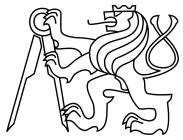
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název:	Stahování a zpracování dat o cenách z internetových obchodů jako nástroj pro zpřesnění predikce inflace
Student:	Bc. Jan Vokoun
Vedoucí:	Ing. Mgr. Pavla Vozárová, Ph.D., M.A.
Studijní program:	Informatika
Studijní obor:	Informační systémy a management
Katedra:	Katedra softwarového inženýrství
Platnost zadání:	Do konce letního semestru 2016/17

Pokyny pro vypracování

1. Seznamte se, formou rešerše, s metodami, kterými predikují krátkodobou inflaci významné finanční instituce (Česká národní banka, Ministerstvo financí ČR, komerční banky). Popište jednotlivé složky inflace.

2. Sestavte přehled webových portálů, na kterých je možné nakoupit zboží odpovídající jednotlivým složkám inflace nebo které alespoň dávají přehled cen tohoto zboží (např. itesco.cz nebo heureka.cz).

3. Navrhněte a implementujte nástroj, který umožní průběžně stahovat a ukládat data o cenách vybraného zboží z alespoň tří těchto webových portálů.

4. Diskutujte, jak je možné tato data využít ke krátkodobé predikci budoucího vývoje inflace a jejích složek.

5. Navrhněte a implementujte webovou aplikaci, která bude pomocí statistického programu R tato data zobrazovat a bude poskytovat krátkodobou prognózu jednotlivých složek inflace.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Michal Valenta, Ph.D.
vedoucí katedry

prof. Ing. Pavel Tvrdík, CSc.
děkan

V Praze dne 8. února 2016

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra softwarového inženýrství

Bakalářská práce

Stahování a zpracování dat o cenách z internetových obchodů jako nástroj pro zpřesnění predikce inflace

Bc. Jan Vokoun

Vedoucí práce: Ing. Mgr. Pavla Vozárová, Ph.D., M.A.

30. června 2017

Poděkování

Na tomto místě bych rád poděkoval Ing. Mgr. Pavle Vozárové, Ph.D, M.A. za veškeré tipy, ochotu a příjemný přístup a své rodině za dlouhodobou podporu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické 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 předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, 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 způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). 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 způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 30. června 2017 .....

České vysoké učení technické v Praze
Fakulta informačních technologií
© 2017 Jan Vokoun. Všechna práva vyhrazena.
Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí, je nezbytný souhlas autora.

Odkaz na tuto práci

Vokoun, Jan. *Stahování a zpracování dat o cenách z internetových obchodů jako nástroj pro zpřesnění predikce inflace*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2017.

Abstrakt

Tato bakalářská práce je zaměřena na možnost využití dat o cenách zboží na internetu pro tvorbu krátkodobé predikce jednotlivých složek inflace. V první části této práce jsou nejprve popsány jednotlivé složky inflace, dále je sestaven seznam webových portálů, které dávají přehled cen zboží odpovídající těmto složkám, a následně je diskutovována možnost jejich využití pro samotnou tvorbu krátkodobé prognózy inflace. V druhé části práce jsou následně navrženy a implementovány dva nástroje, jejichž cílem je tuto prognózu prakticky realizovat.

Klíčová slova inflace, index spotřebitelských cen, predikce inflace, časové řady, ARIMA modely, web scraping

Abstract

This bachelor thesis is focused on using online data prices from web portals in order to create a short term forecast of specific inflation components. This thesis firstly describes components of inflation, offers a list of web portals providing information about prices of corresponding products and finally discusses the possibility of using them for the inflation prediction. The second part of this thesis is focused on the design and implemenation of two applications which, together, aim to forecast the inflation rate of specific inflation compounds.

Keywords inflation, consumer price index, inflation forecast, time series, ARIMA models , web scraping

Obsah

Úvod **1**

1 Inflace **3**
1.1 Činnost České národní banky 3
1.2 Měření 4

2 Spotřební koš **7**
2.1 Potraviny a nealkoholické nápoje 8
2.2 Alkoholické nápoje a tabák 8
2.3 Odívání a obuv 9
2.4 Bydlení, voda, energie, paliva 9
2.5 Bytové vybavení, zařízení domácnosti, opravy 10
2.6 Zdraví 10
2.7 Doprava 10
2.8 Pošty a telekomunikace 11
2.9 Rekreace a kultura 11
2.10 Vzdělávání 12
2.11 Stravování a ubytování 12
2.12 Ostatní zboží a služby 12
2.13 Nespecializované webové portály 13

3 Využitelnost dat získaných online **15**
3.1 Online vs. offline 15
3.2 Nowcasting 15
3.3 Predikce 16

4 Aplikace ukládající data o cenách **21**
4.1 Scraping 21
4.2 Výběr portálů 21
4.3 Použité technologie 25

4.4 Návrh 25
4.5 Funkce 26
4.6 Případy použití 30

5 Webová aplikace **33**
5.1 Použité technologie 33
5.2 Spuštění 33
5.3 Tvorba predikce 34
5.4 Použití 35
5.5 Ukázka funkce 36

Závěr **37**

Literatura **39**

A Seznam použitých zkratek **41**

B Obsah přiloženého CD **43**

Seznam obrázků

2.1 Spotřební koš 7

4.1 Hlavní menu portálu iTesco.cz 22
4.2 Hlavní menu portálu makro.cz 23
4.3 Hlavní menu portálu Heureka.cz 24
4.4 Kategorie portálu Heureka.cz 24
4.5 Hlavní obrazovka aplikace stahující data 26
4.6 Adresářová struktura uložených cen 28
4.7 Vytvoření plánované úlohy 30

5.1 Graf predikované inflace 36

Seznam tabulek

5.1 Hodnoty predikované inflace 36

Úvod

Inflace je významným makroekonomickým ukazatelem ovlivňujícím ekonomickou stabilitu každé země, a proto je pečlivě monitorována téměř všemi významnými finančními institucemi a často také ovlivňována centrálními bankami po celém světě. Problémem, se kterým se mnohé z těchto subjektů potýkají, je fakt, že oficiální míra inflace je zpravidla zveřejňována s několikadenním zpožděním, což může znamenat negativní vliv na jejich další rozhodování.

V České republice je oficiální míra inflace zveřejňována Českým statistickým úřadem, který velkou část podkladových hodnot získává zjišťováním cen zboží v kamenných obchodech. S rostoucí oblibou nakupování po internetu se však nabízí alternativa, kterou je zjišťování cen zboží online a jejich následné využití k získání aktuálních hodnot inflace v okamžitém čase.

Teoretickým cílem této práce je popsání jednotlivých složek inflace a sestavení přehledu portálů, na kterých je možné zboží z těchto složek zakoupit, praktickým pak navržení a implementace dvou aplikací. Úkolem první z nich je průběžné stahování a ukládání cen zboží z internetových obchodů, druhá z nich následně umožní získaná data zobrazit a poskytnout jejich krátkodobou predikci.

Tato práce je formálně členěna do 5 kapitol. První z nich se zabývá inflací, zejména smyslem a způsobem jejího měření. V druhé části jsou popsány jednotlivé složky inflace a je sestaven přehled portálů, na kterých je možné zboží z těchto složek zakoupit. Ve třetí kapitole jsou diskutovány možnosti využití dat získaných online a metody predikce. V předposlední kapitole je nejprve navržena a implementována aplikace sloužící ke stahování a ukládání dat, v poslední části je pak navržena a implementována webová aplikace, díky které je možné získaná data zobrazit a získat predikci jejich budoucího krátkodobého vývoje.

Kapitola 1

Inflace

V této kapitole bude nejprve stručně popsána činnost České národní banky v souvislosti s inflací a následně popsány metody jejího měření.

1.1 Činnost České národní banky

Pojem inflace je v posledních letech často skloňován zejména s kroky České národní banky, což má souvislost s jejím hlavním cílem, kterým je dle §2 zákona o ČNB péče o cenovou stabilitu. Právě k péči o cenovou stabilitu totiž od roku 1998 Česká národní banka využívá měnově politický režim cílování inflace.

Tento režim znamená, že Česká národní banka nejprve s předstihem veřejně vyhlásí svůj inflační cíl, tedy míru inflace, o kterou bude usilovat. Tento cíl je od ledna roku 2010 až do případného vstupu České republiky k eurozóně stanoven na výši 2 % (odpovídající meziročnímu přírůstku indexu spotřebitelských cen) s tolerovaným oboustranným pásmem jednoho procentního bodu. Výše tohoto cíle je dostatečně nízká na to, aby byla zajištěna stabilita cen a zároveň dostatečně vysoká na to, aby České národní bance umožnila provádět měnovou politiku i v případě nemožnosti snížení krátkodobých úrokových sazeb. Již samotné vyhlášení tohoto cíle má dopad na ekonomiku, neboť dojde k ovlivnění inflačních očekávání.[1]

K dosažení tohoto cíle využívá Česká národní banka nástrojů, o kterých rozhoduje sedmičlenná bankovní rada v čele s guvernérem. Tato rada se schází pravidelně osmkrát do roka a rozhoduje na základě podkladů dodaných od jednotlivých útvarů ČNB, především pak od sekce měnové a statistiky, která bankovní radě pro všechna její jednání střídavě poskytuje nově vytvořenou prognózu české ekonomiky a zprávu o jejím vývoji v porovnání s předchozí prognózou.

Dopady kroků učiněných ČNB se na ekonomice zpravidla neprojevují ihned, ale až s odstupem několika měsíců, a proto je prognóza vždy orientována na nadcházející rok a půl. Při její tvorbě je použito mnoho modelů, jejichž

výstupy jsou diskutovány a zohledňovány v dlouhodobějších předpovědích, výsledky se však vždy odvíjejí od dostupných informací o vývoji české i zahraniční ekonomiky.[2]

Výsledky této prognózy se vždy opírají o nové informace popisující vývoj české i zahraniční ekonomiky, které jsou nejprve využity pro tvorbu expertní krátkodobé projekce a ta je následně zohledněna při tvorbě prognózy střednědobé. Celý proces tvorby je poměrně složitý a zároveň komplikovaný okolnostmi, mezi které patří například nutnost zjednodušení komplikovaných ekonomických vztahů, nejistota spojené s vnějšími vlivy (přírodní katastrofy, zahraniční situace) nebo nedokonalost vstupních dat.[3]

1.2 Měření

Inflace je v literatuře obvykle definována jako všeobecný růst cenové hladiny v čase. K jejímu vyjádření se využívá míra inflace, která charakterizuje změnu cenové hladiny během určitého časového období, nejčastěji za pomoci cenových indexů, mezi které patří zejména index spotřebitelských cen, index cen výrobců a deflátor HDP.[4]

1.2.1 Index spotřebitelských cen

V České republice je hlavní pozornost věnována především vývoji indexu spotřebitelských cen (známého také jako CPI - Consumer Price Index). Tento cenový index pracuje s uměle vytvořeným spotřebním košem (v ČR spravovaným Českým statistickým úřadem), který obsahuje statky tvořící významnou část výdajů běžné domácnost a váhy odpovídají podílu příslušných položek na celkových nákladech. S pomocí těchto vah je v prvním sledovaném období nejprve vypočten vážený průměr, který slouží jako referenční hodnota, se kterou jsou následně porovnávány výsledky stejně vypočteného váženého průměru v následujících obdobích.[4] Vzorec pro výpočet CPI mezi 2 obdobími lze zapsat následující formou:

$$CPI = \frac{\sum\limits_{i} \frac{p_{1i}}{p_{0i}} * w_i}{\sum\limits_{i} w_i} * 100 \tag{1.1}$$

kde p_{0i} představuje cenu statku i v prvním období, p_{1i} jeho cenu v druhém období a w_i jeho váhu ve spotřebním koši.

Míru inflace lze ze získaných hodnot CPI ve dvou obdobích získat následovně:

$$\pi = \frac{CPI_1 - CPI_0}{CPI_0} * 100 \tag{1.2}$$

Za tato období bývají běžně dosazovány dva za sebou jdoucí měsíce (pak se jedná o inflaci meziměsíční), nebo dva měsíce s 12 měsíčním odstupem (pak se jedná o inflaci meziroční).

Obdobným způsobem je možné získat míru inflace i za pomoci hodnot jiných cenových indexů.

1.2.2 Index cen výrobců

Dalším populárním indexem je index cen výrobců (PPI - Production Price Index), který ve skutečnosti sleduje hned několik indexů (cen zemědělských výrobců, cen průmyslových výrobců, cen stavebních prací, apod.). Jeho princip je obdobný jako u indexu spotřebitelských cen, avšak spotřební koš je u tohoto indikátoru nahrazen vyrobenými a prodanými průmyslovými výrobky. I tento index je v České republice spravován Českým statistickým úřadem, který měsíčně získává údaje o celkem přibližně 4800 položkách od více než 1200 organizací.[5]

1.2.3 Deflátor HDP

Deflátor na rozdíl od předchozích indexů nesleduje žádnou konkrétní skupinu zboží, ale zachycuje vývoj cen všech služeb a statků vyprodukovaných v ekonomice a jedná se tedy o nejkomplexnější ukazatel inflace.[6] Výpočet tohoto indexu je založen na porovnání výše nominálního (ohodnoceného běžnými cenami sledovaného roku) a reálného (ohodnoceného stálými cenami základního roku):[4]

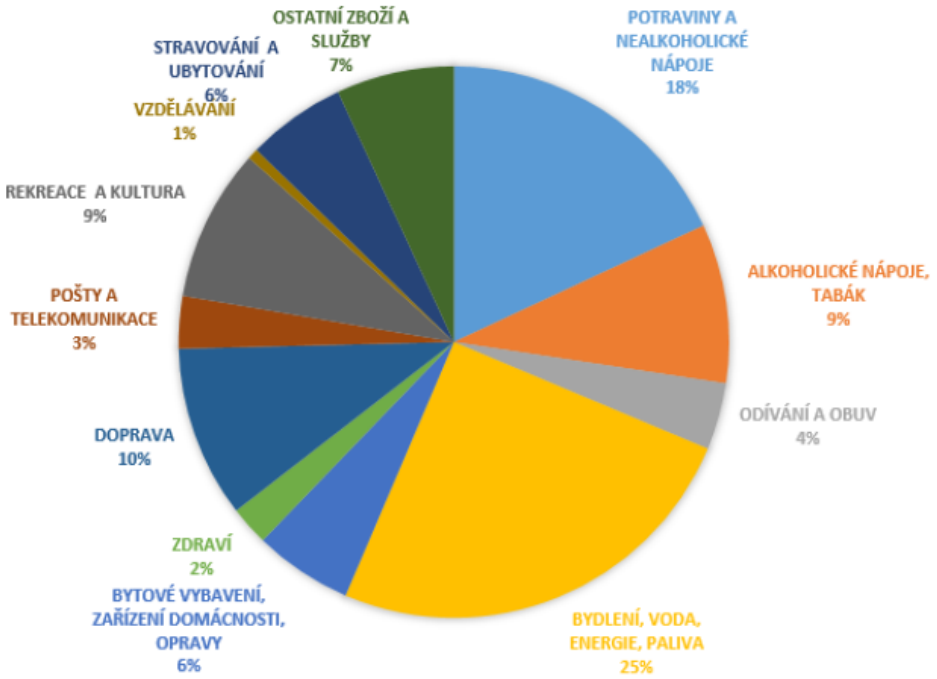
$$deflátor_{HDP} = \frac{nominálníHDP}{reálnýHDP} * 100 \tag{1.3}$$

Kapitola 2

Spotřební koš

Spotřební koš je základním modelem, který je použit pro výpočet indexu spotřebitelských cen. V České republice je vytvořen a sledován Českým statistickým úřadem a v současnosti zahrnuje přibližně 700 položek ve 12 kategoriích.

V této kapitole bude sestaven přehled těchto oddílů společně s webovými portály, které nabízejí zboží spadající do příslušné kategorie. Portály, jejichž sortiment zboží zasahuje do více kategorií, budou zmíněny zvlášť v poslední podkapitole.



Obrázek 2.1: Poměry oddílů spotřebního koše

2.1 Potraviny a nealkoholické nápoje

V tomto oddíle jsou zahrnuty veškeré potraviny a nealkoholické nápoje, jejichž celkový podíl na spotřebním koši tvoří přibližně 18 %. Samotné potraviny tvoří zhruba 90 % váhy této kategorie, zbytek připadá na nealkoholické nápoje. Ceny položek z této koše jsou zjišťovány přímo v terénu.[7]

Nejvýznamnějšími podkategoriemi potravin je Maso (4,34 %), Mléko, sýry a vejce (3,21 %), Pečivo (2,91 %) Uzenářské zboží (1,7 %), Zelenina (1,43 %), Ovoce (1,25 %) a Cukr, marmeláda, med, čokoláda, cukrovinky a cukrářské výrobky (1,25 %).[8]

Portály:

- iTesco.cz

  Portál maloobchodního řetězce Tesco nabízející více než 20 tisíc položek s cenami odpovídajícími cenám v kamenné prodejně. Součástí nabídky je i rozvoz objednaného zboží.

- Rohlik.cz, kosik.cz, kolonial.cz, potravinydomu.cz, z-market.cz a další

  Internetové supermarkety nabízející rozvoz zboží.

- Makro.cz

  Portál velkoobchodního řetězce Makro s více než 30 tisíci položkami. Umožňuje zobrazit sortiment z kteréhokoliv ze 13 vlastních obchodních center v České republice.

- Maneo.cz

  Internetový obchod zaměřený na zásobování gastronomických provozů, občerstvení a provozen.

2.2 Alkoholické nápoje a tabák

Podíl alkoholických nápojů a tabákových výrobků celkově odpovídá přibližně 9,34 % na spotřebním koši, vliv těchto statků na vývoj spotřebního koše je tedy v porovnání s potravinami a nealko nápoji zhruba poloviční.

Alkoholické nápoje jsou dohromady ve spotřebním koši zastoupeny 4,25 %. Tyto nápoje jsou rozděleny do tří kategorií – Lihoviny (1,7 %), Vína (1,15 %) a Piva (1,4 %).[8]

Tabákové výrobky jsou zastoupeny pouze 8 značkami cigaret (s různými vahami), jejich celkový podíl na spotřebním koši je však relativně vysoký a dohromady odpovídá celkem 5,09 %. Ceny těchto výrobků jsou sice regulovány a jejich hodnoty jsou vždy stanoveny v cenových věštnících Ministerstva financí, po jejich změně však dochází k doprodeji výrobků za předchozí ceny, a proto jsou stále sledovány pracovníky ČSÚ přímo v terénu.[7]

Alkoholické nápoje a tabák jsou běžně prodávány společně s potravinami a nealko nápoji, seznam portálů nabízejících zboží z této kategorie je tedy velmi podobný seznamu předchozí složky.

2.3 Odívání a obuv

Váha této kategorie na spotřebním koši je přibližně 3,91 %, přičemž na Odívání připadá celkem 2,92 % a na Obuv (včetně oprav) 0,99 %. Do této kategorie jsou kromě konečných produktů zahrnuty také oděvní materiály nebo služby jako čištění, opravy a půjčování oděvů.[8]

Specifickým problémem při měření inflace této kategorie je fakt, že se její položky poměrně rychle mění. V případě významné změny kvality položky usiluje ČSÚ o výběr obdobného produktu s co nejpodobnějšími měřitelnými parametry (např. výrobce, značka, gramáž, složení materiálu), za změnu kvality však není považován rozdíl ve střihu, vzoru nebo tvaru. Obtížně klasifikovatelné případy posuzuje oddělení statistiky spotřebitelských cen individuálně.[7]

Portály:

- Glami.cz

  Virtuální obchodní galerie, agregátor internetových obchodů s módou.

- Zoot.cz

  Největší internetový obchod s módou v České republice, nabízí přes 300 originálních značek.

2.4 Bydlení, voda, energie, paliva

Tato kategorie odpovídá základním nákladům na bydlení, které zpravidla tvoří velkou část rozpočtu domácností, tudíž je z hlediska podílu na spotřebním koši nejvýznamnější a celkově odpovídá více než čtvrtině (25,12 %).

Nejvýraznější zastoupení v této kategorii mají elektrická a tepelná energie, plyn a ostatní paliva (10,01 %), imputované nájemné (hypotetické nájemné odpovídající ceně, za kterou by bylo možné pronajmout vlastněnou nemovitost – 8,7 %), reálné nájemné (3,53 %), služby spojené s bydlením – vodné, stočné a sběr odpadů (2,02 %) a náklady na běžnou údržbu a drobné opravy bytu (0,87 %).[8]

Pro tuto kategorii neexistují žádné zcela vhodné webové portály. Ke zjištění cen průměrného nájemného by teoreticky bylo možné využít například nabídek realitních kanceláří, dlouhodobé sledování vývoje cen konkrétních nemovitostí tímto způsobem je však z principu téměř nerealizovatelné.

Přímo dostupný není ani kompletní přehled cen energií, informace o cenách je však možné získat s pomocí internetových kalkulátorů.

Portály:

- cenyenergie.cz

  Kalkulačka cen elektřiny a plynu doporučena Ministerstvem financí ČR.

- porovnej24.c

  Srovnávací portál porovnávající ceny energií, volání a finančních služeb

2.5 Bytové vybavení, zařízení domácnosti, opravy

Váha tohoto oddílu je stanovena celkem na 5,78 %. Součástí je především nábytek a podlahové krytiny (2,16 %), domácí přístroje a spotřebiče včetně oprav (1,04 %) a spotřební zboží pro běžnou údržbu domácnosti (1,02 %).[8]

Portály:

- ikea.cz, kika.cz, asko-nabytek.cz, sconto.cz

  Webové stránky řetězců prodejen nábytku a bytových doplňků.

2.6 Zdraví

Váha kategorie zdraví je nastavena na 2,3 %, přičemž většinovou část tvoří Léčiva a zdravotnické pomůcky (1,54 %) a zbytek připadá na různé typy zdravotních péčí.[8]

Portály:

- drmax.cz, benu.cz a pilulka.cz

  Internetové obchody řetězců kamenných lékáren.

- lekarna.cz

  Internetová lékárna.

2.7 Doprava

Doprava se podílí na celkovém spotřebním koši 10,08 %. 3,69 % tvoří nákup nových i ojetých dopravních prostředků (automobily, motocykly, jízdní kola), 4,86 % jejich provoz (zejména pohonné hmoty, které tvoří 3,31 %) a zbylých 1,54 % připadá na dopravní služby.[8]

Dopravní prostředky jsou vybrány na základě podkladů od experta na trh s automobily. Ceny nových automobilů jsou získávány v terénu a ceny ojetých vozů od firmy EurotaxGlass's, která každý měsíc vydává ceník odrážející jejich aktuální cenu na trhu.

Ceny pohonných hmot jsou průběžně zjišťovány v pondělí každého týdne z výkazů získaných od 12 hlavních distributorů v České republice. Průměrná cena za tento týden je vypočtena jako vážený průměr z průměrných cen (z jednotlivých poboček) daných distributorů s váhami odpovídajícími jimi prodanému množství. Průměrná měsíční cena je pak získána jejich prostým aritmetickým průměrem.[7]

Portály:

- mBenzin.cz

  Portál zaznamenávající data o cenách pohonných hmot na čerpacích stanicích po celé ČR získaných od uživatelů.

2.8 Pošty a telekomunikace

Tento oddíl je zastoupen celkovou váhou 3,07 %. Hlavní roli v něm hrají především telefonické a telefaxové služby (2,91 %), na poštovní služby a telefonická zařízení pak připadá zbylých 0,16 %.[8]

Zpracování cen telefonních služeb je vzhledem k častým úpravám tarifů ze strany operátorů poměrně komplikované. Řešením bylo zavedení subindexu, který obsahuje přibližně 60 cen, jejichž váhy jsou určeny na základě několika statistik Českého statistického úřadu i Českého telekomunikačního úřadu.[7]

Obdobně jako v případě spotřebních energií existuje na českém internetu množství portálů umožňující srovnání i tarifů mobilních operátorů, možnost jejich využití pro monitorování průběžného vývoje cen je však s ohledem na počet parametrů a relativně komplikovaný způsob výpočtu poměrně omezená.

Portály:

- tarifomat.cz, srovnanioperatoru.cz a další

  Internetové kalkulačky a porovnávače telefonních tarifů.

2.9 Rekreace a kultura

Tato kategorie se podílí na spotřebním koši celkem 8,96 % a je tvořena poměrně pestrým výběrem položek. Patří do ní ceny rekreačních a kulturních služeb (2,37 %), rekreací včetně vybavení + položky spojené se zahradou a domácími zvířaty (2 %), ceny dovolených v tuzemsku i vybraných zahraničních zemí (1,98 %), audiovizuální a fotografická zařízení (1,6 %) a ceny novin, knih a papírenského zboží (0,85 %).[8]

Portály:

- invia.cz

  Cestovní agentura zprostředkovávající nabídky z více než 300 cestovních kanceláří z ČR i zahraničí.

- kosmas.cz, neoluxor.cz, knihy.cz, martinus.cz

  Kamenná a internetová knihkupectví.

2.10 Vzdělávání

Váha oddílu spojeného se vzděláváním je stanovena na 0,62 % a má na spotřebním koši tedy nejnižší podíl ze všech 12 oddílů. Položky v této kategorii jsou rozděleny do sekcí podle úrovně vzdělávání (preprimární, primární, sekundární, terciární a nedefinované). Předmětem sledování jsou zejména ceny vzdělávacích služeb poskytovaných soukromými školami a výše poplatků na státních vysokých školách a v mateřských školkách.[8]

Portál:

- vysokeskoly.com

  Adresář vysokých škol udávající i přehled školného na soukromých VŠ.

2.11 Stravování a ubytování

Tento oddíl zahrnuje pouze služby a na spotřební koš má vliv v podobě 5,81 %. Převážnou část tvoří stravovací služby, tedy ceny jídel v jídelnách, restauracích a kavárnách (5,05 %), zbytek pak připadá na ubytovací služby v hotelech, penzionech, chatách a kolejích (0,76 %).[8]

Portál:

- DameJidlo.cz

  Nejrozšířenější rozvážková služba hotových jídel České republice, pokrývá jídelní lístky přibližně 1500 restaurací v 10 městech ČR.

2.12 Ostatní zboží a služby

Na poslední oddíl, který je obsahem velmi rozmanitý, ve spotřebním koši připadá váha ve výši 6,92 %. Nejvýznamnější část (3,16 %) připadá pododdíl Osobní péče, který je složen zejména z kadeřnických a kosmetických služeb (1,26 %) a kosmetického a drogistického zboží (1,85 %). Zbylou část spotřebního koše pak tvoří finanční služby (1,36 %), pojištění (0,65 %), sociální péče (0,65 %) a osobní potřeby a doplňky (0,62 %).[8]

Portály:

- xantea.cz, indrogerie.cz, vmd-drogerie a další

  Internetové drogerie.

- chytryhonza.cz, usetreno.cz

  Srovnávací portály porovnávající ceny finančních služeb (mimo jiné).

2.13 Nespecializované webové portály

Ceny zboží na internetu lze zjistit i na portálech, které se z pohledu spotřebního koše neorientují pouze na jednu konkrétní skupinu produktů. Mezi tyto portály lze zařadit zejména obchodní domy, které nabízejí vlastní sortiment nejrůznějšího zboží (elektroniky, oblečení, drogerie, zahradních potřeb a dalších), a vyhledávače zboží, které agregují zboží až z desetitisíců českých internetových obchodů s různým zaměřením. Velmi významná část tržeb na internetu navíc patří obchodům s elektronikou, tu lze však nalézt minimálně ve třech hlavních sekcích spotřebního koše.

Obchodní domy:

- Mall.cz

  Vítěz soutěže ShopRoku v kategorii Obchodní domy za rok 2016, nabízí sortiment s cca 200 000 produkty.

- Eva.cz, kasa.cz

  Zbylí finalisté soutěže ShopRoku.

Cenové vyhledávače:

- Heureka.cz

  Největší vyhledávač a srovnávač zboží na českém internetu s 50% podílem na tomto segmentu. Nabízí přibližně 17 miliónů produktů.

- zbozi.cz

  Druhý největší portál v této kategorii s přibližně 45% zastoupením.

- Hledejceny.cz

  Třetí největší portál s tímto zaměřením a cca 4% podílem na trhu.

- Kupi.cz

  Portál monitorující slevy v supermarketech, hypermarketech, obchodech zaměřených na bydlení, zahradu, nábytek, elektro, drogerii, parfumerii, lekárny a další.

Obchody s elektronikou:

- Alza.cz, CZC.cz, TSBohemia.cz

  Největší české internetové obchody s elektronikou.

Kapitola 3

Využitelnost dat získaných online

V této kapitole je diskutována možnost využití dat o cenách z internetových obchodů a navržen způsob predikce jejich budoucího vývoje.

3.1 Online vs. offline

Základní otázkou, kterou je nutné si položit ještě před samotným zpracováním dat získaných z online zdrojů, je, zda jsou si ceny v internetových a kamenných obchodech vůbec podobné, neboť v případě negativní odpovědi by ztrácelo význam tato data pro zjištění indexu spotřebitelských cen vůbec zkoumat.

Odpověď ve své práci podává Cavallo[9], který porovnává vývoj cen v 56 retailových kamenných i online obchodech z 10 různých zemí (Argentina, Austrálie, Brazílie, Kanada, Čína, Německo, Japonsko, Jihoafrická republika, Spojené království a Spojené státy). Výsledkem jeho studie je zjištění, že ceny se shodují průměrně v přibližně 72 % případů a podává důkaz, že data získaná z webu mohou sloužit jako reprezentativní vzorek odpovídající datům získaným v kamenných obchodech.

3.2 Nowcasting

Now-casting je termín, který vznikl spojením slow *now* a *forecasting*, což lze volným způsobem přeložit jako predikci současnosti. Tento pojem je již relativně dlouhou dobu používán v meteorologii a v posledních letech své místo nachází i v ekonomii, ve které může být uplatněn k predikci statistik, jež jsou publikovány s delší periodou (tedy zároveň prodlevou mezi jednotlivými zveřejněními).

Obecný princip této metody je použití informací, které slouží jako podklad pro očekávanou statistiku a lze je získávat v kratších časových intervalech. Po-

užitím těchto dat je vhodným postupem možné dospět k brzkému odhadu očekávané statistiky ještě dříve před jejím oficiálním zveřejněním, využití online cen k získání aktuální inflace je tedy ukázkovým příkladem.[10]

Předpoklad, že hodnoty inflace získané tímto způsobem, je možné využít i pro *nowcasting* oficiální inflace, potvrzuje další studie primárně zaměřená na komparaci obou inflací v Argentině, kde byly hodnoty naměřené oficiální inflace delší dobu kritizovány.[11]

V rámci této studia byla během let 2007-2011 srovnávána podobnost hodnot inflace zjištěných pomocí obou metod ve čtyřech dalších zemích Jižní Ameriky. Jejím výstupem bylo zjištění, že meziroční vývoj cen na internetových stránkách největších supermarketů v jednotlivých zemích skutečně poměrně přesvědčivě kopíruje trend vývoje reálně zjišťovaného indexu spotřebitelských cen, a to jak v zemi s relativně nízkou inflací (okolo 3 %, Chile), tak i v ekonomicky nestabilní zemi potýkající se s pádivou inflací (okolo 30 %, Venezuela).

Na podobnosti obou výsledků se však významně projevila skutečnost, že data ve všech zemích byla stahována pouze z jediného supermarketu. Markantní rozdíl v přesnosti je patrný zejména při porovnání korelace obou indexů mezi Chile a Brazílií. Zatímco data o cenách v Chile byla získávána u řetězce s 27% podílem na trhu, který zveřejňoval své ceny pro Santiago de Chile (hlavní město tvořící 55 % hodnoty CPI dle místní metodiky), v Brazílii byly ceny stahovány ze supermarketu s pouze 15% podílem na trhu a s cenami platnými pro Rio de Janeiro, které má dle tamní metodiky pro výpočet indexu váhu pouze 17 %. Výsledek byl dle předpokladů o poznání přesnější v případě Chile, což potvrzuje očekávatelné - významnější statistický vzorek znamená vyšší přesnost.[11]

3.3 Predikce

Jedním z cílů této práce je poskytnout krátkodobou prognózu jednotlivých složek inflace využívající dat získaných z online zdrojů. V této kapitole bude diskutováno, zda a jakým způsobem toho lze dosáhnout.

3.3.1 Časové řady

Časové řady jsou v literatuře definovány jako posloupnost srovnatelných a časově (ze směru minulost - přítomnost) jednoznačně uspořádaných pozorování[12], což data o vývoji inflace bez problému splňují. Cílem analýzy časových řad je zpravidla vytvoření odpovídajícího modelu, který umožní řadám porozumět a předpovídat jejich budoucí hodnoty.

3.3.1.1 Kategorie

Časové řady lze rozčlenit dle různých kritérií do následujících kategorií:

- Krátkodobé a dlouhodobé

  Jako dlouhodobé časové řady jsou označovány řady s periodicitou delší než jeden rok, míra inflace tedy spadá mezi časové řady krátkodobé.

- Ekvidistantní a neekvidistantní

  Pojem ekvidistance v časových řadách označuje konstantní rozdíl mezi pozorováními. V této kapitole bude vycházeno z premisy, že data budou získávána pravidelně s periodou jednoho dne (případně týdne), bude se tedy jednat o časovou řadu ekvidistantní, se kterou se obecně snáze manipuluje.

- Deterministické a stochastické

  Determinismus u časových řad říká, že jsou konstruovány dle nějakého modelu, nezahrnují tedy prvek náhody a lze je tedy přesně rekonstruovat. Naprostá většina časových řad vyskytujících se v reálném životě však prvek neurčitosti obsahuje a výjimkou není ani míra inflace.

- Stacionární a nestacionární

  Stacionární řady jsou takové řady, jejichž hodnoty aritmetického průměru a variability jsou v celém průběhu shodné. Opakem jsou řady nestacionární, jejichž hodnoty těchto dvou statistických ukazatelů se průběžně mění. Veškeré časové řady, které například vykazují trend nebo sezónní složku, jsou tedy nestacionární.

3.3.1.2 Metody analýzy

Možností, jak analyzovat časovou řadu je poměrně mnoho. Mezi nejpoužívanější metody a postupy patří následující:[12]

- Dekompozice časové řady
- Box-Jenkinsova metodologie
- Spektrální analýza
- Vícerozměrné metody

Ve zbytku této práce však bude pozornost orientována pouze na Box-Jenkinsonovu metodologii. Modely vzniklé za pomoci této metodologie jsou relativně snadno aplikovatelné a s dobrými výsledky v praxi využívány i mnohými centrálními bankami, v Evropě například islandskou Central Bank of Iceland nebo norskou Norges Banka.[13]

3.3.2 Box-Jenkinsonova metodologie

Box-Jenkinsonova metodologie předpokládá, že jednotlivá pozorování v časové řadě jsou navzájem závislá, a zaměřuje se na jejich korelační analýzu. Modely, které vznikají touto metodou, využívají dva druhy procesů - autoregresní (AR) a metodu klouzavých součtů (MA), podle kterých jsou tyto modely zkráceně nazývány jako ARMA modely. [14]

3.3.2.1 Autoregresivní modely

Pomocí autoregresního procesu lze předpovědět budoucí hodnoty sledované proměnné s využitím jejích hodnot v minulosti. Matematickým zápisem lze tento proces zapsat následujícím způsobem:

$$y_t = c + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + e_t \tag{3.1}$$

kde c značí konstantu a e_t představuje náhodnou složku (nejčastěji bílý šum). Tento model je také nazýván jako $AR(p)$ model, ve kterém p značí řád procesu.[15]

Autoregresní modely jsou velmi flexibilní a zvládají zpracovat mnoho časových řad s různým chováním.

3.3.2.2 Modely klouzavých součtů

Metoda klouzavých součtů na rozdíl od AR nevyužívá pro předpověď předchozí hodnoty, ale chyby v předchozích předpovědích. Tento model lze popsat následující formou:

$$y_t = c + et + b_1 e_{t-1} + b_2 e_{t-2} + ... + b_q e_{t-q} \tag{3.2}$$

kde c opět značí konstantu a e bílý šum (jehož hodnoty nejsou přímo sledovány).[15]

3.3.2.3 ARIMA modely

ARIMA modely jsou modely vzniklé kombinací tří částí:

- AR
- MA
- Integrace

První dvě části odpovídají již popsaným procesům, integrační část pak udává stupeň diferenciace časové řady před jejich aplikací, čehož je využito například pro stacionarizaci nestacionárních řad. Výsledný model je označován jako $ARIMA(p, d, q)$, kde p odpovídá řádu AR, q řádu MA a q je právě zmíněný stupeň diferenciace.

3.3.2.4 SARIMA modely

Speciálním typem ARIMA modelů jsou modely označované jako SARIMA (Seasonal ARIMA), které je vhodné použít na analýzu časových řad vykazujících vliv sezónnosti. Tyto modely jsou zapisovány jako $ARIMA(p, d, q)x(P, D, Q)$, přičemž parametry s velkým písmenem představují parametry ARIMA modelu sezónní složky.[16]

Kapitola 4

Aplikace ukládající data o cenách

Cílem této kapitoly je vytvoření a popis práce nástroje, jenž umožní průběžně stahovat a ukládat data o cenách vybraného zboží z webových portálů.

4.1 Scraping

Pro činnost, která se od nástroje očekává, existuje již poměrně zavedený pojem *webscraping*. který souhrnně označuje proces extrahování a zpracování informací z webových stránek. Obecně se tato technika využívá zejména v případech, kdy existuje web s (pro uživatele) mnoha cennými daty, která jsou obtížně zpracovatelná přímo z prohlížeče, ale jsou uspořádána jasně definovanou strukturou (zpravidla pomocí jazyka HTML).[17]

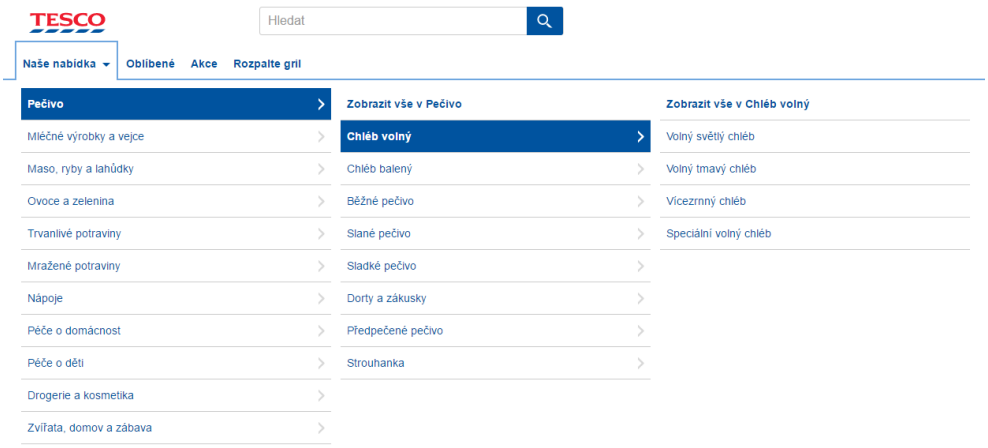
Jako web scraper (nebo crawler, jak je tento typ aplikací také nazýván) je pak označen program, který za uživatele automaticky prochází zadané stránky a na základě zadaných instrukcí stahuje a ukládá relevantní získané informace způsobem, díky kterému je s nimi dále možné snadno manipulovat.

4.2 Výběr portálů

K průběžnému sledování cen produktů byly vybrány tři portály - iTesco.cz, makro.cz a Heureka.cz.

4.2.1 iTesco.cz

Portál iTesco.cz (resp. projekt Potraviny online, dostupný na webové adrese `https://nakup.itesco.cz/groceries/`) je jedním z nejvhodnějších kandidátů pro účely této aplikace, jelikož se jedná o portál nadnárodního řetězce Tesco, který v současnosti stále jako jediný z velkých českých maloobchodů



Obrázek 4.1: Menu portálu iTesco.cz

s rychloobrátkovým zbožím zveřejňuje online kompletní nabídku svého sortimentu. Tento sortiment čítá téměř 20 tisíc položek, z nichž velká část spadá do prvních dvou sekcí spotřebního koše (*Potraviny a nealkoholické nápoje* a *Alkoholické nápoje, tabák*), které dohromady tvoří jeho nejvýznamější část (27 %).

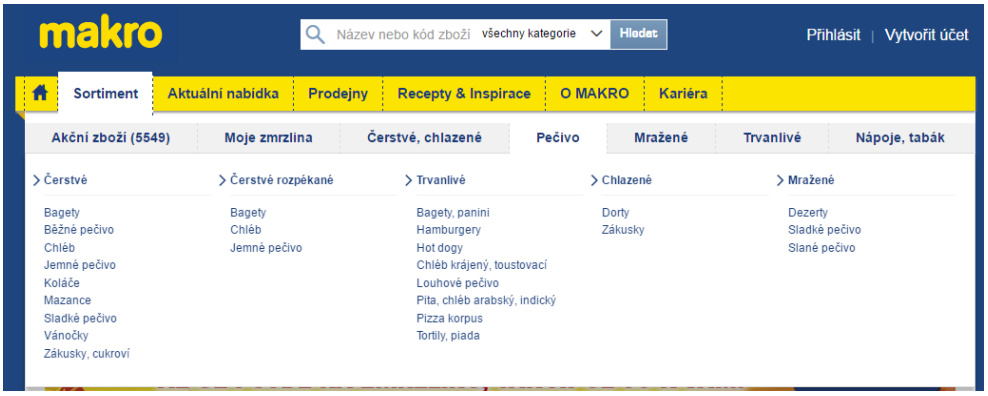
Zboží na tomto portále je organizováno do tříúrovňových kategorií, které jsou v prohlížeči dostupné pomocí menu umístěného na levé straně (4.1). Každá z kategorií nižší úrovně v sobě zahrnuje položky dostupné v kategorii o úroveň výše, součástí kategorie *Pečivo* jsou tedy například všechny produkty z podkategorie *Chléb volný*, které samy obsahují například veškeré položky z podpodkategorie *Volný světlý chléb*.

Ve zdrojovém kódu jsou informace o kategoriích umístěné v hodnotě atributu tagu *<HTML>*. Dolování názvů, odkazů a hierarchie je relativně komplikovanou záležitostí, neboť jsou tyto informace ve zdrojovém kódu často redundantní a ukryté ve velkém množství sémanticky bezvýznamných řetězců. K získání kompletní struktury je tedy využita série regulárních výrazů a operací s textovými řetězci.

Samotné zboží a informace o jeho ceně jsou dostupné přímo ze stránky kategorie a tato data jsou obalena jasně definovanými tagy, díky čemuž je jejich vytěžení s pomocí jazyka XPath poměrně jednoduchou záležitostí. Drobnou nepříjemností je však omezení počtu produktů na 1 stránku (24), což v případě rozsáhlé kategorie znamená nutnost stažení většího počtu stránek.

4.2.2 makro.cz

Druhým zvoleným portálem je portál obchodního řetězce Makro (dostupný na `https://sortiment.makro.cz`), který však na rozdíl od Teska není zaměřen na širokou veřejnost, ale cílí především na podnikatele. I tento řetězec



Obrázek 4.2: Menu portálu makro.cz

má v ČR dlouhou tradici a funguje zde již od roku 1990 (v různých podobách). Sortiment je rovněž orientován především na první dvě hlavní sekce spotřebního koše.

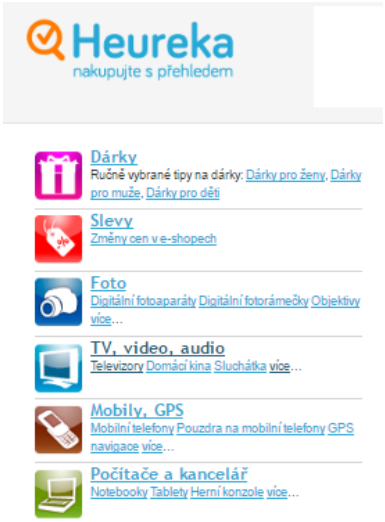
Portál makro.cz má stejně jako portál iTesco.cz kategorie rozmístěné do kategorií tří úrovní, jež jsou rovněž přístupné přímo z titulní stránky a pro které opět platí, že obsahují produkty svých podkategorií. Vytěžení informací o všech kategoriích je v tomto případě však vcelku snadnou záležitostí, neboť veškerá data jsou uchována v přehledné struktuře pomocí HTML tagů a lze je tedy adresovat jazykem XPath.

Zboží (včetně cen) se na tomto portále nachází v kategoriích, kterým lze pomocí parametru v URL adrese přikázat, aby zobrazily až 90 produktů na stránku. U každého produktu jsou zpravidla uvedeny 4 ceny (za kus a za kilo, s DPH a bez DPH), které lze opět získat s využitím jazyka XPath.

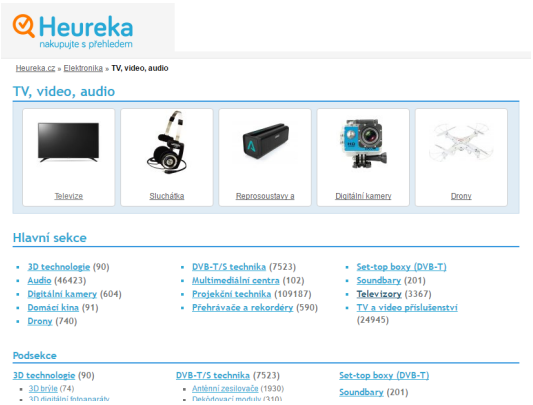
4.2.3 Heureka.cz

Portál Heureka.cz je v České republice nejvyužívanějším vyhledávačem zboží. Jeho fungování je založeno na agregování produktů z několika desítek tisíc internetových obchodů, které následně zobrazuje vlastním uživatelům. Sortiment tohoto portálu přesahuje 20 miliónů položek a pokrývá tedy téměř veškeré části spotřebního koše, které obsahují běžně dostupné, hmatatelné produkty. V rámci aplikace však není řešeno stahování a ukládání dat z kategorie *Oblečení a móda*, neboť její struktura je naprosto odlišná od zbytku portálu.

Stavba portálu Heureka.cz je ve srovnáním s předchozími portály výrazně komplikovanější. Na titulní stránce jsou zobrazeny pouze hlavní kategorie + některé podkategorie (4.3) a jejich kompletní hierarchii není možné získat stažením jediné stránky. Portál člení své kategorie do celkem pěti úrovní, přičemž umožňuje pouze zobrazení podkategorií o 1 nebo 2 kategorie výše (4.4), což vzhledem k jeho velikosti činí samotný proces získávání struktury kategorií



Obrázek 4.3: Hlavní menu portálu Heureka.cz



Obrázek 4.4: Kategorie portálu Heureka.cz

časově poměrně náročným.

Tento fakt byl zohledněn i při tvorbě aplikace, která nejprve stáhne a zobrazí seznam přibližně 15 hlavních kategorií, uživatel si vybere ty, o jejichž podkategorie má zájem a aplikace získá strukturu pouze zvolených kategorií. Tím dojde k významné časové úspoře, neboť pokud uživatele zajímají například ceny produktů z kategorie *Tablety* patřící pod sekci *Elektronika*, není nutné získávat například strukturu podkategorií hlavní sekce *Kosmetika a zdraví*.

Oproti předchozím portálům je Heureka.cz specifická i v dostupnosti produktů, které jsou umístěny pouze v některých kategoriích (nezávisle na úrovni). Hlavní kategorie *Počítače a kancelář* tedy přímo neobsahuje žádné produkty, ale pouze drží odkazy na specifičtější podkategorie.

Dalším rozdílem je i zobrazená cena produktů. Portál Heureka.cz zpravidla nabízí každé zboží z několika internetových obchodů, jejichž ceny se navzájem liší a zboží tedy lze nakoupit za různé ceny. Přímo ze stránky kategorie jsou však viditelné pouze minimální a maximální ceny, za které lze zboží koupit, a pro zobrazení všech cen je nezbytně nutné navštívit vlastní stránku produktu, což samozřejmě znamená vyšší nároky na čas a stahovaná data. Z tohoto důvodu je v aplikaci zahrnuta možnost volby jak přesného (avšak časově náročného), tak rychlého (ale pouze orientačního) způsobu zjištění cen zboží.

Počet produktů v katalogu Heureka.cz přesahuje 20 miliónů, přičemž lze předpokládat, že velká většina z nich není prodávaná s vysokou frekvencí. Například samotná kategorie *Notebooky* v současnoti obsahuje celkem 3120 položek, z většiny se však jedná již o zastaralé stroje, o které již ze strany zákazníků pravděpodobně není velký zájem.

K vyfiltrování relevantních produktů lze využít značky *TOP Produkt*, kterou Heureka.cz uděluje až 50 aktuálně nejprodávanějším položkám ve vybraných populárních kategoriích. Tento filtr je rovněž součástí implementované aplikace, zachována je však i možnost získat data o všech produktech.

4.3 Použité technologie

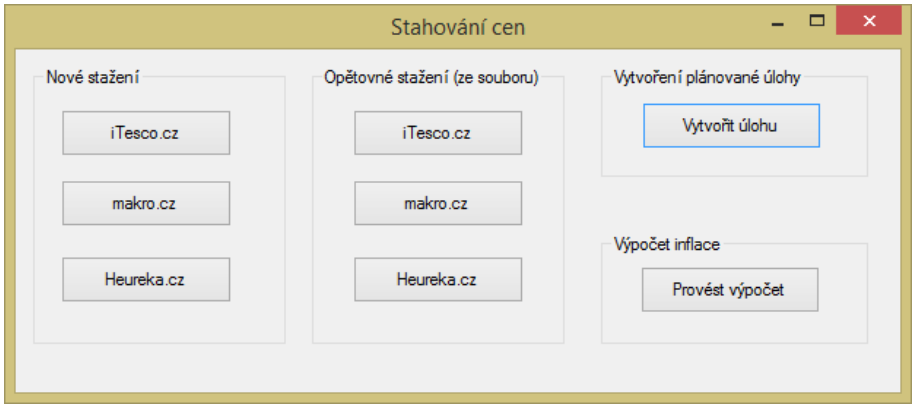
Aplikace byla vytvořena výhradně ve vysokoúrovňovém a objektově orientovaném programovacím jazyce C#. Důvodem k jeho výběru byla především autorova osobní motivace seznámit se s dosud neznámou moderní technologií, možnost relativně jednoduché tvorby grafického rozhraní a existence knihoven pro snadné parsování jazyka HTML.

Ke tvorbě aplikace bylo využito vývojové prostředí Microsoft Visual Studio Community 2015 s verzí jazyka C# 6.0 a .NET Frameworkem 4.6, který je vyžadovaný pro tvorbu a chod aplikací. Testována byla na dvou strojích s operačními systémy Windows 8.1 a Windows 8.0.

V aplikaci je využita volně dostupná externí knihovna HtmlAgilityPack, která obecně slouží jako DOM parser (= syntaktický analyzátor objektových modelů XML a HTML dokumentů). Výhodou této knihovny je její vyšší tolerance vůči nevalidním HTML dokumentům a podpora jazyka XPath, díky kterému lze snadno manipulovat s HTML elementy, jejich hodnotami nebo atributy. Z .NET Framework byly použity například knihovny určené tvorbu formulářů, práci s dotazovacím jazykem LINQ, použití regulárních výrazů nebo vytváření plánovaných úloh.

4.4 Návrh

Aplikace je tvořena přibližně 50 třídami, které jsou logicky členěny do 3 logicky odlišných vrstev - prezentační, aplikační a datové.



Obrázek 4.5: Hlavní obrazovka (MainForm)

Úkolem prezentační vrstvy je zobrazení možných akcí uživateli a zprostředkování uživatelských vstupů vrstvě aplikační. Prezentační vrstva je v aplikaci řešena pomocí formulářů, které jsou součástí .NET Framework.

Aplikační vrstva zajišťuje logiku celé aplikace a v této vrstvě je soustředěna většina všech tříd. Tyto třídy se starají zejména stahování, vytěžování, ukládání a zpracování dat ze stažených stránek.

Datová vrstva je v aplikaci reprezentována objekty kategorií a produktů, které jsou při běhu programu udržovány v paměti a po zpracování uloženy na disk.

4.5 Funkce

4.5.1 Spuštění programu

Hlavní funkce *main*, která je volána při každém spuštění aplikace, se nachází ve třídě *Program*. Při standardním uživatelském spuštění tato funkce vytvoří novou instanci formulářové třídy *MainForm*(4.5), která zobrazí uživateli hlavní možnosti popsané v následujících podsekcích.

Aplikaci je však možné spustit i s parametrem, kterým je adresa na předem vytvořený konfigurační soubor (vysvětleno dále). V případě tohoto způsobu spuštění běží program na pozadí a uživatel je upozorněn až na jeho dokončení.

4.5.2 Stahování dat

K zahájení stahování dat dochází po stisknutí jednoho z tlačítek umístěných v horních dvou boxech (označených *Nové stažení* a *Opětovné stažení (ze souboru)*). Po kliknutí na kterékoliv z těchto tlačítek dojde k vytvoření a spuštění controlleru příslušného portálu, který celý proces organizuje. Samotné staho-

vání dat je pak řešeno třídou *Downloader*, která využívá třídu *WebClient*, jenž je součástí .NET Framework.

4.5.3 Zpracování stažených dat

Zpracování dat ze stažených webových stránek je jádrem celé aplikace a zabývá se jím tedy hned několik tříd. Každý portál má vlastní třídy nazvané *CategoryParser* a *ProductParser* a portál Heureka.cz navíc vlastní ještě třídu *MainPageParser*.

Úlohou těchto tříd je získání informací ze stažených HTML stránek, k čemuž využívají především tří technik - adresování pomocí jazyka XPath, hledání vzorů pomocí regulárních výrazů a operací s textovými řetězci. K adresaci pomocí jazyka XPath je využita již zmíněná knihovna HtmlAgilityPack, díky které je možné z XML (a tedy i HTML) dokumentů vybírat konkrétní elementy a pracovat s jejich hodnotami a atributy.

Analýzou zdrojového kódu úvodní stránky portálu Heureka.cz bylo například zjištěno, že odkazy a názvy hlavních kategorií se nacházejí v blokovém elementu *div* s atributem *"id=hp-categories"* a jsou obaleny nadpisem třetí úrovně *<h3>*. XPath výraz, jehož použitím je možné získat kolekci všech uzlů s informacemi o hlavních kategoriích, může mít podobu *\\div[id='hp-categories']\\h3*. Konkrétní implementace tohoto parsování je v aplikaci řešena statickou třídou *NodeParser*.

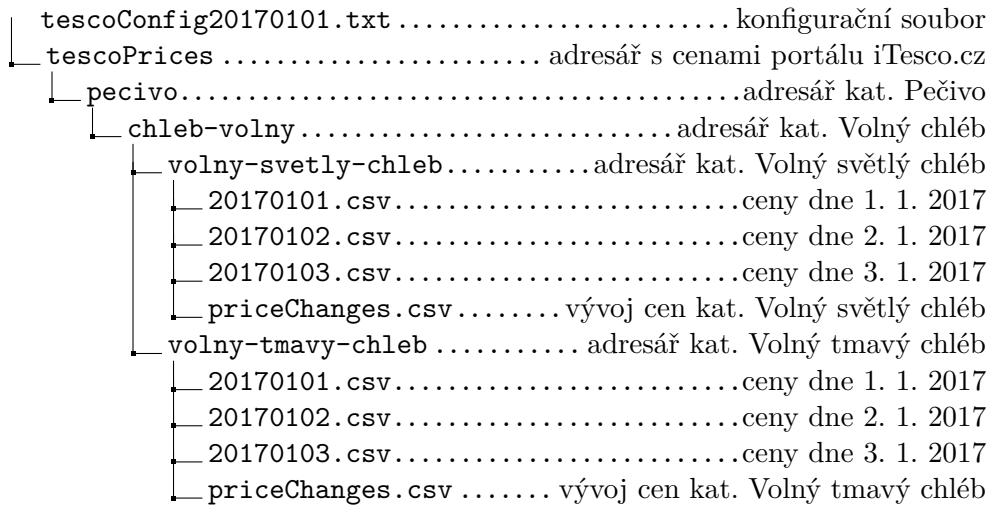
4.5.4 Ukládání zpracovaných dat

Aby bylo možné zpracovaná data dále využít, je nezbytné zajistit jejich perzistenci a pro tento úkol připadaly v úvahu především 2 varianty - ukládání do textových souborů a tradiční SQL databáze.

Hlavní výhodou využití databází je především přehledná a jednoduchá manipulace s daty pomocí SQL dotazů, podmínkou je však samotná přítomnost databázového serveru a jeho správné konfigurace. Z tohoto důvodu bylo zvoleno řešení ukládající data do souborů organizovaných pomocí adresářové struktury do uživatelem zvolené složky.

K ukládání informací dochází po každém stažení a zpracování dat z internetu. V případě, že se uživatel rozhodne jednorázově stáhnout vybrané kategorie některého z portálů, je po jejich výběru nejprve vytvořen konfigurační soubor nesoucí jejich informace (id, název, odkaz, rodič, typ, označení ke stažení) a následně složka nesoucí název internetového portálu s příponou *Prices* (tedy například *heurekaPrices*).

V dalším kroku je v této složce vytvořena adresářová struktura odpovídající všem stahovaných kategoriím, přičemž názvy složek odpovídají jednoznačným identifikátorům daných kategorií. Pokud by například uživatel stahoval vývoj cen kategorie *Volný světlý chléb* (id *volny-svetly-chleb*) nacházející se v hierarchii portálu iTesco.cz pod kategoriemi *Volný chléb* (id *chleb-volny*)



Obrázek 4.6: Struktura adresářů s cenami po 3 staženích

a *Pečivo* (id *pecivo*), nalezl by veškeré ceny kýžené kategorie v souborech umístěných ve složce *user/tescoPrices/pecivo/chleb-volny/volny-svetly-chleb*.

Tyto soubory mají jasně definovaný název i strukturu. Název souboru nese tvar *yyyyMMdd.csv*, kde písmena před koncovkou odpovídají standardním parametrům řetězce formátu data a času, který v reprezentuje den zápisu. V případě každodenního stahování dat během ledna roku 2017 tedy ve složce vznikne v každé stahované kategorii 31 souborů s koncovkou csv, jejichž názvy budou nést hodnoty 20170101 až 20170131. O zápis těchto souborů se stará třída *PriceFileWriter*.

Při každém uložení těchto souborů je však navíc složka automaticky prozkoumána a pokud obsahuje alespoň dva tyto soubory, dojde k vytvoření dalšího souboru s názvem *priceChanges.csv*, ve kterém jsou zaznamenány relativní změny cen celé kategorie po každém stažení. Tyto změny cen jsou vypočteny třídou *PriceChangeFileWriter*, která využívá poměrně intuitivní algoritmus - otevře dva soubory, jejichž názvy odpovídají nejnovějším datům, načte identifikátory i ceny zboží z obou souborů, získá relativní změny cen zboží přítomného v obou souborech a z nich nakonec vypočte geometrický průměr, který zaznamená do souboru *priceChanges.csv*. Výhodou tohoto algoritmu je fakt, že do vývoje cen umožňuje zahrnout i nově dostupné zboží a dokáže pružně reagovat na změnu sortimentu.

Na konci 3. 1. 2017 by adresářová struktura při sledování kategorií *Volný světlý chléb* a *Volný tmavý chléb* vypadala způsobem znázorněným na obrázku 4.6.

4.5.5 Výpočet meziměsíčního růstu cen

Hlavním smyslem této aplikace je zpracování dat do vhodného tvaru pro další použití. Základní časovou jednotkou pro měření inflace bývá zpravidla 1 měsíc, avšak data v souboru *priceChanges.csv* obsahují informace o cenách získané s výrazně kratší periodou, a proto je nutné je dále převést.

Po kliknutí na některé z tlačítek v boxu s názvem *Výpočet inflace* uživatel zvolí konfigurační soubor příslušného portálu, načež mu bude nabídnut seznam kategorií, které byly stahovány. S vybranými kategoriemi následně pracuje třída *InflationCalculator*, která načte obsah souboru *priceChanges.csv*, uspořádá údaje o změnách cen do měsíčních úseků a provede jejich pronásobení.

Pokud by v například v průběhu dubna bylo v různé dny zaznamenáno celkem 30 měření, během kterých by naměřené hodnoty byly každý den vždy o 0,1 % vyšší, získaná meziměsíční inflace by odpovídala přibližně 3,04 %.

Takto získaná data jsou na závěr nabídnuta uživateli k uložení do nového souboru, jenž slouží jako podklad pro webovou aplikaci poskytující krátkodobou predikci.

4.5.6 Opakované stažení

V rámci tvorby aplikace byla snaha o (z pohledu uživatele) maximální zjednodušení procesu opakovaného stahování cen. Z tohoto důvodu aplikace nabízí možnost vynechat celý proces výběru kategorií ke stahování a rovnou umožňuje stáhnout ceny již dříve stahovaných.

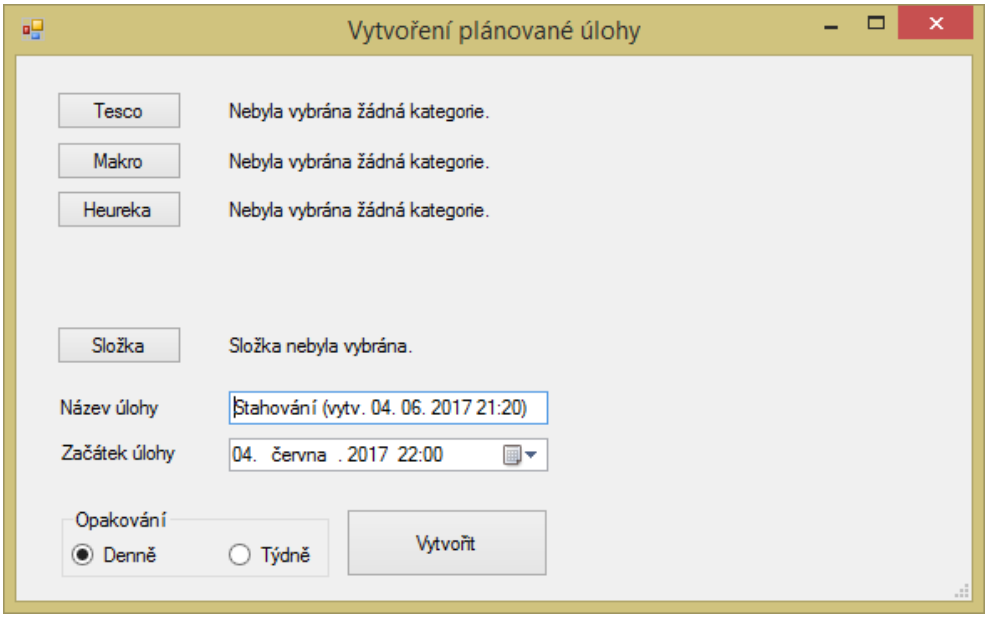
Tohoto scénáře lze docílit kliknutím na některé z tlačítek v boxu *Opětovné stažení (ze souboru)*, které uživateli nabídne k výběru již existující konfigurační soubor (obsahující informace o kategoriích). Obsah tohoto souboru je načten jednou ze tříd nesoucích název *FileLoader* (každý portál má vlastní) a ihned je zahájeno stahování a zpracování dat jako v případě nového stažení.

4.5.7 Automatické stahování

Posledním způsobem, jakým je možné data z internetu získávat, je využití automatického stahování. Operační systém Windows disponuje nástrojem s názvem Plánovač úloh, který slouží k pravidelnému spouštění akcí (často například aktualizací, antivirových kontrol a podobně), čehož lze využít i v této aplikaci.

K vytvoření plánované úlohy dochází po stisknutí tlačítka *Vytvořit úlohu*, po kterém dojde k otevření nového formulářového okna třídy *TaskCreatorForm*, které dává uživateli možnost výběru stahovaných kategorií ze všech portálů zároveň, nastavení počátku spuštění a volby periody opakování (4.7).

O samotné vytvoření této úlohy se stará třída *TaskController*, která nejprve vytvoří nový soubor obsahující jména konfiguračních souborů zvolených portálů a následně založí novou plánovanou úlohu, které pošle tento nově



Obrázek 4.7: Vytvoření plánované úlohy

vzniklý soubor jako parametr. Spuštění aplikace s tímto parametrem bude mít za následek chod na pozadí, během kterého třída *TaskRunner* přečte obsah souboru, načte jednotlivé konfigurační soubory a zadá pokyn ke stahování vybraných kategorií.

Podmínkou pro správný chod této funkcionality je povinnost ponechat spouštěcí soubor i veškeré konfigurační soubory v jejich původním umístění. Pokud dojde k přesunu spouštěcího souboru, úloha proběhne neúspěšně a bez informování uživatele, pokud dojde k přesunu či smazání některého z konfiguračních souborů, bude uživatel při spuštění informován oknem se zprávou.

Takto vytvořenou plánovanou úlohu je možné smazat či upravit přímo pomocí nástroje Plánovač úloh, který je dostupný například ze systémového příkazového řádku příkazem *taskschd.msc*.

4.6 Případy použití

4.6.1 První stažení

Portály makro.cz a iTesco.cz

Prvním scénářem, kterého může uživatel využít, je jednorázové stažení cen všech produktů uživatelem zvolených kategorií. Jeho průběh je shodný pro portály makro.cz a iTesco.cz.

1. Uživatel zahájí akci kliknutím na tlačítko portálu v okně s titulkem Nové stažení.

2. Aplikace zobrazí uživateli dialogové okno s výběrem složky určené pro stahování souborů.
3. Uživatel zvolí cílovou složku.
4. Aplikace stáhne hlavní stránku portálu a získá odkazy a jména všech kategorií, které nabídne uživateli v novém okně.
5. Uživatel vybere kategorie, o které má zájem.
6. Aplikace stáhne ceny zboží z vybraných kategorií a uloží je na disk.
7. Aplikace vytvoří konfigurační soubor obsahující uživatelem zvolené kategorie, díky čemuž bude v budoucnu možné stažení opakovat.

Heureka.cz

1. Uživatel zahájí akci kliknutím na tlačítko Heureka.cz v okně Nové stažení.
2. Aplikace zobrazí uživateli dialogové okno s výběrem složky určené pro stahování souborů.
3. Uživatel zvolí cílovou složku.
4. Aplikace stáhne hlavní stránku portálu a získá odkazy a jména hlavních kategorií, které zobrazí uživateli.
5. Uživatel vybere hlavní kategorie, které mohou obsahovat podkategorie, o které má zájem.
6. Aplikace stáhne a vytvoří strukturu podkategorií zvolených hlavních kategorií.
7. Uživatel vybere podkategorie, o které má zájem, a zároveň zvolí typ produktů (všechny / TOP) a výběr typu ceny (pouze minimální a maximální / kompletní).
8. Aplikace stáhne ceny zboží z vybraných kategorií a uloží je na disk.
9. Aplikace vytvoří konfigurační soubor obsahující uživatelem zvolené kategorie, díky čemuž bude v budoucnu možné stažení opakovat.

4.6.2 Opětovné stažení

1. Uživatel zahájí akci kliknutím na tlačítko Opakovat stahování.
2. Aplikace nabídne uživateli dialogové okno pro výběr konfiguračního souboru.
3. Uživatel vybere dříve vytvořený konfigurační soubor.
4. Aplikace načte kategorie z konfiguračního souboru a stáhne a uloží ceny zboží, které se v kategoriích aktuálně nacházejí.

4.6.3 Plánované stažení

1. Uživatel zahájí akci kliknutím na tlačítko Plánované stažení.
2. Aplikace zobrazí uživateli okno s možnostmi.
3. Uživatel zvolí portál, ze kterého má zájem stahovat data.
4. Aplikace stáhne hlavní stránku portálu a získá odkazy a jména všech kategorií, které nabídne uživateli v novém okně.
5. Uživatel vybere kategorie, o které má zájem.
6. V případě zájmu o pravidelné stahování z dalšího portálu přejde opět ke kroku č. 3.
7. Uživatel klikne na tlačítko Složka.
8. Aplikace zobrazí dialogové okno s výběrem složky.
9. Uživatel vybere cílovou složku.
10. Uživatel volitelně nastaví název úlohy, její počátek a periodu spouštění.
11. Uživatel klikne na tlačítko vytvořit.
12. Aplikace vytvoří v cílové složce konfigurační soubory pro všechny vybrané portály.
13. Aplikace vytvoří soubor taskConfigYYYYMMDD.txt obsahující názvy konfiguračních souborů jednotlivých kategorií.
14. Systém při aktivaci plánované úlohy spustí aplikaci s parametrem odpovídajícím cestě k souboru taskConfigYYYYMMDD.txt.
15. Aplikace načte obsah souboru taskConfigYYYYMMDD.txt a zjistí názvy konfiguračních souborů všech vybraných portálů.
16. Aplikace stáhne a uloží ceny zboží ze všech zvolených kategorií ze všech vybraných portálů.

Kapitola 5

Webová aplikace

Druhou implementační částí této práce bylo vytvoření webové aplikace, jejíž úkolem je zobrazení dat získaných z nástroje ukládajícího data o cenách a využití těchto dat pro tvorbu krátkodobé prognózy jednotlivých složek inflace. Vzhledem k tomu, že data získaná z předchozího nástroje reprezentují vývoj meziměsíční inflace, budou předpovězené hodnoty rovněž odpovídat dalšímu očekávanému meziměsíčnímu vývoji.

5.1 Použité technologie

Webová aplikace zobrazující data byla vytvořena za pomoci webového aplikačního frameworku Shiny. Tento framework využívá ke své práci programovací jazyk R a s jeho využitím lze vytvářet interaktivní webové aplikace i bez znalosti jazyků jako jsou HTML, CSS nebo Javascript.

Aplikace vytvořené s Shiny jsou vždy tvořeny dvěma komponentami - serverovým skriptem a definicí uživatelského rozhraní. V serverové části se nachází logika načítání dat, výpočtů jejich predikce a generování grafů, grafická část pak určuje samotný vzhled aplikace.

5.2 Spuštění

Webová aplikace využívající Shiny framework může být spuštěna dvěma způsoby - na vlastním serveru, nebo v na portále Shinyapps.io. V obou případech je nutnou podmínkou pro spuštění přítomnost vývojového prostředí R (například RStudio).

5.2.1 Shinyapps.io

Shinyapps.io funguje jako platform as a service (PaaS), tedy jako cloudová platforma umožňující spouštění nahraných aplikací. Portál nabízí několik různých plánů lišících se (kromě ceny) maximálním počtem aplikací, jejich ča-

sovým využitím a podporou. Základní balíček, který je zdarma, umožňuje spuštění až pěti aplikací a 25 aktivních hodin měsíčně, což je pro potřeby této práce prozatím postačující.

Po vytvoření zdrojových souborů a založení uživatelského účtu lze postup spuštění následně rozdělit do tří jednoduchých kroků. Prvním z nich je instalace a načtení knihovny *rsconnect* v prostředí R, které lze vyvolat následujícími dvěma příkazy:

```
install.packages('rsconnect')
library(rsconnect)
```

Dalším krokem, je zadání příkazu s vygenerovaným tokenem, který je dostupný po přihlášení do osobního účtu na portálu Shinyapps.io. Podoba tohoto příkazu, jehož zadáním se provede autorizace počítače se serverem, je následující:

```
rsconnect::setAccountInfo(name="<ACCOUNT>",
token="<TOKEN>", secret="<SECRET>")
```

Finálním krokem je odeslání zdrojových souborů na server, čímž dojde k očekávanému spuštění. Toho lze docílit zvoláním tohoto příkazu, jehož parametrem je cesta k adresáři obsahujícího zdrojové soubory:

```
rsconnect::deployApp('path')
```

Vytvořená aplikace spuštěná tímto způsobem je aktuálně dostupná na adrese `https://predpovedinflace.shinyapps.io/shiny/`.

5.2.2 Lokální spuštění

Alternativou, která byla při práci využita zejména pro testování, je spuštění aplikace na lokálním serveru. Toho je možné dosáhnout nastavením aktuálního pracovního adresáře na adresář se zdrojovými soubory a zvoláním funkce *runApp*().

```
setwd('path')
shiny::runApp()
```

5.3 Tvorba predikce

Vytvořená aplikace vytváří předpověď pomocí ARIMA modelů popsaných v předchozí části této práce, k čemuž využívá funkce z balíčku *forecast* využitelného v statistickém programu R.

Samotné vytvoření modelu je řešeno 2 způsoby - vlastním zadáním parametrů *ARIMA(p,d,q)(P,D,Q)* a automatickým nalezením nejvhodnějšího modelu. Pro manuální výběr parametrů slouží funkce *Arima()*, která je využita takto:

```
dm <- read.csv(file)
y <- ts(dm$Value, f=12)

fit <- Arima(y, order=c(p, d, q),
seasonal = list(order = c(P, D, Q),
period = 12))
```

Vhodný model lze získat díky funkci *auto.arima*(), která prochází veškeré modely (s volitelnou maximální hodnotou parametrů) a vrací ten nejlepší z hlediska informačních kritérií.

```
dm <- read.csv(file)
y <- ts(dm$Value, f=12)

fit <- auto.arima(y)
```

Po vytvoření modelu je již možné získat předpověď pro zvolený počet nadcházejících časových období, ke které je využita funkce *forecast.Arima()*:

```
res <- forecast.Arima(fit, h=months)
```

kde h je parametr značící počet období, pro které má být předpověď vytvořena. Součástí výsledku je kromě bodového odhadu i konfidenční interval (defaultně 95% a 80%), který je při vykreslení (například funkcí *plot()*) zobrazen i graficky.

5.4 Použití

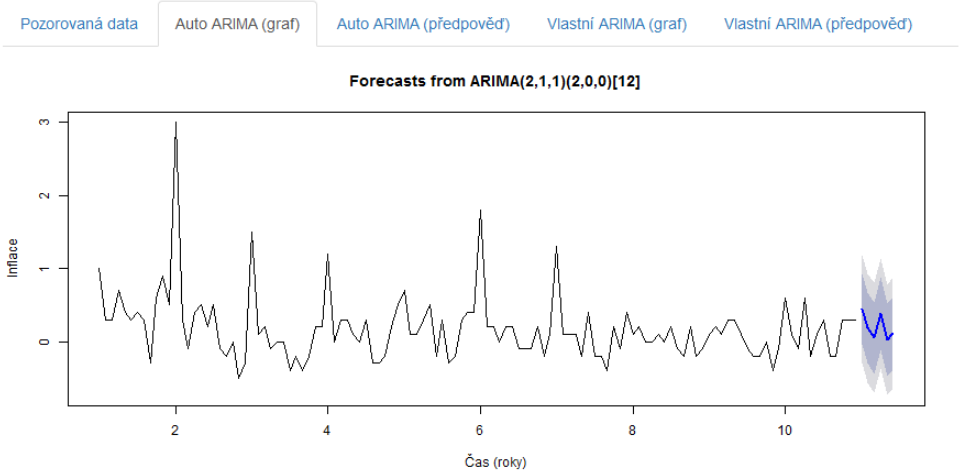
Aplikace je graficky rozdělena na dvě části. Po levé stranně okna se nachází postranní panel s tlačítkem určeným k nahrání souboru a políčka sloužící k zadání parametrů pro předpověď.

Převážnou většinu části však zabírá okno s 5 kartami (záložkami), které po rozkliknutí zobrazí okna ukazující:

- Uživatelem vložená data ve formě grafu
- Krátkodobou předpověď jejich budoucího vývoje ve formě grafu (s automaticky vypočteným modelem)
- Krátkodobou předpověď jejich budoucího vývoje ve formě tabulky (s automaticky vypočteným modelem)
- Krátkodobou předpověď jejich budoucího vývoje ve formě grafu (s modelem získaným dle uživatelem zadaných parametrů)
- Krátkodobou předpověď jejich budoucího vývoje ve formě tabulky (s modelem získaným dle uživatelem zadaných parametrů)

Období	Bodový odhad	Lo 80	Hi 80	Lo 95	Hi 95	Skutečnost
Leden 2017	0.442	-0.040	0.923	-0.294	1.177	0.8
Únor 2017	0.187	-0.301	0.675	-0.559	0.934	0.4
Březen 2017	0.048	-0.446	0.541	-0.707	0.803	0.0
Duben 2017	0.387	-0.107	0.882	-0.369	1.144	0.0
Květen 2017	0.025	-0.471	0.520	-0.733	0.782	?
Červen 2017	0.119	-0.376	0.614	-0.638	0.876	?

Tabulka 5.1: Hodnoty predikované inflace



Obrázek 5.1: Graf predikované inflace

Uživatel stisknutím tlačítka na levé straně vybere a nahraje vygenerovaný soubor ve formátu CSV, čímž automaticky dojde k vykreslení všech karet dle zadaných parametrů. Jejich případná změna po nahrání souboru se projeví až po stisknutí tlačítka Obnovit.

5.5 Ukázka funkce

Pro identifikování sezónní složky je nezbytné, aby statistický vzorek zachycoval minimálně roční vývoj, pro přesnější výsledky je však samozřejmě vhodné, aby vzorek obsahoval data za několik let, která však během tvorby této práce nebylo možné nasbírat.

Na místo toho byla pro demonstraci použita historická oficiální data meziměsíční inflace z období ledna 2007 do prosince 2016, která byla predikována na dalších 6 měsíců. Výsledkem je vytvořená tabulka (Tabulka 5.1) obsahující bodové odhady a 80% a 95% konfidenční intervaly a také graf (Obrázek 5.1), který tyto hodnoty zobrazuje pomocí různých odstínů modré barvy.

Závěr

Cílem teoretické části této práce bylo popsání jednotlivých složek inflace a vytvoření přehledu webových portálů, na kterých je možné získat data o cenách zboží, které do těchto složek patří. Tento cíl byl naplněn ve druhé kapitole, v níž byly stručně rozebrány jednotlivé sekce spotřebního koše, ke kterým byly přiřazeny webové stránkami poskytující informace o cenách odpovídajících produktů.

Hlavním cílem praktické části následně bylo navržení a implementace dvou nástrojů sloužících k zobrazení a tvorbě krátkodobé predikce vývoje cen jednotlivých složek inflace na základě dat stažených z internetu.

Úkolem první aplikace bylo umožnit průběžné stahování a zpracování dat ze tří webových portálů - Heureka.cz, iTesco.cz a makro.cz, čehož bylo v práci dosaženo. Aplikace byla napsána v jazyce C#, disponuje grafickým rozhraním, nabízí uživateli konkrétní volbu kategorií ke stažení, umožňuje jejich rychlé znovustažení a díky systémové funkci Plánovač úloh dokonce zprostředkovává automatické a pravidelné stahování cen produktů na pozadí. Ze získaných dat tento nástroj následně dokáže zjistit měsíční vývoj cen produktů jednotlivých kategorií a exportovat jej do souboru, který slouží jako podklad pro webovou aplikaci, jejíž vytvoření bylo druhým hlavním cílem této práce.

Cílem této webové aplikace je možnost zobrazit data exportovaná z předchozího nástroje a poskytnout jejich krátkodobou prognózu. K naplnění tohoto cíle je využito webového frameworku Shiny, který zajišťuje nahrání souboru, interakci s uživatelem a grafickou stránku. Samotné zobrazení dat a tvorba jejich predikce jsou pak řešené s pomocí statistického programu R.

Navzdory naplněným cílům stanoveným v zadání práce však stále existuje velký prostor pro rozšíření a vylepšení obou aplikací. Do budoucna by bylo vhodné například umožnit sledování vývoje cen z více portálů či lokalit, doplnit webovou aplikaci o další možné pokročilé metody predikce nebo zpříjemnit uživatelské prostředí.

Literatura

[1] Česká národní banka: Cílování inflace v ČR [online]. 2017, [cit. 2017-05-20]. Dostupné z: https://www.cnb.cz/cs/menova_politika/cilovani.html

[2] Česká národní banka: Měnověpolitické nástroje - Zprávy o inflaci [online]. 2017, [cit. 2017-05-20]. Dostupné z: https://www.cnb.cz/cs/menova_politika/zpravy_o_inflaci/

[3] Česká národní banka: Jak vzniká prognóza [online]. 2017, [cit. 2017-05-20]. Dostupné z: https://www.cnb.cz/cs/faq/jak_vznika_prognoza.html

[4] Pavelka, T.: *Mikroekonomie základní kurz*. Praha: Melandrium, třetí vydání, 2007, ISBN 8086175584.

[5] Český statistický úřad: Index cen výrobců - metodické poznámky [online]. [cit. 2017-05-20]. Dostupné z: https://www.czso.cz/csu/czso/ipc_cr

[6] Holman, R.: *Ekonomie*. C. H. Beck, třetí vydání, 2002, ISBN 8071796816.

[7] Český statistický úřad: Indexy spotřebitelských cen (metodická příručka pro uživatele) [online]. 2017, [cit. 2017-05-20]. Dostupné z: https://www.czso.cz/documents/10180/26822363/manual_isc_2017.pdf

[8] Český statistický úřad: Spotřební koš pro výpočet indexu spotřebitelských cen od ledna 2017 [online]. 2017, [cit. 2017-05-20]. Dostupné z: https://www.czso.cz/documents/10180/23195394/spot_kos2017.pdf

[9] Cavallo, A.: *Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers[online]*. [cit. 2017-05-20]. Dostupné z: http://www.mit.edu/~afc/papers/Cavallo_Online_Offline.pdf

[10] Bańbura, M.; Giannone, D.; Modugno, M.; aj.: *Now-casting and the real-time data flow[online]*. [cit. 2017-05-20]. Dostupné z: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf

[11] Cavallo, A.: *Online and official price indexes: Measuring Argentina s inflation[online]*. [cit. 2017-05-20]. Dostupné z: http://www.mit.edu/~afc/papers/Cavallo-Argentina-JME.pdf

[12] Hindls, J.: *Statistika pro ekonomy*. Praha: Professional Publishing, třetí vydání, 2007, ISBN 9788086946436.

[13] Central Bank of Iceland: *Forecast errors in Central Bank of Iceland inflation forecasts[online]*. 2009, [cit. 2017-05-20]. Dostupné z: https://www.cb.is/lisalib/getfile.aspx?itemid=7099

[14] Hančlová, J.; Tvrdý, L.: *Úvod do analýzy časových řad*. Ostrava: VŠB-TU, 2003.

[15] Danel, R.: *Predikce časové řady pomocí autoregresního modelu[online]*. [cit. 2017-05-20]. Dostupné z: http://homel.vsb.cz/~dan11/publikace/Danel_Autoregresni_model_predikce_casovych_rad.pdf

[16] Křivý, I.: *Analýza časových řad [online]*. [cit. 2017-05-20]. Dostupné z: https://publi.cz/download/publication/45?pc=1

[17] Lawson, R.: *Web Scraping with Python*. Packt Publishing, Ltd., první vydání, 2015, ISBN 9781782164364.

Příloha A

Seznam použitých zkratek

CPI Consumer Price Index

PPI Production Price Index

ČNB Česká národní banka

ČSÚ Český statistický úřad

MVC Model-View-Controller

DOM Document Object Model

XPath XML Path Language

GUI Graphical User Interface

XML Extensible Markup Language

AR AutoRegressive

XML Moving Average

ARMA AutoRegressive Moving Average

ARIMA AutoRegressive Integrated Moving Average

SARIMA Seasonal AutoRegressive Integrated Moving Average

HTML Seasonal AutoRegressive Integrated Moving Average

CSS Cascading Style Sheets

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresář se spustitelnou formou implementace
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `doc` ..... dokumentace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
  - `thesis.ps` ..... text práce ve formátu PS