KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Imputace chybějících hodnot v rozsáhlých datových souborech

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2013

Vypracovala:
Bc. Markéta Nárožná
AME, II. ročník

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vytvořila samostatně pod vedením RNDr. Karla Hrona, Ph.D., a že jsem v seznamu literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci, dne 3. dubna 2013

Poděkování

Děkuji vedoucímu diplomové práce, RNDr. Karlu Hronovi, Ph.D., za odborné vedení a čas, který mi věnoval. Dále bych ráda poděkovala konzultantům z Českého statistického úřadu, RNDr. Jaromíru Kalmusovi a Ing. Štěpánu Tourkovi, za cenné rady, jež mi poskytli.

Obsah

Úvod

Téma diplomové práce Imputace chybějících hodnot v rozsáhlých datových souborech jsem si zvolila z několika důvodů. Prvním a také nejdůležitějším důvodem byly jednoznačně mé dosavadní pracovní zkušenosti s výběrovými šetřeními v sociální oblasti. Každý, kdo někdy pracoval s reálnými daty, ví, že jsou chybějící hodnoty nedílnou součástí téměř všech výzkumů. Situace je o mnoho obtížnější, pokud hraje ve výzkumu významnou roli lidský faktor. Během své praxe jsem se na vlastní kůži setkala s neochotou respondentů, kteří často odpověď neznají, nepamatují si ji, a nebo odpovědět jednoduše nechtějí. Při výběru tématu práce mne samozřejmě zaujala avizovaná spolupráce s Českým statistickým úřadem (ČSÚ), zejména potom možnost aplikovat získané teoretické znalosti na reálná data. Při seznamování se s tématem diplomové práce jsem narazila na názor P. D. Alissona, uznávaného odborníka na tuto problematiku, který o chybějících hodnotách hovoří jako o „malém špinavém tajemství statistiky“ [1]. Je všeobecně známo, že ženy mají pro tajemství slabost. Ani já nejsem výjimkou, proto jsem se rozhodla tomuto tajemství „přijít na kloub“.

Cílem diplomové práce je zpracovat ucelený přehled metod imputace, dále potom uvedené metody aplikovat na reálná data z šetření Životní podmínky poskytnutá Českým statistickým úřadem.

Diplomová práce je rozčleněna do šesti kapitol. První kapitola slouží k seznámení s šetřením Životní podmínky, s jeho průběhem i specifiky, jejichž znalost je pro další výklad o chybějících hodnotách a metodách imputace nezbytná.

Následující kapitola pojednává o chybějících hodnotách. V úvodu jsou pro připomenutí stručně zmíněny základní typy proměnných, se kterými se můžeme setkat v rozsáhlých datových souborech, dále jsou zde definovány mechanismy vzniku chybějících hodnot, které jsou pro lepší pochopení ilustrovány jednoduchými příklady z šetření Životní podmínky.

Třetí kapitola je věnována vizualizaci, jakožto modernímu přístupu k rozeznání mechanismů vzniku chybějících hodnot a také ke zvolení správné metody imputace.

V úvodu čtvrté kapitoly, která nese název Imputace chybějících hodnot, najdeme klasifikaci metod imputace, dále jsou tu ve dvou podkapitolách popsány vybrané metody prosté imputace - nejprve metody jednorozměrné, poté vícerozměrné.

Pátá kapitola podává stručný přehled o výskytu chybějících hodnot v šetření Životní podmínky z roku 2010, z něhož pochází data použitá v praktické části diplomové práce. Najdeme zde mimo jiné informace o počtu celkem vyšetřených bytů, domácností i osob.

Poslední kapitola je věnována aplikaci získaných teoretických znalostí na reálná data poskytnutá ČSÚ. Celá kapitola je přitom orientována na knihovnu VIM, jakožto na užitečný nástroj určený k vizualizaci a imputaci chybějících hodnot, který je volně dostupný v statistickém softwaru R. Úvod kapitoly slouží k popisu datových souborů, následuje stručné seznámení s knihovnou VIM. Dále jsou pomocí této knihovny provedeny vizualizace i imputace chybějících údajů v datových souborech týkajících se šetření domácnosti. V závěru kapitoly najdeme vyhodnocení a srovnání úspěšnosti použitých metod imputace.

1 Šetření Životní podmínky

V praktické části diplomové práce budeme pracovat s daty z šetření Životní podmínky, proto nejprve stručně shrneme charakteristiky a průběh tohoto šetření. Tato kapitola byla vytvořena převážně s využitím zdrojů [12] a [27].

K vykonávání šetření Životní podmínky se Česká republika zavázala vstupem do Evropské unie v roce 2004. Český statistický úřad provádí toto šetření každé jaro od roku 2005. Hlavním cílem šetření Životní podmínky je dlouhodobě získávat srovnatelná data o ekonomické a sociální situaci domácností. Jelikož obdobná šetření probíhají každý rok také ve všech zemích Evropské unie, dále na Islandu, v Norsku, Chorvatsku, Švýcarsku a Turecku (pod názvem European Union - Statistics on Income and Living Conditions, zkráceně EU-SILC), slouží výsledky tohoto šetření ke srovnání životních podmínek ve výše zmíněných zemích. Toto výběrové šetření je považováno za velmi důležité, jelikož životní podmínky domácností a jednotlivců reflektují vliv přijatých zákonů a sociální politiky.

Výběrovou jednotkou šetření Životní podmínky je byt. Byty jsou voleny náhodným dvoustupňovým výběrem, přičemž jsou nejdříve vybrány sčítací obvody, což jsou nejmenší existující územní jednotky ČR, a poté je v každém z vybraných obvodů zvoleno deset bytů. Tyto byty jsou jednoznačně identifikovány adresou a číslem popisným, případně číslem bytu v domě. Do výběru jsou zahrnuty všechny kraje tak, aby bylo pokryto celé území ČR. Platí přitom, že rozsah výběru za kraj je přímo úměrný jeho velikosti.

Šetření vykonávají školení tazatelé, jenž navštěvují vybrané byty a zjišťují údaje za všechny osoby, které mají v době šetření na dané adrese obvyklé bydliště. To znamená, že trvalé bydliště osob není pro šetření Životní podmínky rozhodující. Šetření probíhá pomocí sady čtyř provázaných dotazníků: zjišťují se informace za byt, za hospodařící domácnost (hospodařící domácnost tvoří osoby, které se ve zvoleném bytě podílí na hrazení základních a provozních výdajů domácnosti), ale také údaje za jednotlivé členy domácnosti, kteří na konci minulého kalendářního roku dosáhli alespoň 16 let věku. Aby bylo možné zaměřit se na aktuální problematiku související s životními podmínkami, obsahuje sada dotazníků

tzv. modul, neboli část, která se každý rok obměňuje (např. v roce 2010 se modul týkal správy financí v domácnosti a rozhodování o nich). V dotaznících (viz příloha) se objevují otázky zaměřené na ekonomickou aktivitu respondentů, jejich bydlení a náklady s ním spojené, dále na příjmy domácnosti i všech jejích členů, včetně transferů peněz mezi jednotlivými domácnostmi. Samotné šetření probíhá ve dvou módech - prvním je tzv. PAPI (Paper And Pencil Interview), to znamená, že vybrané byty navštěvuje tazatel s papírovými dotazníky, druhý mód nazýváme CAPI (Computer Assisted Personal Interview), zde tazatel přichází do bytu s počítačem, do kterého zapisuje získané informace. Lze říci, že oba zmíněné módy mají v současné době na sběru dat zhruba stejný podíl.

Šetření Životní podmínky má podobu tzv. rotačního panelu. Vybrané byty jsou navštěvovány opakovaně v ročním intervalu po dobu čtyř let, přičemž se každý rok asi jedna čtvrtina obmění. Často se stává, že se v průběhu čtyřletého cyklu přestěhuje celá hospodařící domácnost či jednotliví členové. Aby se soubor neustále nezmenšoval, jsou tyto domácnosti (případně osoby) dohledávány na jejich nových adresách.

Účast respondentů v šetření je zcela dobrovolná. Z tohoto důvodu se může stát, že se respondent odmítne aktivně zúčastnit šetření - nejčastěji se tak děje v první vlně šetření, méně často i v dalších vlnách (např. přijde jiný pověřený tazatel nebo za domácnost odpovídá jiná osoba než ve vlnách předchozích). Domácnosti, které se šetření odmítnou zúčastnit, se opakovaně nenavštěvují.

V průběhu celého šetření je kladen zřetel na anonymitu. Získaná data jsou chráněna zákonem o státní statistické službě č. 89/1995 Sb. a také zákonem o ochraně osobních údajů č. 101/2000 Sb. Všechny osoby, které se nějakým způsobem na šetření podílí, jsou vázány mlčenlivostí.

Výsledky šetření Životní podmínky jsou každoročně publikovány ve výtisku nazvaném Příjmy a životní podmínky domácností ČR. Veškeré zveřejňované údaje jsou odhady vytvořené na základě dat získaných z šetření, která byla následně přepočítána na celou populaci.

2 Chybějící hodnoty

Chybějící hodnoty, neboli missings (z anglického missing values), mohou v datových souborech vznikat z různých důvodů. Ve výběrových šetření stojí za jejich vznikem působení lidského faktoru. Chybějícími údaji se musíme zabývat především z toho důvodu, že drtivá většina používaných statistických softwarů předpokládá úplnou obdélníkovou datovou matici. Pokud tento předpoklad není splněn, jsou zpravidla jednotky, jimž některé hodnoty chybí, ignorovány. Tento přístup, nazývaný analýza kompletních případů (complete-case analysis nebo listwise deletion), se obecně považuje za nevhodný. Je s ním často spojena citelná ztráta informace. Kromě toho jsou chybějící údaje mnohdy typické pro určité skupiny obyvatelstva (např. otázky týkající se příjmu neradi zodpovídají podnikatelé). Jelikož se výsledky výběrového šetření přepočítávají pro celou populaci, získáme tímto vynecháním zkreslené výsledky.

Při tvorbě této kapitoly byly využity zejména zdroje [3], [8], [18] a [24].

2.1 Úvod do problematiky

Je všeobecně známo, že nejlepším způsobem jak zacházet s chybějícími daty, je předcházet jejich vzniku. To je však v případě výběrových šetření, v nichž působí lidský faktor, velmi obtížné. Naším cílem ve výběrových šetřeních je tedy mimo jiné napozorovat u každé výběrové jednotky veškeré sledované veličiny. Zde však často dochází k problémům - jak už bylo řečeno, respondenti mnohdy odpovědět vůbec nechtějí, odpověď neznají, nebo si ji nepamatují. Musíme si však uvědomit, že ne vždy vznikají chybějící hodnoty vinou respondenta. V některých případech se chyba vyskytuje na straně tazatele - může například nesrozumitelně položit otázku, nebo může při dotazování otázku úplně přeskočit.

Rozlišujeme tři hlavní přístupy k analýze chybějících hodnot:

- vyloučení všech neúplných záznamů z dalšího zpracování,
- nahrazení (imputace) chybějících údajů a následné zpracování již kompletních dat,

- zpracování neúplných dat speciálními metodami.

V případě, že pro danou výběrovou jednotku schází zjišťovaná veličina, hovoříme o non-response. Pokud u jednotky chybí veškeré zjišťované veličiny, mluvíme o tzv. jednotkové (unit) non-response. Za typický příklad považujeme situaci, kdy se respondent odmítne zúčastnit šetření. Druhým typem je tzv. položková (item) non-response, kdy postrádáme jen některé ze zjišťovaných veličin. O položkové non-response hovoříme např. tehdy, pokud respondent nechce sdělit výši svých příjmů.

Speciálně školení tazatelé postupují tak, aby od respondentů získali alespoň nějaké informace (např. pokud respondent nechce prozradit výši svých příjmů, snaží se tazatel zjistit alespoň interval, v jakém se příjmy respondenta pohybují). Pokud se ani toto nepodaří, přichází na řadu tzv. imputace chybějících hodnot.

V analýze chybějících údajů používáme obvyklé značení. Máme dánu datovou matici $\mathbf{Y} = (y_{ij})$ o rozměru $(n \times p)$ - řádky datové matice reprezentují statistické jednotky, sloupce představují proměnné zjišťované pro každou statistickou jednotku. Proměnné v datové matici jsou téměř vždy reálná čísla, která zastupují hodnoty různých typů proměnných. Proto nyní krátce připomeneme základní typy proměnných. Rozlišujeme:

- **kvantitativní (numerické) proměnné** - jsou vyjádřeny číselně, lze je měřit a dělíme je na:
  - **diskrétní proměnné** - mají konečné nebo spočetné množství variant (např. věk v letech, počet dětí v domácnosti),
  - **spojité proměnné** - mají nespočetné množství možných realizací z množiny reálných čísel nebo z jejich libovolné podmnožiny (např. průměrný měsíční příjem v Kč),
- **kvalitativní (kategoriální) proměnné** - jsou vyjádřeny slovně, nelze je měřit, můžeme je pouze řadit do tříd (kategorií). Podle počtu kategorií, kterých může proměnná nabývat, rozlišujeme:

- **alternativní (binární) proměnné** - nabývají pouze dvou hodnot (např. pohlaví),
- **množné proměnné** - nabývají více hodnot (např. rodinný stav, zaměstnání).

Podle vztahu mezi jednotlivými kategoriemi lze kvalitativní proměnné rozdělit na:

- **nominální proměnné** - nabývají různých, avšak rovnocenných hodnot, které nelze smysluplně uspořádat (např. pohlaví),
- **ordinální proměnné** - nabývají hodnot, které lze porovnávat a seřadit (např. nejvyšší stupeň dosaženého vzdělání).

Ve výběrových šetřeních se mohou vyskytovat tzv. semi-spojité (polo-spojité) proměnné. Jako semi-spojité označujeme proměnné, kde distribuce jedné části dat je spojitá, zatímco další část dat obsahuje určitý podíl stejných hodnot (nejčastěji nul). Jako příklad z šetření Životní podmínky uveďme příjmové složky.

2.2 Mechanismy vzniku chybějících hodnot

Nahrazování chybějících hodnot, jímž se budu ve své práci zabývat, bývá často problematické. Způsob nahrazování by měl vždy vycházet z mechanismů vedoucích ke vzniku chybějících hodnot. Zajímá nás zejména, zda skutečnost, že je hodnota chybějící, nějakým způsobem souvisí s hodnotami daných proměnných v datovém souboru.

Nechť $\mathbf{Y} = (y_{ij})$ značí úplnou datovou matici a dále nechť $\mathbf{M} = (m_{ij})$ je matice indikující chybějící hodnoty, kde

$$m_{ij} = \begin{cases} 1 & y_{ij} \text{ chybí,} \\ 0 & y_{ij} \text{ je přítomna.} \end{cases}$$

Mechanismus vedoucí ke vzniku chybějících hodnot je dán podmíněným rozdělením pravděpodobnosti matice $\mathbf{M}$ za podmínky $\mathbf{Y}$, tj.

$$f(\mathbf{M}\,|\mathbf{Y}, \boldsymbol{\phi})\,,$$

kde $\boldsymbol{\phi}$ značí neznámé parametry.

Existuje celá řada důvodů, proč chybějící hodnoty vznikají. Podle příčiny jejich vzniku rozlišujeme následující druhy chybějících hodnot. Autorem níže zmíněné klasifikace je Donald B. Rubin, poprvé ji publikoval v roce 1976.

2.2.1 Missing Completely At Random - MCAR

Česky říkáme, že data chybí zcela náhodně. Jestliže hovoříme o Missing Completely At Random (MCAR), máme na mysli, že chybějící hodnoty nezávisí na hodnotách proměnných v $\mathbf{Y}$, ať pozorovaných či chybějících. Jinak řečeno - pravděpodobnost, že hodnota y_{ij} je chybějící, nesouvisí s hodnotou y_{ij}, ani s hodnotami ostatních proměnných. Pomocí podmíněného rozdělení pravděpodobnosti lze vyjádřit MCAR takto

$$f(\mathbf{M}\,|\mathbf{Y}, \boldsymbol{\phi}) = f(\mathbf{M}\,|\boldsymbol{\phi})\,,$$

pro každou složku $\mathbf{Y}$ a $\boldsymbol{\phi}$.

Příkladem z šetření Životní podmínky může být situace, kdy se během vyplňování dotazníku výrazně zhorší zdravotní stav respondenta a ten není z tohoto důvodu schopen dotazování dokončit. Jedná se tedy o situaci, která vůbec nesouvisí s daným šetřením. Dalším příkladem MCAR je, jestliže tazatel v průběhu dotazování neúmyslně zapomene položit jednu otázku. Také v tomto případě platí, že okolnosti vedoucí k výskytu chybějící hodnoty vznikly zcela náhodně. Tato situace však může nastat pouze tehdy, pokud šetření probíhá v módu PAPI. Použitý počítač by takto vzniklou chybějící hodnotu okamžitě odhalil. Chybějící hodnota má také mechanismus MCAR, pokud za vyšetřovanou osobu odpovídá jiný člen domácnosti, který nezná odpověď na některou otázku.

Jak již bylo zmíněno výše, analýza kompletních případů se nepovažuje za vhodný přístup k chybějícím údajům. Na druhou stranu, v případě MCAR dat může vést vynechání nekompletních pozorování k nevychýleným odhadům.

V literatuře se můžeme setkat také s názvem plánované chybějící údaje (Planned Missing Data). Předpoklady na mechanismus MCAR jsou velmi přísné, v praxi se chybějící data s MCAR vyskytují poměrně málo.

2.2.2 Missing At Random - MAR

Další kategorií je náhodné rozložení chybějících dat neboli Missing At Random (MAR). V tomto případě rozdělíme složky datové matice $\mathbf{Y}$ na napozorované hodnoty, označíme $\mathbf{Y}_{obs}$, a chybějící hodnoty, $\mathbf{Y}_{miss}$. Mechanismus nazveme MAR, pokud jsou chybějící hodnoty závislé pouze na napozorovaných složkách $\mathbf{Y}_{obs}$ a nikoliv na složkách chybějících. Matematicky vyjádřeno

$$f(\mathbf{M}\,|\mathbf{Y}, \boldsymbol{\phi}) = f(\mathbf{M}\,|\mathbf{Y}_{obs}, \boldsymbol{\phi})\,,$$

pro všechny složky $\mathbf{Y}_{miss}$ a $\boldsymbol{\phi}$.

V praxi výběrových šetření platí nepsané pravidlo, že má tazatel větší úspěšnost v otázkách týkajících se příjmu, pokud na tyto otázky odpovídá žena. Pravděpodobnost, že je hodnota týkající se příjmu chybějící, tedy závisí na pohlaví dotazovaného, nikoliv na výši jeho příjmu. Hodnoty příjmu považujeme za MAR. Pokud uvažujeme, že ženy sdělí s větší pravděpodobností svůj příjem, jelikož jejich příjem bývá obecně nižší, nejedná se o MAR, nýbrž o NMAR.

O chybějících datech, které mají rozložení MAR, někdy mluvíme jako o ignorovatelných chybějících datech. To znamená, že model dokáže vysvětlit mechanismus chybějících hodnot a chybějící hodnoty mohou být ignorovány, jestliže s nimi model počítá. Rozhodně nelze říci, že lze data kompletně ignorovat a použít analýzu kompletních případů.

2.2.3 Not Missing At Random - NMAR

Jestliže není mechanismus chybějících údajů typu MCAR nebo MAR, klasifikujeme jej jako nenáhodné rozložení chybějících dat. Not Missing At Random nám říká, že rozdělení matice $\mathbf{M}$ závisí na chybějících hodnotách v datové matici $\mathbf{Y}$, tj.

$$f(\mathbf{M}\,|\mathbf{Y}, \boldsymbol{\phi}) = f(\mathbf{M}\,|(\mathbf{Y}_{obs}, \mathbf{Y}_{miss})\,, \boldsymbol{\phi}),$$

pro všechny $\boldsymbol{\phi}$.

Typickou ukázkou NMAR je již několikrát zmíněný příklad týkající se příjmů - čím je příjem respondenta vyšší, tím menší je pravděpodobnost, že svůj

příjem sdělí. Jinými slovy - NMAR nastává, pokud je pravděpodobnost výskytu chybějících hodnot ve spojitosti s hodnotami, které jsou nepřítomny. Při studiu některých zdrojů se můžeme setkat s názvem Missing Not At Random - MNAR.

Tento mechanismus se v praxi vyskytuje nejčastěji a bohužel je také nejproblematičtější.

3 Vizualizace nekompletních dat

Před samotným procesem imputace chybějících hodnot potřebujeme odhalit zákonitosti a získat znalosti o chybějících hodnotách v datovém souboru. Teprve na základě těchto informací můžeme stanovit vhodnou metodu imputace. V praxi bývá obtížné identifikovat jednotlivé mechanismy vzniku chybějících hodnot. Identifikace vyžaduje znalost chybějících hodnot samotných a také zkušenosti s daným šetřením. Situace je ještě komplikovanější, pokud se jedná o rozsáhlé datové soubory, jestliže se datové soubory skládají z různých typů proměnných, nebo pokud se v datovém souboru vyskytují odlehlá pozorování (outliers). Rozlišujeme dva hlavní přístupy k analýze nekompletních dat - statistické testy a vizualizační nástroje. Naši pozornost zaměříme na modernější a užívanější z přístupů, tedy na vizualizaci chybějících hodnot. Tato kapitola byla zpracována pomocí zdrojů [13] a [21].

Vizualizace nekompletních dat umožňuje současné zkoumání datového souboru a struktury chybějících hodnot. Vizualizační nástroje však zpravidla nejsou součástí statistických softwarů, zcela chybí v často užívaných softwarech jako jsou SAS, SPSS nebo STATA.

Ve statistickém softwaru R máme k dispozici knihovnu VIM (Visualization and Imputation of Missing values), která byla navržena tak, aby umožnila identifikovat mechanismy vzniku chybějících hodnot a také tyto chybějící hodnoty nahradit. Kromě toho můžeme podrobit vizuálnímu zkoumání, užitím různých grafických nástrojů, již doplněný datový soubor. Za velkou přednost této knihovny považujeme schopnost vytvářet vysoce kvalitní grafické výstupy.

Nyní se budeme věnovat jednotlivým grafickým nástrojům, které najdeme v knihovně VIM. Jelikož u některých grafů prozatím neexistují české ekvivalenty k jejich anglickým názvům, budeme používat tato anglická označení. Veškeré dále uvedené grafické nástroje lze aplikovat nejen na údaje chybějící, ale také na hodnoty imputované.

Agregační graf (Aggregation plot)

Agregační graf použijeme, pokud nás zajímá, jaké množství chybějících hodnot obsahují jednotlivé proměnné. Tento typ grafu navíc dokáže vykreslit kombinace chybějících, případně imputovaných hodnot v daných proměnných. Agregační graf je dobře čitelný pouze tehdy, když existuje velké množství různých kombinací s nízkou četností výskytu.

Histogram, sloupcový graf (Barplot), spinogram a spine plot

Histogram, jenž považujeme za nejrozšířenější grafický nástroj pro zobrazení hodnot jedné spojité proměnné, nachází své využití také v analýze chybějících dat. Skládá se ze sloupců shodné šířky, jejichž výška reprezentuje četnost sledované proměnné v daném intervalu hodnot. Pokud vykreslujeme histogram pro proměnnou s chybějícími hodnotami, je možné množství chybějících hodnot pro danou proměnnou zobrazit do zvláštního sloupce. Tento sloupec odpovídající nedostupným datům bývá od zbytku grafu oddělen malou mezerou. Dále je možné rozdělit jednotlivé sloupce podle informací o chybějících hodnotách v jiné proměnné. Obdobným způsobem lze zobrazit kategoriální proměnné pomocí sloupcového grafu.

Modifikaci histogramu představuje tzv. spinogram, v němž odráží relativní četnost sledované proměnné šířka sloupců. Obdobný graf můžeme vytvořit pro kategoriální proměnné, potom mluvíme o spine plotu. Stejně jako u histogramu a sloupcového grafu, lze podíl chybějících hodnot sledované proměnné zobrazit pomocí zvláštního sloupce. Na vertikální ose je opět možné vykreslit podíl chybějících a dostupných hodnot v jiné proměnné. Za velkou výhodu těchto grafů považujeme, že je možné srovnávat počty chybějících hodnot v různých sloupcích. Značné rozdíly v těchto poměrech ukazují na mechanismus vzniku MAR.

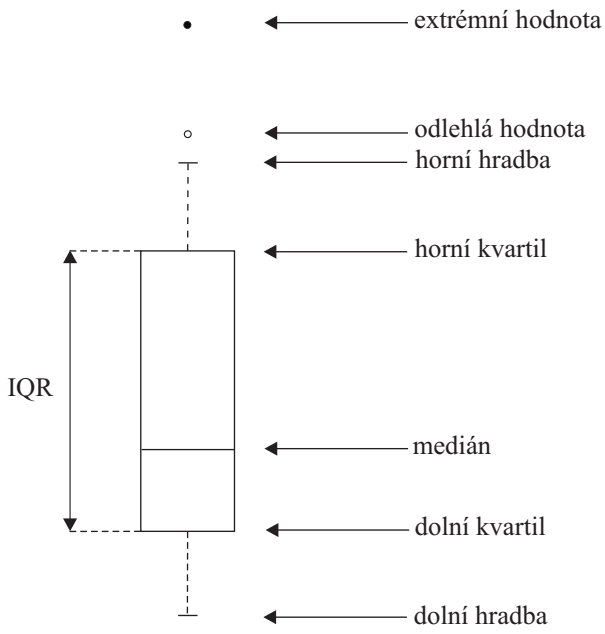
Boxplot a paralelní boxploty (Parallel boxplots)

Za další velmi významný grafický nástroj považujeme boxplot neboli krabicový graf. Funkce krabicového grafu je spíše popisná, umožňuje však posoudit variabilitu a symetrii dat, dále slouží k odhalení odlehlých a extrémních pozorování. Nejprve krátce připomeneme standardní boxplot (viz Obr. 1), ten se skládá

z obdélníku (neboli krabice) a vousů (neboli antén). Obdélník pojímá přibližně polovinu datového souboru, jeho dolní hrana je určena dolním kvartilem, horní hranu vymezuje horní kvartil. Výška krabičky je tedy dána tzv. interkvartilovým rozpětím (Interquartile Range), které získáme ze vztahu

$$IQR = x_{0,25} - x_{0,75},$$

kde $x_{0,25}$ značí dolní kvartil a $x_{0,75}$ horní kvartil. Krabici dělí na dvě části medián, jenž reprezentuje zešikmení. Vousy, které vystupují z krabice, jsou zakončeny tzv. vnitřními hradbami. Dolní hradbu získáme ze vztahu $x_{0,25} - 1,5 \cdot IQR$, horní hradbu potom podobně $x_{0,75} + 1,5 \cdot IQR$. Hodnoty, jež nespadají do intervalu ohraničeného hradbami, považujeme za podezřelé a jsou znázorněny kroužky. Někdy lze v grafu najít i tzv. extrémní hodnoty, které se nachází pod hranicí $x_{0,25} - 3 \cdot IQR$ a nad hranicí $x_{0,75} - 3 \cdot IQR$.



Obr. 1: Boxplot

Při zpracování datových souborů obsahujících chybějící data používáme tzv. paralelní boxploty. Tento typ grafu je užitečné prozkoumat, pokud existuje podezření, že jedna spojitá proměnná vysvětluje rozložení chybějících hodnot v ostat-

ních proměnných.

Bodový graf (Scatter plot) a marginplot

Bodový graf obecně pokládáme za jeden z nejjednodušších a tím pádem i nejužívanějších grafických nástrojů. Kromě tradičního bodového grafu, jenž ukazuje vztah mezi dvěma číselnými proměnnými zobrazením bodů do dvourozměrného grafu, lze tímto způsobem také zobrazit informace o chybějících hodnotách. Jestliže aplikujeme bodový graf na dvě proměnné, z nichž obsahuje chybějící hodnoty pouze jedna, vykreslí se do bodového grafu navíc elipsy spolehlivosti. Elipsy spolehlivosti jsou definovány jako množiny p-rozměrných bodů, jejichž Mahalanobisova vzdálenost od středu (typicky výběrového průměru) se rovná odmocnině z určitého kvantilu χ^2 rozdělení o dvou stupních volnosti. Výhodou takového zobrazení elipsy spolehlivosti je snadnější identifikace odlehlých pozorování - za odlehlá považujeme ta pozorování, která leží vně větší elipsy spolehlivosti. K vykreslení chybějících hodnot v jednotlivých proměnných lze také využít jednorozměrný bodový graf podél souřadnicových os. Knihovna VIM dále umožňuje zobrazení boxplotu pro chybějící i dostupná data. Tento typ grafu budeme dále označovat jako marginplot, jedná se totiž o bodový graf s doplňujícími informaci na okrajích. Pro oba zmíněné typy grafů jsou v knihovně VIM dostupné i jejich maticové modifikace - maticový bodový graf a maticový marginplot.

Dvourozměrný jitter plot (Bivariate jitter plot)

Tento typ grafu slouží k zobrazení počtu kombinací chybějících hodnot dvou proměnných. Graf je nejprve rozdělen do maximálně čtyř čtverců v závislosti na výskytu chybějících hodnot v jedné či obou proměnných. Počty chybějících případně imputovaných a dostupných hodnot jsou vyjádřeny číslem.

Maticový graf (Matrix plot)

Maticový graf vykresluje veškeré údaje z datové matice jako obdélníky. Proměnné jsou v prvé řadě transformovány na interval $[0, 1]$. Také maticový graf lze použít jak pro číselná, tak i pro kategoriální data. Dostupná data jsou zobrazena pomocí kontinuálního barevného schématu. Výpočet barev probíhá pomocí in-

terpolací, chybějící údaje znázorňujeme odlišnou barvou. Nejčastěji se dostupná data vykreslují v odstínech šedi, zatímco chybějící hodnoty kontrastní barvou. Množství informace obsažené v grafu závisí na typu proměnné. Uvažujme např. nominální proměnné, které nemají smysluplné uspořádání kategorií. Pro tyto proměnné dokážeme díky použitým barvám určit, v které kategorii se v ostatních proměnných vyskytují chybějící hodnoty nejhojněji, avšak řazení barev nám neposkytuje v tomto případě relevantní informaci.

Mozaikový graf (Mosaic plot)

Mozaikový graf umožňuje prozkoumat vztah mezi dvěma nebo více kategoriálními proměnnými. Jedná se tedy o grafickou reprezentaci mnohorozměrné kontingenční tabulky. Četnosti jednotlivých buněk kategoriální proměnné jsou zobrazovány na plochu proporcionálních obdélníků (tzv. dlažba). Při konstrukci mozaikového grafu je výchozí obdélník nejprve vertikálně rozdělen na místech odpovídajících četnostem jednotlivých kategoriích dané proměnné. Takto získané menší obdélníky jsou znovu rozčleněny, tentokrát horizontálně podle podmíněné četnosti druhé proměnné. Tímto způsobem lze pokračovat pro další proměnné. Dlaždice mohou být rozděleny podle počtu chybějících hodnot jiné proměnné. Tento typ grafu je užitečný při zkoumání rozdělení chybějících hodnot.

4 Imputace chybějících hodnot

Imputace představuje flexibilní metodu pro řešení problémů s chybějícími daty. Jedná se o postup, kterým jsou chybějící hodnoty jedné nebo více proměnných vyplněny náhradami (náhradními hodnotami). Jak uvidíme později v této kapitole, tyto náhrady mohou být získány různými způsoby. Úplnou datovou matici, kterou získáme použitím imputace, dále analyzujeme standardními statistickými metodami.

Základní literaturou pro tvorbu této kapitoly byly zdroje [1], [8], [9] a [11].

4.1 Klasifikace metod imputace

Existuje celá řada dělení metod imputace, toto dělení není vždy zcela jednoznačné. Nejčastěji rozlišujeme:

- **prostou imputaci (single imputation)**, kdy je každá chybějící hodnota nahrazena jednou, příslušnou metodou vygenerovanou, hodnotou. Tyto metody můžeme dále dělit na:
  - **metody jednorozměrné**,
  - **metody vícerozměrné**,
- **mnohonásobnou imputaci (multiple imputation)**, kdy každou chybějící položku nahradíme více hodnotami na základě jejich předpokládaného rozložení.

Mezi nejmodernější přístupy současnosti řadíme metody mnohonásobné imputace. Těmito metodami se vzhledem k omezenému rozsahu diplomové práce nebudeme zabývat. Zájemce o tuto problematiku odkazujeme na [2] či [26]. Naši pozornost zaměříme na vybrané metody prosté imputace, které nějakým způsobem souvisí s knihovnou VIM.

4.2 Jednorozměrné metody prosté imputace

Tato kapitola bude věnována jednorozměrným metodám prosté imputace. Jednorozměrnost těchto metod znamená, že nahrazujeme chybějící hodnoty pouze v rámci jedné proměnné. Mezi nejjednodušší řadíme tzv. subjektivní metody imputace. Do této skupiny metod patří deduktivní imputace nebo imputace založená na pravidlech, která se používá nejčastěji k opravení zjevné systematické chyby. Další skupinu tvoří metody, kdy imputujeme chybějící hodnotu ve sloupci příslušnou charakteristikou polohy - nejčastěji aritmetickým průměrem, mediánem či modální kategorií. Výběr charakteristiky závisí na typu proměnné, jejíž hodnotu nahrazujeme. Aritmetický průměr a medián je vhodný pro kvantitativní proměnné, modus zvolíme, pokud se jedná o kvalitativní proměnnou. Další variantou může být nahrazení chybějící hodnoty tzv. středem rozpětí (mid-range), který se vypočítá jako aritmetický průměr minimální a maximální hodnoty dané proměnné.

Při tvorbě této podkapitoly byly využity následující zdroje [4], [8], [17] a [23].

4.2.1 Deduktivní imputace

Aplikujeme-li deduktivní imputaci, nahrazujeme chybějící hodnotu na základě odvození logických vztahů mezi jednotlivými proměnnými (např. chybí jedna složka z příjmů, ale známe celkové příjmy, můžeme tedy chybějící složku dopočítat). Tuto metodu řadíme mezi nejjednodušší a také nejlevnější, provádí se v raných stádiích šetření.

4.2.2 Nahrazení průměrnou hodnotou

Jak napovídá název metody, v tomto případě nahrazujeme chybějící hodnoty y_{ij} hodnotou $\bar{y}_j^{(j)}$, kde $\bar{y}_j^{(j)}$ značí aritmetický průměr pozorovaných hodnot příslušné proměnné Y_j. Můžeme se setkat také s pojmenováním imputace nepodmíněnými průměry (unconditional mean imputation) [8].

S tímto přístupem se pojí hned několik problémů. Je zřejmé, že všechny chybějící hodnoty příslušné proměnné budeme nahrazovat stejnou hodnotou, kterou

navíc silně ovlivňují odlehlá pozorování. Lze ukázat, že aritmetický průměr pozorovaných hodnot a aritmetický průměr již doplněných hodnot se rovnají. Použitím této metody tedy nezískáme novou informaci, ale pouze zvýšíme velikost vzorku, což vede k podhodnocení chyb. Toto podhodnocení považujeme za přirozený důsledek imputace chybějících hodnot středem rozdělení (resp. jeho odhadem).

Tato metoda imputace poskytuje dobré výsledky, pokud jsou chybějící hodnoty MCAR. Přes uvedené problémy patří nahrazení průměrnou hodnotou, zejména v sociálních vědách, mezi poměrně oblíbené metody imputace. Její použití se však v dnešní době, kvůli výše zmíněným problémům, nedoporučuje.

4.3 Vícerozměrné metody prosté imputace

Při aplikaci vícerozměrných metod prosté imputace nahrazujeme chybějící hodnoty ve více proměnných současně. Do této skupiny patří celá řada metod, jmenujme např. hot-deck imputaci, cold-deck imputaci, regresní imputaci či metody založené na maximální věrohodnosti.

4.3.1 Regresní imputace

V prvé řadě se budeme věnovat zdokonalení imputace průměrnou hodnotou, tedy regresní imputaci. Při aplikaci tohoto přístupu je totiž imputovaná hodnota, na rozdíl od nahrazení průměrnou hodnotou, nějakým způsobem podmíněna dalšími informacemi, které o daném objektu (např. osobě) máme. Proto se můžeme v literatuře setkat s názvem imputace podmíněnými průměry (conditional mean imputation).

Tato podkapitola vychází ze zdrojů [8], [25] a [26]. Jelikož se jedná o doplnění pomocí metody lineární regrese, připomeneme na úvod základní poznatky z regresní analýzy, které byly přejaty z [7].

Regresní analýza slouží ke studiu závislostí v datovém souboru. Jejím cílem je objasnění vztahu náhodné veličiny Y, kterou nazýváme závisle proměnnou neboli vysvětlovanou proměnnou, na nezávislých neboli vysvětlujících proměnných $\mathbf{x} = (x_1, \ldots, x_p)$. Regresní funkce, jež určuje podmíněnou střední hodnotu veli-

činy Y za podmínky $x_1, \ldots, x_p$, má tvar

$$\mathrm{E}(Y\,|(x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

kde $\beta_0, \ldots, \beta_p$ jsou neznámé parametry.

Potom pro i-té pozorování platí

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde ϵ_i je náhodná chyba při i-tém pozorování.

Maticový zápis vypadá takto

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Přitom předpokládáme, že

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{o} \quad \text{a} \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,$$

kde $\mathbf{o}$ značí nulový sloupcový vektor a $\mathbf{I}_n$ jednotkovou matici řádu n.

Odhad vektorového parametru $\boldsymbol{\beta}$ pomocí metody nejmenších čtverců vypočteme, za předpokladu plné hodnosti matice $\mathbf{X}$, podle známého vztahu

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Regresní imputace se využívá k nahrazení položkové non-response, jestliže máme k dispozici pomocná data. V praxi se získávají odhady chybějících hodnot tak, že s dříve imputovanými hodnotami zacházíme jako s napozorovanými daty (vystupují tedy jako regresory) a využíváme výše uvedeného vztahu pro odhady neznámých parametrů.

Pro srozumitelnost uvažujme nejprve proměnné s kompletně napozorovanými hodnotami $Y_1, \ldots, Y_{k-1}$ a Y_k pozorovanou pro prvních r respondentů a chybějící pro $n - r$ respondentů. Regresní imputace počítá regresi proměnné Y_k na $Y_1, \ldots, Y_{k-1}$ založenou na r kompletních případech. Chybějící hodnoty jsou následně vyplňovány předpověďmi získanými z regrese.

Regresní imputace ve vícerozměrném modelu je o něco složitější, protože metody imputace musíme aplikovat iterativně. Uvažujme nyní, že proměnné s chybějícími hodnotami jsou uspořádány do matice $\mathbf{Y}$, v níž sloupce $Y_1, \ldots, Y_k$ odpovídají jednotlivým proměnným. Dále mějme plně napozorované prediktory $\mathbf{X}$. V první řadě budeme imputovat veškeré chybějící hodnoty v $\mathbf{Y}$ využitím nějakého jednoduššího přístupu (např. nahrazení průměrnou hodnotou nebo mediánem). V druhé fázi budeme imputovat Y_1 danými $Y_2, \ldots, Y_k$ a $\mathbf{X}$, dále Y_2 danými $Y_1, Y_3 \ldots, Y_k$ a $\mathbf{X}$, tj. využijeme již doplněné hodnoty pro Y_1. Tímto způsobem pokračujeme v doplňování každé proměnné, dokud není dosaženo aproximativní konvergence.

4.3.2 Náhodná hot-deck imputace

Dále zaměříme naši pozornost na přístup s názvem hot-deck imputace. Jedná se o jednoduché a intuitivní metody, které ke svému použití nepožadují splnění žádných složitých předpokladů. Přestože se hot-deck metody imputace v praxi hojně využívají, není k nim příslušná teorie tak propracována, jako je tomu u ostatních metod imputace. Při aplikaci hot-deck imputace je chybějící hodnota u daného respondenta (tzv. příjemce) nahrazena hodnotou získanou od jednoho nebo více kompletních pozorování, tzv. dárců. Dárce je tedy respondent, který se nějakým způsobem podobá jednotce, pro kterou tato hodnota chybí. Existuje několik možných cest, jak vybrat vhodného dárce. Těmito cestami se budeme podrobněji zabývat v následujícím textu. Za zmínku stojí také to, jak vznikl poněkud záhadný název této skupiny metod, jenž pochází z doby, kdy se jako paměťová média užívaly děrné štítky. Na těchto štítcích byly uloženy datové soubory. Při nahrazení chybějící hodnoty hodnotou od jiného respondenta byl děrný

štítek opakovaně použit, a proto docházelo k jeho zahřátí - odtud hot (česky horký).

Níže popsané metody řadící se do skupiny hot-deck imputace vychází z [3], [10] a [20].

Do skupiny metod hot-deck imputace řadíme náhodnou hot-deck imputaci (Random Hot-Deck imputation - RHD), kdy dárce náhodně zvolíme z množiny potencionálních dárců (tzv. donor pool). Příkladem může být situace, kdy u respondenta chybí údaj o jeho příjmu. Dárce vybereme náhodně z množiny potencionálních dárců se stejným pohlavím, věkovou kategorií a vzděláním, jako má příjemce.

Tato metoda dosáhla největšího rozmachu v polovině 20. století, kdy ji používalo velké množství statistických úřadů. Musíme si však uvědomit, že v té době přistupovali obyvatelé k šetřením zodpovědněji. Bylo pro ně ctí, že se mohou šetření zúčastnit, proto byl výskyt chybějících hodnot nízký. Tento přístup bývá mnohdy kritizován [18]. Pokud datový soubor obsahuje velmi málo chybějících hodnot, může metoda poskytovat dobré výsledky.

4.3.3 Sekvenční hot-deck imputace

V tomto případě hledáme dárce ve vzorku postupně, tj. začneme od prvního řádku a pokračujeme na další, dokud nenajdeme vhodného dárce. Chybí-li příjmová položka, seřadíme datový soubor v závislosti na věku, pohlaví a povolání, poté prohledáváme takto seřazený datový soubor od začátku. Často se můžeme setkat s uspořádáním datového souboru podle geografického hlediska, jelikož se předpokládá, že respondenti v rámci jedné geografické oblasti jsou si podobnější než respondenti z různých oblastí. Uvádí se, že použití sekvenční hot-deck imputace (Sequentional Hot-Deck imputation - SHD) je výhodné, pokud datový soubor obsahuje malé množství chybějících hodnot.

4.3.4 Hot-deck imputace nejbližším sousedem

Dále lze imputovat metodou nejbližšího souseda (Nearest Neighbour Imputation), v tomto případně identifikujeme jediného dárce, který je danému respondentovi nejvíce „podobný“. Tuto „podobnost“ definujeme pomocí funkce vzdálenosti. Obecně platí, že funkce vzdálenosti závisí na typu proměnné. Jako dárce tedy zvolíme respondenta s nejmenší vzdáleností od příjemce. Jestliže je chybějící položkou vzdělání, najdeme respondenta s nejpodobnějším příjmem a jeho příjem použijeme k imputaci chybějící hodnoty.

Označme i-tý objekt jako $y_i = (y_{i1}, \ldots, y_{ip})^T$ pro $i = 1, \ldots, n$, kde n je počet pozorování a p značí počet proměnných datového souboru $\mathbf{Y}$. Nejpoužívanějším typem vzdálenosti pro spojité proměnné je euklidovská vzdálenost. Euklidovská metrika mezi i-tým a j-tým pozorováním je dána vztahem

$$d_E(i,j) = \sqrt{\sum_{k=1}^{p}(y_{ij} - y_{jk})^2}.$$

Její zobecnění tvoří Minkowského metrika

$$d_{Min}(i,j) = \sqrt[q]{\sum_{k=1}^{p}(y_{ij} - y_{jk})^2},$$

kde q je libovolné reálné číslo větší nebo rovno 1. V některých případech se používá Mahalanobisova vzdálenost, která je dána následujícím vztahem

$$d_{Mah}(i,j) = \sqrt{(y_i - y_j)\mathbf{\Sigma}^{-1}(y_i - y_j)},$$

kde $\mathbf{\Sigma}^{-1}$ značí inverzní matici k (výběrové) varianční matici datové matice $\mathbf{Y}$.

Jedná-li se o binární proměnnou, jejíž prvky mohou nabývat jen dvou hodnot - obvykle 0 a 1, musíme nadefinovat další vzdálenosti. Pro vysvětlení vezměme v úvahu následující i-té a j-té pozorování, které jsme pro přehlednost zapsali do tabulky.

i	1	0	0	1	0	1	0	1	1	0
j	1	0	0	0	1	0	0	1	0	0

Nyní označme písmeny a až d všechny možné kombinace 0 a 1, které mohou nastat. Nechť a označuje počet kombinací $(0,0)^T$, b počet kombinací $(1,0)^T$, c značí počet kombinací $(0,1)^T$ a konečně d počet kombinací $(1,1)^T$. Je zřejmé, že v našem ilustrativním příkladu platí $a = 4, b = 3, c = 1, d = 2$.
Nyní můžeme definovat Jaccardovu vzdálenost pomocí vztahu

$$d_{Jac}(i,j) = \frac{a}{a+b+c},$$

nebo Randovu vzdálenost jako

$$d_{Ran}(i,j) = \frac{a+d}{a+b+c+d}.$$

Je-li sledovaná proměnná nominální, je naším cílem přepsat hodnoty kategoriální proměnné do binární podoby. Potom můžeme postupovat obdobně jako u binárních proměnných. Např. mějme dán vektor $(1,2,3,2)^t$, jeho binární reprezentace vypadá následovně

1	1	0	0
2	0	1	0
3	0	0	1
2	0	1	0

Dosud jsme se zabývali případy, kdy proměnné byly stejného typu. Proto se nabízí otázka, jak postupovat, jestliže máme k dispozici různé typy proměnných? K měření vzdáleností mezi proměnnými různého typu je určena zobecněná Gowerova vzdálenost

$$d_{Gov}(i,j) = \frac{\sum_{k=1}^{p} \delta_{ijk} \cdot d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}},$$

kde d_{ijk} je vzdálenost mezi i-tým a j-tým pozorováním k-té proměnné a δ_{ijk} jsou váhy.
Je-li proměnná nominální, uvažujeme o sloupcích jako o binárních proměnných, tj.

$$d_{ijk} = \begin{cases} 0 & \text{pokud } y_{ik} = y_{jk}, \\ 1 & \text{jinak.} \end{cases}$$

U proměnných spojitého typu uvažujeme sloupce jako spojité proměnné na nějakém intervalu, vzdálenost vypočítáme pomocí vztahu

$$d_{ijk} = \frac{|y_{ik} - y_{jk}|}{r_k},$$

kde r_k značí rozpětí k-té proměnné.

Pro ordinální kategoriální proměnné nahradíme hodnoty odpovídajícím indexem pozice

$$z_{ik} = \frac{q_{ik} - 1}{\max(q_{ik}) - 1},$$

kde q_{ik} je kategorie odpovídající i-tému pozorování k-té proměnné. S proměnnou z_{ik} dále zacházíme jako se spojitou proměnnou na nějakém intervalu.
Váhy jsou definovány tímto způsobem:

$$\delta_{ijk} = \begin{cases} 0 & \text{jestliže chybí jedna z hodnot } y_{ik} \text{ nebo } y_{jk}, \\ 1 & \text{jinak.} \end{cases}$$

V praxi pozorování s chybějící hodnotou do výpočtu vzdálenosti neuvažujeme. Výše uvedená teorie byla převzata z [14]. Více informací o této problematice lze najít v [6].

4.3.5 Algoritmus k nejbližších sousedů

Algoritmus k-nejbližších sousedů, anglicky nazývaný k-Nearest Neighbour algorithm (k-NN), řadíme také mezi jednodušší metody imputace. Jedná se o zobecnění hot-deck imputace. Základy tohoto algoritmu byly položeny již v polovině

20. století, později byl algoritmus přepracován do dnešní podoby. Jedná se o učící se klasifikační metodu, která nachází uplatnění v nejrůznějších oblastech, např. v kriminalistice k rozpoznávání obličeje nebo v lékařství k identifikaci nádorových buněk.

Algoritmus se kromě výše zmíněného využívá právě k imputaci chybějících hodnot, hledáme pomocí něj k „nejpodobnějších“ pozorování v daném datovém souboru. Imputované hodnoty tedy vypočteme pomocí dostupných hodnot od k nejbližších sousedů.

Nejprve zvolíme parametr k jako malé přirozené číslo, s rostoucím k se zvyšuje náročnost výpočtu, na druhou stranu poskytuje metoda zpravidla lepší výsledky. V dalším kroku hledáme k nejbližších sousedů. Za nejbližší sousedy považujeme ta pozorování, jejichž vzdálenost k danému objektu je menší nebo rovna k-té nejmenší vzdálenosti. Pro výpočet vzdáleností lze použít metriky, které byly zmíněny u předchozí metody. Nahrazovanou hodnotu získáme jako aritmetický průměr či modus dostupných hodnot od námi nalezených k nejbližších sousedů.

Za nevýhodu metody k-NN bývá považována právě nutnost stanovit číslo k, tedy počet nejbližších sousedů. V ideálním případě by mělo být k určeno pomocí simulace, kdy bychom v dostupných datech náhodně vytvořili chybějící hodnoty a tyto chybějící hodnoty bychom poté imputovali metodou k-NN pro různá k. Jako optimální bychom zřejmě zvolili takové k, které minimalizuje rozdíl mezi imputovanými a původně dostupnými hodnotami. Jelikož algoritmus k nejbližších sousedů prochází při hledání „podobných“ objektů všechny prvky datového souboru, musíme v případě rozsáhlého datového souboru počítat s velkým množstvím operací a s tím související časovou náročností výpočtu. V praxi se ke snížení časové náročnosti výpočtů využívají různé optimalizační techniky, např. M-tree [19]. Příliš se nedoporučuje použití algoritmu na kategoriální proměnné nominálního typu, mohou se vyskytnout problémy se stanovením nejbližší kategorie.

4.3.6 Algoritmus IRMI

Častým předpokladem užívaným pro mnohorozměrné metody imputace jsou data pocházející z mnohorozměrného normálního rozdělení, nebo alespoň blížící se mnohorozměrnému normálnímu rozdělení. Jakmile datový soubor obsahuje odlehlé hodnoty, nebo je zešikmený, nabývají na významu metody imputace založené na robustních odhadech, které jsou méně ovlivňovány odlehlými pozorováními. Z tohoto důvodu bude závěr této kapitoly věnován algoritmu IRMI, který je k dispozici v knihovně VIM. Teorie k algoritmu IRMI byla převzata z [15].

Algoritmus zvaný IRMI (Iterative Robust Model-based Imputation) je nástroj určený pro iterační klasickou i robustní imputaci založenou na modelu, pomocí kterého je možné současně imputovat různé druhy proměnných. Tento softwarový nástroj je, jako funkce `irmi()`, dostupný v knihovně VIM. V softwaru R existuje celá řada dalších knihoven, pomocí kterých lze chybějící hodnoty nahrazovat, jmenujme knihovny Amelia, imputation, mix, MICE či mi. Zmíněné knihovny však nejsou určeny pro práci s různými druhy proměnných včetně semi-spojitých proměnných.

Dá se říci, že níže popsaný nástroj IRMI staví na základech, které položil algoritmus zvaný IVEware (Imputation and Variance Estimation software), který je dostupný ve statistickém softwaru SAS. Algoritmus IRMI však přináší řadu vylepšení zejména s ohledem na možnou robustnost imputovaných hodnot. Navíc nepotřebuje pro svoje použití žádnou plně napozorovanou proměnnou. V každém kroku iterace je jedna proměnná použita jako vysvětlovaná, zatímco zbývající proměnné mají úlohu vysvětlujících proměnných.

Algoritmus pracuje následujícím způsobem:

Krok 1: Nejprve jsou veškeré chybějící hodnoty doplněny použitím nějaké jednoduché metody imputace (používá se nahrazení mediánem, jako výchozí je nastaveno nahrazení algoritmem k-NN).

Krok 2: V dalším kroku seřadíme proměnné podle počtu původně chybějících hodnot. Abychom předešli problémům se značením, předpokládejme, že pro-

měnné jsou tímto způsobem již setříděny, tj.

$$\mathcal{M}(\mathbf{x}_1) \geq \mathcal{M}(\mathbf{x}_2) \geq \ldots \geq \mathcal{M}(\mathbf{x}_p),$$

kde $\mathcal{M}(\mathbf{x}_j)$ značí počet chybějících hodnot v proměnné $\mathbf{x}_j$. Dále stanovme $I = \{1, \ldots, p\}$.

Krok 3: Položme $l = 1$.

Krok 4: Označme $m_l \subset \{1, \ldots, n\}$ indexy těch pozorování, která byla původně chybějící v proměnné $\mathbf{x}_l$ a $o_l \subset \{1, \ldots, n\} \setminus m_l$ indexy odpovídající pozorovaným hodnotám proměnné $\mathbf{x}_l$; Nechť $|o_l| = r$ a $|m_l| = n - r$ a nechť chybějící hodnoty v l-té proměnné jsou u prvních r pozorování. A dále nechť $\mathbf{X}^{o_l}_{I\setminus\{l\}}$ a $\mathbf{X}^{m_l}_{I\setminus\{l\}}$ značí datové matice ostatních proměnných odpovídající pozorovaným, respektive chybějícím hodnotám proměnné $\mathbf{x}_l$. Je velmi důležité zdůraznit, že jednotlivé proměnné jsou svázány regresním vztahem

$$\mathbf{x}_l^{o_l} = \mathbf{X}^{o_l}_{I\setminus\{l\}}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\beta}$ jsou neznámé regresní koeficienty a $\boldsymbol{\varepsilon}$ značí vektor chyb. Distribuce $\mathbf{x}_l^{o_l}$ závisí na použitém regresním modelu. Uvažujme proměnnou $\mathbf{x}_l^{o_l}$ jako náhodnou a označme podmíněný průměr jako $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_r)^t$. Užitím teorie zobecněných lineárních modelů definujeme spojovací funkci g takovou, že $\boldsymbol{\mu}$ je lineární prediktor, tj. $(g(\mu_1), \ldots g(\mu_r))^t = \mathbf{X}^{o_l}_{I\setminus\{l\}}\boldsymbol{\beta}$. Jestliže je proměnná:

- **spojitá**, spojovací funkcí je $\boldsymbol{\mu}$, použijeme klasickou regresi (odhady parametrů pomocí metody nejmenších čtverců) nebo robustní regresi,
- **kategoriální**, aplikujeme zobecněnou lineární regresi (případně může být použita robustní regrese),
- **alternativní**, spojovací funkcí je $\log(\frac{\mu_i}{1-\mu_i}),$ pro $i = 1, \ldots, r$, tj. použijeme logistickou lineární regresi (případně může být použita robustní regrese),

- **semi-spojitá**, aplikujeme dvoufázový přístup. V první fázi použijeme logistickou regresi (jako výchozí je nastavena klasická logistická regrese) k rozhodnutí, zda bude imputována také konstanta (obvykle nula). V druhé fázi je na spojitou část dat použita klasická nebo robustní regrese,
- **diskrétní**, použijeme zobecněnou Poissonovu klasickou (robustní) lineární regresi se spojovací funkcí $\log(\mu_i)$, pro $i = 1, \ldots, r$.

Volitelně lze použít postupný výběr modelu podle AIC (Akaikeho informační kritérium) - parametr `step` funkce `irmi`, který zahrnuje pouze k nejdůležitějších proměnných, $k \subset I \setminus \{l\}$ v regresním modelu zmíněném výše v tomto kroku. Jinak uvažujeme $k = I \setminus \{l\}$. Více informací o problematice zobecněné lineární regrese lze najít např. v [5].

Krok 5: V tomto kroku odhadneme regresní koeficienty $\boldsymbol{\beta}$ odpovídající modelu z předchozího kroku. Tyto odhadnuté regresní koeficienty $\hat{\boldsymbol{\beta}}$ použijeme k nahrazení chybějících částí $\mathbf{x}_l^{m_l}$ následujícím způsobem

$$\hat{\mathbf{x}}_l^{m_l} = \mathbf{X}_k^{m_l} \hat{\boldsymbol{\beta}}.$$

Krok 6: Provedeme kroky 4 a 5 pro každé $l = 2, \ldots, p$.

Krok 7: Opakujeme kroky 3 až 6 tak dlouho, dokud se nahrazované hodnoty neustálí, tj. dokud

$$\sum_i (\hat{x}_{il}^{m_l} - \tilde{x}_{il}^{m_l})^2 < \delta, \quad \text{pro všechna } i \in m_l \text{ a } l \in I,$$

a pro dostatečně malou konstantu δ. Zde $\hat{x}_{il}^{m_l}$ značí i-tou imputovanou hodnotu l-té proměnné z poslední iterace a $\tilde{x}_{il}^{m_l}$ je i-tá imputovaná hodnota z předchozí iterace.

Autoři algoritmu [15] uvádí, že ačkoliv neexistuje důkaz konvergence, experimenty s nasimulovanými i reálnými daty ukazují, že algoritmus obvykle konverguje po několika iteracích a již po druhé iteraci dochází k výraznému zlepšení kvality imputace.

Pokud jsou odhady regresních koeficientů v kroku 4 provedeny klasickou cestou, potom algoritmus nazýváme `IMI`. Zkratka `IRMI` poukazuje na robustní regresi, jež redukuje vliv odlehlých pozorování na odhady regresních parametrů. Poznamenejme, že robustní regrese pro spojité a semi-spojité proměnné chrání proti špatně inicializovaným chybějícím hodnotám, jelikož jsou odhady regresních koeficientů založeny na většině pozorování. Musíme si ovšem uvědomit, že semi-spojité proměnné jsou zpravidla součástí velmi komplexních datových souborů, což činí při robustní imputaci značné numerické problémy.

5 Chybějící hodnoty v šetření Životní podmínky

Tato kapitola pojednává o šetření Životní podmínky z roku 2010, kterým se budeme v práci dále zabývat. V roce 2010 bylo k účasti v šetření Životní podmínky náhodným dvoustupňovým výběrem vybráno celkem 11 171 bytů - 4 300 bytů nových (tzv. první vlna) a 6 949 bytů z vln předchozích. Ne všechny domácnosti z předchozích vln byly zastiženy na jejich původních adresách, 165 domácností bylo sledováno do jejich nového bydliště. Ukázalo se, že z celkového počtu bytů ve výběru, bylo 4,9 % neobydlených nebo k bydlení nezpůsobilých. Šetření Životní podmínky bylo provedeno v 10 624 bytech a 10 720 domácnostech, celkem se šetření zúčastnilo 18 209 respondentů. Přehled o výskytu jednotkové non-response v šetření poskytuje následující tabulka.

	Odezva (%)		
	Celkem	**První vlna**	**Další vlny**
Response, celkem	**84,9**	**65,7**	**96,3**
Non-response, celkem	**15,1**	**34,3**	**3,7**
Odmítnutí	78,5	79,7	72,2
Domácnost nekontaktována	15,3	15,1	16,5
Domácnost neschopna odpovídat	5,3	4,4	10,1
Jiné důvody	0,9	0,8	1,2

Tab. 1: Jednotková non-response

Z tabulky vidíme, že úspěšnost šetření výrazně stoupá v dalších vlnách. Hlavní příčinou je zkušenost s daným šetřením získaná účastí v první vlně. Non-response dělíme do čtyř základních kategorií. První kategorii nazýváme *odmítnutí*. Musíme si uvědomit, že tato kategorie zahrnuje také situace, kdy se domácnost neodmítne zúčastnit šetření jako takového, ale neposkytne informace týkající se příjmů v takovém rozsahu, který by kvalifikoval domácnost jako úspěšně vyšetřenou. Za úspěšně vyšetřenou se považuje ta domácnost, kde chybí položky

týkající se příjmů pouze u jediné osoby. Tato osoba zároveň nesmí být hlavou domácnosti. Hlavními důvody pro odmítnutí účasti v šetření jsou ochrana osobních údajů a obava z jejich zneužití, nebo strach z kontaktu s cizími lidmi. Do skupiny *domácnost nekontaktována* řadíme domácnosti, se kterými se nepodařilo navázat kontakt navzdory vykonání minimálně tří předepsaných pokusů. Další kategorie, pojmenovaná *domácnost neschopna odpovídat*, obsahuje ty domácnosti, jejíž členové nejsou v důsledku špatného zdravotního stavu způsobilí k účasti v šetření. Mezi *jiné důvody* spadá např. jazyková bariéra, v tabulce vidíme, že se jedná o okrajový problém.

Nejčastěji se položková non-response objevuje bezesporu u otázek týkajících se příjmů a jejich složek. Následující tabulka ilustruje výskyt položkové non-response u čistých peněžních příjmů osob starších 16 let.

	% osob pobírajících daný příjem	**% osob s chybějící hodnotou**
Zaměstnanecké příjmy	47,76	0,16
Příspěvky do soukromého penzijního poj.	44,20	0,05
Příspěvky ze soukromého penzijního poj.	0,49	0,00
Dávky v nezaměstnanosti	3,76	0,29
Starobní důchod	30,85	0,04
Pozůstalostní dávky	9,63	0,06
Nemocenské dávky	6,38	0,09
Invalidní důchod	8,05	0,07
Příspěvky spojené se studiem	0,78	0,00

Tab. 2: Položková non-response: čisté příjmy osob starších 16 let

Druhý sloupec tabulky ukazuje, jaké procento osob starších 16 let pobírá daný příjem. Ve třetím sloupci potom vidíme, jaké je procentuální zastoupení osob, které sdělili, že daný příjem pobírají, ale neprozradili jeho výši ani žádné čás-

tečné informace (např. interval v jakém se daný příjem pohybuje). Je zřejmé, že se nejvíce chybějících hodnot vyskytuje u položky nesoucí název dávky v nezaměstnanosti, dále pak u zaměstnaneckých příjmů. U dalších složek čistého příjmu se pohybuje non-response řádově v setinách procent. Veškeré výše publikované údaje byly převzaty z [16].

6 Praktická část

Tato kapitola je věnována aplikaci vybraných metod imputace na reálná data. Data pro diplomovou práci byla poskytnuta Českým statistickým úřadem a pochází z šetření Životní podmínky z roku 2010. Jedná se o skutečná data, jež nesmí být z důvodů ochrany osobních údajů respondentů dále šířena, proto byla zpracovávána na půdě ČSÚ. Rozsah výběru činí 9 098 domácností. K dispozici máme náhodně nasimulovanou 1% a 5% non-response u vybraných proměnných, dále také původní kompletní datový soubor. Tento soubor nám později poslouží k posouzení kvality nahrazených hodnot.

Náš datový soubor obsahuje informace o bydlení, o velikosti a vybavení domácnosti, dále informace o nákladech na bydlení a také o finanční situaci domácností. Datový soubor čítá na šedesát proměnných různého typu, chybějící hodnoty najdeme v deseti z nich. Informace o označení a interpretaci proměnných, v nichž se chybějící hodnoty vyskytují, poskytuje následující tabulka:

Označení	**Typ**	**Význam**	**Hodnoty**
CELK_M2	spojitá	velikost bytu - podlahová plocha v m^2	
MALY_BYT	binární	nedostatek místa v bytě	1 - ano, 2 - ne
HLUK	binární	hluk od sousedů nebo z ulice	1 - ano, 2 - ne
VAND_KRIMI	binární	vandalismus, kriminalita	1 - ano, 2 - ne
CENABYTU	spojitá	odhad tržní ceny bytu v Kč	
NAKL_ZATEZ	množná	jsou náklady na bydlení finanční zátěží	1 - velkou zátěží 2 - určitou zátěží 3 - žádnou zátěží
PUJC_ZATEZ	množná	je splácení půjček finanční zátěží	1 - velkou zátěží 2 - určitou zátěží 3 - žádnou zátěží

Označení	Typ	Význam	Hodnoty
VYCHAZELA	množná	jak domácnost vycházela s příjmy	1 - s velkými obtížemi 2 - s obtížemi 3 - s menšími obtížemi 4 - docela snadno 5 - snadno 6 - velmi snadno
DOVOLENA	binární	domácnost si může dovolit ročně alespoň týdenní dovolenou mimo domov	1 - ano, 2 - ne
DOST_VYTAP	binární	dostatečné vytápění bytu	1 - ano, 2 - ne

Tab. 3: Proměnné obsahující chybějící hodnoty

6.1 Knihovna VIM

Jak již bylo zmíněno, zvolili jsme pro zpracování dat statistický software R. Hlavními důvody pro tuto volbu jsou volná šiřitelnost a široká využitelnost tohoto softwaru, jenž se používá především ke statistickým analýzám. K tomuto účelu software R využívá také ČSÚ. Volně dostupné instalační soubory lze najít na **http:/r-project.org**. Na této adrese lze stáhnout také tzv. knihovny (packages), což jsou skupiny funkcí, nápovědy a datových souborů, jejichž doinstalováním si můžeme program dle aktuální potřeby „vylepšit“. Naši pozornost zaměříme na již několikrát zmíněnou knihovnu VIM, která byla speciálně vytvořena pro vizualizaci a imputaci chybějících hodnot. Při zpracování praktické části byl využit především zdroj [22].

Knihovnu lze jednoduše doinstalovat přímo v hlavním menu programu R → *Packages* → *Install Package(s)*. Nejprve musíme zvolit zemi (v našich podmínkách se doporučuje nejlépe nějaká evropská země), poté stačí v seznamu vyhledat potřebnou knihovnu - v našem případě knihovnu VIM. Jestliže chceme knihovnu používat, musíme ji následně také načíst. Načtení rovněž provedeme v hlavním menu → *Packages* → *Load Package*, kde vybereme dříve nainstalovanou kni-

hovnu. Knihovnu můžeme nainstalovat a načíst také přímo v konzole pomocí příkazů

```
> install.packages("VIM"),
> library(VIM).
```

Pokud potřebujeme zjistit, jaké funkce a datové soubory jsou v rámci knihovny k dispozici, zadáme příkaz

```
> data (package ="VIM").
```

Jestliže chceme pracovat s vlastními daty, musíme je v prvé řadě načíst ze souboru, v našem případě ze souboru ve formátu .csv. Náš datový soubor týkající se domácností, který obsahuje 5% non-response, nese název P_5PN_RESP.csv a načteme ho tímto způsobem

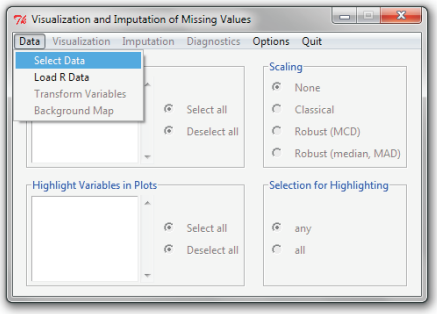
```
> P_5PN=read.csv2("P_5PN_RESP.csv", na.string="-1"),
```

kde funkce `read.csv2` slouží k načtení datové tabulky, která obsahuje hlavičku a jejíž sloupce jsou odděleny středníky, parametr `na.string` určuje jako chybějící ty hodnoty, ve kterých jsou proměnné rovny -1.

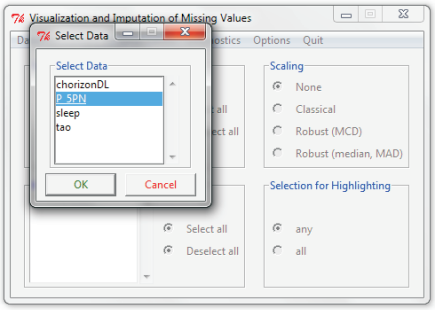
Součástí knihovny je příjemné grafické uživatelské prostředí (Graphical User Interface - GUI), které umožňuje snadnou manipulaci s funkcemi zahrnutými v knihovně `VIM`, vyvoláme je pomocí

```
> vmGUImenu().
```

V uživatelském prostředí GUI můžeme načíst příslušný datový soubor pomocí záložky *Data* → *Select data* (viz. Obr. 2), zde zvolíme z nabídky datových souborů (viz Obr. 3). Kromě souborů, které jsou v knihovně volně k dispozici, zde najdeme pod zkratkou P_5PN data z šetření Životní podmínky, která byla načtena výše zmíněným způsobem. Po úspěšném načtení dat (viz Obr. 4) můžeme přejít k vizualizaci.



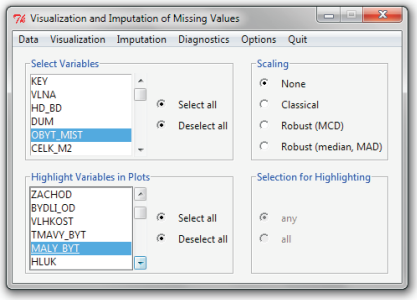
Obr. 2: GUI



Obr. 3: Dialogové okno pro výběr dat

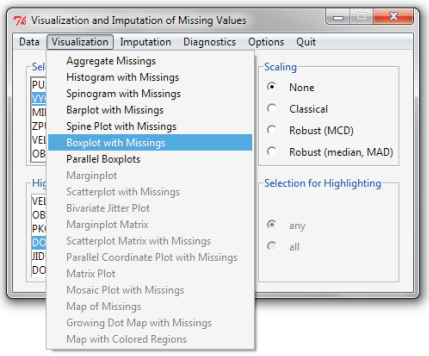
6.2 Vizualizace nekompletních dat

Vizualizaci lze jednoduše provádět pomocí GUI, nejprve zvolíme proměnnou nebo proměnné, pro které má být graf vykreslen. Poznamenejme, že proměnné jsou do grafů zakreslovány v tom pořadí, v jakém byly vybrány. Poté samostatně vybereme proměnnou či proměnné, které mají být v grafu zvýrazněny. Výběr proměnných provádíme v levé části GUI. Pokud chceme zvýraznit dvě nebo více proměnných, je možné v pravé dolní části GUI zvolit, zda budou pozorování s chybějícími hodnotami zvýrazněny pouze v některé, či ve všech označených proměnných. Dále v menu vybereme *Visualization*, kde zvolíme požadovaný typ grafu (viz Obr. 5). Všimněme si, že pokud označíme jednu proměnnou, jsou v nabídce k dispozici pouze jednorozměrné grafy. Velmi důležitou součástí menu související s vizualizací je záložka *Options* → *Preferences*, která slouží mimo jiné k nastavení barevného schématu použitého v grafech. Nyní již k samotné vizualizaci našeho datového souboru. Vizualizační nástroje budeme pro přehlednost



Obr. 4: Výběr proměnných v GUI

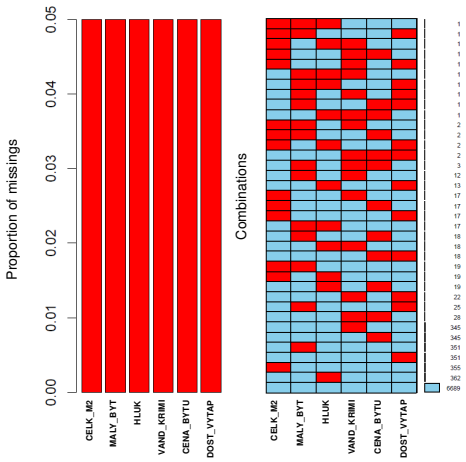
ilustrovat na datovém souboru s 5% non-response, který je uložen pod názvem P_5PN.



Obr. 5: Vizualizace v GUI

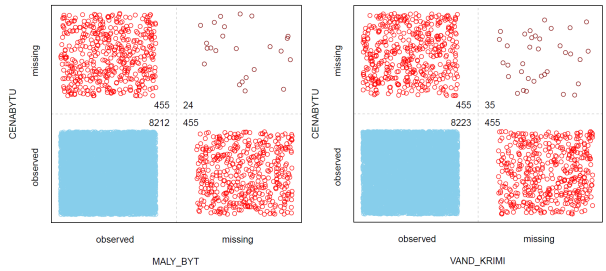
Často nás zajímá, kolik chybějících hodnot se nachází v jednotlivých proměnných, případně jaké existují kombinace chybějících a napozorovaných hodnot. Tyto informace nám poskytuje agregační graf.

Na levé straně Obr. 6 vidíme výskyt non-response ve zvolených proměnných. Je zřejmé, že ve všech námi vybraných proměnných se vyskytuje 5% non-response. Na pravé straně najdeme přehled o výskytu kombinací chybějících a dostupných hodnot - červené obdélníky reprezentují chybějící hodnoty v odpovídající proměnné, modré obdélníky značí dostupná data. Četnost jednotlivých kombinací indikuje malý sloupcový graf na pravé straně. Pro ilustraci byl vykreslen agregační graf sestavený z vybraných proměnných. Agregační graf získáme výše popsaným způsobem v GUI nebo pomocí funkce `aggr()`.



Obr. 6: Agregační graf

Velmi užitečným nástrojem může být tzv. dvourozměrný jitter plot, který zobrazuje kombinace chybějících a dostupných hodnot ve dvou zvolených proměnných. V levé části Obr. 7 vidíme graf zobrazující proměnné MALY_BYT a CENABYTU, který je rozdělen do čtyř čtverců. Je zřejmé, že pro 8 212 domácností máme k dispozici obě proměnné, pro 455 domácností chybí proměnná MALY_BYT, pro stejný počet domácností chybí proměnná CELK_M2. Z tohoto počtu nemáme celkem pro 24 domácností k dispozici ani jednu ze sledovaných proměnných. Na pravé straně vidíme stejný typ grafu pro proměnné VAND_KRIMI a CENABYTU.

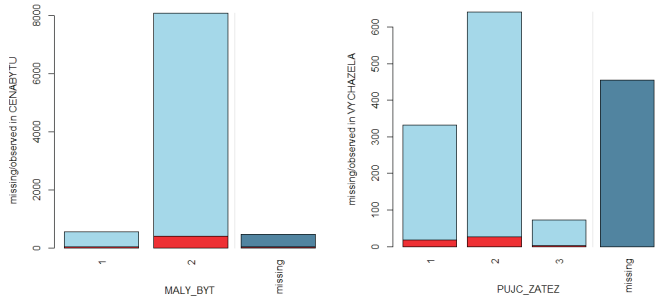


Obr. 7: Jitter ploty

Dále naši pozornost zaměříme na jednorozměrné grafy, v prvé řadě na sloupcový graf, který můžeme opět získat v GUI, nebo použitím funkce `barMiss()`.

V levé části Obr. 8 vidíme sloupcový graf pro proměnnou MALY_BYT se zvýrazněnými chybějícími hodnotami proměnné CENABYTU. Lze zjistit, že za malý svůj byt považuje pouze 560 domácností, naopak 8 083 domácností shledává svůj byt za dostatečně velký. Chybějící hodnoty v proměnné vyjadřující cenu bytu jsou zakresleny červeně. Je patrné, že v prvním sloupci najdeme 26 chybějících hodnot a v druhém sloupci, tedy u domácností, které svůj byt označily jako dostatečně velký, najdeme 405 chybějících hodnot. Dále víme, že proměnná CENABYTU obsahuje celkem 455 chybějících hodnot, z toho vyplývá, že u 24 domácností chybí hodnoty u obou námi sledovaných proměnných. Tmavě modrý sloupec ležící vpravo od grafu indikuje počet chybějících hodnot v proměnné MALY_BYT. Jak již bylo řečeno, v této proměnné chybí 455 hodnot. V pravé části tohoto obrázku najdeme sloupcový graf proměnné PUJC_ZATEZ se zvýrazněnými chybějícími hodnotami v proměnné VYCHAZELA. Všimněme si nízkých absolutních četností u jednotlivých kategorií množné proměnné PUJC_ZATEZ. Tato proměnná totiž na rozdíl od ostatních proměnných, v nichž se vyskytují chybějící hodnoty, obsahuje také nuly. Nula se v této proměnné nachází, pokud příslušná otázka není položena vzhledem k odpovědi na otázku předchozí. Tyto nulové hodnoty se do imputace nezahrnují. První výše popsaný sloupcový graf bychom získali zadáním příkazu

```
> barMiss(P_5PN[,c("MALY_BYT","CENABYTU")]).
```



Obr. 8: Sloupcové grafy

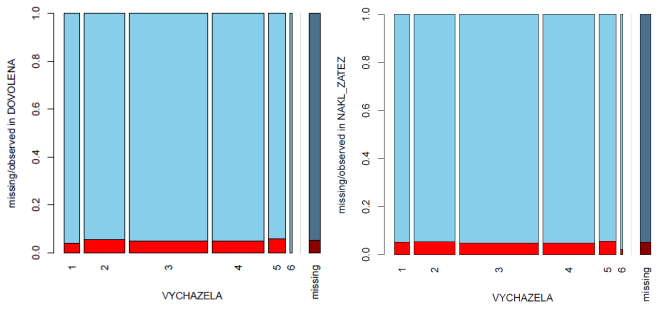
Všechny vizualizační nástroje dostupné v knihovně VIM fungují interaktivně

- jakmile je vykreslen jakýkoliv námi požadovaný graf, můžeme v něm přepínat mezi proměnnými, např. mějme sloupcový graf proměnné MALY_BYT se zobrazením chybějících hodnot pro proměnnou CENABYTU. Kliknutím do levé či pravé části grafu se vykreslí histogram pro proměnnou CENABYTU se zvýrazněnými non-response v proměnné MALY_BYT. Jestliže klikneme jinam, vrátíme se zpátky do GUI a do konzoly.

Nabízí se použití modifikace sloupcového grafu nazývané spine plot. Za přednost tohoto typu grafu považujeme možnost porovnávat podíly chybějících hodnot v jednotlivých kategoriích. V levé části Obr. 9 vidíme vykreslené spine ploty pro proměnnou s názvem VYCHAZELA, jež reprezentuje jak domácnost vycházela se svými příjmy. Chybějící hodnoty v proměnné DOVOLENA jsou odlišeny červenou barvou. V pravé části obrázku najdeme spine plot pro proměnnou VYCHAZELA se zvýrazněnými chybějícími hodnotami proměnné NAKL_ZATEZ. Můžeme říci, že podíly chybějících hodnot k hodnotám dostupným jsou srovnatelné. Všimněme si kategorie s číslem 6 v levé části Obr. , která reprezentuje, že domácnost vychází se svými příjmy velmi snadno, zde nechybí žádné pozorování. Na druhou stranu četnost této odpovědi je pouhých 90. Kdyby byly rozdíly v poměru chybějících a dostupných dat v jednotlivých kategoriích markantní, mohli bychom mluvit o mechanismu vzniku MAR. V našem případě nelze o výrazných rozdílech hovořit, jelikož naše chybějící hodnoty mají mechanismus vzniku MCAR. První zmíněný spine plot jsme získali následujícím příkazem

```
> spineMiss(P_5PN[,c("VYCHAZELA","DOVOLENA")]).
```

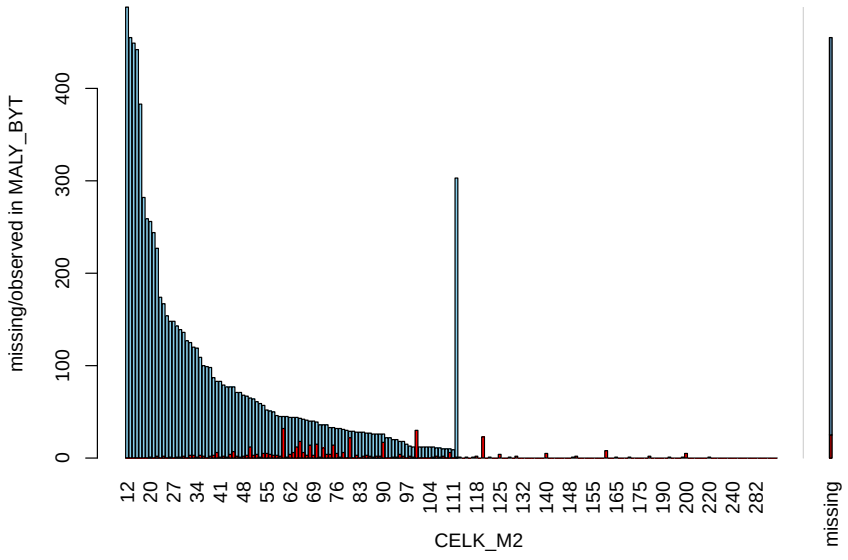
Non-response se v našem datovém souboru vyskytuje ve dvou spojitých proměnných, na tyto proměnné aplikujeme histogram, který nám poskytne informace o jejich rozdělení pravděpodobnosti. Také v histogramu je možné zvýraznit chybějící hodnoty v jiné proměnné. Na Obr. 10 vidíme histogram proměnné CELK_M2 s červeně zvýrazněnými chybějícími hodnotami proměnné MALY_BYT. Chybějící hodnoty proměnné CELK_M2 zobrazuje sloupec na pravé straně. Z histogramu je zřejmé, že zkoumaná proměnná má přibližně exponenciální rozdělení pravdě-



Obr. 9: Spine ploty

podobnosti. Všimněme si vysoké četnosti u plochy bytu 120 m^2 - po detailnějším prozkoumáním datového souboru jsme zjistili, že se v drtivé většině případů jedná o byty v samostatně stojících rodinných domech, dvojdomcích nebo řadových domech. Plocha 120 m^2 odpovídá bytům s minimálně třemi obytnými místnostmi. Tento histogram jsme získali použitím příkazu

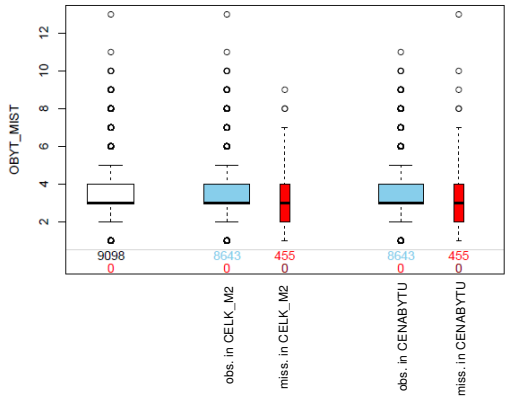
```
> histMiss(P_5PN[,c("CELK_M2","MALY_BYT")]).
```



Obr. 10: Histogram

Dalším nástrojem sloužícím k vizualizaci non-response je boxplot, který zís-

káme užitím funkce `pbox()`. Boxplot umožňuje posoudit variabilitu a symetrii našich dat a také existenci odlehlých hodnot. Na Obr. 11 vidíme paralelní boxploty proměnné OBYT_MIST, která vyjadřuje počet obytných místností v bytě, a v níž se nevyskytují žádné chybějící hodnoty. Medián této proměnné, který uvnitř krabice znázorňuje tučná čára, je roven třem. Odlehlé hodnoty vyskytující se pouze na horní straně upozorňují na nesymetrickou distribuci dat. Na obrázku můžeme dále rozeznat modré krabice značící dostupná data v proměnných CELK_M2 a CENABYTU, červené krabice znázorňující chybějící hodnoty v těchto proměnných. Informace o chybějících hodnotách doplňují čísla v dolní části grafu, červená čísla reprezentují počet chybějících hodnot v dané proměnné, modrá čísla naopak počet dostupných informací.



Obr. 11: Paralelní boxploty

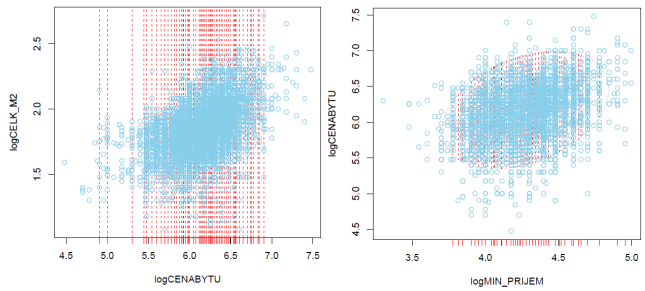
Mezi velmi užitečné nástroje pro zkoumání chybějících hodnot patří bodový graf, v knihovně VIM jej vykreslíme funkcí `scattMiss()`. Před vytvořením grafu jsme pro lepší zobrazení na data aplikovali logaritmickou transformaci. Transformaci můžeme provézt v GUI → *Data* → *Transform Variables*, v levé části okna zvolíme proměnné a v pravé části typ transformace, potvrdíme. Druhou možností je použití funkce `prepare()`, konkrétně parametru `transformation`.

Nyní již k bodovým grafům - v levé části Obr. 12 vidíme bodový graf proměnných logCELK_M2 a logCENABYTU, chybějící hodnoty proměnné logCELK_M2 jsou opět zvýrazněny konstrastující červenou barvou. Z bodového grafu lze usu-

zovat o vzájemné závislosti proměnných, v tomto případě můžeme tedy hovořit o přímé závislosti. V pravé části Obr. 12 vidíme vykreslený bodový graf, v němž obsahuje chybějící hodnoty pouze jedna proměnná - ve spojité proměnné logMIN_PRIJEM se totiž nevyskytují žádné chybějící hodnoty. Tato proměnná vyjadřuje, s jakým nejnižším čistým příjmem by domácnost byla schopna vyjít. Do grafu se v tomto případě zobrazí také elipsa spolehlivosti, pomocí které můžeme identifikovat odlehlá pozorování. Zmíněné grafy jsme získali zadáním následujících příkazů

```
> scattMiss(logP_5PN[,c("logCENABYTU","logCELK_M2")]),
> scattMiss(logP_5PN[,c("logMIN_PRIJEM","logCENABYTU")],
  inEllipse=TRUE),
```

kde logP_5PN je název souboru s transformovanými proměnnými. Pomocí parametru `inEllipse` nastavíme vykreslení přerušovaných čar pouze uvnitř elipsy spolehlivosti. Hlavním důvodem pro tuto volbu je lepší orientace v grafu.

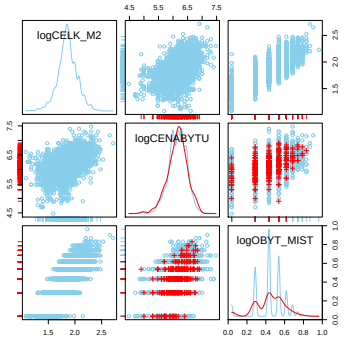


Obr. 12: Bodové grafy

Další typ grafu, na který zaměříme naši pozornost, je známý pod názvem maticový bodový graf. Na Obr. 13 vidíme maticový bodový graf transformovaných proměnných logCELK_M2, logCENABYTU a logOBYT_MIST. Poslední zmíněná proměnná udává počet obytných místností v bytě, jedná se tedy o proměnnou diskrétního typu. Chybějící hodnoty jsou tradičně zvýrazněny červenou barvou, v tomto případě pro proměnnou logCELK_M2. Na „diagonále“ grafu vidíme rozdělení pravděpodobnosti jednotlivých proměnných, mimo „diagonálu“

leží bodové grafy pro všechny kombinace námi vybraných proměnných. Všimněme si, že pokud je na bodovém grafu jednou ze zobrazovaných proměnných proměnná, pro kterou chceme nedostupná data zvýraznit, jsou chybějící hodnoty vykresleny pomocí červených značek u příslušné souřadnicové osy. Jestliže se podíváme na bodové grafy složené z proměnných logCENABYTU a logOBYT_MIST, jsou nedostupná data zobrazena červenými křížky přímo v příslušném bodovém grafu. Tento maticový bodový graf vykreslíme zadáním příkazu

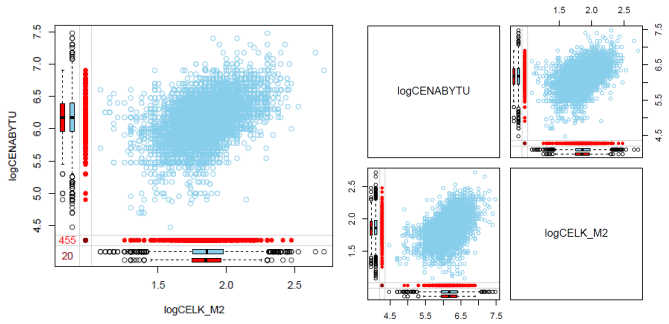
```
> scattmatrixMiss(logP_5PN[,c("logCELK_M2","logCENABYTU",
  "logOBYT_MIST")], highlight="logCELK_M2".)
```



Obr. 13: Maticový bodový graf

Nyní budeme pracovat s modifikací bodového grafu, která nese název marginplot. V levé části Obr. 14 vidíme marginplot pro transformované proměnné logCELK_M2 a logCENABYTU. U souřadnicových os jsou červeně zobrazeny jednorozměrné bodové grafy pro chybějící hodnoty v jednotlivých proměnných, dále jsou zde vykresleny boxploty pro dostupná i chybějící data. Z grafu je zřejmé, že v jednotlivých proměnných chybí hodnoty u 455 domácností, u 20 z nich není k dispozici ani jedna z uvažovaných proměnných. V pravé části obrázku vidíme modifikaci marginplotu zvanou maticový marginplot. Tyto grafy jsme získali následujícím způsobem

```
> marginplot(logP_5PN[,c("logCELK_M2","logCENABYTU")]),
> marginmatrix(logP_5PN[,c("logCELK_M2","logCENABYTU")]).
```

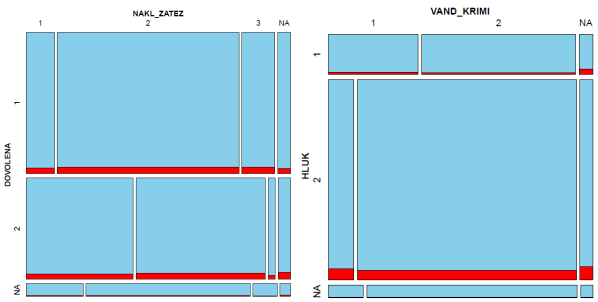


Obr. 14: Marginplot (vlevo) a maticový marginplot (vpravo)

Dále budeme zkoumat mozaikový graf, který představuje interpretaci mnohorozměrné kontingenční tabulky. V levé části Obr. 15 vidíme mozaikový graf proměnných DOVOLENA a NAKL_ZATEZ se zvýrazněnými chybějícími hodnotami v proměnné VYCHAZELA. Plocha každého obdélníku proporcionálně odpovídá četnosti daného prvku v souboru. Jestliže se zaměříme na non-response, vidíme, že poměry chybějících hodnot k hodnotám dostupným jsou ve všech kategoriích zhruba srovnatelné. Výjimku tvoří skupina domácností, které odpověděly, že si nemohou jednou ročně dovolit alespoň týdenní dovolenou mimo domov, a současně pro tyto domácnosti nepředstavují náklady na bydlení žádnou zátěž. Graf na pravé straně zobrazuje proměnné HLUK a VAND_KRIMI se zvýrazněnými chybějícími hodnotami v proměnné CENABYTU. První zmíněný mozaikový graf vykreslíme zadáním následujícího příkazu

```
> mosaicMiss(P_5PN[,c("DOVOLENA","NAKL_ZATEZ")],
  highlight="VYCHAZELA").
```

Poslední vizualizační nástroj, o němž se zmíníme, se nazývá maticový graf. Maticový graf pro proměnné logCENABYTU a logCELK_M2 seřazený podle proměnné logOBYT_MIST je vykreslen na Obr. 16. Dostupná data jsou tentokrát zobrazena v odstínech šedi, tyto odstíny reprezentují jednotlivé kategorie, platí, že s rostoucím počtem pozorování se zvětšuje výška sloupce. Chybějící hodnoty jsou odlišeny kontrastující červenou barvou. Z grafu je zřejmé, že chybějící hodnoty v proměnných logCENABYTU a logCELK_M2 nejsou závislé na proměnné

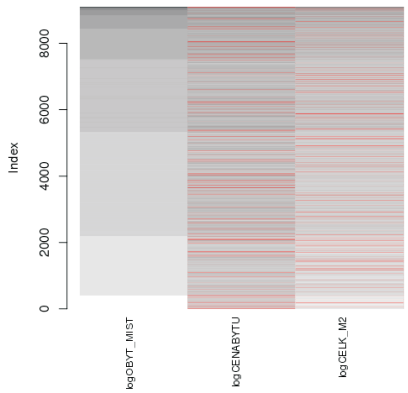


Obr. 15: Mozaikové grafy

vyjadřující počet obytných místností v bytě. Také tento maticový graf potvrzuje, že chybějící údaje v našem datovém souboru mají mechanismus vzniku MCAR. Popsaný graf můžeme opět získat přímo v GUI nebo zadáním následujícího příkazu

```
> matrixplot(logP_5PN[,c("logOBYT_MIST","logCENABYTU",
  "logCELK_M2")]);
```

proměnnou, podle které chceme zbývající proměnné seřadit, vybereme v již vykresleném grafu kliknutím na zvolenou proměnnou.



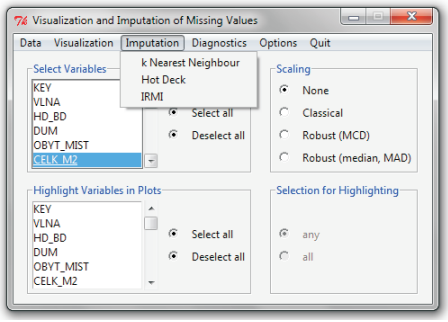
Obr. 16: Maticový graf

Ke stejným závěrům bychom dospěli, kdybychom provedli vizualizaci datového souboru s 1% non-response. Vzhledem k omezenému rozsahu diplomové

práce uvádíme vizualizaci pouze pro datový soubor s 5% výskytem non-response.

6.3 Vybrané metody imputace chybějících hodnot

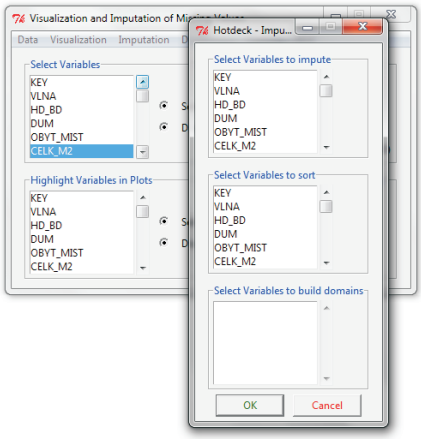
Nyní se budeme věnovat samotné imputaci, kterou ukážeme na datovém souboru obsahujícím 5% non-response. Stejným způsobem doplníme také datový soubor s 1% non-response. Dříve popsané uživatelské prostředí GUI totiž můžeme použít také k doplňování datových souborů, v nichž se vyskytují chybějící data. Pro nahrazování chybějících hodnot slouží v GUI záložka → *Imputation*, nabízené metody jsou algoritmus k-NN, hot-deck imputace a algoritmus IRMI (viz. Obr. 17).



Obr. 17: Imputace pomocí GUI

Nejprve se zaměříme na hot-deck imputaci, zvolíme-li v záložce *Imputation* → *Hot Deck*, otevře se nové okno (viz Obr. 18). Nejjednodušší metoda hot-deck imputace, se nazývá náhodná hot-deck imputace. Chceme-li provést náhodnou hot-deck imputaci, vybereme nejprve v horní části okna proměnné, které mají být imputovány a v dolní části okna potom zvolíme proměnné pro tvorbu tzv. domén neboli skupin, v rámci kterých se bude hledat vhodný dárce.
Imputaci lze samozřejmě provést také zadáním příslušných příkazů do konzoly

```
> P_5PN_RHD=hotdeck(P_5PN,variable=c("CELK_M2","MALY_BYT","HLUK",
  "VAND_KRIMI","CENABYTU","NAKL_ZATEZ","VYCHAZELA","DOVOLENA",
  "DOST_VYTAP"),domain_var=c("VEL","OBLAST","OBYT_MIST",
  "NEOCEK_VYD")).
> write.csv2(P_5PN_RHD,"P_5PN_RHD.csv").
```



Obr. 18: Hot-deck imputace v GUI

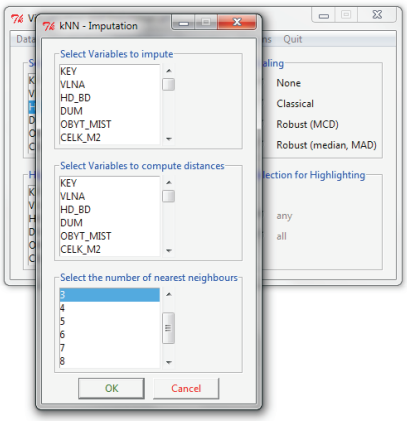
Prvním příkazem provedeme náhodnou hot-deck imputaci chybějících hodnot u všech nekompletních proměnných. K vytvoření domén, neboli skupin v rámci kterých se bude hledat vhodný dárce, použijeme parametr `domain_var`. V našem případě bereme v úvahu velikost obce, stupeň urbanizace, počet obytných místností v bytě a také, zda by si domácnost mohla dovolit neočekávaný výdaj ve výši šest tisíc Kč. Důvodem pro tuto volbu je souvislost prvních zmíněných proměnných s chybějícími hodnotami týkajícími se bydlení, proměnná NEOCEK_VYDAJ zase souvisí s chybějícími hodnotami týkajícími se příjmů domácností. Druhým příkazem uložíme již doplněný datový soubor pod názvem P_5PN_RHD.csv. Pomocí tohoto příkazu budeme pod příslušnými jmény ukládat veškeré kompletní datové soubory.

Další metodou hot-deck imputace, kterou knihovna VIM nabízí, je sekvenční hot-deck imputace. Opět lze využít výše zmíněné okno pro hot-deck imputaci, tentokrát vybereme proměnné pro imputaci a dále ve střední části okna zvolíme proměnné, podle kterých se datový soubor setřídí. Alternativou je použití níže zmíněného příkazu

```
> P_5PN_SHD=hotdeck(P_5PN,variable=c("CELK_M2","MALY_BYT","HLUK",
  "VAND_KRIMI","CENABYTU","NAKL_ZATEZ","VYCHAZELA","DOVOLENA",
  "DOST_VYTAP"),ord_var="VEL").
```

Opět nahrazujeme veškeré chybějící hodnoty v datovém souboru, k setřídění slouží parametr `ord_var`. Provedeme tedy sekvenční hot-deck imputaci u všech nekompletních proměnných se seřazením podle proměnné VEL, která značí velikost obce. Tuto proměnnou jsme vybrali na základě úvahy, že domácnosti žijící ve stejně velkých obcích si jsou určitým způsobem podobné.

Nyní přejdeme k imputaci metodou k-nejbližších sousedů, kterou považujeme za zdokonalení metod hot-deck imputace. Stejně jako v předchozím případě můžeme k nahrazování použít uživatelské prostředí GUI → *Imputation* → *k Nearest Neighbour*, otevře se nové okno. V jeho horní části vybereme proměnné, jejichž chybějící hodnoty chceme nahrazovat, dále ve střední části okna zvolíme proměnné, pro které se mají počítat vzdálenosti, a nakonec určíme číslo k jako počet nejbližších sousedů (viz Obr. 19). Nejprve se zaměříme na imputaci jedním sousedem, kterou provedeme velmi jednoduše zadáním následujícího příkazu



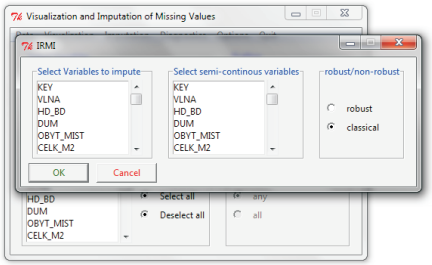
Obr. 19: Hot-deck imputace v GUI

```
> P_5PN_1NN=kNN(P_5PN,k=1).
```

Zdokonalení této metody spočívá v nalezení většího počtu nejbližších sousedů, přičemž imputovanou hodnotu získáme jako průměr hodnot od těchto sousedů. Ve výše zmíněném příkazu stačí zvětšit číslo k. V našem případě jsme zvolili $k = 3, 5, 7$ a 9. Obecně platí, že čím je počet nejbližších sousedů vyšší, tím lepší imputované hodnoty metoda přináší. Dalším volitelným parametrem je parametr

`dist_var()`, pomocí kterého nastavíme proměnné pro výpočet vzdáleností. My jsme pro výpočet vzdáleností použili všechny proměnné z datového souboru, které přímo souvisí s respondenty, tj. např. s výjimkou pořadové proměnné nebo proměnné udávající způsob, jakým šetření proběhlo. Zbývá doplnit, že algoritmus k-NN dostupný v knihovně VIM, používá pro výpočet vzdáleností Gowerovu vzdálenost.

Nyní se budeme věnovat algoritmu `irmi`, také tento algoritmus najdeme v GUI menu *Imputation* → *IRMI*. Na Obr. 20 vidíme, že po otevření příslušného okna stačí zvolit proměnné, které chceme imputovat, dále vybrat semi-spojité proměnné a rozhodnout se mezi klasickou či robustní verzí algoritmu. Stejný výsledek získáme zadáním následujícího příkazu



Obr. 20: Hot-deck imputace v GUI

```
> P_5PN_IMI=irmi(P_5PN,maxit=5),
```

kde parametr `maxit` značí maximální počet iterací (jako výchozí je nastaveno `maxit=100`). Vzhledem k rozsahu datového souboru a časové náročnosti výpočtu jsme zvolili pouhých pět iterací. Tímto příkazem získáme klasickou verzi algoritmu, požadujeme-li robustní verzi, přidáme parametr `robust=TRUE`.

Po aplikaci tohoto algoritmu jsme dostali u spojitých proměnných CELK_M2 a CENABYTU několik záporných hodnot. Výskyt těchto záporných hodnot může být způsoben komplexností algoritmu. Proto jsme se rozhodli na tato data před samotnou imputací aplikovat logaritmickou transformaci, po imputaci jsme na již doplněné datové soubory použili exponenciální funkci.

Výše zmíněnými postupy jsme doplnili také chybějící hodnoty v proměnné PUJC_ZATEZ. V této proměnné se, jak již bylo zmíněno, vyskytuje velké množství nul. Z tohoto důvodu jsme metody imputace neaplikovali na celý datový soubor, ale pouze na jeho podmnožinu, v níž proměnná PUJC_ZATEZ neobsahuje žádné nuly.

6.4 Porovnání použitých metod imputace

K vyhodnocení úspěšnosti jednotlivých metod imputace využijeme původní kompletní datový soubor, který nám byl poskytnut ČSÚ. Pro všechny výše zmíněné metody vypočítáme součet čtverců odchylek pro proměnné s chybějícími hodnotami, který dále vydělíme počtem chybějících hodnot v dané proměnné. Tedy pro k-tou proměnnou postupujeme dle následující vzorce

$$err_k = \frac{\sum_{i=1}^{n} \left(y_{ik}^{kom} - y_{ik}^{imp}\right)^2}{m_k},$$

kde y_{ik}^{kom} značí hodnotu k-té proměnné pro i-té pozorování z původního kompletního datového souboru, zatímco y_{ik}^{imp} označuje odpovídající hodnotu z doplněného datového souboru a m_k je počet chybějících hodnot v k-té proměnné.

Informace o úspěšnosti jednotlivých metod imputace u datového souboru s 5% non-response podává Tab. 4. V prvé řadě zaměřme naši pozornost na metody hot-deck imputace. Všimněme si, že náhodná hot-deck imputace (RHD) vychází ve všech sledovaných proměnných lépe než sekvenční hot-deck imputace (SHD). Za hlavní důvod považujeme fakt, že jsme při náhodné hot-deck imputaci zvolili pro tvorbu domén několik proměnných, zatímco v sekvenční hot-deck imputaci byla k setřídění pozorování vybrána jediná proměnná.

Dále jsme na datový soubor aplikovali algoritmus k-nejbližších sousedů (k-NN), nejprve jsme doplňovali nejbližším sousedem, tj. zvolili jsme $k = 1$. Porovnáním výsledků této metody s náhodnou hot-deck imputací, zjistíme, že imputace nejbližším sousedem přináší lepší výsledky ve třech proměnných, výrazně lépe vychází pouze v proměnné CENABYTU. U proměnné MALY_BYT poskytují obě

	CELK_M2	MALY_BYT	HLUK	VAND_KRIMI	CENABYTU	NAKL_ZATEZ	PUJC_ZATEZ	VYCHAZELA	DOVOLENA	DOST_VYTAP
RHD	$1,148 \times 10^3$	0,101	0,281	0,224	$2,135 \times 10^{12}$	0,541	0,651	1,701	0,332	0,127
SHD	$2,440 \times 10^3$	0,119	0,305	0,253	$3,502 \times 10^{12}$	0,596	0,719	2,321	0,501	0,132
1-NN	$1,319 \times 10^3$	0,101	0,334	0,213	$1,947 \times 10^{12}$	0,565	0,730	2,222	0,455	0,097
3-NN	$1,067 \times 10^3$	0,055	0,246	0,158	$1,403 \times 10^{12}$	0,420	0,567	1,804	0,429	0,079
5-NN	$0,930 \times 10^3$	0,051	0,213	0,149	$1,337 \times 10^{12}$	0,402	0,538	1,648	0,420	0,073
7-NN	$0,871 \times 10^3$	0,041	0,209	0,147	$1,185 \times 10^{12}$	0,378	0,523	1,534	0,418	0,070
9-NN	$0,834 \times 10^3$	0,051	0,202	0,141	$1,132 \times 10^{12}$	0,378	0,512	1,442	0,404	0,070
IMI	$0,982 \times 10^3$	0,051	0,178	0,142	$2,188 \times 10^{12}$	0,266	0,347	0,835	0,176	0,075

Tab. 4: Porovnání metod imputace pro datový soubor s 5% non-response

metody (RHD i 1-NN) totožné výsledky. Dále jsme datový soubor doplnili pomocí metody tří nejbližších sousedů (3-NN). Porovnáním s náhodnou hot-deck imputací vidíme, že doplněním pomocí 3-NN získáváme oproti předchozím metodám lepší výsledky v osmi proměnných, výjimkami jsou proměnné VYCHAZELA a DOVOLENA. Na druhou stranu upozorněme na poměrně výrazné zlepšení u proměnné vyjadřující odhad tržní ceny bytu, která nese označení CENABYTU. Srovnáme-li 1-NN a 3-NN, vidíme u 3-NN zlepšení kvality doplňovaných hodnot. Abychom ověřili předběžnou hypotézu, že se s rostoucím k zvyšuje také kvalita imputace, doplňovali jsme také pěti nejbližšími sousedy (5-NN). Provedením imputace 5-NN získáme znovu lepší výsledky, u proměnné DOVOLENA však stále vítězí náhodná hot-deck imputace. Aplikací algoritmu sedmi nejbližších sousedů (7-NN) dojde k opětovnému zlepšení doplněných hodnot. Zaměříme-li naši pozornost na imputaci devíti nejbližšími sousedy (9-NN), odhalíme, že kvalita imputací opět nepatrně vzrostla. Všimněme si, že se ukazatel kvality imputace u proměnné MALY_BYT se od $k = 5$ dále nezlepšuje, obdobně pro proměnnou DOST_VYTAP od $k = 7$.

Poslední metodou imputace, kterou jsme vyzkoušeli na praktických datech, je algoritmus IMI. Ve srovnání s výsledky ostatních metod vychází IMI lépe u pěti proměnných - jedná se o binární proměnné HLUK a DOVOLENA a množné proměnné s označením NAKL_ZATEZ, PUJC_ZATEZ a VYCHAZELA. V proměnné MALY_BYT dosahuje algoritmus IMI stejných výsledků jako algoritmus k-NN pro $k = 5, 7$ a 9. Všimněme si, že u proměnných VAND_KRIMI a DOST_VYTAP, v níž vychází algoritmus IMI hůře než algoritmus k-nejbližších sousedů, není mezi IMI a k-NN příliš velký rozdíl. Zaměříme-li naši pozornost na spojité proměnné, vidíme oproti algoritmu k-NN značné zhoršení, zejména pak v proměnné CENABYTU. Další zlepšení výsledků by mohlo přinést použití robustní verze tohoto algoritmu, která eliminuje vliv odlehlých pozorování. Tento postup jsme však kvůli omezenému rozsahu diplomové práce a určité numerické nestabilitě robustního algoritmu, která by zřejmě vyžadovala dodatečnou simulační studii, na data neaplikovali.

Nyní se budeme zabývat výsledky získanými pro datový soubor s 1% non-response, které popisuje Tab. 5.

V předchozím případě vycházela náhodná hot-deck imputace ve všech proměnných lépe než sekvenční hot-imputace. Situace u tohoto datového souboru je poněkud odlišná. Srovnáme-li tyto dvě metody, vidíme, že SHD vítězí u pěti proměnných. Všimněme si zejména značného zlepšení u proměnné s názvem CENABYTU. Zaměříme-li se na imputaci nejbližším sousedem, zjistíme, že 1-NN překonává SHD celkem u šesti proměnných. Porovnáním metod 1-NN a RHD, vidíme, že metoda nejbližšího souseda poskytuje lepší výsledky u sedmi proměnných, výjimkami jsou proměnné VYCHAZELA a DOVOLENA. U proměnné HLUK vychází obě metody (1-NN a RHD) totožně. Také u tohoto datového souboru obecně platí, že s rostoucím k, tedy počtem nejbližších sousedů, roste kvalita imputace. Z Tab. 5 je zřejmé, že toto pravidlo neplatí u spojitých proměnných CELK_M2 a CENABYTU pro $k = 7$ a $k = 9$. Také u algoritmu IMI můžeme pozorovat obdobné závěry, jako v předchozím případě - u šesti proměnných poskytuje lepší výsledky než všechny aplikované metody imputace, ve dvou proměnných je srovnatelný s algoritmem k-NN a opět zaostává ve spojitých proměnných.

Porovnáním výsledků v obou datových souborech jsme dospěli k obdobným závěrům. Překvapující mohou být poměrně dobré výsledky metod ze skupiny hot-deck imputace. Nicméně musíme mít na paměti, že tyto metody poskytují dobré výsledky pro chybějící hodnoty s mechanismem vzniku MCAR a pro non-response 5% a nižší, což je náš případ. Dále je zřejmé, že u metody k-nejbližších sousedů roste s číslem k také kvalita imputace. O algoritmu IMI lze říci, že poskytuje lepší výsledky než k-NN u kategoriálních proměnných, naopak u spojitých proměnných za algoritmem k-NN oproti očekáváním poněkud zaostává. Jestliže srovnáme kvalitu doplňovaných hodnot v datovém souboru s 5% a 1% non-response, zjistíme, že jsme obecně obdrželi lepší výsledky pro datový soubor s 1% výskytem non-response. Jedná se o běžný jev, jelikož úspěšnost metod imputace s rostoucím počtem chybějících hodnot klesá.

	CELK_M2	MALY_BYT	HLUK	VAND_KRIMI	CENABYTU	NAKL_ZATEZ	PUJC_ZATEZ	VYCHAZELA	DOVOLENA	DOST_VYTAP
RHD	$0,930 \times 10^3$	0,154	0,286	0,286	$3,248 \times 10^{12}$	0,791	0,736	1,440	0,275	0,165
SHD	$2,229 \times 10^3$	0,110	0,352	0,198	$2,285 \times 10^{12}$	0,659	0,780	2,088	0,418	0,087
1-NN	$0,833 \times 10^3$	0,055	0,286	0,220	$2,702 \times 10^{12}$	0,571	0,670	2,396	0,473	0,099
3-NN	$0,745 \times 10^3$	0,044	0,253	0,198	$1,464 \times 10^{12}$	0,505	0,648	1,473	0,462	0,066
5-NN	$0,666 \times 10^3$	0,044	0,242	0,176	$1,601 \times 10^{12}$	0,451	0,560	1,374	0,440	0,066
7-NN	$0,573 \times 10^3$	0,033	0,242	0,187	$1,288 \times 10^{12}$	0,418	0,527	1,363	0,451	0,055
9-NN	$0,621 \times 10^3$	0,033	0,231	0,165	$1,420 \times 10^{12}$	0,407	0,451	1,242	0,429	0,055
IMI	$1,170 \times 10^3$	0,033	0,176	0,132	$2,100 \times 10^{12}$	0,231	0,352	0,978	0,165	0,055

Tab. 5: Porovnání metod imputace pro datový soubor s 1% non-response

Závěr

Diplomová práce pojednává o chybějících hodnotách a o způsobech jejich nahrazování. Jelikož je problematika imputace chybějících hodnot poměrně široké téma, bylo by prakticky nemožné dodržet doporučený rozsah práce a zároveň se věnovat všem známým metodám imputace. Z tohoto důvodu jsem se omezila na metody imputace, které jsou dostupné v knihovně VIM, či s těmito metodami souvisí. V praktické části byly zjištěné teoretické poznatky aplikovány na reálná data z šetření Životní podmínky. Následným vyhodnocením použitých metod imputace jsme dospěli vesměs k očekávaným závěrům. Nejhorší výsledky poskytovaly jednoznačně metody patřící do skupiny hot-deck imputace. Značné zlepšení výsledků přinesla aplikace algoritmu k-nejbližších sousedů. Zde jsme potvrdili hypotézu, že s rostoucím počtem nejbližších sousedů roste také kvalita doplňovaných hodnot. Tento algoritmus dosahoval nejlepších výsledků pro spojité proměnné, zatímco pro kategoriální proměnné byl překonán algoritmem IMI.

Kromě datových souborů za domácnosti, kterými jsme se zabývala v praktické části diplomové práce, jsem na ČSÚ mohla analyzovat také příslušné soubory týkající se osob. U osob je situace o mnoho komplikovanější - jedná se o velmi komplexní soubory, v nichž se vyskytuje velké množství skrytých vazeb a filtrů, a které navíc obsahují semi-spojité proměnné. Pro práci se semi-spojitými proměnnými je určen algoritmus `irmi`. Jeho použitím jsem však nedospěla k uspokojivým výsledkům, což může být dáno již zmíněnou komplexností těchto datových souborů. Správné provedení imputací by vyžadovalo detailní prozkoumání všech vztahů mezi proměnnými, na které nebylo v diplomové práci dostatek prostoru. Přes všechny zmíněné problémy mne tato problematika velmi zaujala, za výzvu považuji zejména další studium komplexních datových souborů, v nichž se nachází semi-spojité proměnné.

Při psaní této práce jsem musela překonat několik překážek. Jelikož o problematice týkající se chybějících hodnot a metod imputace prakticky neexistuje literatura v českém jazyce, čerpala jsem informace téměř výhradně z anglicky psané literatury. Velkým přínosem pro mne byla, po překonání počátečních ne-

zdarů, práce s reálnými daty v softwaru R. Za nejproblematičtější považuji jednoznačně práci s knihovnou VIM. Bezpochyby se jedná o velmi užitečný nástroj sloužící k vizualizaci chybějících, případně imputovaných hodnot, a k samotné imputaci, který poskytuje velmi kvalitní grafické výstupy. Knihovna však trpí určitými nedostatky, zejména rozhodneme-li se pro užívání grafického prostředí GUI.

Literatura

[1] Allison, P. D., *Missing Data*, The SAGE Handbook of Quantitative Methods in Psychology, 2009, 72 - 89

[2] Allison, P. D., *Multiple imputation for missing data - A cautionary tale*, Sociological Methods & Research **28**, 301 - 309 (2000)

[3] Andridge, R. R., Little, R. J. A., *A review of hot deck imputation for survey non-response*, Int Stat Rev. **78**, 40 - 64 (2010)

[4] Baraldi, A. N., Enders C. K., *An introduction to modern missing data analyses*, Journal of School Psychology **48**, 5 - 37 (2010)

[5] Dobson, A. J., Barnett, A. J., *An introduction to generalized linear models*, 3rd edition, Boca Raton, Fla.: CRC Press, Taylor & Francis Group, 2008

[6] Gower, J. C., *A general coefficient of similarity and some of its properties*, Biometrics **27**, 857-871 (1971)

[7] Hron, K., Kunderová, P., *Základy počtu pravděpodobnosti a metod matematické statistiky*, Olomouc: Vydavatelství Univerzity Palackého, 2013

[8] Little, R. J. A., Rubin, D. B., *Statistical Analysis with Missing data*, 2nd edition, Hoboken: John Wiley & Sons, Inc., 2002

[9] Lohr, S. L., *Sampling: Design and Analysis*, 2nd edition, Boston: Cengange Learning, 2010

[10] Peterson, L. E., *K-nearest neighbor*, Scholarpedia **2**, 1883 (2009)

[11] Särndal, C. E., Lundström, S., *Estimation in Surveys with Nonresponse*, Padstow: John Wiley & Sons Ldt, 2005

[12] Sequensová, V., *Životní podmínky v ČR 2010*, 1. vydání, Praha: Český statistický úřad, 2011

[13] Templ, M., Alfons, A., Filzmoser, P., *Exploring incomplete data using visualization techniques*, Advances in Data Analysis and Classification **6**, 29 - 47 (2012)

[14] Templ, M., *Imputation*, interní materiál, 2013

[15] Templ, M., Kowarik, A., Filzmoser, P., *Iterative stepwise regression imputation using standard and robust methods*, Computational Statistics and Data Analysis **55**, 2793 - 2806 (2011)

[16] *EU-SILK 2010, Intermediate quality report - Czech Republic*, interní materiál ČSÚ, 2011

[17] Durrant, G. B., *Imputation methods for handlich item-nonresponse in social sciences: A methodological review*, [online], dostupné z: *http://eprints.ncrm.ac.uk/86/1/MethodsReviewPaperNCRM-002.pdf* [citováno 18. 10. 2012]

[18] Howell, D. C., *Treatment of missing data* [online], dostupné z: *http://www.uvm.edu/∼/dhowell/StatPages/More_Stuff/Missing_Data*, [citováno 5. 10. 2012]

[19] Pejčoch, D., *Metody řešení problematiky neúplných dat* [online], dostupné z: *http://www.dataquality.cz/tutorial/tutorial_04.pdf*, [citováno 15. 2. 2013]

[20] Teknomo, K., *K nearest neighbors tutorial* [online], dostupné z: *http://people.revoledu.com/kardi/tutorial/KNN/index.html*, [citováno 23. 2. 2013]

[21] Templ, M., Alfons, A., *An application of VIM, the R package for visualization of missing values, to EU-SILC data* [online], dostupné z: *http://cran.r-project.org/web/packages/VIM/vignettes/VIM-EU-SILC.pdf*, [citováno 14. 11. 2012]

[22] Templ, M. et al., *Package 'VIM'*, [online], dostupné z: *http://cran.r-project.org/web/packages/VIM/VIM.pdf*, [citováno 25. 2. 2012]

[23] Imputation [online], dostupné z: *http://www.statcan.gc.ca/pub/12-539-x/steps-etapes/4058328-eng.htm*, [citováno 27. 10. 2012]

[24] Introduction to missing data [online], dostupné z: *http://missingdata.lshtm.ac.uk/*, [citováno 4. 10. 2012]

[25] Missing-data imputation [online], dostupné z: *http://www.stat.columbia.edu/∼gelman/arm/missing.pdf*, [citováno 25. 2. 2013]

[26] Multiple imputation for missing data: Concepts and new development [online], dostupné z:
http://www.math.montana.edu/∼jimrc/classes/stat506/notes/multipleimputation-SAS.pdf,
[citováno 26. 2. 2013]

[27] Životní podmínky českých domácností [online], dostupné z:
http://www.czso.cz/csu/tz.nsf/i/zivotni_podminky_ceskych_domacnosti,
[citováno 10. 9. 2012]

Přílohy

Příloha č. 1: Životní podmínky 2010, dotazník za byt

Příloha č. 2: Životní podmínky 2010, dotazník za hospodařící domácnost

Příloha č. 3: Životní podmínky 2010, dotazník za osobu

Příloha č. 1:

Český statistický úřad

ŽIVOTNÍ PODMÍNKY 2010
DOTAZNÍK ZA BYT

SILC - A

Registrováno
ČSÚ ČV 209/10
ze dne 8.12.2009
IKF 753110

Účast v šetření je dobrovolná. Zjištěné individuální údaje jsou chráněny podle zákona č. 89/1995 Sb., o státní statistické službě, ve znění pozdějších předpisů.

Tazatel zaznamená datum návštěvy a celkový čas zahájení a ukončení rozhovoru nad všemi dotazníky.	den ☐☐ měsíc ☐☐
zahájení ☐☐:☐☐	ukončení ☐☐:☐☐

Adresa nešetřena - administrativní odpad (pouze 1. vlna) ☐

A1 Údaje o vyšetření bytu/domácnosti

	HD1	HD2	HD3	HD4
1. **Domácnost vyšetřena**	☐	☐	☐	☐
Domácnost nevyšetřena				
2. odmítnutí šetření (neochota sdělovat informace)	☐	☐	☐	☐
3. domácnost nezastižena, dočasně nepřítomná	☐	☐	☐	☐
4. domácnost neschopna účasti (zdrav. důvody, vysoký věk)	☐	☐	☐	☐
5. jiné důvody (jazyková bariéra)	☐	☐	☐	☐
šetření nepřichází v úvahu (pouze 2. až 4. vlna)				
6. - celá HD odstěhovaná	☐	☐	☐	☐
7. - celá HD v kolektivní domácnosti nebo instituci	☐	☐	☐	☐
8. - celá HD v zahraničí	☐	☐	☐	☐
9. - žádný člen HD již nežije	☐	☐	☐	☐
10. - HD již neobsahuje panelovou osobu	☐	☐	☐	☐
11. - HD sloučena s jinou HD v bytě	☐	☐	☐	☐

A2 Změny v domácnosti
(pouze vyšetřené domácnosti ze 2. až 4. vlny)

	HD1	HD2	HD3	HD4
1. Některá panelová osoba se z HD odstěhovala	☐	☐	☐	☐
2. Domácnost rozdělena v rámci bytu	☐	☐	☐	☐
3. Jiná změna nebo beze změny	☐	☐	☐	☐

Poznámky:

Kód CZ NUTS **CZ0** ☐☐☐

AI Identifikační údaje

1. Území	☐☐☐☐
2. Pagina	☐☐☐☐
3. Číslo SO	☐☐☐☐
4. Vlna šetření	☐
5. Aktuální počet hospodařících domácností (jen u vyšetřeného bytu)	☐
6. Číslo tazatele	☐☐☐☐☐
AIK. Kontrolní součet	☐☐☐☐☐
Obec:	

Vzory číslic 0123456789 Vyznačování ☒

A3 Druh domu

1. samostatně stojící rodinný dům	☐
2. dvojdomek, řadový dům	☐
3. bytový dům s méně než 10 byty	☐
4. bytový dům s 10 a více byty	☐
5. jiný	☐

A4 Počet dotazníků

A byt ☐	B HD ☐	C osoby ☐☐
		za byt celkem ☐☐

A5 Osoby zúčastněné na šetření

	Datum	Jméno	Podpis
Tazatel předal			
Kontroloval, vyznačil			

A6 Údaje o osobách

Zapisují se následující osoby:
V domácnostech, které jsou zahrnuty do šetření **prvním rokem**, se zapisují osoby, které ve vybraném bytě obvykle bydlí, dále podnájemníci a hosté, jejichž zamýšlená délka pobytu v dané domácnosti je delší než 6 měsíců, a rovněž osoby dočasně nepřítomné, které však nejsou členy žádné jiné bytové domácnosti, mají jasnou finanční vazbu na vybranou domácnost a jejich nepřítomnost nepřekročí dobu 6 měsíců. Výjimku tvoří osoby studující nebo pracující mimo domov, u nichž nezáleží na délce nepřítomnosti, avšak nesmí mít žádnou jinou soukromou adresu a musí mít úzké finanční vazby na vybranou domácnost.

V domácnostech, které jsou zahrnuty do šetření **opakovaně**, se zapíší všechny osoby z Výpisu osob z předchozí vlny šetření, a to jak současní členové domácnosti, tak i odstěhované a zemřelé osoby. Zapisují se i nově přistěhovaní a narození, kteří jsou členy HD s alespoň jednou panelovou osobou.

Nezapisují se osoby, které jsou jen dočasně přítomné, ale mají svou vlastní domácnost jinde (návštěvy, osoby jinde v nájmu atd.), a dále osoby dlouhodobě nepřítomné bez existenčních vazeb na vybranou domácnost, jejichž doba nepřítomnosti je delší než 6 měsíců. Při opakované návštěvě domácnosti se nezapisují ani osoby, které spolu tvoří samostatnou HD složenou pouze z osob mimo panel, tj. členové HD, která neobsahuje žádnou panelovou osobu.

Pořadové číslo		IČ osoby	Křestní jméno	Vztah k uživateli bytu, osobě v čele HD			Identifikační číslo (ze sl. 3)		
osoby	HD						otce	matky	partnera
			Slouží ke snadnější identifikaci osob při rozhovoru. Jako první se zapíše uživatel/ka bytu, pak jeho/její partner/ka, děti a ostatní osoby.		BD	HD			
1	2	3	4		5	6	7	8	9
1.									
2.									
3.									
4.									
5.									
6.									
7.									
8.									
9.									
10.									

A7 Dohledané osoby

Původní identifikace dohledaných osob z předchozí vlny šetření	území	pagina	IČ osoby	

	Vztah k uživateli bytu, osobě v čele HD
1	uživatel/ka bytu, osoba v čele HD
2	manželka, družka (ve sl. 5 též manžel, druh)
3	syn, dcera
4	zeť, snacha
5	vnuk, vnučka
6	otec, matka, tchán, tchyně
7	bratr, sestra
8	jiná příbuzná/blízká osoba
9	podnájemník, členové jeho HD (pouze sl. 5)

A8 Údaje o společném hospodaření

Uveďte prosím pořadová čísla osob, které spolu v rámci bytu/domu tvoří samostatně hospodařící domácnosti. Např. 1 + 2, 3 + 4 + 5, 6.

Pořadová čísla osob	HD č. 1	HD č. 2	HD č. 3	HD č. 4	HD č. 5

A6 Údaje o osobách

Měsíc a rok narození		Pohlaví	Rodinný stav	Rok sňatku	Přítomnost/nepřítomnost v domácnosti (demografický pohyb, stěhování)								Kontrolní součet
								Měsíc od	Rok od	Měsíc do	Rok do		
měsíc	rok (poslední dvojčíslí)	1 muž 2 žena		Zapíše se dvojčíslí roku posledního sňatku*	1 panelová osoba 2 osoba mimo panel	Status osoby (2. až 4. vlna)	1 přítomný 2 dočasně nepřítomný	Zapíše se měsíc a dvojčíslí roku, kdy se osoba do domácnosti přistěhovala		Zapíše se měsíc a dvojčíslí roku, kdy se osoba z domácnosti odstěhovala, popř. zemřela		Místo odstěhování	Součet číselných hodnot ve sloupcích 2, 3, 5 až 22
10	11	12	13	14	15	16	17	18	19	20	21	22	23

	Rodinný/osobní stav
1	svobodný(-á)
2	ženatý, vdaná, registrované partnerství
3	ovdovělý(-á), registrované partnerství zaniklé smrtí
4	rozvedený(-á), registrované partnerství zaniklé rozhodnutím

* U registrovaného partnerství se zapíše dvojčíslí roku poslední registrace

	Status osoby (pouze 2. až 4. vlna)
1	osoba šetřena v tomto bytě i v minulém roce
2	přistěhovaný(-á)
3	narozený(-á)
4	odstěhovaný(-á)
5	zemřelý(-á)
6	bývalý člen domácnosti přítomný v minulém roce více než 3 měsíce

	Místo odstěhování
1	do jiné soukromé domácnosti v ČR
2	do kolektivní domácnosti či instituce
3	do zahraničí
4	nezjištěno

Poznámky:

A6 Údaje o osobách

Pořadové číslo osoby	Nejvyšší dokončené vzdělání	Dvojčíslí roku dokončení nejvyššího vzdělání	Současné studium	Dotazník C za osobu: Výsledek vyšetření dotazníku C	Dotazník C za osobu: Způsob vyplnění	Dotazník C za osobu: IČ osoby, která poskytla údaje za jinou osobu	Kontrolní součet: Součet číselných hodnot ve sloupcích 24 až 29
	24	25	26	27	28	29	30
1.							
2.							
3.							
4.							
5.							
6.							
7.							
8.							
9.							
10.							

	Nejvyšší dokončené vzdělání / Současné studium
0	předškolní děti, neukončený 1. stupeň ZŠ (sl. 24)
1	první stupeň ZŠ
2	druhý stupeň ZŠ
3	vyučen(a), nižší střední
4	úplné střední s maturitou
5	nástavbové studium, pomaturitní kurzy, absolvování dvou nebo více SŠ
6	vyšší odborné (DiS.)
7	vysokoškolské bakalářské
8	vysokoškolské magisterské či inženýrské
9	doktorské (Ph.D., CSc., DrSc.)

	Výsledek vyšetření dotazníku C
1	kompletní
2	částečně vyplněný - chybí příjmy
3	částečně vyplněný - příjmy jsou, chybí jiné údaje
4	nevyplněný
5	osoba mladší 16 let
6	osoba již není členem domácnosti

	Způsob vyplnění
1	osobně s tazatelem
2	jinou osobou
3	samovyplnění

Příloha č. 2:

Český statistický úřad

ŽIVOTNÍ PODMÍNKY 2010
DOTAZNÍK ZA HOSPODAŘÍCÍ DOMÁCNOST

SILC - B

Registrováno
ČSÚ ČV 209/10
ze dne 8.12.2009
IKF 753210

Při opětovné návštěvě tazatel nejprve dotazem u respondenta ověří, zda se jedná o byt stejné velikosti a uspořádání jako v předchozím roce šetření, tj. bez zásadních změn co do počtu místností, rozlohy bytu nebo jeho příslušenství. Dále se ujistí, že nedošlo ke změně počtu hospodařících domácností v bytě (sloučení/rozdělení domácnosti).

taková změna nastala ☐ ↓ | stejný byt jako loni ☐ → **B5**

BI Identifikační údaje

1. Území
2. Pagina
3. Číslo HD
4. IČ osoby, která poskytla informace pro dotazník B

BIK. Kontrolní součet

Vzory číslic 0 1 2 3 4 5 6 7 8 9 — Vyznačování ☒

Bydlení

B1 Kolik obytných místností využívá Vaše domácnost?

počet místností

B2 Jaká je celková plocha Vašeho bytu?

plocha v m²

B3 Je ve Vašem bytě následující příslušenství?

	ano	ano, s další HD v bytě	ne
1. koupelna, sprchový kout	☐	☐	☐
2. splachovací WC	☐	☐	☐

B4 Kdy jste se nastěhovali do tohoto bytu? Pokud došlo ke změně právního vztahu Vaší domácnosti k tomuto bytu, uveďte rok této změny.

rok

B5 Vnímáte některý z následujících problémů spojených s Vaším bydlením?

	ano	ne
1. zatékání střechou, vlhké zdi, podlahy, základy nebo shnilá okna, rámy, podlahy	☐	☐
2. příliš tmavý byt, nedostatek denního světla	☐	☐
3. příliš malý byt, nedostatek místa	☐	☐
4. hluk od sousedů nebo hluk z ulice (doprava, obchody, továrny atd.)	☐	☐
5. znečištění, špína nebo jiné problémy se životním prostředím	☐	☐
6. kriminalita, násilí nebo vandalismus v okolí	☐	☐

B6 Jakou právní formou Vaše domácnost užívá tento byt?

Vlastníci	1. byt ve vlastním domě	☐	→ **B7**
	2. byt v osobním vlastnictví	☐	
	3. družstevní byt (SBD)	☐	
Nájemci	4. nájemní (pronajatý) byt	☐	→ **B10**
	5. podnájem (část bytu)	☐	→ **B11**
Bezplatné užívání	6. služební, domovnický byt	☐	
	7. bydlení u příbuzných apod.	☐	

Vlastníci

B7 Splácíte nebo spláceli jste v minulém roce na tento dům/byt hypotéku nebo úvěr ze stavebního spoření?

1. ano, alespoň jedna splátka byla v minulém roce ☐
2. ano, splácíme až od letošního roku ☐ → **B11**
3. ne ☐ → **B11**

B8 Můžete mi sdělit celkovou měsíční splátku Vaší hypotéky/úvěru zahrnující splátku jistiny i úroků? Uveďte prosím také počet měsíců, po které jste v minulém roce hypotéku/úvěr spláceli.

měsíční splátka v Kč

počet měsíců

B9 Kolik činil součet zaplacených úroků za minulý kalendářní rok?

součet úroků v Kč → **B11**

neznám výši úroků ☐ → **B11**

Nájemci

B10 Jaký typ nájemného platíte?

1. tržní ☐ | 2. regulované ☐

Tržní cena bytu/domu

B11 Následující otázka slouží k odhadu současných cen bydlení v ČR. Pokuste se prosím odhadnout tržní cenu bytu/domu, ve kterém bydlíte.

cena v Kč

BK1. Kontrolní součet za otázky B1, B2, B4, B8, B9 a B11

Náklady na bydlení

B12 Uveďte prosím náklady na Vaše bydlení dle jednotlivých výdajových položek:

měsíční - současné

	neplatí	zahrnuto v položce 1	částka v Kč
1. nájemné, úhrada za užívání bytu, fond oprav atd.	☐		
2. společné služby pro celý dům	☐	☐	
3. elektřina	☐	☐	
4. ústřední vytápění a teplá voda	☐	☐	
5. plyn z dálkového zdroje	☐	☐	
6. vodné a stočné	☐	☐	
roční - minulý kalendářní rok			
7. odvoz odpadků	☐	☐	
8. tuhá a tekutá paliva	☐	☐	
9. ostatní náklady (pojištění domu/bytu, běžná údržba)	☐	☐	
B12K. Kontrolní součet za tab. B12			

B13 Vezmete-li v úvahu své celkové náklady na bydlení a dále platby úroků z hypotéky/úvěru (nikoliv však splátky jistiny), řekli byste, že jsou tyto výdaje pro Vaši domácnost:

1. velkou zátěží ☐
2. určitou zátěží ☐
3. žádnou zátěží ☐

Vybavení domácnosti

B14 Uveďte prosím, které z následujících předmětů máte ve Vaší domácnosti. U předmětů, které nemáte, označte důvod, proč je nemáte:

	má	nemá: nemůže si dovolit	nemá: jiný důvod
1. telefon (mobilní, pevná linka)	☐	☐	☐
2. barevný televizor	☐	☐	☐
3. počítač	☐	☐	☐
4. přístup na internet	☐	☐	☐
5. pračka	☐	☐	☐
6. automobil (nikoliv výhradně služební)	☐	☐	☐

Finanční situace

B15 Může si Vaše domácnost dovolit zaplatit z vlastních zdrojů neočekávaný výdaj ve výši 8 500 Kč?

1. ano ☐ 2. ne ☐

B16 Může si Vaše domácnost dovolit uvedené služby nebo výrobky?

	ano	ne
1. zaplatit ročně všem členům HD alespoň týdenní dovolenou mimo domov	☐	☐
2. jíst maso, drůbež nebo ryby každý druhý den (nebo jejich vegetariánské náhražky)	☐	☐
3. dostatečně vytápět byt	☐	☐

B17 Splácíte nějaké půjčky z nákupů na splátky nebo leasing či spotřebitelský úvěr? (Nezahrnujte již úvěry spojené s Vaším bydlením.)

1. ano ☐ 2. ne ☐ → **B19**

B18 Do jaké míry představuje splácení těchto dluhů a úroků z nich finanční zátěž pro Vaši domácnost?

1. velkou zátěž ☐
2. určitou zátěž ☐
3. žádnou zátěž ☐

B19 Dostala se Vaše domácnost někdy během posledních 12 měsíců do takových finančních problémů, že nebyla schopna zaplatit v termínu některou z následujících plateb?

	ano, jednou	ano, vícekrát	ne	netýká se
1. nájemné za byt	☐	☐	☐	☐
2. platby za teplo, elektřinu, plyn, vodu za tento byt	☐	☐	☐	☐
3. splátka hypotéky nebo půjčky na tento dům/byt	☐	☐	☐	☐
4. splátky ostatních půjček a úvěrů	☐	☐	☐	☐

B20 Jak vychází Vaše domácnost s celkovým měsíčním příjmem?

1. s velkými obtížemi ☐
2. s obtížemi ☐
3. s menšími obtížemi ☐
4. docela snadno ☐
5. snadno ☐
6. velmi snadno ☐

B21 Jaký nejnižší možný čistý měsíční příjem by musela mít Vaše domácnost, aby s ním vyšla? Odpovězte prosím s přihlédnutím k současnému složení a podmínkám ve Vaší domácnosti.

měsíční částka v Kč

MODUL 2010

TAZATEL: *Žijí v domácnosti alespoň 2 osoby narozené 1993 a dříve?*

1. ano ☐ 2. ne ☐ → **B24**

B22 Jakým způsobem nahlížíte na příjmy Vaší domácnosti?

1. všechny příjmy považujeme za společné ☐
2. část příjmů považujeme za společné a část příjmů bereme jako soukromé ☐
3. všechny příjmy považujeme za soukromé ☐
4. žádné příjmy nemáme ☐

B23 Kdo je zodpovědný za hospodaření se společnými financemi (příjmy, platbami, úsporami) domácnosti? (Lze označit více odpovědí.)

1. jeden nebo více členů domácnosti ☐

	IČ	Křestní jméno

2. jedna nebo více osob mimo domácnost (účetní, daňový nebo finanční poradce, sociální pracovník, příbuzní) ☐

3. nemáme žádné společné finance domácnosti ☐

→

Spotřeba z vlastního hospodářství nebo podniku

B24 Odhadněte prosím množství, resp. hodnotu spotřebovaných výrobků, které Vaše domácnost spotřebovala z vlastního hospodaření nebo podniku, který vlastníte nebo provozujete (nezahrnujte spotřebu pro krmení zvířat).
Uveďte množství, případně částku za celý minulý kalendářní rok.

1. maso a masné výrobky (kg)
2. vejce (ks)
3. brambory (kg)
4. ovoce (kg)
5. zelenina (kg)
6. ostatní potraviny a nápoje včetně stravování (Kč)
7. průmyslové výrobky a služby (Kč)

nemáme ☐

Transfery mezi domácnostmi

B25 Mnoho lidí poskytuje peněžní nebo naturální výpomoc osobám žijícím v jiné domácnosti nebo v nějaké instituci.
Uveďte prosím částku příjmů/výdajů, kterou jste Vy nebo někdo jiný z Vaší domácnosti v minulém kalendářním roce dostával/poskytoval pravidelně (opakovaně) a u jednorázových částek uveďte součet za celý rok. Pokuste se také odhadnout hodnotu přijatých, resp. věnovaných darů za celý minulý kalendářní rok.

	transfery nebyly		přijaté v Kč	vydané v Kč
1. výživné (na děti i bývalého partnera)	☐	☐ M ☐ R		
2. další opakované peněžní transfery (podpora studentů, blízkých osob v jiných domácnostech)	☐	☐ M ☐ R		
3. jednorázové a mimořádné částky	☐	R		
4. naturální transfery (přijaté/darované produkty z vlastního hospodářství nebo podniku, bezplatné stravování u příbuzných, přijaté/darované výrobky a služby)	☐	R		

BK3. Kontrolní součet za otázky B23 až B25

Dávky státní sociální podpory a sociální péče

B26 Pobírala Vaše domácnost v minulém kalendářním roce některý z těchto sociálních příjmů?
Uveďte prosím počet měsíců pobírání a měsíční částku, resp. částku za celý rok.

	nepobírala		počet měsíců	Kč	počet měsíců	Kč
1. přídavky na děti	☐	M				
2. sociální příplatek	☐	M				
3. příspěvek na bydlení	☐	M				
4. dávky pěstounské péče	☐	M				
5. pomoc v hmotné nouzi	☐	M				
6. jiné sociální dávky	☐	R				
7. porodné	nepobírala ☐		Kč	8. pohřebné	nepobírala ☐	Kč

B26K. Kontrolní součet za tab. B26

Příjmy z pronájmu

B27 Měl(a) jste Vy nebo někdo jiný z Vaší domácnosti v minulém kalendářním roce příjem z pronájmu nemovitosti (bytu nebo jeho části, domu, nebytových prostor, chaty, pozemku), popř. movitých věcí (auta, strojů atd.)?

1. ano ☐ | 2. ne ☐ → **B30**

B28 Můžete prosím uvést, kolik činil příjem z tohoto pronájmu po odečtení nákladů (údržba, opravy, úroky z úvěru, pojištění a jiné poplatky)? Označte, zda uvádíte hrubou nebo čistou částku a zda jde o měsíční nebo roční příjem.

příjem z pronájmu v Kč ☐☐☐☐☐☐☐

1.	hrubá částka ☐	měsíční ☐	→ **B30**
2.	čistá částka ☐	roční ☐	
3.	neznám přesnou výši ☐ ↓		

B29 Odhadněte prosím alespoň interval, do kterého by patřil hrubý roční příjem Vaší domácnosti z tohoto pronájmu.

A.	méně než 20 000 Kč ☐	D.	100 001 – 200 000 ☐
B.	20 001 – 50 000 ☐	E.	200 001 – 500 000 ☐
C.	50 001 – 100 000 ☐	F.	500 001 a více ☐

→

Daně z nemovitostí

B30 Platil(a) jste Vy nebo někdo jiný z Vaší domácnosti v minulém roce daň z nemovitosti?

1. ano ☐ | 2. ne ☐ → **B33**

B31 Uveďte prosím částku, kterou jste musel(a) zaplatit.

částka v Kč ☐☐☐☐☐ → **B33**

nevím ☐ ↓

B32 Odhadněte prosím alespoň interval, do kterého by zaplacená částka pravděpodobně patřila.

A.	méně než 1 000 Kč ☐	D.	3 001 – 4 000 ☐
B.	1 001 – 2 000 ☐	E.	4 001 – 5 000 ☐
C.	2 001 – 3 000 ☐	F.	5 001 a více ☐

BK4. Kontrolní součet za otázky B28 a B31 ☐☐☐☐☐☐☐

Péče o děti do 12 let (narozené v roce 1997 a později)

B33 Jakým způsobem je ve Vaší domácnosti zajištěna péče o děti do 12 let (kromě péče samotných rodičů případně pěstounů)? Uveďte prosím, zda je Vaše dítě předškolák či školák a kolik hodin týdně tráví v uvedených zařízeních nebo v péči jiné osoby.

Identifikační číslo dítěte	předškolák	školák	1. předškolní zařízení	2. povinná školní docházka	3. školní družina, dětská centra	4. denní zařízení, stacionáře	5. chůva, au-pair	6. prarodiče, příbuzní a známí
☐☐	☐	☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐
☐☐	☐	☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐
☐☐	☐	☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐
☐☐	☐	☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐
☐☐	☐	☐	☐☐	☐☐	☐☐	☐☐	☐☐	☐☐

v HD nejsou děti do 12 let ☐ | B33K. Kontrolní součet za tab. B33 ☐☐☐☐

Péče o nesoběstačné osoby

B34 Pečuje některý člen Vaší domácnosti o nesoběstačnou osobu (z části nebo zcela, trvale či dočasně) - ať už z Vaší domácnosti nebo z jiné soukromé domácnosti? Pokud ano, uveďte prosím, kolik hodin týdně obvykle věnuje jednotlivým druhům péče.

Kdo poskytuje	Komu je poskytována				Jaká péče je poskytována - počet hodin týdně				
					v domácnosti příjemce péče			mimo domácnost příjemce péče	
IČ osoby	IČ osoby	pohlaví	věk	stupeň závislosti	zdravotní péče	ostatní péče o osobu	péče o domácnost	návštěvy ve zdravotnických zařízeních	ostatní pomoc
☐☐	☐☐	☐	☐☐	☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
☐☐	☐☐	☐	☐☐	☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
☐☐	☐☐	☐	☐☐	☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
☐☐	☐☐	☐	☐☐	☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐

nikdo z HD nepečuje o takové osoby ☐ | B34K. Kontrolní součet za tab. B34 ☐☐☐☐

Mnohokrát děkujeme za účast na tomto rozhovoru

Příloha č. 3:

ČSÚ
Český statistický úřad

ŽIVOTNÍ PODMÍNKY 2010
DOTAZNÍK ZA OSOBU

SILC - C

Registrováno
ČSÚ ČV 209/10
ze dne 8.12.2009
IKF 753310

Tazatel zaznamená datum, čas zahájení a ukončení rozhovoru nad dotazníkem C	den	měsíc
zahájení :	ukončení :	

Vzory číslic 0 1 2 3 4 5 6 7 8 9 — Vyznačování ☒

CI Identifikační údaje

1.	Území	
2.	Pagina	
3.	IČ osoby	
CIK.	Kontrolní součet	

Pracovní aktivita

C1 Uveďte prosím Vaši pracovní aktivitu v jednotlivých měsících minulého roku a v současnosti. V případě, že jste v měsíci pracoval(a) na částečný úvazek, vyznačte do hlavní činnosti jeden z kódů 1 až 3 a zatrhněte políčko v řádku částečný úvazek.

	1. leden 2009	2. únor 2009	3. březen 2009	4. duben 2009	5. květen 2009	6. červen 2009	7. červenec 2009	8. srpen 2009	9. září 2009	10. říjen 2009	11. listopad 2009	12. prosinec 2009	13. celý rok 2009	14. aktuální stav
A. **hlavní činnost**														
částečný úvazek	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
souběžné činnosti														
B. zaměstnanec	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
C. samostatně činný(-á)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
D. žák, učeň, student (denní studium)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
důchody a sociální dávky														
E. starobní důchod	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
F. invalidní důchod	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
G. rodičovský příspěvek	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

C1K. Kontrolní součet za tab. C1

Pracovní aktivita - hlavní činnost	
1 zaměstnanec	6 žák, učeň, student (denní studium)
2 samostatně činný(-á)	7 ve starobním důchodu
3 na placené mateřské dovolené	8 v invalidním důchodu
4 na rodičovské dovolené	9 v domácnosti, péče o děti nebo péče o blízkou osobu
5 nezaměstnaný(-á)	0 ostatní ekonomicky neaktivní

CT1 Na základě odpovědí na otázku C1A zaškrtněte odpovídající možnost.

1.	v C1A odpověděl v posledním sloupci kódy 1 až 3	☐ → **C13**
2.	v C1A odpověděl v posledním sloupci kódem 5	☐ → **C2**
3.	v C1A odpověděl v posledním sloupci kódy 4, 6 až 9 nebo 0	☐ → **C4**

Nezaměstnaní

C2 Jak dlouho jste nezaměstnaný(-á)?

1. méně než 1 rok ☐ počet měsíců ☐☐
2. 1 rok a déle ☐

C3 Jste evidován(a) na úřadu práce a pobíráte podporu v nezaměstnanosti?

1. jsem evidován(a) a pobírám podporu ☐
2. jsem evidován(a) a nepobírám podporu ☐
3. nejsem evidován(a) ☐

Nepracující

C4 Vykonával(a) jste během posledních 7 dní nějakou placenou práci nebo podnikatelskou činnost, i kdyby se jednalo pouze o jednu hodinu?

1. ano, pravidelnou činnost ☐
2. ano, jednorázovou činnost ☐
3. ne ☐

C5 Hledal(a) jste v posledních 4 týdnech práci?

1. ano ☐
2. ne, vyřizuji založení vlastní firmy ☐
3. ne, mám již sjednané zaměstnání a nejpozději do 3 měsíců do něj nastoupím ☐ → **C8**
4. ne, z jiných důvodů ☐

C6 Jakým způsobem jste v uplynulých 4 týdnech hledal(a) práci?

		ano	ne
1.	nabídka pracovních míst na úřadu práce	☐	☐
2.	žádost o místo přímo u zaměstnavatele	☐	☐
3.	pomoc zprostředkovatelské agentury	☐	☐
4.	dotaz u příbuzných a přátel	☐	☐
5.	inzerce v novinách nebo na internetu	☐	☐
6.	účast na pohovoru, zkouškách či testu	☐	☐
7.	čekání na výsledek žádosti o místo nebo na výsledek konkurzu	☐	☐
8.	jiný aktivní způsob hledání	☐	☐

C7 Jste připraven(a) nastoupit do zaměstnání během 2 týdnů?

1. ano ☐ 2. ne ☐

C8 Byl(a) jste už někdy zaměstnán(a) nebo jste podnikal(a)? (Mělo by se jednat o pravidelnou činnost trvající alespoň 6 měsíců.)

1. ano ☐
2. pracuji ve svém prvním zaměstnání ☐ → **C13**
3. ne ☐ → **C26**

Bývalé hlavní zaměstnání

C9 Jaké bylo Vaše poslední ukončené hlavní zaměstnání (profese)? Popište prosím co nejpodrobněji práci, kterou jste vykonával(a).

→ dvoumístný kód KZAM: ☐☐

C10 Jaké postavení jste měl(a) ve svém posledním hlavním zaměstnání?

1. zaměstnanec ☐
2. společník, jednatel s.r.o. ☐
3. osoba samostatně výdělečně činná se zaměstnanci ☐ → **C13**
4. osoba samostatně výdělečně činná bez zaměstnanců ☐
5. pomáhající rodinný příslušník ☐

C11 Jaký typ pracovní smlouvy jste měl(a) ve svém posledním hlavním zaměstnání?

1. smlouva na dobu neurčitou ☐
2. smlouva na dobu určitou ☐
3. dohoda o pracovní činnosti nebo o provedení práce ☐
4. práce bez smlouvy ☐

C12 Měl(a) jste ve svém posledním hlavním zaměstnání nějaké podřízené zaměstnance?

1. ano ☐ 2. ne ☐

C13 Kdy jste nastoupil(a) do svého prvního řádného zaměstnání, popř. začal(a) podnikat?

první zaměstnání v roce ☐☐☐☐

C14 Kolik let jste od té doby odpracoval(a)?

počet roků ☐☐

TAZATEL: *Pokud respondent odpověděl v otázce C4 kódem 2 (ano, jednorázovou činnost) nebo kódem 3 (ne), přejděte na otázku C23, jinak pokračujte následující otázkou.*

CK2. Kontrolní součet za otázky C2, C9, C13 a C14 ☐☐☐☐

Současné hlavní zaměstnání

C15 Jaké je Vaše současné hlavní zaměstnání (profese)? Popište prosím co nejpodrobněji práci, kterou vykonáváte. (Pokud máte více zaměstnání, popište to, ve kterém odpracujete nejvíce hodin.)

→ dvoumístný kód KZAM:

C16 Popište hlavní činnost místní jednotky firmy nebo organizace, kde pracujete.

→ dvoumístný kód CZ-NACE:

C17 Kolik lidí má místní jednotka firmy nebo organizace, v níž pracujete? (Do počtu zahrňte také sebe.)

A.	přesný počet osob, pokud je od 1 do 10	
B.	11 – 19 osob	☐
C.	20 – 49 osob	☐
D.	50 a více osob	☐
E.	nevím přesně, ale max. 10 osob	☐
F.	nevím přesně, ale více než 10 osob	☐

C18 Kolik hodin týdně obvykle odpracujete ve svém hlavním zaměstnání/podnikání nebo při práci pro rodinnou firmu? (U zaměstnanců se zahrne i neplacená práce přesčas.)

obvyklý počet hodin týdně

C19 Jaké postavení máte ve svém současném hlavním zaměstnání?

1.	zaměstnanec	☐
2.	společník, jednatel s.r.o.	☐
3.	osoba samostatně výdělečně činná se zaměstnanci	☐
4.	osoba samostatně výdělečně činná bez zaměstnanců	☐ → **C22**
5.	pomáhající rodinný příslušník	☐

CI Identifikační údaje

1.	Území	
2.	Pagina	
3.	IČ osoby	
CIK.	Kontrolní součet	

C20 Jaký typ pracovní smlouvy máte ve svém zaměstnání?

1.	smlouva na dobu neurčitou	☐
2.	smlouva na dobu určitou	☐
3.	dohoda o pracovní činnosti nebo o provedení práce	☐
4.	práce bez smlouvy	☐

C21 Máte ve svém současném zaměstnání nějaké podřízené zaměstnance?

1.	ano ☐	2.	ne ☐

C22 Změnil(a) jste od 1.1.2009 své hlavní zaměstnání? Pokud ano, uveďte měsíc a rok ukončení předchozího zaměstnání. (V případě více změn uveďte poslední.)

1.	ano ☐	měsíc	rok
2.	ne ☐ → **C24**		

C23 Proč jste změnil(a) své předchozí zaměstnání, resp. ukončil(a) své poslední zaměstnání?

1.	získání nebo hledání lepšího místa	☐
2.	konec zaměstnání na dobu určitou	☐
3.	donucen(a) k odchodu zaměstnavatelem (propuštění z práce, uzavření firmy atd.)	☐
4.	prodej nebo uzavření vlastní či rodinné firmy	☐
5.	potřeba pečovat o děti nebo o blízkou osobu	☐
6.	přestěhování kvůli zaměstnání partnera nebo z důvodu sňatku	☐
7.	vlastní nemoc nebo invalidita	☐
8.	všechny ostatní důvody (odchod do starobního nebo předčasného důchodu, změna bydliště, další studium, rozhodnutí žít z úspor atd.)	☐

CK3. Kontrolní součet za otázky C15 až C18, C22

Další činnost

TAZATEL: *Osoby, které nemají současné hlavní zaměstnání, pokračují otázkou C26.*

C24 Máte vedle svého hlavního zaměstnání/podnikání ještě nějaké jiné zaměstnání/podnikání?
Pokud ano, o jakou činnost se jedná a kolik hodin týdně v tomto dalším zaměstnání/podnikání obvykle odpracujete?

1.	ano, závislá činnost	☐	počet hodin týdně ☐☐
2.	ano, podnikání	☐	
3.	ano, obojí	☐	
4.	ne	☐	

→

TAZATEL: *Pokud je součet odpracovaných hodin v hlavním a vedlejším zaměstnání (otázky C18 a C24) menší než 30 hodin, pokračujte následující otázkou C25, jinak přejděte na otázku C26.*

C25 Uveďte důvod, proč pracujete méně než 30 hodin týdně.

1.	vlastní nemoc nebo invalidita	☐
2.	chtěl(a) bych pracovat více hodin, ale nemohu najít práci na plný úvazek	☐
3.	studium nebo proškolování	☐
4.	nechci pracovat více hodin	☐
5.	péče o děti nebo jiné osoby	☐
6.	počet hodin za všechna moje zaměstnání je považován za plný úvazek	☐
7.	jiné důvody	☐

PŘÍJMY ZA ROK 2009

Příjmy ze závislé činnosti

C26 Měl(a) jste v minulém kalendářním roce příjem ze závislého pracovního poměru? Vedle příjmu z hlavního pracovního poměru uveďte případný příjem z dalších pracovních poměrů, prací na dohody a ostatních jednorázových a příležitostných prací.

1.	ano	☐	
2.	ne	☐	→ **C33**

C27 Uveďte prosím, kolik činil v minulém kalendářním roce Váš hrubý, resp. čistý příjem ze zaměstnání (dle uvedených položek). Můžete uvést buď pravidelný měsíční příjem (průměrný) nebo celkový roční příjem zahrnující i veškeré příplatky a mimořádné příjmy.

			počet měsíců	Kč	počet měsíců	Kč
hlavní pracovní poměr	hrubý ☐	M	☐☐	☐☐☐☐☐☐	☐☐	☐☐☐☐☐☐
	čistý ☐	R	☐☐	☐☐☐☐☐☐		
další pracovní poměr	hrubý ☐	M	☐☐	☐☐☐☐☐☐	☐☐	☐☐☐☐☐☐
	čistý ☐	R	☐☐	☐☐☐☐☐☐		
dohoda(-y) o provedení práce	hrubý	R		☐☐☐☐☐☐		
dohoda(-y) o pracovní činnosti	hrubý	R		☐☐☐☐☐☐		

C28 Obdržel(a) jste v minulém kalendářním roce některou z následujících položek jako příjem navíc?
- náhrada za přesčasy; 13. a 14. plat; mimořádné odměny, prémie; podíly na výsledku hospodaření firmy, bonusy; příplatek na dovolenou, příspěvek FKSP; odstupné; spropitné; jiné platby (ošatné, diety, provize atd.)

1.	ano	☐	→ **C29**
2.	ne	☐	→ **C30**

C29 Pokud jste nezahrnul(a) některou z těchto položek do výše uvedených příjmů, uveďte prosím hrubou, resp. čistou částku těchto dodatečných plateb za celý minulý kalendářní rok.

☐ hrubá
☐ čistá
částka v Kč ☐☐☐☐☐☐☐

všechny platby jsou již zahrnuty ☐

CK4. Kontrolní součet za otázky C24, C27 a C29

☐☐☐☐☐☐☐☐

Požitky od zaměstnavatele

C30 Poskytoval Vám Váš zaměstnavatel v minulém kalendářním roce automobil, dodávku či jiné motorové vozidlo, které jste mohl(a) využívat i pro soukromé účely? Pokud ano, uveďte kolik měsíců jste jej využíval(a).

1. ano, využíval(a) jsem ☐ počet měsíců
2. neposkytoval / nevyužíval(a) jsem ☐

C31 Poskytoval Vám Váš zaměstnavatel v minulém kalendářním roce příspěvky na stravování? Pokud ano, uveďte prosím počet měsíců, po které jste tyto příspěvky pobíral(a), jejich počet za měsíc a dále jejich hodnotu a cenu, za jakou jste je koupil(a).

1. **závodní stravování** ☐ 2. **stravenky** ☐

počet měsíců | ks/měsíc

hodnota jídla / hodnota stravenky | za cenu

3. **neposkytoval / nevyužíval(a) jsem** ☐

C32 Poskytoval Vám Váš zaměstnavatel některé další výhody a nepeněžní služby, ať už zdarma nebo za částečnou úhradu? Uveďte prosím pouze ty, které jste v minulém kalendářním roce využíval(a).

		ano	ne
1.	mobilní nebo pevný telefon	☐	☐
2.	jazykové kurzy	☐	☐
3.	příspěvek na benzín, dopravu	☐	☐
4.	sleva na firemní zboží či služby	☐	☐
5.	příspěvek na sportovní vyžití, dovolenou	☐	☐
6.	příspěvek na penzijní či životní pojištění	☐	☐
7.	bezúročná půjčka	☐	☐

Příjmy z podnikání a jiné samostatně výdělečné činnosti

C33 Měl(a) jste v minulém kalendářním roce nějaké příjmy ze živnostenského podnikání, zemědělské výroby, z podnikání podle zvláštních předpisů (lékaři, advokáti), z výkonu nezávislého povolání (profesionální sportovci, umělci, tlumočníci), z podílu společníků veřejné obchodní společnosti a komplementářů komanditní společnosti nebo příjmy z autorských práv?

1. ano ☐ → **C35**
2. ano, jako spolupracující osoba ☐
3. ne ☐ → **C38**

C34 Uveďte prosím částku rozepsanou na Vaši osobu jako podíl z úhrnné společné částky.

částka v Kč → **C38**

C35 Jaký byl Váš výsledek hospodaření (tj. rozdíl mezi příjmy a výdaji) za minulý kalendářní rok? Vyberte prosím některou z následujících možností pro vyčíslení Vašich příjmů z podnikání.

1. **daňové přiznání** (ř. 113) - hrubý zisk/ztráta v Kč snížený o případný podíl přerozdělený na spolupracující osoby
2. **přehled o příjmech a výdajích OSVČ za rok 2009** (ř. 24) - rozdíl příjmů a výdajů vykázaný pro ČSSZ
3. **roční účetní závěrka** - hospodářský výsledek
4. **žádný dokument** - vlastní odhad zisku/ztráty v Kč — hrubý / čistý

C36 Vybíral(a) jste opakovaně peněžní prostředky z příjmů z Vašeho podnikání na soukromé účely, jako např. na provoz domácnosti či uspokojování běžných osobních potřeb členů domácnosti? Pokud ano, uveďte prosím počet měsíců a pravidelné měsíční částky.

1. ano ☐ 2. ne ☐

počet měsíců | měsíční částka v Kč

C37 Použil(a) jste nějaké peněžní prostředky z příjmů z Vašeho podnikání na jednorázové výdaje typu dovolená, školné, zařízení domácnosti, rekonstrukce bytu atd.? Pokud ano, uveďte prosím součet všech částek použitých pro domácnost za celý minulý kalendářní rok.

1. ano ☐ 2. ne ☐

celková roční částka v Kč

C38 Měl(a) jste v minulém kalendářním roce příjmy za příspěvky do novin, časopisů, rozhlasu nebo televize jako drobné autorské honoráře nepřesahující 7 000 Kč měsíčně od téhož plátce? Pokud ano, uveďte prosím jejich roční úhrn.

1. ano ☐ 2. ne ☐

celková roční částka v Kč

CK5. Kontrolní součet za otázky C30, C31, C34 až C38

H

Ostatní příjmy

C39 Měl(a) jste v minulém kalendářním roce některé z následujících příjmů, ať už pravidelných nebo jednorázových? Uveďte prosím vždy čistý celoroční příjem.

		neměl(a)	roční částka v Kč
1.	**příjmy z kapitálového majetku** úroky z účtů v bance, z vkladů a vkladových listů; podíly na zisku kapitálových společností a družstev, dividendy z akcií; výnosy z cenných papírů, dluhopisy, obligace; příjmy ze zdrojů v zahraničí	☐	
2.	**příjmy z prodeje** příležitostná domácí samovýroba; prodej zemědělských výrobků jako přebytků z osobního hospodářství	☐	
3.	**příjmy ze životního pojištění** pojistné plnění pro případ dožití určitého věku nebo v případě úmrtí jiné osoby	☐	
4.	**příjmy z neživotního pojištění** plnění z jiných druhů pojištění (úrazové, nemovitosti, domácnosti, motor. vozidel atd.)	☐	
5.	**stipendia, kapesné učňů** prospěchové, sociální i ubytovací stipendium	☐	
6.	**jiné příjmy** náhrady majetkových křivd (restituce; totální nasazení, PTP); výhry z loterií, sázek, sportovních a jiných soutěží; dědictví, odstupné za uvolnění bytu; státní podpora a úroky ze stavebního spoření (pouze v případě jednorázové výplaty)	☐	
	C39K. Kontrolní součet za tab. C39		

Sociální příjmy

+

C40 Pobíral(a) jste v minulém kalendářním roce některý z uvedených druhů sociálních příspěvků a dávek? Uveďte prosím počet měsíců pobírání a měsíční částku. U vybraných dávek uveďte také počet dní nemoci, resp. OČR, pokud se jednalo o dobu kratší než 1 měsíc.

			nepobíral(a)	počet měsíců	Kč	počet měsíců	Kč
1.	podpora v nezaměstnanosti	rekvalifikace ☐	☐				
2.	rodičovský příspěvek		☐				
3.	příspěvek na péči		☐				
4.	výsluhový příspěvek, odchodné		☐				
	dávky nemocenského pojištění						
5.	nemocenské	počet dní nemoci	☐				
6.	ošetřovné	počet dní OČR	☐				
7.	peněžitá pomoc v mateřství		☐				
8.	vyrovnávací příspěvek v těhotenství a mateřství		☐				
	dávky důchodového pojištění						
9.	starobní důchod		☐				
10.	vdovský/vdovecký důchod		☐				
11.	částečný invalidní důchod		☐				
12.	plný invalidní důchod		☐				
13.	sirotčí důchod	počet dětí	☐				
14.	příspěvky pro těžce zdravotně postižené občany (na opatření zvláštních pomůcek, na úpravu bytu, na zakoupení a provoz motorového vozidla atd.)		☐				
	C40K. Kontrolní součet za tab. C40						

Daně z příjmu

C41 Měl(a) jste v minulém kalendářním roce nějaké zdanitelné příjmy, ze kterých Vy resp. Váš zaměstnavatel odvádíte daň?

1. ano ☐ | 2. ne ☐ → **C43**

C42 Uplatňujete za rok 2009 u svého zaměstnavatele nebo ve svém daňovém přiznání nárok na následující úlevy na dani?

1.	na vyživované dítě	počet dětí ☐☐
2.	na vyživované dítě ZTP-P	počet dětí ☐☐
3.	na manželku/manžela	☐
4.	na manželku/manžela, který je držitelem ZTP-P	☐
5.	na poživatele částečného invalidního důchodu	☐
6.	na poživatele plného invalidního důchodu	☐
7.	na držitele ZTP-P	☐
8.	na studium	☐
9.	hodnota daru	☐☐☐☐☐☐☐
10.	odečet úroků	☐☐☐☐☐☐☐
11.	penzijní připojištění	☐☐☐☐☐☐☐
12.	životní pojištění	☐☐☐☐☐☐☐
13.	členský příspěvek odborů	☐☐☐☐☐☐☐
14.	odpočet ztráty	☐☐☐☐☐☐☐
15.	další položky podle §34	☐☐☐☐☐☐☐

Penzijní připojištění

C43 Platil(a) jste si v minulém kalendářním roce příspěvky na penzijní připojištění?
Pokud ano, uveďte prosím počet měsíců a měsíční částku Vašich (vlastních) příspěvků v minulém kalendářním roce.

1. ano ☐ | 2. ne ☐

počet měsíců ☐☐ | měsíční částka v Kč ☐☐☐☐☐

C44 Přispíval Vám v minulém kalendářním roce Váš zaměstnavatel na soukromé penzijní připojištění?
Pokud ano, uveďte prosím počet měsíců a měsíční částku těchto příspěvků od zaměstnavatele.

1. ano ☐ | 2. ne ☐

počet měsíců ☐☐ | měsíční částka v Kč ☐☐☐☐☐

C45 Pobíral(a) jste v minulém kalendářním roce pravidelný důchod z Vašeho penzijního připojištění?
Pokud ano, uveďte prosím počet měsíců a měsíční částku pobíraného důchodu.

1. ano ☐ | 2. ne ☐

počet měsíců ☐☐ | měsíční částka v Kč ☐☐☐☐☐

Biografické informace

C46 Ve které zemi jste se narodil(a)?

1. ČR (území dnešní ČR) ☐
2. jiná země: ☐☐

C47 Jaká je Vaše státní příslušnost? Pokud máte dvojí občanství, uveďte prosím obě.

1. státní příslušnost 1:
   - ČR ☐
   - jiná země: ☐☐
2. státní příslušnost 2:
   - země: ☐☐

C48 Jestliže jste pobýval(a) dlouhodobě mimo ČR, uveďte prosím rok, ve kterém jste se do ČR přistěhoval(a) nebo vrátil(a).

1. rok přistěhování ☐☐☐☐ | 2. netýká se mě ☐

Modul 2010

TAZATEL: *Žijí v domácnosti alespoň 2 osoby narozené 1993 a dříve?*

1. ano ☐ | 2. ne ☐ → **C56**

C49 S jakou částí Vašeho příjmu hospodaříte sám/sama odděleně od společných příjmů Vaší domácnosti?

1. s celým svým příjmem ☐
2. s více než polovinou mého příjmu ☐
3. zhruba s polovinou mého příjmu ☐
4. s méně než polovinou mého příjmu ☐
5. celý můj příjem je společný ☐
6. nemám příjem ☐

C50 Můžete utrácet peníze pro své potřeby (např. za oblečení, obuv, své koníčky, sportovní potřeby, lístek do kina, knihy atd.), aniž byste to musel(a) s někým předem konzultovat?

1. ano mohu, vždy nebo téměř vždy ☐
2. ano mohu, někdy ☐
3. nemohu nikdy nebo téměř nikdy ☐

C51 Máte přístup k bankovnímu účtu, se kterým můžete disponovat - vybírat, vkládat peníze, uskutečňovat převody, apod.? (Nemusí jít o účet na Vaše jméno, stačí mít k účtu podpisový vzor.)

1. ano ☐ | 2. ne ☐

C52 Kolik let žijete a hospodaříte se svým partnerem/manželem, svojí partnerkou/manželkou ve společné domácnosti?

1. počet let ☐☐
2. nemám v domácnosti partnera/partnerku, manžela/manželku ☐ → **C55**

CK7. Kontrolní součet za otázky C42 až C48, C52 ☐☐☐☐☐☐☐☐

H

L

C53 V následujících otázkách uvažujte sebe a Vašeho partnera / Vaši partnerku.					
Uveďte prosím, kdo z vás většinou rozhoduje:	spíše já	oba dva stejně	spíše partner/ka	situace ještě nenastala	nemáme úspory
1. o výši vydání na každodenní nákupy	☐	☐	☐		
2. při koupi nákladnějšího zboží dlouhodobé spotřeby (lednička, pračka, televize), popř. nábytku	☐	☐	☐	☐	
3. zda si domácnost půjčí peníze, požádá o hypotéku, nakoupí zboží na spotřebitelský úvěr nebo využije jiné půjčky	☐	☐	☐	☐	
4. zda si domácnost vybere a utratí část společných úspor	☐	☐	☐	☐	☐
5. při důležitých rozhodnutích (nejen finančních, ale i životních, pracovních, aj.)	☐	☐	☐		

C54 Kdo z vás většinou rozhoduje o nákladnějších výdajích pro vaše děti (vydání na vzdělání, kroužky, sportovní potřeby, apod.)?	
1. spíše já	☐
2. oba dva stejně	☐
3. spíše partner/ka	☐
4. netýká se nás	☐ → **C56**

C55 Můžete utrácet peníze pro potřeby dětí, které žijí ve Vaší domácnosti, dávat jim kapesné apod., aniž byste to musel(a) s někým předem konzultovat?	
1. ano mohu, vždy nebo téměř vždy	☐
2. ano mohu, někdy	☐
3. nemohu nikdy nebo téměř nikdy	☐
4. netýká se mě	☐

Zdraví

C56 Jak celkově hodnotíte svůj zdravotní stav?			
1. velmi dobrý	☐	4. špatný	☐
2. dobrý	☐	5. velmi špatný	☐
3. přijatelný	☐		

C57 Máte nějakou dlouhodobou nemoc nebo dlouhodobý zdravotní problém? (Problém, který již trvá nebo bude trvat 6 a více měsíců.)			
1. ano	☐	2. ne	☐

C58 Byl(a) jste kvůli zdravotním problémům nejméně po dobu posledních 6 měsíců omezen(a) v činnostech, které lidé obvykle dělají?	
1. ano, velmi omezen(a)	☐
2. ano, omezen(a)	☐
3. neomezen(a)	☐

C59 Kolikrát jste za posledních 12 měsíců navštívil(a) praktického lékaře nebo specialistu s výjimkou zubaře a očního lékaře?	
počet návštěv	☐☐

C60 Potřeboval(a) jste během posledních 12 měsíců navštívit zubaře, a přesto jste k němu nešel(-a)?	
1. ano, taková situace minimálně jednou nastala a k zubaři jsem nešel(-a)	☐ ↓
2. ne, taková situace nenastala	☐ → **C62**

C61 Proč jste k zubaři nešel(-a)? Uveďte hlavní důvod.	
1. nemohl(a) jsem si to dovolit (příliš drahé, nehradí pojišťovna)	☐
2. čekací seznamy, nutnost objednávat se ve velkém předstihu	☐
3. nemohl(a) jsem se uvolnit z práce nebo od péče o děti či jinou osobu	☐
4. daleké cestování, nebyl vhodný způsob dopravy	☐
5. strach ze zubaře, vyšetření, léčby apod.	☐
6. chtěl(a) jsem počkat, zda se zdravotní problém sám nezlepší	☐
7. neznal(a) jsem žádného dobrého zubaře	☐
8. z jiného důvodu	☐

C62 Potřeboval(a) jste během posledních 12 měsíců navštívit lékaře (kromě zubaře), a přesto jste k němu nešel(-a)?	
1. ano, taková situace minimálně jednou nastala a k lékaři jsem nešel(-a)	☐ ↓
2. ne, taková situace nenastala	☐ → **konec**

C63 Proč jste k lékaři nešel(-a)? Uveďte hlavní důvod.	
1. nemohl(a) jsem si to dovolit (příliš drahé, nehradí pojišťovna)	☐
2. čekací seznamy, nutnost objednávat se ve velkém předstihu	☐
3. nemohl(a) jsem se uvolnit z práce nebo od péče o děti či jinou osobu	☐
4. daleké cestování, nebyl vhodný způsob dopravy	☐
5. strach z lékařů, nemocnic, vyšetření, léčby apod.	☐
6. chtěl(a) jsem počkat, zda se zdravotní problém sám nezlepší	☐
7. neznal(a) jsem žádného dobrého lékaře	☐
8. z jiného důvodu	☐

Mnohokrát děkujeme za účast na tomto rozhovoru