UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VYUŽITIE METÓD DATA MININGU NA PREDPOVEDANIE INFLÁCIE S VYUŽITÍM DÁT Z E-SHOPOV

DIPLOMOVÁ PRÁCA

2019 Bc. Patrik BACHAN

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VYUŽITIE METÓD DATA MININGU NA PREDPOVEDANIE INFLÁCIE S VYUŽITÍM DÁT Z E-SHOPOV

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ing. Eva Hromádková, PhD.

Bratislava 2019 **Bc. Patrik BACHAN**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Využitie metód data miningu na predpovedanie inflácie s využitím dát z e-shopov
Alternative nowcast of inflation using prices from e-shops

Anotácia: V posledných rokoch sa veľká časť realizácie spotrebiteľských nákupov presunula do online prostredia. Viacero národných štatistických úradov tento trend spozorovalo a v tejto súvislosti začalo experimentovať s použitím dát z e-shopov na meranie inflácie (Breton et al. 2016, Horrigan 2013, alebo Nygaard 2015). V rámci navrhovanej diplomovej práce bude študent pracovať na empirickej analýze možnosti využitia týchto dát pre zlepšenie krátkodobej out-of-sample predpovede inflácie vo vybraných kategóriách tovarov. Tri navrhované potenciálne metodické prístupy zahŕňajú : (1) dynamické faktorové modely kombinované s odhadom koeficientov pomocou LASSO metódy (Li and Chen, 2014), (2) široko spektrálne VAR modely s použitím Bayesian shrinkage (Banbura et al., 2010), alebo (3) elastic net regresia založená na algoritmoch strojového učenia (Tiffin, 2016).

Vedúci: Mgr. Ing. Eva Hromádková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 24.01.2018

Dátum schválenia: 25.01.2018 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcel poďakovať vedúcej diplomovej práce Mgr. Ing. Eve Hromádkovej, PhD. za umožnenie pracovať na zaujímavej téme, sprostredkovanie dát z ČNB, odborné vedenie, cenné myšlienky a v neposlednom rade za nesmiernu ochotu pri konzultáciách. Ďalej by som sa chcel poďakovať mame za jazykovú úpravu a priateľke za pomoc pri formálnej úprave práce. Na záver by som chcel vyjadriť vďaku celej rodine a všetkým priateľom, ktorí ma pri písaní práce podporovali.

Abstrakt

BACHAN, Patrik: Využitie metód data miningu na predpovedanie inflácie s využitím dát z e-shopov [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ing. Eva Hromádková, PhD., Bratislava, 2019, 78 s.

Cieľom tejto diplomovej práce je zostrojiť niekoľko alternatívnych spotrebiteľských indexov aplikovaním regularizácie a kategorizácie a pomocou nich sledovať cenový vývoj tovarov predávaných online. Ďalším z cieľov je replikovať súčasný CPI a na základe tejto replikácie spraviť predikciu za kratši čas, ako pri určovaní oficiálneho CPI definovaného Českým štatistickým úradom. Autori v práci poukazujú na niektoré neefektívne kroky v procese tvorby a výpočtu indexu spotrebiteľských cien (Consumer Price Index - CPI) a pomocou moderných metód data miningu a strojového učenia ponúkajú alternatívny postup merania CPI v kategóriách "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak". Kvôli vyšším podielom internetových predajov v maloobchodnom sektore a vyššej transparentnosti oficiálne používanej metodológie bol vybratý trh v Českej republike. Práca v poslednej kapitole približuje tri alternatívne postupy zostrojenia CPI na základe dát stiahnutých z troch e-shopov. Postup pre konštrukciu CPI získaný agregáciou produktov po kategóriách má po praktickej stránke najlepšie využitie.

Kľúčové slová: Inflácia, Predikcia, Data mining, Strojové učenie, LASSO Regresia, Metódy oporného bodu

Abstract

BACHAN, Patrik: Alternative nowcast of inflation using prices from e-shops [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ing. Eva Hromádková, PhD., Bratislava, 2019, 78 p.

The aim of the thesis is construction of various alternative consumer indices using data from e-shops and comparison of these results to the actual indices measured by Czech Statistical Bureau. Other goal is to replicate official CPI and, based on this replication, to nowcast the prediction faster. First, authors introduce and describe the process of construction Consumer Price Index (CPI) with the focus on its ineffective elements. Using modern methods of datamining and machine learning algorithms, they are creating an alternative way of measuring CPI in the categories "Foods and non-alcoholic beverages" and "Alcoholic beverages and tobacco". Market in Czech republic has been chosen due to bigger market share of retail purchases in online environment and better transparency of official methodology. There are three alternative approaches of development CPI in the last chapter, that are based on data scraped from three different e-shops. The best approach of development CPI for practical application is the CPI based on aggregation by categories.

Keywords: Inflation, Nowcast, Datamining, Machine learning, LASSO Regression, Support vector machines

Obsah

Zoznam obrázkov **9**

Zoznam tabuliek **11**

Úvod **13**

1 Inflácia **15**

1.1 Motivácia pre skúmanie alternatívnych metód merania inflácie 15

1.2 Definícia inflácie a jej historický vývoj 16

1.3 Súčasná metodológia merania inflácie v Českej republike 17

1.3.1 Index spotrebiteľských cien - CPI 17

1.3.2 Klasifikácia a výber reprezentantov 18

1.3.3 Váhový systém 18

1.3.4 Výber respondentov a zisťovanie spotrebiteľských cien 20

1.3.5 Spôsob výpočtov indexov spotrebiteľských cien 20

1.3.6 Kvalitatívne očistenie 21

1.3.7 Harmonizovaný index spotrebiteľských cien - HICP 25

1.3.8 Jadrová inflácia 25

2 Dáta **27**

2.1 Ceny položiek v spotrebnom koši z merania ČSÚ 27

2.2 Aplikácia na sťahovanie dát z online obchodov 28

2.3 Dáta z Českej národnej banky 29

3 Kategorizácia tovarov **33**

3.1 Metóda oporného bodu 34

3.1.1 Definícia úlohy 35
3.2 Kategorizácia tovarov do ECOICOP kategórií 37

4 Regularizácia **41**
4.1 Vzťah medzi vychýlením (bias) a varianciou v lineárnej regresii 41
4.2 Hrebeňová (ridge) regresia 44
4.2.1 Vzťah medzi vychýlením a varianciou v hrebeňovej regresii . . . 46
4.3 Voľba regularizačného parametra λ 47
4.3.1 Krížová validácia 47
4.4 LASSO regresia 47
4.5 Elastic net regresia 48

5 Alternatívne metódy merania inflácie **51**
5.1 Tvorba spotrebného koša podľa metodiky ČSÚ 51
5.1.1 Určenie reprezentantov 51
5.1.2 Tvorba spotrebného koša na základe dát z troch e-shopov 53
5.1.3 Tvorba spotrebného koša na základe dát z ČNB 57
5.2 Regularizácia pre vytvorenie spotrebného koša 58
5.3 Alternatívny spotrebný kôš podľa kategórií 64

Záver **67**

Literatúra **69**

Príloha A **71**

Príloha B **75**

Zoznam obrázkov

1.1 Náčrt tvorby CPI. Zdroj: [13]. 19
1.2 Metódy kvalitatívneho očistenia. 22

2.1 Počty stiahnutých cenových záznamov v jednotlivých týždňoch od 20. januára 2019 do 22. apríla 2019 29
2.2 Počty unikátnych stiahnutých položiek v jednotlivých mesiacoch v ČNB. 30

3.1 Náčrt metódy oporného bodu v dvojrozmernom priestore. 34
3.2 Trik s jadrom. Zdroj: [19]. 35

4.1 Vzťahy medzi vychýlením a varianciou. Zdroj: [14]. 44
4.2 Vzťahy medzi vychýlením, varianciou a komplexnosťou modelu. Zdroj: [14]. 45

5.1 Histogramy pomeru cien v kategórii "Potraviny a nealkoholické nápoje". 54
5.2 Boxploty pomeru cien v kategórii "Potraviny a nealkoholické nápoje". . 55
5.3 Histogramy pomeru cien v kategórii "Alkoholické nápoje a tabak". . . . 56
5.4 Boxploty pomeru cien v kategórii "Alkoholické nápoje a tabak". 56
5.5 Hodnoty spotrebného koša tvoreného položkami z *itesco.cz* a spotrebného koša ČSÚ. 57
5.6 Porovnanie CPI a inflácie alternatívneho koša z e-shopu *itesco.cz* a oficiálneho spotrebného koša ČSÚ. 58
5.7 Vývin cien spotrebného koša za obdobie január 2017 - marec 2019, hore mesačné súčty, dole ich týždňová interpolácia. 59
5.8 Chýbajúce hodnoty v datasete. 60
5.9 Vývin hodnoty spotrebného koša pomocou LASSO regresie. 61

5.10 Porovnanie CPI a inflácie medzi ČSÚ a modelom LASSO regresie. . . . 63
5.11 Porovnanie CPI a inflácie medzi ČSÚ a modelom hrebeňovej regresie. . 63
5.12 Porovnanie CPI a inflácie medzi ČSÚ a modelom Elastic net regresie. . 64
5.13 Porovnanie CPI a inflácie pre oficiálny a alternatívny spotrebný kôš agregovaný po skupinách v kategórii "Potraviny a nealkoholické nápoje". 66
5.14 Porovnanie CPI a inflácie pre oficiálny a alternatívny spotrebný kôš agregovaný po skupinách v kategórii "Alkoholické nápoje a tabak". . . . 66

Zoznam tabuliek

1.1 Výber reprezentantov. Zdroj: [4]. 19
1.2 Očistenie metódou “BRIDGE OVERLAP“. 24

2.1 Príklad niektorých položiek spotrebného koša definovaného ČSÚ. 27
2.2 Príklad štruktúry dát z e-shopu *kosik.cz*. 28
2.3 Príklad štruktúry dát z e-shopu *itesco.cz* stiahnutých v ČNB. 30
2.4 Popisné charakteristiky po kategóriách na dátach z ČNB. 31

3.1 Príklad kategorizácie vybratých produktov. 38

5.1 Problémové položky pri tvorbe koša z troch e-shopov. 53
5.2 Porovnanie reálneho a nami dopočítaného CPI v kategórii ”Potraviny a nealkoholické nápoje”. 54
5.3 Porovnanie reálneho a nami dopočítaného CPI v kategórii ”Alkoholické nápoje a tabak”. 55
A1 Produkty s príslušnými váhami vybraté LASSO regresiou. 71
B1 Členenie spotrebného koša podľa ČSÚ. 75
B2 Prevod medzi kategóriami na stránke itesco.cz a ECOICOP kategorizáciou. 76

Úvod

Inflácia ako miera rastu cenovej úrovne vybratých tovarov a služieb je jednou z najdôležitejších makroekonomických štatistík pre potreby rozhodovania sa v oblasti hospodárskej politiky. Významne ovplyvňuje monetárnu politiku centrálnych bánk, rozhoduje pri určovaní výšky dôchodkov, sociálnych dávok a slúži mimo iné aj ako ukazovateľ pri vyjednávaní o zvyšovanie platov a miezd.

Samotné meranie inflácie zohráva v makroekonomickom kontexte dôležitú úlohu. Ako najlepšie zachytiť tento rast a aproximovať ho v rámci jedného ukazovateľa? Ako rozumná voľba sa javí Index spotrebiteľských cien (Consumer Price Index - CPI), ktorý udáva percentuálnu zmenu cien na preddefinovanom spotrebnom koši tovarov a služieb.

Proces tvorby takéhoto indexu je komplexný problém, ktorý je v súčasnosti prepracovaný do detailov. Avšak, niektoré prvky tvorby tohto indexu, ako napríklad zbieranie dát, kde väčšina cien je zisťovaná manuálne v predajniach, sú nákladné a ťažkopádne.

Svet sa každým rokom viac digitalizuje, čo platí aj pri správaní spotrebiteľov, ktorí stále viac nákupov realizujú online [5]. Tento trend zachytilo viacero štatistických inštitúcií, ktoré prišli s návrhmi rôznych metodológií, ako online dáta využiť pre inflačné metriky [1], [3]. V našej práci sme sa zamerali na určenie CPI v dvoch kategóriách: "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak". Jedna z hlavných pridaných hodnôt našej práce je zrýchlenie výpočtu inflačného ukazovateľa; zatiaľ, čo zber dát pri určení oficiálneho CPI prebieha prvých 20 dní v mesiaci [4], my vieme pomocou knižníc určených pre dynamické sťahovanie dát v softvéri RStudio stiahnuť dáta a vypočítať relatívne presný odhad (nowcast) index v priebehu pár hodín.

V prvej kapitole sa budeme venovať hlavne úvodu do problematiky a následne rozpracovaniu aktuálnej metodiky tvorby CPI v Českej republike [4].

V druhej kapitole uvádzame popis dát, ktoré sme v diplomovej práci používali.

Ďalej budeme pokračovať spracovaním témy kategorizácie tovarov. Ako vysvetlíme v úvode kapitoly, rozčlenenie na skupiny je dôležité hlavne z dôvodu predpísanej kategorizácie ECOICOP, ktorá vyžaduje podrobné členenie a následne je na toto členenie naviazaný váhový systém. Kvôli tejto problematike predstavíme koncept Metód oporného bodu, ktoré sa pri textovej kategorizácii ukazujú ako spoľahlivý nástroj [12].

V štvrtej kapitole sme rozpracovali tému regularizácie. Jedná sa o regresnú techniku, ktorá efektívne umožňuje zjednodušiť model lineárnej regresie znížením alebo eliminovaním jednotlivých koeficientov [14]. Táto metóda sa ukazuje ako užitočná pri tvorbe alternatívneho spotrebného koša - z množstva vstupných položiek dokáže vybrať reprezentatívnu vzorku produktov, ktoré najlepšie korešpondujú s hodnotou oficiálneho spotrebného koša definovaného Českým štatistickým úradom.

Na záver navrhneme tri potenciálne prístupy pre tvorbu alternatívnych košov a ich vyhodnotenie v kontexte prípravy nowcastu. Budeme vychádzať z popísaných techník v predchádzajúcich kapitolách.

Kapitola 1

Inflácia

1.1 Motivácia pre skúmanie alternatívnych metód merania inflácie

Základný proces merania inflácie vo väčšine štátov ostal nezmenený niekoľko dekád. Veľký počet ľudí pracujúcich pre štatistické úrady navštevuje stovky obchodov na mesačnej alebo polmesačnej báze, aby zozbierali ceny tovarov a služieb z preddefinovaného spotrebného koša. Tieto dáta sú následne spracované a použité na výpočet Indexu spotrebiteľských cien (Consumer Price Index - CPI) alebo iné podobné štatistické indikátory. Navzdory tomu, že je tento proces detailne prepracovaný, v skutočnosti je do značnej miery nákladný, ťažkopádny a často príliš pomalý pre niektorých užívateľov.

Cestou k zlepšeniu metodiky v tomto kontexte je používanie cien zozbieraných online. Zatiaľ čo dáta sú na internete rozptýlené po tisíckach stránok, pokrok v automatizovaných "scraping" softvéroch umožňuje užívateľom sťahovať veľké dátové súbory a používať ich na rozličné aplikácie. Dáta takto zozbierané sú lacné, rýchle a presné, čím sa stávajú ideálnym doplnkom pre tradičné metódy zbierania dát, obzvlášť v kategóriách tovarov, ktoré sú dobre reprezentované online.

V našej práci sa zameriavame na zefektívnenie určenia inflačnej štatistiky CPI v kategóriách "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak" pomocou zberu dát z najväčších internetových obchodov v Českej republike.

1.2 Definícia inflácie a jej historický vývoj

V ekonomickej teórii sa inflácia definuje ako kvantitatívna metrika rastu priemernej cenovej úrovne vybratých tovarov a služieb v ekonomike za určité obdobie [10]. Takáto jednoduchá definícia však skrýva viacero možností, čo vlastne chceme merať a ako to chceme merať. V našej práci sa zameriavame skôr na metodologickú časť. Technologický rozvoj za posledné desaťročia umožňuje zhromažďovať a spracovávať obrovské množstvá dát a z nášho pohľadu je tu priestor pre využitie tohto potenciálu na zefektívnenie určenia inflačných štatistík.

Prečo je inflácia dôležitá? Inflácia nie je iba štatistický koncept. Na inflačných štatistikách záleží viacero vecí, dokonca sú jednými z najdôležitejších oficiálnych štatistík. Markantne ovplyvňujú monetárne politiky centrálnych bánk, určujú príjmy miliónov ľudí, rozhodujú pri určovaní výšok sociálnych dávok a dôchodkov. Sú oporným bodom pri vyjednávaní platu. Vo Veľkej Británii sú dokonca inflačné štatistiky používané na trhoch o objeme niekoľko miliárd, kde sú cenné papiere priamo naviazané na inflačnú mieru (index-linked gilts) [13].

Výraz infácia mal však v minulosti iný význam. Často sa spájal skôr so znehodnotením meny. Podľa [16] by sa ako prvé príklady inflácie mohli uviesť prípady z čias, keď sa platilo zlatom. Kráľ (vláda) nariadil pozbierať všetky zlaté mince, roztaviť ich a zmiešať s inými kovmi, napríklad striebrom alebo meďou. Takto mohol panovník vyrobiť z rovnakého množstva zlata väčší počet mincí, vyplatiť svoje dlhy a následne vrátiť znečistené mince do obehu. Nominálna cena mincí by ostala rovnaká. Týmto by sa zvýšila ponuka peňazí, ale zároveň by klesla relatívna hodnota jednotlivých mincí a teda ľud by potreboval viac mincí na kúpu tovarov a služieb ako pred vydaním nových mincí.

V histórii sa tiež objavili situácie, keď znehodnotenie meny spôsobilo veľké množstvo zlata alebo striebra, ktoré sa naraz dostalo do ekonomiky. Príkladom môže byť publikácia [7] o cenovej revolúcii, keď sa ceny na niektorých miestach v západnej Európe zdvihli v priebehu 150 rokov až 7-násobne. Dôvodom bolo masívne dovážanie zlata z novoobjavených oblastí na západnej pologuli - Ameriky.

V 19. storočí, konkrétne od 1830 až do Americkej občianskej vojny sa prvýkrat v literatúre začína objavovať slovo inflácia. V tomto prípade, ako sme spomenuli vyššie,

však inflácia znamená znehodnocovanie meny. Vyplývalo to z rozvoja bánk v amerických kolóniách, kde banky vydávali bankovky, ktoré bolo možné zameniť za zlato. Stávalo sa však, že banky vydali viac bankoviek ako mohli v skutočnosti preplatiť a tento fakt znepokojoval ekonómov, ktorí tento jav komentovali ako infláciu [2].

S príchodom Keynesovho pohľadu na ekonómiu v roku 1936, ktorý dominoval nasledovných 40 rokov, sa na infáciu začalo pozerať jednoznačnejšie ako na nárast cien, nakoľko boli spochybnené vzťahy medzi ponukou peňazí a cenovou úrovňou [2].

1.3 Súčasná metodológia merania inflácie v Českej republike

Vzhľadom na potreby našej práce sme si na alternatívne meranie inflácie vybrali Českú republiku. Oproti Slovensku majú v Českej republike vyšší podiel predajov na internete [5], transparentnejšiu prezentáciu použitej metodológie a jednoduchší prístup k nameraným dátam. V celej podkapitole sme čerpali z publikácie [4].

1.3.1 Index spotrebiteľských cien - CPI

Na vyjadrenie inflácie vo svete sa pre menovopolitické situácie primárne používa index spotrebiteľských cien (Consumer Price Index - CPI) ([15], str. 20). Podobne je to aj v Českej republike. Reprezentatívnym spôsobom meria v časovom vývoji relatívne zmeny konečných spotrebiteľských cien (vrátane daní) tovarov a služieb platených obyvateľstvom.

Metodika tvorby indexov spotrebiteľských cien je pravidelne aktualizovaná. Najčastejšími zmenami sú úprava váh, výber cenových reprezentatov (viď 1.3.2) a úprava výpočtu cenových indexov. Z dôvodu čo najväčšej aktuálnosti sa váhy menia každé dva roky, výber cenových reprezentantov každý rok a metodika podľa reálnych potrieb alebo požiadaviek Európskej únie.

V roku 2017 prebehla komplexná revízia váhového systému. Váhy sú od januára 2018 zaktualizované na základe výdajov domácností podľa národných účtov z roku 2016, teda sú vypočítané na základe štatistiky rodinných účtov z roku 2016 a ďalších dostupných zdrojov a prieskumov.

1.3.2 Klasifikácia a výber reprezentantov

Od januára 2018 su detailné indexy spotrebiteľských cien počítané na základe novozavedenej klasifikácie ECOICOP (európska klasifikácia individuálnej spotreby podľa účelu), ktorá nahradzuje pôvodnú klasifikáciu (CZ-COICOP – Classification of Individual Consumption by Purpose) [4]. Táto nová klasifikácia závádza podrobnejšie členenie, triedenie výrobkov a služieb do 12 kategórií je však rovnaké ako v predchádzajúcej klasifikácii.

Z dlhodobých pozorovaní výplýva, že ceny určitých homogénnych skupín, vyrobených z rovnakých surovín a materiálov a zhodných úžitkových vlastností sa vyvíjajú vo vysokej vzájomnej korelácii. Z tohto poznatku vychádza princíp spotrebného koša a jeho vzťah k nižšie popísanej váhovej schéme. Každý cenový reprezentant teda predstavuje rovnaký alebo veľmi podobný vývoj tej časti výdajov obyvateľov, ktorá je ako váha reprezentantovi priradená.

Za cenových reprezentantov sú preto vyberané výrobky a služby, ktoré kupuje priemerný obyvateľ Českej republiky. Váhový systém dáva tomuto výberu reprezentatívny rozmer, nakoľko váhy odzkradľujú zastúpenie produktov v priemernej spotrebe.

Pri výbere sa prihliada najmä k váhovému podielu v štruktúre spotreby. Táto informácia sa získava zo štatistík rodinných účtov, čo sú záznamy vybratých rodín o ich spotrebe. Ďalej sa prihliada k stálosti produktov na trhu a ich zameniteľnosti v budúcnosti. Revízia reprezentantov sa robí každý rok a preto aj počet v jednotlivých kategóriách ECOICOP sa môže z roka na rok líšiť.

V Tabuľke 1.1 uvádzame počet reprezentantov v jednotlivých kategóriách pre rok 2018.

1.3.3 Váhový systém

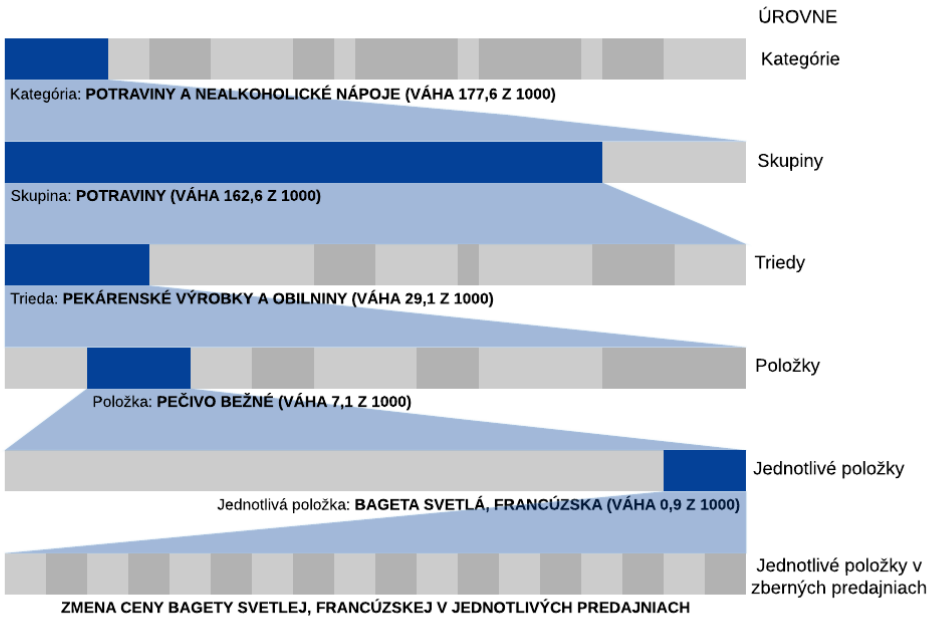
Od januára 2019 sa CPI počíta na aktualizovaných váhach, ktoré vychádzajú z národného konceptu konečnej peňažnej spotreby domácností. Na určenie váh detailných cenových reprezentantov slúži hlavne štatistika rodinných účtov z roku 2016, ale takisto aj iné zdroje a prieskumy.

Aktuálne váhy uvádzame v Tabuľke 1.1 spolu s absolútnymi počtami reprezentantov a na Obrázku 1.1 grafické znázornenie spotrebného koša.

Kategória	Názov kategórie (ECOICOP)	Stále váhy v ‰	Počet reprezentantov
01	Potraviny a nealkoholické nápoje	177,6	158
02	Alkoholické nápoje a tabak	92,2	23
03	Odevy a obuv	41,5	65
04	Bývanie, voda, energie, palivá	257,2	41
05	Nábytok, bytové zariadenie a bežná údržba domácnosti	58,1	79
06	Zdravotníctvo	24,5	20
07	Doprava	101,7	88
08	Pošty a telekomunikácie	28,9	7
09	Rekreácia a kultúra	86,3	107
10	Vzdelávanie	5,7	12
11	Stravovanie a ubytovanie	59,1	42
12	Ostatné tovary a služby	67,2	51
	Úhrnný index spotrebiteľských cien	1000,000	693

Tabuľka 1.1: Výber reprezentantov. Zdroj: [4].

Okrem celkového CPI pre domácnosti sa počítajú individuálne indexy spotrebiteľských cien (životných nákladov) aj pre domácnosti dôchodcov a domácnosti žijúce v Prahe. Reprezentanti v týchto prípadoch ostávaju rovnakí, menia sa váhy.



Obr. 1.1: Náčrt tvorby CPI. Zdroj: [13].

1.3.4 Výber respondentov a zisťovanie spotrebiteľských cien

Spravodajskú sieť respondentov tvoria predajne, hypermarkety, supermarkety obchodné reťazce a ostatné inštitúcie poskytujúce služby. Všetky tieto inštitúcie sú monitorované v 35 okresoch naprieč Českou republikou a v hlavnom meste Prahe. Iba malý zlomok cien sa zisťuje z internetových obchodov[1].

Ceny sa monitorujú mesačne, prvých 20 dní. Ceny sa zbierajú dvomi spôsobmi:

1. **centrálne sledované ceny** - rovnaké ceny pre celú Českú republiku, priemerné ceny zistené z rôznych výkazov ale aj ceny zistené z internetu (napr. poštovné, pohonné hmoty alebo dialničná známka),

2. **miestne sledované ceny** - prevažná väčšina tovarov a služieb je monitorovaná práve týmto spôsobom. 47 pracovníkov v teréne zisťuje ceny osobnou návštevou.

Celkovo sa mesačne zozbiera až 50 000 cien.

1.3.5 Spôsob výpočtov indexov spotrebiteľských cien

Priemerná cena každého reprezentanta sa najprv určuje ako aritmetický priemer na úrovni krajov. Následne sa pre celú Českú republiku vypočíta vážený aritmetický priemer, kde váhy používajú stredné stavy obyvateľstva v krajoch z roku 2016. U niektorých reprezentantov sa priemerná cena určuje odlišným spôsobom, kde sa cena zisťuje centrálne a nepoužívajú sa váhy pre jednotlivé kraje, ale zvolí sa obyčajný aritmetický priemer. Jedná sa napríklad o nájomné v nájomných bytoch, pohonné hmoty alebo lístky na lyžiarsky vlek.

Samotný výpočet indexu spotrebiteľských cien sa následne vypočíta pomocou Laspayresovho vzorca:

$$I = \frac{\sum_{i=1}^{n} \frac{p_{1,i}}{p_{0,i}} p_{0,i} q_{0,i}}{\sum_{i=1}^{n} p_{0,i} q_{0,i}} \cdot 100, \tag{1.1}$$

kde $p_{1,i}$ je cena i-teho tovaru (služby) v sledovanom (bežnom) období, $p_{0,i}$ je cena i-teho tovaru (služby) v základnom období a $q_{0,i}$ zodpovedá váhe i-teho tovaru (služby) v základnom období.

[1] Z internetových obchodov sa podľa ČSÚ zisťuje iba 16 položiek z celkového počtu 692 (medzi takto zisťované položky patrí napr. biela technika, elektronika alebo pobyt v hoteli).

1.3.6 Kvalitatívne očistenie

Kvalitatívne očistenie cenového indexu sa aplikuje pri zámene cenového reprezentanta, ktorý bol zaradený do cenového prieskumu, za iný. Tento prístup sa aplikuje hlavne v troch prípadoch:

1. **Zámena cenového reprezentanta podľa terénneho pracovníka** - Najčastejší prípad výmeny, nastáva, keď pracovník terénneho prieskumu nenájde produkt zodpovedajúci popisu v skúmanej predajni, teda sa prestane v danej predajni vyskytovať, respektíve dôjde k uzavretiu predajne. V tomto prípade tak musí pracovník (po informácii od vedúceho predajne, že ide o trvalú zmenu) zistiť cenu iného, podobného výrobku, alebo cenu zistiť z inej predajne.

2. **Zámena cenového reprezentanta kvôli aktualizácii spotrebného koša** - Revízia spotrebného koša sa koná každoročne, za zmenu môže byť zodpovedný pokles nákupov daného tovaru alebo absencia na trhu.

3. **Kvantitatívne očistenie** - V tomto prípade sa jedná o prípad, keď položka skúmaná v sledovanom období má cenu určenú za inú mernú jednotku ako v predchádzajúcom období.

Metódy kvalitatívneho očistenia

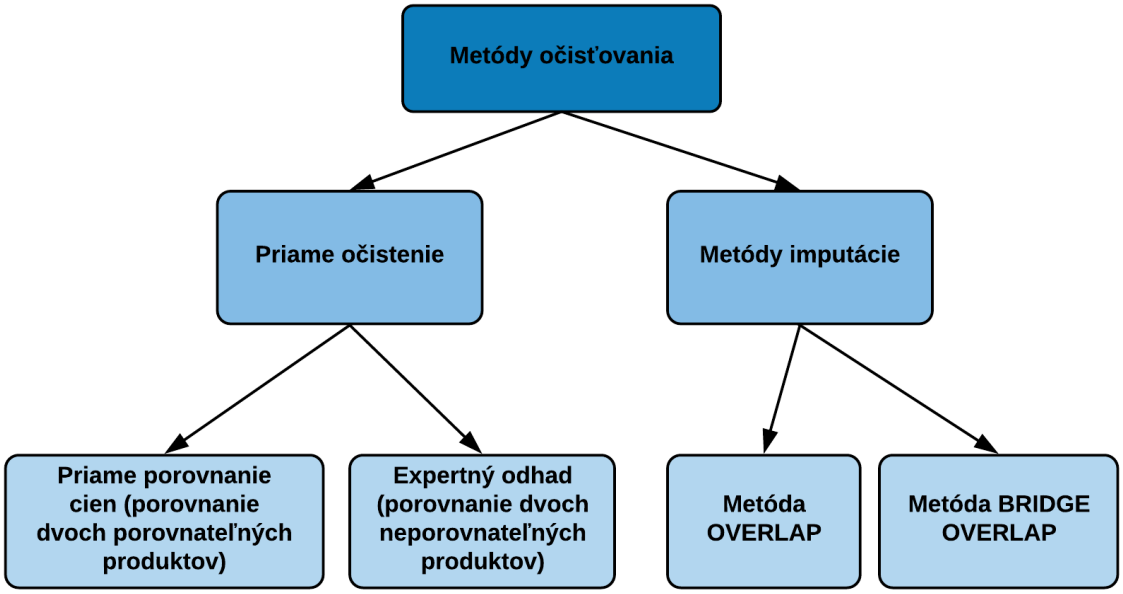
Na Obrázku 1.2 ilustrujeme rozdelenie metód, ktoré sa používajú na kvalitatívne očistenie. Jednotlivé metódy predstavíme v nasledujúcej sekcii.

1. **Priame očistenie**

   (a) Priame porovnanie cien (porovnanie dvoch porovnateľných produktov)

   Nová položka sa považuje za porovnateľnú, ak:

   - je vyrobená rovnakým výrobcom,
   - je vyrobená z rovnakých materiálov a má rovnaké alebo podobné technické parametre,
   - má rovnakú mernú jednotku,
   - má rovnaký typ balenia,



Obr. 1.2: Metódy kvalitatívneho očistenia.

- rozdiely medzi pôvodnou a novou položkou sú nepodstatné a môžu sa týkať vkusu a subjektívnych preferencií zákazníkov, napríklad: farba, vzor látky, tvar, atď.

Pokiaľ sú výrobky považované za porovnateľné z hľadiska kvality, potom sa rozdiel v ich cene premietne ako zdraženie alebo zľava v cenovom indexe.

(b) Expertný odhad (porovnanie dvoch neporovnateľných produktov)

Vo väčšine prípadov náhrad má pôvodný a nový výrobok rozdielnu kvalitu a ich ceny nemôžu byť pri zámene priamo porovnané. V tomto prípade je nutné rozdiel v ich cene očistiť o zmenu kvality pôvodného a nového produktu. To znamená vyčísliť, aká časť z rozdielu cien oboch produktov je spôsobená zmenou kvality a aká časť predstavuje zdraženie, respektíve zľavu.

Informácie pre toto rozhodnutie sa dajú získať:

- odborným odhadom pracovníka v teréne v spolupráci s personálom predajne,
- odborným odhadom externých expertov,
- použitím hedonických regresných modelov.

ČSÚ používa služby externých odborníkov najmä pri zámene reprezentan-

tov v spotrebnom koši. Poslednú možnosť ČSÚ zatiaľ nepoužíva, pretože hedonické modely vyžadujú objemnú databázu so širokým rozsahom charakteristík výrobkov, čo vyžaduje značné náklady nielen na vytvorenie takej databázy, ale aj na jej údržbu.

2. **Metódy imputácie (porovnanie dvoch neporovnateľných produktov)**:

   (a) Metóda "OVERLAP" (prekrývanie)

   Metóda prekrývania sa používa, keď je možné sledovať cenu pôvodného a nového výrobku v rovnakom období (mesiaci), ktoré je označené ako prekrývané. V prekrývanom mesiaci sa zaznamenajú ceny oboch výrobkov. Cenový vývoj je v tomto mesiaci meraný na pôvodnom výrobku, ďalší mesiac už na novom výrobku a s využitím dát z prekrývaného mesiaca. Rozdiel v cene pôvodného a nového výrobku sa v cene neprejaví. Túto metódu používa ČSÚ pri zámene reprezentantov v spotrebnom koši.

   (b) Metóda "BRIDGE OVERLAP" (mostové prekrývanie)

   Rozdiel oproti "OVERLAP" metóde je ten, že v tomto prípade nie sú k dispozícii dáta o oboch tovaroch v jednom prekrývanom mesiaci. Novému tovaru sa priradí cenový vývoj podobných produktov, ktoré sa vyskytovali v sledovanom aj v predchádzajúcom mesiaci.

Príklad kvalitatívneho očistenia

Pri kvalitatívnom očistení cenových indexov predstavuje najväčší problém rozhodnutie, do akej miery je nový výrobok porovnateľný s pôvodným, teda kedy použiť priame porovnanie cien, kedy a akou metódou ceny kvalitatívne očistiť. Nižšie uvádzame príklady možných postupov:

1. **Zámena cenového reprezentanta podľa terénneho pracovníka**

   Hlavným arbitrom pri posúdení porovnateľnosti pri zmene druhu je pracovník v teréne, ktorý najviac pozná oba druhy výrobkov a môže získať informácie od personálu predajne. V prípade, že považuje výrobok za porovnateľný, označí jeho cenu kódom "V" (podobný výrobok). V opačnom prípade doplní kód "3" (iný

druh). Následne sa záznamy o cene posunú na krajské pracovisko, kde sa informácie môžu overiť telefonicky. Poslednú supervíziu zmien a metódu očistenia "BRIDGE OVERLAP" aplikujú pracovníci oddelenia štatistiky spotrebiteľských cien.

2. **Zámena cenového reprezentanta kvôli aktualizácii spotrebného koša**

   Ako už bolo spomenuté v 1.3.2, každoročne dochádza k revízii spotrebného koša a teda aj k aktualizácii cenových reprezentantov. U nových výrobkov je veľmi obtiažne (prakticky nemožné) zistiť cenu v základnom období, pretože často výrobok na trhu ešte neexistoval. V takom prípade nový reprezentant naviaže na doterajší vývoj pôvodného reprezentanta s využitím metód "OVERLAP" a "BRIDGE OVERLAP". V Tabuľke 1.2 uvádzame príklad kvalitatívneho očistenia metódou "BRIDGE OVERLAP", ktorú sme použili v našej aplikácii pri náhrade reprezentanta "Bravčová pečeň s kosťou" reprezentantom "Bravčová pečeň", ku ktorej došlo na prelome rokov 2018 a 2019.

	Cena 12/2018	Cena 01/2019 pôvodná	CPI 01/2019 v kategórii Potravín	Prevod BRIDGE OVERLAP	Cena 01/2019 očistená
Bravčová pečeň s kosťou	115,9		100,9		
Bravčová pečeň		118,3	100,9	$\frac{118,3}{118,3/115,9}$ $\frac{100,9}{100}$	116,9

Tabuľka 1.2: Očistenie metódou "BRIDGE OVERLAP".

3. **Kvantitatívne očistenie**

   Kvantitatívne očistenie sa týka položiek, ktorých cena v období t bola zistená v inej mernej jednotke ako v období $t-1$. Očistenie sa aplikuje lineárnym prepočtom na rovnakú mernú jednotku. Ak sa po tomto prepočte zmení cena, ďalší postup závisí na tom, či:

   - pôvodné balenie na trhu prestalo existovať (napríklad zníženie váhy čokolády zo 125 gramov na 100 gramov),
   - pôvodné balenie nie je k dispozícii len v danom období t (napríklad prací prášok o váhe 1 kg je dočasne nahradený 3 kg balením).

V prvom prípade je zmena ceny zahrnutá do indexu ako zdraženie alebo zľava. V druhom prípade sa položky považujú za neporovnateľné a postup kvalitatívneho očisťovania je rovnaký ako pri inom druhu. To znamená rovnaký postup ako bol spomenutý vyššie pri náhrade produktu terénnym pracovníkom - pracovník označí produkt kódom "3" (iný druh alebo zmena gramáže) a tento cenový rozdiel sa do indexu nezapočíta.

1.3.7 Harmonizovaný index spotrebiteľských cien - HICP

Harmonizovaný index spotrebiteľských cien (HICP - Harmonized Index of Consumer Prices) vznikol kvôli zabezpečeniu možnosti porovnania indexov spotrebiteľských cien v rámci krajín Európskej únie. Motivácia bola zmeranie trendov inflácie členských štátov Európskej únie ako kritérium vstupu do menovej únie. Veľa technických aspektov výpočtu HICP sa používa aj v národných indexoch. Rozdiely sú napríklad zahrnutie tržieb za nákupy cudzincov na území Českej republiky do HICP alebo naopak, v HICP nie je zahrnuté imputované nájomné [4].

HICP sa ďalej používa pre výpočet MUICP (Index spotrebiteľských cien menovej únie), ktorý meria infáciu štátov, ktoré sú v menovej únii. Takisto sa používa pre výpočet EICP (Európsky index spotrebiteľských cien), ktorý meria infláciu 28 štátov Európskej únie.

Jeden z hlavných cieľov Európskej centrálnej banky (ECB) je cenová stabilita, ktorá sa definuje ako ročné tempo rastu HICP 2% a menej. HICP a MUICP štatistiky sú preto kritické pre ECB v zmysle určovania monetárnej politiky v eurozóne [11].

1.3.8 Jadrová inflácia

Dôležitou inflačnou štatistikou je jadrová inflácia. V Českej republike sa počíta na celom spotrebnom koši po vylúčení vplyvu zmien regulovaných cien, daňových úprav a iných administratívnych opatrení. Takáto inflácia očistená o sezónne vplyvy je publikovaná ako jadrová inflácia po sezónnom očistení [4]. Vyjadruje teda akúsi trhovú infláciu, kde sa neberú do úvahy zásahy štátu.

Kapitola 2

Dáta

2.1 Ceny položiek v spotrebnom koši z merania ČSÚ

Z ČSÚ sme si vyžiadali časový rad mesačných cien reprezentantov, ktorý vstupujú do tvorby oficiálneho CPI. ČSÚ nám poskytli tieto dáta v období január 2012 až marec 2019 (viď Tabuľka 2.1 pre príklad štruktúry). Vzhľadom na to, že v januári 2017 sa výraznejšie zmenili reprezentanti a váhový systém, pracovali sme kvôli dátovej konzistencii len s časovým úsekom január 2017 až marec 2019.

ECOICOP	Názov	Merná jednotka	Váha	Cena 02/2019	Cena 03/2019
E01	POTRAVINY A NEALKOHOLICKÉ NÁPOJE		177,6		
E01.1	Potraviny		160,3		
E01.11	Pekárenské výrobky; obiloviny		28,3		
E01.111	RÝŽE		0,8		
E01.111.01	RÝŽE LOUPANÁ DLOUHOZRNNÁ	1 kg	0,8	35,7	35,2
E01.112	MOUKA A JINÉ OBILOVINY		1,2		
E01.112.01	PŠENIČNÁ MOUKA HLADKÁ	1 kg	0,6	12,2	12,4
E01.112.02	PŠENIČNÁ MOUKA HRUBÁ	1 kg	0,6	12,0	11,5

Tabuľka 2.1: Príklad niektorých položiek spotrebného koša definovaného ČSÚ.

Z dát poskytnutých ČSÚ sme vypočítali súčty všetkých položiek prenásobené príslušnými váhami spotrebného koša v kategóriách "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak" za každý mesiac od januára 2017. Problém bol len s dvomi položkami zo spotrebného koša. Položka *Bravčová pečeň s kosťou* sa od januára

nahradila položkou *Bravčová pečeň* a v tomto prípade sme použili kvalitatívne očistenie metódou BRIDGE OVERLAP, ktorú sme opísali v 1.3.6. Druhá problémová položka bola položka *Kapor chladený, mrazený*, ktorá sa od januára 2018 nahradila položkou *Kapor chladený*. V tomto prípade sa nám od ČSÚ nepodarilo získať dáta pre položku *Kapor chladený, mrazený* z roku 2017 a teda nebolo možné aplikovať kvalitatívne očistenie. Túto položku sme preto do súčtu nezahrnuli. Vypustenie jednej položky malo na súčty minimálny vplyv, keďže sme súčty tvorili zo 157 položiek.

2.2 Aplikácia na sťahovanie dát z online obchodov

Pre potreby našej analýzy sme potrebovali pravidelne zhromažďovať dáta sledujúce cenový vývoj v kategóriách "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak" na online portáloch. Pre tento cieľ sme vyvinuli skript v softvéri RStudio, ktorý po spustení stiahne všetky položky v ponuke troch najväčších online nákupných portálov v Českej republike, a to *rohlik.cz*, *kosik.cz* a *itesco.cz*[1].

Skript bol automaticky spúšťaný raz do týždňa v pondelok, aby sme mali dostatočné množstvo dát pre analytickú prácu. V Tabuľke 2.2 pre ilustráciu uvádzame príklad štruktúry stiahnutých dát.

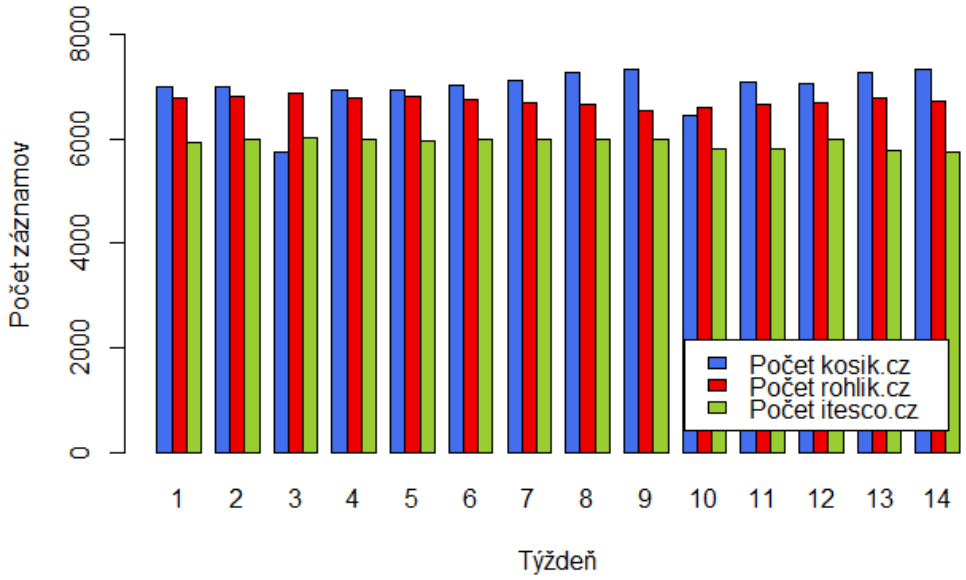
Dátum	Kód	Názov	Cena v Kč	Zľava v %	Dostupnosť	Cena za jednotku
22.4.2019	73018	Figaro Čokoláda na vaření hořká 100g	55.8	19		(132,06 Kč / kg)
22.4.2019	29783	Bohemia Kotlíkové brambůrky Mořská sůl a rozmarýn 120g	49.9	0		(415,84 Kč / kg)
22.4.2019	71742	Rio Mare Insalatissime Tuňákový salát Texana 160g	76.9	0		(480,63 Kč / kg)
22.4.2019	71136	Schär Snackers - bezlepkové krekry 115g	59.9	0		(520,87 Kč / kg)
22.4.2019	25552	Diana Arašídy ve wasabi těstíčku 100g	NA	0	Vyprodáno.	NA

Tabuľka 2.2: Príklad štruktúry dát z e-shopu *kosik.cz*.

Na Obrázku 2.1 ilustrujeme počet pozorovaní, ktoré sme týždenne stiahli v ča-

[1]Na stiahnutie dát priamo zo stránok sme použili knižnice `rvest` a `RSelenium`. `Rvest` umožňuje stiahnutie html kódu zo stránky a následnú prácu s ním, zatiaľ čo pomocou `RSelenium` softvéru RStudio otvorí automatizovaný prehliadač, ktorý je možné navigovať príkazmi cez RStudio. Program prejde iteratívne cez všetky zadané URL a ukladá zozbierané dáta do CSV súborov.

sovom intervale od 20. januára 2019 do 22. apríla 2019. Na obrázku uvádzame počty po odstránení duplicít (v každom obchode sa vyskytovali duplicity, nakoľko jeden produkt sa môže vyskytovať vo viacerých kategóriách - napríklad Lieskovce v čokoláde sa môžu vyskytovať v kategórii "Trvanlivé" alebo "Ovocie"). Online obchod *kosik.cz* mal najširší sortiment tovarov až na dve výnimky.



Obr. 2.1: Počty stiahnutých cenových záznamov v jednotlivých týždňoch od 20. januára 2019 do 22. apríla 2019

2.3 Dáta z Českej národnej banky

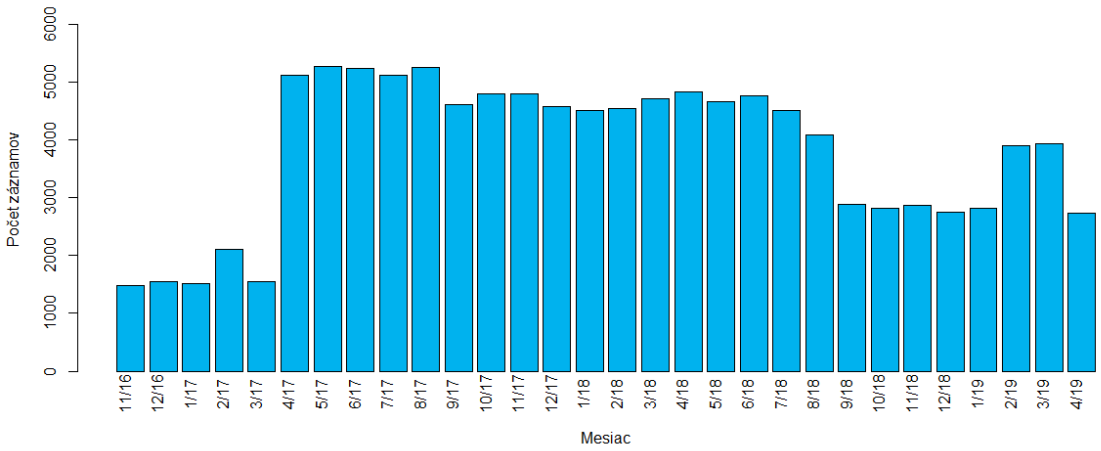
Pre naše modely však 14 týždňov pozorovaní nestačilo. Nakoľko v Českej národnej banke (ČNB) experimentujú so sťahovaním dát online kvôli zlepšeniu inflačných metrík, na účely našej diplomovej práce nám poskytli ich historické merania. Jednalo sa konkrétne o denné stiahnutia položiek z e-shopu *itesco.cz*. Príklad štruktúry týchto dát uvádzame v Tabuľke 2.3.

V ČNB začali dáta sťahovať 18. novembra 2016 a nám poskytli všetky záznamy ku dňu 30. apríl 2019. Jednalo sa spolu o 457 CSV súborov, periodicita sa počas nám poskytnutého časového intervalu líšila - v niektorých týždňoch bolo zaznamenaných 5

Dátum	Akcia	Ceny	Ceny za jednotku	Názvy	Typ	Jednotka
16.01.2017		5.11	29.9	Banány	01.162	Kč/kg
16.01.2017		49.9	49.9	Ananas	01.162	Kč/Kus
16.01.2017	-22% bežná cena 21,90 nyní 16,90	16.9	16.9	Avokádo voľné ks	01.162	Kč/Kus
16.01.2017		79.9	443.89	DATLE 180G	01.162	Kč/kg
16.01.2017		16.9	16.9	Granátové jablko	01.162	Kč/Kus

Tabuľka 2.3: Príklad štruktúry dát z e-shopu *itesco.cz* stiahnutých v ČNB.

stiahnutí, v niektorých iba 1 a v poslednom týždni každého roku neboli k dispozícii žiadne merania. Na Obrázku 2.2 sme znázornili, koľko unikátnych produktov sme mali k dispozícii každý mesiac - súbory sme spájali plným spájaním (v angl. full join), teda ak sa v mesiaci produkt stiahol minimálne jedenkrát, je zahrnutý v počte pre daný mesiac.



Obr. 2.2: Počty unikátnych stiahnutých položiek v jednotlivých mesiacoch v ČNB.

Pre potreby našich modelov sme potrebovali vytvoriť dataset, ktorý by obsahoval názvy položiek a ich cenu v jednotlivých mesiacoch. Zvolili sme si teda základný mesiac marec 2019, a spätne sme na položky, ktoré boli v marci 2019 dostupné, mapovali položky z predchádzajúcich mesiacov až po obdobie január 2017[2]. Toto mapovanie sme robili na základe názvu a štýlom ľavého spájania (tzv. left join) - teda zobrali sme všetky položky v marci 2019 a pokiaľ sa v predchádzajúcich mesiacoch položka nachá-

[2]Obdobia november 2016 a december 2016 sme vypustili, nakoľko od roku 2017 prišlo ku výraznejšej obmene oficiálneho spotrebného koša a revízii váh.

dzala, namapovali sme ju. Pokiaľ nie, hodnota na jej mieste ostala prázdna. Takto sme získali prvotný dataset, ktorý obsahoval 3939 pozorovaní. Nakoľko ale až 40% hodnôt chýbalo, rozhodli sme sa pozorovania, ktoré mali viac ako polovicu záznamov prázdnych, odstrániť. Dospeli sme ku 2490 pozorovaniam, kde chýbalo už len 18% hodnôt. Tieto hodnoty sme doplnili tak, že sme im priradili hodnotu ceny z nadchádzajúceho mesiaca. Analogickým postupom sme si pripravili dáta aj s týždňovou granularitou, kde sme ako základný týždeň zobrali posledný marcový týždeň. Takto sme sa dostali k 2663 pozorovaniam a po vymazaní pozorovaní z nadpolovičnou väčšinou chýbajúcich záznamov sme sa dopracovali k číslu 1694.

Keďže sme nemali v originálnom datasete zahrnutú kategorizáciu položiek, stiahli sme túto informáciu z *itesco.cz* pre dostupné produkty. Niektoré položky z nášho datasetu sa na stránkach už nenachádzali a tie sme dourčili pomocou metódy oporného bodu (viď 3.2). Na takto rozkategorizovanom súbore sme spravili súhrnné štatistiky. Výsledky uvádzame v Tabuľke 2.4. Najväčšiu cenovú fluktuáciu vykazujú kategórie "Vína" a "Liehovín". V týchto kategóriách sa to dalo očakávať, nakoľko majú zvyčajne najvyššiu maržu. Na druhej strane, najstabilnejšie ceny sa javia v kategórii "Mäso".

Názov kategórie	**Priemerná cena**	**Štandardná odchýlka**	**Variancia**	**Počet**
Cukr, marmeláda, med, čokoláda a cukrovinky	47,1	0,93	0,87	478
Káva, čaj a kakao	71,1	1,22	1,49	76
Lihoviny	202,0	10,86	117,93	22
Maso	24,6	0,41	0,16	87
Minerální vody, nealkoholické nápoje, ovocné a zeleninové šťávy	42,4	0,77	0,59	139
Mléko, sýry a vejce	27,7	0,47	0,22	365
Oleje a tuky	50,6	1,87	3,48	26
Ovoce	38,8	0,6	0,36	120
Pekárenské výrobky, obiloviny	32,5	0,44	0,19	411
Piva	21,9	0,85	0,73	40
Potravinářské výrobky jinde neuvedené	18,6	0,47	0,22	157
Ryby a mořské plody	56,2	2,14	4,56	68
Tabák	91,0	1,43	2,04	30
Vína	135,6	6,88	47,34	130
Zelenina	31,4	0,66	0,43	341

Tabuľka 2.4: Popisné charakteristiky po kategóriách na dátach z ČNB.

Kapitola 3

Kategorizácia tovarov

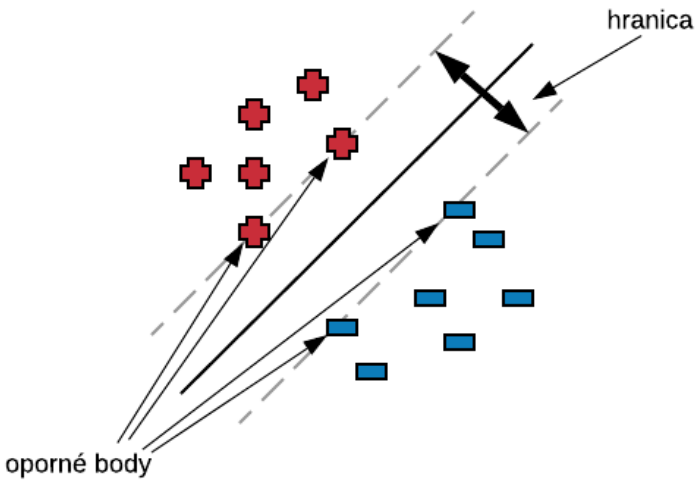
Ako sme uviedli v podkapitole 1.3.2, štruktúra CPI zodpovedá európskej klasifikácii ECOICOP, ktorá závadza podrobné členenie výrobkov do kategórií, skupín a tried. Pre potreby zostrojenia oficiálnych inflačných metrík je preto potrebná klasifikácia produktov do jednotlivých kategórií.

Pri scrapovaní dát z webstránok e-shopov sa zväčša dá dopracovať k určitej klasifikácii, ktorú však určuje predajca a teda takéto členenie väčšinou nespĺňa štruktúru ECOICOP. Takisto ale môže nastať prípad, keď z rôznych dôvodov nevieme kategóriu zistiť (predajca takúto informáciu neponúka alebo ju neponúka pre všetky sťahované položky). Tiež môže nastať situácia, že kategória položky nie je konzistentná s definíciou produktu. Napríklad, Ovocná tyčinka môže byť zaradená v kategórii "Ovocie a zelenina" alebo v kategórii, ktorú chce predajca promovať, napríklad "Vianočné sladkosti".

Na základe vyššie uvedených príkladov teda pri tvorbe cenových indexov nie je vhodné spoliehať sa na kategorizáciu predajcov. Môžeme ale predajcovu klasifikačnú štruktúru využiť ako počiatočný bod našej klasifikačnej metódy pre korektné zatriedenie do správnej ECOICOP kategórie. Pre dané použitie sa ako najlepšie klasifikačné algoritmy ukazujú Metódy oporného bodu (z angl. Support Vector Machines) [1].

3.1 Metóda oporného bodu

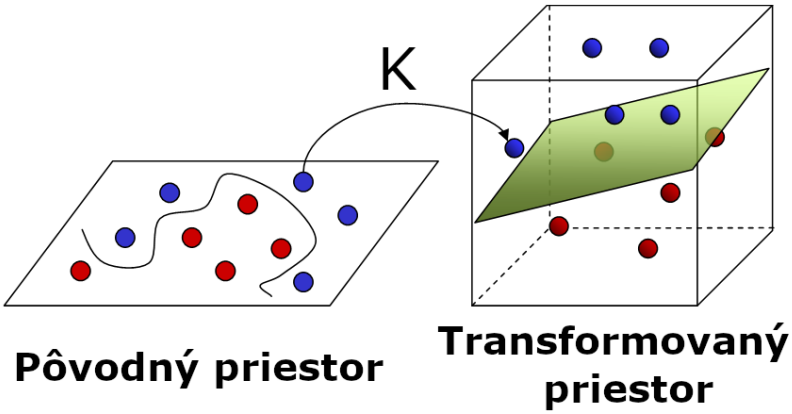
Metódy oporného bodu sú známym klasifikačným nástrojom. Originálna metóda rieši binárny klasifikačný problém. V priestore vstupných premenných sa snaží preložiť podpriestor tak, aby čo najlepšie (v zmysle najširšej hranice) oddelil dve cieľové skupiny. Na Obrázku 3.1 prikladáme ilustračný príklad v dvojrozmernom priestore. Deliaci podpriestor je v tomto prípade priamka.



Obr. 3.1: Náčrt metódy oporného bodu v dvojrozmernom priestore.

Často však nie je možné určiť hranicu priamo. V takýchto prípadoch sa používa "Trik s jadrom"(z angl. "Kernel trick"). Trik s jadrom transformuje dáta do viacrozmerného priestoru, kde sa klasifikácia spraví jednoduchšie (viď Obrázok 3.2). V prípade, že dokážeme dáta klasifikovať bez triku s jadrom, dáta voláme lineárne separovateľné a v opačnom prípade lineárne neseparovateľné dáta.

Ako sme spomenuli vyššie, originálna metóda oporného bodu rieši binárny problém (cieľová premenná nadobúda hodnoty len dvoch kategórií - pozitívne a negatívne). V prípade viacerých kategórií cieľovej premennej sa používajú prístupy ako napríklad jeden proti všetkým (z angl. one-against-all) alebo jeden proti jednému (z angl. one-against-one). V prípade jeden proti všetkým sa model trénuje pre každú kategóriu osobitne tak, že prvky kategórie sú pozitívne hodnoty a všetky ostatné, ktoré nepatria



Obr. 3.2: Trik s jadrom. Zdroj: [19].

do danej kategórie, sú negatívne. Takto sa natrénuje k (k je počet rôznych hodnôt klasifikátora) modelov a klasifikácia je určená vyšším skóre (vyššia hodnota účelovej funkcie) jednotlivých modelov pre individuálne pozorovania. V prípade jeden proti jednému sa trénuje model vždy len s dvomi hodnotami cieľovej premennej a teda jedna je pozitívna a druhá negatívna. Natrénuje sa $k(k-1)/2$ modelov. Následne sú pri klasifikácii pre každé pozorovanie použité všetky modely a predikovaný klasifikátor je ten, ktorý je zvolený vo väčšine modelov - dostane najviac hlasov.

3.1.1 Definícia úlohy

Majme dátový súbor, kde $(X_i, y_i), i = 1, ..., n, X_i \in R^m$ sú páry vektora charakteristík X_i a cieľovej premennej y_i, ktorá nadobúda hodnoty $+1$ alebo -1. Ako sme uviedli v úvode sekcie, snažíme sa určiť najširšiu hranicu medzi bodmi, aby ich čo najlepšie klasifikovala [8]. Uvažujme teda nasledovné rovnice pre podpriestory pozitívnych a negatívnych záznamov:

$$\begin{aligned} W^T X_+ + b &= +1, \\ W^T X_- + b &= -1, \end{aligned} \tag{3.1}$$

kde W sú koeficienty X a b zodpovedá posunutiu podpriestoru. Po odpočítaní rovníc v 3.1 a následnom predelení dlžkou vektora W, ktorá je definovaná ako $= \sqrt{\sum_{j=1}^{m} W_j^2}$ dostaneme výraz:

$$\frac{W^T(X_+ - X_-)}{\|W\|} = \frac{2}{\|W\|}. \tag{3.2}$$

Výraz v 3.2 na ľavej strane je vzdialenosť medzi pozitívnym a negatívnym podpriestorom. Maximalizačná úloha sa potom dá sformulovať nasledovne:

$$\begin{aligned} &\max_{W,b} \frac{2}{\|W\|}, \\ &\text{s.t. } y_i(W^T X_i + b) \geq 1, \end{aligned} \tag{3.3}$$

kde podmienka $y_i(W^T X_i + b) \geqslant 1$ zahŕňa predpoklad, že dáta sú dokonale lineárne separovateľné (viď Obrázok 3.1). Hranica, ktorú sa snažíme maximalizovať, sa zvykne nazývať aj "tvrdá" (z angl. hard margin). Predpoklad lineárnej separovateľnosti ale v praxi takmer nikdy nie je splnený. Pre tento prípad bola vyvinutá úloha tzv. "mäkkej" hranice (z angl. soft margin), ktorá je obmenou úlohy 3.3:

$$\begin{aligned} &\min_{W,b} \frac{\|W\|^2}{2} + C * \frac{1}{n} * \sum_{i=1}^{n} \xi_i, \\ &\text{s.t. } y_i(W^T X_i + b) \geq 1 - \xi_i, \\ &\xi_i \geq 0, \forall i, \end{aligned} \tag{3.4}$$

kde konštanta C je parameter, ktorý reguluje pretrénovanie modelu a ξ_i sú tzv. "slack variables", ktoré vyjadrujú chybu zle zaklasifikovaných bodov.

Chyba zle zaklasifikovaných bodov ξ_i sa dá interpretovať ako stratová funkcia a najčastejšie sa používajú dva prípady - L1 stratová funkcia $\xi_i = Max(0, 1 - y_i(W^T X_i + b))$ alebo L2 stratová funkcia $\xi_i = Max(0, 1 - y_i(W^T X_i + b))^2$.

Zadefinovanú úlohu vo vzorci 3.4 dokážeme previesť na duálnu úlohu [8].

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j X_i^T X_j = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, X_j) \\ &\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, \\ &0 \leq \alpha \leq C 1_n. \end{aligned} \tag{3.5}$$

Výhoda takéhoto zápisu je, že sa tu dá použiť trik s jadrom. Takisto si treba uvedomiť, že α_i bude nenulová iba pre oporné vektory [8]. Jadro $K(X_i, X_j)$ sa používa pre transformovanie vektora vstupných premenných do viacrozmerného priestoru (viď Obrázok 3.2).

Príklady jadra:

1. Lineárne jadro - $K(X, Y) = X^T Y + c$, kde c je konštanta
2. Polynomiálne jadro - $K(X, Y) = (X^T Y + c)^d$, kde c,d sú konštanty
3. Gausovské jadro - $K(X, Y) = exp(-\frac{\|X-Y\|^2}{2\delta})$, kde δ je konštanta

3.2 Kategorizácia tovarov do ECOICOP kategórií

Pri scrapovaní stránok nastala pri niektorých položkách situácia, že sa nám nepodarilo stiahnuť kategóriu položky. Dôvodom mohla byť chyba v kóde na stránke predajcu alebo nedostupnosť takejto informácie na sťahovanej stránke. Tieto skutočnosti nás motivovali vyvinúť proces, ktorý by takúto kategorizáciu vykonal v prípade nedostupnosti informácie u predajcu.

Metódy oporného bodu sa ukazujú ako veľmi efektívna metóda pri kategorizácii textu [12]. Cieľom kategorizácie textu je klasifikácia vstupného vektora obsahujúceho textové reťazce do pevnej množiny predefinovaných kategórií. Každá zložka vektora je teda začlenená do jednej kategórie (napr. Banány do kategórie "Ovocie").

Dáta od ČNB, s ktorými sme pracovali, sú rovnaké, aké sme opísali v kapitole 2.3. Nakoľko nedisponovali premennou, ktorá by určovala kategóriu produktov, dodatočne sme kategórie stiahli a napárovali. Z celkového počtu 1694 sa nám poradilo priradiť kategóriu 1614 produktom. Nezaradených 80 produktov sme zakategorizovali na konci za pomoci natrénovaného modelu. Aby sme vedeli overiť úspešnosť klasifikácie, rozdelili sme dáta na trénovaciu a testovaciu množinu. Model sme trénovali na trénovacej a presnosť kategorizácie overovali na testovacej množine.

Prvým krokom pri textovej kategorizácii je transformácia vstupného vektora s textovými reťazcami do vhodného zobrazenia pre klasifikačnú úlohu. Podľa [12] sa ako vhodná metóda ukazuje rozloženie vstupného textu (v našom prípade názvu položky)

na jednotlivé slová s tým, že každé slovo vyskytujúce sa vo vektore s textovými reťazcami reprezentuje jednu premennú. Hodnota premennej je potom počet výskytov slova v danom reťazci[1].

Aby sme sa vyhli zbytočne vysokému počtu premenných a vysokodimenzionálnej matici, nevytvárali sme novú premennú pre slová obsahujúce číslovky (napr. 100%, 100g) a takisto pre interpunkčné znamienka[2].

Na upravenom datasete sme potom natrénovali model L2 regularizovanej (stratová funkcia je $\xi_i = Max(0, 1 - y_i(W^T X_i + b))^2$) metódy oporného bodu s lineárnym jadrom. Po použití modelu na testovaciu sadu sme správne klasifikovali 368 z 404 záznamov, čo predstavuje 91,09 % úspešnosť. Príklad kategorizácie uvádzame v Tabuľke 3.1. Tieto kategórie sme následne previedli do kategorizácie ECOICOP podľa Tabuľky B2, ktorú prikladáme v Prílohe B.

Názov položky	Predpovedaná kategória	Skutočná kategória
3Bit Sušenka s mléčnou čokoládou a mléčnou náplní 46g	Čokoládové tyčinky	Čokoládové tyčinky
Adriana Linguine těstoviny semolinové sušené 500g	Těstoviny	Těstoviny
AGRO-LA Jogurt bílý 200g	Bílé jogurty	Bílé jogurty
Alibona Třešně celé 360g	Konzervované a sterilizované ovoce	Konzervované a sterilizované ovoce
Alka Dupetky solené 80g	Krekry	Krekry
Almar Losos uzený premium 100g	Rybí speciality	Rybí speciality
Anglická slanina	Chipsy	Pultové uzeniny
Apetito S hermelínem 150g	Tavené a termizované sýry	Tavené a termizované sýry
Apetito Se sýrem s modrou plísní 150g	Plísňové a zrající sýry	Tavené a termizované sýry
Arašídy neloupané pražené	Ořechy	Ořechy

Tabuľka 3.1: Príklad kategorizácie vybratých produktov.

Nakoľko sme boli spokojní s presnosťou modelu, použili sme ho aj na klasifikáciu produktov, pri ktorých sa nám nepodarlo zistiť kategóriu. V tomto prípade však

[1]Na tento účel sme použili R knižnicu `tm`, ktorá umožňuje požadovanú transformáciu. Základnou štruktúrou pre prácu s dátami v rámci knižnice je tzv. "Corpus", ktorý reprezentuje kolekciu vstupných dokumentov. V našom prípade to znamená, že "Corpus" obsahuje zoznam hodnôt vstupného reťazca a počet slov v každom reťazci.

[2]V rámci knižnice `tm` slúži na tento účel funkcia `DocumentTermMatrix`, ktorej sa ako vstupné parametre môžu dať ďalšie kritéria zostrojenia premenných. V našom postupe sme vylúčili slová obsahujúce číslovky a interpunkčné znamienka. Takisto sú automaticky vylúčené reťazce, ktoré sú kratšie ako 3 znaky. V textovom miningu sa ešte často vylučujú tzv. "stop-words", ako napríklad "alebo","a" alebo "aj".

nevieme overiť presnosť, nakoľko nemáme informáciu o skutočnej kategórii.

Kapitola 4

Regularizácia

Ako sme uviedli v podkapitole 1.3.4, manuálny výber a zisťovanie cien jednotlivých tovarov je časovo náročný proces - len v kategórii "Potraviny a nealkoholické nápoje" je potrebné identifikovať dokopy 158 položiek. Tento prístup by sa však dal do značnej miery automatizovať s použitím techniky regularizácie (z angl. regularization). CPI pre jednotlivé kategórie sa vypočíta ako pomer dvoch súčtov, ktoré obsahujú ceny všetkých položiek danej kategórie prenásobené objemovými váhami určenými štatistikou rodinných účtov. K týmto súčtom sa vieme dopracovať na základe dát od ČSÚ, ktorý nám poskytol ceny jednotlivých reprezentantov na mesačnej báze od januára 2017. Tieto súčty teda budú predstavovať cieľovú premennú. Naším cieľom je nájsť optimálnu lineárnu kombináciu tovarov z e-shopov tak, aby čo najlepšie replikovala vývin súčtov vstupujúcich do CPI a bola schopná ich predikovať do budúcnosti. Naša úloha spočíva v dvoch krokoch - nájsť dané produkty (prediktory) a ich váhy (koeficienty). V nasledujúcej kapitole si priblížime metódy, ktoré umožňujú riešenie úlohy tohoto typu.

4.1 Vzťah medzi vychýlením (bias) a varianciou v lineárnej regresii

Predpokladajme jednoduchý model lineárnej regresie, kde cieľom je vysvetliť n pozorovaní cieľovej premennej y pomocou lineárnej kombinácie m prediktorov - X a normálne rozdelených chýb ϵ s varianciou σ^2:

$$\begin{aligned} y &= X\beta + \epsilon, \\ \epsilon &\sim N(0, \sigma^2). \end{aligned} \tag{4.1}$$

Keďže nepoznáme skutočné hodnoty parametra β, musíme ich odhadnúť pomocou metódy najmenších štvorcov (OLS - Ordinary Least Squares z angl.). Odhad $\hat{\beta}$ teda dostaneme riešením minimalizačnej úlohy stratovej funkcie $L_{OLS}(\beta)$:

$$\min_{\beta} L_{OLS}(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \min_{\beta} \|y - X\beta\|^2. \tag{4.2}$$

Riešením tejto minimalizácie dostaneme známy odhad $\hat{\beta}_{OLS} = (X^T X)^{-1}(X^T y)$.

Pri vyhodnocovaní modelu sú dva dôležité prvky - vychýlenie (z angl. bias) a variancia. Vychýlenie definujeme ako rozdiel medzi skutočnou hodnotou a strednou hodnotou odhadnutého parametra:

$$Bias(\hat{\beta}_{OLS}) = E(\hat{\beta}_{OLS}) - \beta. \tag{4.3}$$

Vychýlenie určuje presnosť odhadu. OLS odhad je vzhľadom na svoju konštrukciu nevychýlený odhad [17] (viz. 4.4 pre dôkaz).

$$\begin{aligned} E(\hat{\beta}_{OLS}) &= \underbrace{E[(X^T X)^{-1} X^T y]}_{y = X\beta + \epsilon} = E[\underbrace{(X^T X)^{-1} X^T X}_{I} \beta + (X^T X)^{-1} X^T \epsilon] = \\ &= E[\beta + (X^T X)^{-1} X^T \epsilon] = \beta + (X^T X)^{-1} X^T \underbrace{E(\epsilon)}_{E(\epsilon) = 0} = \beta, \end{aligned} \tag{4.4}$$

Na druhej strane variancia meria rozptyl alebo neistotu odhadov. Odvodiť sa dá do nasledovného tvaru [17]:

$$\begin{aligned} Var(\hat{\beta}_{OLS}) = \underbrace{Var[\beta + (X^T X)^{-1} X^T \epsilon]}_{Var(A\epsilon) = AVar(\epsilon)A^T,\, Var(\beta) = 0} &= (X^T X)^{-1} X^T Var(\epsilon)[(X^T X)^{-1} X^T]^T = \\ = \underbrace{\sigma^2 (X^T X)^{-1} X^T [(X^T X)^{-1} X^T]^T}_{((X^T X)^{-1})^T = (X^T X)^{-1}} &= \sigma^2 (X^T X)^{-1}, \end{aligned} \tag{4.5}$$

kde sme použili rovnaký rozklad ako pri odvodení vzorca 4.4, následne sme využili vlastnosť rozkladu variancie $Var(A\epsilon) = AVar(\epsilon)A^T$ a symetrickosť matice $(X^T X^{-1})$.

Neznámu varianciu chýb σ^2 je možné odhadnúť pomocou rezíduí z regresie nasledovne [17]:

$$\sigma^2 = E\Big(\frac{\hat{\epsilon}^T\hat{\epsilon}}{n-m}\Big), \qquad (4.6)$$
$$\text{kde } \hat{\epsilon} = y - X\hat{\beta}_{OLS}.$$

Dôkaz.

$$\hat{\epsilon}^T\hat{\epsilon} = \underbrace{(M_X y)^T M_X y}_{M_X = I_n - X(X^TX)^{-1}X^T} =$$

$$= \underbrace{y^T M_X^T M_X y}_{M_X^T M_X = M_X} = y^T M_X y = \underbrace{(X\beta+\epsilon)^T M_X (X\beta+\epsilon)}_{M_X X = 0} = \epsilon^T M_X \epsilon$$

$$E(\hat{\epsilon}^T\hat{\epsilon}) = E(\underbrace{\epsilon^T M_X \epsilon}_{\text{matica } 1x1}) = \underbrace{E[tr(\epsilon^T M_X \epsilon)]}_{tr(AB)=tr(BA),\ tr() \text{ je stopa matice}} = \underbrace{E[tr(M_X \epsilon\epsilon^T)]}_{E[tr(A)]=tr[E(A)]} =$$

$$= tr[E(M_X\epsilon\epsilon^T)] = \underbrace{tr[M_X E(\epsilon\epsilon^T)]}_{E(\epsilon\epsilon^T)=E((\epsilon-E(\epsilon))(\epsilon-E(\epsilon))^T)=Var(\epsilon)=\sigma^2 I_n} =$$

$$= tr(M_X\sigma^2 I_n) = \sigma^2 tr(M_X) = \underbrace{\sigma^2 tr[I_n - X(X^TX)^{-1}X^T]}_{tr(A+B)=tr(A)+tr(B)} = \sigma^2(n-m)$$

□

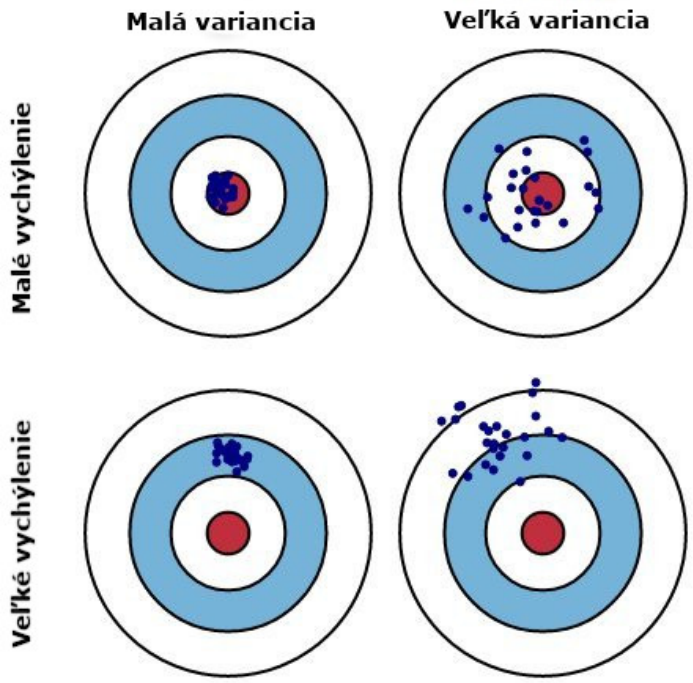
Matica M_X sa nazýva ortogonálny projektor X, t.j. zobrazuje maticu X do nuly.

Na Obrázku 4.1 ilustrujeme podstatu vychýlenia a variancie. Stred terča symbolizuje skutočný parameter β a zásahy sú hodnoty odhadov, ktoré vyplývajú zo štyroch rozličných modelov - model s malým vychýlením a malou varianciou, veľkým vychýlením a malou varianciou, malým vychýlením a veľkou varianciou a veľkým vychýlením a veľkou varianciou. Cieľom pre dobrý model je malé vychýlenie a nízka variancia.

Chyba modelu sa dá rozdeliť na tri časti: chyba z dôvodu veľkej variancie, chyba z dôvodu významného vychýlenia a zvyšok, t.j. náhodný rozptyl:

$$E(\epsilon) = [E(X\hat{\beta}) - X\beta]^2 + E[X\hat{\beta} - E(X\hat{\beta})]^2 + \sigma^2 = Bias^2 + Var + \sigma^2 \qquad (4.7)$$

Ako sme odvodili v 4.4 vyššie, OLS odhad je vzhľadom na svoju konštrukciu nevychýlený odhad. Môže mať ale veľkú varianciu, špecificky v nasledovných prípadoch:



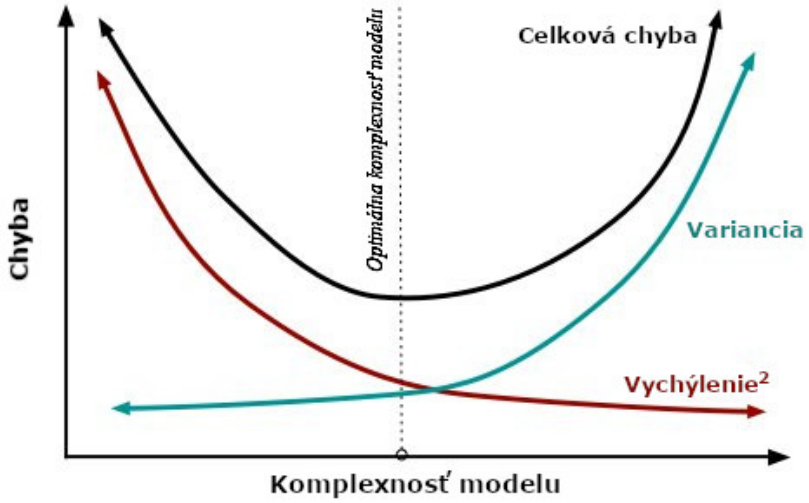
Obr. 4.1: Vzťahy medzi vychýlením a varianciou. Zdroj: [14].

- Premenné v matici X (prediktory) sú navzájom vysoko korelované (t.j. multikolinearita);
- Prediktorov je príliš veľa. Tento fakt vyplýva aj zo vzorca 4.6: ak sa m blíži k n, variancia sa limitne blíži k nekonečnu.

Všeobecné riešenie k týmto problémom je znížiť varianciu za cenu pridania menšieho vychýlenia. Tento prístup sa volá regularizácia a je takmer vždy užitočný pri prediktívnej analýze. Na Obrázku 4.2 je znázornené, aký vplyv má na chybu predikcie variancia spolu s vychýlením. Pri zvýšení komplexnosti modelu (čomu sa pri lineárnej regresii dá rozumieť ako zvýšeniu počtu prediktorov) variancia odhadu stúpa, ale klesá vychýlenie. Perfektne replikujúci odhad by sa nachádzal na pravej strane obrázku, čo je ďaleko od optimálneho riešenia. Preto použijeme regularizáciu: za cenu zvýšenia vychýlenia zjednodušíme model a posunieme sa do stredu - do optimálneho bodu.

4.2 Hrebeňová (ridge) regresia

Z úvahy v predchádzajúcej podkapitole vyplýva, že pokiaľ chceme znížiť komplexnosť modelu, musíme znížiť počet prediktorov. Môžeme použiť postupy ako výber smerom



Obr. 4.2: Vzťahy medzi vychýlením, varianciou a komplexnosťou modelu. Zdroj: [14].

dopredu alebo spätné odstraňovanie [1], lenže v tomto prípade nedokážeme povedať nič o vplyve odstránených premenných na výsledok. Na odstránenie premenných sa ale môžeme pozerať aj ako na nastavenie ich koeficientov rovných nule. Toto je prípad LASSO regresie, ku ktorej sa vrátime v ďalšej podkapitole. V hrebeňovej (z angl. ridge) regresii koeficienty, ktoré sú príliš ďaleko od nuly, penalizujeme a teda umelo zmenšíme. Takto znížime komplexnosť modelu, zatiaľ čo všetky premenné v modeli ostávajú. Stratová funkcia v prípade hrebeňovej regresie je na rozdiel od OLS (4.2) doplnená o penalizáciu koeficientov:

$$\min_{\beta} L_{ridge}(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 = \min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2. \tag{4.8}$$

Po vyriešení tohto problému dostaneme odhad $\hat{\beta}(\lambda)$, ktorý je pre hrebeňovú regresiu $\hat{\beta}_{ridge}(\lambda) = (X^T X + \lambda I)^{-1}(X^T y)$, kde I je matica s jednotkami na diagonále.

Parameter λ je regularizačný parameter. Jeho vhodnému zvoleniu sa budeme venovať na ďalších stranách. Zatiaľ upozorníme na jeho jednu dôležitú vlastnosť v limitných situáciách:

[1] Výber smerom dopredu (z angl. forward selection) začína s prázdnou rovnicou. Postupne po jednom sa pridávajú prediktory začínajúc tým, ktorý má s cieľovou premennou najväčšiu koreláciu. Prednostne sa potom pridávajú prvky s väčšou teoretickou dôležitosťou, pokiaľ zlepšujú model. Spätné odstraňovanie (s angl. backward deletion) je opačný proces. Začína sa s modelom so všetkými dostupnými premennými a postupne sa po jednej odstraňujú, pokiaľ neprispievajú k zlepšeniu. [6]

- Pre $\lambda \to 0$, $\hat{\beta}_{ridge}(\lambda) \to \hat{\beta}_{OLS}$;
- Pre $\lambda \to \infty$, $\hat{\beta}_{ridge}(\lambda) \to 0$.

Zvolením λ rovnej 0 teda dostaneme rovnaký model ako OLS, zatiaľ čo zvyšovaním parametra zvýšime aj silu penalizácie.

4.2.1 Vzťah medzi vychýlením a varianciou v hrebeňovej regresii

Po rozšírení stratovej funkcie v hrebeňovej regresii o regularizačný parameter vieme odvodiť momenty koeficientov. Stredná hodnota sa dá vyjadriť pomocou $\hat{\beta}_{OLS}$ [18]:

$$\begin{aligned} E[\hat{\beta}_{ridge}(\lambda)] &= E[(X^TX + \lambda I)^{-1}(X^Ty)] = \\ &= E[(X^TX + \lambda I)^{-1}(X^TX)(X^TX)^{-1}(X^Ty)] = E[(X^TX + \lambda I)^{-1}(X^TX)\hat{\beta}_{OLS}] = \\ &= (X^TX + \lambda I)^{-1}(X^TX)E(\hat{\beta}_{OLS}) = \underbrace{(X^TX + \lambda I)^{-1}(X^TX)}_{W}\beta. \end{aligned} \tag{4.9}$$

Maticu $(X^TX + \lambda I)^{-1}X^TX$ sme v poslednom kroku vo vzorci 4.9 označili ako W. Pomocou tejto matice odvodíme varianciu koeficientov pri hrebeňovej regresii [18]:

$$\begin{aligned} W\hat{\beta}_{OLS} = W(X^TX)^{-1}X^Ty = (X^TX + \lambda I)^{-1}\underbrace{X^TX(X^TX)^{-1}}_{I}X^Ty = \\ = (X^TX + \lambda I)^{-1}X^Ty = \hat{\beta}_{ridge}(\lambda). \end{aligned}$$

$$\begin{aligned} Var[\hat{\beta}_{ridge}(\lambda)] = \underbrace{Var[W\hat{\beta}_{OLS}]}_{Var(AX) = AVar(X)A^T} = WVar[\hat{\beta}_{OLS}]W^T = \sigma^2 W(X^TX)^{-1}W^T = \\ = \sigma^2(X^TX + \lambda I)^{-1}X^TX[(X^TX + \lambda I)^{-1}]^T. \end{aligned} \tag{4.10}$$

Zo vzorcov 4.3 a 4.9 vyplýva, že s rastúcou λ sa bude znižovať $E[\hat{\beta}_{ridge}(\lambda)]$ a teda vychýlenie bude rásť. Naopak, variancia bude na základe vzorca 4.10 s rastúcou λ klesať. Z tejto skutočnosti vzniká otázka: do akej miery sme ochotní vychýliť odhad s cieľom znížiť varianciu? Alebo: aká je optimálna hodnota parametra λ ?

4.3 Voľba regularizačného parametra λ

Existuje viacero spôsobov, ako zvoliť vhodný regularizačný parameter. Tradičnejší prístup by bol zvoliť λ pomocou nejakého informačného kritéria (Akaikeho alebo Bayesovo informačné kritérium) a vybrať hodnotu λ, ktorá toto informačné kritérium minimalizuje. Takýto prístup zlepší replikáciu modelu na dáta [14], tzv. in-sample. My použijeme techniku, ktorá je viac zameraná na prediktívny výkon [14] - krížovú validáciu.

4.3.1 Krížová validácia

Na vykonanie krížovej validácie je potrebné zvoliť si P diskrétnych hodnôt λ, rozdeliť dataset do K skupín a postupovať podľa algoritmu:

Krížová validácia

- Pre p (1 až P):
  - pre k (1 až K):
    - vyčleň k-tu skupinu ako testovaciu sadu,
    - ostatné skupiny použi spolu ako trénovaciu sadu s $\lambda = \lambda_p$ na odhad $\hat{\beta}_{ridge}$,
    - predikuj ciel'ovú premennú na testovacej sade $y_{test,k} = X_{test,k}\hat{\beta}_{ridge}$,
    - spočítaj súčet štvorcov rezíduí: $SSR_k = \|y - y_{test,k}\|^2$,
  - koniec k,
  - vypočítaj priemerné SSR pre p-tu iteráciu: $SSR_p = \frac{1}{K}\sum_{k=1}^{K} SSR_k$,
- koniec p,
- vyber optimálnu hodnotu $\lambda_{opt} = argmin_p SSR_p$.

4.4 LASSO regresia

LASSO regresia (z angl. Least Absolute Shrinkage and Selection Operator) je koncepčne podobná hrebeňovej regresii. Takisto pridáva penalizáciu nenulovým koeficientom, ale na rozdiel od hrebeňovej regresie, ktorá penalizuje súčet štvorcov koeficientov (L2 penalizácia), LASSO regresia penalizuje súčet ich absolútnych hodnôt (L1 penalizácia). Dôsledkom je, že pre väčšie hodnoty parametra λ sú niektoré koeficienty rovné presne nule, čo nenastáva pri hrebeňovej regresii, kde sa koeficienty blížia k nule iba asymptoticky. Tento dôsledok je pre náš zámer kľúčový. Z veľkého počtu prediktorov, ktoré v našom prípade predstavujú všetky dostupné produkty zozbierané online, ostanú

ibe tie, ktoré najlepšie replikujú CPI za kategórie "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak".

Minimalizačná úloha je v prípade LASSO regresie definovaná nasledovne:

$$\min_{\beta} L_{lasso}(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{m} |\beta_j|. \tag{4.11}$$

4.5 Elastic net regresia

LASSO regresia dobre zovšeobecní model, keď je v modeli málo prediktorov so štatisticky významným koeficientom a ostatné sú blízke nule. Teda z celkového počtu prediktorov iba malý počet signifikantne ovplyvňuje cieľovú premennú. Hrebeňová regresia sa zase ukazuje ako dobrá voľba v prípadoch, keď existuje veľa koeficientov s približne rovnakou hodnotou a teda, keď veľa prediktorov ovplyvňuje výsledok [14]. V praxi ale skutočnú hodnotu koeficientov nevieme a teda takéto porovnanie použiteľnosti oboch metód je pomerne teoretické. Preto bol vyvinutý ďalší koncept a to Elastic Net regresia.

Pôvodne sa táto technika objavila ako riešenie kritiky LASSO regresie, kde výber premenných v prípade príliš veľa prediktorov je obmedzený vzhľadom na počet pozorovaní kvôli konvexnej povahe optimalizačného problému [20]. Riešenie spočíva v skombinovaní hrebeňovej a LASSO regresie s cieľom dostať to najlepšie z oboch metód. Minimalizačná úloha je teda sformulovaná ako kombinácia týchto dvoch metód:

$$\min_{\beta} L_{enet}(\beta) = \min_{\beta} \frac{\sum_{i=1}^{n} (y_i - x_i^T \beta)^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \beta_j^2 + \alpha \sum_{j=1}^{m} |\beta_j| \right), \tag{4.12}$$

kde α je parameter miešania medzi hrebeňovou regresiou ($\alpha = 0$) a LASSO regresiou ($\alpha = 1$). Pre jeho optimálne zvolenie môžeme znovu použiť krížovú validáciu. Otázku môže vzbudzovať predelenie prvého člena výrazom $2n$. Takto upravený výraz hovorí o priemernej hodnote štvorcov chýb (v tomto prípade ešte predelených dvomi), čo je v praxi rozšírená metrika pre porovnanie presnosti modelu. Bez takejto úpravy by vo vzorci vystupoval iba súčet štvorcov chýb, ktorý sa môže výrazne líšiť v závislosti od počtu záznamov v datasete a teda na jeho základe je ťažšie modely porovnávať (špeciálne ak sú budované nad množinami s rôznymi počtami záznamov). Dôvod pre-

delenia číslom 2 je, že pri derivácii pomôže zrušiť číslo 2, ktoré vznikne z exponentu pri derivácii. Rovnaký dôvod platí aj pri člene hrebeňovej regresie, ktorý je takisto predelený číslom 2. Vzhľadom na to, že parameter λ je voľný, tieto zmeny nemajú na finálny výsledok žiadny vplyv. Pri krížovej validácii je lambda určená iteratívne a teda finálna hodnota je preškálovaná, aby poskytla rovnaké riešenie ako v prípade, ak by sme tieto úpravy v rovnici nepoužili[2].

[2]Vzorec udávame z dôvodu, že v softvéri `RStudio` a knižnici `glmnet`, ktorú sme používali pre našu prácu, sa pracuje práve s takto definovanou úlohou.

Kapitola 5

Alternatívne metódy merania inflácie

V predchádzajúcich kapitolách sme načrtli niekoľko metód, ako zefektívniť proces tvorby inflačných štatistík. V tejto kapitole uvedieme tri konkrétne prístupy, ako zostrojiť alternatívny spotrebný kôš pre potreby merania CPI.

5.1 Tvorba spotrebného koša podľa metodiky ČSÚ

5.1.1 Určenie reprezentantov

V prvom prístupe pre tvorbu alternatívneho spotrebného koša sme spracovali replikáciu spotrebného koša určeného Českým štatistickým úradom (Český Statistický Úřad - ČSÚ), t.j. z množiny produktov zozbierných v rámci scrapovania online obchodov sa snažíme vybrať tie, ktoré svojimi vlastnosťami najviac zodpovedajú oficiálnemu spotrebnému košu ČSÚ. Postupovali sme v zmysle oficiálnej metodiky popísanej v podkapitole 1.3. Takýto tradičný prístup zahŕňal manuálne dohľadávanie produktov podľa definovaného spotrebného koša ČSÚ, kvantitatívne očistenie a následné vypočítanie alternatívneho CPI na základe cien položiek v nami vybratom spotrebnom koši. Očakávaný výsledok bol, že medzimesačná zmena hodnoty CPI, teda inflácia, bude rovnaká, resp. veľmi podobná ako pri výsledkoch ČSÚ. Pridaná hodnota tohto prístupu je nízka cena získaných dát, rýchlosť získania dát a teda celkovo efektívnejší proces.

Kategória "Potraviny a nealkoholické nápoje" obsahuje 158 položiek, "Alkoholické

nápoje a tabak" obsahuje 23 položiek. My sme tieto položky dohľadávali na všetkých troch uvedených e-shopoch. Jediné dve kritéria výberu bol názov a cena[1] z minulých období nameraných ČSÚ, nakoľko ČSÚ nešpecifikuje jednotlivé položky na úroveň producenta. Postupovali sme podľa nasledovného algoritmu, kde i symbolizuje i-tu položku:

Algoritmus výberu produktu

- Pre i v (1 až 181):
  - nájdi i-ty tovar z oficiálneho spotrebného koša podl'a názvu na portáloch *rohlik.cz*, *kosik.cz* a *itesco.cz*,
  - zohl'adni cenu ČSÚ a vyber tovar, ktorý najlepšie zodpovedá popisu a cena sa zároveň výrazne nelíši od ceny nameranej ČSÚ (tolerancia ±50% ceny),
  - ak sa rovnaký tovar spĺňajúci kritériá nachádza na všetkých portáloch, vyber tento tovar
  - ak sa rovnaký tovar spĺňajúci kritériá nachádza na dvoch portáloch, vyber tento tovar,
  - ak sa tovar spĺňajúci kritériá nachádza len na jednom portáli, vyber tento tovar,
  - urči koeficient kvantitatívneho očistenia (napr. ak sa tovar predáva v balení 500g, ale v špecifikácii spotrebného koša je uvedená cena pre 250g, koeficient bude 0,5),
- koniec i.

Snažili sme sa v prvom rade o nájdenie takého kandidáta, ktorý najlepšie vystihuje kandidáta zo spotrebného koša a v prípade, že bolo takýchto tovarov viac, vybrali sme ten, ktorý sa vyskytoval vo všetkých e-shopoch[2]. Cena bola určená ako jednoduchý priemer zo zistených cien. Niektoré položky sa nevyskytovali vo všetkých e-shopoch

[1]Cena z meraní ČSÚ bola smerodajná v zmysle, že sme nevyberali príliš kvalitné (drahé), alebo príliš nekvalitné (lacné) produkty. Po zohľadnení tejto ceny sme teda vyberali produkty približne rovnakej kvality ako ČSÚ.

[2]Optimálnym kritériom výberu by bol objem predaja danej položky, no takúto informáciu sme nemali k dispozícii.

a teda ceny boli zisťované len z dvoch alebo jedného obchodu. Po zosumarizovaní jednotlivých položiek sme ešte tieto ceny upravili tak, aby ceny boli za rovnakú mernú jednotku, akú udáva ČSÚ. CPI je následne počítaný ako pomer súčtov týchto cien prenásobených oficiálnymi váhami za sledované a bázické obdobia.

5.1.2 Tvorba spotrebného koša na základe dát z troch e-shopov

V prvom prístupe sme vytvorili spotrebný kôš použitím dát, ktoré sme scrapovali v období 14 týždňov. Tieto dáta mali výhodu presnejších cien, nakoľko brali do úvahy tri e-shopy. Nevýhodou je kratší časový interval, ktorý nam neumožňuje porovnávať CPI na dlhom časovom horizonte. Z tohto dôvodu sme spracovali medzimesačné indexy pre 3 mesiace a porovnali hodnoty s metrikou určenou ČSÚ.

V Tabuľke 5.1 sú uvedené dve problematické položky, kde sa produkt buď vôbec nenachádzal na trhu v danom období alebo sa nepodarila nájsť cenovo porovnateľná alternatíva. Pri oboch produktoch sme teda prebrali cenu ČSÚ. Nedostupnosť, resp. vysoká cena broskýň a nektarínok sa dá vysvetliť sezónnosťou, nakoľko v zimných a skorých jarných mesiacoch sa tento produkt skoro vôbec na trhu nenachádza. V prípade piškótovej rolády takýto produkt začali e-shopy ponúkať až v marci, alternatívy dostupné do tohto obdobia sa nám nezdali ako vhodná náhrada.

Produkt		Január	Február	Marec
BROSKVE/NEKTARINKY	Priemer cena e-shop	NA	134.90 Kč/Kg	130,52 Kč/Kg
	Cena ČSÚ	45,95 Kč/Kg	45,95 Kč/Kg	45,95 Kč/Kg
PIŠKOTOVÁ ROLÁDA	Priemer cena e-shop	NA	NA	102,83 Kč/Kg
	Cena ČSÚ	112,01 Kč/Kg	107,92 Kč/Kg	109,21 Kč/Kg

Tabuľka 5.1: Problémové položky pri tvorbe koša z troch e-shopov.

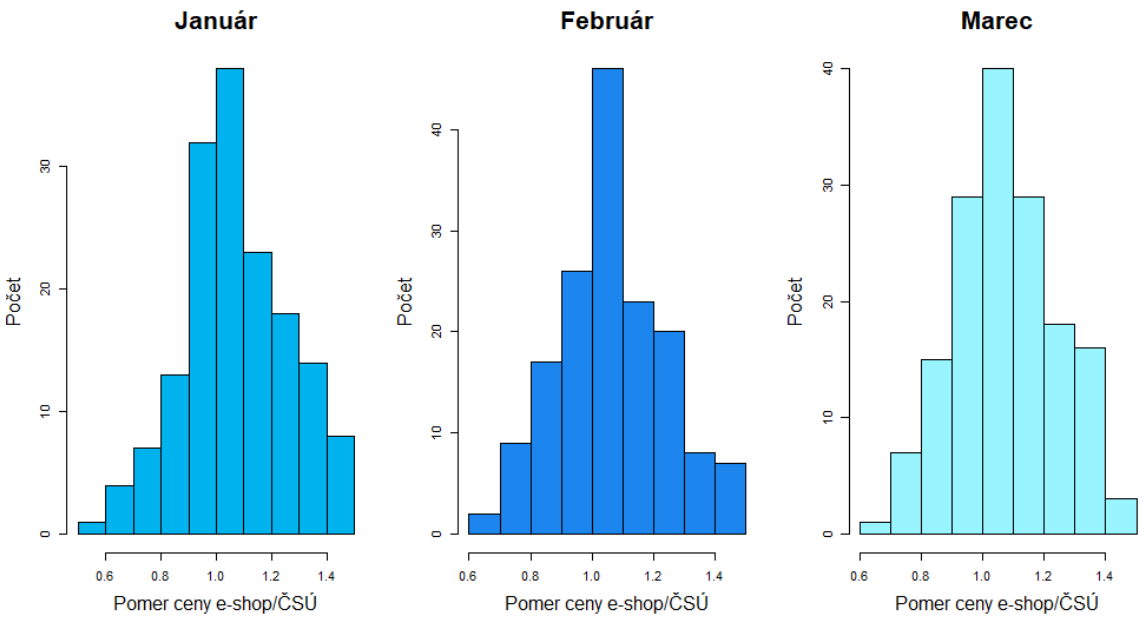
Medzimesačné CPI na rôznych košoch a obdobiach pre kategóriu "Potraviny a nealkoholické nápoje" sú v Tabuľke 5.2. Z týchto pozorovaní sa môže javiť, že ceny na internete su citlivejšie na zmenu, nakoľko CPI z e-shopov je v marci vyššie o 0,57 % a vo februári zase nižšie o 1,76 % ako štatistika ČSÚ. V čase dokončovania diplomovej práce ešte neboli dostupné dáta z ČSÚ za apríl a preto sme nevedeli porovnať metriky, ale náš kôš indikuje mierne zlacnenie oproti cenám v marci. Pozítivna je zhoda smeru

CPI v porovnávaných košoch; teda pokles vo februári a rast v marci.

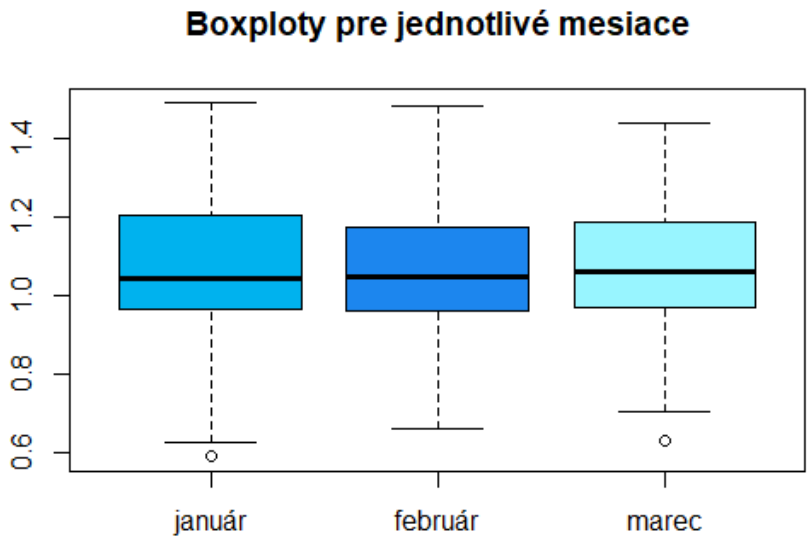
Mesiac	Medzimesačný CPI ČSÚ	Medzimesačný CPI alternatívneho spotrebného koša	Absolútna chyba	Relatívna chyba
apríl	NA	99,30	NA	NA
marec	100,13	100,70	0,57	0,5693 %
február	99,89	98,13	1,76	1,7935 %

Tabuľka 5.2: Porovnanie reálneho a nami dopočítaného CPI v kategórii "Potraviny a nealkoholické nápoje".

Obrázky 5.1 a 5.2 ilustrujú vzťah medzi cenami 158 produktov, ktoré sme zozbierali my a ktoré zozbieral ČSÚ. Histogramy vykazujú normálne rozdelenie, čo sme otestovali Shapiro-Wilk testom, pomocou ktorého sme hypotézu normality nezamietli. Z boxplotových grafov sa dá usudzovať, že tovary v online predaji majú v priemere vyššie ceny. Po otestovaní jednovýberovým t-testom na hladine 5% sa táto úvaha potvrdila, keď vo všetkých troch mesiacoch bola p-hodnota takmer nulová a teda nulovú hypotézu, ktorá hovorila o tom, že ceny v kamenných predajniach sú oproti e-shopom v priemere menšie, sme zamietli.



Obr. 5.1: Histogramy pomeru cien v kategórii "Potraviny a nealkoholické nápoje".



Obr. 5.2: Boxploty pomeru cien v kategórii "Potraviny a nealkoholické nápoje".

O poznanie horšie výsledky sme dosiahli v kategórii "Alkoholické nápoje a tabak". Porovnanie indexov zhŕňame v Tabuľke 5.3. Nekonzistencia s metrikou ČSÚ sa dá vysvetliť nízkym počtom reprezentantov, ktoré majú všeobecne vyššie váhy ako v kategórii "Potraviny a nealkoholické nápoje"[3] a teda cenový výkyv už len jedného reprezentanta má na celkový výsledok signifikantnejší vplyv.

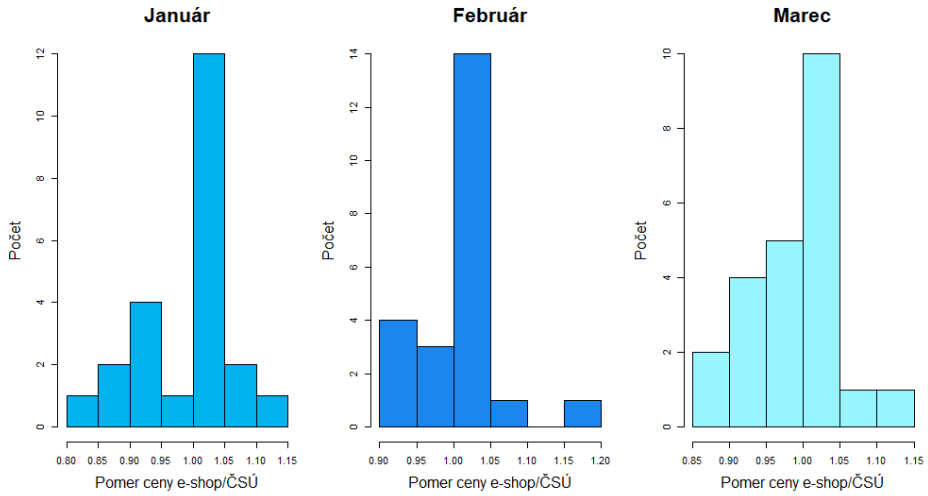
Mesiac	Medzimesačný CPI ČSÚ	Medzimesačný CPI alternatívneho spotrebného koša	Absolútna chyba	Relatívna chyba
apríl	NA	99,46	NA	NA
marec	100,86	99,41	1,45	1,4586 %
február	98,24	100,87	2,63	2,6771 %

Tabuľka 5.3: Porovnanie reálneho a nami dopočítaného CPI v kategórii "Alkoholické nápoje a tabak".

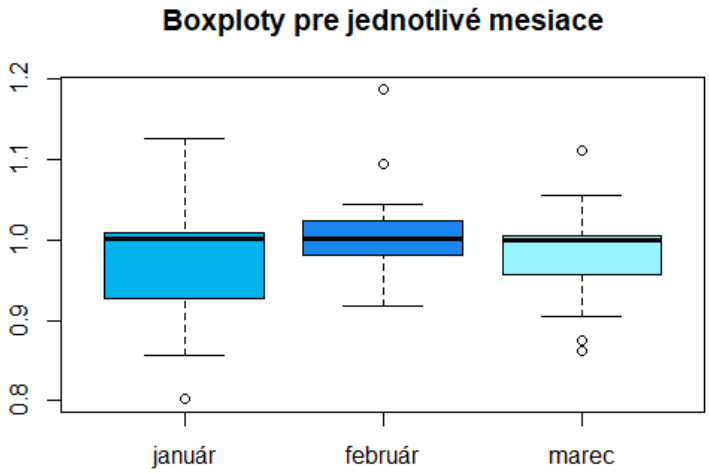
Po otestovaní Shapiro-Wilk testom sme normalitu zamietli len pre februárové

[3]Kategória "Potraviny a nealkoholické nápoje" je tvorená 158 reprezentantami, ktoré v celom spotrebnom koši reprezentujú váhu 177,6 ‰, zatiaľ čo kategória "Alkoholické nápoje a tabak" je tvorená 23 reprezentantami s váhou 92,2 ‰.

pomery cien. Preto sme na otestovanie hypotézy, či sú ceny v e-shopoch drahšie, použili pre február jednovýberový Wilcoxonov test, kde sme nulovú hypotézu, ktorá hovorí o tom, že ceny v e-shopoch nie sú vyššie, nezamietli. Pri januári a marci sme skrz normalitu dát mohli použiť jednovýberový t-test a v oboch prípadoch sme nulovú hypotézu takisto nezamietli. Podobne, ako pri kategórii "Potraviny a nealkoholické nápoje", aj tu ilustrujeme rozdelenie pomerov cien pomocou histogramov (Obrázok 5.3) a boxplotov (Obrázok 5.4).



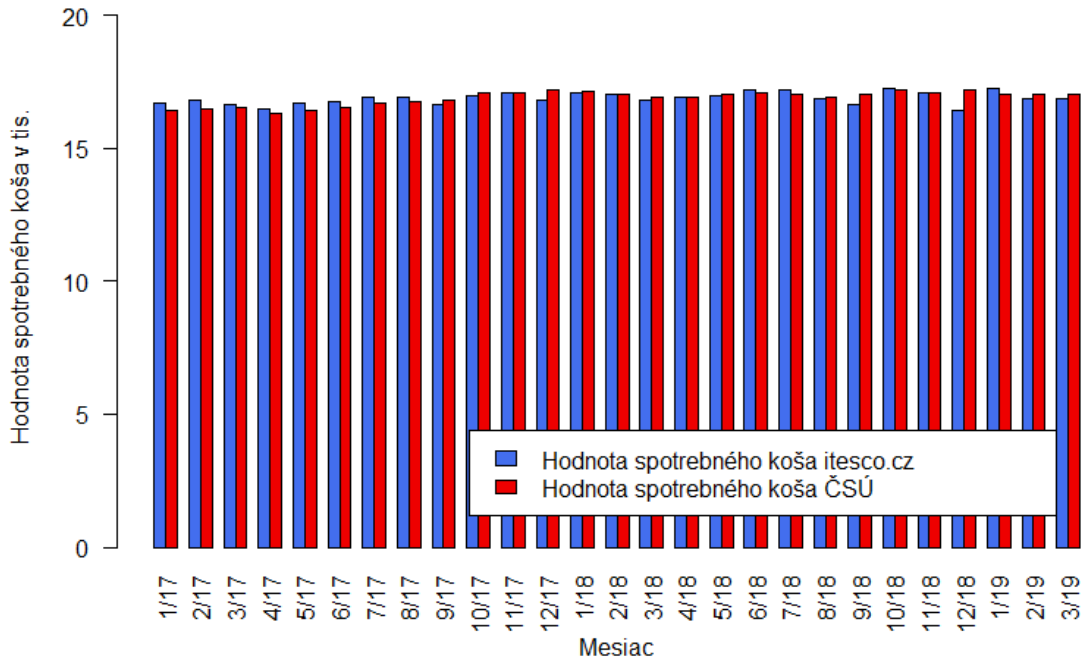
Obr. 5.3: Histogramy pomeru cien v kategórii "Alkoholické nápoje a tabak".



Obr. 5.4: Boxploty pomeru cien v kategórii "Alkoholické nápoje a tabak".

5.1.3 Tvorba spotrebného koša na základe dát z ČNB

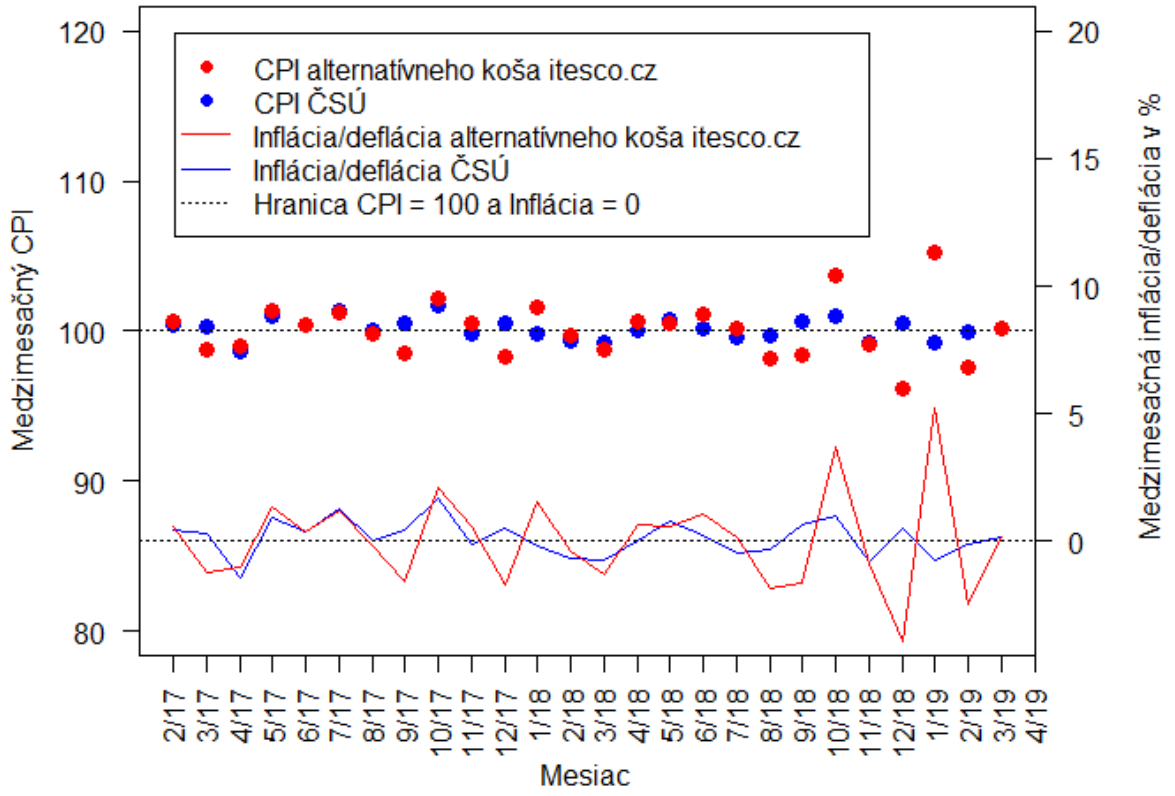
V druhom prístupe sme použili sme dáta z ČNB, konkrétne upravený dataset popísaný v 2.3. Väčší rozsah pozorovaní nám umožnil porovnať CPI na dlhšom časovom horizonte. Spotrebný kôš sme vytvárali pre kategóriu "Potraviny a nealkoholické nápoje" rovnakou metodikou, akú sme popísali v sekcii 5.1.1. Vzhľadom na to, že dáta boli iba z jedného e-shopu, častejšie sme mali problém vyhľadať vhodného cenového reprezentanta, konkrétne sa nám to nepodarilo pri 48 záznamoch. Po zohľadnení váh tieto položky tvorili 20,8% z ceny spotrebného koša a preto sme v ďalšom kroku zhodnotili, že ich ceny preberieme z ČSÚ. Následne sme vypočítali hodnotu takto zostrojeného spotrebného koša (súčet cien položiek prenásobených prísušnými váhami) a porovnali sme jeho hodnotu s hodnotou spotrebného koša zostrojeného ČSÚ, čo ilustrujeme na Obrázku 5.5.



Obr. 5.5: Hodnoty spotrebného koša tvoreného položkami z *itesco.cz* a spotrebného koša ČSÚ.

Na základe týchto súčtov potom vieme vypočítať medzimesačný CPI, ktorý dostaneme vydelením hodnôt spotrebných košov po sebe idúcich mesiacov. Tento priebeh sme znázornili na Obrázku 5.6. Opäť môžeme vidieť, že internetový CPI je o niečo

volatilnejší ako oficiálny CPI a to najmä v poslednom období.



Obr. 5.6: Porovnanie CPI a inflácie alternatívneho koša z e-shopu *itesco.cz* a oficiálneho spotrebného koša ČSÚ.

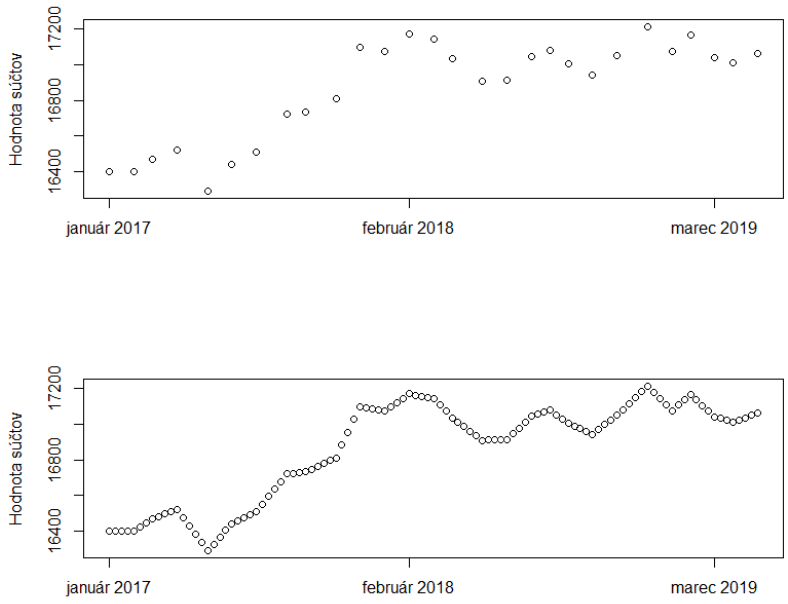
5.2 Regularizácia pre vytvorenie spotrebného koša

Ako sme uviedli v úvode kapitoly 4, regularizácia môže byť pomocná pri znížení variancie modelu za cenu určitého vychýlenia. Použijeme túto vlastnosť na navrhnutie optimálneho spotrebného koša na základe dát zozbieraných v ČNB. V kapitole 1.3.5 sme uviedli, ako ČSÚ počíta CPI. CPI je pomer dvoch súčtov za sledované a bázické obdobie. Tieto súčty vyjadrujú cenu spotrebného koša v danom období a obsahujú položky definované ČSÚ prenásobené váhami určenými štatistikou rodinných účtov. K týmto súčtom sa vieme dopracovať na základe dát od ČSÚ, v ktorých nájdeme mesačné ceny od januára 2017. Z nich sme vypočítali súčet pre každý mesiac, ktorý slúžil ako

cieľová premenná. Naším hlavným cieľom bolo nájsť optimálnu lineárnu kombináciu položiek z online obchodov tak, aby čo najlepšie aproximovala vývin súčtov vstupujúcich do CPI - nájsť dané produkty (prediktory) a ich váhy (koeficienty).

Ako sme uviedli v 2.3, dáta z ČNB obsahujú 457 pozorovaní web-scrapnutých dát z *itesco.cz* v časovom úseku november 2016 až apríl 2019 a tieto dáta nám ČNB poskytla na účely tvorby alternatívneho koša v rámci našej diplomovej práce.

Nakoľko dáta zozbierané v ČNB mali týždňovú periodicitu, pre zvýšenie počtu záznamov sme mesačné indexy ČSÚ lineárne interpolovali. Takto sme sa z 28 pozorovaní dostali na 112 pozorovaní. Ako ťažiskové týždne - týždne, ktorým sme priradili mesačný súčet, sme vzhľadom na časovanie zberu dát ČSÚ (ČSÚ meria ceny do 20. dňa v mesiaci) zvolili vždy posledné týždne v mesiacoch. V mesiaci január 2017 (prvý záznam) sme každému týždňu pridali januárový súčet, keďže sme nemali hodnotu predchádzajúceho mesiaca a teda nemohli sme tak presnejšie interpolovať. Na Obrázku 5.7 ilustrujeme vývin súčtov za obdobie január 2017 - marec 2019.



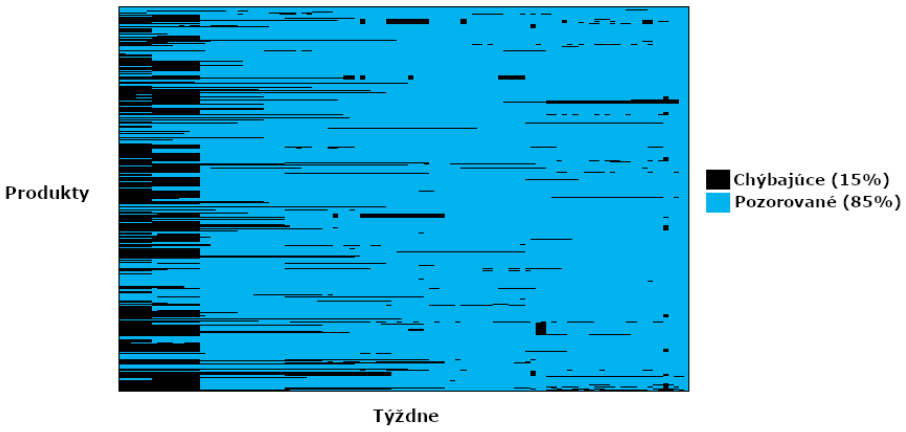
Obr. 5.7: Vývin cien spotrebného koša za obdobie január 2017 - marec 2019, hore mesačné súčty, dole ich týždňová interpolácia.

Dáta s cenami od ČNB sme najprv pospájali do týždňov. Pre každý týždeň sme robili plné spájanie (tzv. full join) na základe názvu, t.j. pokiaľ sa v jednom týždni

zistila cena položky aspon raz, zahrnuli sme túto položku do zoznamu. Pokiaľ sa zistilo viac cien ako jedna, zahrnuli sme priemernú týždennú hodnotu.

Následne sme všetky týždne pospájali štýlom ľavého spájania (tzv. left join), kde sme ako základný týždeň brali posledný týždeň v marci 2019 a na tento týždeň sme "nalepili" ostatné týždne na základe názvu položiek. Zvolili sme takýto prístup kvôli aktuálnosti dát, pri experimentoch s chronologickým párovaním (ako základný týždeň použitý prvý januárový 2017) sa ukazovalo, že veľa položiek nebolo v posledných týždňoch našich dát dostupných.

Po takomto pospájani sme mali k dispozícii 2633 riadkov (jeden riadok pre každý produkt) a 112 stĺpcov (jeden stĺpec pre každý týždeň). Avšak z celého datasetu predstavovalo až 38% chýbajúce hodnoty. Preto sme produkty, ktoré mali viac ako polovicu meraní nedostupných, odstránili. Po tejto úprave sme sa dopracovali ku 1694 produktom, ktorým ostalo v datasete z celkového počtu záznamov už len 15% chýbajúcich hodnôt, čo ilustrujeme na Obrázku 5.8. Týždne sú zoradené zľava doprava v chronologickom poradí (od najstarších po najnovšie).



Obr. 5.8: Chýbajúce hodnoty v datasete.

Keďže pre posledný týždeň máme všetky pozorovania, rozhodli sme sa chýbajúce hodnoty doplniť vždy hodnotou z nadchádzajúcho týždňa.

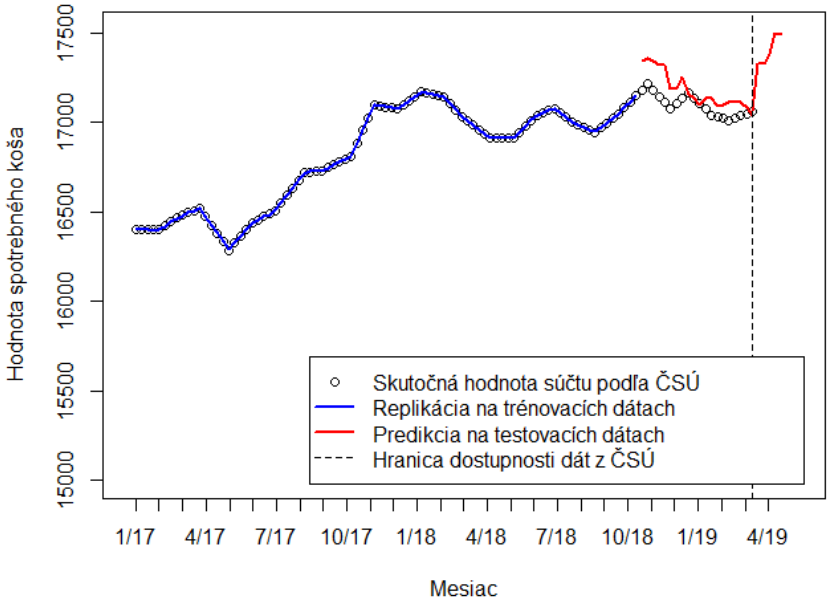
Pre spresnenie modelu sme zahrnuli tri binárne (v angl. dummy) premenné: Vianoce, Veľká noc a Black Friday. Dummy premenné nadobúdajú hodnotu 1 pre týždeň pred Vianocami a Veľkou nocou a počas týždňa, v ktorom je Black Friday[4]. Keďže

[4]Black Friday je špecifický piatok, keď predajcovia ponúkajú extrémne zľavy. Zvyčajne je vyhlásený

spotrebitelia v týchto týždňoch nakupujú viac ako obvykle, chceli sme zachytiť tento dočasný šok a pozorovať tak, či predajca *itesco.cz* v týchto týždňoch ponúka zľavy alebo zdvihne ceny.

Pre odhad a vyhodnotenie modelu sme si dáta rozdelili na trénovaciu (87 týždňov) a testovaciu množinu (25 týždňov). Následne sme stále na rovnakej trénovacej množine 100x natrénovali a vyhodnotili model - LASSO regresiu. Pretože parameter λ získame pomocou krížovej validácie, môže nadobúdať rôzne hodnoty a teda aj model sa môže medzi iteráciami líšiť. Ako kritérium kvality sme zvolili koeficient determinácie $R^2 = \rho(y_{test}, \hat{y}_{test})^2$, teda umocnený korelačný koeficient medzi cieľovou premennou v testovacej množine a predikovanou cieľovou premennou v testovacej množine. $R^2 \in (0, 1)$, čiže čím bližšie je hodnota pri 1, tým lepšiu replikáciu dostávame a model je kvalitnejší. Keďže váhy v oficiálnom spotrebnom koši sú nezáporné, pre všetky premenné (okrem dummy premenných) sme pridali takéto ohraničenie aj do modelu. Takisto sme pridali ohraničenie, aby dummy premenné vždy ostali v modeli.

V našom prípade sme sa pri LASSO regresii dostali na maximálnu hodnotu $R^2 = 0,5831$. Na Obrázku 5.9 ilustrujeme vývin hodnoty spotrebného koša.



Obr. 5.9: Vývin hodnoty spotrebného koša pomocou LASSO regresie.

jeden krát v roku a často práve v novembri.

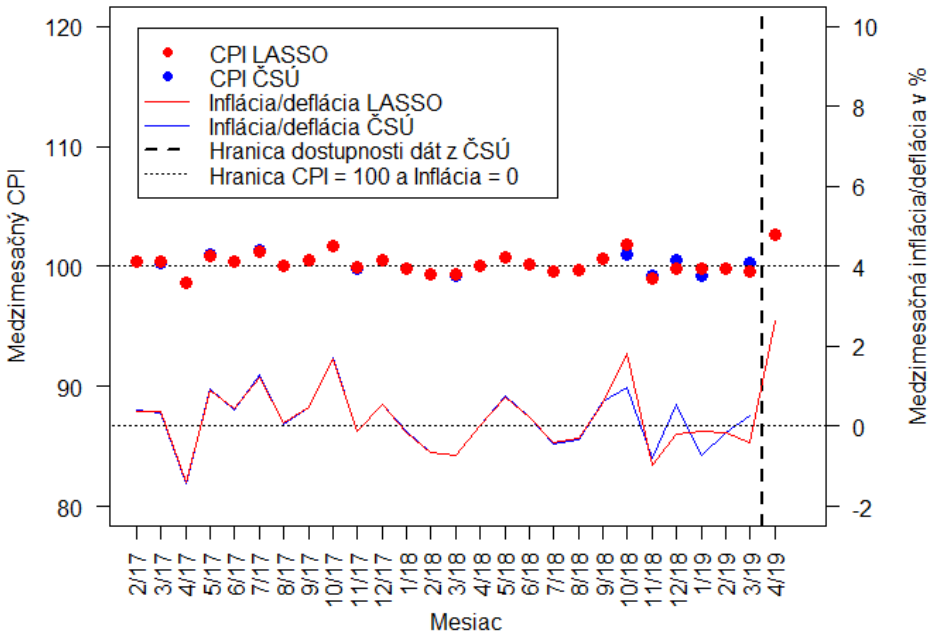
Keďže ku dňu 8.5.2019 sme stále nemali výsledky merania cien za apríl od ČSÚ, záznamy pre posledných 5 aprílových týždňov v obrázku chýbajú. Vedeli sme ale pre toto obdobie spraviť predikciu, ktorá naznačuje, že ceny v apríli rástli.

LASSO regresia nám vybrala 111 položiek[5] a 3 dummy premenné, ktoré aj s váhami - koeficientami prikladáme v Prílohe A. Hodnoty dummy premenných, ktoré sú v tejto prílohe na konci, indikujú, že pred Vianocami a Veľkou nocou sú ceny vyššie. Naopak, počas týždňa, kedy je Black Friday, naozaj prichádza k jemnému zníženiu cien. Black Friday akcia sa uplatňuje hlavne v elektronike, ale v posledných rokoch sa rozšíril aj do iných odvetví. Keďže ale pracujeme v sektore potravín, nie je táto zľava až taká výrazná.

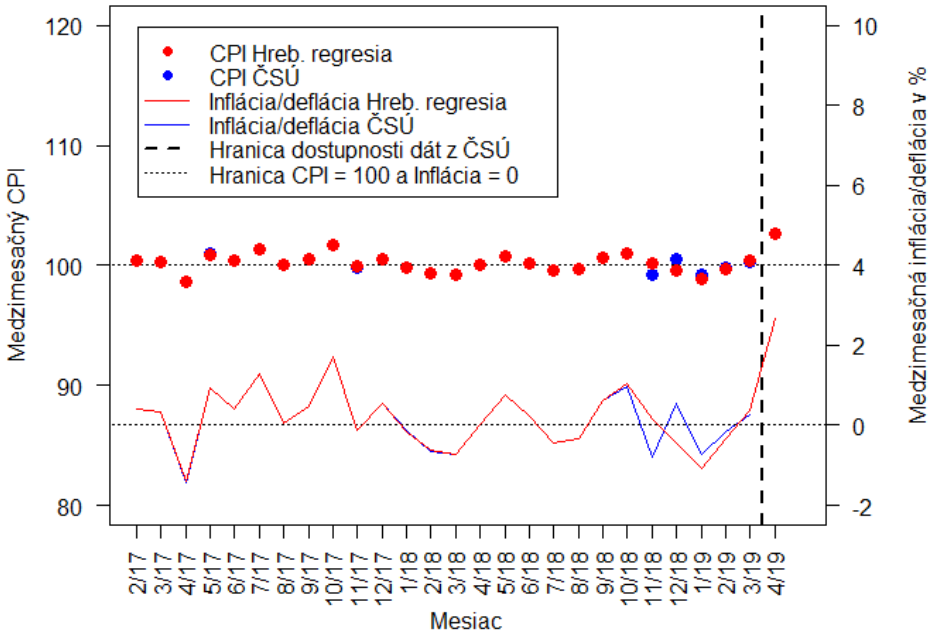
Vzhľadom na to, že sa v našej práci zameriavame na inflačnú metriku, na Obrázku 5.10 ilustrujeme vývin CPI v čase, v ktorom sme mali k dispozícii dáta. K hodnote CPI sme sa dostali jednoducho predelením hodnoty spotrebného koša hodnotou z predchádzajúcho mesiaca a následným prenásobením tejto hodnoty číslom 100. Takisto sme na obrázku zobrazili infláciu, resp. deflaciu, ktorá je určená rozdielom medzi CPI a hodnotou 100, ktorá reprezentuje nezmenené ceny. Hodnoty inflácie sme doplnili do grafu na pravú os. Na základe nášho modelu teda očakávame, že v apríli bude nameraná medzimesačná inflácia nad 2 %.

Výsledky modelu sú z nášho pohľadu uspokojivé. Vidíme, že na trénovacích dátach model dokáže dokonale replikovať CPI nameraný ČSÚ a predikcia tiež vykazuje známky správneho aproximovania vývoja indexu. Rozhodli sme sa tiež experimentovať s ostatnými regularizačnými regresiami. Na Obrázku 5.11 ilustrujeme replikáciu CPI pomocou hrebeňovej regresie. Podľa koeficientu determinácie $R^2 = 0,6016$ sme dostali jemne lepšiu zhodu pri predikovaní ako pri LASSO regresii. Model je však tvorený príliš veľkým počtom produktov a teda sa nedá efektívne replikovať štruktúra oficiálneho spotrebného koša. Takisto ako pri LASSO regresii model indikuje v apríli medzimesačnú inflácia nad 2 %.

[5] 12 z týchto 111 položiek sa vyskytovalo aj v spotrebnom koši, ktorý sme vytvorili podľa metodiky ČSÚ na základe dát z ČNB, ktorý opisujeme v sekcii 5.1.3.

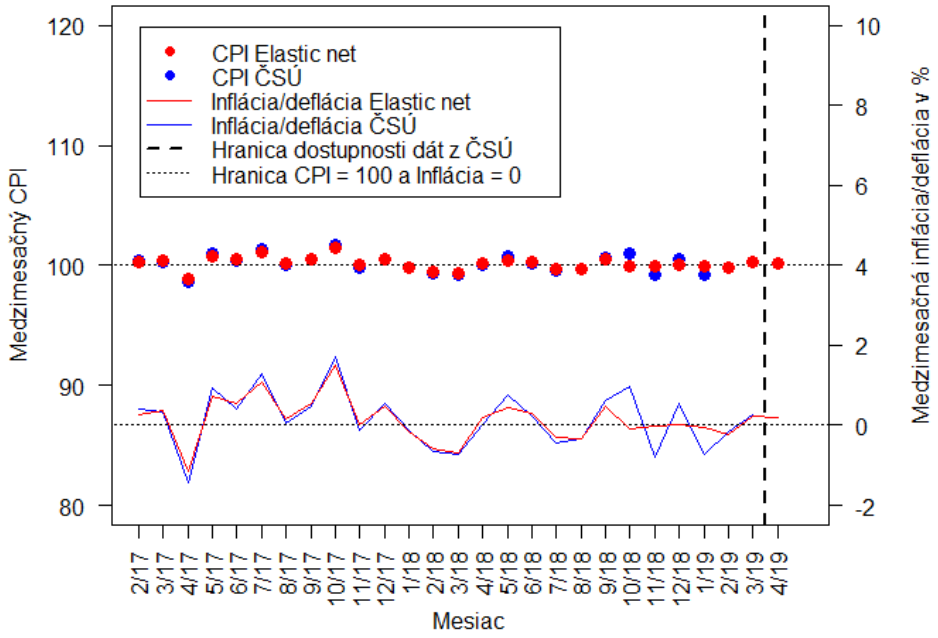


Obr. 5.10: Porovnanie CPI a inflácie medzi ČSÚ a modelom LASSO regresie.



Obr. 5.11: Porovnanie CPI a inflácie medzi ČSÚ a modelom hrebeňovej regresie.

Na záver sme ešte implementovali model Elastic Net regresie, ktorá je kombináciou LASSO regresie a hrebeňovej regresie. Na Obrázku 5.12 môžeme vidieť porovnanie CPI podľa ČSÚ a CPI podľa Elastic Net modelu. Elastic Net model nám z troch regularizačných regresií poskytuje najkonzervatívnejšiu predikciu. Takisto sa viac líši aj replikácia na trénovacích dátach, aj keď stále poskytuje pomerne presnú zhodu. Na základe tohto modelu by sme v apríli očakávali len jemne pozitívnu infláciu.



Obr. 5.12: Porovnanie CPI a inflácie medzi ČSÚ a modelom Elastic net regresie.

5.3 Alternatívny spotrebný kôš podľa kategórií

V poslednom alternatívnom spotrebnom koši sme vychádzali z agregácie zozbieraných produktov do kategórií, na ktorých sme tento kôš postavili. Predpokladali sme, že skupiny výrobkov s podobným zložením a vlastnosťami sa budú vyvíjať vo vzájomnej korelácii.

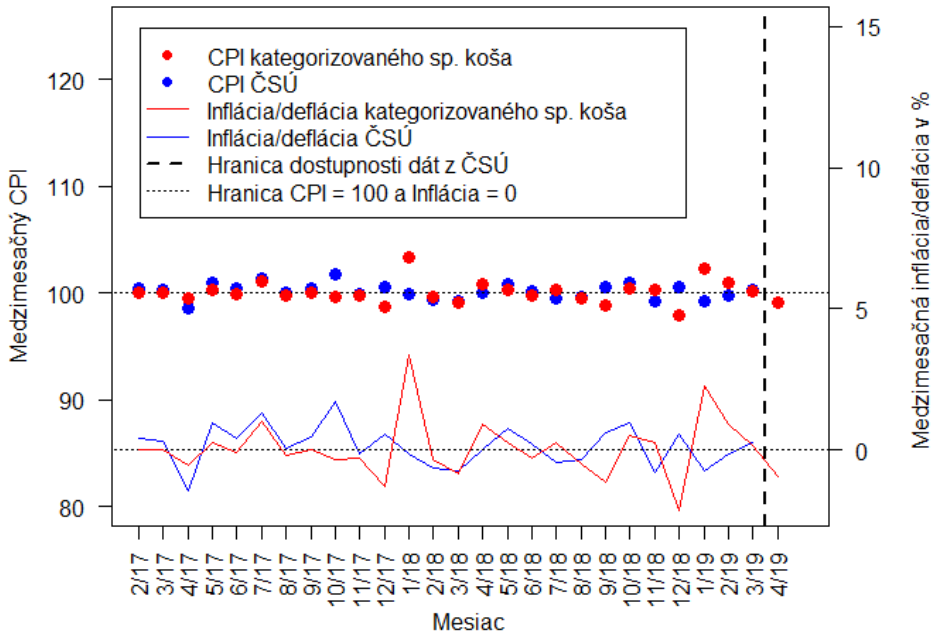
Ako sme uviedli v kapitole 3.2, vieme pomerne presne zakategorizovať položky získané stiahnutím z e-shopov, pokiaľ vychádzame z kategorizácie predajcu. V Tabuľke B2, ktorú prikladáme v Prílohe B, uvádzame spôsob prevedenia kategórii používaných

na portáli *itesco.cz* na ECOICOP kategóriu, ktoré sa používajú v členení oficiálneho spotrebného koša definovaného ČSÚ.

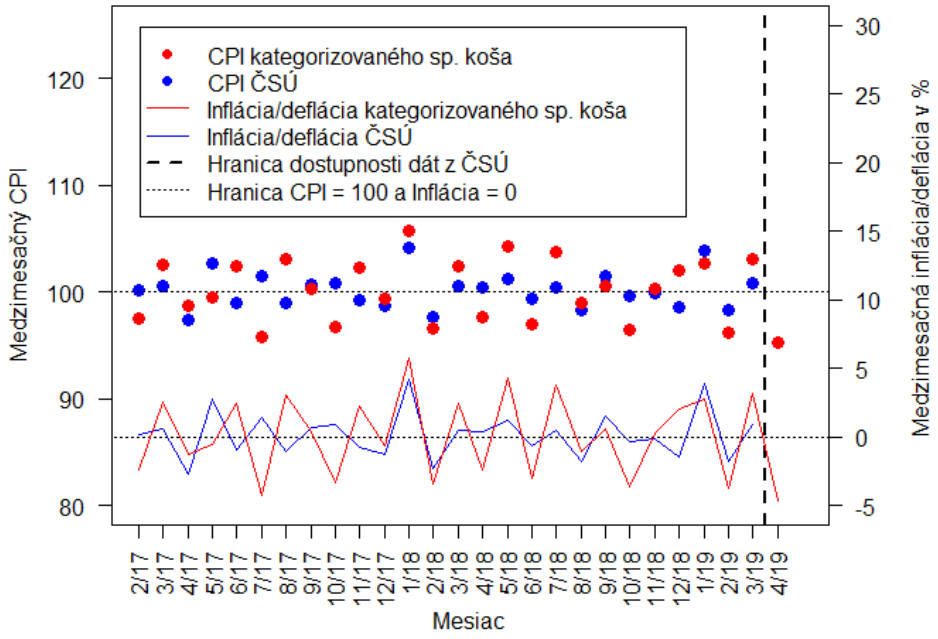
Vychádzali sme z rovnakého datasetu, ako sme opísali v 2.3 (t.j. dáta ČNB). Pre pozorovania, ktoré sú v tomto súbore, sme zo stránky *itesco.cz* stiahli kategorizáciu predajcu a previedli sme ju na kategorizáciu ECOICOP. Pre položky, pri ktorých sa nám nepodarilo kategóriu stiahnuť, sme použili metódu oporného bodu, ako sme opísali v 3.2. Následne sme v každom mesiaci zlúčili položky pre jednotlivé kategórie na priemernú cenu pre danú kategóriu. Priemerné ceny sme prenásobili váhami aké udáva spotrebný kôš ČSÚ (váhy jednotlivých skupín uvádzame v Tabuľke B1, ktorú prikladáme v Prílohe B) a súčet takto prenásobených hodnôt nám dal pre každý mesiac hodnotu alternatívneho spotrebného koša. Medzimesačný CPI sme následne dostali spätným predelením.

Na Obrázku 5.13 ilustrujeme, ako korešponduje CPI ČSÚ s CPI nášho alternatívneho koša v kategórii "Potraviny a nealkoholické nápoje". Vidíme, že vo viacerých mesiacoch sme dostali takmer totožnú hodnotu a preto sa dá táto metóda považovať za veľmi dobrú alternatívu. Výrazné výkyvy vznikajú hlavne v období na prelome rokov, kde to môže byť spôsobené nedostupnosťou dát v tomto období (posledný týždeň v každom roku ČNB dáta nesťahovala) a prítomnosťou sviatkov ako Vianoce a Nový rok. Táto metóda je ale z nášho pohľadu najlepším využitím dát z e-shopov, nakoľko pracuje so všetkými stiahnutými položkami.

Naopak, pri kategórii "Alkoholické nápoje a tabak" sa táto metóda neukazuje ako správna voľba. Na Obrázku 5.14 môžeme sledovať, že hodnoty CPI ČSÚ a CPI alternatívneho koša sú markantne odlišné. Je to pochopiteľné, keďže kategória "Alkoholické nápoje a tabak" obsahuje len 4 podskupiny s pomerne veľkými váhami a cenová zmena v jednej má veľký efekt na hodnotu celého koša.



Obr. 5.13: Porovnanie CPI a inflácie pre oficiálny a alternatívny spotrebný kôš agregovaný po skupinách v kategórii "Potraviny a nealkoholické nápoje".



Obr. 5.14: Porovnanie CPI a inflácie pre oficiálny a alternatívny spotrebný kôš agregovaný po skupinách v kategórii "Alkoholické nápoje a tabak".

Záver

V diplomovej práci sme sa venovali alternatívnym prístupom pre určenie CPI a jeho predikcii na krátkom časovom horizonte.

V prvej kapitole sme predstavili koncept inflácie. Ako sme uviedli v 1.1, vzhľadom na súčasnú digitalizáciu je proces tvorby CPI do značnej miery ťažkopádny. Tento fakt nás motivoval vyvinúť efektívnejší spôsob tvorby inflačných indexov a to využitím dát z e-shopov, konkrétne *rohlik.cz*, *kosik.cz* a *itesco.cz*. Predstavili sme preto metodiku určenia inflačných štatistík v Českej republike, ktorou sme sa riadili pri zostavovaní alternatívnych indexov. Vzhľadom na sortiment zmienených e-shopov sme CPI spracovávali pre dve kategórie: "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak".

Na účely diplomovej práce sme potrebovali pravidelene sťahovať dáta zo spomenutých e-shopov. Vyvinuli sme preto aplikáciu v softvéri `RStudio`, ktorá na pravidelnej báze raz za týždeň sťahovala všetky dostupné dáta z troch e-shopov. Nakoľko sme nesťahovali dáta dostatočne dlhú dobu, mali sme k dispozícii málo pozorovaní a teda tieto dáta pre tvorbu modelov nestačili. Obrátili sme sa preto na ČNB, ktorá experimentovala so sťahovaním dát z e-shopu *itesco.cz* od novembra 2016. Pre účely našej práce nám ČNB tieto dáta poskytla. Taktiež sme si vyžiadali od ČSÚ mesačné ceny položiek zo spotrebného koša nameraných pre Českú repuliku, aby sme vedeli replikovať indexy počítané ČSÚ a tieto výsledky porovnať s našimi.

V práci sme pokračovali predstavením kategorizácie produktov. Na tieto účely sme použili metódu oporného bodu, ktorej teóriu sme priblížili v 3.1. Použitím tejto metódy sme pri kategorizácii produktov dosiahli úspešnosť zaradenia do správnej kategórie na úrovni 91,09%. Metódu sme následne použili, aby sme zakategorizovali položky, pri ktorých sa nám nepodarilo stiahnuť kategóriu z e-shopu.

V štvrtej kapitole sme sa venovali regularizácii. Táto technika pomáha znižovať varianciu predikcie pri lineárnej regresii penalizovaním koeficientov za cenu zvýšenia vychýlenia. Z hľadiska našej praktickej aplikácie bola najefektívnejšia LASSO regresia, ktorá pri správne nastavenom parametri λ koeficientom, ktoré nie sú signifikantné, nastaví nulovú hodnotu. Takto vyfiltruje iba podstatné premenné, ktoré v našom prípade tvorili produkty alternatívneho spotrebného koša.

V poslednej kapitole sme všetky nadobudnuté vedomosti aplikovali v praxi, kde sme zostrojili tri indexy odlišnými metódami. V prvej metóde, kde sme pracovali s nami stiahnutými dátami a kde sme určili medzimesačný index v kategóriách "Potraviny a nealkoholické nápoje" a "Alkoholické nápoje a tabak", sme dostali uspokojivé výsledky, hlavne v kategórii "Potraviny a nealkoholické nápoje", keď sme správne určili infláciu v marci a deflaciu vo februári. Rovnakou metodikou sme následne zostrojili kôš aj z dát od ČNB. Vzhľadom na to, že sme mali k dispozícii len jeden dátový zdroj a nepodarilo sa nám dohľadať všetky produkty, v 48 prípadoch, čomu zodpovedá 20,8% ceny spotrebného koša, sme ceny prebrali z meraní ČSÚ.

Ďalší alternatívny spotrebný kôš sme vytvorili použitím LASSO regresnej techniky. Opäť sme vychádzali z dát z ČNB. Model LASSO regresie vybral 111 cenových reprezentantov, ktorých uvádzame v Prílohe A. Replikácia indexu bola obzvlášť dobrá na in-sample vzorke, ale model takisto pomerne dobre predikoval aj budúce indexy. V tejto kapitole sme ešte experimentovali s využitím hrebeňovej a Elastic net regresie, kde sme dosiahli podobné výsledky ako pri LASSO regresii.

Ako najspoľahlivejší sa ukázal alternatívny index pri zostavení spotrebného koša na základe kategórií. Najlepšiu kvalitu mu pripisujeme preto, pretože využíva všetky dáta stiahnuté z e-shopu a preto je najstabilnejší. Jediné výkyvy od CPI určeného ČSÚ nastávali opakovane v decembri a januári. Nezhoda v decembrovom a januárovom indexe môže byť spôsobená nedostupnosťou dát v poslednom týždni v roku a tak isto cenovými fluktuáciami počas sviatkov (Vianoce a Nový rok).

Naša práca môže slúžiť ako základ pre budúce rozpracovanie metodiky využitia online dát pre výpočet CPI. Veľkou výhodou dát z e-shopov je časová dostupnosť a rýchlosť ich zozbierania. Pri budúcom použití takýchto dát na obohatenie súčasného CPI ale treba zohľadniť niektoré odlišnosti online trhu, napríklad vyššie ceny.

Literatúra

[1] Breton, R. a kol.: Research indices using web scraped data: May 2016 update. *Newport: Office for National Statistics* (2016)

[2] Bryan, M. F.: On the origin and evolution of the word inflation. *Public Administration and Public Policy* 97 (2002), 593-600

[3] Cavallo, A., Roberto R.: The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30.2 (2016), 151-178

[4] Český statistický úřad: Indexy spotřebitelských cen (metodická příručka pro uživatele) 2019. Dostupné na internete (15.2.2019): http://bit.ly/2W0iIN8

[5] Eurostat: E-commerce statistics. Dostupné na internete (1.11.2018): https://bit.ly/2xIAhUt

[6] Halinski, R. S., Feldt, L. S.: The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, 7 (3) (1970), 151-157

[7] Hamilton, E. J.: *American treasure and the price revolution in Spain.* Cambridge, MA: Harvard University Press, 1934

[8] Harman, R.: Prednášky z Viacrozmerných štatistických analýz II, FMFI UK, Bratislava, 2018

[9] Huang, T.: Comparison of L2-Regularized Multi-Class Linear Classifiers. Master thesis, College of Electrical Engineering and Computer Science, National Taiwan University, 2010

[10] Investopedia: Inflation. Dostupné na internete (23.1.2019): https://bit.ly/2moZU6T

[11] Investopedia: Monetary Union Index of Consumer Prices - MUICP. Dostupné na internete (23.1.2019): https://bit.ly/2sVskbV

[12] Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *European conference on machine learning*. Springer, Berlin, Heidelberg (1998)

[13] Johnson, P.: UK consumer price statistics: A review. *UKSA* (2015)

[14] Oleszak, M.: Regularization: Ridge, Lasso and Elastic Net. Dostupné na internete (20.4.2019): http://bit.ly/2W2gUzW

[15] Patel, N., Villar, A.: Measuring inflation. *BIS Paper*, 89b (2016)

[16] Shostak, F.: Commodity Prices and Inflation: What is the Connection?. Dostupné na internete (1.11.2018): https://bit.ly/2CW3EEz

[17] Somorčík, J.: Prednášky z Ekonometrie, FMFI UK, Bratislava, 2018

[18] Wieringen, W. N.: Lecture notes on ridge regression. Dostupné na internete (20.4.2019): http://bit.ly/2VBHgM8

[19] Wilimitis, D.: The Kernel Trick in Support Vector Classification. Dostupné na internete (20.4.2019): http://bit.ly/2YgQ7R3

[20] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67 (2) (2005), 301-320

Príloha A

Tabuľka A1: Produkty s príslušnými váhami vybraté LASSO regresiou.

Názov tovaru	$\hat{\beta}_{lasso}$
Hruška abate	4.920
Tesco Mandle jádra pražená v hořké čokoládě a bílé čokoládě 100g	2.436
Tesco Tenké bramborové hranolky 750g	4.527
Tesco Bramborové hranolky předsmažené hluboce zmrazené 2kg	5.662
Svoboda Knedlík exclusive 500g	18.263
Tesco Vitakids Cherry rajčata datlová 250g	0.978
Tesco Paprika červená chilli 50g	3.278
Paprika kapie 500g	1.980
Okurka hadovka	0.412
Křen praný	6.481
Cibule červená 3 ks	1.586
České Houby Hlíva ústřičná volná	0.480
Hollandia Selský jogurt bílý s kulturou BiFi 500g	24.775

Názov tovaru	$\hat{\beta}_{lasso}$
Choceňská Mlékárna Choceňský smetanový jogurt bílý 150g	2.168
Danone Activia Bílá 4 x 120g	1.051
Mlékárna Kunín Termix s příchutí vanilka 90g	0.017
Danone Activia Jogurt s bifidokulturou jahoda/meruňka 4 x 120g	0.002
Mlékárna Valašské Meziříčí Kefírové mléko nízkotučné jahodové 450g	6.479
Tesco Kefírové mléko 500ml	1.344
Tesco Kefírové mléko jahodové 450g	3.821
Tatra Farmářské máslo 200g	0.448
Tesco Máslo 250g	8.051
Moravia Bylinkové máslo 125g	8.736
Camargue Máslo solené s krystalky mořské soli 250g	0.001

Názov tovaru	$\hat{\beta}_{lasso}$
Choceňská Mlékárna Choceňské tradiční pomazánkové neochucené 200g	0.393
Hochland Almette Nadýchaný tvarohový sýr 150g	7.195
Krémový sýr žervé 90g	6.865
Zálesák Klasik pikantní se zeleninou jemný smetanový sýr 125g	0.015
Gervais Natur Sýr termizovaný 200g	9.270
Želetava Nový smetanový se smetanou 3 ks 150g	0.011
Arla Čerstvý smetanový sýr 150g	0.045
A.W. Olomoucké tvarůžky malé 100g	0.918
Tesco Niva sýr s plísní uvnitř hmoty 100g	0.886
Tesco Finest Bleu d'Auvergne sýr plnotučný poloměkký 150g	0.209
Jaroměřická Mlékárna Jaroměřický měkký tvaroh 250g	0.064
Jaroměřická Mlékárna Jaroměřický tvaroh tvrdý na strouhání 200 g	21.166
Kostelecké Uzeniny Debrecínská pečeně	0.016
Kostelecké Uzeniny Javořické párky	0.857

Názov tovaru	$\hat{\beta}_{lasso}$
Pejskar Vysočina	0.836
Tesco Bramborový salát 500g	0.663
Tesco Vlašský salát 500g	0.313
Gastro Holandský salát 140g	0.338
Melites Listové těsto 500g	0.989
Vepřová plec bez kosti	0.279
Vepřové koleno s kostí	0.368
Vepřová krkovice s kostí volná	3.335
Hovězí krk podplecí volný	0.111
České kuřecí prsní řízky volné	0.125
Kuřecí játra volná	8.175
Emco Plněné polštářky s čokoládovou příchutí 175g	0.994
Milka Waffelini 31g	0.961
Dr. Ensa Mandle v mléčné čokoládě a skořici 80g	0.245
Dr. Ensa Kešu v mléčné čokoládě 80g	2.049
Figaro Tatiana Bonbony z mléčné čokolády s lískooříškovou náplní 194g	0.134
Figaro Karamelky z mléčné čokolády 121g	0.893
Milka Milkinis tyčinky z mléčné čokolády s mléčnou náplní 8 ks 87,5g	0.001
Skittles Fruits 38g	2.314
Casali Čokoládová cukrovinka s alkoholovou náplní 100g	0.541

Názov tovaru	$\hat{\beta}_{lasso}$
Kinder Happy Hippo Oplatka s mléčnou a lískooříškovou náplní 20,7g	3.729
Chupa Chups Drops s jahodovou příchutí 15g	5.658
Skittles Crazy sours 38g	0.013
Skittles Wild berry 38g	0.017
Maoam Žvýkací cukrovinka 110g	1.897
Medovník Original Classic 1600g	0.365
Opavia Bebe Dobré ráno naměkko s kousky čokolády 50g	3.614
Opavia Bebe Dobré ráno kakaové cereální sušenky s čokoládovými pečičkami 8 x 50g	0.003
Opavia BeBe Dobré ráno cereální sušenky s jogurtovou náplní 50,6g	3.468
Hamé Sweet Extra Jam Jahoda, méně sladký 790g	4.215
Lagris Čočka červená 500g	0.156
Lagris Čočka 500g	0.003
Vitana Parboiled rýže 4 x 100g	2.028
Vitana Rýže dlouhozrnná 4 x 100g	0.020
Barilla Spaghetti 500g	0.665
Zátkovy Vaječné Těstoviny Torti 500g	0.005
Barilla Fusilli 500g	0.110

Názov tovaru	$\hat{\beta}_{lasso}$
Barilla Penne Rigate 500g	0.001
Druid Babiččiny Vaječné fleky 330g	0.009
Adriana Spaghetti těstoviny semolinové sušené 500g	0.470
efko Kysané zelí 500g	0.310
Znojmia Kysané zelí 320g	0.699
Znojemská okurka 660g	0.204
efko Červené zelí 500g	0.201
Viva Carne Utopenci ve sladkokyselém nálevu 670g	0.527
Kostelecké Uzeniny Utopenci 680g	0.322
Kostelecké Uzeniny Pikantní utopenci 380g	0.182
Dr.Rashid Cukr třtinový 500g	0.860
Solné Mlýny Sůl jedlá mořská s jodem 1kg	6.813
Tesco Mouka hrubá 1kg	3.270
Palma Heliol Slunečnicový olej 1l	1.628
Mokate Tigo Instantní směs pro přípravu kakaového nápoje 150g	0.855
Jacobs Aroma Standard káva pražená mletá 250g	0.001
Jihlavanka Extra special pražená mletá káva 150g	0.092
Tesco Cereal & Coffee instantní směs 200g	0.504

Názov tovaru	$\hat{\beta}_{lasso}$
Aquila Aqualinea Pramenitá voda jemně perlivá 1,5l	0.496
Mirinda Pomeranč 2,25l	1.878
Vinea 1,5l	1.763
Mirinda Pomeranč 1,5l	0.664
Zlatopramen 11 pivo ležák světlý 1,5l	0.818
Habánské Sklepy Sauvignon suché bílé víno 0,75l	0.301
Víno Mikulov Svatý Urban bílé víno polosladké 0,75l	0.078
Rybex Treska ala losos drcená 120g	1.142
Houska staročeská sůl-kmín 70g	2.576
Koláč tvarohový 65g	3.507
Makovka 88g	0.815
Olma Dobré Mléko Selské 1l	0.109
Mlékárna Kunín Selské mléko čerstvé 1l	0.005
Mlékárna Kunín Čerstvé mléko polotučné 1l	0.008
Bohemilk Condé Zahuštěné plnotučné mléko 410g	2.714
Tesco Vejce M 30 ks	0.230
Hamé Sweet Extra Jam Meruňka, méně sladký 790g	3.183
Lay's Stix Smažené bramborové lupínky s příchutí kečupu 70g	0.382

Názov tovaru	$\hat{\beta}_{lasso}$
Binárna premenná Vianoce	7.393
Binárna premenná Veľká noc	21.759
Binárna premenná Black friday	-2.654

Príloha B

Tabuľka B1: Členenie spotrebného koša podľa ČSÚ.

ECOICOP	**Názov**	**Váha v ‰**
E01	**POTRAVINY A NEALKOHOLICKÉ NÁPOJE**	**177,627679**
E01.1	Potraviny	160,326214
E01.11	Pekárenské výrobky; obiloviny	28,313605
E01.12	Maso	42,134077
E01.13	Ryby a mořské plody	4,443854
E01.14	Mléko, sýry a vejce	31,614573
E01.15	Oleje a tuky	7,605283
E01.16	Ovoce	12,973513
E01.17	Zelenina	14,436646
E01.18	Cukr, marmeláda, med, čokoláda a cukrovinky	12,061109
E01.19	Potravinářské výrobky jinde neuvedené	6,743554
E01.2	Nealkoholické nápoje	17,301465
E01.21	Káva, čaj a kakao	6,483030
E01.22	Minerální vody, nealkoholické nápoje, ovocné a zeleninové šťávy	10,818435
E02	**ALKOHOLICKÉ NÁPOJE, TABÁK**	**92,144617**
E02.1	Alkoholické nápoje	41,271880
E02.11	Lihoviny	16,140281
E02.12	Vína	11,426417
E02.13	Piva	13,705182
E02.2	Tabák	50,872737
E02.20	Tabák	50,872737

Tabuľka B2: Prevod medzi kategóriami na stránke itesco.cz a ECOICOP kategorizáciou.

Kategória podľa itesco.cz	**Kategória podľa ECOICOP**
100% džusy	Minerální vody, nealkoholické nápoje, ovocné a zeleninové šťávy
Americké brambory a krokety	Zelenina
Balené ryby	Ryby a mořské plody
Balené saláty	Zelenina
Banány a exotické ovoce	Ovoce
Bílé jogurty	Mléko, sýry a vejce
Bonboniéry	Cukr, marmeláda, med, čokoláda a cukrovinky
Bonbóny a lízátka	Cukr, marmeláda, med, čokoláda a cukrovinky
Brambory	Zelenina
Cereálie a kukuřičné lupínky	Pekárenské výrobky; obiloviny
Citrusy	Ovoce
Cukr	Cukr, marmeláda, med, čokoláda a cukrovinky
Černé čaje	Káva, čaj a kakao
Čerstvé mléko	Mléko, sýry a vejce
Česnek a cibule	Zelenina
Čokoládové tyčinky	Cukr, marmeláda, med, čokoláda a cukrovinky
Dorty	Pekárenské výrobky; obiloviny
Droždí	Potravinářské výrobky jinde neuvedené
Drůbeží maso	Maso
Džemy a marmelády	Cukr, marmeláda, med, čokoláda a cukrovinky
Hotová těsta	Pekárenské výrobky; obiloviny
Hotové přílohy a knedlíky	Pekárenské výrobky; obiloviny
Houby	Zelenina
Hovězí maso	Maso
Chipsy	Zelenina
Jablka a hrušky	Ovoce
Jahody a lesní plody	Ovoce

Kategória podľa itesco.cz	**Kategória podľa ECOICOP**
Jemně perlivé	Minerální vody, nealkoholické nápoje, ovocné a zeleninové šťávy
Jogurtové a mléčné nápoje	Mléko, sýry a vejce
Kakao a Granko	Káva, čaj a kakao
Knedlíky v prášku	Pekárenské výrobky; obiloviny
Kondenzované mléko	Mléko, sýry a vejce
Konzervovaná a nakládaná zelenina	Zelenina
Konzervované a sterilizované ovoce	Ovoce
Konzervované maso	Maso
Kořenová zelenina	Zelenina
Košťálová zelenina	Zelenina
Krekry	Pekárenské výrobky; obiloviny
Ležák pivo	Piva
Luštěniny	Zelenina
Máslo	Oleje a tuky
Medy	Cukr, marmeláda, med, čokoláda a cukrovinky
Mletá káva	Káva, čaj a kakao
Mouka	Pekárenské výrobky; obiloviny
Mražené hranolky	Zelenina
Náhražky kávy	Káva, čaj a kakao
Ochucené jogurty	Mléko, sýry a vejce
Okurky a saláty	Zelenina
Ořechy	Ovoce
Ostatní příchutě	Minerální vody, nealkoholické nápoje, ovocné a zeleninové šťávy
Ovocné destiláty	Lihoviny
Peckoviny	Ovoce
Pizza	Pekárenské výrobky; obiloviny
Plísňové a zrající sýry	Mléko, sýry a vejce
Polévky	Potravinářské výrobky jinde neuvedené
Pomazánky	Mléko, sýry a vejce
Pultové uzeniny	Maso

Kategória podľa itesco.cz	**Kategória podľa ECOICOP**
Rajčata a papriky	Zelenina
Rostlinné oleje	Oleje a tuky
Rybí speciality	Ryby a mořské plody
Rýže	Pekárenské výrobky; obiloviny
Sirupy	Minerální vody, nealkoholické nápoje, ovocné a zeleninové šťávy
Sladké pečivo volné	Pekárenské výrobky; obiloviny
Sůl	Potravinářské výrobky jinde neuvedené
Sušenky a oplatky	Pekárenské výrobky; obiloviny
Světlé pečivo	Pekárenské výrobky; obiloviny
Tabulkové čokolády	Cukr, marmeláda, med, čokoláda a cukrovinky
Tavené a termizované sýry	Mléko, sýry a vejce
Tekuté koření	Potravinářské výrobky jinde neuvedené
Těstoviny	Pekárenské výrobky; obiloviny
Tuzemská bílá vína	Vína
Tvarohy	Mléko, sýry a vejce
Vejce	Mléko, sýry a vejce
Vepřové maso	Maso
Volný světlý chléb	Pekárenské výrobky; obiloviny
Cigarety	Tabák