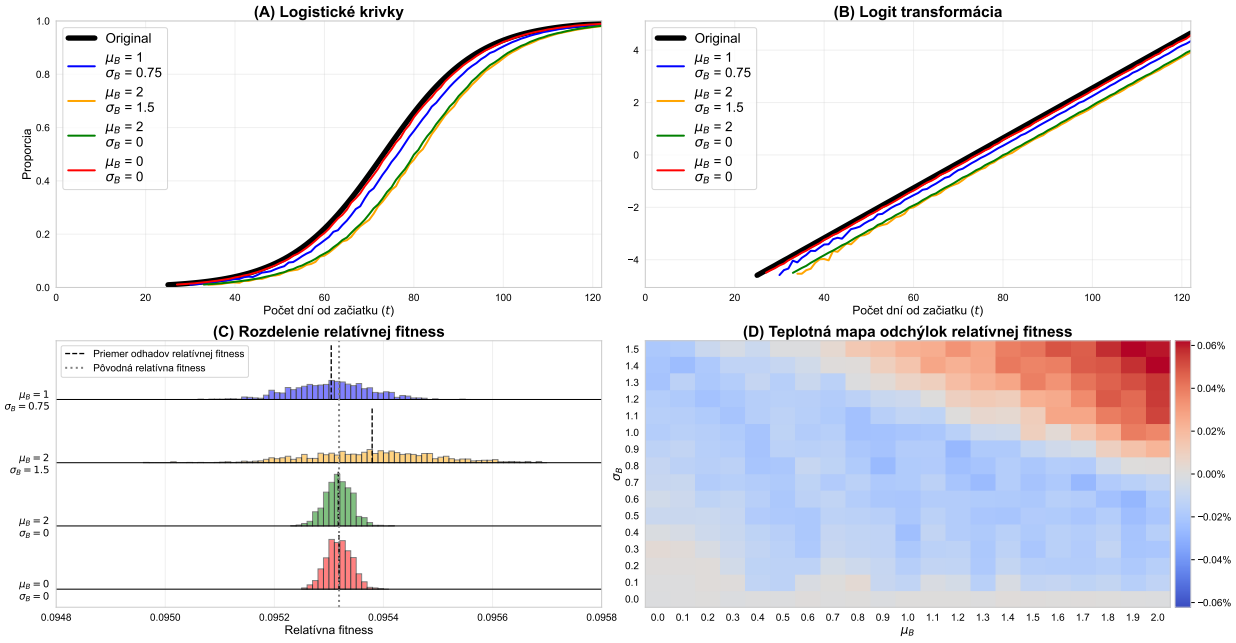
V tejto časti budeme skúmať, či zmena normálneho rozdelenia na log-normálne bude mať výrazný vplyv na odhad relatívnej fitness. Opäť budeme náhodné oneskorenie modelovať samostatne pre variant A aj variant B pomocou rozdelenia s určitou strednou hodnotou a disperziou:

$$\begin{aligned} \tau_A \sim Lognormal(\mu_A, \sigma_A^2)\,, \ln \tau_A \sim \mathcal{N}(\mu_A, \sigma_A^2)\,, \\ \tau_B \sim Lognormal(\mu_B, \sigma_B^2)\,, \ln \tau_B \sim \mathcal{N}(\mu_B, \sigma_B^2)\,. \end{aligned} \tag{43}$$

V časti 3.2 sme normálne rozdelenie diskretizovali pomocou denných binov, aby sme vedeli efektívne prerozdeliť počet prípadov v nejakom dni podľa pravdepodobnosti padnutia do určitého binu. Náhodné oneskorenie sme limitovali na interval $(\mu_A - 2\sigma_A, \mu_A + 2\sigma_A)$, resp. $(\mu_B - 2\sigma_B, \mu_B + 2\sigma_B)$. Využijeme ale vlastnosť $\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$ a pozmeníme interval na $(\exp(\mu_A - 2\sigma_A), \exp(\mu_A + 2\sigma_A))$. Toto nám dovolí rozumne ohraničiť hodnoty a zistiť, ktoré kombinácie μ a σ majú priťažký chvost a veľa hodnôt výrazne väčších ako medián. Hranice okrajových binov sme zaokrúhlili na celé čísla, aby sme vedeli jednoznačne prisúdiť každý bin určitému dňu. Keďže log-normálne rozdelenie nie je symetrické, tento krok

neovplyvní naše výsledky. Následne sme pomocou kumulatívnej distribučnej funkcie log-normálneho rozdelenia rozdelili biny rovnako ako pri normálnom rozdelení a vypočítali pravdepodobnosti padnutia realizácie do jednotlivých binov.

Na simuláciu sme si vybrali parametre $r_A = 1$ a $r_B = 1.1$ ako pri normálnom rozdelení. Keďže τ_A a τ_B sú teraz log-normálne rozdelené, vybrali sme si iné hodnoty parametrov pre variant A, a to $\mu_A = 1$ a $\sigma_A = 0.5$. Táto simulácia bude teda obsahovať výraznejšie oneskorenia prípadov variantu B. Parametre log-normálneho rozdelenia sme vyberali tak, aby realizácie náhodných premenných τ_A a τ_B boli približne v intervale $(0, 20)$. Takto budeme môcť porovnať výsledky s Obr. 6. Podobne ako predtým sme si vybrali niekoľko vzorových režimov, kde otestujeme robustnosť výpočtu relatívnej fitness, a následne sme vytvorili teplotnú mapu pre veľa kombinácií parametrov μ_B a σ_B. Výsledky sme zakreslili do Obr. 7 a Tabuľky 6.



Obr. 7: Aplikácia log-normálne rozdeleného náhodného oneskorenia na inferenciu relatívnej fitness. (A) Logistické krivky pre jednotlivé režimy parametrov. **(B)** Logit transformácie logistických kriviek z grafu (A). **(C)** Histogram rozdelenia odhadov relatívnej fitness pre každý režim parametrov so zaznačenou strednou hodnotou (priemerom) odhadu. **(D)** Teplotná mapa pre všetky kombinácie μ_B a σ_B pri fixnom μ_A a σ_A.

Parametre	Pôvodná rel. fitness	Rel. fitness po pridaní oneskorenia	Rozdiel
$\mu_B = 1$, $\sigma_B = 0.75$	0.095319	0.095304	-0.0151%
$\mu_B = 2$, $\sigma_B = 1.5$	0.095319	0.095379	0.0636%
$\mu_B = 2$, $\sigma_B = 0$	0.095319	0.095317	-0.0015%
$\mu_B = 0$, $\sigma_B = 0$	0.095319	0.095318	-0.0002%

Tabuľka 6: Hodnoty relatívnej fitness pre rôzne režimy parametrov a ich odchýlka od pôvodnej hodnoty relatívnej fitness.

Priemer odhadnutých hodnôt relatívnej fitness v grafe (C) sme opäť zaznačili do Tabuľky 6. Vidíme, že sme dostali podobné výsledky ako pri normálnom rozdelení. Kvôli vstupu variantu B až v čase $t = 25$ opäť eliminujeme prípadné nekonzistentnosti a odhad relatívnej fitness je pri všetkých režimoch v grafe (D) relatívne presný. Môžeme pozorovať, že pre nízke hodnoty σ_B registrujeme najmenšie odchýlky a pre kombinácie vysokého μ_B a σ_B registrujeme malé kladné odchýlky. Existuje potenciál na jemné zvyšovanie odchýlky relatívnej fitness pri zvyšujúcom sa μ_B a σ_B, ale samotné oneskorenia by už vtedy boli extrémne veľké, čo je veľmi nerealistické, preto existencia tohto fenoménu nie je relevantná.

Exponenciálny model je najjednoduchší zmysluplný model na modelovanie šírenia vírusov vďaka dobrej interpretovateľnosti výsledkov a možnosti modelovania proporcie variantu B logistickou krivkou. V nasledujúcej kapitole sa budeme venovať realistickejšiemu modelu definovanému systémom obyčajných diferenciálnych rovníc. Našou hlavnou otázkou bude to, či sa naďalej budú dať využiť princípy logistickej funkcie na inferenciu relatívnej fitness a ako budú na jej odhad vplývať oneskorenia prípadov.

4 SIR model

V tejto kapitole sa budeme zaoberať modelom priamo určeným na šírenie vírusov, a to SIR modelom. Vysvetlíme, v čom sa tento model líši oproti exponenciálnemu modelu a aké jeho aspekty ho robia realistickejším. Pomocou numerických experimentov v Pythone zistíme, za akých podmienok sa dá výhoda invadujúceho variantu oproti udomácnenému variantu modelovať logistickou krivkou podobne ako pri exponenciálnom modeli. Následne do základného modelu zahrnieme náhodné oneskorenie podľa vzoru z predchádzajúcej kapitoly. Nakoniec sa pozrieme na viacero modifikácií SIR modelu, ktoré ovplyvňujú imunitu voči jednotlivým variantom.

4.1 Základný model

SIR model bol vytvorený na začiatku 20. storočia R. Rossom a W. Hamerom [24]. Na ich prácu následne v roku 1927 nadviazali W. O. Kermack a A. G. McKendrick [14]. SIR model pozostáva z troch nelineárnych obyčajných diferenciálnych rovníc, ktoré reprezentujú zmenu v populácii vzhľadom na šírenie vírusu. Nech máme populáciu konštantnej veľkosti N. Túto populáciu následne rozdelíme na 3 skupiny s variabilnou veľkosťou na základe času t:

- $S(t)$ - susceptible (náchylný na infekciu);
- $I(t)$ - infectious (infekčný);
- $R(t)$ - recovered (uzdravený).

Pre tieto 3 skupiny (z angl. *compartments*) platí

$$S(t) + I(t) + R(t) = N\,. \tag{44}$$

Základný SIR model prináša so sebou viacero predpokladov:

- predpoklad veľkej a uzavretej populácie;
- v populácii sa nerodia ani neumierajú jedinci, teda populácia má konštantnú veľkosť;
- infekcia má nulovú inkubačnú dobu, teda jedinec je po infikovaní okamžite infekčný;

- jedinec má po prekonaní infekcie nekonečnú imunitu;
- každý jedinec má rovnakú šancu sa dostať do kontaktu s každým iným jedincom.

Posledný predpoklad je založený na proporciách jednotlivých skupín a vopred určených koeficientoch infekčnosti. Intuitívne, každý jedinec sa stretne s každým s rovnakou pravdepodobnosťou [24].

SIR model ako systém troch obyčajných diferenciálnych rovníc pre jeden variant vírusu vieme podľa [24] zapísať ako

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N} IS\,, \\ \frac{dI}{dt} &= \frac{\beta}{N} IS - \gamma I\,, \\ \frac{dR}{dt} &= \gamma I\,. \end{aligned} \tag{45}$$

Rovnice modelujú presun jedincov medzi skupinami S, I a R v čase. Parameter β predstavuje mieru infekčnosti nových jedincov (z angl. *infection rate*) a parameter γ predstavuje mieru uzdravenia (z angl. *recovery rate*). Podiel $\frac{1}{\gamma}$ vyjadruje počet časových jednotiek, ktoré jedinec strávi v skupine I. Ak sčítame všetky rovnice dokopy, dostaneme nulovú zmenu v čase. Toto je v súlade s predpokladom, že veľkosť populácie je nemenná, jedinci sa len presúvajú medzi jednotlivými skupinami.

Nasledujúce výpočty sú prevzaté zo zdroja [24]. Predpokladajme, že každý infikovaný jedinec má κ kontaktov schopných preniesť infekciu za jednotku času a κ je nezávislé od veľkosti populácie. Potom z κ kontaktov je $\kappa\frac{S}{N}$ kontaktov s jedincami náchylnými na infekciu (skupina S). Ak podiel adekvátnych kontaktov η spôsobí prenos infekcie, tak každy infikovaný jedinec infikuje $\kappa\eta\frac{S}{N}$ jedincov náchylných na infekciu za jednotku času. Preto ak definujeme

$$\xi = \frac{\beta}{N}\,, \tag{46}$$

tak $\beta = \kappa\eta$. Niekedy sa v literatúre označuje práve ξ ako β. Následne vieme definovať efektívne reprodukčné číslo R_e a bežné reprodukčné číslo R_0 ako

$$\begin{aligned} R_e &= \frac{S(0)}{N}\frac{\beta}{\gamma}\,, \\ R_0 &= \frac{\beta}{\gamma}\,. \end{aligned} \tag{47}$$

Ak celá populácia okrem jedného jedinca je v skupine S (t. j. $S(0) = N - 1\,, I(0) = 1\,, R(0) = 0$) a populácia je veľká, tak $R_e = \frac{N-1}{N}\frac{\beta}{\gamma}$ je približne rovné R_0.

Systém obyčajných diferenciálnych rovníc SIR modelu nemá explicitné riešenie, ale dá sa vyjadriť analytické riešenie pre maximálny počet infikovaných jedincov v I, ktoré označíme I^{max}. Riešenie ukážeme podľa zdroja [24]. Vydelením prvej rovnice druhou rovnicou v systéme (45) dostaneme

$$\frac{dS}{dI} = \frac{-\xi SI}{\xi SI - \gamma I}\,. \tag{48}$$

Táto obyčajná diferenciálna rovnica je separovateľná a prepíšeme ju ako

$$\int \frac{\xi S - \gamma}{\xi S} dS = -\int dI\,. \tag{49}$$

Z toho vyplýva, že bude platiť $C = -I - S + \frac{\gamma}{\xi}\ln S$. Inak povedané, $\forall t \geq 0$:

$$I(t) + S(t) - \frac{\gamma}{\xi}\ln S(t) = I(0) + S(0) - \frac{\gamma}{\xi}\ln S(0)\,. \tag{50}$$

I^{max} nastane vtedy, keď $\frac{dI}{dt} = 0$, čo podľa (54) nastáva vtedy, keď $S = \frac{\gamma}{\xi}$. Aplikovaním (50) dostaneme

$$I^{max} + \frac{\gamma}{\xi} - \frac{\gamma}{\xi}\ln\left(\frac{\gamma}{\xi}\right) = I(0) + S(0) - \frac{\gamma}{\xi}\ln S(0)\,. \tag{51}$$

Vyjadríme I^{max}:

$$I^{max} = I(0) + S(0) - \frac{\gamma}{\xi}\ln S(0) - \frac{\gamma}{\xi} + \frac{\gamma}{\xi}\ln\left(\frac{\gamma}{\xi}\right)\,. \tag{52}$$

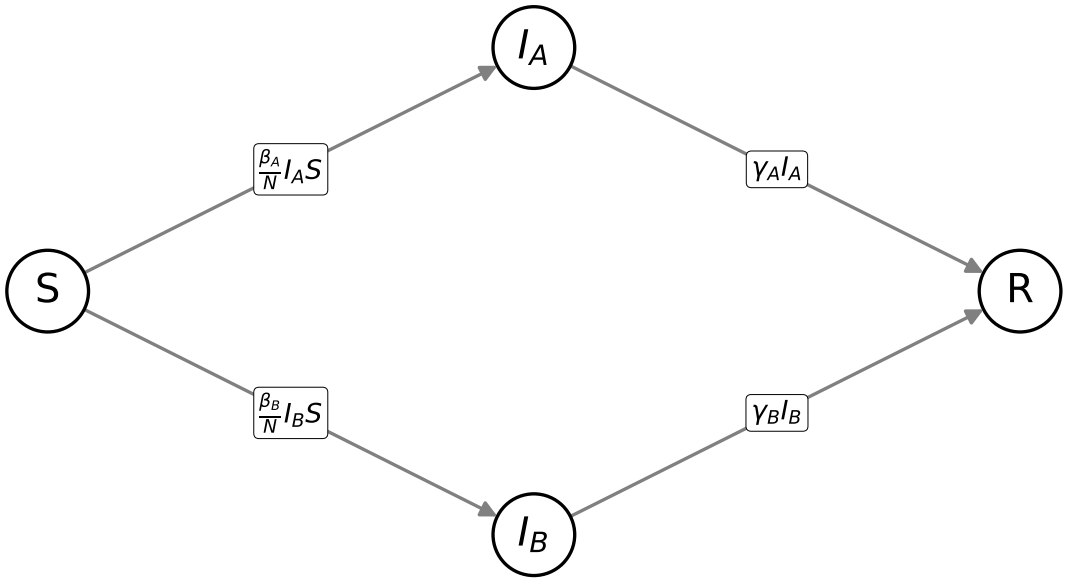
Doteraz sme písali o SIR modeli s jedným variantom vírusu. V tejto práci uvažujeme dva varianty vírusu, ktoré označíme A a B. V základnom modeli budeme predpokladať, že jedinec po prekonaní infekcie spôsobenej variantom A alebo B získa trvalú imunitu proti obom variantom. Následne definujeme SIR model pre dva varianty ako

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta_A}{N} I_A S - \frac{\beta_B}{N} I_B S\,,\\
\frac{dI_A}{dt} &= \frac{\beta_A}{N} I_A S - \gamma_A I_A\,,\\
\frac{dI_B}{dt} &= \frac{\beta_B}{N} I_B S - \gamma_B I_B\,,\\
\frac{dR}{dt} &= \gamma_A I_A + \gamma_B I_B\,,
\end{aligned} \tag{53}$$

čo sa dá ekvivalentne prepísať pomocou $\xi_A = \frac{\beta_A}{N}$ a $\xi_B = \frac{\beta_B}{N}$ podľa (46) ako

$$\begin{aligned} \frac{dS}{dt} &= -\xi_A I_A S - \xi_A I_B S\,, \\ \frac{dI_A}{dt} &= \xi_A I_A S - \gamma_A I_A\,, \\ \frac{dI_B}{dt} &= \xi_B I_B S - \gamma_B I_B\,, \\ \frac{dR}{dt} &= \gamma_A I_A + \gamma_B I_B\,. \end{aligned} \tag{54}$$

V modeli (53) skupinu I modelujeme samostatne pre oba varianty ako I_A a I_B. Každý variant má samostatnú mieru infekčnosti β_A, resp. β_B a mieru uzdravenia γ_A, resp. γ_B. Diagram systému (53) sme znázornili na Obr. 8.



Obr. 8: Diagram systému (53) s vyznačením pohybov medzi skupinami.

V nasledujúcej časti implementujeme základný SIR model pre varianty A a B v Pythone podľa systému (53) a budeme simulovať šírenie oboch variantov v populácii. Pre účely simulácií budeme uvažovať $I_B^{max} > I_A^{max}$.

4.1.1 Analýza vstupných parametrov

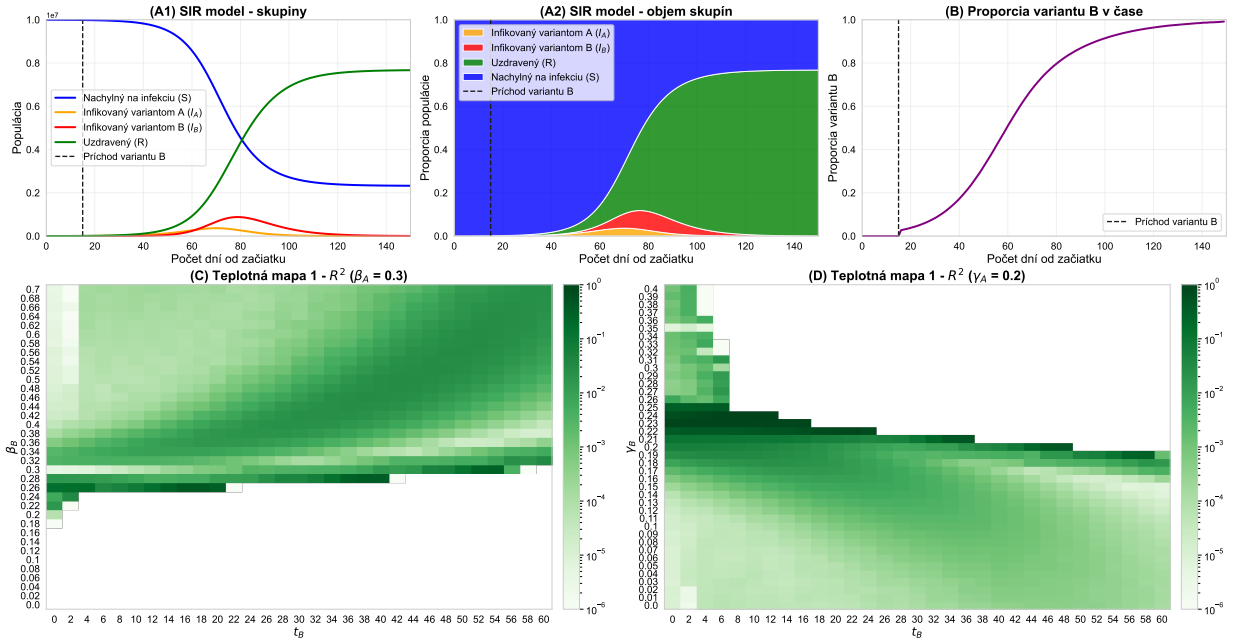
V tejto časti vykonáme simulácie pre SIR model v Pythone, pričom využijeme knižnice `numpy`, `matplotlib`, `sklearn` a `scipy`. Na rozdiel od exponenciálneho modelu nemáme pri SIR modeli garanciu, že vývoj proporcie variantu B v čase bude konzistentný s logistickou

funkciou. Toto je spôsobené obmedzenou veľkosťou populácie, ktorá nesie so sebou riziko vyčerpania populácie náchylnej na infekciu hneď prvým variantom.

Preto najprv začneme analýzou výberu parametrov $\beta_A, \beta_B, \gamma_A, \gamma_B$ a taktiež času vstupu variantu B do populácie, ktorý označíme t_B. Chceme zistiť, pre aké kombinácie parametrov je výrazne narušený predpoklad logistickej formy proporcie variantu B v populácii. Pre účely simulácie sme veľkosť populácie zafixovali na hodnotu $N = 10\ 000\ 000$ a taktiež počet prípadov variantu A a B pri vstupe do populácie (1000 pre variant A, resp. 100 pre variant B). Systém diferenciálnych rovníc (53) budeme riešiť v spojitom čase pomocou funkcie `odeint()` z knižnice `scipy`. Funkcia `odeint()` v základnom nastavení využíva Adamsovu metódu, ale automaticky prepína medzi Adamsovou metódou a BDF (implicitnou matódou na báze spätných diferencií) v prípade numericky nestabilných diferenciálnych rovníc.

Šírenie variantu A a B priamo ovplyvňujú parametre β_A a γ_A, resp. β_B a γ_B. Veľmi dôležitý je tiež parameter t_B, ktorý nepriamo ovplyvňuje schopnosť variantu B sa rozšíriť, pretože v skorší čas t bude väčšia časť populácie v skupine S. Rozhodli sme sa preto vyhodnotiť vzťah medzi β_A a β_B, resp. γ_A a γ_B od vstupu variantu B do populácie. Vykonali sme analýzu spolu so vzorovým príkladom, ktorá je viditeľná na Obr. 9.

Pre vzorový príklad sme zvolili hodnotu parametrov $\beta_A = 0.3, \beta_B = 0.35, \gamma_A = \frac{1}{5}, \gamma_B = \frac{1}{6}$. Variant B vstúpi do populácie v čase $t_B = 15$. Na grafe (A1) môžeme vidieť priebeh variantu A a B. Keďže parametre β_A, β_B neboli nastavené na vysoké hodnoty, tak po skončení pandémie ostala časť populácie ešte v skupine S. Taktiež pozorujeme, že variant B celkovo infikoval oveľa väčšiu časť populácie ako variant A. Toto nám potvrdzuje aj graf (A2), na ktorom môžeme vidieť vývoj objemu populácie v každej skupine. Ak by varianty A a B boli úplne nezávislé a obidva začínali v čase $t = 0$, tak by sme maximum krivky I_A dosiahli v čase $t = 80$ a maximum krivky I_B v čase $t = 50$. Na grafe (A1) je šírenie variantu A do istej miery zastavené variantom B a variant B sa kvôli menšej časti populácie v skupine S tiež rozšíri o niečo menej.



Obr. 9: Simulácie SIR modelu pri rôznych hodnotách jeho parametrov. (A1) Prehľad vývoja skupín S, I_A, I_B, R na konkrétnom príklade. **(A2)** Proporcie skupín S, I_A, I_B, R na konkrétnom príklade. **(B)** Ukážka proporcie variantu B v čase na konkrétnom príklade. **(C)** Výpočet konzistencie proporcie variantu B s logistickou funkciou cez $1-R^2$ pre rôzne hodnoty parametrov t_B a β_B. **(D)** Výpočet konzistencie proporcie variantu B s logistickou funkciou cez $1-R^2$ pre rôzne hodnoty parametrov t_B a γ_B.

Nakoniec na grafe (B) je vyobrazený vývoj podielu variantu B v populácii, ktorý nápadne pripomína logistickú krivku. Nie je zrejme perfektná, lebo horný koniec krivky je málo strmý na logistickú krivku. Aby sme boli konzistentní s reálnymi dátami z kapitoly 2, pri výpočte proporcie variantu B používame dennú incidenciu (počty prípadov), nie dennú prevalenciu (aktívne prípady) z kriviek I_A a I_B. Za predpokladu dostatočne jemného delenia časového rozsahu, cez ktorý riešime systém diferenciálnych rovníc, vieme spočítať okamžitú dennú incidenciu ako

$$\begin{aligned} i^* &= \underset{i\in\{0,\dots,n-1\}}{\arg\min} \ |t_i - d| \ , \\ incidencia_A(d) &= \beta_A \frac{S(t_{i^*})I_A(t_{i^*})}{N} \ , \\ incidencia_B(d) &= \beta_B \frac{S(t_{i^*})I_B(t_{i^*})}{N} \ , \end{aligned} \tag{55}$$

kde d je index konkrétneho dňa a i sú časové body, v ktorých bol vyriešený systém (53).

Ako už bolo napísané vyššie, v exponenciálnom modeli bez aplikovania oneskorení sme mali garanciu konzistencie proporcie variantu B s logistickou funkciou. Keďže pri SIR modeli túto garanciu nemáme, je potrebné kvantitatívne vyhodnotiť spomínanú konzistenciu. Keďže relatívnu fitness počítame ako smernicu nafitovanej lineárnej krivky na body z logit transformácie, vieme pomocou fitu lineárnej regresie zhodnotiť kvalitu modelu. Využijeme pritom koeficient determinácie, známky ako R^2. Koeficient determinácie sa všeobecne počíta vzorcom

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2}\,, \tag{56}$$

kde RSS je suma štvorcov rezíduí (z angl. *residual sum of squares*) a TSS je celková suma štvorcov (z angl. *total sum of squares*). Hodnota $R^2 = 1$ určuje perfektný model, pričom v našom kontexte vyhodnocovania logistiky kriviek budeme za ideálne považovať hodnoty väčšie ako 0.99.

Na grafe (C) môžeme vidieť závislosť kvality logistickej krivky na hodnotách t_B a β_B. Najtmavšie oblasti predstavujú najväčšie hodnoty $1 - R^2$, teda najmenšie hodnoty R^2. Uvedomme si, že parameter β_A bol nastavený na hodnotu 0.3. Ak bude β_B menšie ako β_A, tak bez ohľadu na hodnotu parametra t_B sa nebude vedieť variant B v populácii rozšíriť, lebo variant A bude mať prevahu. Preto máme v dolnej časti grafu (C) až na pár výnimiek prázdne polia. Ak je β_B naopak príliš veľké, tak sa zase variant A nestihne rozšíriť a variant B ho celý pohltí. Čo sa týka hodnôt t_B, ak variant B vstúpi do populácie neskoro, tak sa môže stať, že variant A stihne infikovať veľkú časť populácie a variant B bude mať obmedzený dosah. Napriek tomu pre isté hodnoty β_B môžeme pozorovať malé $1 - R^2$, ale proporcia variantu B v čase nebude mať veľkú výpovednú hodnotu kvôli absencii súboja s variantom A a celkovo veľmi malému počtu nových prípadov. Na grafe (C) môžeme celkovo pozorovať dve plochy s malým $1 - R^2$ (vľavo hore a vpravo dole), ale len prvá menovaná bude mať štatistický význam kvôli dostatočne nízkemu t_B. Všeobecne môžeme konštatovať, že mierne vyššie β_B ako β_A a včasný príchod variantu B do populácie zabezpečia logistiku proporcie variantu B konzistentnú s logistickou funkciou.

Graf (D) taktiež objasňuje kvalitu logistiky proporcie variantu B, ale pri voľnom γ_B

namiesto β_B. Pripomeňme si, hodnota parametra γ_A je rovná 0.2. Vyššie hodnoty γ_B znamenajú menšiu konkurencieschopnosť variantu B, preto môžeme očakávať inverzný vzťah s t_B ako pri grafe (C), čo sa nám aj potvrdilo. Pozorujeme, že ak $\gamma_B > \gamma_A$, tak veľmi veľa kombinácií γ_B a t_B nemá výsledky, lebo variant B sa nestihol rozšíriť kvôli rýchlemu uzdraveniu infikovanej časti populácie. Variant B by nikdy nezískal potrebnú kritickú masu na to, aby dosiahol exponenciálny rast. Hodnota γ_B by tiež realisticky nemala byť blízko nule, lebo z definície γ_B vyplýva, že $\frac{1}{\gamma_B}$ je počet dní trvania infekcie, a v tom prípade by sa hodnota zlomku blížila k nekonečnu. Tiež si všimnime prítomnosť dvoch oblastí s malým $1 - R^2$ podobne ako v grafe (C), ale len oblasť pre menšie t_B, γ_B bude relevantná. Môžeme teda konštatovať, že miera výhody variantu B nad variantom A v podobe vhodne zvoleného $\gamma_B < \gamma_A$ a vstupu variantu B do populácie dostatočne včas zabezpečí konzistentnosť proporcie variantu B s logistickou krivkou.

V nasledujúcej časti budeme po vzore exponenciálneho modelu pridávať do základného SIR modelu náhodné oneskorenie pre variant A a B a budeme skúmať jeho vplyv na inferenciu relatívnej fitness a logistiku proporcie variantu B.

4.1.2 Aplikácia oneskorenia

V tejto časti zahrnieme do základného SIR modelu náhodné oneskorenie v podobe τ_A a τ_B generovaných z normálneho a log-normálneho rozdelenia.

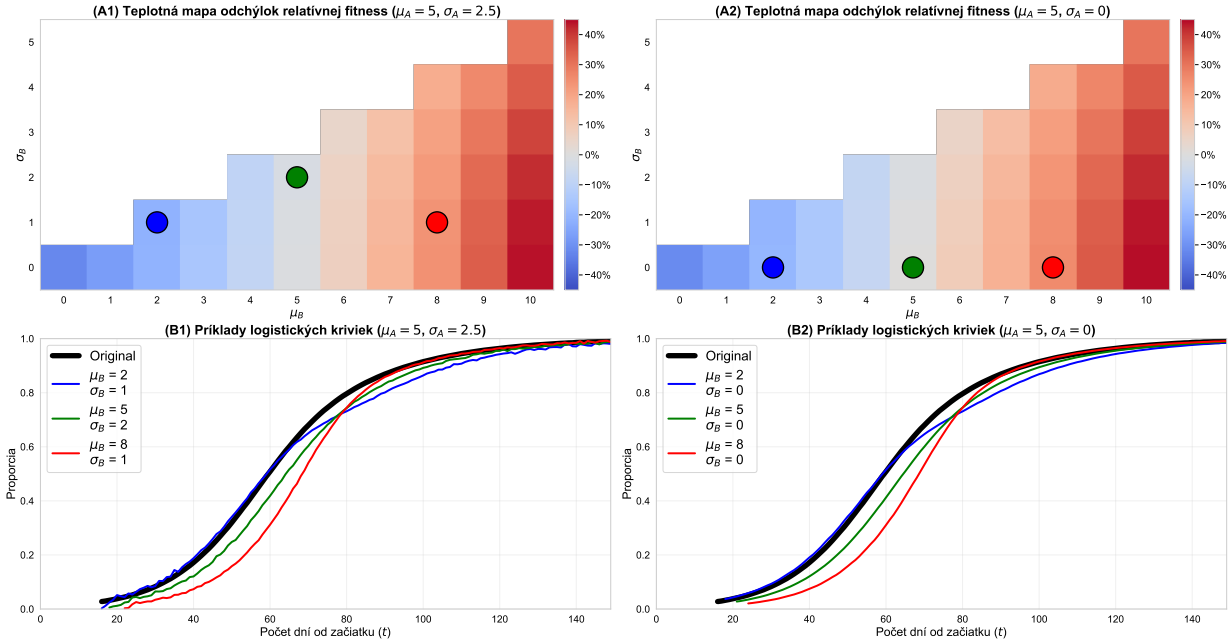
Začneme normálnym rozdelením, pričom pre variant A aj variant B budeme nezávisle generovať realizácie náhodných premenných

$$\begin{aligned} \tau_A &\sim \mathcal{N}(\mu_A, \sigma_A^2)\,, \\ \tau_B &\sim \mathcal{N}(\mu_B, \sigma_B^2)\,. \end{aligned} \tag{57}$$

SIR model sme riešili v spojitom priestore, ale náhodné oneskorenia budeme aplikovať diskrétne. Preto pri simuláciách budeme uvažovať diskrétne rozdelenie možných hodnôt τ_A a τ_B na biny o šírke 1, podobne ako pri exponenciálnom modeli. To znamená, že budeme mať denné biny rozdelené v intervale $(\mu_A - 2\sigma_A, \mu_A + 2\sigma_A)$, resp. $(\mu_B - 2\sigma_B, \mu_B + 2\sigma_B)$, čo nám umožňuje spoľahlivo pokryť skoro celú hustotu normálneho rozdelenia. Je tiež dôležité opäť poznamenať, že hodnoty krivky I_1 a I_2 nie sú počty denných prípadov, ale počty

aktívnych prípadov v daný deň. Preto ich pred aplikovaním oneskorení prekonvertujeme na denné prípady, aby sme dodržali metodiku stanovenú v časti 4.1.1.

Z časti 4.1.1 vieme, že čas, kedy variant B vstúpi do populácie, je kritický. Preto nám pri náhodných oneskoreniach vzniklo podozrenie, či na odhalenie najvýraznejších odchýlok relatívnej fitness stochasticitu vlastne potrebujeme. Preto sme najprv vykonali dve analýzy, jednu pre stochastické oneskorenie τ_A a jednu pre deterministické rozdelenie τ_A, pod ktorým budeme rozumieť iba μ_A, pričom $\sigma_A = 0$. Parametre μ_A, σ_A sme zafixovali na hodnotách $\mu_A = 5, \sigma_A = 2.5$ a voľné μ_B, σ_B a výsledky predstavili na Obr. 10.



Obr. 10: Výpočty odchýlok odhadu relatívnej fitness pre normálne rozdelené oneskorenie. (A1) Teplotná mapa odchýlok relatívnej fitness pre $\mu_A = 5, \sigma_A = 2.5$. **(A2)** Teplotná mapa odchýlok relatívnej fitness pre $\mu_A = 5, \sigma_A = 0$. **(B1)** Vybrané kombinácie z grafu (A1) a ich reprezentácia proporciou variantu B v čase. Zakreslené sú len tie kombinácie, ktoré spĺňajú predpoklad $\mu_B \geq 2\sigma_B$. **(B2)** Vybrané kombinácie z grafu (A2) a ich reprezentácia proporciou variantu B v čase. Zakreslené sú len tie kombinácie, ktoré spĺňajú predpoklad $\mu_B \geq 2\sigma_B$.

Parametre	Pôvodná rel. fitness	Rel. fitness po pridaní oneskorenia	Rozdiel
$\mu_B = 2$, $\sigma_B = 1$	0.069720	0.054753	-21.4681%
$\mu_B = 5$, $\sigma_B = 2$	0.069720	0.068356	-1.9566%
$\mu_B = 8$, $\sigma_B = 1$	0.069720	0.086274	23.7439%
$\mu_B = 2$, $\sigma_B = 0$	0.069720	0.055740	-20.0524%
$\mu_B = 5$, $\sigma_B = 0$	0.069720	0.069720	0%
$\mu_B = 8$, $\sigma_B = 0$	0.069720	0.087461	25.4459%

Tabuľka 7: Hodnoty relatívnej fitness pre rôzne režimy parametrov a ich odchýlka od pôvodnej hodnoty relatívnej fitness podľa Obr. 10.

Na grafe (A1) môžeme vidieť odchýlky odhadu priemernej relatívnej fitness od základnej hodnoty. Všimnime si, že pre hodnoty $\mu_B > \mu_A$ registrujeme zvyšujúce sa kladné odchýlky a pre hodnoty $\mu_B < \mu_A$ zvyšujúce sa záporné odchýlky. Parameter σ_B veľa na situácii nemení, ale pri väčších hodnotách μ_B a σ_B sa odchýlka zmenšuje.

Na grafe (B1) máme vybrané 3 kombinácie μ_B a σ_B z grafu (A1), pre ktoré sme vyobrazili vývoj proporcie variantu B v porovnaní s krivkou bez náhodných oneskorení. Môžeme vidieť, že modrá krivka, ktorá znázorňuje príklad zápornej odchýlky relatívnej fitness, sa v polovici viditeľne zlomí a celkový rast krivky je menej strmý. Zelená krivka, ktorá bola vybraná ako príklad zo stredu grafu (A1), neprejavuje okrem posunutia v čase dopredu výrazné zmeny oproti pôvodnej čiernej krivke. Červená krivka, ktorá znázorňuje príklad kladnej odchýlky relatívnej fitness, má zase viditeľný strmší rast, odkiaľ pramení aj vyššia odhadnutá hodnota relatívnej fitness.

Na grafoch (A2) a (B2) sme vykonali simuláciu za takých istých predpokladov, ale parametre μ_A, σ_A sme zafixovali na $\mu_A = 5, \sigma_A = 0$, teda máme konštantné oneskorenie. Taktiež sme príklady do grafu (B2) vybrali bez náhodnosti oneskorenia, teda s hodnotou $\sigma_B = 0$. Môžeme pozorovať, že odchýlok odhadov relatívnej fitness sa to veľmi nedotklo a sú vizuálne veľmi podobné ako pri nastavení $\mu_A = 5, \sigma_A = 2.5$. Taktiež pozorujeme na grafe (A2) skutočnosť, že hlavný zdroj odchýlok relatívnej fitness je spôsobený posunom hodnoty μ_B, nie σ_B. Túto skutočnosť nám numericky potvrdzujú odhadnuté hodnoty fit-

ness v Tabuľke 7. Na grafe (B2) vidíme hladší priebeh variantu B v čase, ale samotný sklon aproximácie logistickej krivky je rovnaký.

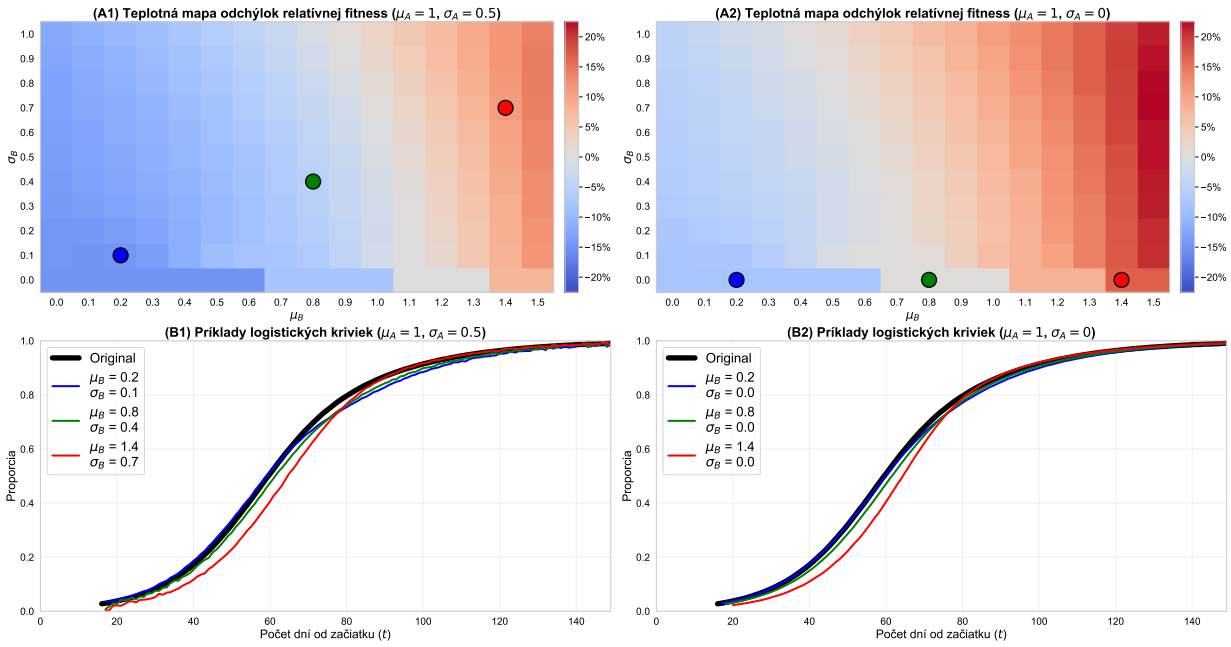
Pre potvrdenie tohto pozorovania sme danú analýzu vykonali okrem normálneho rozdelenia aj pre log-normálne rozdelenie. Pre pripomenutie, pre variant A a variant B budeme nezávisle generovať realizácie náhodných premenných

$$\begin{aligned} \tau_A &\sim Lognormal(\mu_A, \sigma_A^2)\,, \\ \tau_B &\sim Lognormal(\mu_B, \sigma_B^2)\,, \end{aligned} \tag{58}$$

ekvivalentne prepísané ako

$$\begin{aligned} \ln(\tau_A) &\sim \mathcal{N}(\mu_A, \sigma_A^2)\,, \\ \ln(\tau_B) &\sim \mathcal{N}(\mu_B, \sigma_B^2)\,. \end{aligned} \tag{59}$$

Parametre μ_A, σ_A sme zafixovali na hodnotách $\mu_A = 1, \sigma_A = 0.5$ a parametre μ_B, σ_B sme nechali voľné. Následne sme zopakovali analýzu pre upravené pole kombinácií μ_B, σ_B a výsledky sme vykreslili na Obr. 11 a zapísali do Tabuľky 8.



Obr. 11: Výpočty odchýlok odhadu relatívnej fitness pre log-normálne rozdelené oneskorenie. (A1) Teplotná mapa odchýlok relatívnej fitness pre $\mu_A = 1, \sigma_A = 0.5$. **(A2)** Teplotná mapa odchýlok relatívnej fitness pre $\mu_A = 1, \sigma_A = 0$. **(B1)** Vybrané kombinácie z grafu (A1) a ich reprezentácia proporciou variantu B v čase. **(B2)** Vybrané kombinácie z grafu (A2) a ich reprezentácia proporciou variantu B v čase.

Parametre	Pôvodná rel. fitness	Rel. fitness po pridaní oneskorenia	Rozdiel
$\mu_B = 0.2$, $\sigma_B = 0.1$	0.069720	0.059630	-14.4721%
$\mu_B = 0.8$, $\sigma_B = 0.4$	0.069720	0.066038	-5.2810%
$\mu_B = 1.4$, $\sigma_B = 0.7$	0.069720	0.077238	10.7834%
$\mu_B = 0.2$, $\sigma_B = 0$	0.069720	0.064447	-7.5630%
$\mu_B = 0.8$, $\sigma_B = 0$	0.069720	0.069720	0%
$\mu_B = 1.4$, $\sigma_B = 0$	0.069720	0.081476	16.8612%

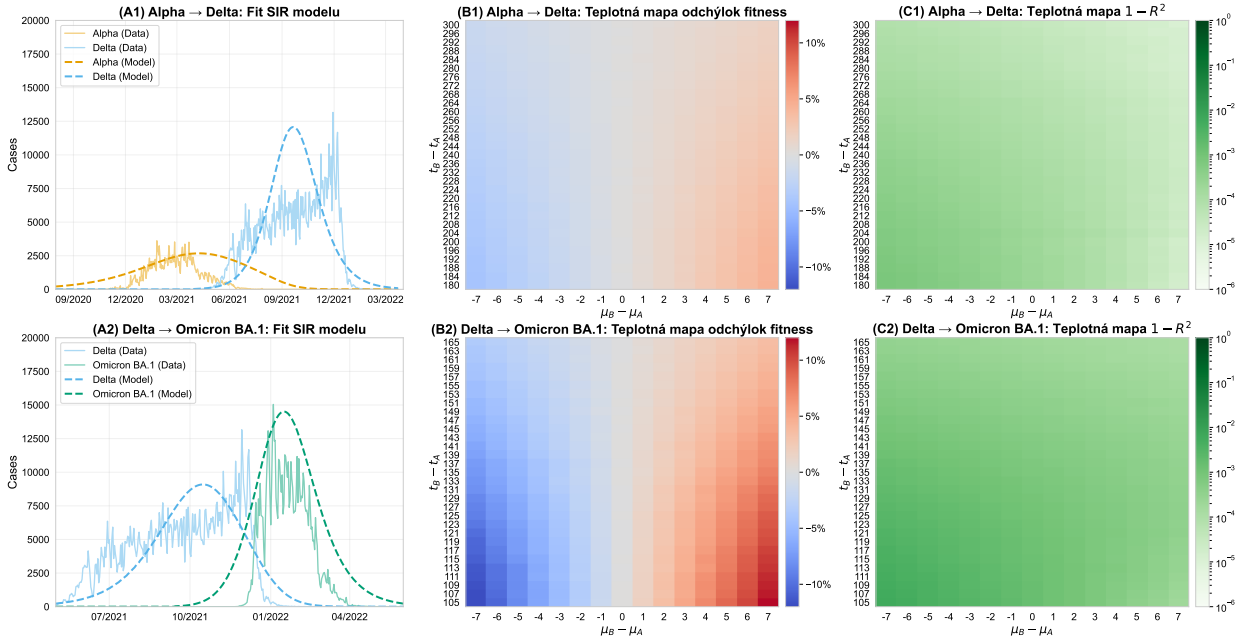
Tabuľka 8: Hodnoty relatívnej fitness pre rôzne režimy parametrov a ich odchýlka od pôvodnej hodnoty relatívnej fitness podľa Obr. 11.

Podobne ako pri Obr. 7, parametre log-normálneho rozdelenia sme vybrali také, aby realizácie náhodných premenných τ_A a τ_B boli približne v intervale $(0, 20)$ a aby boli výsledky porovnateľné s Obr. 10. Na grafoch (A1) a (A2) môžeme vidieť, že pre kombinácie menších μ_B, σ_B registrujeme záporné odchýlky relatívnej fitness. Pre kombinácie väčších μ_B, σ_B naopak registrujeme kladné odchýlky. Na grafe (A2), kde variant A má deterministické oneskorenie, pozorujeme všeobecne kladnejšie odchýlky odhadu relatívnej fitness, a v absolútnej hodnote aj trochu väčšie ako na grafe (A1) podľa Tabuľky 8. Čo sa týka grafov (B1) a (B2), sú viditeľne podobné, pričom modrá krivka je opäť príkladom odhadu so zápornou odchýlkou, zelená krivka príklad odhadu s odchýlkou blízkej nule a červená krivka príklad odhadu s kladnou odchýlkou. Aj po vybraní príkladov s $\sigma_B = 0$ v prípade grafu (B2) nepozorujeme menej extrémne odchýlky ako pri grafe (B1).

Z Obr. 10 aj Obr. 11 si vieme vyvodiť záver, že stochasticita oneskorenia daná parametrami σ_A a σ_B má na odchýlky odhadu relatívnej fitness výrazne menší vplyv, ako stredná hodnota oneskorenia daná parametrami μ_A a μ_B. V pokračovaní tejto časti aj v nasledujúcich častiach preto pre zjednodušenie interpretácie a sprehľadnenie výsledkov budeme pracovať s deterministickým, resp. konštantným oneskorením daným parametrami μ_A a μ_B.

V časti 4.1.1 sme ukázali, že parameter t_B určujúci čas vstupu variantu B do populácie má veľký vplyv na kvalitu logistiky proporcie variantu B aj na výsledný odhad relatív-

nej fitness. Ak si spomenieme na kapitolu 2, tak v nej sme venovali osobitnú pozornosť analyzovaniu vĺn variantov Alpha → Delta a Delta → Omicron BA.1. Taktiež sme zistili, že kým variant Alpha už bol na ústupe v dobe objavenia variantu Delta, tak príchod variantu Omicron BA.1 sprevádzal súboj s variantom Delta. Rozhodli sme sa tieto 2 vlny použiť ako podklad pre SIR model a vykonali sme analýzu odchýlky relatívnej fitness kvality logistických kriviek (proporcií variantu B v populácii). Výsledky je možné vidieť na Obr. 12.



Obr. 12: Výpočty odchýlok odhadu relatívnej fitness a kvality logistických kriviek s podkladom reálnych vĺn Alpha → Delta a Delta → Omicron BA.1. (A1) Reálne denné dáta z vlny Alpha → Delta v porovnaní s SIR modelom pre vhodné hodnoty parametrov. **(B1)** Teplotná mapa odchýlok odhadu relatívnej fitness pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(C1)** Teplotná mapa výpočtov $1 - R^2$ pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(A2)** Analogicky podľa (A1) pre vlnu Delta → Omicron BA.1. **(B2)** Analogicky podľa (B1) pre vlnu Delta → Omicron BA.1. **(C2)** Analogicky podľa (C1) pre vlnu Delta → Omicron BA.1.

Na grafe (A1) a (A2) môžeme vidieť dennú incidenciu variantov Alpha, Delta a Omicron BA.1. Tieto dáta sme porovnali s SIR modelom zvolením vhodných parametrov $\beta_A, \beta_A, \gamma_A, \gamma_B$. Vyobrazené krivky nie sú krivky I_A a I_B, ale ich konverzia na dennú incidenciu, aby boli porovnateľné s reálnymi dátami z kapitoly 2.

Na grafoch (B1) a (B2) následne vidíme vplyv oneskorenia na odhad relatívnej fitness. Ako sme napísali vyššie, počítame už s deterministickým oneskorením, ktoré budeme charakterizovať rozdielom medzi oneskorením variantu B (μ_B) a oneskorením variantu A (μ_A). Kladná hodnota výrazu $\mu_B - \mu_A$ bude teda znamenať väčšie oneskorenie variantu B a záporná hodnota naopak väčšie oneskorenie variantu A. Podobne budeme uvažovať aj pri vstupe variantu B do populácie (t_B), kde nás tiež bude zaujímať práve rozdiel medzi objavením variantu A a variantu B. Preto budeme pracovať s výrazom $t_B - t_A$, ktorý považujeme za kladný.

Pri vlne Delta → Omicron BA.1 registrujeme všeobecne väčšie odchýlky relatívnej fitness ako pri vlne Alpha → Delta. Toto je spôsobené uvažovaním dlhšieho časového okna, pričom mediánová hodnota uvažovaného vstupu druhého variantu do populácie je 240 dní po prvom variante v prípade vlny Alpha → Delta oproti len 135 dňom v prípade vlny Delta → Omicron BA.1. Ak na obe vlny aplikujeme rovnaké oneskorenie $t_B - t_A$, je zrejmé, že pri vlne Delta → Omicron BA.1 to spôsobí väčší relatívny odklon od základnej hodnoty, pretože prvý aj druhý variant sú v čase od seba menej vzdialené. Všimnime si, že väčší odstup variantov od seba (väčšia hodnota $t_B - t_A$) má za následok pokles odchýlky relatívnej fitness. Taktiež je vhodné spomenúť, že keby sme počítali aj s parametrami σ_A, σ_B, tak by sme kvôli veľkým hodnotám $t_B - t_A$ nevideli ich vplyv. Keďže vplyv $\mu_B - \mu_A$ stráca na dôležitosti s rastúcim $t_B - t_A$, tak vplyv σ_A, σ_B by bol už úplne zanedbateľný.

Na grafoch (C1) a (C2) pozorujeme kvalitu logistických kriviek daných proporciou variantu B v čase pre každú kombináciu $\mu_B - \mu_A$ a $t_B - t_A$ z grafov (B1) a (B2). Keďže uvažujeme výraz $1 - R^2$, chceme mať hodnoty čo najbližšie 0. Môžeme konštatovať, že na oboch grafoch nevidíme výrazné odchýlky od perfektnej logistickej krivky. Najnižšie hodnoty $1 - R^2$, ktoré sú ale stále dostatočne vysoké, môžeme pozorovať v ľavej dolnej časti grafov, teda v oblasti najväčších záporných odchýlok relatívnej fitness.

Pre ďalšie simulácie v tejto aj nasledujúcich častiach budeme uvažovať dva rovnaké režimy šírenia variantu B, a to v čase poklesu variantu A (neskorý príchod) a v čase maxima variantu A (skorý príchod), modelované po vzore vĺn Alpha → Delta a Delta → Omicron

BA.1. Vytvorili sme také dve kombinácie parametrov, ktoré budú vhodné pre uvažované dva režimy a zapísali sme ich do Tabuľky 9.

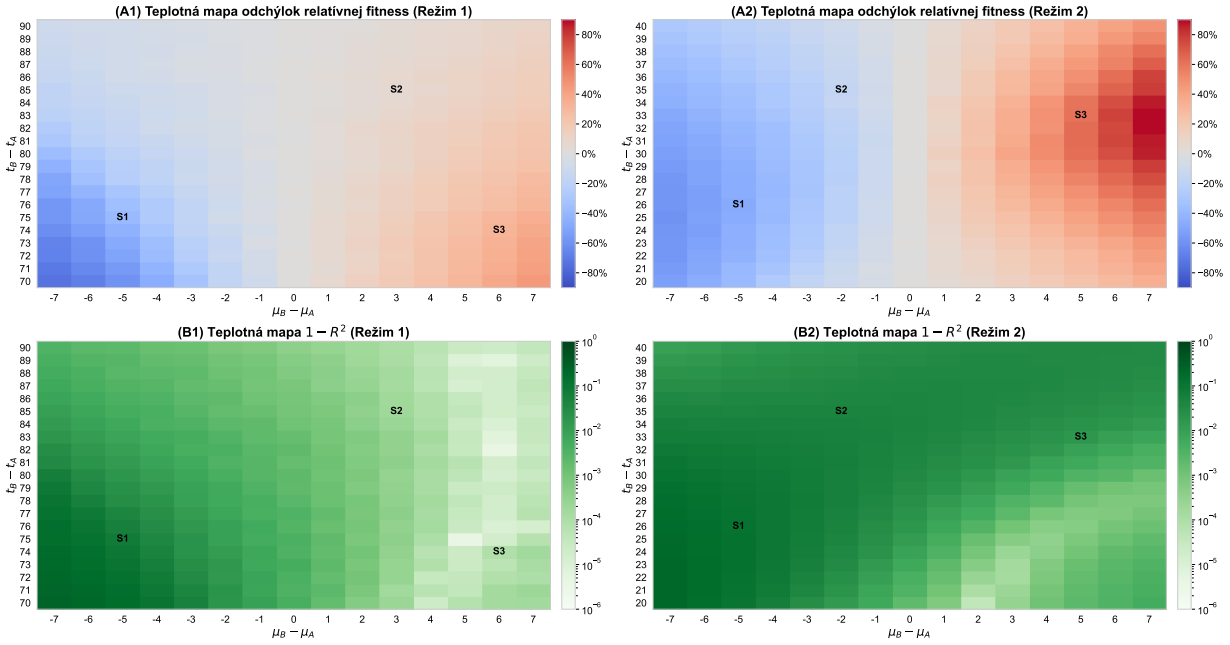
Parametre	Neskorý príchod variantu B (režim 1)	Skorý príchod variantu B (režim 2)
β_A	0.3	0.3
β_B	1.0	0.45
γ_A	$\frac{1}{5}$	$\frac{1}{5}$
γ_B	$\frac{1}{5}$	$\frac{1}{5}$

Tabuľka 9: Výber parametrov pre simulácie SIR modelov.

Všimnime si, že pri neskorom príchode variantu B uvažujeme oveľa vyššiu hodnotu β_B, teda priamo väčšiu infekčnosť variantu B. Na prvý pohľad to môže pôsobiť neintuitívne, ale treba si uvedomiť, že variant B sa v populácii objaví až po tom, ako mal variant A čas infikovať veľkú časť populácie, ktorá je následne už imúnna. Preto je potrebné nastaviť β_B na vysokú hodnotu, aby variant B mal možnosť ešte infikovať čo najväčšiu časť zostávajúcej neimúnnej časti populácie. Pri skorom príchode variantu B nám stačí aj menšia hodnota β_B, lebo vďaka predpokladu $\beta_B > \beta_A$ variant B preváži nad variantom A.

Následne sme vykonali simulácie opäť pre rôzne hodnoty výrazov $\mu_B - \mu_A$ a $t_B - t_A$ podľa Obr. 12, pričom sa zaujímame o odchýlky v odhade relatívnej fitness a schopnosť interpretovať proporciu variantu B logistickou krivkou. Vizualizáciu je možné vidieť na Obr. 13.

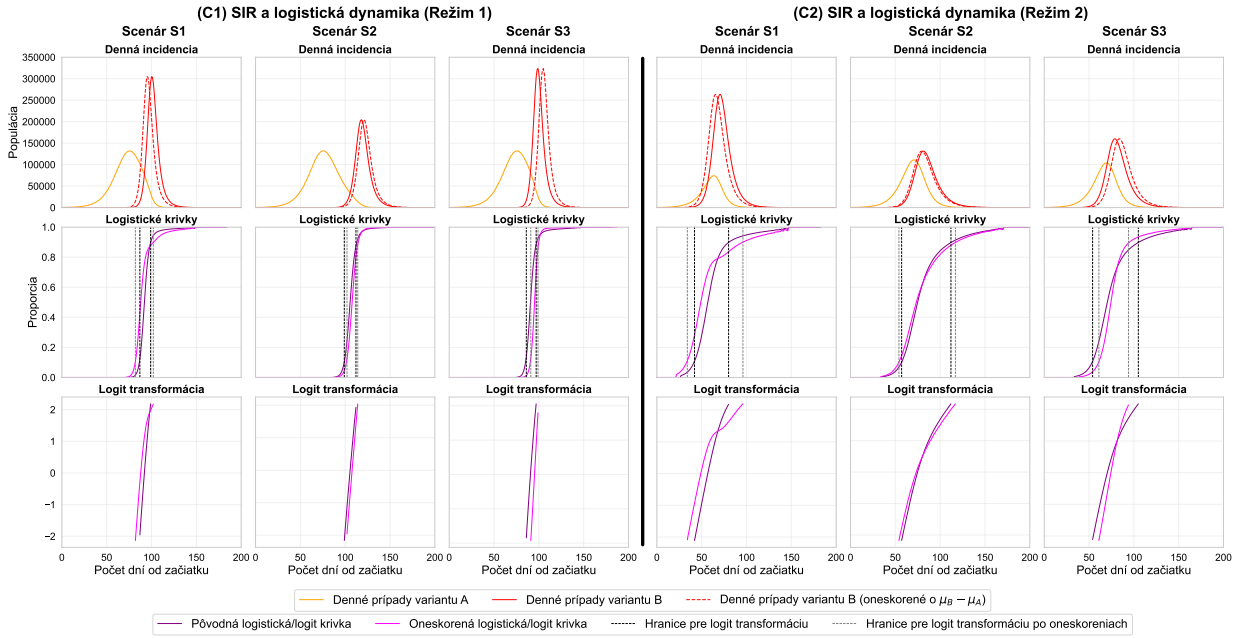
Grafy (A1) a (B1) zodpovedajú režimu uvažujúcemu neskorý príchod variantu B, grafy (A2) a (B2) zodpovedajú režimu uvažujúcemu skorý príchod variantu B. Na grafoch (A1) a (A2) môžeme pozorovať podobnú situáciu ako na Obr. 12, teda záporné odchýlky odhadu relatívnej fitness sú spôsobené zápornou hodnotou $\mu_B - \mu_A$, kým kladné odchýlky sú spôsobené kladnou hodnotou $\mu_B - \mu_A$. Taktiež sa opakuje jav, kedy pri rastúcich hodnotách $t_B - t_A$ sa zmenšuje odchýlka relatívnej fitness.



Obr. 13: Výpočty odchýlok odhadu relatívnej fitness a kvality logistických kriviek pre dva režimy parametrov podľa modelu (53). (A1) Teplotná mapa odchýlok odhadu relatívnej fitness pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(B1)** Teplotná mapa výpočtov $1 - R^2$ pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(A2)** Analogicky podľa (A1) pre druhý režim. **(B2)** Analogicky podľa (B1) pre druhý režim.

Na grafe (B1) pozorujeme najhoršiu kvalitu logistických kriviek pri najzápornejších hodnotách $\mu_B - \mu_A$ podobne ako na Obr. 12. Na grafe (B2) kvôli výrazne väčšiemu súboju variantu B s variantom A už pozorujeme na viacerých miestach horšiu logistiku proporcie variantu B, ale opäť najviac pri záporných odchýlkach relatívnej fitness.

Pre lepšie pochopenie situácie sme si označili viacero kombinácií $\mu_B - \mu_A$ a $t_B - t_A$ písmenami S1, S2 a S3. Scenár S1 predstavuje výraznú zápornú odchýlku relatívnej fitness, scenár S2 malú odchýlku a scenár S3 výraznú kladnú odchýlku. Tieto tri kombinácie sme následne podrobnejšie preskúmali a vizualizovali na Obr. 14. Konkrétne odchýlky odhadu fitness a hodnoty R^2 pre každý scenár sme zaznačili do Tabuľky.



Obr. 14: Vybrané príklady kombinácií výrazov $\mu_B - \mu_A$ a $t_B - t_A$ pre oba režimy podľa Obr. 13. (C1) Prehľad dennej incidencie, proporcie variantu B a jej logit transformácia pre tri príklady z grafov (A1) a (B1). **(C2)** Analogicky podľa grafov (A2) a (B2).

Scenár	Odchýlka rel. fitness	Hodnota R^2
S1 (Režim 1)	-38.3478%	0.925404
S2 (Režim 1)	5.4628%	0.999807
S3 (Režim 1)	31.1417%	0.999910
S1 (Režim 2)	-47.3383%	0.881218
S2 (Režim 2)	-15.6606%	0.954236
S3 (Režim 2)	60.7153%	0.984087

Tabuľka 10: Odchýlky relatívnej fitness a hodnoty R^2 pre scenáre z Obr. 13.

Graf (C1) je pokračovanie grafov (A1) a (B1) z Obr. 13, graf (C2) je pokračovanie grafov (A2) a (B2). V prvom riadku môžeme vidieť krivku reprezentujúcu denný počet prípadov pre variant A aj variant B. Časové oneskorenie $\mu_B - \mu_A$ sme pre jednoduchosť znázornili ako posun denných prípadov variantu B. Môžeme vidieť, že kým pre prvý režim máme krivky vizuálne konzistentné s logistickou krivkou v druhom riadku a s lineárnou krivkou v treťom riadku, pre druhý režim pozorujeme odchýlky od perfektnej logistiky proporcie variantu B. Toto už bolo čiastočne naznačené grafom (B2) na Obr. 13 a najvýraznejšiu

odchýlku pozorujeme u scenára S1. Všimnime si tiež „zahnutie“ krivky po logit transformácii pre druhý režim. Vyzerá to tak, že sa láme priamo v inflexnom bode logistickej krivky nad ňou, a pomalší rast logistickej krivky za inflexným bodom má za následok väčšie zahnutie krivky po logit transformácii.

V nasledujúcich častiach budeme podobným spôsobom skúmať určité modifikácie SIR modelu s využitím deterministického oneskorenia a porovnáme výsledky s cieľom identifikovať režim, pri ktorom máme výrazne menšiu konzistenciu proporcie variantu B s logistickou funkciou.

4.2 Model s dočasnou imunitou

V tejto časti sa pozrieme na prvú z modifikácií SIR modelu. Ako bolo napísané v časti 4.1 v modeli (53), jedinec infikovaný hocijakým variantom vírusu získa po jeho prekonaní imunitu voči všetkým variantom. Tento jav sa nazýva krížová imunita (z angl. *cross-immunity*).

Existuje veľmi veľa modelov, v ktorých sa vyskytujú iné skupiny ako S, I a R. My sa budeme konkrétne venovať zmenám v krížovej imunite. Sú tri základné spôsoby, ako zásadne zmeniť základný SIR model (53) len upravením nastavenia krížovej imunity:

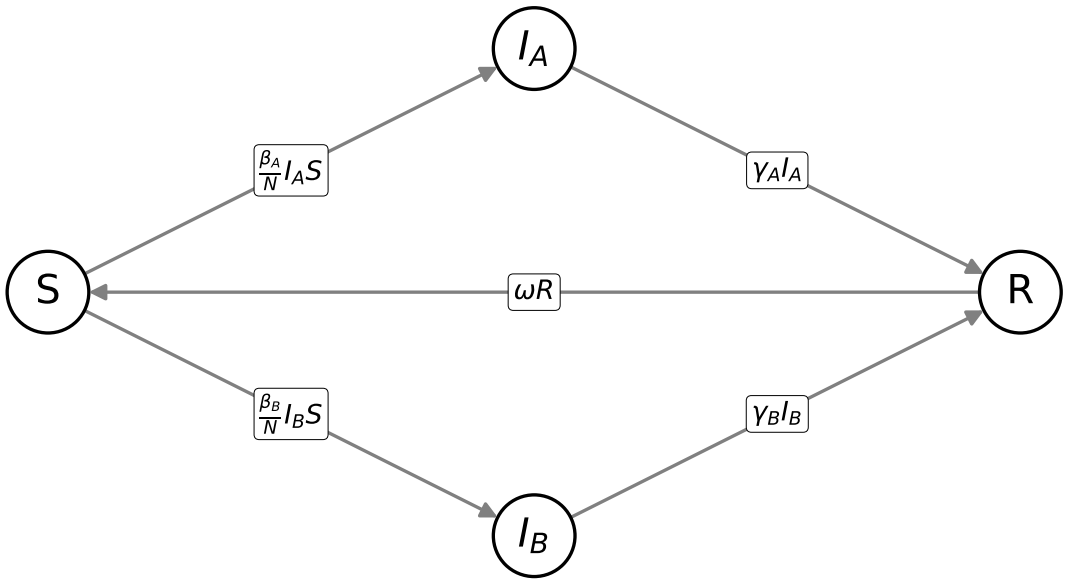
- Dočasná imunita: získaná imunita je úplná, ale trvá iba určitú dobu
- Neúplná imunita: získaná imunita je len voči variantu, ktorého infekcia bola prekonaná
- Žiadna imunita: jedinec nenadobudne imunitu po prekonaní žiadneho variantu

Posledný spôsob sa dá všeobecne zapísať ako SIS model (bez skupiny R) a nebudeme sa ním v tejto práci zaoberať. Prvé dva spôsoby rozoberieme v tejto a nasledujúcej časti.

Prvá modifikácia SIR modelu, ktorej sa budeme venovať, sa nazýva SIR model s dočasnou alebo miznúcou imunitou (z angl. *waning immunity*). Literatúra tento model označuje aj ako SIRS model [9]. Uvažujeme dva varianty vírusu, A a B. Ak jedinec prekoná infekciu

variantom A alebo B, získa perfektnú imunitu voči obom variantom na dobu T dní. Parameter, ktorý túto imunitu bude určovať, označíme $\omega \coloneqq \frac{1}{T}$. Následne vieme prepísať model (53) na model (60). Jeho diagram je možné vidieť na Obr. 15.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta_A}{N} I_A S - \frac{\beta_B}{N} I_B S + \omega R\,, \\
\frac{dI_A}{dt} &= \frac{\beta_A}{N} I_A S - \gamma_A I_A\,, \\
\frac{dI_B}{dt} &= \frac{\beta_B}{N} I_B S - \gamma_B I_B\,, \\
\frac{dR}{dt} &= \gamma_A I_A + \gamma_B I_B - \omega R\,,
\end{aligned} \tag{60}$$

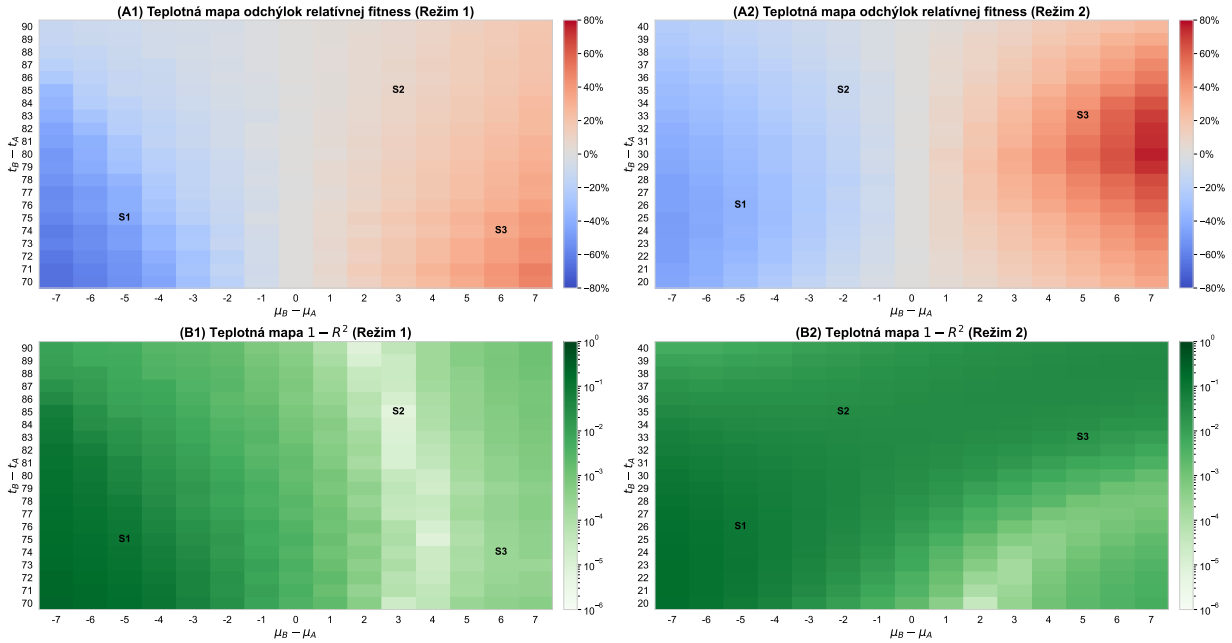


Obr. 15: Diagram systému (60) s vyznačením pohybov medzi skupinami.

Všimnime si, že sa v modeli zmenili predpisy $\frac{dS}{dt}$ a $\frac{dS}{dt}$. Zo skupiny R ubúdajú jedinci rýchlosťou danou parametrom ω a rovnakou rýchlosťou pribúdajú do skupiny S, ide teda len o presun časti jedincov zo skupiny R do skupiny S. Predpoklad o nemennosti veľkosti populácie zostáva zachovaný. Vieme povedať, že ak $\gamma_A I_A + \gamma_B I_B < \omega R$, tak počet jedincov v skupine R bude klesať, čo v základnom modeli (53) nie je možné. Rovnako za platnosti $-\frac{\beta_A}{N} I_A S - \frac{\beta_B}{N} I_B S < \omega R$ môžeme pozorovať zvyšovanie počtu jedincov v skupine S.

Následne si kladieme otázku, či zahrnutie miznúcej imunity do SIR modelu bude mať zásadný vplyv na odhad relatívnej fitness a či budeme vedieť pozorovať výrazné zmeny

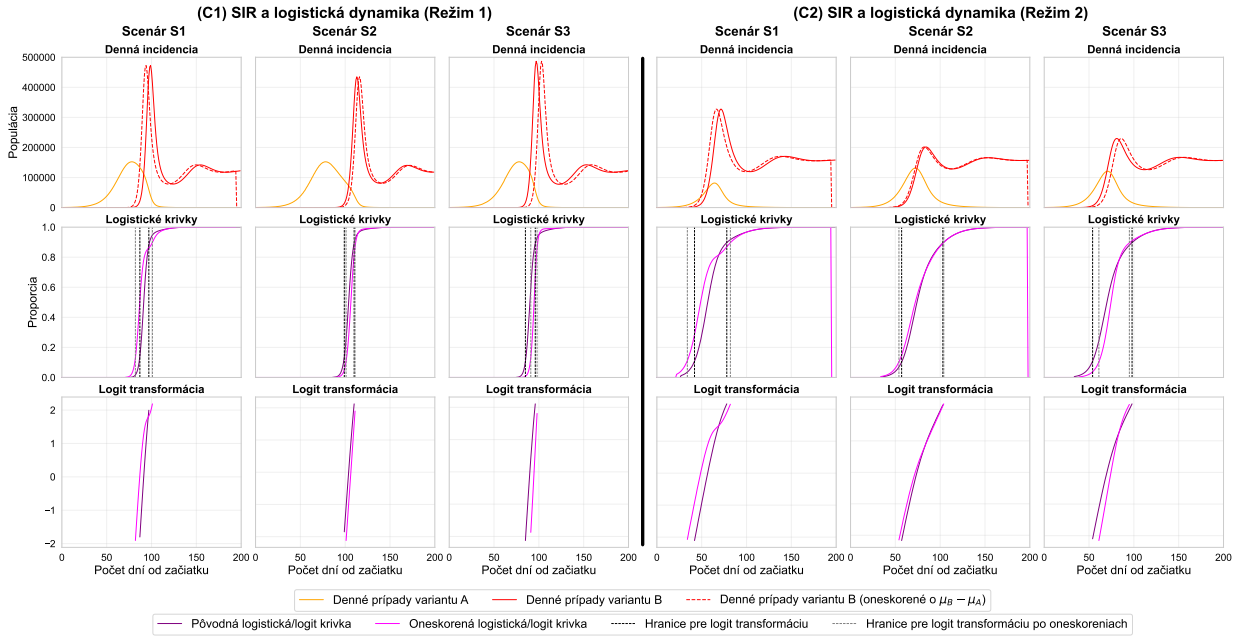
oproti základnému SIR modelu. Zopakovali sme preto simulácie vykonané v časti 4.1 na základe výberu parametrov z Tabuľky 9 a výsledky analýzy vizualizovali na Obr. 16.



Obr. 16: Výpočty odchýlok odhadu relatívnej fitness a kvality logistických kriviek pre dva režimy parametrov podľa modelu (60). (A1) Teplotná mapa odchýlok odhadu relatívnej fitness pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(B1)** Teplotná mapa výpočtov $1 - R^2$ pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(A2)** Analogicky podľa (A1) pre druhý režim. **(B2)** Analogicky podľa (B1) pre druhý režim.

Hodnotu parametra ω sme nastavili na $\frac{1}{90}$, teda predpokladáme 90-dňovú imunitu. V porovnaní s Obr. 13 nevidíme veľa zmien, grafy (A1) a (A2) vyzerajú vizuálne rovnako. Na grafe (B1) vidíme mierne zmeny, ale iba v pásme najlepších výsledkov, teda hodnôt s $1 - R^2 < 10^{-3}$. Na grafoch (B1) aj (B2) opäť pozorujeme najhoršie výsledky pre najzápornejšie odchýlky v odhade relatívnej fitness.

Keďže máme veľmi podobné výsledky ako pri základnom SIR modeli, rozhodli sme sa využiť tie isté tri scenáre ako na Obr. 14 a vykreslili výsledky simulácií na Obr. 17 nižšie. Taktiež sme zaznačili odchýlky odhadu fitness a hodnoty R^2 pre scenáre do Tabuľky 11.



Obr. 17: Vybrané príklady kombinácií výrazov $\mu_B - \mu_A$ a $t_B - t_A$ pre oba režimy podľa Obr. 16. (C1) Prehľad dennej incidencie, proporcie variantu B a jej logit transformácia pre tri príklady z grafov (A1) a (B1). **(C2)** Analogicky podľa grafov (A2) a (B2).

Scenár	Odchýlka rel. fitness	Hodnota R^2
S1 (Režim 1)	-43.049%	0.905000
S2 (Režim 1)	8.1563%	0.999994
S3 (Režim 1)	33.1650%	0.999777
S1 (Režim 2)	-37.9337%	0.912283
S2 (Režim 2)	-11.5161%	0.971348
S3 (Režim 2)	47.2391%	0.982671

Tabuľka 11: Odchýlky relatívnej fitness a hodnoty R^2 pre scenáre z Obr. 16.

Všimnime si, že počet denných prípadov po určitom čase začne znova rásť. Toto je v súlade s definíciou modelu (60), kedy odliv jedincov zo skupiny R prevýši ich príliv a sú opäť náchylní na infekciu. Mimo tohto pozorovania sa veľa vecí nemení. Keby sme nastavili parameter ω na nižšiu hodnotu (napr. 60), evidovali by sme na Obr. 17 po prvej hlavnej vlne infekcií viacero menších vĺn, ktoré by sa časom stabilizovali. Tento fenomén pozorujeme aj pri $\omega = \frac{1}{90}$ (hlavne pri prvom režime), ale v menšej miere. Môžeme teda usúdiť, že inklúzia parametra ω neprehĺbi odchýlky v odhade relatívnej fitness a tiež nezhorší našu

schopnosť intepretovať proporciu variantu B v čase ako logistickú krivku.

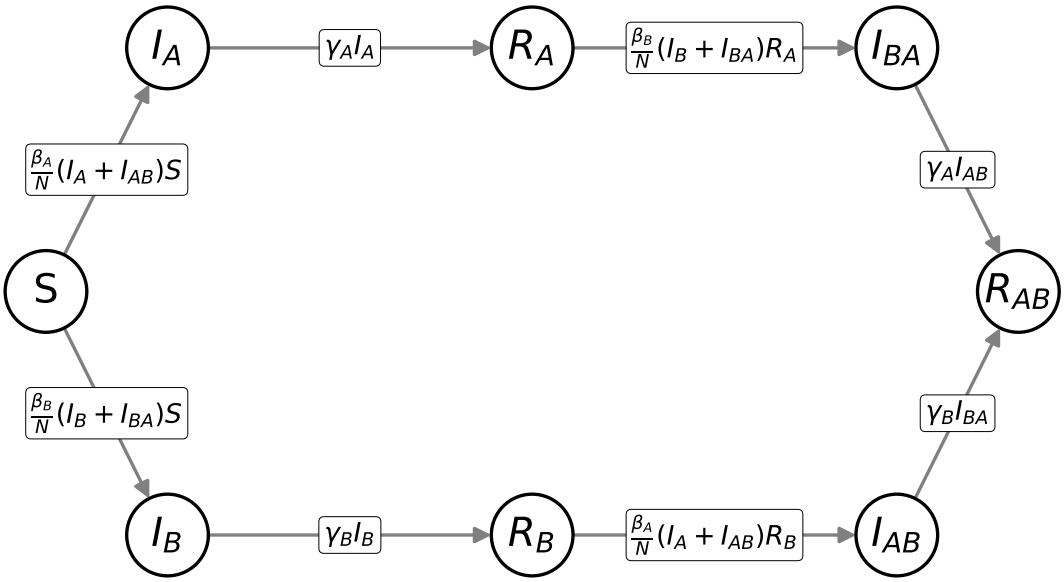
V nasledujúcej časti sa budeme zaoberať druhým z troch prístupov opísaných na začiatku tejto časti, a to neúplnou imunitou, teda absenciou krížovej imunity.

4.3 Model bez krížovej imunity

V tejto časti budeme rozoberať druhý spôsob modifikácie SIR modelu, a to odstránením krížovej imunity medzi variantmi. Opäť uvažujme dva varianty vírusu A a B. Ak sa jedinec infikuje variantom A, po jeho prekonaní nezíska imunitu voči obom variantom, ale len voči variantu A. Bude stále plne náchylný na infekciu variantom B. Pri prvotnej infekcii variantom B bude analogicky mať po prekonaní imunitu len voči variantu B. Ak prekoná infekcie variantom A aj B, stane sa imúnny voči obom variantom, teda efektívne bude v skupine R.

Keďže budeme musieť modelovať prechod jedincov aj v stave, kedy majú imunitu len proti jednému variantu, budeme potrebovať viac diferenciálnych rovníc. Základný SIR model (53) prepíšeme ako model (61) na základe zdroja [16]. Diagram tohto modelu sme vizualizovali na Obr. 18.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta_A}{N}(I_A + I_{AB})S - \frac{\beta_B}{N}(I_B + I_{BA})S\,,\\
\frac{dI_A}{dt} &= \frac{\beta_A}{N}(I_A + I_{AB})S - \gamma_A I_A\,,\\
\frac{dI_B}{dt} &= \frac{\beta_B}{N}(I_B + I_{BA})S - \gamma_B I_B\,,\\
\frac{dR_A}{dt} &= \gamma_A I_A - \frac{\beta_B}{N}(I_B + I_{BA})R_A\,,\\
\frac{dR_B}{dt} &= \gamma_B I_B - \frac{\beta_A}{N}(I_A + I_{AB})R_B\,,\\
\frac{dI_{BA}}{dt} &= \frac{\beta_B}{N}(I_B + I_{BA})R_A - \gamma_B I_{BA}\,,\\
\frac{dI_{AB}}{dt} &= \frac{\beta_A}{N}(I_A + I_{AB})R_B - \gamma_A I_{AB}\,,\\
\frac{dR_{AB}}{dt} &= \gamma_A I_{AB} + \gamma_B I_{BA}\,,
\end{aligned} \tag{61}$$



Obr. 18: Diagram systému (61) s vyznačením pohybov medzi skupinami.

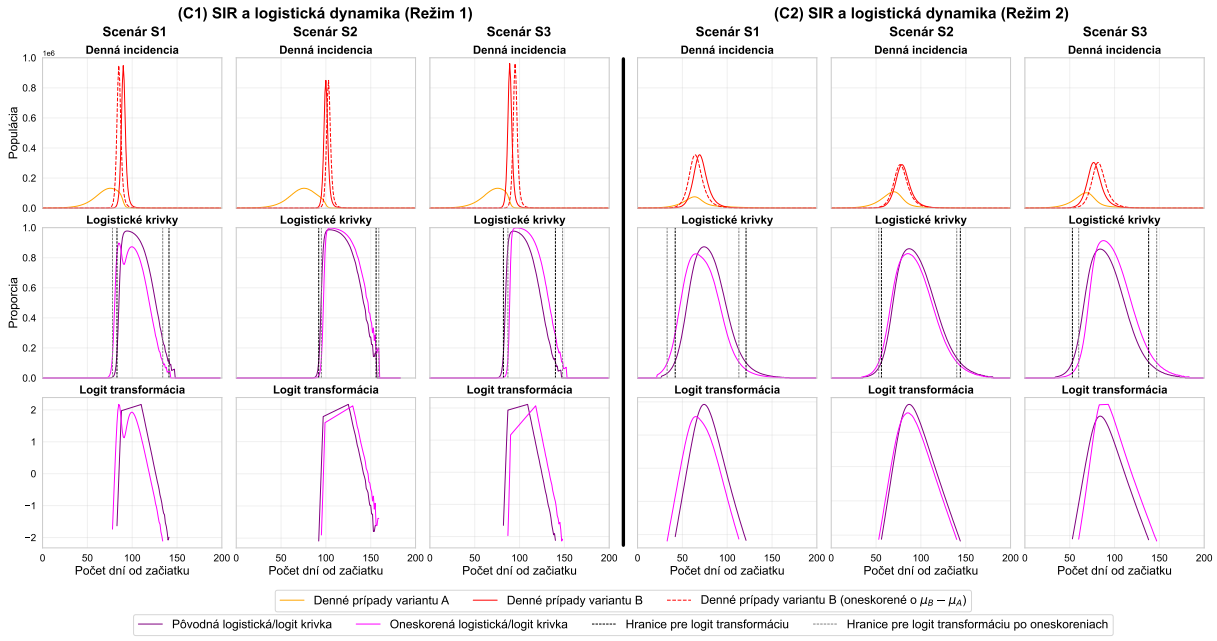
Všimnime si, že máme oproti základnému modelu o 4 rovnice naviac. Je to preto, lebo potrebujeme vedieť aj o jedincoch, ktorí prekonali len variant A alebo len variant B, a taktiež aj o jedincoch, ktorí sa druhýkrát infikovali buď variantom A alebo variantom B. Zaviedli sme preto nové označenia skupín:

- R_A: jedinec prekonal variant A, ale je náchylný na infekciu variantom B
- R_B: jedinec prekonal variant B, ale je náchylný na infekciu variantom A
- I_{BA}: jedinec je infikovaný variantom B potom, ako bol v minulosti infikovaný variantom A
- I_{AB}: jedinec je infikovaný variantom A potom, ako bol v minulosti infikovaný variantom B

Výraz $\frac{dR_{AB}}{dt}$ je epidemiologicky ekvivalentný výrazu $\frac{dR}{dt}$ z modelu (53), lebo signalizuje úplnú imunitu voči obom variantom, aj keď jedinec prekonal dve infekcie namiesto jednej. Taktiež je potrebné zohľadniť reformuláciu rovníc, v ktorých sa vyskytuje I_A alebo I_B v kontexte šírenia infekcie (nie uzdravenia sa). Keď chceme modelovať šírenie variantu A, nesmieme zabudnúť na to, že aj jedinci najprv infikovaní variantom B a následne variantom A sú rovnako infekční a schopní infikovať iných jedincov variantom A. Preto sme

v modeli (61) prepísali I_A ako $(I_A + I_{AB})$ a I_B ako $(I_B + I_{BA})$ všade, kde sa vyskytuje parameter β_A alebo β_B.

Následne prejdeme k simuláciám, ale ako neskôr ukážeme, použitie tých istých hodnôt parametrov uvedených v Tabuľke 9 nezabezpečí konzistenciu proporcie variantu B s logistickou funkciou. Napriek tomu pre referenciu ukážeme, ako by výsledky simulácií vyzerali s použitím hodnôt z Tabuľky 9. Simulácie s teplotnými mapami nateraz preskočíme a rovno vykreslíme tri príklady na Obr. 19, pričom použijeme rovnaké kombinácie výrazov $\mu_B - \mu_A$ a $t_B - t_A$.



Obr. 19: Testovacie príklady kombinácií výrazov $\mu_B - \mu_A$ a $t_B - t_A$ pre oba režimy podľa modelu (61). (C1) Prehľad dennej incidencie, proporcie variantu B a jej logit transformácia pre tri príklady z grafov (A1) a (B1). **(C2)** Analogicky podľa grafov (A2) a (B2).

Ako môžeme vidieť, žiaden z príkladov pre oba režimy vstupu variantu B nedovoľuje modelovať proporciu variantu B ako logistickú krivku. Všimnime si, že pri prvom režime má variant B výrazne vyšší počet prípadov vďaka vysokej β_B. Pri druhom režime nevidíme obrovský rozdiel medzi dynamikou variantu A a B, ale zdanlivo sa variant A udrží v populácii pridlho, čo následne spôsobí jeho zvýšenú proporciu a eliminuje konzistenciu proporcie variantu B s logistickou funkciou.

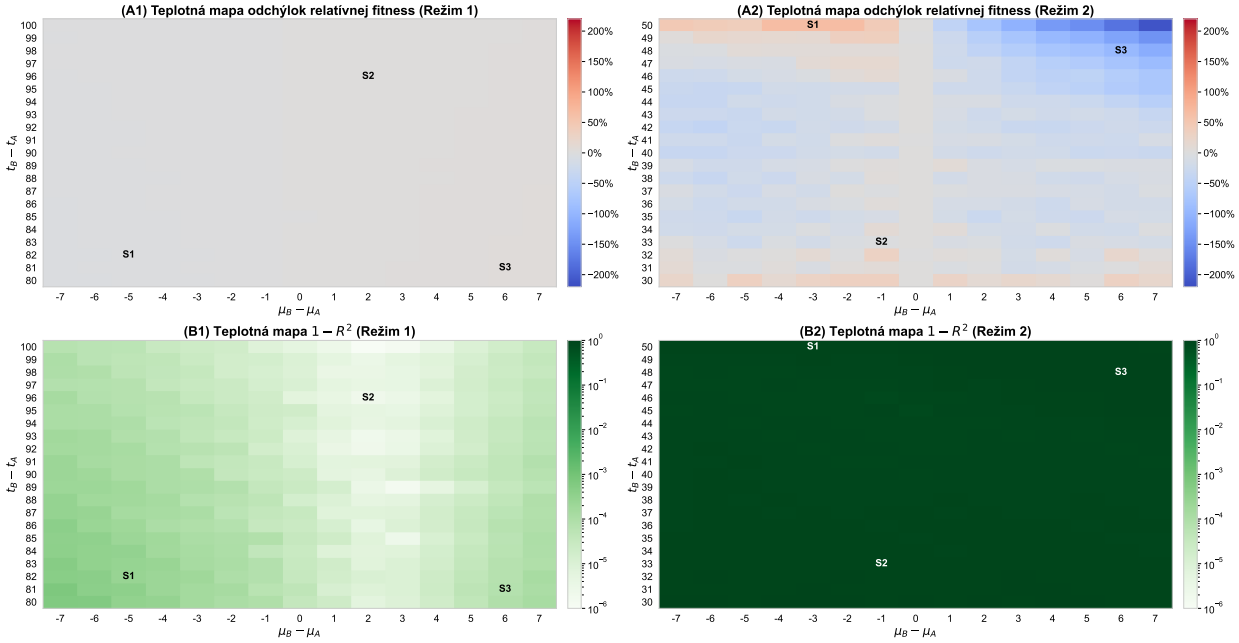
Odpoveď na tento problém sa nachádza vo formulácii systému (61). Uvedomme si, že každý jedinec sa môže infikovať variantom A aj variantom B. To znamená, že po infikovaní jedným variantom sa nestanú imúnnymi, ale z pohľadu druhého variantu sa vrátia do skupiny S. To znamená, že každý variant má k dispozícii celú skupinu S a môže sa šíriť prakticky nezávisle od ostatných variantov. Jediná výnimka je vtedy, keď je daný jedinec infikovaný iným variantom, pretože neuvažujeme možnosť infikovania oboma variantmi naraz. Taktiež sa vieme na danú situáciu pozrieť z pohľadu udomácneného variantu A. Tým, že sa variant B objaví v populácii, tak na čas kvôli vyššiemu β_B výrazne zredukuje počet novo infikovaných jedincov variantom A. Keď sa ale z variantu B uzdravia, tak variant A sa opäť môže šíriť ďalej. Prítomnosť variantu B v istom zmysle predĺži životnosť variantu A a preto nevidíme pri proporcii variantu B konzistenciu s logistickou funkciou.

Je teda zrejmé, že veľmi vysoká β_B spôsobuje rýchly nástup a následne pokles variantu B, pretože oveľa rýchlejšie infikuje populáciu ako variant A. Ak je medzi vstupom variantu A a B do populácie malý rozdiel (t. j. máme malé $t_B - t_A$), tak sa variant A môže udržať v populácii dlhšie ako variant B. Ak chceme mať konzistenciu proporcie variantu B s logistickou krivkou, musíme upraviť nasledujúce dve veci:

- Pomalšie šírenie variantu B: Ak zmenšíme parameter β_B, tak sa variant B bude pomalšie šíriť, čím sa udrží v populácii dlhšie nezávisle od variantu A;
- Neskorší nástup variantu B: Ak zväčšíme výraz $t_B - t_A$, tak variant A bude mať dostatočný čas na rozšírenie a zníži sa riziko toho, že na konci bude proporčne silnejší ako variant B.

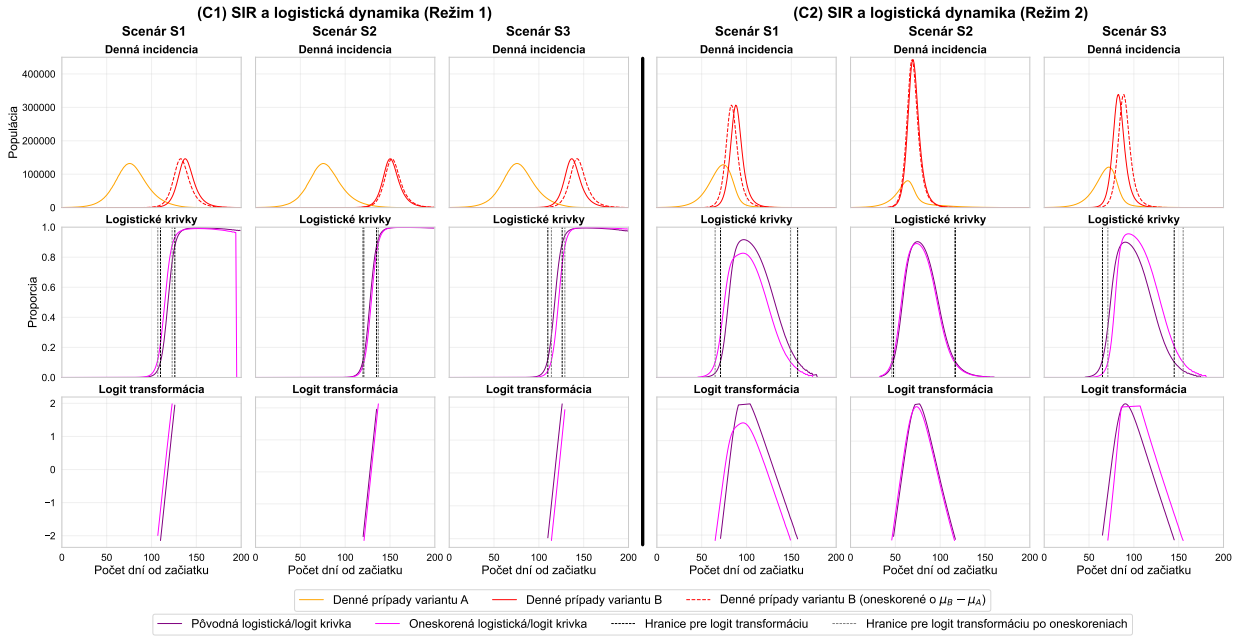
Následne sa vieme vrátiť k simuláciám a vykonať zmeny v parametroch s cieľom dosiahnutia realistickejších výsledkov. V prvom režime nastavíme $\beta_B = 0.4$, aby sa variant B nerozšíril príliš rýchlo. Nedostatočnej časti populácie v skupine S pre infekciu variantom B sa nemusíme obávať práve vďaka absencii krížovej imunity, preto môžeme voliť β_B relatívne nízke, ale samozrejme stále väčšie ako β_A. Taktiež zvolíme $t_B - t_A = 90$. V druhom režime sa pokúsime upraviť β_B a $t_B - t_A$ tak, aby sme stále spĺňali podstatu druhého režimu (variant B príde skoro po variante A a získa prevahu). Rozhodli sme sa pre hodnoty $\beta_B = 0.5$ a $t_B - t_A = 50$. Idea za týmto riešením je, že zvýšenie hodnoty β_B

kompenzujeme neskorším príchodom variantu B, aby sa variant B nerozšíril príliš skoro. Následne sme vykonali simulácie po vzore častí 4.1 a 4.2 a výsledky vykreslili na Obr. 20.



Obr. 20: Výpočty odchýlok odhadu relatívnej fitness a kvality logistických kriviek pre dva režimy parametrov podľa modelu (61). (A1) Teplotná mapa odchýlok odhadu relatívnej fitness pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(B1)** Teplotná mapa výpočtov $1 - R^2$ pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(A2)** Analogicky podľa (A1) pre druhý režim. **(B2)** Analogicky podľa (B1) pre druhý režim.

Všimnime si veľký rozdiel medzi grafmi (A1) a (A2). Na grafe (A1) registrujeme veľmi malé odchýlky odhadu fitness. Taktiež môžeme pozorovať dobrú konzistenciu proporcie B s logistickou funkciou na grafe (B1). Na grafe (B2) naopak pozorujeme veľmi slabú konzistenciu s logistickou krivkou, čo má za následok veľmi nepresné odhady relatívnej fitness na grafe (A1). Taktiež nevidíme žiadnu podobnosť druhého režimu s prvým režimom, ale práve kvôli veľmi nízkym hodnotám R^2 nemôžeme brať hodnoty na grafe (A2) ako štatisticky významné. Pre bližší pohľad na výsledky sme si vybrali tri nové scenáre, ktoré sme zakreslili do Obr. 21 a aj sme ich zaznačili do Tabuľky 12.



Obr. 21: Vybrané príklady kombinácií výrazov $\mu_B - \mu_A$ a $t_B - t_A$ pre oba režimy podľa Obr. 20. (C1) Prehľad dennej incidencie, proporcie variantu B a jej logit transformácia pre tri príklady z grafov (A1) a (B1). **(C2)** Analogicky podľa grafov (A2) a (B2).

Scenár	Odchýlka rel. fitness	Hodnota R^2
S1 (Režim 1)	-2.7484%	0.999741
S2 (Režim 1)	0.5194%	0.999998
S3 (Režim 1)	3.7125%	0.999926
S1 (Režim 2)	64.1941%	0.091529
S2 (Režim 2)	-1.8802%	0.074559
S3 (Režim 2)	-96.1263%	0.000259

Tabuľka 12: Odchýlky relatívnej fitness a hodnoty R^2 pre scenáre z Obr. 20.

Na grafe (C1) môžeme vidieť, že proporcie variantu B pre všetky tri scenáre sú konzistentné s logistickou krivkou. Taktiež pozorujeme vizuálne identickú smernicu logit transformácií. Na grafe (C2) sa nám opakujú výsledky z Obr. 19, kedy variant B príde príliš skoro a variant A má ešte vysoký podiel v populácii. Tento problém by sa dal vyriešiť oddialením vstupu variantu B do populácie, teda väčšou hodnotou $t_B - t_A$. Uvedomme si ale, že celá podstata druhého režimu je otestovať scenáre, kedy variant B príde skoro po variante A, aby sme mohli sledovať jeho prevahu v súboji s variantom A. Keďže ale

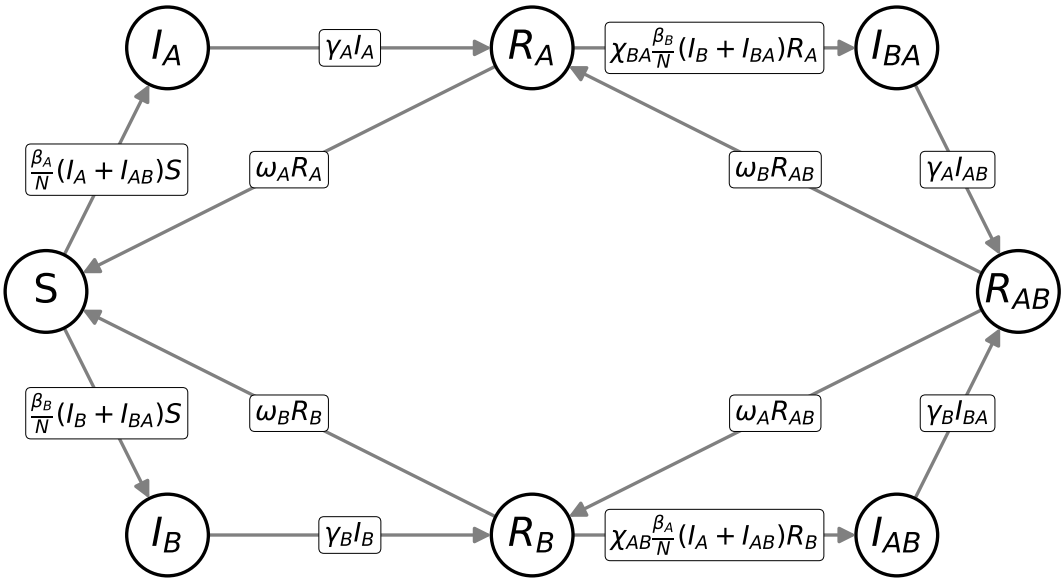
model (61) svojím zápisom efektívne eliminuje priamy súboj medzi variantmi, tak záleží skoro výlučne na čase vstupu variantu B. Ak by sme $t_B - t_A$ zväčšili, tak by sme síce dostali proporciu variantu B viac konzistentnú s logistickou krivkou, ale už by sme mali efektívne príklad prvého režimu, kedy variant B vstupuje do populácie neskoro. Môžeme preto konštatovať, že pre druhý režim sa nedá nájsť kombinácia parametrov taká, ktorá by umožnila modelovať proporciu variantu B ako logistickú funkciu.

V nasledujúcej časti využijeme poznatky z tejto aj z predchádzajúcich častí a vytvoríme kombinovaný SIR model, kde naraz aplikujeme viaceré prístupy k imunite po prekonaní infekcie.

4.4 Kombinovaný SIR model

V tejto časti sa budeme venovať poslednej modifikácii SIR modelu, pričom skombinujeme prístupy z častí 4.2 a 4.3 s cieľom dosiahnuť ešte realistickejší model. V novom modeli bude prítomná neúplná krížová imunita, ktorá má obmedzené trvanie a taktiež predstavíme princíp asymetrickej imunity. Pri uvažovaní dvoch variantov A a B budeme asymetrickú imunitu chápať ako jav, kedy jedinec po prekonaní prvého variantu má inú úroveň imunity voči druhému variantu ako úroveň imunity voči prvému variantu, ktorú by získal po prekonaní druhého variantu [16]. Skombinujeme teda modely (60) a (61) a prepíšeme ich ako model (62). Jeho diagram sme vykreslili na Obr. 22.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta_A}{N}(I_A + I_{AB})S - \frac{\beta_B}{N}(I_B + I_{BA})S + \omega_A R_A + \omega_B R_B\,,\\
\frac{dI_A}{dt} &= \frac{\beta_A}{N}(I_A + I_{AB})S - \gamma_A I_A\,,\\
\frac{dI_B}{dt} &= \frac{\beta_B}{N}(I_B + I_{BA})S - \gamma_B I_B\,,\\
\frac{dR_A}{dt} &= \gamma_A I_A - \chi_{BA}\frac{\beta_B}{N}(I_B + I_{BA})R_A - \omega_A R_A + \omega_B R_{AB}\,,\\
\frac{dR_B}{dt} &= \gamma_B I_B - \chi_{AB}\frac{\beta_A}{N}(I_A + I_{AB})R_B - \omega_B R_B + \omega_A R_{AB}\,,\\
\frac{dI_{BA}}{dt} &= \chi_{BA}\frac{\beta_B}{N}(I_B + I_{BA})R_A - \gamma_B I_{BA}\,,\\
\frac{dI_{AB}}{dt} &= \chi_{AB}\frac{\beta_A}{N}(I_A + I_{AB})R_B - \gamma_A I_{AB}\,,\\
\frac{dR_{AB}}{dt} &= \gamma_A I_{AB} + \gamma_B I_{BA} - \omega_A R_{AB} - \omega_B R_{AB}\,,
\end{aligned} \tag{62}$$



Obr. 22: Diagram systému (62) s vyznačením pohybov medzi skupinami.

V systéme (62) máme rovnaký počet rovníc ako v systéme (61), ale pridali sme viaceré parametre ovplyvňujúce imunitu po prekonaní infekcie. Parameter ω_A určuje rýchlosť straty imunity voči variantu A a parameter ω_B rýchlosť straty imunity voči variantu B. Uvedomme si, že strata imunity je sekvenčná, t. j. jedinec po strate imunity sa nepresunie z R_{AB} rovno do S, ale najprv do R_A alebo R_B a následne do S. Parameter χ_{BA} určuje mieru imunity voči variantu B po prekonaní variantu A, pričom $\chi_{AB} = 0$ vyjadruje úplnú imunitu a $\chi_{BA} = 1$ žiadnu imunitu. χ_{AB} analogicky určuje mieru imunity voči variantu A po prekonaní variantu B. Nastavenie parametrov χ_{BA} a χ_{AB} na rôzne hodnoty predstavuje asymetrickú imunitu.

Na simulácie sme využili upravený zoznam parametrov z Tabuľky 9. Keďže sa v modeli (62) vyskytuje viacero nových parametrov, vytvorili sme novú Tabuľku 13, do ktorej sme zapísali hodnoty parametrov, ktoré budeme používať v simuláciách.

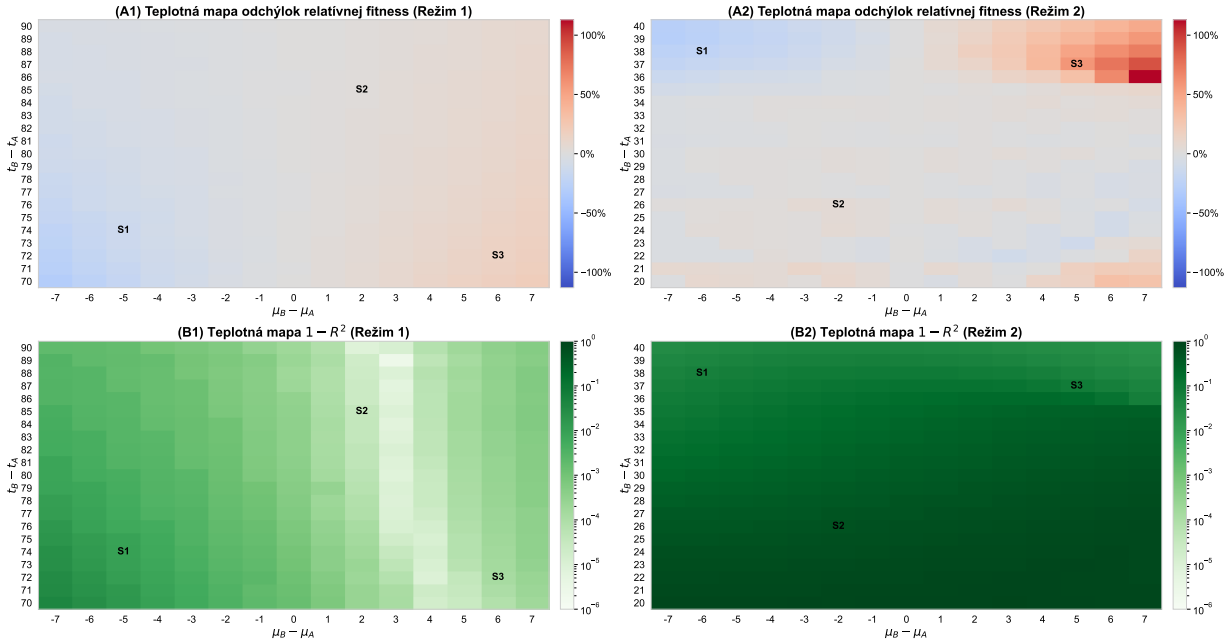
Parametre	Neskorý príchod variantu B (režim 1)	Skorý príchod variantu B (režim 2)
β_A	0.3	0.3
β_B	0.6	0.43
γ_A	$\frac{1}{5}$	$\frac{1}{5}$
γ_B	$\frac{1}{5}$	$\frac{1}{5}$
t_B	30	80
χ_{BA}	0.4	0.4
χ_{AB}	0.6	0.6
ω_A	$\frac{1}{90}$	$\frac{1}{90}$
ω_B	$\frac{1}{75}$	$\frac{1}{75}$

Tabuľka 13: Výber parametrov pre simulácie SIR modelu (62).

Pre prvý režim sme zvýšili hodnotu β_B a pre druhý režim ju znížili, čím sme sa viac priblížili k hodnotám pre základný model (53) v Tabuľke 9. Zároveň sme t_B upravili späť na pôvodné hodnoty. Tieto zmeny sme uskutočnili preto, lebo zavedením čiastočnej krížovej imunity danej χ_{AB} a χ_{BA} namiesto úplnej krížovej imunity očakávame menej extrémne hodnoty ako pre model (61). Zároveň uvažujeme $\chi_{AB} > \chi_{BA}$, pretože variant B považujeme za silnejší variant, teda by jedinec po jeho prekonaní mal mať aj vyššiu mieru imunitu voči slabšiemu variantu A. Nakoniec definujeme $\omega_A = \frac{1}{90}$ a $\omega_B = \frac{1}{75}$, teda imunita voči silnejšiemu variantu B trvá kratšie.

Následne sa posunieme k simuláciám. Pomocou parametrov z Tabuľky 13 sme vykonali simulácie modelu (62) podľa vzoru z predchádzajúcich častí a výsledky sme vykreslili na Obr. 23.

Na grafe (A1) môžeme pozorovať v tomto čase už známe výsledky, kedy pre vyššie hodnoty $\mu_B - \mu_A$ pozorujeme kladné odchýlky relatívnej fitness a pre nižšie hodnoty $\mu_B - \mu_A$ záporné odchýlky. Graf (B1), podobne ako (A1) je tiež konzistentný so simuláciami základného SIR modelu na Obr. 13. Na grafe (A2) pozorujeme podobný fenomén s odchýlkami relatívnej fitness, ale zaujímavé je primárne prepojenie s grafom (B2). Všeobecne pozorujeme menšie hodnoty R^2, pričom primárne vidíme hodnoty $R^2 < 0.99$. Pre oblasti

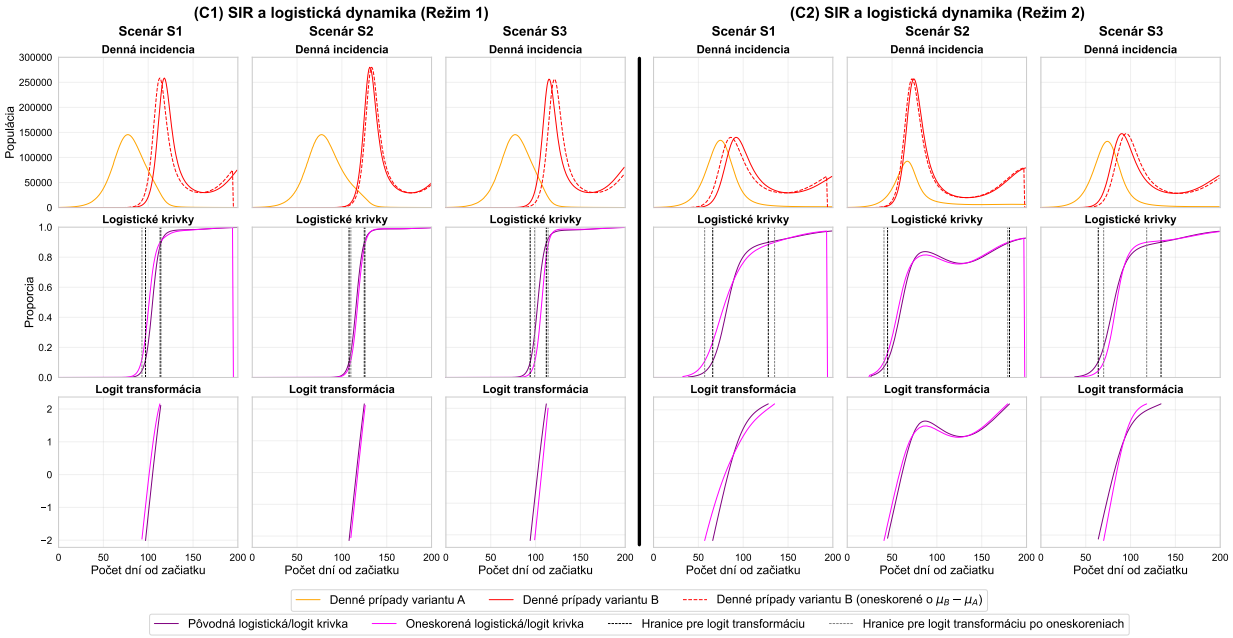


Obr. 23: Výpočty odchýlok odhadu relatívnej fitness a kvality logistických kriviek pre dva režimy parametrov podľa modelu (62). (A1) Teplotná mapa odchýlok odhadu relatívnej fitness pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(B1)** Teplotná mapa výpočtov $1 - R^2$ pre rôzne hodnoty $\mu_B - \mu_A$ a $t_B - t_A$. **(A2)** Analogicky podľa (A1) pre druhý režim. **(B2)** Analogicky podľa (B1) pre druhý režim.

s vyššou hodnotou $t_B - t_A$ pozorujeme väčšie odchýlky relatívnej fitness, ale zároveň aj vyššiu konzistenciu proporcie variantu B s logistickou funkciou. Pre nižšie $t_B - t_A$ vidíme menšie hodnoty R^2. Následne sme si vybrali tri scenáre rovnakým spôsobom ako v predchádzajúcich častiach a priblížili sme si ich na Obr. 24 a v Tabuľke 14.

Scenár	Odchýlka rel. fitness	Hodnota R^2
S1 (Režim 1)	-13.6182%	0.990528
S2 (Režim 1)	2.5447%	0.999965
S3 (Režim 1)	14.7091%	0.999866
S1 (Režim 2)	-22.6717%	0.945268
S2 (Režim 2)	4.7589%	0.471756
S3 (Režim 2)	57.4632%	0.041368

Tabuľka 14: Odchýlky relatívnej fitness a hodnoty R^2 pre scenáre z Obr. 23.



Obr. 24: Vybrané príklady kombinácií výrazov $\mu_B - \mu_A$ a $t_B - t_A$ pre oba režimy podľa Obr. 23. (C1) Prehľad dennej incidencie, proporcie variantu B a jej logit transformácia pre tri príklady z grafov (A1) a (B1). **(C2)** Analogicky podľa grafov (A2) a (B2).

Pri prvom režime môžeme pozorovať konzistenciu proporcie variantu B s logistickou krivkou aj po aplikovaní oneskorenia. Kvôli prítomnosti straty imunity danej parametrami ω_A a ω_B môžeme pozorovať aj druhú vlnu variantu B podobne ako na Obr. 17. Pri druhom režime pozorujeme čiastočnú konzistenciu proporcie variantu B s logistickou krivkou. Odchýlky od perfektnej podoby krivky nie sú také výrazné ako na Obr. 21, ale krivky po stredovom inflexnom bode rastú oveľa pomalšie ako pred ním. Za zmienku stojí scenár S2 druhého režimu, kde proporcia variantu B dokonca poklesne kvôli koncu prvej vlny variantu B, ale následne so začiatkom druhej vlny začne znova rásť.

Záver

Problematika populačnej genetiky sa po globálnej pandémii COVID-19 stala viac aktuálnou než kedykoľvek predtým. Aj v dnešnej dobe vedci neustále skúmajú dáta z obdobia pandémie a hľadajú nové poznatky, ktoré by sa dali využiť v budúcnosti a zabezpečili by lepšiu pripravenosť na ďalšie pandémie.

V tejto práci sme sa venovali modelovaniu šírenia vírusov pomocou rôznych modelov. Najprv sme predstavili princípy populačnej genetiky a zhrnuli sme doterajšie dôležité znalosti z tohto odvetvia. Predostreli sme motiváciu na modelovanie virálnej dynamiky SARS-CoV-2 pomocou dát z databázy GISAID zachytávajúcich priebeh šírenia vírusu v Spojenom kráľovstve v rokoch 2020 - 2024. Vykonali sme dôslednú analýzu dvoch vĺn s veľkou incidenciu, a to Alpha $\rightarrow$ Delta a Delta $\rightarrow$ Omicron BA.1. Zistili sme, že podiel invadujúceho variantu oproti udomácnenému variantu sa dá modelovať logistickou funkciou, z ktorej vieme inferovať hodnotu relatívnej fitness pomocou logit transformácie. Robustnosť tejto hodnoty sme overili pomocou neparametrického bootstrapu, ktorý nám potvrdil konzistentnosť výsledkov. V práci si po výsledkoch tejto kapitoly kladieme dve otázky: Kedy ide inferovať hodnotu relatívnej fitness? Ako presne vieme odhadnúť relatívnu fitness pri prítomnosti oneskorení?

Následne sme študovali schopnosť správne identifikovať relatívnu fitness z logit transformácie dát pomocou modelov dynamiky šírenia. Ako prvý sme analyzovali exponenciálny model. Pomocou syntetických dát sme najprv ukázali jeho funkčnosť, pričom podiel invadujúceho variantu oproti udomácnenému variantu sa dal aproximovať pomocou logistickej krivky, pri absencii náhodnosti v posune bola aproximácia dokonca exaktná. Následne sme na tento model aplikovali náhodné oneskorenia z normálneho rozdelenia. Na základe analytického výpočtu sme zistili, že presnosť odhadu relatívnej fitness invadujúceho variantu poklesne pri režime, kde je štandardná odchýlka náhodného oneskorenia invadujúceho variantu výrazne menšia ako u udomácneného variantu. Túto skutočnosť sme prvotne sledovali aj v numerických experimentoch, pričom udomácnený variant sme označili ako variant A a invadujúci variant ako variant B. Sledované odchýlky relatívnej fitness boli však spôsobené takou kombináciou oneskorenia a vstupu variantu B do populácie, ktorá

výrazne ovplyvnila začiatočnú proporciu variantu B oproti variantu A a hraničné porušovala predpoklad, že variant B nezistíme skôr ako variant A. Preto sme simulácie upravili a deterministicky posunuli vstup variantu B do populácie, čo podľa kapitoly 3 neovplyvnilo odhad fitness. Týmto sme eliminovali nekonzistentnosti na začiatku sledovaného obdobia a následne sme už nenašli režim, kde by sme registrovali výrazné odchýlky v odhade relatívnej fitness. Použitie realistickejšieho log-normálneho rozdelenia nám potvrdilo tento výsledok.

Za účelom realistickejšieho modelovania šírenia vírusu sme vybrali a analyzovali SIR model pre dva varianty (53). Náš cieľ bol zistiť, do akej miery bude proporcia variantu B stále konzistentná s logistickou funkciou a ako to bude vplývať na hodnoty odhadovanej relatívnej fitness. Pri analýze parametrov modelu (ešte pred aplikovaním oneskorenia) sme zistili, že pri fixne nastavených parametroch β_A a γ_A nemôžeme mať priveľmi malé β_B alebo priveľmi veľké γ_B. Ak by sa tak stalo, variant B sa nevie dostatočne rozšíriť a proporcia variantu B nebude konzistentná s logistickou funkciou ako v exponenciálnom modeli. Preto sme si vybrali zmysluplnú kombináciu vstupných parametrov a následne sme na model (53) aplikovali normálne a log-normálne rozdelené oneskorenia a zisťovali sme ich vplyv na odhad relatívnej fitness. Zistili sme, že pri stredných hodnotách oneskorenia pre variant B μ_B vyšších ako pre variant A μ_A registrujeme kladné odchýlky relatívnej fitness, teda odhadnutá fitness je väčšia ako fitness bez oneskorenia. Naopak, pri hodnotách μ_B nižších ako μ_A registrujeme záporné odchýlky, teda odhadnutá fitness je menšia. Taktiež sme dospeli k záveru, že štandardná odchýlka náhodných oneskorení má malý vplyv na odchýlky relatívnej fitness, dokonca za istých okolností ich aj zmenšuje. Preto sme v práci pokračovali s deterministickým oneskorením určeným len μ_A a μ_B, čo nám poskytlo priestor sa náležite venovať veľmi dôležitému faktoru - času vstupu variantu B do populácie. Experimentálne sme nafitovali SIR model na reálne denné dáta z vĺn Alpha $\rightarrow$ Delta a Delta $\rightarrow$ Omicron BA.1 a objasnili sme, že skorší príchod variantu B do populácie, resp. malý rozdiel medzi začiatkami variantov A a B spôsobuje väčšie odchýlky relatívnej fitness. Na základe týchto pozorovaní sme vytvorili dva režimy, jeden s neskorým a druhý so skorým príchodom variantu B a využili sme ich na pokračovanie práce, kde sme sa venovali modifikáciám SIR modelu.

Ako prvú modifikáciu sme si vybrali SIR model s dočasnou imunitou (60), ktorý nám umožnil modelovať koniec imunity jedincov a návrat k náchylnosti na infekciu. Simulácie modelu sme vykonali na dvoch režimoch spomenutých vyššie a získali sme veľmi podobné výsledky ako pre základný model (53). Obidva varianty boli schopné v populácii spôsobiť niekoľko vĺn infekcií, ale variant B vďaka vyššej infekčnosti prevládol nad variantom A a proporcia variantu B ostala konzistentnou s logistickou funkciou. Následne sme prešli k druhej modifikácii, a to SIR modelu bez krížovej imunity (61). Toto fundamentálne zmenilo naše vnímanie šírenia variantov A a B, pretože vyššia infekčnosť variantu B už nezabránila variantu A infikovať veľkú časť populácie. Ukázali sme, že ak je infekčnosť variantu B príliš vysoká, tak príliš rýchlo nainfikuje populáciu a zanikne, kým variant A bude ešte stále existovať. Taktiež sme poukázali na fakt, že príliš skorý príchod variantu B do populácie má za následok veľké odchýlky relatívnej fitness a nekonzistentnosť proporcie variantu B s logistickou funkciou. Nakoniec sme všetky naše vedomosti zlúčili do kombinovaného SIR modelu (62). Ukázali sme, že sa nám zlepšili hodnoty R^2, dokonca pre prvý režim vyzerali veľmi podobne ako pri základnom modeli (53). Pri druhom režime ale stále bolo R^2 slabšie, čo bolo spôsobené menším časovým odstupom variantov ako pri prvom režime.

Fenomén oneskorenia pozorovaných prípadov infekcie je prítomný a reálny problém, s ktorým sa musia vedci vedieť vysporiadať a následne vydať odporúčania pre daný vírus. Oneskorenie v realite spôsobuje veľa faktorov, napr. testovacie kapacity, asymptomatické infekcie aj nedostatočný reporting zo strany vládnych subjektov. Sústredili sme sa práve na oneskorenie, pretože na takejto škále ešte nebolo veľmi preskúmané. Z našej práce je viditeľné, že záporná odchýlka v odhade relatívnej fitness môže navodiť dojem, že vírus je menej nebezpečný než v realite je. Uvedomujeme si, že preskúmať SIR model pre úplne všetky kombinácie parametrov je ťažko uskutočniteľné a výber niekoľkých príkladov nikdy nebude odrážať úplne celé spektrum výsledkov. Vybrali sme ale príklady tak, aby sme za zachovania zrozumiteľnosti práce priniesli čo najobšírnejší prehľad numerickej stránky modelov, ktorý sme podložili dôkladne spravenými vizualizáciami. Vedeli by sme si predstaviť potenciálne pokračovanie tejto práce napríklad formou dôkladnej analytic-

kej a numerickej analýzy jedného špecifického modelu, prípadne použitia iného známeho modelu ako napríklad SEIR. Naša motivácia boli hlavne dve vlny vírusu SARS-CoV-2 analyzované v kapitole 2, ale existuje možnosť podobne dôkladnej analýzy aj ďalších vĺn, napríklad aj v porovnaní s inými krajinami. Správna interpretácia zašumených dát vplyvom časového oneskorenia je dôležitým faktorom pri rozhodovaní o opatreniach proti vírusom a je užitočné vedieť o možných úskaliach často používaných spôsobov modelovania šírenia vírusov.

Zoznam použitej literatúry

[1] Berrar, D.: *Introduction to the Non-Parametric Bootstrap*, Encyclopedia of Bioinformatics and Computational Biology, 2nd Edition, Volume 1 (2025), 339-348

[2] Charniga. K., Park, S. W., Akhmetzhanov, A. R., et al.: *Best practices for estimating and reporting epidemiological delay distributions of infectious diseases*, PLOS Computational Biology 10 (2024)

[3] Chen, Z., Azman, A. S., Chen, X., et al.: *Global landscape of SARS-CoV-2 genomic surveillance and data sharing*, Nature Genetics 54 (2022), 499-507

[4] Davies, N. G., Abbott, S., Barnard, E. C., et al.: *Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England*, Science 372 (2021)

[5] Dokumentácia Python knižnice matplotlib.pyplot, dostupné na internete (3.4.2026): https://matplotlib.org/stable/api/pyplot_summary.html

[6] Dokumentácia Python knižnice multiprocessing, dostupné na internete (3.4.2026): https://docs.python.org/3/library/multiprocessing.html

[7] Dokumentácia Python knižnice scipy, dostupné na internete (3.4.2026): https://docs.scipy.org/doc/scipy/reference/index.html#scipy-api

[8] Earnest, R., Uddin, R., Matluk, N., et al.: *Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha in New England, USA*, Cell Reports Medicine 3 (2022)

[9] Ehrhardt, M., Gašper, J., Kilianová, S.: *SIR-based mathematical modeling of infectious diseases with vaccination and waning immunity*, Journal of Computational Science 37 (2019)

[10] Elliott, P., Eales, O., Bodinier, B., et al.: *Dynamics of a national Omicron SARS-CoV-2 epidemic during January 2022 in England*, Nature Communications 13 (2022)

[11] GISAID, dostupné na internete (11.3.2025): https://gisaid.org/

[12] Hart, W. S., Miller, E., Andrews, N. J., et al.: *Generation time of the alpha and delta SARS-CoV-2 variants: an epidemiological analysis*, The Lancet Infectious Diseases (2022), 603-610

[13] Hartl, D. L., Clark, A. G.: *Principles of Population Genetics: Third Edition*, Sinauer Associates, Inc., Canada, 1997

[14] Kermack, W. O., McKendrick, A. G.: *A contribution to the mathematical theory of epidemics*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 115 (1927), 700-721

[15] Kretzschmar, M., Waalinga, J.: Mathematical Models in Infectious Disease Epidemiology, dostupné na internete (1.1.2009): https://link.springer.com/chapter/10.1007/978-0-387-93835-6_12

[16] Kucharski, A. J., Andreasen, V., Gog, J. R.: *Capturing the dynamics of pathogens with many strains*, Journal of Mathematical Biology 72 (2016), 1-24

[17] Liu, C., Lu, J., Li, P., et al.: *A comparative study on epidemiological characteristics, transmissibility, and pathogenicity of three COVID-19 outbreaks caused by different variants*, International Journal of Infectious Diseases 134 (2023), 78-87

[18] Marée, A. F. M., Keulen, W., Boucher, C. A. B., et al.: *Estimating Relative Fitness in Viral Competition Experiments*, Journal of Virology Volume 74/23, 11067 - 11072

[19] McCrone, J. T., Hill, V., Bajaj, S., et al.: *Context-specific emergence and growth of the SARS-CoV-2 Delta variant*, Nature 610 (2022), 154-160

[20] Otto, S. P., Day, T., Arino, J., et al.: *The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic*, Current Biology 31 (2021), 918-928

[21] S., Yo: Estimation of a Continuous Distribution on a Real Line by Discretization Methods–Complete Version–, dostupné na internete (November 2017): https://arxiv.org/pdf/1709.09520

[22] Shu, Y., McCauley, J.: *GISAID: Global initiative on sharing all influenza data - from vision to reality*, Eurosurveillance 22 (2017), 2-4

[23] Volz, E.: *Fitness, growth and transmissibility of SARS-CoV-2 genetic variants*, Nature Reviews Genetics 24 (2023), 724-734

[24] Weiss, H.: *The SIR model and the Foundations of Public Health*, Materials matemàtics (2013), 1-17

[25] Wu, Y., Kang, L., Guo, Z., et al.: *Incubation Period of COVID-19 Caused by Unique SARS-CoV-2 Strains*, JAMA Network Open 5 (2022)

Príloha A - Skripty

Zoznam Python skriptov na generovanie vizualizácií (každý skript vytvorí jeden špecifický obrázok v práci):

- Obr. 1: `image_data_overview_sk.py`
- Obr. 2: `image_data_fitness_sk.py`
- Obr. 3: `image_data_fitness_bonus_sk.py`
- Obr. 4: `image_exp_sk.py`
- Obr. 5: `image_exp_normal_intro_sk.py`
- Obr. 6: `image_exp_normal_sk.py`
- Obr. 7: `image_exp_lognormal_sk.py`
- Obr. 9: `image_sir_logistic_sk.py`
- Obr. 10: `image_sir_normal_intro_sk.py`
- Obr. 11: `image_sir_lognormal_intro_sk.py`
- Obr. 12: `image_sir_covid_sk.py`
- Obr. 13: `image_sir_sk.py`
- Obr. 14: `image_sir_scenarios_sk.py`
- Obr. 16: `image_sir_wi_sk.py`
- Obr. 17: `image_sir_wi_scenarios_sk.py`
- Obr. 19: `image_sir_nci_intro_sk.py`
- Obr. 20: `image_sir_nci_sk.py`
- Obr. 21: `image_sir_nci_scenarios_sk.py`
- Obr. 23: `image_sir_vci_sk.py`

- Obr. 24: `image_sir_vci_scenarios_sk.py`

Ďalšie skripty a súbory:

- Spracovanie surových dát z GISAID: `uk_data_read.py`
- Generalizácia a zoskupenie variantov z GISAID: `pango_groups_final.json`
- Generovanie SIR diagramov pre Obr. 8, 15, 18, 22: `diagram_sir.py`

Poznámky:

- Skripty boli spúšťané na QHD/2K monitore (rozlíšenie 2560x1440). Na inom rozlíšení sa môže vo vizualizáciách v režime na celú obrazovku vyskytnúť prelínanie prvkov.
- Skripty s využitím knižnice `multiprocessing` sú veľmi náročné na procesor (CPU), pričom na slabších procesoroch s menej jadrami môžu skripty bežať neúmerne dlho. Skripty boli spúšťané na procesore AMD Ryzen 5 7600.

Link na GitHub repozitár so skriptami: https://github.com/Robert-Valaska/DP