FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Hledání a práce s veličinami souvisejícími s TFR
Student:	Daniel Brotz
Vedoucí:	PhDr. Ing. Tomáš Evan, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

Popis:
Nastudujte problematiku total fertility rate (TFR) a získejte vhodné datové zdroje pro vytvoření přehledné webové aplikace, která bude zobrazovat TFR v různých státech v čase s projekcí do blízké budoucnosti. Zobrazte také prokázané veličiny, které mají na tuto hodnotu vliv. Vyberte několik online zdrojů (Google Trends, open data atd.) a pokuste se najít nové veličiny s významnou korelací.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 30. prosince 2021 v Praze.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Hledání a práce s veličinami souvisejícími s TFR

Daniel Brotz

Katedra aplikované matematiky
Vedoucí práce: PhDr. Ing. Tomáš Evan, Ph.D.

12. května 2022

Poděkování

V prvé řadě bych chtěl poděkovat PhDr. Ing. Tomáši Evanovi, Ph.D. za vedení této práce a motivaci při jejím zpracování. Za konzultace děkuji také Ing. Marku Sušickému. Svým nejbližším pak děkuji za podporu, kterou mi věnovali po celou dobu studia.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principu při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisu. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisu, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programu, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli zpusobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelum). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným zpusobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 12. května 2022 ……

České vysoké učení technické v Praze
Fakulta informačních technologií

Odkaz na tuto práci

Brotz, Daniel. *Hledání a práce s veličinami souvisejícími s TFR*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2022.

Abstrakt

Práce se zabývá demografickým ukazatelem plodnosti (TFR) a jeho významným poklesem v posledních desetiletích. Popisuje jeho vývoj a podává shrnutí literatury na téma veličin s plodností souvisejících. Na základě této literatury vybírá online datové zdroje, z nichž v praktické části získává data a nalézá nové korelace. Výsledky jsou prezentovány ve webové aplikaci, která rovněž nabízí přehled vývoje TFR ve vybraných zemích Evropy a jeho předpověď do blízké budoucnosti. Většina výsledků analýzy je v souladu s existující literaturou, nové korelace jsou poskytnuty k interpretaci.

Klíčová slova TFR, plodnost, demografie, otevřená data, korelace, předpověď časových řad

Abstract

This thesis deals with the problem of decline of the demographical indicator of fertility (TFR). Firstly, it summarizes existing work concerning known indicators related to TFR. Based on the presented literature, online data sources are selected and new correlations are found between indicators from these sources and TFR. The results are made available in a web-based application. This application also provides an overview and a prediction of the development of TFR in selected European countries. The majority of results are consistent with the presented literature, while new correlations are provided for further analysis.

Keywords TFR, fertility, demographics, open data, correlation, time series prediction

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Total fertility rate (TFR) či česky úhrnná plodnost je demografický ukazatel popisující počet dětí, které by se ve sledované společnosti mohly narodit jedné ženě. Aby se počet obyvatel dlouhodobě udržel na stejné hodnotě, TFR by mělo dosahovat hodnoty odhadované pro rozvinuté země na 2,1. Následkem populační exploze ve 20. století vznikla celosvětová snaha plodnost regulovat. Mnoho takových programů svůj cíl úspěšně splnilo a hodnota začala prudce klesat. Jedním z možných scénářů vývoje proto je, že se ještě v tomto století začne světová populace zmenšovat.

Zejména v Evropě populace stárne a cesta ke zvýšení plodnosti se zdá být složitější než opačný proces. Více než kdy dříve proto vznikají práce zkoumající demografické, socioekonomické a další faktory ve vztahu s TFR. Stejný důraz ale není kladen na výzkum zabývající se ukazateli, které jsou dostupné online. Internet se díky svému dnešním masovému rozšíření stal zrcadlem lidského chování, ale data zde dostupná zůstavají stále z velké části nevyužitá. Tato práce proto online zdroje využívá a nachází nové souvislosti, které mohou při dalším výzkumu TFR pomoci.

Práce se zaměřuje zejména na statistiky internetového vyhledávání získané z Google Trends. Využívá také stále se rozšiřující nabídky otevřených dat, které různé státní i mezinárodní insituce zvěřejňují. Jelikož je u otevřených dat kladen důraz na možnost strojového zpracování, jedná se o ideální vstup pro softwarové statistické šetření.

Čtenář je nejprve seznámen s demografií, jejími základními pojmy a ukazateli. Důraz je kladen na TFR, jeho měření a odvozené ukazatele. Nechybí historický kontext popisující vývoj růstu populace a s tím související rozvoj demografie. Začleněn je také přehled literatury, která diskutuje souvislosti mezi TFR a jinými ukazateli, které se v tomto kontextu běžně zkoumají. Následuje přehled matematických metod používaných pro zpracování dat s charakterem časových řad, kterými se práce zabývá.

V praktické části jsou využity online datové zdroje poskytující jak tradiční ukazatele zmiňované ve vybraných studiích, tak i nové, o nichž existuje minimální množství literatury. Práce popisuje proces získávání a následného statistického zpracování dat pomocí dříve popsaných metod. Součástí je také tvorba webové aplikace, která získané výsledky vizualizuje a zpřístupňuje interaktivní formou.

Kapitola 1

Cíle práce

Cílem teoretické části práce je seznámit čtenáře s TFR (total fertility rate) a souvisejícími demografickými ukazateli. Uveden bude i jejich vývoj v historii s důrazem na poslední čtyři dekády, které zároveň ohraničují rozsah dat zkoumaných v praktické části. Práce analyzuje vybrané stávající studie TFR a v těchto studiích nalezne prokázané veličiny, které na TFR mají vliv. Popsány budou také matematické metody vhodné k ověřování korelace časových řad a jejich projekci do budoucnosti.

V praktické části práce je cílem implementovat webovou aplikaci, která přehledně zobrazí vývoj TFR ve vybraných geografických oblastech s jeho předpovědí do blízké budoucnosti. Dalším cílem je nalezení nových ukazatelů z vybraných online zdrojů jako jsou Google Trends či open data státních institucí, které s vývojem TFR významně korelují. Zároveň aplikace poskytne přehled veličin s prokázaným vlivem na TFR a vlastních nalezených ukazatelů.

Přínosem bakalářské práce je jednotné zobrazení veličin prokazatelně souvisejících s TFR, které mohou být využity k jeho ovlivňování. Práce také poskytne nové významně korelující veličiny.

Kapitola 2

Analýza

Tato kapitola představuje základní demografické termíny a ukazatele. Na jejich základě definuje TFR a předkládá historický kontext, který se k tomuto ukazateli váže. Podává přehled literatury diskutující souvislost různých ukazatelů TFR. Zavádí také matematické metody užité při implementaci výsledného software.

2.1 Demografie a ukazatel plodnosti

Demografie je věda zabývající se populacemi, jejich vlastnostmi a proměnou v čase. Z hlediska lidské populace je hlavním objektem zájmu obnova obyvatelstva a s tím související procesy jako porodnost a úmrtnost. Základní údaje, tedy čísla získaná například při sčítání lidu či studiích na vybraných skupinách, obvykle kombinujeme, abychom získali nové, agregované ukazatele. Tyto nové hodnoty pak poskytují výstižnější odpovědi na konkrétní otázky o zkoumané populaci. Takto získané hodnoty dělíme do tří skupin:

Poměrná čísla extenzitní (ukazatele) vznikají porovnáváním dvou údajů měřících veličinu stejné povahy ve stejném prostoru a čase. Udávají se obvykle v procentech. V následujícím textu bude však pojem ukazatel užíván obecně jako označení demografického údaje či jiné měřitelné veličiny.

Poměrná čísla intenzitní (míry, kvocienty) jsou výsledkem podílu nastalých jevů s počtem jejich nositelů. Počet jevů se získává jako suma za určené období. Počet nositelů je pak vybraným způsobem aproximovaná hodnota reflektující změnu mezi začátkem a koncem daného období. Jako příklad lze uvést obecnou míru plodnosti popsanou níže.

Poměrná čísla srovnávací (indexy) srovnávají časově či prostorově nesouvisející hodnoty. Hodí se tam, kde chybí data pro výpočet míry – místo nich se použijí jiná data, která jsou k dispozici a empiricky o nich byl potvrzen vztah ke kýžené míře. Příkladem může být index plodnosti, který použijeme, pokud nemáme data o živě narozených dětech, ale chceme odhadnout plodnost. [1]

2.1.1 Základní demografické ukazatele

Elementární demografické ukazatele se vztahují k obnově obyvatelstva a jeho velikosti. Jedním ze základních ukazatelů je *porodnost (natalita)*, která popisuje počet narozených za určité časové období na 1 000 jedinců a uvádí se v promilích. Porodnost můžeme dále specifikovat jako *hrubou míru porodnosti* uvádějící počet všech živě narozených N^v na 1 000 jedinců za jeden rok. Počet narození se vztahuje ke střednímu stavu obyvatelstva P (obvykle aritmetický průměr z počtu obyvatel na začátku a konci roku). Míru pak lze získat ze vztahu

$$hmp = \frac{N}{P} * 1000.$$

Ukazatelem doplňujícím porodnost je *úmrtnost (mortalita)*. Podobně jako porodnost se udává v promilích a popisuje počet zemřelých na 1 000 jedinců za sledované období. Základním ukazatelem úmrtnosti je *hrubá míra úmrtnosti* vypočítaná jako podíl počtu zemřelých a středního stavu obyvatel obdobně jako hrubá míra porodnosti. Jelikož úmrtnost v populaci prudce roste s věkem, výstižnější je *míra úmrtnosti dle věku*, která dává do poměru počty zemřelých a celkové počty žijících vždy pro jednotlivé věkové skupiny.

Pro pochopení významu TFR je nutné se nejprve seznámit s ukazatelem plodnosti. *Plodnost (fertility rate)* udává průměrný počet potomků narozených jedné ženě za celý její život. Od plodnosti odlišujeme termín *fekundita (plodivost)*, který označuje potenciál plození potomků jakožto biologickou schopnost. Nejjednodušším ukazatelem plodnosti je *obecná míra plodnosti*. Pro její výpočet je třeba počet živě narozených dětí N^v a počet žen v reprodukčním věku (15–49 let) P^z_{15-49}. Míra je výsledkem dosazení do vztahu

$$f = \frac{N^v}{P^z_{15-49}} * 1000.$$

Nedostatkem tohoto jednoduchého ukazatele je ale nerozlišení různých věkových skupin, jelikož mezi nimi bývají v počtu dětí významné rozdíly. Podobně jako u úmrtnosti lze proto vypočítat i *míru plodnosti podle věku*. Do poměru dává počet živě narozených dětí ženám ve věku x N^v_x a střední stav žen ve věku x P^z_x. Míra se získává jako výsledek výpočtu

$$f_x = \frac{N^v_x}{P^z_x} * 1000.$$

V případě nedostupnosti dat je možné použít počet všech narozených N_x. [1]

2.1.2 TFR

TFR je anglickou zkratkou pro Total fertility rate, český název veličiny je *úhrnná plodnost*. Cílem této veličiny je odpovědět na otázku, kolik dětí připadá na jednu ženu. Vychází z tabulky plodnosti, která sestává z řady měr

plodnosti podle věku f_x. Úhrnnou plodnost získáme sečtením těchto měr pro jedno období. TFR tedy vyjadřuje počet dětí, které by se narodily jedné ženě, pokud by přežila celé své reprodukční období a po jeho dobu by platily míry plodnosti f_x stejné jako v období, za které míry plodnosti sčítáme. Zkoumáme tedy fiktivní generaci, jejíž míry plodnosti jsou složeny ze skutečných měr plodnosti 35 navazujících generací.

Součtem měr plodnosti v jedné generaci, která je již za koncem svého reprodukčního období, získáme *konečnou plodnost*. Ta říká, kolik dětí se průměrně narodilo ženám této generace. Vynásobením úhrnné plodnosti podílem dívek mezi narozenými získáme jednoduchý ukazatel *hrubé míry reprodukce*. Pokud dosahuje hodnoty vyšší než 1, dá se zhruba říci, že daná generace zajišťuje svou obměnu, ovšem se zanedbáním úmrtnosti.

Za povšimnutí stojí rozdíl mezi vývojem úhrnné a konečné plodnosti. Jelikož úhrnná plodnost sleduje fiktivní generaci vytvořenou z hodnot získaných obvykle za jeden rok, je mnohem citlivější na vnější vlivy, které ovlivňují porody. Například období nepříznivé ekonomické situace může totiž ovlivnit pouze časování porodů. V takovém období tedy TFR prudce poklesne, ale po odeznění negativního vlivu se ztráta na plodnosti značně vykompenzuje. Konečná plodnost generací, které byly ve svém reprodukčním období danou situací ovlivněny, tedy nemusí utrpět takovou ztrátu. Proto se TFR hodí ke zkoumání souvislosti externích vlivů s plodností. [1]

2.1.3 Historický kontext

Před průmyslovou revolucí byl vysoký počet dětí v rodině vyvážen vysokou úmrtností. Ženy trávily přes dvě třetiny svého života v dospělosti péčí o děti, kterých na průměrnou rodinu připadalo šest. V roce 1798 vyšla Esej o principu populace, podle jejíhož autora se teorie v ní obsažená nazývá jako *malthusiánství*. Velikost populace byla podle něj držena v šachu dvěma vlivy, které by vyvážily případný nadměrný růst, a sice zvýšením mortality kvůli sníženým mzdám doprovázeným hladomorem, válkou či epidemií na jedné straně a zpožděním vstupu do manželství vzhledem k nepříznivé ekonomické situaci, což by zapříčinilo rozmach prostituce a dalších neřestí na straně druhé. [2]

Mortalita ale začala klesat, ať už následkem zlepšení ekonomické situace či rozvojem medicíny a světová populace se tak rozrůstala. Tento proces demografického přechodu je dnes v různých fázích pozorovatelný na celém světě, přestože jeho kolébkou byla západní Evropa. Další fází, kterou dodnes prošla většina rozvinutých států, je snížení plodnosti související s opětovným zpomalením růstu populace. Pro zbytek světa ale přišel demografický přechod později, například v Indii bylo ještě před bezmála sto lety běžně šest dětí v jedné domácnosti. [2]

Po druhé světové válce lze ale pozorovat prudké zrychlení růstu populace téměř na celém světě. Objevily se četné obavy o nerovnováhu mezi produkcí a spotřebou a v padesátých letech vznikly první návrhy na zavedení programů

na snížení plodnosti zejména v rozvojových zemích. V sedmdesátých letech už se pak téměř sto států z celého světa aktivně podílelo na ovlivňování obyvatel v rozhodnutích týkajících se rodičovství ať už pozitivním či negativním směrem. Programy založené jednak na změně společenské představy o ideální velikosti rodiny, jednak na distribuci a zvyšování povědomí o antikoncepci a např. také na regulaci ekonomickými prostředky byly velmi úspěšné. Realizovány přitom byly napříč všemi druhy státních zřízení, politickými orentacemi či náboženskými vyznáními. [3]

Úspěch snahy o snížení plodnosti ale vystřídalo rychlé vystřízlivění a po prudkém propadu se TFR odrazilo ode dna na hodnotách pod jedním dítětem na ženu. Zvláštním úkazem jsou pak postkomunistické země devadesátých let, kdy zřejmě lidé odložili rodičovství pod vlivem rozšířených možností v důsledku pronikající západní kultury. V dnešní době je TFR v polovině států světa pod hodnotou 2,1 považovanou za hranici reprodukce. [4]

Rozhodnutí mít méně dětí či vyhnout se rodičovství úplně je třeba na jedné straně považovat za racionální úsudek jednotlivce jakožto výsledek zvážení vlastní osobní situace, vyhlídek na budoucnost a ideálů. Filosofickou otázkou je pak zvážení bezpodmínečnosti přítomnosti přirozené potřeby reprodukce u každého jedince. Z praktického hlediska ale velmi nízká plodnost vede ke stárnutí populace a hrozící ekonomické nestabilitě na poli důchodů, výdajů na zdravotnictví a dalším problémům. Důkazem nutnosti studování této problematiky pak je čím dál rychlejší tempo vzniku literatury na téma plodnosti a souvisejících veličin.

2.2 Výzkum veličin souvisejících s TFR

Tato podkapitola se zaměřuje na veličiny s prokázaným vlivem na TFR či s potvrzenou silnou korelací. Veličiny spadají do časového rozsahu posledních čtyř dekád a jsou analyzovány v praktické části práce. Jejich vizualizace v porovnání s vývojem TFR je dostupná ve webové aplikaci. Vybraným tématům, jejichž ukazatelům byla prokázána významnější tvrzení o korelaci, jsou věnovány samostatné podkapitoly. Ostatní ukazatele jsou pak shrnuty ve společné kapitole.

2.2.1 Výše důchodů

Většina evropských států provozuje sociální systém typu PAYGO (pay-as-you-go), kdy výdělečně činní financují penzi těch, kteří jsou ve stejném čase v důchodu. Vzhledem ke klesající plodnosti se přitom snižuje podíl ekonomicky aktivního obyvatelstva a podíl příjemců důchodu naopak narůstá [5]. Možným řešením by mohlo být zvýšení věku odchodu do důchodu, [6] argumentuje, že i mírné zvýšení postačí pro vyrovnání ztrát ve státním příjmu určeném na důchody. Navíc by podle [7] stárnutí populace mohlo snížit výdaje v jiných

oblastech jako je školství, i když změna by byla patrná zejména v méně rozvinutých zemích.

Pokud však výše přijímaného důchodu ovlivňuje plánování rodičovství obyvatel v produktivním věku, je třeba zvážit, jestli snaha o zachování současné výše důchodů není v konfliktu se snahou o zvýšení či udržení plodnosti. Podle [4] existuje významná negativní korelace mezi výší důchodů a plodností. Vysvětlení podává [8] na rozdílu mezi tradiční a moderní společností, kdy dříve probíhalo mezigenerační zajištění uvnitř rodiny. Pozdější oslabení rodinných vazeb dalo vzniknout sociálnímu státu a lidé tak přestali mít potřebu vychovat tolik dětí. Studie [9] argumentuje naopak tím, že důchody kompenzují cenu příležitosti výchovy dítěte a při modelování realokace prostředků z důchodů na vzdělávání pozoruje snížení plodnosti, ovšem pouze za určitých daňových podmínek.

Jako možná alternativa ke protichůdným postupům snižování důchodů a snahy o jejich udržení na stejné hladině se nabízí podpora sociálního systému odměňujícího důchodce za děti, které vychovali. Na možnou nestabilitu takového systému upozorňuje [5]. Podle [10] pak jeho funkce úzce závisí na předpokladu, že všichni rodičové se o své dětí starají stejně a že rodičovství lze dokonale ovlivňovat. Tyto obavy podporuje [11], ovlivnění výše důchodu na základě počtu dětí totiž může penalizovat rodiče, kteří do dětí investují víc a mají jich proto méně.

Penzijní systémy se běžně rozlišují na Bismarckovský a Beveridgeovský typ. První jmenovaný odměňuje příjemce proporcionálně k jejich předchozímu výdělku, druhý prostředky získané od pracujících přerozděluje. Kombinaci klasického Bismarckovského typu (i) se zmíněným systémem odměňování na základě výchovy dětí (ii) navrhuje [12]. Řeší tak zmíněnou námitku [10] ohledně funkčnosti (ii) tím, že jedinec se může sám rozhodnout podle vlastní situace, jestli chce mít děti a zvolí si (ii) či upřednostní vlastní výdělek a na něj navázanou odměnu (i).

Z historického pohledu prezentovaného v [13] je nicméně na datech z období vzniku Bismarckova penzijního systému v Německu ke konci 19. století znatelný negativní vztah mezi plodností a důchody. Na maďarských datech zavedení důchodového systému modeluje [14] se zjištěním, že plodnost se v takové situaci snižuje. Podporu přidává [15] s výzkumem na vzorku 28 zemí OECD, zavedení penzí navíc přičítá i snížení motivace k zakládání rodin. Z toho může plynout snížení mezigeneračního zajištění, které vede k dalšímu posílení výdajů na důchody.

Nabízí se proto otázka, jestli je vůbec penzijní systém potřebný. Protiargumentem pak je [12] s tvrzením, že přímá podpora uvnitř rodiny není vynutitelná. Pokud by ale byli v rámci kompromisu podporování jen nízkopříjmoví jedinci, mohl by se objevit problém se zneužíváním takového systému, kdy by zmínění ztráceli motivaci k individuálnímu spoření. Na mezigenerační problematiku navazuje [6] uvádějící, že mladší populace přestává považovat přímou podporu starší generace za nutnou. Tvrzení, že důchodový systém přerozdělu-

jící prostředky pojišťuje příjem jedince ve stáří lépe [12], je tak podpořeno. Zesilování důrazu na důchody ale vzhledem k uvedené literatuře [8, 14, 15] může efekt mizící přímé podpory v rodině spolu se snižující se plodností prohlubovat. Přesto ale jedním z faktorů ovlivňujících rozhodování potenciálních rodičů ohledně výchovy dětí může být očekávání budoucí podpory, která nemusí nutně nabývat peněžní podoby.

V podmínkách systému typu PAYGO ale přináší přivedení nové pracovní síly na svět výhodu i všem ostatním vrstevníkům rodičů [11]. Bez přímé kompenzace tak jedinec může ztratit motivaci vychovat potomky, pokud se započítá altruismus v rámci rodiny [15].

2.2.2 Zaměstanost žen

Ukazatelem participace obyvatelstva v pracovním procesu je je *LFP (labour force participation)*. Při zaměření pouze na ženy se pak jedná o *FLFP (female LFP)*, v některé literatuře zkracováno jako *FLP*. Způsob měření se napříč státy může lišit, ne vždy se například započítává práce na částečný úvazek či různé formy brigády. Naopak ženy na mateřské dovolené mohou být započítávány jako součást pracovní síly a proto je problematické dokonale porovnat výsledky různých studií zkoumajících vliv FLP na plodnost. [4]

Pro vysvětlení souvislosti plodnosti a zaměstanosti žen existuje tzv. hypotéza nekompatibility, která zdůrazňuje kolizi výkonu zaměstnání mimo domov a potřeby starat se doma o potomky. Spekuluje se naopak, že v zemědělsky zaměřených zemích ženy pracují blíže domovu a problém proto není tak výrazný [16]. Na druhé straně stojí teorie potřeby většího příjmu, když do domácnosti přibyde dítě, což by indikovalo větší poptávku matek po zaměstnání [17]. Druhou teorii [17] vyvrací s panelovým výzkumem na zemích G7 mezi lety 1960 a 2006, kde nalézá negativní vztah FLFP a TFR. Do modelu však nezahrnuje žádné další ukazatele, jejichž vliv by mohl zůstat skrytý.

Vztah se pokouší odhalit s využitím japonských dat [18], přidávající do analýzy mzdu žen a ukazatel intenzity státní podpory na podporu plodnosti. Zjišťuje však zásadní negativní vliv mezd jak na TFR, tak na FLFP, kde zejména druhý vztah je těžko interpretovatelný. Pozorovaná pozitivní korelace mezi TFR a FLFP mizí, když je do modelu přidána státní podpora plodnosti. Efekt FLFP na plodnost zkoumá také [4] nacházející slabou, avšak pozitivní korelaci na vzorku 34 zemí OECD. Tu však považuje za efekt společenských změn východní Evropy v 90. letech, kde tento výsledek pozoruje, nikoli za obecný důkaz pozitivního vztahu. V období 2000–2013 je již asociace nevýznamná. Situace v USA podle výzkumu [19] na intervalu mezi lety 1948 a 2007 je odlišná, oboustranná nepřímá závislost přes další ukazatele je mezi TFR a FLFP nalezena.

Interpretaci zkoumaného vztahu nabízí kromě hypotézy nekompatibility i další teorie. Jednak se jedná o přístup školy New Home Economics zavedený G. Beckerem v 60. letech, kdy je mzda ženy považována za cenu příležitosti

výchovy potomka, z čehož vyplývá negativní vztah mezi výší mzdy a plodností [19]. Jednak je zmiňována hypotéza, kterou představil R. Easterlin, vysvětlující plodnost jako veličinu závislou na „relativním příjmu muže“. Myšlenka spočívá v porovnání příjmu syna a rodičů. Je-li syn schopen dosáhnout stejné nebo lepší životní úrovně ve srovnání s tou, v níž vyrůstal, plodnost vzniknuvší rodiny bude vyšší a naopak. Do souvislosti pak bývá dávána i relativní velikost kohort s odůvodněním, že menší kohorta znamená menší konkurenci a lepší mzdové vyhlídky, potažmo vyšší plodnost [20]. Podporu Easterlinovy hypotézy nabízí [19]. Přidává se [21] s modelem pracujícím s několika dalšími ukazateli ve Spojených státech.

Vzájemně si odporující výsledky lze nalézt v litaratuře na téma změny znaménka korelace mezi TFR a FLFP ke konci minulého století. 22 zemí s nízkou plodností bere v úvahu [22]. Objevuje změnu v polovině 80. let, kdy se korelace mění na pozitivní a přisuzuje ji strukturálním společenským změnám. Obdobný výsledek nalézá [23] na panelu 23 států OECD, na němž sleduje vývoj korelace napříč státy pro každý rok (cross-sectional). Již zmiňovaný výzkum [4] sice nalézá (na geograficky širším vzorku) také změnu, avšak mnohem později a s varováním, že při individuálním zkoumání jednotlivých zemí se ukazuje, že vztah vůbec nemusí být přítomen, případně podléhá jinému neznámému ukazateli. Tyto ukazatele se pokouší identifikovat [24] s odkazem na [25], kde je změna znaménka korelace přisuzována také vzájemná různorodost sledovaných zemí. Potvrzení přináší také [26] na datech sahajících do roku 2017, kde se také potvrzuje tím pozitivnější korelace mezi TFR a FLFP, čím novější data jsou uvažována. Naproti tomu [16] naléza významnou negativní korelaci mezi zaměstnaností žen a plodností, nicméně výsledky se liší v závislosti na typu zaměstnání.

2.2.3 Státní prorodinná politika

Zatímco ještě před padesáti lety byla problémem vysoká plodnost, zejména rozvinuté země dnes trápí opačná otázka. Státní opatření na podporu plodnosti mohou mít mnoho podob, jedná se například o daňová zvýhodnění, podporu pracovních příležitosti, započítání vychovaných dětí do výše starobního důchodu nebo přímo „mzdu pro matky“. Názory týkající se síly či dokonce polarity vlivu jednotlivých možností na TFR se však různí. [4, 12]

Rozsáhlou studii provádí [27], když porovnává vliv prorodinných faktorů na plodnost ve 20 evropských státech. Mezi možné vlivy zahrnuje nejen institucionální podporu, ale také pracovní podmínky, rodinnou podporu a další. Silný pozitivní vliv těchto činitelů nachází zejména vzhledem k druhým a dalším narozením. V souladu stojí studie [4], která bere v úvahu státní podporu rodin měřenou podílem na HDP a nalézá rovněž pozitivní vliv na plodnost.

V opozici stojí například [3] se svým výzkumem téměř 200 států, kdy ukazuje, že od 70. let narůstá podíl těch, které aplikují prorodinnou politiku, nicméně jejich průměrná plodnost sice pomaleji ale stále klesá. Není však jasné,

jestli průměr nesnižují postupně se přidávající státy, když začnou takovou politiku zavádět. Příklady na datech z konkrétních zemí totiž potvrzují pozitivní vazbu. Jedná se např. o studii [28] zkoumající vliv finančních příspěvků pro rodiny s dětmi (vázaných na předchozí zaměstnanost) na plodnost ve Švédsku. Prezentovaná data ukazují, že příspěvky mohly pomoci zpomalit pokles plodnosti. Rozvoj mnoha různých způsobů podpory rodin v Kanadě začínající už před sto lety analyzuje [29]. Jejich pozitivní vliv na TFR je potvrzen, přestože není tak silný, jak by se mohlo očekávat. Mimo jiné také zmiňuje silnější vliv výše mužských než ženských mezd.

Institucionální péče o děti se dá považovat za další úlevu rodičům, kteří v důsledku získávají čas na ekonomickou aktivitu v zaměstnání. Studie [30] k tomuto fenoménu přistupuje z pohledu migrace obyvatelstva mezi městským a venkovským prostředím a změny v politice podpory. Hlavním zjištěním je, že tato forma podpory skutečně plodnost zvyšuje, dalším efektem ale může být přesun domácností na venkov. Vzhledem k rozdílům mezi plodností městské a venkovské populace, kterou dokazuje například [31], může zmíněný přesun populace ještě posílit zvyšování plodnosti.

Za porovnání stojí také lokální studie v různých kulturách. Německá reforma příspěvků na děti v 90. letech podle [32] měla za následek zejména zvýšení počtu druhých a dalších dětí ve vysokopříjmových a vzdělanějších rodinách. Výzkum na datech z podobného období v Izraeli popsaný v [33] německé zjištění podporuje, když nalézá silnější vliv příspěvků na děti především na plodnost žen s vyšší úrovní vzdělání. V opozici stojí již zmíněný článek [28], který považuje za převažující vliv na vyšší plodnost nízké vzdělání a ani státní podpora jej údajně nedokáže kompenzovat. Vzhledem k velmi nízké plodnosti, již zažívá Jižní Korea, se nabízí analyzovat vliv příspěvků i zde. Na početných datech z 230 municipalit postihujících třináctileté období od roku 2001 se podle [34] ukazuje jasný pozitivní efekt, přestože z ekonomického hlediska se zdá nereálné jen pomocí příspěvků dostatečně zvýšit TFR. Pro Finsko kolem přelomu tisíciletí přináší další důkazy [35]. Na vzorku tamnějších rodin nachází větší pravděpodobnost narození druhého (se slabším vlivem třetího) dítěte těm ženám, které přijímají finanční příspěvky.

Z pohledu rozdílů mezi státy na poli ceny výchovy dítěte souvisejících s plodností nahlíží problematiku [36]. Zaměřuje se na několik zemí západní Evropy spolu s USA a přestože je obtížné dobře porovnat celkovou cenu rodičovství, konflikt pracovního zatížení s časem stráveným v domácnosti a další faktory patrně dobře rozdíly v plodnosti vysvětlují. Proto se zdá smyslupné cenu dítěte zejména z pohledu ušlého zisku kvůli pauze v zaměstnání vyrovnávat státní podporou. Přehled literatury, který vyznívá spíše ve prospěch slabšího až nevýznamného vlivu státních opatření přináší [37]. Kritizuje zejména studie nezapočítávající dostatečné množství dalších možných vlivů do svých modelů.

2.2.4 Ostatní vlivy

Vzdělání Od průmyslové revoluce se začala rapidněji zvyšovat kvalifikovanost práce a rodiče proto investovali více do vzdělání dětí a začínali se orientovat na kvalitu, nikoliv kvantitu [3]. Dnes je zejména úroveň vzdělanosti žen většinou asociována s nižší plodností. Vysvětlením může být například opoždění sňatku, pokud se potenciální matka rozhodne déle studovat [38], za další možnou interpretaci lze považovat zlepšení pracovních podmínek a související zvýšení ceny příležitosti rodičovství u žen. V případě pozorování vlivu vzdělání na další ukazatele jako LFP se totiž ukazuje pozitivní asociace například v případě Itálie 90. let [39], nelze tedy vyloučit i nepřímé vlivy vzdělání na plodnost. S důkazem přímého vlivu vzdělanosti obou rodičů na TFR přichází studie z Taiwanu [40], u matek je navíc efekt silnější než u otců. Při pohledu na méně rozvinuté země je patrný vliv jak základního [41], tak středního vzdělání [42]. Je ale třeba podotknout, že v těchto zemích zůstávají nejsilnějším faktorem srážejícím plodnost různé státní programy vzniklé přímo za tímto účelem [3, 43].

Ekonomika a související rozvoj Jedním ze základních ukazatelů úrovně ekonomiky je HDP. Z historického hlediska lze pozorovat růst ekonomiky současně s poklesem plodnosti, již na začátku demografického přechodu se jedná o dobře pozorovatelný proces [44]. Zvyšování příjmu jednotlivců vyvažuje nutné zvýšení investice do dítěte, které zmiňuje [3], ekonomika se tak dále rozvíjí a z rodičovství se stává překážka zaměstnání [38]. Přerod popisuje také [45], zaměřuje se ale na opačný efekt, kdy demografickému přechodu připisuje podíl na růstu ekonomiky a přechodu k dnešnímu systému. Stejně tak v pozdější době je možné pozorovat podobný efekt, navíc jsou dostupná podrobnější data, pomocí nichž lze rozvoj popsat. Studie [46] na 114 státech v druhé polovině minulého století rozeznává dvojí směr vazby mezi příjmem a plodností. Snížení plodnosti podle jejích zjištění ústí ve zvýšení příjmů a to naopak ve snižování plodnosti.

Urbanizace Urbanizace je spojena se snižováním plodnosti již od počátku demografického přechodu, studování tohoto fenoménu začalo již před druhou světovou válkou. Přesun obyvatel do měst provázejí výrazné společenské i ekonomické změny od zlepšení dostupnosti zdravotní péče k lepšímu přístupu ke vzdělání, jehož efekt byl popsán výše [47]. Mezi jedny z prvních důkazů vlivu urbanizace patří [48]. Výzkum dat z přelomu tisíciletí [49] se snaží vysvětlit stejný rozdíl na příkladu Číny a přes výrazný vliv politiky jednoho dítěte nachází významný negativní vliv urbanizace na plodnost, který se ještě umocňuje na počátku 21. století. Studie pokrývající více než posledních 60 let [50] historie Atén a jejich okolí podchycuje změny v distribuci obyvatelstva rámci urbanizačního cyklu a nalézá přímou návaznost na fluktuace plodnosti.

2.3 Matematické metody

Časové řady, kterými se tato práce zabývá, jsou posloupnostmi po sobě jdoucích hodnot s pevnou frekvencí vzorkování. Pro účely analýzy se jedná o náhodné vektory s přidanými vlastnostmi, jež jsou popsány níže.

Uvažuje-li se vektor $\mathbf{x} \in \mathbb{R}_n$ a pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$, pak je náhodným vektorem reálná vektorová funkce $\mathbf{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))^T$, kde $\omega \in \Omega$, pro niž platí

$$\mathbf{x} \in \mathbb{R}_n \Rightarrow [\mathbf{X} < \mathbf{X}] \in \mathcal{A}.$$

[51]

2.3.1 Testování hypotéz

Testování hypotéz je ve statistice rozhodováním o platnosti nějakého tvrzení týkajícího se rozdělení distribuční funkce náhodného vektoru. Stanoví se *nulová hypotéza* H_0 a její negace – *alternativní hypotéza* H_A. Rozhodnutím pak je zamítnutí či nezamítnutí H_0. Jelikož rozhodování probíhá na základě hodnot náhodného vektoru, mohou nastat dva druhy chyb. Jednak *chyba prvního druhu*, tedy neoprávněné zamítnutí H_0, jednak *chyba druhého druhu*, tedy nezamítnutí H_0, ačkoli neplatí. [51]

Pro rozhodnutí sestavíme množinu možných výsledků *(konfidenční interval)* náhodného pokusu, do níž by měly co nejčastěji spadat náhodné vektory, pro něž platí H_A. Velikost konfidenčního intervalu je zvolena tak, aby chyba prvního druhu nastávala nejvýše s $(100 * \alpha)\%$ pravděpodobností. Pravděpodobnost „falešného obvinění“ je tedy pro takto sestavený test pevně pod kontrolou. Naopak síla testu, tedy pravděpodobnost zamítnutí H_0, pokud platí H_A je neznámá, ale testy se konstruují tak, aby byla co možná nejvyšší. [51]

2.3.2 Lineární regrese

Lineární regrese vysvětluje závislost náhodné veličiny y na vysvětlující proměné pomocí modelu ve tvaru

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{2.1}$$

kde α a β jsou parametry a ε_i jsou i.i.d. náhodné veličiny s nulovou střední hodnotou.

Pro hledání parametrů modelu se nejčastěji využívá *metoda nejmenších čtverců (OLS)*, která poskytuje nestranné odhady s minimálním rozptylem a známým rozdělením. Tato metoda minimalizuje reziduální součet čtverců

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{2.2}$$

kde $\hat{y}_i$ je výsledkem dosazení odhadů parametrů a, b do rovnice

$$\hat{y}_i = a + bx_i. \tag{2.3}$$

Derivací reziduálního součtu čtverců podle $\hat{\alpha}$, $\hat{\beta}$ je nalezen vztah, pro nějž je součet minimální, nejlepší odhady se pak rovnají

$$a = \bar{y} - b\bar{x}, \tag{2.4}$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \tag{2.5}$$

kde $\bar{x}$, $\bar{y}$ jsou výběrové průměry daných veličin. [52]

Zjištěné odhady parametrů lze využít k testování lineární závislosti, tedy nulové hypotézy

$$H_0 : \beta = 0 \tag{2.6}$$

proti alternativní hypotéze

$$H_A : \beta \neq 0. \tag{2.7}$$

Reziduální rozptyl je nestranným odhadem rozptylu vysvětlované proměnné definovaný jako

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y - \hat{y})^2 \tag{2.8}$$

Za předpokladu normálního rozdělení chybových členů se zkonstruuje konfidenční interval o $100(1-\alpha)\%$ spolehlivosti ve tvaru

$$b \pm t_{\alpha/2,n-2} \frac{\sqrt{s^2}}{\sqrt{\sum (x - \bar{x})^2}}, \tag{2.9}$$

kde $t_{\alpha/2,n-2}$ je hodnota Studentova t-rozdělení s $n-2$ stupni volnosti. Na základě p-hodnoty testu pak lze vyvrátit nulovou hypotézu a přijmout alternativní, že vysvětlující proměnná je dobrým prediktorem vysvětlované proměnné. [53]

Kvalitu modelu lze měřit také *koeficientem determinace* R^2 definovaným vztahem

$$R^2 = \frac{\sum (\hat{y}_t - \bar{y})^2}{\sum (y_t - \bar{y})^2}. \tag{2.10}$$

Model odpovídá skutečným hodnotám tím přesněji, čím blíže je R^2 blíže jedné. V případě regrese s více vysvětlujícími proměnnými se ale hodnota tohoto ukazatele zvyšuje vždy při přidání další vysvětlující proměnné. Proto nelze takový model hodnotit pouze na základě koeficientu determinace, aby nedošlo k jeho přeučení. [54]

2.3.3 Korelace

Dalším způsobem popsání míry lineární závislosti mezi dvěma náhodnými veličinami je korelační koeficient. Pro jeho definici je však nutné zavést nejprve *střední hodnotu* náhodné veličiny $\mathrm{E}X$, tedy typickou hodnotu, kolem níž se realizace náhodné veličiny pohybují. Střední hodnota je definována v závislosti na rozdělení veličiny, které je přiřazována. [51]

Mírou kolísání hodnot náhodné veličiny kolem střední hodnoty je pak *rozptyl (variance)*, jenž se definuje jako

$$\mathrm{var}X = \mathrm{E}(X - \mathrm{E}X)^2. \tag{2.11}$$

Odmocnina z rozptylu je pak *směrodatnou odchylkou* náhodné veličiny. [51]

Jsou-li definovány pro dvě náhodné veličiny X, Y rozptyly, zavádí se *kovariance* těchto veličin

$$\mathrm{cov}(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y). \tag{2.12}$$

Je-li kovariance nulová, pak jsou náhodné veličiny nezávislé. Pro porovnání kovariancí napříč dvojicemi náhodných veličin je ale vhodné hodnotu normalizovat na společný rozsah. Právě tak vzniká *korelační koeficient* definovaný jako

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}X\,\mathrm{var}Y}}. \tag{2.13}$$

Pro jeho hodnotu platí $-1 \leq \rho_{X,Y} \leq 1$ a čím blíže je v absolutní hodnotě jedné, tím silnější lineární závislost náhodných veličin potvrzuje. Korelace však neznamená, že mezi dvěma veličinami existuje příčinná souvislost – kauzalita. Pro účely analýzy korelace náhodných vektorů lze uvést, že pokud je hodnota kladná, nastává *pozitivní korelace* a čím jsou jednotlivé hodnoty jednoho vektoru větší, tím větší jsou i hodnoty druhého vektoru. Záporný korelační koeficient indikuje *negativní korelaci*, kdy jsou větší hodnoty jednoho vektoru asociovány s nižšími hodnotami vektoru druhého. [51]

2.3.4 Stacionarita

Proti běžnému vektoru náhodných veličin mají časové řady vlastnost vzájemné návaznosti jednotlivých hodnot, což přináší nové závislosti mezi hodnotami a znemožňuje použití některých metod kvůli vlivu na rozdělení.

Hlavním problémem, který vykazuje většina dat používaných v této práci, je absence stacionarity. Nejběžnějšími rozpoznávacími znameními nestacionarity je přítomnost trendu nebo sezónního opakování. Stacionární časová řada vykazuje stejné vlastnosti nezávisle na čase, ve kterém je pozorována. Tato vlastnost je definována pro časovou řadu y_t s minulými hodnotami $y_1, \ldots, y_n$. Stacionární je pak taková časová řada y_t, pro níž rozdělení $(y_t, \ldots, y_{t+s})$ pro všechna s nezáleží na t. [54]

Jednoduchým způsobem jak odstranit nestacionaritu je diferencování, při němž z původní časové řady vznikne řada rozdílů každé hodnoty v daném čase s hodnotou přímo předcházející. Diferencovaná časová řada se pak značí

$$y'_t = y_t - y_{t-1}. \tag{2.14}$$

Tuto transformaci lze provést i vícekrát, většinou však stačí k zajištění stacionarity provést nejvýše dvojitou diferenciaci. Diferencováním lze totiž stabilizovat střední hodnotu časové řady, jelikož dojde k odebrání trendu. Sezónní opakování o periodě m pak lze odebrat obdobně pomocí výrazu

$$y'_t = y_t - y_{t-m}. \tag{2.15}$$

Pro stabilizaci rozptylu je pak možné transformovat hodnoty logaritmizací. [54]

Zavedení nástroje používaného k ověření stacionarity vyžaduje definici autoregresního modelu. Jedná se o model vysvětlující hodnotu v daném čase pomocí p předchozích hodnot stejného procesu (řádu p). Označuje se pak jako AR(p) model s rovnicí

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{2.16}$$

kde ϕ_n jsou parametry a ε_t náhodné veličiny reprezentující chybu, často pak i.i.d. z rozdělení $N(0, \sigma^2)$. [54]

AR(p) model je pak stacionární, pokud pro největší kořen θ charakteristické rovnice

$$z^p = \alpha_1 z^{p-1} + \alpha_2 z^{p-2} + \cdots + \alpha_p z + \alpha_p \tag{2.17}$$

platí $|\theta| < 1$. Pokud je kořenem $\theta = 1$, modelovaná časová řada má jednotkový kořen a není stacionární. [55]

Problém pak nastává při pokusu o modelování regrese nestacionárních procesů. V případě dvou nezávislých časových řad y_t a x_t může OLS odhad regresní přímky dát pro dostatečně velký vzorek dat statisticky významný sklon. [56] Pokud například střední hodnoty dvou řad podléhají nějakému (ne nutně stejnému) trendu, regrese pak porovnává přítomnost trendu a ne změn ve směru vývoje obou procesů a je proto sporná.

Pro manuální rozpoznání nestacionární časové řady se často používá pohled na *graf autokorelací*. Autokorelace v bodě p má hodnotu korelačního koeficientu časové řady se zpožděním sebe sama o p hodnot. Pokud autokorelace se vzrůstajícím zpožděním klesá pomalu, proces lze označit za nestacionární. Jedná se však o nepříliš objektivní metodu založenou pouze na zkušenosti pozorovatele.

2.3.5 Dickey-Fuller test

Algoritmické rozpoznání stacionarity lze provést pomocí testů stacionarity, které se zaměřují na vybrané parametry. Často takovým parametrem je právě

přítomnost jednotkového kořene. Jednoduchým testem stacionarity je *Dickey–Fuller test (DF)*. Tento test používá pojem *náhodné procházky*, což je časová řada definovaná jako

$$y_t = y_{t-1} + \varepsilon_t. \tag{2.18}$$

DF test pak porovnává nulovou hypotézu, že proces odpovídá náhodné procházce

$$H_0 : x_t = x_{t-1} + \varepsilon_t \tag{2.19}$$

s alternativní hypotézou popisující stacionární $AR(1)$ model

$$H_A : x_t = c + \rho x_{t-1} + \varepsilon_t, \tag{2.20}$$

kde c a $|\rho| < 1$ jsou konstantní. Jelikož premisa nestacionarity je nulovou hypotézou (koeficient x_{t-1} je tehdy rovný 1), DF dává přednost označení dat za nestacionární a vyžaduje dostatečně silný vyvracející důkaz, aby H_0 zamítl. Tento předpoklad je v souladu s praktickým poznatkem, že mnoho časových řad vzniklých měřením procesů kolem nás je nestacionárních. [55]

DF se realizuje natrénováním modelu lineární regrese. Z hodnot vývoje časové řady $(x_1, \ldots, x_n)$ se získá vysvětlovaná proměnná $(x_2, \ldots, x_n)$ a vysvětlující proměnná $(x_1, \ldots, x_{n-1})$. Pro získaný model ve tvaru odpovídajícím rovnici 2.20 je získána t-statistika

$$\tau_\mu = \frac{\hat{\rho} - 1}{s_{\hat{\rho}}}, \tag{2.21}$$

kde $\hat{\rho}$ je odhad sklonu regresní přímky a $s_{\hat{\rho}}$ odhad jeho standardní odchylky. Podle tabelovaných hodnot se pak rozliší, jesli lze H_0 zamítnout. [55]

DF test ale existuje i v rozšířených variantách. Výše popsané hypotézy totiž předpokládjí, že vývoj časové řady nepodléhá deterministickému trendu. Často je ale trend přítomný a pro zvolení správné varianty je třeba rozlišit způsoby, jakými se trend na vývoji hodnot podílí. První možností je model náhodné procházky s konstatním členem posunu

$$x_t = c + x_{t-1} + \varepsilon_t, \tag{2.22}$$

kde ε_t je z $N(0, \sigma^2)$ a řada má jednotkový kořen, její hodnoty se s časem neomezeně zvyšují. Druhou možností je model odpovídající lineární regresi, který jednotkový kořen neobsahuje. Je proto možné použít DF obdobným způsobem jako v předchozím případě pro rozlišení typu trendu, který řada obsahuje a případnému zamítnutí nestacionarity. Jako nulová hypotéza se tehdy použije model odpovídající náhodné procházce s konstantou podle rovnice 2.22 a jako alternativní hypotéza druhý jmenovaný model. Používá se pak ale statistika τ_τ, který je závislá i na přidané vysvětlující proměnné, a to na čase. [55]

Odvozený od DF je *Augmented Dickey-Fuller test (ADF)*, který se od DF liší zvýšením řádu autoregresního modelu. Místo pouhých přímo předcházejících hodnot jako vysvětlující proměnné se v ADF používá p vysvětlujících

proměnných se zpožděním od 1 do p. Počet zpoždění se odhaduje minimalizací zvoleného informačního kritéria (AIC, BIC – popsáno níže). Stejně jako u základního DF je možné přidat člen kompenzující trend či konstantu posunu v závislosti na charakteru testované časové řady. Testový model včetně těchto kompenzačních členů je pak definován jako

$$\Delta y_t = \alpha_0 + \theta y_{t-1} + \beta t + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_p \Delta y_{t-p} + \varepsilon_t, \tag{2.23}$$

kde α_0 je konstanta a β je koeficient časově závislého trendu. Nulová hypotéza je pak

$$H_0 : \theta = 0 \tag{2.24}$$

a alternativní hypotéza

$$H_A : \theta < 0, \tag{2.25}$$

přičemž k vyvrácení H_0 se používá hodnota Dickey-Fuller t-statistiky z odhadu θ. Pokud se uplatní model bez členu časové závislého trendu a nulová hypotéza je vyvrácena, časová řada je stacionární. Použije-li se i člen trendu, časová řada je stacionární kolem trendu a stejně jako v případě obyčejného DF testu je třeba ji modelovat včetně trendu, diferencování nemá význam. [57]

2.3.6 KPSS test

Kwiatkowski–Phillips–Schmidt–Shin test (KPSS) je druhým běžně používaným testem stacionarity. Podle jeho tvůrců DF až příliš často nevyvrátí nulovou hypotézu přítomnosti nulového kořene a označuje tak mnoho časových řad za nestacionární. KPSS proto modeluje řadu jako součet deterministického trendu, náhodné procházky a stacionárního chybového členu

$$y_t = \beta t + r_t + \varepsilon_t, \tag{2.26}$$

kde ε_t je i.i.d. náhodná veličina z $N(0, \sigma^2)$. Nultý člen náhodné procházky r_0 je považován za intercept. Nulovou hypotézou je narozdíl od DF stacionarita kolem trendu, vyjádřená jako

$$H_0 : \sigma^2 = 0. \tag{2.27}$$

Alternativní hypotézou je pak přítomnost nulového kořene. Výklad výsledné p-hodnoty je tak oproti DF opačný. Pokud je $\beta = 0$, nulová hypotéza odpovídá stacionaritě kolem konstatní hodnoty. [58]

Pro vyhodnocení stacionarity vzhledem k rozdílným hypotézám KPSS a DF není možné tyto dva testy libovolně zaměňovat. Pokud na zvolené hladině KPSS vyvrací stacionaritu a DF naopak vyvrací nestacionaritu, časová řada odpovídá modelu popsanému rovnicí 2.22 a je tedy třeba ji diferencovat pro zajištění stacionarity. V opačném případě KPSS nevyvrací stacionaritu kolem trendu, kterou ADF bez příslušné kompenzace nerozpozná a z časové řady je nutné odečíst trend. [59]

2.3.7 Předpověď časových řad

Nejjednoduššími metodami předpovídání budoucího vývoje časové řady jsou předpovědi podle

- střední hodnoty,
- poslední známé hodnoty,
- poslední známé hodnoty z minulého sezónního opakování,
- přímky s krajními body počátku a konce pozorování (celkové změny).

Většinou je ale možné sestrojit lepší matematický model, podle něhož by bylo možné vývoj předpovědět. Tyto naivní metody proto slouží k porovnání – jakýkoli jiný model by měl být lepší než ony. [54]

Jednou z možností je nasadit na data lineární regresní model podle času či jiné vysvětlující proměnné, ale ta nemusí být vždy k dispozici. Podobným řešením, které řeší tento problém a z lineární regrese vychází, je proto *autoregrese*. Autoregresní model AR(p) používá k vysvětlení hodnoty v daném čase lineární kombinaci p předchozích hodnot vysvětlované proměnné. Definuje se vztahem 2.16. Parametry modelu se běžně omezují na modelování stacionárních časových řad, pro AR(1) tedy $|\phi_1| < 1$. [54]

Druhou možností je model *klouzavého průměru*. Narozdíl od autoregresního modelu nevyužívá přímo pozorované hodnoty, hodnoty vysvětluje na základě předešlých odchylek předpovědi, tedy

$$y_t = c + \varepsilon_y + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \cdots + \phi_q y_{t-q} \tag{2.28}$$

značený jako MA(q) čili model klouzavého průměru řádu q. Tento model by neměl být zaměňován s vyhlazováním procesu pomocí klouzavého průměru, které je používáno ke zviditelnění změny v trendu. Každý stacionární AR(p) model lze zapsat jako MA(∞) model. Postupnou substitucí pro AR(1) například takto:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_t \\ &= \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1\varepsilon_{t-1} + \varepsilon_t \\ &\vdots \end{aligned} \tag{2.29}$$

Díky podmínce stacionarity AR modelu se přitom ϕ_1 se zvyšující mocninou bude zmenšovat. Naopak pokud je $|\phi_1| > 1$, každá pozdější hodnota má na aktuální hodnotu větší a větší vliv. Je-li možné provést opačný proces k substituci uvedené výše, MA model je *invertibilní*. [54]

Zkombinováním obou uvedených modelů vznikne ARIMA model, kde I znamená integraci, inverzní proces k diferenciaci. ARIMA(p, d, q) model se definuje jako

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{2.30}$$

kde

- y' je diferencovaná časová řada,
- p je řád AR části modelu,
- q je řád MA části modelu,
- d je počet diferenciací časové řady.

Stejně jako u předchozích modelů je vyžadována stacionarita modelovaného procesu. AR, resp. MA model je možné zavést také jako speciální případ ARIMA modelu ve tvaru ARIMA$(p, 0, 0)$, resp. ARIMA$(0, 0, q)$. Vynechá-li se diferenciace, vznikne ARMA(p, q) model. [54]

Pro nalezení vhodných parametrů lze použít různé strategie jako MLE, Hannah–Rissanen či Yule–Walker s případnými omezeními pouze na AR či jiné speciální případy ARIMA modelu. Nejběžnější metrikou je pak MLE *(metoda maximální věrohodnosti)*, která volí takové parametry, pro něž by model s maximální pravděpodobností vracel hodnoty odpovídající původní časové řadě. Narozdíl od lineární regrese totiž nelze v obecném případě použít OLS vzhledem k přítomnosti odchylek minulých předpovědí jakožto vysvětlující proměnné, jejíž hodnota není známa. Metodu nejmenších čtverců je proto možné použít pouze pro AR model. [54, 60].

Stejně jako u ADF je i pro tento model třeba určit, kolik předchozích hodnot (a diferenciací) se má pro vysvětlování použít. K tomuto účelu se využívají různá informační kritéria. Jejich výhodou oproti přístupu pouze podle odchylky předpovídaných hodnot od skutečných je započítání informace o vnitřní složitosti modelu, čímž se předchází přeučení. [54]

Prvním z běžně užívaných kritérií je AIC *(Akaike information criterion)*, stavějící na informační teorii a souvisejícím přístupu, že žádný model není dokonalý, jedná se pouze o odhad skutečnosti. Kritérium je založené na Kullback–Leiblerově informační ztrátě, tedy vzdálenosti modelu od reality, která ale k výpočtu musí být známa a proto je nahrazena MLE. Další složkou je pak počet parametrů K v posuzovaném modelu. AIC v základní formě se zavádí jako

$$AIC = -2\log(L) + 2K, \tag{2.31}$$

kde L značí pravděpodobnost, že získaná data odpovídají skutečným. Model, pro nějž je hodnota kritéria nejnižší, je zvolen jako nejlepší. Pro modelování procesů s malým množstvím vzorků n se pak používá upravená verze AICc. Tu

je vhodné používat pro časové řady, kde $\frac{n}{K} <$ zhruba 40. Základní kritérium rozšiřuje následovně:

$$AIC = -2\log(L) + 2K + \frac{2K(K+1)}{n-K-1}. \tag{2.32}$$

[61]

Druhým kritériem je BIC *(Bayesian information criterion)* definovaný vztahem

$$BIC = -2\log(L) + K\log(n), \tag{2.33}$$

kde n je počet pozorování v časové řadě. Narozdíl od AIC tedy penalizuje složitost modelu víc se vzrůstajícím množstvím dat. Může tedy zvolit příliš jednoduchý model a místo přeučení dojde k nedostatečnému natrénování. [61]

Kapitola 3

Návrh

V této kapitole jsou představeny požadavky na výstupy práce. Vychází z nich členění architektury softwarového řešení. Součástí je také přehled použitých technologií. V druhé části kapitoly je čtenář seznámen s výběrem datových zdrojů použitých jako vstup pro software.

3.1 Softwarový návrh

3.1.1 Specifikace požadavků

Hlavním výstupem práce je ověření korelace vybraných datových sad s TFR a prezentace výsledků ve webové aplikaci. Aplikace má zároveň sloužit pro přehled historického a odhadovaného vývoje TFR.

Nejprve je proto nutné získat data. Software prochází při prvním spuštění či na explicitní žádost znovu později zadané datové zdroje, získá aktuální data pro jednotlivé ukazatele a uloží je v jednotném formátu. Zjistí-li v datech mezery, doplní je interpolací. Získaná data nemusí nutně pokrývat celé zadané období. Stejným způsobem jako u ostatních ukazatelů získá aplikace data i pro TFR.

Časové řady vývoje jednotlivých ukazatelů jsou následně podrobeny testu korelace s TFR. Pokud je datových bodů příliš málo, pouze se uloží k zobrazení uživatelem, ale jejich korelace se nevyhodnocuje. Výsledky výpočtů jsou rovněž uloženy, aby mohly být prezentovány uživatelovi.

Veškerá perzistentní data, jako jsou časové řady vývoje ukazatelů a výsledky výpočtů, se ukládají na společné úložiště s jasně definovanou strukturou. Tento požadavek splňuje relační databáze. Úložiště je přístupné jednak interně pro sběr a analýzu dat, jednak veřejně pro klienty prostřednictvím API. Klientem, který je součástí tohoto řešení, je webová aplikace.

Získané výsledky poté zobrazuje webová aplikace. Není potřeba, aby výpočty prováděla každá instance aplikace u každého uživatele sama. Zdrojová data jsou totiž pro všechny klienty stejná a obsah vizualizace také. Inter-

aktivita webové aplikace tedy spočívá v možnosti individuálního procházení dostupných dat. Tato data jsou vhodným způsobem prezentována. Pro časové řady je zobrazen graf vývoje a základní údaje, jako je název ukazatele, popis a měrná jednotka. Časové řady jsou hierarchicky uspořádány jednak podle regionu, jednak podle datového zdroje, z nějž pochází.

3.1.2 Komponenty softwarového systému

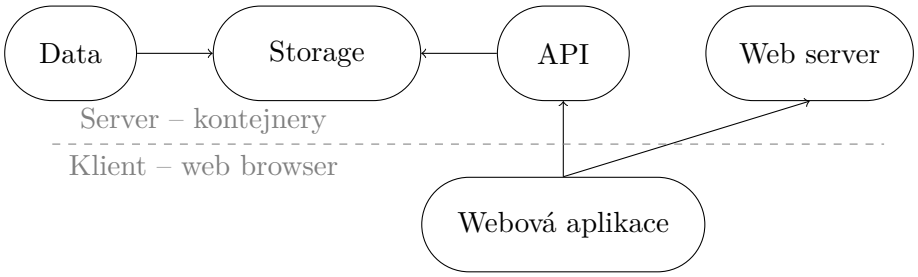
Ze specifikace požadavků vyplývá rozdělení softwaru na následující samostatné celky:

- Sběr dat a výpočty nad nimi *(Data)*
- Úložiště *(Storage)*
- Veřejné zpřístupnění úložiště *(API)*
- Webová aplikace *(App)*

Pro nasazení softwaru je použitý nástroj pro správu linuxových kontejnerů Docker. Mezi jeho hlavní výhody patří zjednodušení distribuce softwarových produktů a zlepšení spravovatelnosti díky vzájemné izolaci jednotlivých součástí. Docker pracuje s kontejnery, tedy vrstvenými souborovými systémy izolovanými od hostujícího operačního systému. Jelikož ale procesy běžící v kontejneru využívají jádro hostovského systému, kontejnery jsou méně prostorově náročné než virtuální stroje. [62]

Výše vyjmenované celky vyjma webové aplikace odpovídají jednotlivým kontejnerům, z nichž se software skládá. Vzájemnou datovou závislost těchto celků popisuje obrázek 3.1. Kontejnery *data* a *api* komunikují na lokální virtuální síti. Oba se připojují ke kontejneru *storage*. První tři kontejnery v seznamu spadají do části *backend*, tedy interních systémů přístupných veřejnosti jen pomocí API. Webová aplikace naopak tvoří *frontend*. Samotný kontejner pro ni je pak tvořený webovým serverem, který poskytuje kód aplikace klientům.

Obrázek 3.1: Závislosti softwarových celků



Orchestraci kontejnerů zajišťuje nástroj Docker compose. Citlivá data, jako jsou přístupová hesla, pomocí nichž se jednotlivé celky při vzájemné komunikaci autentizují, nejsou součástí konfiguračního souboru. Namísto toho jsou

předávána do docker compose pomocí proměnných prostředí (environment variables). Lze je tedy mít uložené odděleně s požadovanou úrovní zabezpečení.

Veškerý software a konfigurace vyjma výše popsaných citlivých dat jsou uložena a jejich vývoj je spravován pomocí aplikace Git [63]. Tento systém pro správu verzí umožňuje na úrovni řádků sledovat změny v textových souborech a anotovat je. Pro případ pozdějího rozvoje aplikace také je také užitečná jeho schopnost zpracovávat změny provedené více různými vývojáři a jejich práci synchronizovat. K tomuto účelu je využitý portál GitHub [64], kde je celý výstup praktické části práce dostupný pod open-source licencí.

3.1.3 Úložiště a jeho zpřístupnění

Pro uložení získaných datových sad a vypočtených informací o nich je použitá relační databáze PostgreSQL [65]. Jedná se o open source databázový stroj s více než třicetiletou historií. Poskytuje potřebné funkce jako jsou rozmanité datové typy, zachování integrity dat a dostatečnou rychlost [66].

Správu uživatelů a přistupových práv výhodně využívá projekt PostgREST [67], který z databáze PostgreSQL automaticky vytváří REST API. Databáze se tak stává „jediným zdrojem pravdy“, podle něhož API poskytuje data a spravuje oprávnění k přístupu. Pro aplikace, kde není potřeba složitá serverová logika mezi datovým úložištěm a klientem, je proto PostgREST vhodným nástrojem.

Pro oba softwarové celky jsou k dispozici oficiální obrazy dockerových kontejnerů [68, 69], které tato práce využívá.

3.1.4 Sběr dat a výpočty nad nimi

Modul pro sběr a zpracování dat je implementovaný v jazyce Python [70], který je vhodný zejména pro svou širokou nabídku knihoven. Pro předzpracování a analýzu dat nabízí balíček *pandas* [71]. Jeho součástí je datový typ tabulky `DataFrame` a vektoru `Series` podporující indexaci, anotované sloupce a transformační operace. Implementace statistických algoritmů včetně lineární regrese nabízí balíčky *scikit-learn* [72] či *SciPy* [73]. Propojení s databází zprostředkovává knihovna Psycopg [74], která umožňuje volat příkazy jazyka SQL na připojené instanci PostgreSQL. Python poskytuje také sadu oficiálních dockerových obrazů. Pro rychlý vývoj a možnost doplňování kódu o formátované poznámky je použit nástroj Jupyter Notebook [75].

3.1.5 Webová aplikace

Webová aplikace je implementovaná pomocí frameworku Flutter [76], který primárně vyvíjí pod open source licencí Google. Za pět let vývoje získal podporu běhu na všech majoritních platformách, včetně Androidu a iOS pro mobilní aplikace i Windows, linuxových distribucí a webu pro desktop. Umožňuje vývoj aplikací pro všechny tyto platformy pomocí jednoho kódu. Zejména

funkce hot reload, která obnoví uživatelské rozhraní i funkcionalitu podle změn v kódu bez nutnosti restartu aplikace a ztráty jejího stavu, je nápomocná pro rychlý vývoj.

Aplikace ve Flutteru jsou sestaveny z widgetů, tedy vícenásobně použitelných jednotek seskupujících jednotlivé části uživatelského rozhraní se související funkcionalitou. Pro správu stavu aplikace používá knihovnu Riverpod [77], která poskytuje framework k distribuci a zpracování dat mezi jednotlivými widgety.

Pro zpřístupnění aplikace uživatelům je využitý modul `http.server` dodávaný spolu s Python runtime. Tento server je v produkčním prostředí možné spolu s API skrýt za reverse proxy, které může přidat šifrovanou komunikaci pomocí HTTPS.

3.2 Datové zdroje

3.2.1 Sledované geografické oblasti

Geograficky jsou data omezená na území Evropské unie, popř. státy, které byly její součástí či s EU sousedí a spolupracují, jako je Spojené či Norské království. Důvodem ohraničení je blízkost států v této oblasti nejen z geografického, ale také z kulturního hlediska. Některé datové zdroje obsahují i data pro celosvětový průměr či průměr v rámci EU. Pro možnost porovnání se světovým vývojem TFR byly tyto datové sady také zahrnuty. V následujících tabulkách je pak počet států a průměrů, pro které jsou ukazatele dostupné souhrnně uváděn jako počet oblastí.

3.2.2 World Bank

World Bank Group je organizace sdružující 189 členských států za účelem poskytování finančních produktů a prostředků zejména rozvojovým zemím. Její součástí je Mezinárodní banka pro obnovu a rozvoj, vznikající v roce 1944 na podporu zemí zdevastovaných druhou světovou válkou a později přecházející k poskytování úvěrů v rozvíjejících se státech. V roce 1960 ji doplňuje Mezinárodní asociace pro rozvoj s cílem financovat rozvoj v nejchudších zemích. [78]

Pro podporu poskytování těchto služeb vznikla Development Data Group, která spravuje množství databází shromažďujících data ze statistických šetření ve členských zemích. Poskytuje rovněž asistenci při rozvoji sběru dat na národní úrovni. Výsledkem je kvalitní otevřená kolekce databází s mnoha indikátory dostupnými pro široký územní i časový rozsah. [79]

Pro účely této práce byla zvolena databáze World Development Indicators [80] poskytující data související s úrovní rozvoje jednotlivých států pod licencí *Creative Commons Attribution (CC-BY) 4.0*. Z této databáze byly vybrány především ekonomické ukazatele, o nichž existuje literatura dokazující kore-

Tabulka 3.1: Ukazatele z databáze World Bank

Název	Časový rozsah	Jednotka	Oblasti
Emise CO_2	1980–2018	kt	31
HDP per capita	1980–2020	mezinárodní dolar	31
Inflace	1980–2020	%	31
Participace mužů 15–24 let v pracovním procesu	1980–2020	%	31
Participace mužů 15+ v pracovním procesu	1980–2020	%	31
Participace osob 15–24 let v pracovním procesu	1980–2020	%	31
Participace osob 15+ v pracovním procesu	1980–2020	%	31
Participace žen 15–24 let v pracovním procesu	1980–2020	%	31
Participace žen 15+ v pracovním procesu	1980–2020	%	31
Podíl žen v parlamentu	1980–2020	%	31
Poměr participace žen vůči mužům v pracovním procesu	1980–2020	%	31
Sebevražednost	1980–2019	počet sebevražd	31
Total fertility rate	1980–2019	počet dětí	31
Úrok z vkladu	různý	%	10

laci. Vzhledem k dostupnosti dat od roku 1960 pro všechny zvolené geografické oblasti byla databáze také zvolena za zdroj časových řad vývoje TFR. Vybrané ukazatele popisuje tabulka 3.1.

Pro účely strojového zpracování je k dispozici API, pro niž existuje volně dostupná obalující knihova pro Python *world-bank-data* [81]. Tato knihovna obsahuje funkci, která na základě kódu ukazatele vrací jeho vývoj pro všechny dostupné státy či průměry napříč jimi v tabulce knihovny pandas.

3.2.3 Eurostat

Statistický úřad Evropské unie sbírá data o členských a vybraných dalších státech rozdělená na makroekonomické, obchodní, regionální, sociální statistiky a statistiky vládních financí [82]. Časové řady hodnot jednotlivých ukazatelů jsou většinou dostupné počínaje rokem vstupu daného státu do EU. Ukazatele nejsou vždy k dispozici pro všechny členské státy, totéž platí o evropském průměru. Výběr ukazatelů byl inspirován literaturou uvedenou v rešerši. Z ukazatelů, jejichž vliv na TFR tyto studie diskutují, se jedná zejména o státní

Tabulka 3.2: Ukazatele z databáze Eurostat

Název	Časový rozsah	Jednotka	Oblasti
Počet vězňů	2008–2019	počet vězňů	28
Podíl žen mezi vyučujícími prvního stupně	2013–2019	%	29
Podíl žen mezi vyučujícími středního vzdělání	2013–2019	%	27
Průměrný věk osamostatnění potomků – mužů	2000–2020	věk	28
Průměrný věk osamostatnění potomků – žen	2000–2020	věk	28
První sňatky mužů	1990–2019	počet sňatků	29
První sňatky žen	1990–2019	počet sňatků	29
Státní výdaje na vzdělávání	2012–2018	% HDP	29
Výdaje na podporu rodin	2008–2019	% soc. výdajů	29
Výdaje na podporu v nezaměstnanosti	2008–2019	% soc. výdajů	29
Výdaje na starobní důchody	2008–2019	% soc. výdajů	29
Výdaje na zdravotní péči	2008–2019	% soc. výdajů	29

výdaje na sociální zabezpečení a podporu. Celkový výčet předkládá tabulka 3.2, časový rozsah je uveden konkrétní pouze tehdy, pokud je podobný pro většinu časových řad. Data jsou přístupná pomocí aplikace Data Browser [83] a podléhají licenci podobající se již zmíněné CC-BY.

Stejně jako u databází World Bank existuje API pro automatizované získávání dat. Ta na základě kódu ukazatele vrací soubor ve formátu TSV, tedy hodnoty oddělené tabulátorem. Pro načtení dat a jejich úpravu do jednotného formátu je použita knihovna pandas.

3.2.4 NKOD

Národní katalog otevřených dat [84] je českým rejstříkem datasetů, které zpřístupňují státní instituce jako ministerstva či úřady, obce, školy a další. Otevřená data jsou podle § 3 odst. 11 zákona č. 106/1999 Sb. o svobodném přístupu k informacím *„informace zveřejňované způsobem umožňujícím dálkový přístup v otevřeném a strojově čitelném formátu, jejichž způsob ani účel následného využití není omezen a které jsou evidovány v národním katalogu otevřených dat.“* Datové sady použité v této práci jsou chráněny opět licencí CC-BY 4.0.

Vybrány byly především datové sady související s tématy zpracované literatury. Jejich výčet podává tabulka 3.3. Při hledání vhodných datových sad bylo zjištěno, že jejich časový rozsah a struktura je narozdíl od předcho-

Tabulka 3.3: Ukazatele z NKOD

Název	Časový rozsah	Jednotka
Dokončené byty	1997–2020	počet bytů
Medián mezd mužů	2011–2020	Kč
Medián mezd žen	2011–2020	Kč
Míra nezaměstnanosti mužů	1993–2021	%
Míra nezaměstnanosti žen	1993–2021	%
Počet důchodů	2008–2020	počet důchodů
Počet tříd	2007–2021	počet tříd

zích dvou zdrojů značně variabilní – pochází totiž z různých insititucí. Každá datová sada tedy byla individuálně vyfiltrována, případně agregována do požadovaného formátu za použití knihovny pandas.

3.2.5 Google Trends

Google Trends [85] poskytuje statistiku využívání vyhledávače Google. Zpracovává téměř surová data o vyhledáváních uskutečněných na celém světě. Statistika je anonymizovaná, rozdělená do kategorií podle tématu vyhledáváných klíčových slova a agregovaná podle času a geografické oblasti. Data jsou dostupná od roku 2004 dál.

Vzhledem k objemu vyhledávání je jejich vzorek používaný k tvorbě statistiky považovaný Googlem za dostatečně reprezentativní. Jeho redukce proti celkovému objemu dat umožňuje zpracování nových dat v řádu minut. Data jsou normalizovaná na geografické a časové úrovni, získaná časová řada sestává z čísel mezi 0 a 100, která vznikají jako přeškálovaný podíl sledovaného klíčového slova či tématu na celkovém počtu uskutečněných vyhledávání. Zkreslení může nastat jednak kvůli nemožnosti dokonale filtrovat vyhledávání uskutečněné automatizovaně, jednak protože při nízké proporci sledovaného klíčového slova je hodnota vynulována. Proto data nemusí být dostupná pro všechny geografické oblasti, když je zde frekvence vyhledávání příliš nízká. Do statistiky také nejsou zahrnuty vyhledávání obsahující některé speciální znaky. [86]

Experimentální studie spolehlivosti této služby [87] však nalezla velké rozdíly v datech získaných v různých časech na identický dotaz. Největší rozdíly byly patrné při dotazech na vývoj frekvence vyhledávání za jeden den po hodinách, kdy data někdy dokonce vykazovala opačný trend. Významné problémy s kvalitou dat byly nalezeny při analýze vývojů o délce osmi měsícu a kratších. Předpokládá se proto, že pro tuto práci, která využívá celou dostupnou délku dat, by s reprezentativností neměl být problém.

Pro účely této práce bylo použito seskupování souvisejících klíčových slov do témat. Předchází se tak nutnosti specifikovat klíčová slova v různých jazycích a variacích. Témata se týkají rodičovství a souvisejících faktorů, seznam je uveden v tabulce 3.4. Google Trends neposkytuje veřejnou API, ale pro Py-

Tabulka 3.4: Ukazatele z Google Trends

Antikoncepce	Hlídání dětí
Hypoteční kalkulačka	Kočárek
Kojenecká láhev	Kojenec
Mateřská škola	Podpora v nezaměstnanosti
Porodnice	Porod
Potrat	Rodičovská dovolená
Rodičovský příspěvek	Rozvod
Stres	Svatba
Těhotenský test	Těhotenství
Umělé oplodnění	Výpověď ze zaměstnání

thon existuje knihovna *pytrends*, která automatizuje získávání dat zpracováním webových stránek. Toto řešení trpí omezením možného počtu požadavků v krátkém čase.

Kapitola 4

Implementace

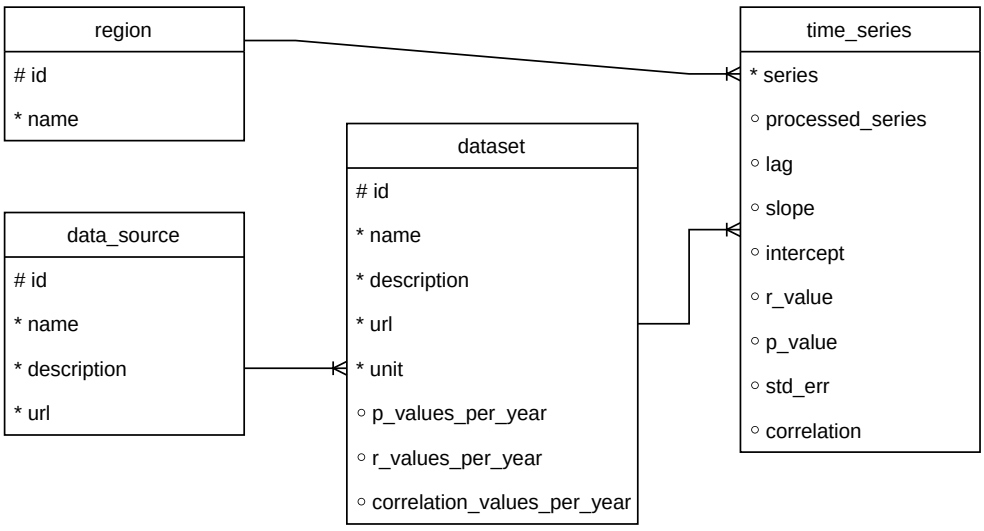
Tato kapitola popisuje proces implementace výše navrženého softwarového systému. Představuje detaily návrhu jednotlivých celků, které systém tvoří a diskutuje problémy, jež byly v průběhu vývoje objeveny a řešeny.

4.1 Databáze

4.1.1 Model

Na základě struktury datových sad ze zdrojů popsaných v minulé kapitole byl vytvořen relační model databáze, který je vložen na obrázku 4.1. Obsahuje jednak entity a pole pro informace získané přímo ze zdrojů, jednak vypočítané dříve popsanými metodami.

Obrázek 4.1: Koceptuální schéma databázových entit



region Tato entita reprezentuje stát nebo jinou geografickou oblast (např. celou EU), pro niž mohou být dostupné časové řady vývojů některých ukazatelů. Skládá se ze dvou polí:

- **id**: unikátní třípísmenný identifikátor podle normy ISO 3166-1 alpha-3,
- **name**: název regionu v češtině.

data_source Tato entita nese informace o datovém zdroji, z něhož pochází jednotlivé datové sady. Obsahuje následující pole:

- **id**: unikátní identifikátor datového zdroje,
- **name**: plný název,
- **description**: krátký popis o délce nejvýše jednoho odstavce,
- **url**: URL domovské webové stránky datového zdroje, je-li dostupná.

dataset Entita datové sady popisuje konkrétní ukazatel, který náleží právě jednomu datovému zdroji. Vzhledem k tomu, že jedním z výstupů výpočetního modulu je cross-sectional korelace, tedy korelace napříč regiony, obsahuje tato entita také výsledky vyhodnocování této korelace. Pokud ale ukazatel není dostupný v dostatečném počtu regionů a časovém rozsahu, pole pro výsledky výpočtů zůstanou prázdná. Entita je tvořena těmito položkami:

- **id**: unikátní identifikátor datové sady,
- **data_source**: unikátní identifikátor rodičovské datové sady,
- **name**: plný název,
- **description**: krátký popis o délce nejvýše jednoho odstavce,
- **url**: URL webové stránky datové sady, je-li dostupná,
- **unit**: jednotka, v níž jsou hodnoty časové řady měřeny,
- **p_values_per_year**: časová řada p-hodnot testu lineární závislosti pro cross-sectional korelaci,
- **r_values_per_year**: časová řada korelačních koeficientů pro cross-sectional korelaci,
- **correlation_values_per_year**: časová řada pravdivostních hodnot potvrzení lineární závislosti při cross-sectional korelaci.

time_series Tato entita reprezentuje jednotlivou časovou řadu, tedy realizaci vývoje daného ukazatele v konkrétním regionu. Obsahuje jednak surové hodnoty časové řady, jednak výsledky výpočtů obdobně, jako je to u datové sady. Výpočty se však liší, protože časová řada je porovnávána pouze s časovou řadou TFR.

- **dataset**: unikátní identifikátor datové sady,
- **region**: unikátní identifikátor regionu,
- **series**: časová řada hodnot ukazatele v souvisejícím regionu,
- **processed_series**: jednou diferencovaná instance téže časové řady,
- **lag**: zpoždění TFR, pro nějž byly nalezeny následující hodnoty,
- **slope**: sklon regresní přímky,
- **intercept**: intercept regresní přímky,
- **r_value**: korelační koeficient této časové řady a vývoje TFR,
- **p_value**: p-hodnota testu lineární závislosti týchž veličin,
- **std_err**: standardní odchylka sklonu regresní přímky,
- **correlation**: pravdivostní hodnota potvrzení testu lineární závislosti.

4.1.2 Pohled pro analýzu výsledků

Pro manuální analýzu výsledků vznikl pohled, který seskupuje časové řady podle ukazatele, k němuž náleží. Poskytuje zejména základní statistiky korelačních koeficientů porovnávajících jednotlivé časové řady s vývojem TFR a představu o tom, jaké znaménko má většina korelací. Časové řady filtruje podle p-hodnoty na 95% hladině. Skládá se z následujících polí:

- **dataset**: unikátní identifikátor datové sady,
- **p_avg**: průměr p-hodnot vybraných časových řad,
- **r_max**: nejvyšší hodnota korelačního koeficientu mezi vybranými řadami,
- **r_min**: nejnižší hodnota korelačního koeficientu mezi vybranými řadami,
- **n_positive**: počet vybraných řad s pozitivní korelací,
- **n_series**: celkový počet vybraných řad.

4.1.3 Inicializace

Výše popsaný model je třeba při tvorbě kontejneru databázového stroje inicializovat. K tomuto účelu jsou v Docker obraze připravena místa, kam je možné napojit buď soubor obsahující přímo výrazy jazyka SQL nebo Bash skript. Obojí je automaticky spuštěno při startu kontejneru, pokud je adresář se soubory databáze prázdný.

Pro účely této práce byla nejprve inicializace řešena pomocí SQL skriptu, ale po přidání API se ukázalo, že bude nutné vytvářet databázové role pro přístup API serveru k databázi. K vytvoření rolí je totiž potřebné zadat přihlašovací údaje a pro zachování bezpečnosti a automatizace nepřipadá v úvahu je psát přímo do inicializačního souboru manuálně. Inicializace proto probíhá pomocí Bash skriptu, který interpoluje proměnné prostředí poskytnuté prostřednictvím nástroje Docker compose.

4.2 API

REST rozhraní pro přístup klientů k databázi je tvořené automaticky. Jak již bylo nastíněno, nutná je však konfigurace rolí pro autentizaci API serveru a následnou autorizaci přístupu klientů. Role, jejímž prostřednictvím přistupuje API server k databázi, je zabezpečená heslem. Vzhledem k otevřené podstatě dat, které API poskytuje, je pro veškeré klientské požadavky používána role anonymního uživatele. Na úrovni databáze je pak tomuto anonymnímu uživateli dovoleno číst veškeré tabulky ve výše popsaném schématu.

4.3 Hledání korelujících ukazatelů

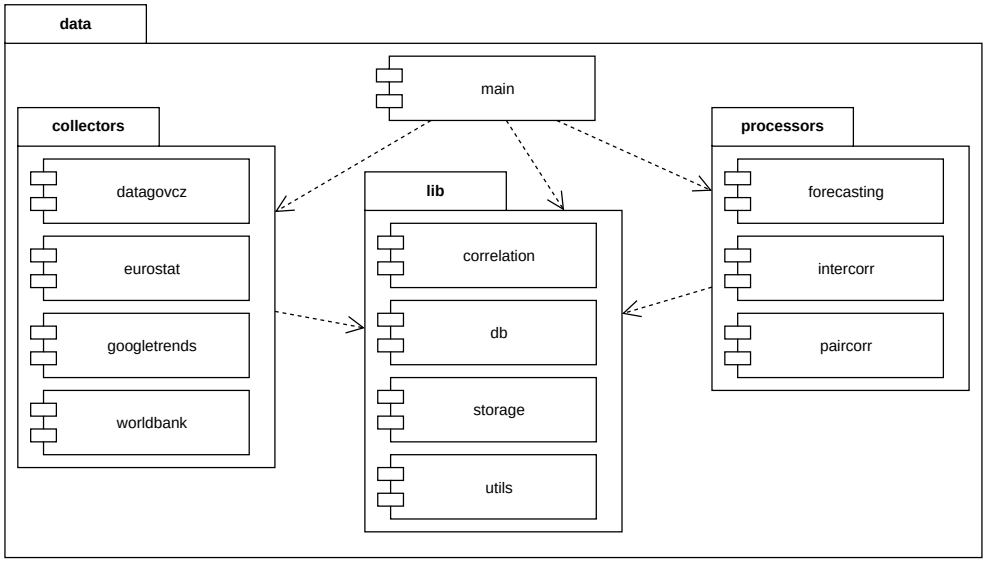
Celek pro sběr a zpracování dat je složený z několika dílčích balíčků, z nichž každý má jasně vymezenou zodpovědnost za každou část funkcionality. Architekturu na úrovni koncových modulů – zdrojových souborů Pythonu – popisuje včetně vnitřních závislostí obrázek 4.2.

4.3.1 Společná funkcionalita

Balíček společných tříd a funkcí používaných na více místech v kódu **lib** poskytuje jednak datový model a jeho persistenci, jednak nástroje k úpravě dat a vyhodnocování výsledků výpočtů. Skládá se z následujících modulů.

storage Obsahuje třídy zrcadlící entity databázového modelu, které jsou vzájemně propojené. Tyto třídy poskytují metody pro přístup k jednotlivým polím a kde je to vhodné, tam také jednochou logiku pro zápis výsledků výpočtů do instančních polí. Obalující třídou je stejnojmenná `Storage`, která drží odkazy na instance všech regionů a datových zdrojů. Struktura se od datových zdrojů dál větví v duchu databázového modelu.

Obrázek 4.2: Diagram balíčku pro sběr a zpracování dat



db Modul pro přístup k databázi obsahuje na míru vytvořené metody, jež serializují atributy objektů z předešlého modulu do příkazů jazyka SQL. Tyto příkazy prostřednictvím knihovy Psycopg pak odesílají databázovému serveru. Konfigurace připojení k databázi je realizována pomocí proměnných prostředí, které mohou být, jak již bylo popsáno, poskytnuty zvenčí. Přístup k databázi je implementován jako třída `Connection`, jejíž metoda `save_storage` jako argument přijímá instanci třídy `Storage` a rekurzivním průchodem jejích dceřiných atributů ukládá v souladu s integritními omezeními veškeré informace do databáze.

utils, correlation Jedná se o pomocné moduly vyčleňující funkce, které jsou v ostatním kódu použity vícekrát, v souladu s principem DRY (*don't repeat yourself*). Pro balíček **collectors** se jedná o funkci `strip_nans`, která ořízne od začátku a konce časové řady všechny body, jejichž hodnota je nedefinovaná. Kód balíčku **processors** pak používá funkci `is_correlation`. Ta na základě vstupní p-hodnoty vrací pravdivostní hodnotu, kterou klient interpretuje jako potvrzení lineární závislosti mezi dvěma ukazateli. Implementace této metody původně rozlišovala přítomnost závislosti i na absolutní hodnotě korelačního koeficientu. Od tohoto přístupu však bylo upuštěno, aby uživatel webové aplikace mohl interpretovat korelační koeficient zcela dle svého uvážení.

4.3.2 Získávání a předzpracování dat

Ukazatele pro statistické šetření jsou získávány ze zdrojů, které poskytují data v rozdílných formátech. Každý datový zdroj má proto vlastní modul, jež individuálně zpracovává vstupní data a vytváří z nich instance příslušných tříd modulu **storage**.

worldbank Datový zdroj World Bank poskytuje nejen oficiální API, ale dostupná je i neoficiální knihovna pro získávání dat o jednotlivých ukazatelích. Modul proto obsahuje definici ukazatelů včetně jejich identifikátoru podle World Bank a dalších pevných atributů jako popisu, které jsou použity pro tvorbu instance třídy datové sady. Dále modul definuje mapu kódů regionů, pro něž jsou získávány časové řady daných ukazatelů. V cyklu je pak použita funkce zmíněné knihovny, která do tabulky ve formě Pandas `DataFrame` nahraje časové řady každého ukazatele pro všechny dostupné regiony. Z tabulky jsou následně vybrány jednotlivé řady, dojde k odstranění krajních časových bodů s chybějící hodnotou a interpolaci vnitřních chybějících hodnot. Vzniklé instance datových tříd jsou uloženy do instance třídy `Storage`, kterou hlavní funkce modulu přijímá jako argument.

eurostat Data ze zdroje Eurostat jsou získávána obdobným způsobem jako z výše popsaného World Bank. Rozdíl je pouze v nutnosti přímého použití API, jež poskytuje datové tabulky ve formátu TSV. Načtení do instance `DataFrame` sice knihovna Pandas dokáže implicitně provést, ale vzhledem k nestandardní struktuře datových polí je třeba každou takovou tabulku ještě předzpracovat, než může být rozdělena na jednotlivé časové řady a ty podrobeny již popsané proceduře.

datagovcz NKOD se liší od předchozích datových zdrojů tím, že se jedná pouze o katalog a konkrétní datové sady poskytují přímo jejich tvůrci. Výsledkem je množina dat v často naprosto odlišných formátech. Každý ukazatel je proto zpracován individuálně, data musí být často manuálně agregována a kód je proto značně nesourodý.

googletrends Podobně jako pro World Bank je dostupná knihovna, která poskytuje datové sady jako objekty knihovny Pandas. Google Trends však nemá oficiální API a knihovna proto získává data z interních zdrojů, odkud data čerpá i oficiální webový klient. Následkem tohoto obcházení oficiálního klienta dochází často k zablokování požadavků ze strany Google, protože interní zdroje jsou limitované na počet požadavků za určitý časový interval. Proto byla do hlavního modulu **main**, který postupně výše popsané moduly spouští, přidána možnost vyřadit je definováním proměnné prostředí. Narozdíl od ostatních datových zdrojů Google Trends nabízí vývoj ukazatelů po měsících. Proto jsou všechny časové řady převzorkovány na roční frekvenci

průměrováním měsíčních hodnot, aby jejich frekvence souhlasila s ostatními. Zbytek implementace modulu pro stahování dat z Google Trends je obdobný jako u výše popsaných.

4.3.3 Korelace datových sad

Algoritmus hledání korelace byl nejprve vyvíjen pomocí nástroje Jupyter Notebook. Pro zjišťování korelace byly navrženy dva postupy. Jednak korelování dvojic časových řad pro stejný region, kdy jednou z nich byla řada vývoje TFR a druhou vývoj jiného ukazatele. Pohled na korelace daného ukazatele s TFR ve všech regionech pak umožňuje vyvodit závěry o jeho vlivu na plodnost napříč regiony. Druhým přístupem je tzv. cross-sectional korelace, tedy postup, který rovnou poskytuje ohodnocení souvislosti sledovaného ukazatele s TFR napříč regiony v rámci jediného výsledku a není proto třeba porovnávat manuálně více korelací dvojic jako u předešlého postupu.

4.3.3.1 Párová korelace

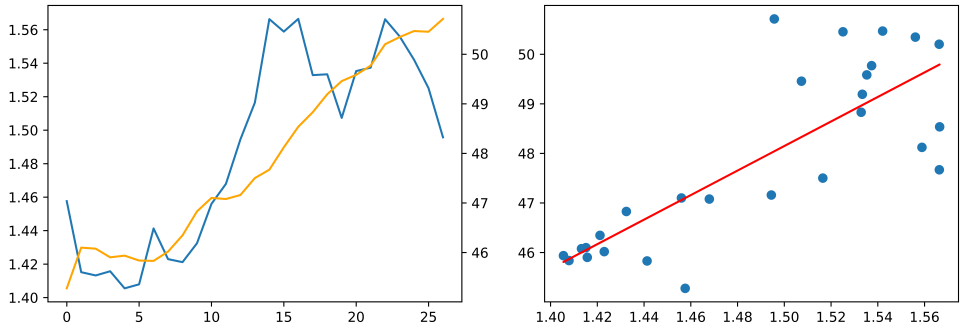
Pro nalezení optimálního posutupu ověřování párové korelace je vybrán vzorek dat z ukazatele World Bank. Toto omezení pouze na několik ukazatelů umožňuje rychlejší přepočítávání upraveného kódu, ale vzorek by měl být dostatečně reprezentativní.

Nejprve je ověřena stacionarita dat a jelikož většina jich podle vizuální inspekce podléhá nějakému trendu, který se případně kolem jednoho bodu láme, nepřekvapivým zjištěním je nestacionarita většiny časových řad.

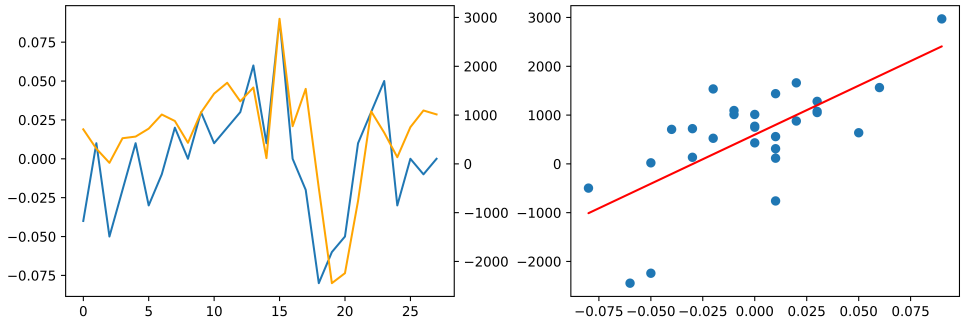
Pro stacionární řady je možné použít přímo lineární regresi a vypočítat korelační koeficient. Tuto funkcionalitu nabízí balíček SciPy. Jelikož ale často dochází k efektu zpoždění (*lag*) reakce jednoho socioekonomického ukazatele na druhý, je třeba porovnávat časové řady nejen v jejich surové podobě. TFR je proti druhému ukazateli opožďováno postupně do minulosti i budoucnosti a vybrány jsou výsledky s nejvyšší hodnotou korelačního koeficientu. Díky tomu uživatel posléze získá hodnoty v takovém relativním zpoždění, o němž existuje tvrzení o nejsilnější korelaci.

Výsledkem však jsou korelace jako na obrázku 4.3, která byla nalezena při zpoždění -4. Vlevo jsou vykresleny oba ukazatele (TFR v modré barvě) a vpravo pak dvojice jejich hodnot ve stejných časových okamžicích proložené regresní přímkou. Je patrné, že se porovnává spíše shoda v celkovém trendu, ale při zaměření na změny směru vývoje obou ukazatelů již souvislost není znát. Korelace je tedy evidentně pochybná, je totiž možné, že výřez časových řad vzniklých při jejich vzájemném opožďování již stacionární není, ačkoliv původní řada byla jako stacionární označena. Časové řady jsou proto diferencovány a výsledky se již jeví smyslupnými. Stejný postup je aplikován na řady, jež byly již v úvodu označeny jako nestacionární a pohled na původní hod-

Obrázek 4.3: Korelace participace žen 15+ v pracovním procesu s TFR v EU



Obrázek 4.4: Korelace diference HDP per capita s TFR v Řecku



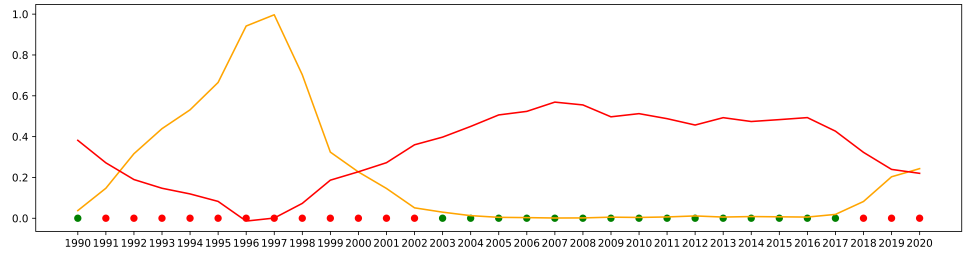
noty ukazatelů, pro něž bylo nalezeno významné tvrzení o korelaci, potvrzuje správnost přístupu. Příkladem výstupu jsou pak grafy na obrázku 4.4.

4.3.3.2 Korelace napříč regiony

K získání jediného grafu, který by popisoval vzájemnou závislost sledovaného ukazatele a TFR napříč regiony, je možné přistoupit více způsoby. První z nich je naivní spočítání průměru časových řad jednoho ukazatele za všechny regiony. Takto vzniklé průměry jsou pak opět porovnány s případnou diferenciací, je-li to nutné. Nevýhodou je však poměrně nízká vypovídací hodnota, výsledkem je totiž jediná korelace, vývoj její síly pro jednotlivé roky není znám.

Jako alternativní přístup se nabízí korelovat vektory, v nichž jednotlivé hodnoty odpovídají vývoji jednoho ukazatele v jednom časovém bodě pro všechny dostupné regiony. Výsledkem je tolik korelací, kolik mají všechny časové řady společných časových bodů. Díky tomu lze rozlišit období, kdy korelace napříč regiony byla silná, případně kdy došlo k jejímu obrácení z pozitivní na negativní či naopak. Je tak možné například ověřit hypotézu o změně vlivu FLP na TFR, jež byla zmíněna v rešerši. Výsledkem jsou grafy vývoje p-hodnot (oranžových) a korelačních koeficientů (červených) jako na obrázku

Obrázek 4.5: Korelace participace žen 15+ v pracovním procesu s TFR napříč regiony



4.5. Pro usnadnění čtení obsahuje body, jejichž barva je zelená, pokud p-hodnota odpovídá alespoň 95% hladině testu lineární závislosti.

4.3.4 Projekce TFR do budoucnosti

Na základě testu stacionarity byly časové řady vývoje TFR v jednotlivých regionech rozděleny. Pro stacionární časové řady se při testu stacionarity pomocí ADF jako nejlepší modely jeví AR(3), respektive ARMA(3, 0), oba bez korekce trendu, které jsou však ekvivalentní. Pro nalezení tohoto optima je použito testování pomocí dělení dat na trénovací a testovací množinu. Původní délka řady o 41 hodnotách je rozdělena na prvních 31 a zbylých 10. Na větší, trénovací množině je natrénován příslušný model a vytvořena předpověď na 10 let dopředu. Ta je následně porovnána se skutečnými hodnotami pomocí MSE. Optimalizované hyperparametry modelů zahrnují řád modelu a přítomnost členů pro kompenzaci trendu, z jejich kombinací je vytvořen několikarozměrný prostor, jehož všechny prvky jsou následně procházeny.

Pro všechny časové řady dohromady je pak vyzkoušen ARIMA model. Počet diferenciací je zvolen pomocí ADF, kdy je v případě nestacionarity řada diferencována a znovu otestována. Pokud je stále nestacionární, diferenciace proběhne ještě jednou. Ostatní hyperparametry jsou zvoleny minimalizací AIC. Následně se porovnává úspěšnost testů stacionarity, modely jsou ohodnoceny AIC a ukazuje se, že KPSS minimalizuje toto kritérium lépe. Pomocí takto nalezeného modelu je pak vývoj TFR předpovídán na 10 let dopředu.

4.4 Webová aplikace

4.4.1 Architektura

Pro implementaci byla zvolena architektura DDD (*domain driven development*), tedy vývoje podle funkčních domén. Tento přístup umožňuje škálování při přidávání funkcionality a zároveň zachovává organizaci kódu. Principem přístupu je rozdělení aplikace na jednotlivé domény zodpovědnosti, kdy je

každá doména složena z několika vrstev sahajících od uživatelského rozhraní až ke zpracování surových dat. Vrstvy jsou obvykle definovány následovně:

- **presentation**: Prezentační vrstva se stará o uživatelské rozhraní – přijímání požadavků od uživatele a zobrazování výstupů aplikace.
- **application**: Do aplikační vrstvy jsou odesílány požadavky uživatele, dochází zde ke zpracování aplikační logiky, která nezahrnuje nízkoúrovňovou práci s daty.
- **domain**: Doménová vrstva obsahuje třídy obalující surová data a dodávající integritní omezení.
- **infrastructure**: Vrstva infrastruktury interaguje s externím rozhraním posktujícím data.

Možný je také opačný přístup, kdy se aplikace dělí na popsané vrstvy a v rámci nich jsou rozlišeny funkční domény, nicméně dělení podle domén umožňuje vyčleňovat jednotlivé domény do samostatných balíčků a přenášet je snadněji napříč aplikacemi. Každá doména má totiž pevně dané rozhraní, pomocí něhož komunikuje s okolím.

Pro správu stavu aplikace je použito paradigma objektů typu *provider* v implementaci dříve popsané knihovny Riverpod. Tato knihovna umožňuje kdekoli v aplikaci (vzhledem k DDD však platí omezení na aplikační vrstvu) definovat instanci speciálního objektu, který poskytuje stav ve formě instance nějakého objektu nesoucího data. Změny v tomto stavu mohou být poslouchány libovolným prvkem uživatelského rozhraní, při jejich zaznamenání dojde k překreslení rozhraní podle nového stavu. Objekty typu provider mohou získávat stav z instancí, které si vytvoří, či které jsou globálně přístupné (např. z objektu obalujícího požadavky na API) anebo z jiných instancí typu provider. Zpracování stavu je tak možné řetězit a větvit bez opakování kódu.

4.4.2 Funkční domény

app Hlavní doména, která sdružuje jednotlivé bloky uživatelského rozhraní a definuje pomocí nich úvodní obrazovku aplikace. Poskytuje celé aplikaci základní kontext potřebný pro distribuci stavu. Obsahuje také logiku navigace v rozhraní.

data Tato doména se stará o zpracování dat a zobrazení detailních informací o nich. Obsahuje většinu funkcionality aplikace jelikož jedinou datovou doménou jsou datové zdroje a jejich datové sady.

- **presentation**: Prezentační vrstva obsahuje vykreslování grafů, informací o dostupných datech a sdružování těchto prvků do jednotlivých obrazovek.

- **application**: Aplikační vrstva poskytuje objekty typu provider, které zpřístupňují data prezentační vrstvě s ohledem na možné prodlevy a chyby požadavků na API.
- **domain**: Doménová vrstva obsahuje třídy zrcadlící doménový model databáze.
- **infrastructure**: Vrstva infrastruktury definuje třídu pro volání požadavků na API, odpovědi poté deserializuje a vrací instance z doménové vrstvy.

config Úlohou této domény je zobrazování rozhraní pro konfiguraci aplikace, poskytování aktuální konfigurace ostatním doménám a persistence tohoto stavu. Pro účely této aplikace se jedná o nastavení barevného schématu.

theming Pro zachování vizuální integrity grafických prvků aplikace definuje tato doména palety barev, sémanticky pojmenované rozměry a styly textů či některých ovládacích prvků.

4.4.3 Uživatelské rozhraní

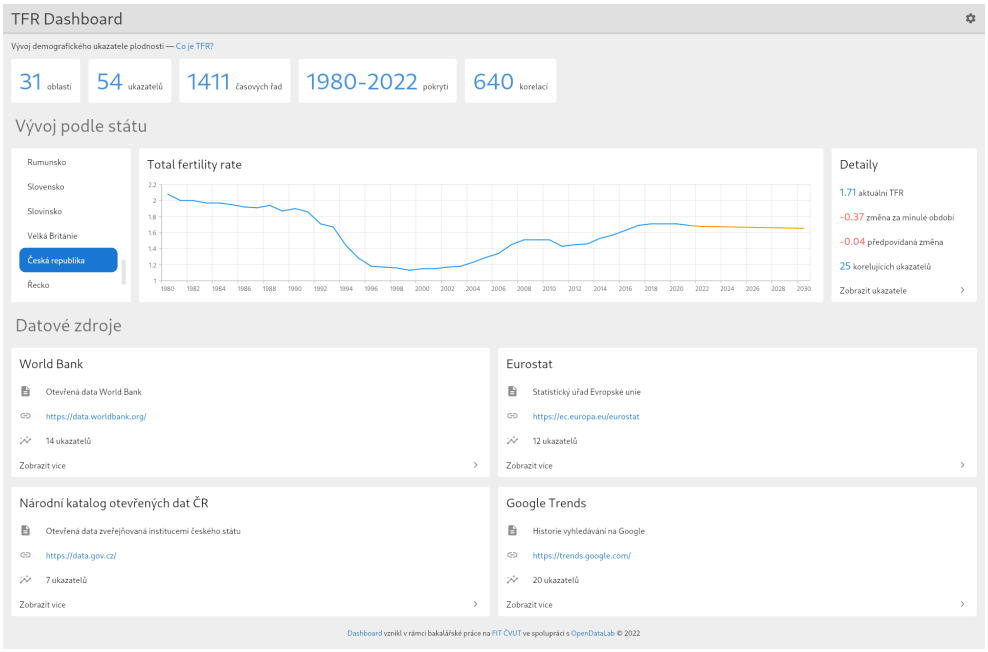
Rozhraní bylo navrženo s ohledem na maximální přístupnost dat vizualizovaných formou grafů. Cílem však zároveň je neodradit uživatele přehlcením rozhraní přílišným množstvím informací. Rozraní je proto vertikálně děleno nadpisy do jednotlivých sekcí a důraz je kladen na postup od obecných informací ke konkrétním ve směru čtení textu. Následuje popis jednotlivých obrazovek aplikace.

Domovskou obrazovku na obrázku 4.6 charakterizuje graf vývoje TFR, který nabízí okamžitý přehled situace ve zvolené oblasti. Základní statistiky o počtu dat a jejich rozsahu zobrazuje pruh karet s čísly v úvodu. Nechybí vysvětlení pojmu TFR, jehož detail je skrytý pod odkazem. Uživatel má dvě možnosti, jak se dostat ke grafům dalších ukazatelů. Buď si nechá zobrazit ukazatele korelující s TFR v aktuálně zvolené oblasti, nebo přejde na detail datového zdroje.

Detail datového zdroje 4.7 zobrazuje popisné informace o zdroji a nabízí seznam příslušných ukazatelů. Po zvolení jednoho z nich se zobrazí přehled včetně časové řady pro region, který byl zvolen na domovské obrazovce. Podle dostupných informací se uživatel může rozhodnout pokračovat na detail některého ukazatele.

Detail ukazatele 4.8 nabízí dělení časových řad podle oblasti na skupinu těch, pro něž bylo nalezeno silné tvrzení o korelaci a ostatních. Hlavním elementem je graf cross-sectional korelace, který narozdíl od grafu používaného při vývoji algoritmu zvýrazňuje období s významnou korelací podbarvením grafu v primární barvě, jež je napříč aplikací užívána vždy při vykreslování

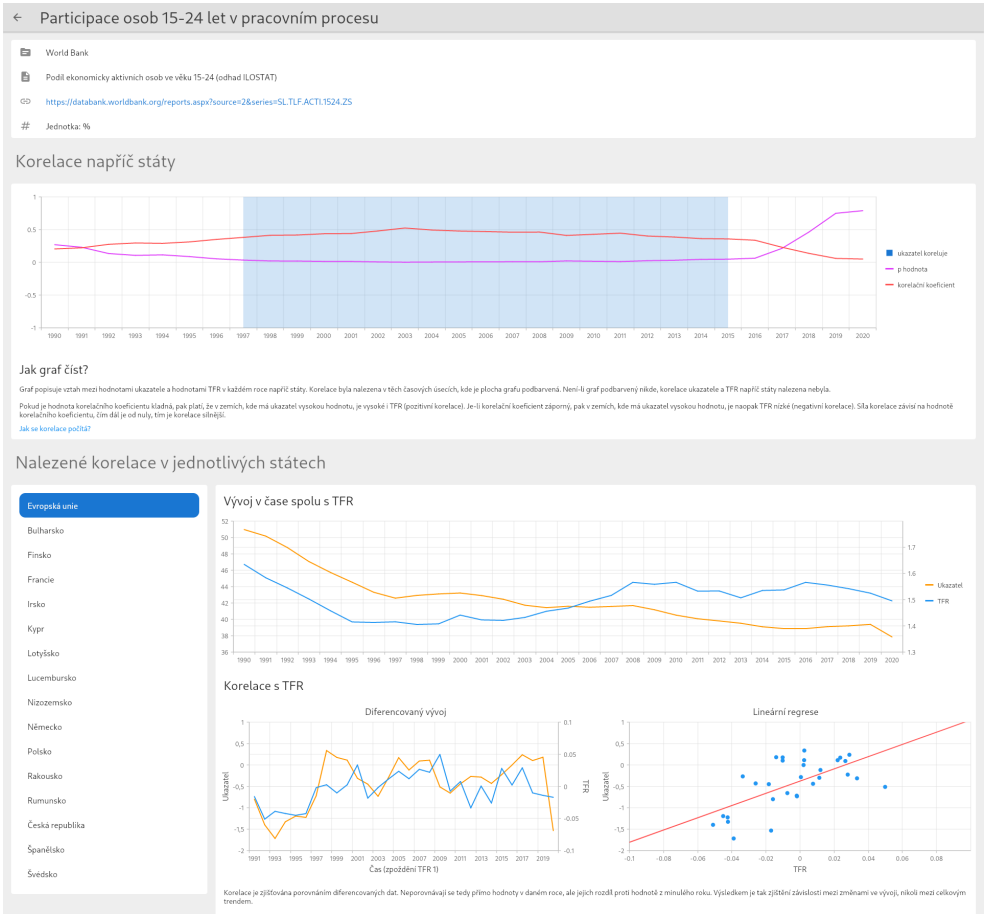
Obrázek 4.6: Domovská obrazovka webové aplikace



Obrázek 4.7: Detail datového zdroje ve webové aplikaci

TFR. Barevná konzistence platí i pro ostatní ukazatele. Na této obrazovce je také dostupný podrobný popis jednotlivých veličin, aby uživatel mohl učinit závěr o zobrazovaných datech.

Obrázek 4.8: Detail ukazatele ve webové aplikaci, zdola oříznuto



Kapitola 5

Vybrané výsledky analýzy

Pomocí modulu pro získávání a analýzu dat bylo shromážděno 54 ukazatelů. Časové řady jejich vývoje jsou dostupné dohromady v 29 státech, případně i pro průměr v rámci EU a celého světa. Data pokrývají období mezi lety 1980 a 2022 a bylo v nich nalezeno 640 dvojic, pro něž je dostupné významné tvrzení o korelaci. Jelikož toto samotné tvrzení nevypovídá o síle vztahu mezi ukazatelem a TFR, tabulka 5.1 poskytuje konkrétnější přehled. Jsou v ní sestupně seřazené ukazatele podle počtu jejich časových řad, pro něž bylo nalazeno významné tvrzení o korelaci na 95% hladině. Zároveň je k dispozici počet těch z nich, jejichž korelační koeficient je kladný a také maximální a minimální přítomná hodnota korelačního koeficientu. Data jsou omezena na prvních několik ukazatelů, celá tabulka je k dispozici v příloze práce spolu s vybranými grafy.

Tabulka 5.1: Nejčastější ukazatele se silným tvrzením o korelaci s TFR

Ukazatel	Min	Max	Kladných	Celkem
Participace osob 15+ v pracovním procesu	-0.58	0.73	15	20
Participace mužů 15+ v pracovním procesu	-0.55	0.62	14	19
Mateřská škola – Téma	-0.80	0.77	13	19
První sňatky mužů	-0.49	0.81	13	18
První sňatky žen	-0.52	0.82	13	18
Kočárek – Téma	-0.82	0.68	8	18
Participace žen 15+ v pracovním procesu	-0.54	0.77	11	17
Umělé oplodnění – Téma	-0.75	0.74	7	17
Výdaje na zdravotní péči	-0.93	0.70	7	17
Stres – Téma	-0.84	0.69	4	17
HDP per capita	-0.42	0.69	14	16

V souladu s prezentovanou literaturou je nalezeno velké množství korelací týkajících se zaměstnanosti. Objevují se také témata z Google Trends, přestože překvapivě například téma „Kočárek“ koreluje spíše negativně, což by bylo možné přičítat nekvalitnímu vzorku, z něhož byla data získáná. Naopak u tématu „Umělé oplodnění“ převažující negativní korelace není tak překvapivá, čím více párů hledá tuto proceduru, tím méně dětí se pravděpodobně má šanci narodit, přestože bylo možné očekávat také pozitivní korelaci se zpožděním, když se nakonec umělé oplodnění zdaří. Literatuře odpovídá také majoritní podíl pozitivních korelací prvních sňatků obou pohlaví, který je dobře vysvětlitelný. Stejně tak HDP koreluje v drtivé většině případů pozitivně. Z některých dalších významných korelací lze uvést ještě pozitivní asociaci tématu „Těhotenství“ z Google Trends, která se jeví smysluplnou, či relativně nerozhodný směr korelace státních výdajů v nezaměstnanosti, jenž ale uvedené literatuře spíše odporuje – očekávaná byla spíše majorita negativních korelací. Dobře vysvětlitelná je ale například velmi častá negativní korelace tématu „Antikoncepce“ na Google Trends, která prodlužuje do 21. století vliv antikoncepce na snižování plodnosti projevující se ve století minulém.

Mezi nalezené cross-sectional korelace patří dobře vysvětlitelné vyhledávání „Hlídání dětí“ na Google, rodiče tuto službu nejen kvůli pracovnímu vytížení potřebují. Počet prvních sňatků koreluje pozitivně pouze v první polovině devadesátých let a později už se souvislost napříč státy významně rozchází, což stojí v kontrastu s pozitivními korelacemi uvnitř jednotlivých států. Naopak průměrný věk, kdy ženy opouštějí rodičovskou domácnost, s plodností koreluje negativně prakticky po celou dobu pozorování, tj. mezi lety 2002–2020. Toto zjištění je dobře vysvětlitelné založením vlastní rodiny ve chvíli, kdy žena začne žít ve vlastní domácnosti. Zaměstnanost mužů koreluje pozitivně rovněž v majoritním intervalu pozorování, naopak zaměstnanost žen tuto asociaci vykazuje jen v krátkém úseku, což potvrzuje jednu z názorových polovin studií prezentovaných v rešerši.

Výsledky jsou tedy ve velké míře podpořeny stávající literaturou. Nové korelace jsou většinou dobře interpretovatelné, výjimky lze přičítat rozdílné metodice výzkumů v literatuře a v této práci. Přesto práce nabízí velké množství nových dat, jejichž možné příčinné souvislosti s plodností by bylo prospěšné odhalit.

Závěr

Hlavním výstupem práce je tvorba webové aplikace, jež umožňuje zobrazit jednak historii a předpověď vývoje demografického ukazatele plodnosti (TFR), jednak oveřené i nově nalezené související ukazatele. Aplikace je svižná, poskytuje rychlý přístup k potřebným datům a vizualizuje výsledky uživatelsky přívětivým způsobem. Její architektura přitom umožňuje přistupovat k datům i strojově a navázat tak na výsledky této práce.

Práce přináší přehled odborné literatury na téma veličin souvisejících s TFR a na jejím základě vybírá nové datové zdroje. World Bank a Eurostat jsou poměrně tradičními zdroji užívanými i ve zmíněné literatuře, naopak využití otevřených dat a statistik vyhledávání na internetu zatím nemá takové zastoupení ve výzkumu. Aplikace, jež je součástí této práce, tak rozšiřuje v dnešní době vznikající kategorii veřejně dostupných vizualizací dat s přidanou hodnotou.

Z navržených metod pro hledání souvislostí mezi získanými ukazateli a TFR je nalezeno optimální řešení sestávající z vyhodnocování korelací diferencí časových řad vývojů ukazatelů. Pro předpověď vývoje TFR na deset let dopředu jsou využity ARIMA modely, jejichž hyperparametry se vybírají automaticky pro každou časovou řadu zvlášť za použití vybrané metriky.

Výstupy, které práce přináší, z většiny potvrzují výzkum ve stávající litaratuře. Objevují se však některá nová zjištění jako například výrazná negativní korelace TFR s četností vyhledávání klíčových slov souvisejícím se stresem na Google anebo napříč různými státy nekonzistentně polarizovaná korelace plodnosti s počtem vězňů. Nevysvětlených závislostí však zůstává více a práce proto poskytuje podnět pro další výzkum na poli ovlivňování plodnosti ve společnosti, které je v dnešní době stále významnějším tématem.

Literatura

[1] Pavlík, Z.; Rychtaříková, J.; Šubrtová, A.: *Základy demografie*. Praha: Academia, první vydání, 1986.

[2] Lee, R.: The Demographic Transition: Three Centuries of Fundamental Change. *Journal of Economic Perspectives*, ročník 17, 02 2003: s. 167–190, doi:10.1257/089533003772034943.

[3] de Silva, T.; Tenreyro, S.: Population Control Policies and Fertility Convergence. *The Journal of Economic Perspectives*, ročník 31, č. 4, 2017: s. 205–228, ISSN 08953309. Dostupné z: `http://www.jstor.org/stable/44425388`

[4] Evan, T.; Vozárová, P.: Influence of women's workforce participation and pensions on total fertility rate: a theoretical and econometric study. *Eurasian Economic Review*, 07 2017, doi:10.1007/s40822-017-0074-0.

[5] Fanti, L.; Gori, L.: Fertility-related pensions and cyclical instability. *Journal of Population Economics*, ročník 26, 07 2013, doi:10.1007/s00148-012-0462-4.

[6] Calwell, J.; Caldwell, P.; Mcdonald, P.: Policy Responses to Low Fertility and Its Consequences: A Global Survey. *Journal of Population Research*, ročník 19, 03 2002: s. 1–24, doi:10.1007/BF03031966.

[7] Clements, B.; Dybczak, K.; Gaspar, V.; aj.: The Fiscal Consequences of Shrinking and Ageing Populations. *Ageing International*, ročník 43, 12 2018: s. 1–24, doi:10.1007/s12126-017-9306-6.

[8] Galasso, V.; Gatti, R.; Profeta, P.: Investing for the old age: Pensions, children and savings. *International Tax and Public Finance*, ročník 16, 08 2009: s. 538–559, doi:10.1007/s10797-009-9104-5.

[9] Omori, T.: Effects of public education and social security on fertility. *Journal of Population Economics*, ročník 22, 07 2009: s. 585–601, doi: 10.1007/s00148-009-0244-9.

[10] Gahvari, F.: Pensions and fertility: In search of a link. *International Tax and Public Finance*, ročník 16, 08 2009: s. 418–442, doi:10.1007/s10797-009-9114-3.

[11] Cremer, H.; Pestieau, P.; Gahvari, F.: Pensions with Heterogenous Individuals and Endogenous Fertility. *Journal of Population Economics*, ročník 21, 02 2008: s. 961–981, doi:10.1007/s00148-006-0114-7.

[12] Cigno, A.: How to Avoid a Pension Crisis: A Question of Intelligent System Design. *CESifo Economic Studies*, ročník 56, 03 2009, doi: 10.1093/cesifo/ifp024.

[13] Fenge, R.; Scheubel, B.: Pensions and fertility: back to the roots: Bismarck's Pension Scheme and the first demographic transition. *Journal of Population Economics*, ročník 30, 08 2016, doi:10.1007/s00148-016-0608-x.

[14] Regős, G.: Can Fertility be Increased With a Pension Reform? *Ageing International*, ročník 40, 05 2015: s. 117–137, doi:10.1007/s12126-014-9206-y.

[15] Ehrlich, I.; Kim, J.: Has Social Security Influenced Family Formation and Fertility in OECD Countries? An Economic and Econometric Analysis. *NBER Working Paper Series*, 2007.

[16] Behrman, J.; Gonalons-Pons, P.: Women's employment and fertility in a global perspective (1960–2015). *Demographic Research*, ročník 43, 2020: s. 707–744, ISSN 14359871, 23637064.

[17] Mishra, V.; Nielsen, I.; Smyth, R.: On the relationship between female labour force participation and fertility in G7 countries: Evidence from panel cointegration and Granger causality. *Empirical Economics*, ročník 38, 04 2009: s. 361–372, doi:10.1007/s00181-009-0270-1.

[18] Kato, H.: Does a relationship between fertility and labor participation of women really exist? Perspectives from time series analysis. *International Journal of Economic Policy Studies*, ročník 14, 02 2020: s. 3–23, doi: 10.1007/s42495-020-00033-2.

[19] Salamaliki, P.; Venetis, I.; Giannakopoulos, N.: The causal relationship between female labor supply and fertility in the USA: Updated evidence via a time series multi-horizon approach. *Journal of Population Economics*, ročník 26, 01 2013: s. 109–145, doi:10.1007/s00148-012-0418-8.

[20] Macunovich, D. J.: Fertility and the Easterlin Hypothesis: An Assessment of the Literature. *Journal of Population Economics*, ročník 11, č. 1, 1998: s. 53–111, ISSN 09331433, 14321475.

[21] McNown, R.; Rajbhandary, S.: Time series analysis of fertility and female labor market behavior. *Journal of Population Economics*, ročník 16, 02 2003: s. 501–523, doi:10.1007/s00148-003-0107-8.

[22] Rindfuss, R.; Guzzo, K.; Morgan, S.: The Changing Institutional Context of Low Fertility. *Population Research and Policy Review*, ročník 22, 12 2003: s. 411–438, doi:10.1023/B:POPU.0000020877.96401.b3.

[23] Adsera, A.: Changing Fertility Rates in Developed Countries. The Impact of Labor Market Institutions. *Journal of Population Economics*, ročník 17, 02 2004: s. 17–43, doi:10.1007/s00148-003-0166-x.

[24] Engelhardt, H.; Prskawetz, A.: On the Changing Correlation Between Fertility and Female Employment Over Space and Time. *European Journal of Population*, ročník 20, 01 2003, doi:10.1023/B:EUJP.0000014543.95571.3b.

[25] Kögel, T.: Did the Association between Fertility and Female Employment within OECD Countries Really Change Its Sign? *Journal of Population Economics*, ročník 17, 02 2004: s. 45–65, doi:10.1007/s00148-003-0180-z.

[26] Oshio, T.: Is a positive association between female employment and fertility still spurious in developed countries? *Demographic Research*, ročník 41, 2019: s. 1277–1288, ISSN 14359871, 23637064.

[27] Harknett, K.; Billari, F.; Medalia, C.: Do Family Support Environments Influence Fertility? Evidence from 20 European Countries. *European Journal of Population*, ročník 30, 02 2014: s. 1–33, doi:10.1007/s10680-013-9308-3.

[28] Bjorklund, A.: Does Family Policy Affect Fertility? Lessons From Sweden. *Journal of Population Economics*, ročník 19, 02 2006: s. 3–24, doi:10.1007/s00148-005-0024-0.

[29] Mcnown, R.; Ridao-Cano, C.: The Effect of Child Benefit Policies on Fertility and Female Labor Force Participation in Canada. *Review of Economics of the Household*, ročník 2, 02 2004: s. 237–254, doi:10.1007/s11150-004-5646-6.

[30] Yakita, S.: Fertility, child care policy, urbanization, and economic growth. *Letters in Spatial and Resource Sciences*, ročník 12, 04 2019, doi:10.1007/s12076-019-00226-0.

[31] Sato, Y.; Yamamoto, K.: Population concentration, urbanization, and demographic transition. *Journal of Urban Economics*, ročník 58, č. 1, 2005: s. 45–61, ISSN 0094-1190, doi:https://doi.org/10.1016/j.jue.2005.01.004. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0094119005000057

[32] Riphahn, R.; Wiynck, F.: Fertility effects of child benefits. *Journal of Population Economics*, ročník 30, 10 2017, doi:10.1007/s00148-017-0647-y.

[33] Schellekens, J.: Family Allowances and Fertility: Socioeconomic Differences. *Demography*, ročník 46, 08 2009: s. 451–68, doi:10.1353/dem.0.0067.

[34] Son, Y.: Do childbirth grants increase the fertility rate? Policy impacts in South Korea. *Review of Economics of the Household*, ročník 16, 09 2018, doi:10.1007/s11150-017-9383-z.

[35] Erlandsson, A.: Child Home Care Allowance and the Transition to Second- and Third-Order Births in Finland. *Population Research and Policy Review*, ročník 36, 08 2017: s. 1–24, doi:10.1007/s11113-017-9437-1.

[36] Pacalova, H.; Engelhardt, H.; Morgan, S.; aj.: Do Cross-National Differences in the Costs of Children Generate Cross-National Differences in Fertility Rates? *Population Research and Policy Review*, ročník 22, 12 2003, doi:10.1023/B:POPU.0000020961.89068.91.

[37] Gauthier, A.: The Impact of Family Policies on Fertility in Industrialized Countries: A Review of the Literature. *Population Research and Policy Review*, ročník 26, 02 2007: s. 323–346, doi:10.1007/s11113-007-9033-x.

[38] Götmark, F.; Andersson, M.: Human fertility in relation to education, economy, religion, contraception, and family planning programs. *BMC Public Health*, ročník 20, č. 1, 2020: s. 1–17, ISSN 14712458. Dostupné z: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=142383499&site=ehost-live&scope=site

[39] Bratti, M.: Labour force participation and marital fertility of Italian women: The role of education. *Journal of Population Economics*, ročník 16, 08 2003: s. 525–554, doi:10.1007/s00148-003-0142-5.

[40] Chen, I.-C.: Parental Education and Fertility: An Empirical Investigation Based on Evidence from Taiwan. *Journal of Family and Economic Issues*, ročník 37, 05 2015, doi:10.1007/s10834-015-9448-1.

[41] Kravdal, .: Main and Interaction Effects of Women's Education and Status on Fertility: The Case of Tanzania. *European Journal of Population*, ročník 17, 06 2001: s. 107–135, doi:10.1023/A:1010725506916.

[42] Hess, P.: Public policy and the total fertility rate: cross-sectional evidence from the LDCs. *Population Research and Policy Review*, ročník 5, 01 1986: s. 253–275, doi:10.1007/BF00136787.

[43] Mauldin, W.; Berelson, B.; Sykes, Z.: Conditions of Fertility Decline in Developing Countries, 1965-75. *Studies in family planning*, ročník 9, 06 1978: s. 90–147, doi:10.2307/1965523.

[44] Doepke, M.: Accounting for Fertility Decline During the Transition to Growth. *Journal of Economic Growth*, ročník 9, 09 2004: s. 347–383, doi: 10.2139/ssrn.279519.

[45] Madsen, J. B.; Islam, M.; Tang, X.; aj.: Was the post-1870 fertility transition a key contributor to growth in the West in the twentieth century? *Journal of Economic Growth*, ročník 25, č. 4, 2020: s. 431–454.

[46] Ahituv, A.: Be Fruitful or Multiply: On the Interplay Between Fertility and Economic Development. *Journal of Population Economics*, ročník 14, 05 2001: s. 51–71, doi:10.1007/s001480050159.

[47] Martine, G.; Alves, J. E.; Cavenaghi, S.: Urbanization and fertility decline: Cashing in on structural change. Technická zpráva, International Institute for Environment and Development, 2013. Dostupné z: http://www.jstor.org/stable/resrep01293

[48] Jaffe, A. J.: Urbanization and fertility. *American Journal of Sociology*, ročník 48, č. 1, 1942: s. 48–60.

[49] Guo, Z.; Wu, Z.; Schimmele, C. M.; aj.: The effect of urbanization on China's fertility. *Population Research and Policy Review*, ročník 31, č. 3, 2012: s. 417–434.

[50] Vinci, S.; Egidi, G.; Salvia, R.; aj.: Natural population growth and urban management in metropolitan regions: Insights from pre-crisis and post-crisis Athens, Greece. *Urban Studies*, 2021.

[51] Zvára, K.; Štěpán, J.: *Pravděpodobnost a matematická statistika*. Praha: Matfyzpress, 1997, ISBN 80-85863-24-3.

[52] Kalu, E.; Kaw, A.; Nguyen, C.: Linear Regression. [online], srpen 2012, [cit. 2022-05-07]. Dostupné z: https://nm.mathforcollege.com/mws/gen/06reg/mws_gen_reg_txt_straightline.pdf

[53] Hartmann, K.; Krois, J.; Waske, B.: E-Learning Project SOGA: Statistics and Geospatial Data Analysis. [online], 2018, [cit. 2022-05-07]. Dostupné z: https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Hypothesis-Tests/Inferential-Methods-

in-Regression-and-Correlation/Inferences-About-the-Slope/index.html

[54] Hyndman, R. J.; Athanasopoulos, G.: *Forecasting: principles and practice*. Melbourne, Australia: OTexts, třetí vydání, 2021, [cit. 2022-05-07]. Dostupné z: https://otexts.com/fpp3/

[55] Hurvich, C.: Differencing and unit root tests. [online], duben 2004, [cit. 2022-05-07]. Dostupné z: https://pages.stern.nyu.edu/~churvich/Forecasting/Handouts/UnitRoot.pdf

[56] Bierens, H. J.: Unit roots. [online], říjen 2007, [cit. 2022-05-07]. Dostupné z: https://personal.psu.edu/hxb11/UNITROOT.PDF

[57] Fomby, T. B.: Augmented Dickey-Fuller Unit Root Tests. [online], září 2006, [cit. 2022-05-07]. Dostupné z: https://s2.smu.edu/tfomby/eco6375/BJ%20Notes/ADF%20Notes.pdf

[58] Kwiatkowski, D.; Phillips, P. C.; Schmidt, P.; aj.: Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, ročník 54, č. 1, 1992: s. 159–178, ISSN 0304-4076, doi: 10.1016/0304-4076(92)90104-Y.

[59] Perktold, J.; Seabold, S.; Taylor, J.; aj.: Stationarity and detrending (ADF/KPSS). [online], květen 2022, [cit. 2022-05-07]. Dostupné z: https://www.statsmodels.org/devel/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

[60] Perktold, J.; Seabold, S.; Taylor, J.; aj.: statsmodels.tsa.arima.model.ARIMA.fit. [online], únor 2022, [cit. 2022-05-07]. Dostupné z: https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.fit.html

[61] Burnham, K. P.; Anderson, D. R.: Multimodel Inference. *Sociological Methods & Research*, ročník 33, 2004: s. 261 – 304.

[62] Matthias, K.; Kane, S. P.: *Docker - Up and Running*. Sebastopol, CA: O'Reilly Media, Červen 2015.

[63] Git community: Git. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://git-scm.com

[64] GitHub, Inc.: GitHub. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://github.com

[65] The PostgreSQL Global Development Group: PostgreSQL. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://www.postgresql.org

[66] The PostgreSQL Global Development Group: PostgreSQL: About. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://www.postgresql.org/about/

[67] Joe Nelson, S. C.: PostgREST Documentation. [online], leden 2022, [cit. 2022-05-06]. Dostupné z: https://postgrest.org/en/stable/

[68] Docker Hub: Postgres – Official Image. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://hub.docker.com/_/postgres

[69] Docker Hub: postgrest/postgrest – Docker Image. [online], únor 2022, [cit. 2022-05-06]. Dostupné z: https://hub.docker.com/r/postgrest/postgrest/

[70] Python Software Foundation: Python. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://www.python.org

[71] pandas – Python Data Analysis Library. [online], duben 2022, [cit. 2022-05-06]. Dostupné z: https://pandas.pydata.org

[72] scikit-learn: machine learning in Python. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://scikit-learn.org/stable/

[73] SciPy. [online], květen 2022, [cit. 2022-05-06]. Dostupné z: https://scipy.org

[74] Varrazzo, D.: PostgreSQL driver for Python – Psycopg. [online], říjen 2021, [cit. 2022-05-06]. Dostupné z: https://www.psycopg.org

[75] Jupyter community: Project Jupyter. [online], 2022, [cit. 2022-05-07]. Dostupné z: https://jupyter.org

[76] Google: Flutter – Build apps for any screen. [online], duben 2022, [cit. 2022-05-06]. Dostupné z: https://flutter.dev

[77] Rousselet, R.: Riverpod. [online], 2022, [cit. 2022-05-06]. Dostupné z: https://riverpod.dev

[78] The World Bank Group: History. [online], 2022, [cit. 2022-04-28]. Dostupné z: https://www.worldbank.org/en/about/history

[79] The World Bank Group: About us – Data. [online], 2022, [cit. 2022-04-28]. Dostupné z: https://data.worldbank.org/about

[80] The World Bank Group: World Development Indicators. [online], duben 2022, [cit. 2022-04-28]. Dostupné z: https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators

[81] Wouts, M.: World Bank Data. [online], červenec 2020, [cit. 2022-04-28]. Dostupné z: https://github.com/mwouts/world_bank_data

[82] Eurostat: Who we are. [online], [cit. 2022-04-28]. Dostupné z: https://ec.europa.eu/eurostat/web/main/about/who-we-are

[83] Eurostat: Statistics. [online], duben 2022, [cit. 2022-04-28]. Dostupné z: https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes

[84] Ministerstvo vnitra České republiky: Datové sady. [online], duben 2022, [cit. 2022-04-28]. Dostupné z: https://data.gov.cz/datové-sady

[85] Google: Google Trends. [online], duben 2022, [cit. 2022-04-28]. Dostupné z: https://trends.google.com

[86] Google: FAQ about Google Trends data. [online], duben 2022, [cit. 2022-04-28]. Dostupné z: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052

[87] Behnen, P.; Kessler, R.; Kruse, F.; aj.: White Paper: Evidence, Scale, and Patterns of Systematic Inconsistencies in Google Trends data. červen 2020, doi:10.13140/RG.2.2.26974.66880.

Příloha **A**

Seznam použitých zkratek

ADF Augmented Dickey-Fuller test

AIC Akaike information criterion

API Application programming interface

AR Autoregression

ARIMA Autoregressive integrated moving average

Bash Bourne-again shell

BIC Bayesian information criterion

CC Creative commons

DF Dickey-Fuller test

DRY Don't repeat yourself

ER Entity relation

FLFP/FLP Female labour force participation

GUI Graphical user interface

HDP hrubý domácí produkt

hmp hrubá míra porodnosti

HTTPS Hypertext Transfer Protocol Secure

i.i.d. independent, identically distributed

KPSS Kwiatkowski–Phillips–Schmidt–Shin test

LFP Labour force participation

MA Moving average

OLS Ordinary least squares

PAYGO Pay-as-you-go

SQL Structured query language

TFR Total fertility rate

TSV Tab separated values

URL Uniform resource locator

Příloha B

Výsledky analýzy ukazatelů

Následující tabulka popisuje ukazatele seřazené sestupně podle počtu jejich časových řad, pro něž bylo nalazeno významné tvrzení o korelaci na 95% hladině. Zároveň je k dispozici počet těch z nich, jejichž korelační koeficient je kladný a také maximální a minimální přítomná hodnota korelačního koeficientu.

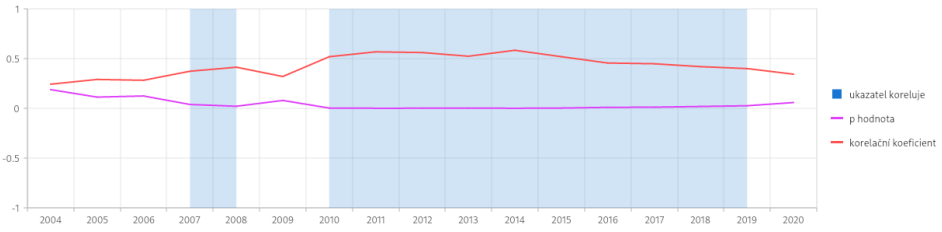
Tabulka B.1: Ukazatele se silným tvrzením o korelaci s TFR

Ukazatel	Min	Max	Kladných	Celkem
Participace osob 15+ v pracovním procesu	-0.58	0.73	15	20
Participace mužů 15+ v pracovním procesu	-0.55	0.62	14	19
Mateřská škola – Téma	-0.80	0.77	13	19
První sňatky mužů	-0.49	0.81	13	18
První sňatky žen	-0.52	0.82	13	18
Kočárek – Téma	-0.82	0.68	8	18
Participace žen 15+ v pracovním procesu	-0.54	0.77	11	17
Umělé oplodnění – Téma	-0.75	0.74	7	17
Výdaje na zdravotní péči	-0.93	0.70	7	17
Stres – Téma	-0.84	0.69	4	17
HDP per capita	-0.42	0.69	14	16
Inflace	-0.51	0.72	14	16
Těhotenství – Téma	-0.69	0.77	14	16
Participace osob 15-24 let v pracovním procesu	-0.64	0.64	11	16
Počet vězňů	-0.82	0.91	11	16

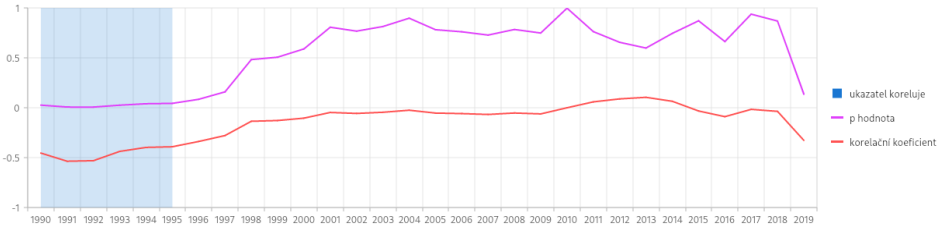
Ukazatel	Min	Max	Kladných	Celkem
Výdaje na podporu v nezaměstnanosti	-0.91	0.94	10	16
Emise CO2	-0.45	0.73	12	15
Porod – Téma	-0.67	0.64	10	15
Hypoteční kalkulačka – Téma	-0.84	0.58	5	15
Participace mužů 15-24 let v pracovním procesu	-0.58	0.58	10	14
Participace žen 15-24 let v pracovním procesu	-0.51	0.67	10	14
Výdaje na podporu rodin	-0.86	0.88	9	14
Rodičovský příspěvek – Téma	-0.71	0.64	8	14
Poměr participace žen vůči mužům v pracovním procesu	-0.59	0.55	7	14
Antikoncepce – Téma	-0.76	0.63	2	14
Podpora v nezaměstnanosti – Téma	-0.84	0.70	8	13
Podíl žen mezi vyučujícími středního vzdělání	-0.91	0.91	7	13
Kojenecká láhev – Téma	-0.69	0.57	6	13
Výdaje na starobní důchody	-0.84	0.87	5	13
Těhotenský test - Téma	-0.63	0.70	10	12
Podíl žen v parlamentu	-0.70	0.58	8	12
Podíl žen mezi vyučujícími prvního stupně	-0.96	0.99	6	12
Potrat – Téma	-0.73	0.79	5	12
Rodičovská dovolená – Téma	-0.73	0.69	3	12
Hlídání dětí – Téma	-0.63	0.70	9	11
Porodnice – Téma	-0.68	0.69	6	10
Výpověď ze zaměstnání – Téma	-0.84	0.68	4	10
Svatba – Téma	-0.79	0.73	4	10
Státní výdaje na vzdělávání	-0.95	0.90	3	10
Rozvod – Téma	-0.61	0.61	5	9
Kojenec – Téma	0.50	0.72	8	8
Průměrný věk osamostatnění potomků – žen	-0.79	0.73	3	8
Sebevražednost	-0.64	0.58	4	6
Průměrný věk osamostatnění potomků – mužů	-0.79	-0.48	0	4
Úrok z vkladu	-0.52	0.77	2	3
Dokončené byty	0.49	0.49	1	1
Medián mezd mužů	-0.97	-0.97	0	1
Medián mezd žen	-0.89	-0.89	0	1

Následující vybrané grafy popisují významné cross-sectional korelace ukazatelů s TFR napříč státy.

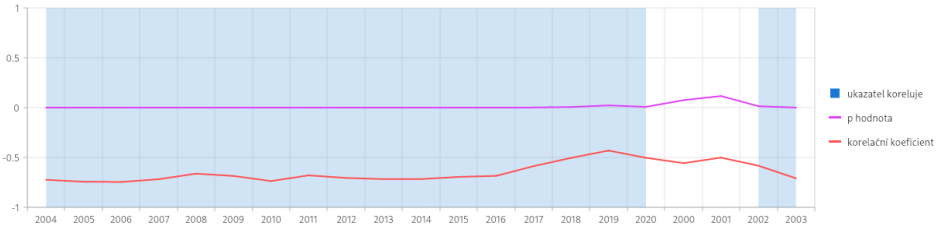
Obrázek B.1: Téma „hlídání dětí“ podle Google Trends



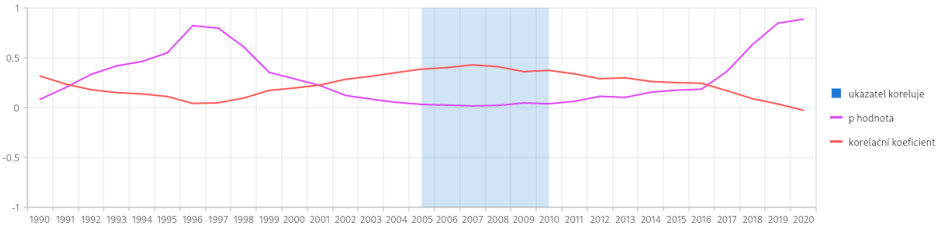
Obrázek B.2: První sňatky žen podle Eurostat



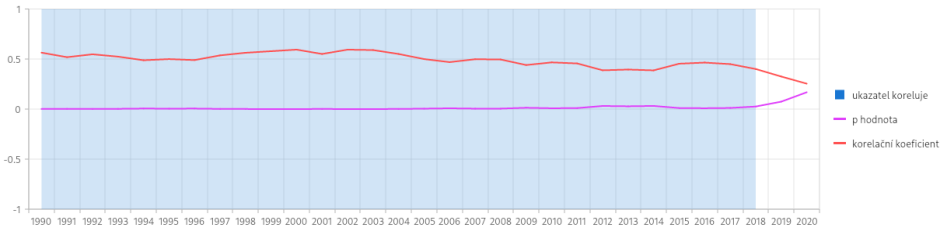
Obrázek B.3: Průměrný věk žen opouštějících rodičovskou domácnost podle Eurostat



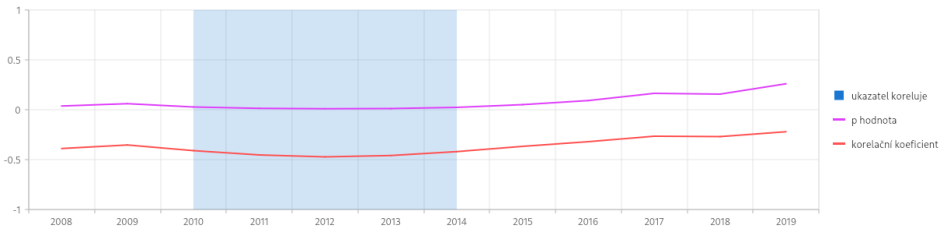
Obrázek B.4: Participace žen 15+ v pracovním procesu



Obrázek B.5: Participace mužů 15+ v pracovním procesu



Obrázek B.6: Státní výdaje na důchody



Příloha C

Obsah přiloženého média

- `readme.txt` ..... stručný popis obsahu média
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF

Příloha D

Instalační příručka

D.1 Předpoklady pro spuštění

Repozitář se zdrojovým kódem je dostupný jednak v digitální příloze práce, jednak na portálu GitHub zde: `https://github.com/opendatalabcz/tfr-dashboard`. Pro spuštění je třeba mít nainstalovaný Docker a Docker Compose.

D.2 Konfigurace

Po získání repozitáře přejděte do jeho adresáře. Přejmenujte přítomný soubor `.env.example` na `.env`. Tento soubor slouží ke konfiguraci jednotlivých služeb, při běhu v lokálním testovacím prostředí v něm lze ponechat výchozí hodnoty.

Pokud chcete definovat vlastní porty, na kterých má být zpřístupněn webový server a API, změňte hodnoty v těchto sekcích souboru `docker-compose.yml`:

- services: api: ports
- services: api: environment: PGRST_OPENAPI_SERVER_PROXY_URI
- services: web: build: args: API_URL
- services: web: ports

D.3 Spuštění

1. Aktivujte kontejner databáze na pozadí, inicializace schématu proběhne automaticky a použijí se proměnné v `.env`:

   ```
   docker-compose up -d postgres
   ```

2. Spusťte sběr a zpracování dat z online zdrojů na popředí a vyčkejte, dokud se data nezpracují. Toto může trvat několik minut. Pokud sběr selže při získávání dat z Google Trends, postupujte podle řešení v podkapitole níže a spusťte poté sběr znovu.

   ```
   docker-compose up data
   ```

3. Poté, co byla data stažena a zpracována, můžete spustit API a webový server, který bude poskytovat klientům aplikaci.

   ```
   docker-compose up -d api web
   ```

S výchozí konfigurací `docker-compose.yml` je nyní dashboard dostupný na `http://127.0.0.1:5053`.

D.3.1 Selhání sběru dat z Google Trends

Google Trends API může dočasně zablokovat přístup kvůli opakovaným požadavkům, které software provádí. V takovém případě je možné dočasně deaktivovat sběr dat z Google Trends přidáním následující proměnné prostředí do souboru `docker-compose.yml`:

```
...
services:
  ...
  data:
    ...
    environment:
      ...
      EXCLUDE_GOOGLETRENDS: 1
...
```

D.4 Prohlížení dat napřímo

Pro manuální prohlížení databáze napřímo bez použití webové aplikace jsou dostupné předkonfigurované kontejnery.

Databázi lze prohlížet pomocí nástroje *pgAdmin*, který je dostupný ve výchozím nastavení po spuštění `docker-compose up -d pgadmin` na adrese `http://127.0.0.1:5050`. Přihlašovací údaje zadejte podle obsahu `.env`.

Další možností je použít interaktivní klient k API, který lze spustit příkazem `docker-compose up -d swagger` a poté jej navštívit na `http://127.0.0.1:5052`.