VÚPSV rilsa
VÝZKUMNÝ ÚSTAV PRÁCE A SOCIÁLNÍCH VĚCÍ
RESEARCH INSTITUTE FOR LABOUR AND SOCIAL AFFAIRS

Report o úpravě datových souborů pro projekt „Vliv minimální mzdy na trh práce v ČR“

Filip Červenka
Vlastimil Beran
Diana Bílková

VÚPSV, v. v. i., Praha
2021

Studie je výstupem projektu *„Vliv minimální mzdy na trh práce v ČR"*, který je řešen v rámci Dlouhodobé koncepce rozvoje výzkumné organizace na období 2018 až 2022 schválené MPSV na základě Rozhodnutí č. 7-RVO/2018 o poskytnutí institucionální podpory na dlouhodobý koncepční rozvoj výzkumné organizace.

Vydal Výzkumný ústav práce a sociálních věcí, v. v. i.
Dělnická 213/12, Praha 7, 170 00
Vyšlo v roce 2021, 1. vydání, počet stran 21
Tisk: VÚPSV, v. v. i.

https://www.vupsv.cz

Abstrakt

Cílem „Reportu o úpravě dat" je představit základní kontroly a úpravy datových souborů ISPV, které byly nezbytným předpokladem pro využití těchto dat k účelům dalšího analytického zkoumání. Report popisuje strukturu a obsah dat ISPV za mzdovou a platovou sféru v rozmezí let 2014 až 2020 – popisuje aplikované kontroly, zjištěnou chybovost, příslušná navržená řešení a další nutné dopočty a doplnění. Provedené kontroly se zaměřovaly na identifikaci neplatných hodnot podle dostupných oficiálních číselníků a podle logicky přípustných hodnot.

Bylo zjištěno, že data ISPV vykazují poměrně vysokou chybovost projevující se u některých proměnných neplatnými hodnotami v řádech desítek až stovek tisíc záznamů v jednotlivých obdobích. Zároveň bylo zjištěno, že struktura chybovosti v datech mzdové sféry se znatelně odlišuje mezi staršími obdobími let 2014 až 2015 a novějšími obdobími let 2016 až 2020. Tento předěl naznačuje možnou změnu v přístupu k úpravě dat ze strany dodavatele. Nicméně popis úpravy dat ze strany dodavatele není veřejně dostupnou informací, zároveň není součástí ani přílohou datových souborů. Tento fakt zásadním způsobem ztěžoval práci na opravě chybových hodnot. Vedle provedených kontrol byly rovněž dopočítány a doplněny proměnné pro hrubé měsíční mzdy a platy a pro přepočtené placené měsíce, jelikož tyto proměnné nejsou součástí původních souborů. Dále byly doplněny specifické umělé proměnné, které umožňují filtrovat data podle zadaných kritérií vzešlých z odborné diskuse. Popsané kroky byly nezbytným předpokladem pro další práci s popsanými daty.

Klíčová slova: mzdová sféra; platová sféra; data; úprava dat; chybovost.

Abstract

The aim of the "Data Modification Report" is to present the basic checks and modifications of the ISPV data files that formed a necessary prerequisite for the use of such data for the purposes of the subsequent analytical investigation. The report provides a description of the structure and content of the ISPV data on the wage and salary sphere between 2014 and 2020, i.e. the checks applied, the detected error rate, the relevant proposed solutions and other necessary additions and supplements. The checks performed focused on the identification of invalid values according to the official numbers available and logically permissible values.

It was found that the ISPV data evinced a relatively high error rate that was manifested for certain variables by invalid values in the order of tens to hundreds of thousands of records over various periods of time. At the same time, it was found that the structure of the error rate in the wage data differed markedly between the older periods of 2014 to 2015 and the newer periods of 2016 to 2020. This divide indicates a change in terms of the approach to the editing of the data by the supplier. However, the description of the modification of the data by the supplier is not publicly available information, nor does it form a part of, or attachment to, the data files. This fact significantly complicated the work involved in the correction of the error values. In addition to the checks outlined above, variables concerning gross monthly wages and salaries and for recalculated paid months were calculated and supplemented since these variables did not form a part of the original files. Furthermore, specific artificial variables were added, which allowed for the filtering of the data according to specified criteria that were proposed

during the expert discussion. The various procedures described above formed necessary prerequisites for subsequent research using the described data.

Key words: wage sphere; salary sphere; data; data modification; error rate.

Obsah

Úvod

Tento report popisuje přístup výzkumného týmu k úpravě dat. V první části textu jsou charakterizována dostupná data a provedené kontroly. Následně jsou popsány odhalené chyby a zvolený přístup k jejich řešení. Poslední část reportu se věnuje úpravám a doplněním datových souborů, které byly nezbytné před přistoupením k samotné analýze.

1. Základní přehled

Výzkumný tým disponuje daty pro mzdovou a platovou sféru pro období let 2014 až 2020. Záznamy obsahují osobní i pracovní charakteristiky zaměstnanců a rovněž jejich charakteristiky základní. Jednotlivé proměnné jsou však definovány pro mzdovou a platovou sféru odlišně – kontrola a úprava dat si proto vyžádala v obou případech specifický přístup.

Data jsou dostupná vždy za pololetí a za celý rok. Mzdová sféra je dále rozdělena na soubory typu MI (údaje zaměstnavatelů) a soubory typu MP (údaje zaměstnanců). Celkově se jedná o 42 samostatných souborů.

Rozsah hlavních datových souborů mzdové sféry se pohybuje mezi 1,5 až 1,7 milióny řádků, což vylučuje použití programů s nižší maximální kapacitou, jakým je například Excel. K úpravě dat byl proto zvolen program MS Access.

Dále je třeba upozornit na fakt, že dostupná data nelze považovat za zcela reprezentativní vzorek, jelikož firmy různých velikosti jsou zastoupeny disproporcionálně. V případě mzdové sféry a sektoru nefinančních podniků a domácností (podnikatelů – fyzických osob) jsou ze 100 % obsaženy podniky, které mají alespoň 250 zaměstnanců a více. Menší podniky jsou podreprezentovány. Středně velké podniky v rozmezí 50 až 249 zaměstnanců jsou zastoupeny pouze z 15 % a podniky v rozmezí 10 až 49 zaměstnanců dokonce jen ze 4,5 %. Tyto výběry se "obměňují metodou rotujícího panelu s periodou rotace devět let – to znamená, že k plné obměně dojde za devět let". Nejmenší podniky, které zaměstnávají od 1 do 9 zaměstnanců, jsou zastoupeny z 1,2 %. Sběr dat u těchto podniků je prováděn pouze jednou za 4 roky. V oblasti finančních a vládních institucí jsou do výběru zahrnuty všechny subjekty s více než 50 zaměstnanci. Menší subjekty v těchto sektorech do výběru nevstupují (ISPV, 2020).

2. Provedené kontroly

Na soubory byly rámcově aplikovány dva druhy kontrol: kontroly pro číselníky a kontroly logických chyb. Rozsah i typ kontrol se v jednotlivých souborech lišil, jelikož soubory z platové a mzdové sféry jsou dostupné v odlišné formě. Rámcově sice zachycují stejné nebo obdobné informace, obsahují však rozdílně označené a především také jinak definované proměnné. Výskyt proměnné, která je stejně definována pro mzdovou i platovou sféru, je spíše výjimkou (např. okres: AA0179 – MISTOVP, nebo státní občanství AA0033 – STOBC). Odlišnosti mezi platovou a mzdovou sférou proces kontroly znatelně komplikovaly.

Následující podkapitoly detailně popisují, jak byly ve mzdové a platové sféře všechny kontroly nastaveny a následně provedeny.

2.1 Provedené kontroly v datech mzdové sféry

Zdrojem informací o specifikaci jednotlivých proměnných v rámci mzdové sféry byla oficiální příručka společnosti Trexima, která pro MPSV statistické zjišťování zpracovává (Trexima, 2020). Tato příručka obsahuje základní informace o datových souborech včetně úplných definic proměnných. Nicméně neobsahuje kompletní číselníky, které bylo potřeba doplnit z jiných zdrojů. Pro proměnnou VZDELANI byl číselník získán z webových stránek Ministerstva školství, mládeže a tělovýchovy (MŠMT, 2021), pro proměnnou CZ-ISCO byl číselník získán z webových stránek Českého statistického úřadu (ČSÚ, 2020a), stejně jako pro proměnnou CZ-ICSE (ČSÚ, 2014a), LAU1 (ČSÚ, 2021a) a STOBC (ČSÚ, 2020b).

Kontrola číselníků se v každém pololetí mzdové sféry týkala 10 proměnných (LAU1, POHLAVI, STOBC, VZDELANI, INVALID, MISTOVP, CZICSE, CZISCO, VEDOUCI a KONTOPD) a všechny záznamy, které obsahovaly jiné než povolené hodnoty, byly evidovány.

Zatímco kontroly na číselníky byly jasně ohraničené daným typem proměnné, u logických kontrol tomu tak nebylo. Veškeré provedené logické kontroly byly proto odborně prodiskutovány, přičemž klíčové bylo dodržet vlastní vnitřní logiku každé proměnné. V souborech mzdové sféry se jednalo v každém pololetí o 7 proměnných (ROKNAR, DOBAZAM, EVIDDNY, KONECEP, ODPRACD, MZDA a FONDJE).

Konkrétně se zjišťovalo, zda jsou splněny následující podmínky:

a) hodnoty ROKNAR (roku narození) patří do intervalu <T-85, T-15>;

b) hodnoty DOBAZAM (doba zaměstnání u současného zaměstnavatele v letech) patří do intervalu (0, 70);

c) hodnoty EVIDDNY (počet kalendářních dnů od počátku roku do konce sledovaného období, ve kterých byl zaměstnanec v evidenčním počtu zaměstnanců) patří do intervalu (0, celkový počet kalendářních dnů ve sledovaném období, tj. 181/182 a 365/366);

d) hodnoty KONECEP (poslední den, kdy byl zaměstnanec v evidenčním počtu zaměstnanců) jsou buď rovny nule (= zaměstnanec byl v evidenčním počtu po celou dobu), nebo se jedná o data ve formátu RRRR-MM-DD, která spadají do sledovaného období;

e) hodnoty MZDA (mzda zúčtovaná v jednotlivých měsících od počátku roku do konce sledovaného období) jsou >0;

f) hodnoty ODPRACD (odpracovaná doba ve sledovaném období v hodinách) patří pro pololetní data do intervalu (0, 1 270) a pro roční data do intervalu (0, 2 230);[1]

g) hodnoty FONDSJE s přičtením hodnot PRESCAS a odečtením hodnot ABSCELK se rovnají hodnotám ODPRACD (a následně byly označeny údaje, které vykazovaly rozdíl větší než +/-5 % maximální povolené hodnoty ODPRACD).

2.2 Provedené kontroly v datech platové sféry

Kontrola pro číselníky v platové sféře probíhala obdobně jako u mzdové sféry, s tím rozdílem, že se přirozeně vyskytovaly proměnné, které u mzdové sféry neevidujeme (platová třída, platový stupeň, druh platového tarifu apod.). Rozdíl byl rovněž v tom, že některé proměnné, které byly významově shodné s proměnnými ze mzdové sféry, byly popsány více či méně odlišným číselníkem (stupeň invalidity, kód zaměstnání, pohlaví apod.) – jejich platnost bylo tudíž potřeba ověřovat.

Údaje o specifikaci proměnných v platové sféře byly získávány z databáze Informačního systému o datových prvcích Ministerstva vnitra (ISPD, 2006–2018). U značné části proměnných číselník (či seznam povolených hodnot) ale uveden nebyl a externí odkaz byl nefunkční. V těchto případech bylo, obdobně jako u mzdové sféry, potřeba dohledávat číselníky individuálně z jiných zdrojů. Pro proměnnou AA0013 – kód zaměstnání byl číselník získán z Českého statistického úřadu (ČSÚ, 2014b), stejně jako pro proměnné AA0019 (ČSÚ, 2021b), AA0690 – místo obvyklého výkonu práce, AA0179 – kód NUTS okresu (ČSÚ, 2021a), AA0033 (ČSÚ 2020b) a AA0091 (ČSÚ, 2014a). Číselník pro proměnnou AA0223 – stupeň nejvyššího dosaženého vzdělání byl získán z webových stránek Ministerstva školství, mládeže a tělovýchovy (MŠMT, 2021).

Kontrolou pro číselníky bylo v každém souboru platové sféry kontrolováno 20 proměnných (AA0019, AA0085, AA0179, AA0676, AA0013, AA0033, AA0091, AA0125, AA0127, AA0128, AA0129, AA0130, AA0136, AA0137, AA0211, AA0223, AA0690, AA1586, AA1591, AA1594).

Na všechny soubory platové sféry byly také aplikovány logické kontroly u dalších šesti proměnných (AA0226, ROK, AA0110, AA0225, AA0231, AA0568).

Konkrétně se zjišťovalo, zda jsou splněny následující podmínky:

a) hodnoty AA0226 (sledované období v měsících) splňují povolený formát (1-6 / 1-12) a svým výskytem odpovídají pololetním/ročním souborům;

b) hodnoty ROK vyjadřují odpovídající rok ve formátu RRRR;

c) hodnoty AA0110 (datum sběru dat) nenabývaly hodnot značících dřívější data předcházející období, která daná data popisují;

d) hodnoty AA0225 (doba zaměstnání v letech u současného zaměstnavatele) patří do intervalu (0, 70);

e) hodnoty AA0231 (rok – ročník narození) patří do intervalu <T-85, T-15>;

[1] Hranice pro první pololetí vznikla jako součet šesti odpracovaných měsíců (6*160 h), zákoníkem práce daného maxima 150 h povolených přesčasů a jednoho dodatečného měsíce (+160 h) pro případy, kdy si zaměstnanec nadpracuje hodiny a později si vybírá náhradní volno. Hranice pro celý rok byla stanovena obdobně s rozdílem v součtu odpracovaných měsíců (12*160).

f) hodnoty AA0568 (datum vynětí z evidenčního stavu) jsou buď prázdné (= zaměstnanec byl v evidenčním stavu po celou dobu), nebo se jedná o datum ve formátu RRRR-MM-DD, které spadá do sledovaného období.

Mimo kontroly pro číselníky a logické kontroly byly na platovou sféru aplikovány ještě kontroly formátu některých proměnných. V každém souboru platové sféry šlo o 5 proměnných (AA0115, AA0116, AA0118, AA0119, AA0123), které mají nabývat pouze formátu celého nezáporného čísla.

Provedené kontroly odhalily jak ve mzdové, tak i v platové sféře řadu chyb, které jsou detailně popsány v následující kapitole.

3. Odhalené chyby

Z hlediska chybovosti lze rozlišit tři části celého datového souboru: první roky mzdové sféry (do roku 2015), následující roky mzdové sféry (od roku 2016) a data platové sféry jako taková.

V datech prvních dvou let sledovaného období mzdové sféry evidujeme chyby u většího množství proměnných. U pozdějších dat jsou některé proměnné, které byly dříve chybové, zcela v pořádku. Jiné proměnné však vykazují výrazně větší chybovost, která je zároveň tvořena menším množstvím unikátních chyb. Při celkovém pohledu na data se nelze ubránit dojmu, že starší data byla ponechána bez větších úprav, zatímco data novější byla před jejich zasláním upravena – nelze však jednoznačně říci, zda to bylo ku prospěchu věci. K tomuto závěru přispívá i pohled na nejvyšší mzdy. Maxima měsíčních mezd u prvních dvou let se pohybují kolem 4 miliónů korun měsíčně (což odpovídá i veřejně dostupným informacím o mzdách nejvyššího managementu některých českých firem), zatímco v novějších datech se nejvyšší mzdy pohybují jen kolem 1,5 miliónu korun měsíčně. Z toho vyplývá nejen samotný velký rozdíl, ale také výrazný rozpor mezi novějšími daty a veřejně známými a dostupnými informacemi.

V datech za platovou sféru je chybovost ve starších i novějších datech víceméně konzistentní a není v nich patrný žádný předěl, jak tomu je u sféry mzdové.

Ke každé z vyjmenovaných částí bylo potřeba přistupovat individuálně. Jejich chybovost je podrobně popsána v níže uvedených podkapitolách.

3.1 První část mzdové sféry

V prvních dvou letech data mzdové sféry obsahují až na jednu výjimku všechny kontrolované proměnné chybné záznamy. U některých proměnných se jedná jen o několik desítek chybných záznamů (např. INVALID, KONTOPD, VEDOUCI), u dalších proměnných však chybné záznamy šly až do řádů stovek (např. ROKNAR, POHLAVI, STOBC). Nicméně vzhledem k velikosti souboru se stále jedná o nízké množství chybných záznamů, které tak pro analýzu dat nepředstavují větší ohrožení.

Za zmínku však stojí forma těchto chyb. Například pro ROKNAR (rok narození) se vyskytují hodnoty jako “0”, “19”, “985”, “1687”, “1856”, “1899” nebo “2055” či “4947”. Pro POHLAVI (pohlaví osoby) se vyjma povolených znaků M – Z (respektive 1 – 2) vyskytuje řada nepovolených kódů, například "×", "—", "Ä" nebo "N". Obdobně i STOBC (státní občanství) obsahuje podobné znaky, které nejsou součástí číselníku, například "--", "╚e", "╚R", nebo "B╠". Řada ostatních objevených chyb představuje zřejmé překlepy a dobře identifikovatelné chyby. Nemalá část neplatných záznamů se však vyskytuje v podobě zvláštních klávesových znaků nebo hodnot nejen nelogických, ale zcela absurdních, jak je ostatně ilustrováno výše.

U dalších proměnných už je situace s chybovostí vážnější – svou četností totiž dosahují řádově tisíců, desetitisíců i statisíců chybných záznamů (např. MISTOVP, CZ-ISCO, VZDELANI nebo FONDJSE).

Dalším rozměrem chybovosti je také fakt, že unikátní chyby se objevují v řádu desítek až stovek, a to u proměnných, jejichž číselníky mají definované přípustné hodnoty v řádu jednotek, výjimečně desítek. Jinými slovy to znamená, že počet unikátních chyb místy i několikanásobně převyšuje rozsah číselníku. Například pro proměnnou MISTOVP v roce 2014 bylo identifikováno 616 unikátních chybných kódů, byť samotný číselník má jen 78 položek. Obdobně i u dalších proměnných jsou unikátní chyby natolik rozmanité a nahodilé, že svým rozsahem výrazně převyšují samotný rozsah číselníku.

3.2 Druhá část mzdové sféry

Data za mzdovou sféru se od roku 2016 značně mění. Řada kontrolovaných proměnných již nevykazuje žádné chyby. Zcela bez chyb jsou například proměnné POHLAVI, INVALID, KONTOPD, EVIDDNY a rovněž poměrně důležité MISTOVP. Výrazně také klesá počet unikátních chyb. Na druhou stranu ale neklesá celkové množství chyb. Stagnace je například u proměnných CZICSE či VEDOUCI. U některých proměnných, například STOBC, VZDELANI, CZISCO a DOBAZAM, však dochází k jejich extrémnímu nárůstu.

Největším příkladem této změny je proměnná VZDELANI. V letech 2014 a 2015 bylo identifikováno v průměru přibližně 1 700 chybných záznamů na soubor. Tyto chybné záznamy se skládaly z 19 až 20 unikátních chybných kódů. V letech 2016 až 2020 se situace zrcadlově otáčí – průměrný počet chybných záznamů se dostává na úroveň 40 000 na soubor. Nicméně tyto chybové záznamy jsou tvořeny pouze dvěma unikátními chybovými kódy.[2]

Při srovnání se staršími daty je zřejmé, že chybné hodnoty z původně sebraných dat byly v novějších datech přepsány, zatímco ve starších byly ponechány v původní podobě. Pro kontrolu a následnou analýzu je to nepříznivý fakt. Dostupná data máme rozdělena na dvě části v odlišné formě. Zároveň můžeme pouze odhadovat, jakým způsobem byly změny a úpravy v novějších souborech provedeny. Informace o tom, jak poskytovatel datových souborů se záznamy zacházel, nejsou součástí oficiální příručky ani jiného dostupného dokumentu. Výzkumný tým se na tuto situaci snažil reagovat tak, že úpravy ve starších souborech přizpůsoboval formě novějších dat (do té míry, do jaké to bylo možné), aby co nejvíce redukoval riziko nekonzistentnosti.

3.3 Platová sféra

Na rozdíl od mzdové sféry ve sféře platové není ve struktuře dat viditelný žádný předěl a chybovost je ve všech letech víceméně konzistentní. Vyskytují se zde proměnné zcela bez chyb (např. AA0226, AA0085, AA0179, AA0676, AA0091, AA0125), které tvoří přibližně polovinu všech kontrolovaných sloupců. Další skupina proměnných obsahuje pravidelně chybné záznamy, ale jen v řádu jednotek, popřípadě stovek (např. AA1586, AA0033, AA0231).

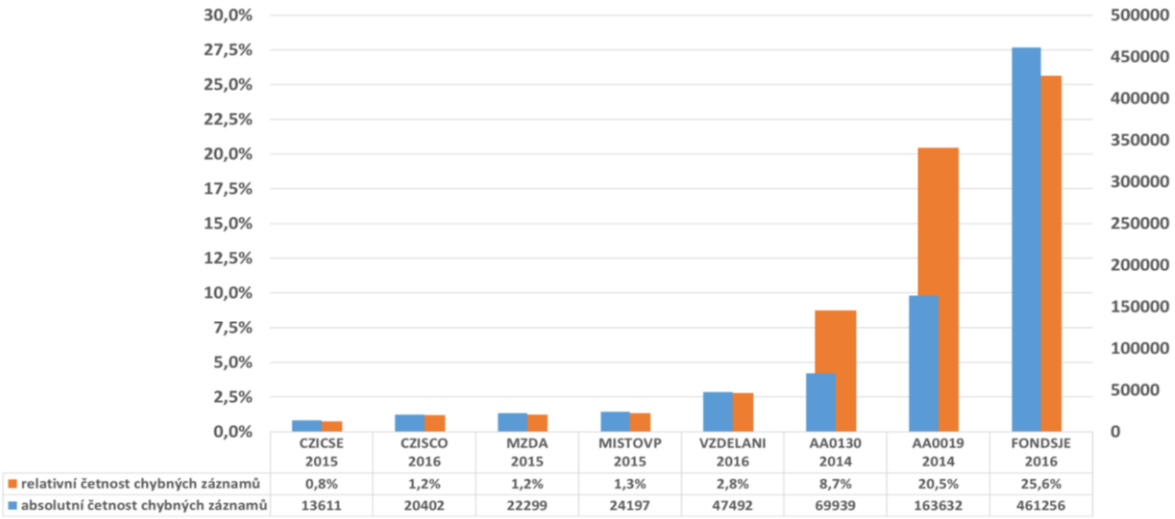
Obdobně jako ve mzdové sféře se i ve sféře platové objevují chyby absurdního charakteru. Jako takové můžeme například označit některé hodnoty proměnné AA0568 (datum vynětí z evidenčního stavu). Pravidelně a v nemalé míře se vyskytují chybné hodnoty jako například "2030", "2046" a poměrně často také "3999". Podobné případy mohou sloužit nejen pro ilustraci kvality dat, ale také náročnosti jejich kontroly a případných oprav.

Vůbec nejproblematičtější proměnné jsou AA0019 (právní forma ekonomického subjektu) a AA0013 (kód zaměstnání) – obě se pohybují řádově ve stovkách tisíc chybových řádků, a tím až v desítkách procent rozsahu celého datového souboru. Tyto chyby nejsou tvořeny pouze neplatnými kódy, ale mnohdy kódy zrušenými a v současnosti nepoužívanými (to je především případ proměnné AA0019 – právní forma ekonomického subjektu).

Příklady nejvíce chybových proměnných ve všech třech popsaných oblastech datových souborů ilustruje níže uvedený graf č. 1.

[2] Unikátní chybové kódy u proměnné VZDELANI jsou "x" a "-". Zdá se pravděpodobné, že byly do této podoby přepsány původní hodnoty (které v dřívějších souborech vidíme jako širší množinu unikátních chybových kódů). Není však jasné, podle jakého klíče byly úpravy v souborech z let 2016 až 2020 provedeny. Z tohoto důvodu nebylo možné tyto chybné hodnoty nijak dál opravovat či měnit.

Graf č. 1 **Nejvíce chybové proměnné v platové a mzdové sféře**



Zdroj: vlastní zpracování evidence kontroly dat

Graf ukazuje nejvíce chybové proměnné v platové i obou částech mzdové sféry. Vidíme, že se vyskytuje řada proměnných, jejichž chybovost se v absolutních číslech pohybuje v řádu desítek tisíc. Některé se však pohybují dokonce ve stovkách tisíc. V relativním vyjádření se jedná o řadu proměnných, jejichž relativní chyba tvoří jednotky procent souboru, až po extrémy, kde chyba tvoří pětinu, respektive čtvrtinu celého daného souboru.

Vzhledem k výrazné chybovosti dat, která jsou sama o sobě velmi rozsáhlá, bylo třeba věnovat jejich řešení značné úsilí. Výzkumný tým se proto pravidelně scházel, aby diskutoval o možnostech přístupu k jednotlivým proměnným. Mimo řešení zjištěných chyb bylo v souborech potřeba provést ještě další úpravy a doplnění. Všechny tyto změny jsou popsány v následující kapitole.

4. Úprava dat

Zjištěné chyby byly nejprve zaznamenány, a kde to bylo možné a vhodné, také opraveny. Veškerá evidence chyb a následných oprav je součástí přílohy. Valná část oprav se týká pouze dat mzdové sféry, jelikož chyby ze sféry platové svým charakterem v mnohých případech opravu neumožňovaly. V jiných případech by jejich oprava byla jednak časově velice náročná, jednak by neslibovala zvýšení kvality souboru. Mnohé chyby byly proto pouze zaznačeny.

Rozšířením souboru bylo dopočítání dalších proměnných, a to přepočtené placené měsíce zaměstnance v kumulaci za sledované období (MESPLAC_prep) a hrubé měsíční mzdy zaměstnance (HMM_mes) – obě dle vzorce Informačního systému o průměrném výdělku (ISPV, 2009). Analogickým způsobem byl spočítán i průměrný měsíční plat pro data z platové sféry.

Dále byl doplněn sloupec s umělou proměnnou nazvaný VSTUP_ANALÝZA (1–3), který umožňuje filtrovat data podle různých kritérií pro usnadnění jejich následné analýzy.

Nadto byl do každého souboru z platové i mzdové sféry doplněn ještě sloupec IDZAM_UNIKATNI (unikátní ID zaměstnance), aby bylo možné provádět analýzu časových řad.

Všechny provedené opravy a doplnění jsou jednotlivě popsány v následujících podkapitolách.

4.1 Oprava chyb

Oprava chyb probíhala pouze tam, kde bylo možné s velkou mírou jistoty identifikovat, jaké měly být správné hodnoty namísto původních chybových (zjevné překlepy), nebo kde bylo možné ověřit správnost informací jiným způsobem.

V souborech MI ze mzdové sféry byly opravovány chyby v proměnné LAU1 (okresy). Správné hodnoty byly primárně dohledány v předchozím souboru podle proměnné IČO, případně bylo čerpáno z externího zdroje (např. z webové stránky daného ekonomického subjektu dohledaného dle IČO).

Další opravy byly provedeny v rámci souborů typu MP (údaje zaměstnanců). U proměnné ROKNAR bylo opraveno jen několik hodnot, například "985" –> "1985", "1856" –> "1956" apod. Většina chyb však představovala hodnoty ze začátku 20. století (zaměstnanci hypoteticky ve věku kolem 100 let), které nebylo možné jednoduše opravit. Některá data byla proto jen smazána a označena přímo v souboru.

Obdobným způsobem byla opravována také proměnná STOBC, například "ČR" –> "CZ", nebo "CS" –> "ČR". Většinu chyb bylo ale možné pouze označit a smazat, jelikož nebylo jednoznačně a spolehlivě určitelné, jaké správné hodnoty by měly nabýt.

Zcela analogicky se postupovalo u proměnných INVALID, CZ-ICSE, VEDOUCI, KONTOPD a KONECEP – tj. opraveny byly pouze zjevné překlepy, zbytek byl zaevidován. Veškeré zásahy do původního souboru byly pečlivě označeny v kontrolním sloupci pro danou proměnnou, vždy pojmenovanou jako "XXX_ZMENA", kde XXX značí název relevantní proměnné.

Odlišný způsob oprav byl zvolen pouze u proměnné MISTOVP, která obsahovala velké množství chyb (jež ale nebylo možné označit za překlepy a přepsat na hodnoty správné) a přitom byla pro následnou analýzu naprosto stěžejní. Zvolen byl následující postup řešení: ze souboru s údaji zaměstnavatelů (soubor MI) byly vyexportovány unikátní kombinace IČ a kódů okresů, z čehož po úpravě vznikl svého druhu klíč, který byl následně přes IČO navázán zpět do souboru s daty zaměstnanců (soubor MP). Tímto způsobem se v každém ze 14 souborů ze mzdové sféry podařilo opravit kolem 90 % identifikovaných chyb – to představuje řádově desetitisíce záznamů v každém jednom souboru.

Třetí skupinu představují proměnné, kde byly chyby pouze smazány a evidovány v kontrolním sloupci (VZDELANI, CZ-ISCO, DOBAZAM, EVIDDNY, FONDSJE, ODPRACD a MZDA), případně evidovány a nemazány (zejména zbylé proměnné v platové sféře).

4.2 Dopočet placených měsíců a měsíčních mezd

Informace o hrubých měsíčních mzdách (platech) jako taková v souborech absentuje. Dostupné jsou pouze proměnné MZDA, tj. "mzda zúčtovaná v jednotlivých měsících od počátku roku do konce sledovaného období" (Trexima, 2020) a AA0144, tj. "celkový plat zaměstnance zúčtovaný k výplatě za sledované období" (ISDP, 2013). To znamená, že k dispozici máme jen sumu za dané pololetí, případně za rok. Pro přepočet byly v rámci mzdové sféry použity vzorce uvedené na webových stránkách Informačního systému o průměrném výdělku (ISPV, 2009).

Vzorec č. 1 **Výpočet hrubé měsíční mzdy zaměstnance**

$$\mathrm{HMM}_{\mathrm{i}}^{\mathrm{MES}} = \frac{\mathrm{MZDA}_{\mathrm{i}} + \mathrm{NAHRADY}_{\mathrm{i}} + \mathrm{POHOTOV}_{\mathrm{i}}}{\mathrm{MESPLAC\_prep}_{\mathrm{i}}}$$

Proměnná NAHRADYi představuje "náhrady mzdy zaměstnance v kumulaci za sledované období", proměnná POHOTOVi představuje "odměny za pracovní pohotovost zaměstnance v kumulaci za sledované období" a MESPLAC_prepi představuje "počet přepočtených placených měsíců zaměstnance v kumulaci za sledované období"; proměnnou je potřeba dopočítat podle dalšího vzorce (ISPV, 2009).

Vzorec č. 2 **Výpočet přepočtených placených měsíců zaměstnance**

$$\mathrm{MESPLAC\_prep}_{\mathrm{i}} = \frac{\mathrm{PRACDNY}_{\mathrm{i}}}{\mathrm{PRACDNY}} * \frac{\mathrm{ODPRACD}_{\mathrm{i}} - \mathrm{PRESCAS}_{\mathrm{i}} + \mathrm{ABSPLAC}_{\mathrm{i}}}{\mathrm{FONDSTA}_{\mathrm{i}}} * \mathrm{POCMES}$$

Proměnná PRACDNYi představuje "celkový počet pracovních dnů zaměstnance", proměnná PRACDNY představuje "počet pracovních dnů za sledované období", proměnná ODPRACDi představuje "celkový počet odpracovaných hodin zaměstnance", proměnná PRESCASi představuje "celkový počet přesčasových hodin zaměstnance", proměnná ABSPLACi představuje "celkový počet neodpracovaných hodin s náhradou mzdy zaměstnance", proměnná FONDSTAi představuje "celkový stanovený hodinový fond zaměstnance" a POCMES představuje "počet měsíců za sledované období", v našem případě tedy 6 nebo 12 (ISPV, 2009).

Pro platovou sféru obdobné vzorce nebyly dostupné. Vzhledem k tomu, že se v ní vyskytují jiné proměnné, výzkumný tým sestavil nový vzorec analogický tomu ze sféry mzdové.

Vzorec č. 3 **Hrubý měsíční plat zaměstnance**

$$\mathrm{HMP} = \frac{\mathrm{AA0144}}{\mathrm{POCMES} * \mathrm{A0136}}$$

Proměnná A0144 představuje „celkový plat zaměstnance zúčtovaný k výplatě za sledované období" a proměnná A0136 představuje „Koeficient průměrného evidenčního stavu zaměstnance přepočtený".

Vzhledem k definici proměnných lze proměnnou AA0144 považovat za analogickou k čitateli ze vzorce pro mzdovou sféru (MZDA + POHOTOV + NAHRADY). Stejně tak součin POCMES a A0136 je možné vzhledem k definici proměnných považovat za analogický ke jmenovateli vzorce ze sféry mzdové (MESPLAC_prep).

Pro zjednodušení se v úvodní fázi analýzy počítalo pouze se zaměstnanci, kteří pracovali po celé sledované období (tj. 181/182 a 365/366 dní v závislosti na typu a období souboru). To fakticky znamená, že podíl PRACDNY a PRACDNYi je vždy roven 1 a není jej tak ve mzdové sféře třeba do výpočtu zahrnovat a v platové sféře není třeba hledat jeho obdobu.

4.3 Doplnění proměnné IDZAM_UNIKATNI

Jak mzdová, tak platová sféra obsahují proměnnou pro identifikaci zaměstnanců. Ve mzdové sféře se jedná o proměnnou IDZAM, tj. "jednoznačný a v čase stálý kód (po dobu šetření v ISPV) fyzické osoby (zaměstnance) v rámci ekonomického subjektu". V platové sféře je to proměnná AA0126 (Fyzická osoba – kód), tj. "jednoznačný kód fyzické osoby (zaměstnance) v rámci zaměstnavatele."

Problémem obou zmíněných proměnných je, že mohou sloužit pouze v rámci zaměstnavatele, respektive v rámci ekonomického subjektu. Může se tak lehce stát, že několik zaměstnanců bude mít stejný kód. Výzkumný tým se proto rozhodl doplnit ID zaměstnance o IČO ekonomického subjektu. Nicméně ani tato úprava nezajišťuje jednoznačnou a unikátní identifikaci zaměstnance v čase – mohlo by totiž dojít k situaci, že se ID v rámci jedné firmy přesune na jiného zaměstnance (např. při propouštění a dalším náboru). Unikátní ID bylo proto ještě rozšířeno o dvě osobní charakteristiky zaměstnaných osob, které jsou v čase vždy konzistentní – rok narození a pohlaví. Proměnná IDZAM_UNIKATNI je tak tvořena celkem 4 složkami (viz vzorce č. 4 a č. 5).

Vzorec č. 4 **Unikátní ID zaměstnance ve mzdové sféře**

$$IDZAM\_UNIKATNI = ICO\_IDZAM\_ROKNAR\_POHLAVI$$

Vzorec č. 5 **Unikátní ID zaměstnance v platové sféře**

$$IDZAM\_UNIKATNI = AA0017\_AA0126\_AA0231\_AA0211$$

Ačkoliv jsou proměnné odlišně označeny, v obou případech se jedná o totožnou definici. Obecně se v obou případech jedná o "ID zaměstnance"&_&"IČO ekonomického subjektu"&_&"rok narození"&_&"pohlaví v kódovém označení 1–2".

4.4 Doplnění proměnné VSTUP_ANALÝZA (1–3)

Přestože byly soubory očištěny od chyb, které bylo možné nalézt kontrolami popsanými v úvodu textu, stále zde přetrvávaly záznamy, jež byly na první pohled nesprávné. Viditelné to bylo především z

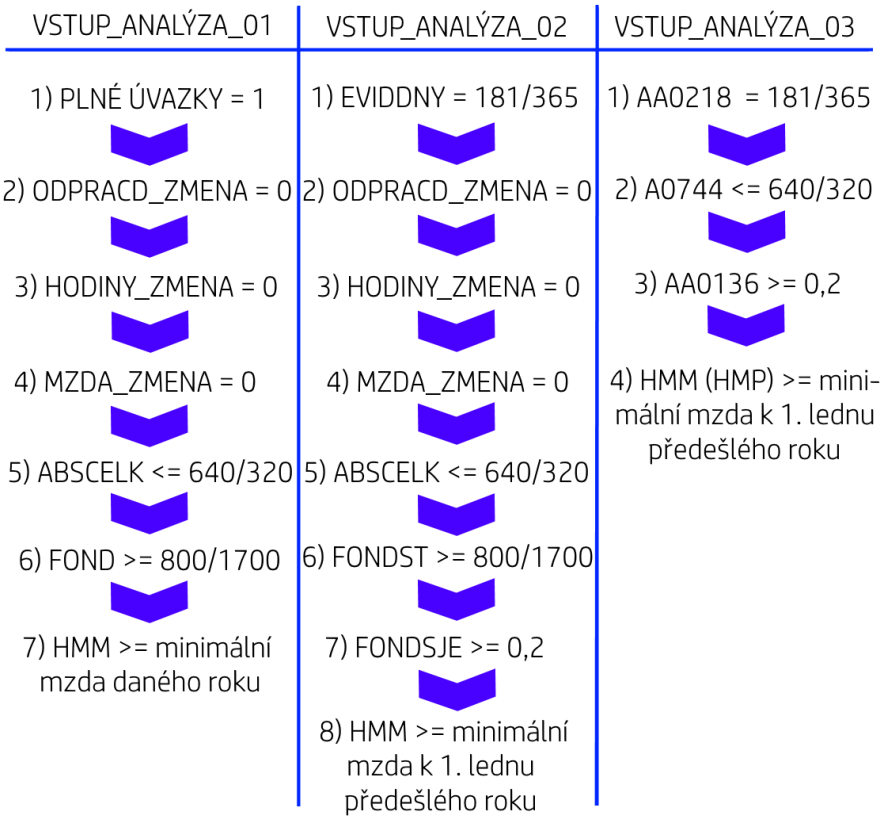
pohledu na hrubé měsíční mzdy. Řádově až tisícům zaměstnanců (v každém souboru) byla vyplacena jejich měsíční odměna za práci hluboko pod úrovní minimální mzdy, včetně nemalého počtu případů, kdy odměna za práci byla ve výši jednotek, desítek či stovek korun. Na druhé straně se vyskytly i ojedinělé případy, kdy mzdy vyskočily do velmi odlehlých hodnot, například v souboru MP_Q2 2015 – zaměstnancům středně velké firmy (Brno-venkov) vycházela průměrná hrubá měsíční mzda na 10,5 miliónu měsíčně.

Tyto anomálie byly zpravidla zapříčiněny způsobem výpočtu mzdy. U velmi nízkých hodnot se často jednalo o zaměstnance s dlouhodobými absencemi. Pokud je zaměstnanec delší dobu v pracovní neschopnosti, přechází povinnost platit nemocenskou na ČSSZ a tyto částky (byť jsou reálným příjmem zaměstnance) se v souboru nepromítnou. U několika případů s velmi vysokými mzdami byl příčinou anomálií samotný jmenovatel vzorce – z různých důvodů vycházel v určitých případech velmi malý a výsledná mzda pak byla extrémně vysoká.

Na základě těchto a dalších obdobných případů se výzkumný tým rozhodl zařadit další kritéria, aby jich co nejvíce odfiltroval, ale zároveň zachoval maximum dat. Byl proto doplněn další sloupek nazvaný "VSTUP_ANALÝZA", který představuje umělou proměnnou, v níž "1" znamená záznam přípustný k analýze a "0" nepřípustný.

Kritéria byla nejprve stanovena pro mzdovou sféru (VSTUP_ANALÝZA_01), následně upravena (VSTUP_ANALÝZA_02) a analogicky sestavena pro sféru platovou (VSTUP_ANALÝZA_03). Složení zahrnutých kritérií je dáno jednak dříve provedenými kontrolami, jednak kritérii orientovanými na absence, výši pracovního poměru a výši mzdy.

Schéma č. 1 **Postup pro tvorbu proměnné VSTUP_ANALÝZA (1–3)**



Proměnná VSTUP_ANALÝZA_01 byla později nahrazena proměnnou VSTUP_ ANALÝZA_02. V textu je uvedena nejen pro informaci, ale rovněž proto, aby proměnná zůstala v datových souborech zachována pro případ budoucí potřeby. Kritéria užitá v proměnné VSTUP_ANALÝZA_01 jsou stanovena striktněji. Zjištění z těchto dat poskytnou odpovědi za zaměstnance pracující na plný úvazek.

Pokud u proměnné VSTUP_ANALÝZA je hodnota rovna jedné, znamená to splnění všech kritérií uvedených ve schématu č. 1.

Proměnná PLNÉ ÚVAZKY (značící, že FONSTA = FONDJSE & EVIDDNY = 181/365) byla zavedena pro zaměstnance s plným úvazkem, kteří pracovali po celou dobu v rámci daného souboru (celý půlrok či rok). Od zařazení toho kritéria si výzkumný tým sliboval zmírnění výše uvedených chyb (odlehlých hodnot mezd).

Položky schématu 2–4, ODPRACD_ZMENA, HODINY_ZMENA (kontrola ODPRACD na FODJSE + PRESCAS – ABSCELK)[3] a MZDA_ZMENA, reflektují úvodní kontroly. Fakt, že mají být rovny nule, znamená, že v těchto záznamech nedocházelo k žádnému mazání dat a původní údaje byly vyhodnoceny jako správné. Položka 5 omezuje vzorek přípustný pro analýzu na ty zaměstnance, kteří měli celkové absence nižší než 2, respektive 4 měsíce (pro pololetí, respektive rok). Tato hranice byla stanovena s přihlédnutím k tomu, že proměnná ABSECLK může být tvořena dovolenými, propustkami anebo kratšími obdobími pracovní neschopnosti. Výzkumný tým je do této výše zhodnotil jako relevantní.

Položka 6 je pouze logickou podmínkou, která by z definice měla platit vždy – bez ohledu na to, zda se jedná o FONDSJE nebo FONDSTA, protože pro plné úvazky jsou si tyto dvě proměnné vždy rovny.

Poslední položka je podmínkou, že zaměstnanec dostával v průměru alespoň minimální mzdu (platnou v daném roce).

Na základě diskuze výzkumný tým dospěl k několika změnám, z nichž byla v první řadě aplikována proměnná VSTUP_ANALÝZA_02. Namísto kritéria na plné úvazky byla hranice stanovena na 0,2 úvazky. Do analýzy tak vstupují pouze zaměstnanci s 0,2 a vyšším úvazkem (tj. kritérium FONDSTA >= 800/1700 & FONSJE >= 0,2*(800/1700)). Zároveň bylo zachováno kritérium na EVIDDNY, tj. „zaměstnanec pracoval po celé sledované období (dané pololetí nebo dané rok)“, které má za úkol jednak pomoci odfiltrovat odlehlé hodnoty, jednak má technické opodstatnění, jelikož usnadňuje výpočet HMM (HMP). Poslední změnou bylo částečné uvolnění hranice pro minimální mzdy, která byla posunuta z platné minimální mzdy na úroveň platnou k 1. lednu 2020. Tato úprava umožnila do analýzy začlenit řádově o několik tisíc záznamů v každém období navíc. Především však umožnila posouzení toho, v jaké míře někteří zaměstnavatelé porušují či neporušují platnou právní úpravu.

Zcela analogicky byla kritéria nastavena v proměnné VSTUP_ANALÝZA_03, která je zaměřena na platovou sféru. Rozdíly jsou pouze formální. Odfiltrování nižších než 0,2 úvazků bylo jednodušší, protože na rozdíl od mzdové sféry ve sféře platové existuje proměnná, která přímo výši úvazku vyjadřuje. Kromě toho absentují některé položky, které byly navázány na kontrolu a v platové sféře nebyly nutné (položky 2–4) (viz schéma č. 1).

Po dokončení všech výše uvedených úprav, doplnění, evidence a oprav chyb byla data připravena pro následující analýzu. Nicméně, aby bylo možné s daty pohodlně pracovat jako s časovými řadami, bylo ještě zapotřebí provést export původních dat do souhrnných souborů.

[3] Tam, kde kontrola odhalila, že ODPRACD ≠ FONDSJE + PRESCAS – ABSCELK (o víc než 5 %), byly tyto záznamy v agregovaném sloupci HODINY_ZMENA označeny a hodnoty z uvedených proměnných byly smazány.

5. Export dat pro analýzu

Poslední úpravou před samotnou analýzou dat, byl jejich export a kompletace do jednoho souboru. Skupinou, která je hlavním předmětem naší analýzy jsou zaměstnanci pobírající minimální mzdu. Byly proto nejprve dohledáni všichni zaměstnanci, kteří alespoň jednou během sledovaného období minimální mzdu pobírali (v pásmu minimální mzdy platné k 1. lednu 2020 až aktuální platné minimální mzdy + 10 %) a zároveň jejich záznamy představují přípustné hodnoty (VSTUP_ANALÝZA = 1). Ze seznamu těchto zaměstnanců byl vytvořen kompilát, který byl zpětně v datových souborech použit jako filtr (skrze proměnnou IDZAM_UNIKATNI).

Před exportem byla do všech souborů doplněna proměnná "ROK&KVARTÁL" (např. 2015Q2, 2015Q4 apod.).

Do samotného exportu byly zahrnuty i ty případy, kdy zaměstnanec stanovená kritéria neplnil (VSTUP_ANALÝZA = 0), aby bylo možné zkoumat důvody, proč zaměstnanec začal či přestal minimální mzdu pobírat.

Aby bylo možné provádět další analýzy, postupovalo se obdobně i v případě dalších skupin zaměstnanců – například zaměstnanci, kteří brali mzdu v rozmezí minimální mzdy +10 % až +20 % (průzkum tzv. spill-over efektu apod.).

Při kontrole vytvořených exportů byla zjištěna další chyba datového souboru, která výše uvedenými kontrolami objevena nebyla. Jednalo se o duplicitní záznamy zaměstnanců (dva a více záznamů pro jedno unikátní ID zaměstnance ve stejném období). V některých případech bylo zřejmé, že pravděpodobným důvodem bylo chybné zadávání vstupních informací (jeden řádek vyplněných hodnot a druhý řádek se stejným IDZAM_UNIKATNI, ale s prázdnými hodnotami ve všech buňkách). Jinde byla interpretace složitější (mohlo se např. jednat o jednoho zaměstnance, který měl u stejného zaměstnavatele více pracovních smluv). Řádově šlo o jednotky případů. Vzhledem k heterogenní formě těchto chyb i jejich nízkému počtu se členové výzkumného týmu shodli na řešení v podobě smazání všech záznamů, které vykazovaly duplicitní hodnoty IDZAM_UNIKATNI v rámci stejného období.

Shrnutí

Datový set, kterým výzkumný tým disponuje, je velmi rozsáhlý a umožňuje unikátní pohled na český trh práce. Jeho využití představuje velkou příležitost, zároveň je však spojeno s řadou úskalí.

Dostupná data mzdové sféry nejsou reprezentativním vzorkem. Obsahují totiž disproporčně více informací o zaměstnancích z velkých firem oproti zaměstnancům z firem malých, což je dáno způsobem sběru dat.

Data jsou poměrně chybová – chybovost některých proměnných se šplhá k desítkám až stovkám tisíc řádků, to představuje až desítky procent celých souborů (např. MISTOVP v letech 2014 a 2015 mzdové sféry nebo proměnná AA0019 ve všech obdobích platové sféry). V některých případech byly chyby opraveny (např. s pomocí souborů typu MI byla opravena většina chyb v proměnné MISTOVP, řádově desítky tisíc záznamů). Kde to nebylo možné, byly chyby alespoň zdokumentovány a zaevidovány, případně smazány.

Celý proces kontrol a oprav byl komplikován nekonzistencí dat ve mzdové sféře, kde je jasně patrný předěl mezi obdobími 2014 až 2015 a 2016 až 2020. Je nanejvýš pravděpodobné, že novější data byla upravována, výzkumný tým však neměl a nemá k dispozici žádné bližší informace o tom, jak přesně byly tyto úpravy provedeny. Další překážkou v kontrolách byla častá absence číselníků, které bylo nutné individuálně dohledávat. Komplikací byly také rozdíly mezi platovou a mzdovou sférou, kde byly obsahově stejné sloupce často vymezeny odlišnými číselníky.

Kromě kontrol a oprav bylo také nutné doplnit proměnnou pro hrubé měsíční mzdy a platy. Ve mzdové sféře byly dostupné oficiální vzorce, v platové však nikoliv – byly proto vytvořeny analogicky k těm ze mzdové sféry. Zároveň bylo nutné doplnit proměnnou VSTUP_ANALÝZA, umožňující filtrovat data podle zadaných kritérií.

Datový soubor je svým rozsahem zcela výjimečný, tím je však také velice náročný na kontroly, opravy i dodatečné úpravy. Se zvýšenou opatrností je také nutné přistupovat k interpretaci pozdějších výsledků, jelikož dostupná data (ač zahrnují miliony zaměstnanců) nejsou reprezentativním vzorkem – údaje o zaměstnancích z malých a středních firem jsou v datech v nedostatečném zastoupení, popřípadě nejsou zahrnuty vůbec.

Zdroje

TREXIMA, 2020. *Příručka pro přípravu vstupních dat*.
Dostupné z: https://www.ispv.cz/cz/Pro-respondenty-setreni/Prirucka.aspx

ČSÚ – Český statistický úřad, 2020a. *Klasifikace zaměstnání (CZ-ISCO)*.
Dostupné z: https://www.czso.cz/csu/czso/klasifikace_zamestnani_-cz_isco-

ČSÚ – Český statistický úřad, 2020b. *Klasifikace zemí (CZ-GEONOM)*.
Dostupné z: https://www.czso.cz/csu/czso/klasifikace_zemi_-cz_geonom-

ČSÚ – Český statistický úřad, 2014a. *Klasifikace postavení v zaměstnání (CZ-ICSE)*. Dostupné na:
https://www.czso.cz/csu/czso/klasifikace_postaveni_v_zamestnani_-cz_icse-

ČSÚ – Český statistický úřad, 2014b. *KZAM – systematická část*.
Dostupné z: https://www.czso.cz/csu/czso/kzam_systematicka_cast

ČSÚ – Český statistický úřad, 2021a. *Číselník okresů (OKRES_LAU)*.
Dostupné z: https://www.czso.cz/csu/czso/ciselnik_okresu_-okres_lau-

ČSÚ – Český statistický úřad, 2021b. *Právní forma organizace – agregace*.
Dostupné z: http://apl.czso.cz/iSMS/cisdet.jsp?kodcis=2755

MŠMT - Ministerstvo školství, mládeže a tělovýchovy, 2021. *Používané číselníky v resortu školství*.
Dostupné z: http://stistko.uiv.cz/katalog/ciselnika.asp

ISDP – Informační systém o datových prvcích Ministerstva vnitra ČR, 2006–2018. *Seznam datových prvků*.
Dostupné z: https://www.sluzby-isvs.cz/isdp/DefaultSSL.aspx

ISPV – Informační systém o průměrném výdělku, 2009. *Výpočetní algoritmy v ISPV a RSCP pro strukturu ISPV2009*. Dostupné z: https://www.ispv.cz/cz/Vysledky-setreni/ Metodika.aspx

ISPV – Informační systém o průměrném výdělku, 2020. *Metodika*.
Dostupné z: https://ispv.cz/cz/Vysledky-setreni/Metodika.aspx

Přílohy

A) Kontrola a opravy dat ze mzdové sféry (pracovní dokument):
https://docs.google.com/spreadsheets/d/1pIFj-BgqXyCRxI7vdAj7ZsVKNPawNs4oHTe293UX7K8/edit?usp=sharing

B) Kontrola a opravy dat z platové sféry (pracovní dokument):
https://docs.google.com/spreadsheets/d/15zTU048Se10lztegylcKSeuVt71OuZCsGdL9RNaEYNg/edit?usp=sharing