Život během pandemie, vlny 1–26

Dokumentace k veřejnému datovému souboru (verze 1.0)

/ Publikováno
7. 6. 2021

// Autoři
PAQ Research
Michaela Kudrnáčová

PAQ RESEARCH

Obsah

Výzkum a jeho účel

O projektu

Cílem projektu Život během pandemie bylo shromažďovat data o vývoji epidemiologicky významného chování českých dospělých během pandemie nemoci covid-19 (zejména počty kontaktů, styk s nakaženým, testování, dodržování preventivních opatření, sociální aktivity, docházka na pracoviště, symptomy onemocnění). Sledovány byly také změny v pracovní aktivitě dotazovaných, ekonomická situace domácností a míry duševního zdraví s cílem vyhodnocovat dopady epidemie na uvedené oblasti života. Dále byly zjišťovány postoje veřejnosti k vládním opatřením, testování a očkování. K základní sadě socio-demografických proměnných dotázaných v první vlně šetření byly v jednotlivých vlnách postupně doplňovány socioekonomické charakteristiky respondentů a domácností, včetně retrospektivního zjišťování stavu před začátkem epidemie. Součástí šetření byly rovněž ad hoc moduly týkající se distanční výuky nebo práce z domova. Obsah šetření vytvořily výzkumná společnost PAQ Research a iniciativa IDEA AntiCovid.

Sběr dat byl financován především z následujících zdrojů:

* Max Planck Institute for Tax Law and Public Finance,
* Akademie věd ČR (ERC-CZ/AV-B (ERC300851901)),
* German Science Foundation CRC TRR 190,
* Technologická agentura České republiky (ETA Město pro lidi, ne pro virus, č. TL04000282),
* Sociologický ústav AV ČR.

Sběr dat ve vybraných vlnách podpořil také Člověk v tísni a soukromí zadavatelé ad-hoc tematických bloků.

Časové řady agregovaných ukazatelů byly pravidelně publikovány na webu zivotbehempandemie.cz (příprava tohoto datového dashboardu byla podpořena grantem TAČR ETA Mapování dopadů ekonomické krize a optimalizace systémů daní, dávek, exekucí a insolvencí pro zmírnění jejích nepříznivých vlivů (TL04000332)).

Klíčová slova

Covid-19, epidemie, pandemie, chování během epidemie, dopady epidemie

Obsah šetření

Základ šetření tvořilo **kontinuální epidemiologické jádro** zaměřující se na charakteristiky respondentů a chování, které může souviset s šířením nákazy. Od druhé vlny šetření byly opakovaně zařazovány otázky na pracovní život a ekonomiku domácností, tedy **kontinuální ekonomický modul**. Většina proměnných z těchto modulů byla dotazována již od první, resp. druhé vlny, část však byla doplněna později. U několika proměnných byla mezi vlnami provedena úprava filtrování či jiného parametru otázky. Na zařazení proměnné jen v části vln nebo úpravy v parametrech proměnných upozorňují *tabulky 1 a 2*. Podrobný přehled všech proměnných a vln, ve kterých se objevovaly, poskytuje seznam proměnných tvořící přílohu této dokumentace (viz část Evidence otázek).

Ve vybraných vlnách byly dále dotazovány **postoje k vládním opatřením** proti šíření epidemie (souhlas s jednotlivými opatřeními, zda představují pro život respondenta komplikace, vnímaná účinnost). Ve vybraných vlnách bylo zařazeno několik otázek na respondentův **odhad ohledně aktuálního a budoucího vývoje epidemie**.

Tabulka 1: Kontinuální jádro epidemiologického modulu

Kód	Popis	Poznámka
q012	kontakt s nakaženou osobou	dotazováno ve všech vlnách (v několika vybraných vlnách zařazena upřesňující otázka na dobu kontaktu q243)
q027	testování respondenta a člena domácnosti na koronavirus	v první vlně odlišná nabídka kategorií než v následujících vlnách (rozlišení na proměnné a = 1. vlna, b = ostatní vlny),
q028, q030	symptomy nemoci respondenta a případných dalších členů jeho domácnosti	v první vlně byl použit filtr více omezující množinu respondentů k dotázání než v následující vlnách
q018, q022	počet přímých kontaktů v předchozích dvou týdnech	dotazováno ve všech vlnách
q266, q224	počet kontaktů delších než 15 minut bez roušky v předchozích dvou týdnech	kontakt v týdnu 2 dotazován od 17. vlny, kontakty v týdnu 1 ve 20. vlně a poté od 24. vlny
q016, q020	společenské aktivity vykonávané v předchozích dvou týdnech (respondent a členové domácnosti)	část aktivit nebyla dotazována v prvních vlnách, část aktivit nebyla dotazována na podzim 2020
q023	podnikaná preventivní opatření (nošení ochrany dýchacích cest, nevychází z domu atd.)	dotazováno ve všech vlnách, pouze nevycházení z domu doplněno až ve 2. vlně

q038	absolvování testu, jeho termín, výsledek, typ	dotazováno od 2. vlny, ve 21. vlně změna způsobu dotazování na termín a typ testu
q017, q021	docházka na pracoviště / do školy v předchozích dvou týdnech	v 1. vlně odlišná nabídka kategorií než v následujících vlnách (rozlišení na proměnné a = 1. vlna, b = ostatní vlny), ve vlnách 3 a 4 nedotazováni na rozdíl od ostatních vln studenti
q200, q277	ochota nechat se očkovat	dotázáno ve vlnách 13, 18, 20; od 21. vlny dotazováno, jen pokud dosud nebyl/a očkován/a

Tabulka 2: Kontinuální jádro ekonomického modulu

Kód	Popis	Poznámka
q046, q278	změna v pracovní aktivitě respondenta od počátku epidemie / v období posledních týdnů	od 2. do 20. vlny zjišťována změna od počátku epidemie, od 21. vlny dotazován stav v posledních týdnech oproti stavu před epidemií u q046 postupné úpravy ve znění nabízených možností ve vlnách 3 až 6 a 19
q106, q122, q279	vybraná omezení pracovní aktivity	od 2. do 20. vlny zjišťována zvlášť omezení, která respondenta postihla od začátku epidemie, a poté stále platná omezení (přičemž v 16. vlně jen aktuálně platná); od 21. vlny dotazována aktuálně platná omezení ve srovnání se stavem před epidemií
q049, q050	počet odpracovaných hodin v předchozích dvou týdnech	ve 2. vlně nedotázáni na rozdíl od ostatních vln nezaměstnaní
q173, q174	dovolená v předchozích dvou týdnech	od 21. vlny zúžení filtrační podmínky
q118	zaměstnanec na překážkách práce	zúžení filtrační podmínky od 10. vlny
q051	obava ze ztráty práce	
q052	hodnocení finanční situace domácnosti	
q053	omezení příjmu domácnosti ve srovnání se stavem před epidemií	v 5. vlně a od 21. vlny dále doplněny otázky specifikující případné navýšení příjmů
q055	zamýšlená a přijatá opatření k řešení finanční situace domácnosti	přijatá opatření dotazována od 5. vlny, omezení výdajů na služby dotazováno od 7. vlny
q056	finanční situace domácnosti (problémy s platbami)	
q057	příjem domácnosti (kategorie)	**vzhledem k rozdílným velikostem domácnosti není vhodné s proměnnou pracovat bez dalších úprav**

Šetření zahrnovalo také pravidelný modul týkající se duševního zdraví a několik ad hoc modulů týkajících se distanční výuky, práce z domova, volebního chování a sociálně-politických postojů, které nejsou z důvodu exkluzivity dat zařazeny do veřejně přístupného souboru.

Sběr dat

Mód sběru

Sběr dat realizovala podle zadání PAQ Research agentura NMS Market Research (člen SIMAR). Data byla sbírána pomocí **CAWI dotazování** na stabilní skupině respondentů vybraných z **Českého národního panelu**. Český národní panel je společný projekt výzkumných agentur Nielsen Admosphere, NMS Market Research a STEM/MARK, který představuje online panel registrovaných respondentů účastnících se výzkumů trhu a veřejného mínění. Do panelu se mohou registrovat lidé starší 15 let trvale žijící na území České republiky. Při registraci je ověřována totožnost respondentů. Respondenti jsou za vyplňování dotazníků finančně odměňováni.

Online metodika sběru dat byla pro výzkum Život během pandemie zvolena vzhledem k potřebě rychlého uspořádání počátečních vln šetření během nástupu první vlny epidemie i citlivý charakter dotazovaných témat.

Výzkum je kvótně reprezentativní pro populaci ČR, ale kvůli módu sběru dat se ho mohli účastnit jen respondenti s připojením k internetu. Výstupy pro starší generaci (65+) lze v důsledku online sběru považovat pouze za orientační.

Délka dotazování

Průměrná délka dotazování se v jednotlivých vlnách pohybovala obvykle **od 14 do 18 minut** (spočteno s vyloučením spodních a horních pěti procent respondentů dle délky vyplňování, tento tzv. „ořezaný“ průměr byl k vyčíslování průměrné délky dotazníku zvolen z toho důvodu, že v online dotazování nemusí čas otevření dotazníku odpovídat času jeho reálného vyplňování).

Strategie pro udržování vysoké míry retence

Pro maximalizaci míry retence mezi vlnami bylo přijato několik opatření:

* zasílání **avizních e-mailů** respondentům několik dnů před spuštěním další vlny dotazování (běžně v pátek před pondělním spuštěním sběru dat), vedle informace o termínu další vlny tyto zprávy obsahovaly poděkování za účast a krátkou informaci o dosavadním využívání dat
* před 21. vlnou a 28. vlnou zaslání **rozsáhlejšího e-mailu s poděkováním**, informací o výši odměn dosud předaných za účast respondentům a prosbou o další spolupráci

* ve většině vln **odměňování** jednoho náhodně vybraného respondenta **tabletem** (nad rámec běžné finanční odměny za vyplněný dotazník v Českém národním panelu)

* udržování **průměrné délky** dotazování mezi 14 až 18 minutami

V jednotlivých vlnách byl při zpracování dat kontrolován vliv atrice (podrobnosti v části Kontrola a zpracování dat) a její vliv je částečně redukován při post-stratifikačním vážení (podrobnosti v části Vážení).

Nejnižší míra retence mezi vlnami byla zaznamenána mezi první a druhou vlnou (85 %), v níž z původních cca 3 100 respondentů vyplnilo dotazník 2 639 jednotlivců. Poté se absolutní velikost vzorku snižovala ve výrazně omezenější míře. *Tabulka 3* vyčísluje míru retence jak k první vlně, tak ke druhé vlně, neboť množství časových řad (ekonomický modul a kvůli změně způsobu zjišťování také část epidemiologického) má svůj počátek až ve druhé vlně. Pro případné posuny v časových řadách jsou tak relevantnější případné změny ve složení vzorku mezi druhou a následujícími vlnami než vůči první vlně.

Tabulka 3: Parametry jednotlivých vln sběru dat

Vlna	n	Retence vůči vlně 1	Retence vůči vlně 2	Začátek sběru	Konec sběru	Rozestup mezi vlnami v týdnech	Délka vyplňování (5% "ořezaný" průměr)
1	3101			18.3.2020	19.3.2020		9,66
2	2639	85,1 %		30.3.2020	1.4.2020	2	15,95
3	2567	82,8 %	91,2 %	14.4.2020	20.4.2020	2	19,05
4	2610	84,2 %	91,1 %	27.4.2020	4.5.2020	2	14,16
5	2470	79,7 %	87,2 %	11.5.2020	16.5.2020	2	16,62
6	2492	80,4 %	87,5 %	25.5.2020	1.6.2020	2	16,90
7	2439	78,7 %	85,2 %	8.6.2020	15.6.2020	2	15,68
8	2319	74,8 %	81,3 %	29.6.2020	7.7.2020	3	14,40
9	2255	72,7 %	79,8 %	20.7.2020	27.7.2020	3	13,31
10	2201	71,0 %	77,7 %	10.8.2020	17.8.2020	3	13,89
11	2240	72,2 %	78,7 %	31.8.2020	7.9.2020	3	15,68
12	2185	70,5 %	77,2 %	14.9.2020	21.9.2020	2	14,26
13	2167	69,9 %	76,8 %	29.9.2020	5.10.2020	2	15,77
14	2246	72,4 %	79,0 %	12.10.2020	19.10.2020	2	18,82
15	2299	74,1 %	80,4 %	26.10.2020	2.11.2020	2	14,52
16	2225	71,8 %	78,4 %	9.11.2020	16.11.2020	2	12,81
17	2292	73,9 %	80,4 %	23.11.2020	30.11.2020	2	15,30
18	2155	69,5 %	76,1 %	7.12.2020	14.12.2020	2	17,74
19	2051	66,1 %	73,1 %	21.12.2020	28.12.2020	2	14,42
20	2186	70,5 %	76,8 %	4.1.2021	11.1.2021	2	13,53
21	2131	68,7 %	74,7 %	25.1.2021	1.2.2021	3	14,24
22	2120	68,4 %	74,6 %	15.2.2021	22.2.2021	3	16,67

23	2130	68,7 %	74,8 %	1.3.2021	8.3.2021	2	14,18
24	2101	67,8 %	74,2 %	15.3.2021	22.3.2021	2	17,66
25	2061	66,5 %	72,5 %	29.3.2021	6.4.2021	2	23,52
26	2059	66,4 %	72,5 %	12.4.2021	19.4.2021	2	18,40

Termíny sběru dat

První vlna šetření proběhla v polovině března 2020 a dotazování bylo poté opakováno v intervalu dvou až tří týdnů. Termíny dotazování shrnuje *tabulka 3*. Do konce března 2021 bylo provedeno 26 vln šetření.

Sběr dat byl zahajován zpravidla v pondělí daného týdne sběru dat a uzavírán byl následující pondělí v dopoledních hodinách. Přibližně dvě třetiny výsledných dotazníků bylo získáno během prvních dvou dnů sběru dat.

Populace, výběrový soubor

Cílová populace

Cílovou populaci šetření představují **jednotlivci ve věku 18 let a vyšším trvale žijící na území České republiky**. Vzhledem k online sběru dat se výzkumu mohli účastnit jen **respondenti s připojením k internetu** a závěry je tak vhodné vztahovat jen k internetové populaci (především ve starších věkových kategoriích).

Výběrový soubor

V první vlně šetření (březen 2020) byli k účasti oslovováni členové Českého národního panelu tak, aby výsledný dotázaný vzorek **kvótně odpovídal dospělé populaci České republiky** dle kombinace pohlaví a věku (18–24 let, 25–34 let, 35–44 let, 45–54 let, 55–64 let, 65 a více let), vzdělání (základní a středoškolské bez maturity, středoškolské s maturitou a vysokoškolské) a kraje.

Pro možnost robustnějšího modelování vztahu sociálních aktivit, protektivního chování a promořenosti byly v hrubém vzorku **nadhodnoceny** rizikové kategorie **měst od 50 tisíc obyvatel**, tedy zastoupení obyvatel těchto měst bylo záměrně vůči populační struktuře navýšeno. Velikost místa bydliště byla podobně jako pohlaví či kraj používána jako kvótní charakteristika (do 49 999, 50 000 a více obyvatel), avšak její cílová distribuce byla upravena tak, aby v hrubém vzorku byl obsažen větší počet obyvatel větších měst, než jakého by mohlo být dosaženo při využití populační distribuce. V souladu s tím byla navýšena proporce Hlavního města Prahy a rovněž u vzdělání byla mírně nadhodnocena kategorie s maturitou a vysokoškolským vzděláním.

Vzorek byl tedy sestaven tak, aby by byl kvótně reprezentativní na dospělou populaci ČR s tzv. **nadvýběrem měst s 50 tisíci a více obyvateli**, který je při zpracování dat **statisticky redukován vážením dat**. Podrobnější informace k vážení (včetně distribuce využitých populačních četností) obsahuje sekce Vážení.

Základní velikost cílového souboru se strukturou danou kvótním předpisem činila 3000 jednotek, dalších až 150 jednotek mohlo mít z hlediska kvótních proměnných libovolné charakteristiky. První vlny se zúčastnilo 3 108 respondentů. Kvótní proměnné a jejich distribuci zahrnuje *tabulka 4*.

Tabulka 4: Kvótní charakteristiky využité k sestavení vzorku v první vlně a jejich rozložení

Pohlaví a věk	**n**	**%**
Muž, 18–24 let	119	4,0
Muž, 25–34 let	242	8,0
Muž, 35–44 let	299	9,9
Muž, 45–54 let	257	8,5
Muž, 55–64 let	227	7,5
Muž, 65 a více let	361	12,0
Žena, 18–24 let	119	4,0
Žena, 25–34 let	242	8,0
Žena, 35–44 let	299	9,9
Žena, 45–54 let	257	8,5
Žena, 55–64 let	227	7,5
Žena, 65 a více let	361	12,0
Velikost místa bydliště (nadvýběr 200 % 50+)	**n**	**%**
Do 49 999 obyvatel	1050	34,9
50 000 a více obyvatel	1950	64,8
Kraj (nadvýběr 200 % Praha)	**n**	**%**
Hlavní město Praha	751	25,0
Středočeský kraj	323	10,7
Jihočeský kraj	155	5,1
Plzeňský kraj	142	4,7
Karlovarský kraj	72	2,4
Ústecký kraj	197	6,5
Liberecký kraj	107	3,6
Královéhradecký kraj	134	4,5
Pardubický kraj	126	4,2
Kraj Vysočina	124	4,1
Jihomoravský kraj	287	9,5
Olomoucký kraj	154	5,1
Zlínský kraj	142	4,7
Moravskoslezský kraj	293	9,7
Nejvyšší dosažené vzdělání	**n**	**%**
Základní a středoškolské bez maturity	1200	39,9
Středoškolské s maturitou a vysokoškolské	1800	59,8
+ 150 jednotek s libovolným rozdělením v kvótních charakteristikách		

První vlny se zúčastnilo 3 108 respondentů. Sedm z nich však nebylo zařazeno do zpracování longitudinálních dat z důvodu nekonzistence uváděných sociodemografik v prvních deseti vlnách šetření. Výsledný vzorek z první vlny tak čítá 3 101 pozorování.

Skupina účastníků z první vlny byla opakovaně oslovována k účasti v dalších vlnách. Oslovení v druhé a další vlně tak záviselo pouze na účasti v první vlně a nebylo podmíněno účastí ve všech předchozích vlnách. Panel nebyl v následujících vlnách doplňován, ve druhé a dalších vlnách se tak účastní vždy jen ti, kteří se zapojili do první vlny.

Z panelu byly postupně vyřazeny jednotky respondentů z důvodu opakovaných problémů v kvalitě vyplnění (rychlý průchod dotazníkem, chybně zodpovězené kontrolní otázky testující pozornost respondenta).

Statistická chyba

Statistická odchylka u výsledků se pro celý vzorek výzkumu (při n = 2100 a vyšším) pohybuje přibližně mezi +/- 1 p. b. u jevů s malou četností a +/- 2 p. b. u jevů s vyšší četností.

Dotazový instrument

Elektronický dotazník

Dotazování bylo realizováno na standardní platformě agentury NMS. Respondent musel u každé otázky navolit odpověď, než bylo možné přesunout se dál. Respondent se mohl vracet k předchozím otázkám.

V otázkách s horizontálními škálami volili respondenti odpověď obvykle s pomocí posuvníku, který tažením umisťovali mezi dvěma popsanými kraji škály s danou jemností. Pokyn vyzýval k výběru odpovědi tažením posuvníku na škále x až y. Popsané byly jen okraje posuvníku, nikoli body mezi (popisky okrajů škály specifikuje dotazník). Posuvník se ve svém výchozím stavu obvykle nacházel ve středu škály, ale u některých proměnných byl nastaven v jiné hodnotě (zejména snížení příjmu domácnosti, které mělo jako výchozí hodnotu nastaveno 100 % = tedy že příjem se domácnosti nesnížil). Volba odpovědi vyžadovala vždy interakci respondenta s posuvníkem, tedy nebylo možné nechat posuvník ve výchozím stavu a přejít na další otázku. Pokud chtěl respondent navolit výchozí odpověď, bylo nutné na posuvník poklepat. Využití posuvníku je v dotazníku značeno textem „slider“.

Evidence otázek

K veřejnému datovému souboru byl připraven **přehled všech otázek**, které jsou v něm obsaženy. Řazení otázek v tomto souboru vychází z kombinovaného řazení otázek v jednotlivých vlnách, tj. nová otázka je v určité vlně umístěna za tu otázku, po níž v dotazníku dané vlny následovala a která se zároveň objevila i v některé z předchozích vln. Čísla obsažená v kódu otázek odpovídají pořadí otázky, v němž byla do šetření zařazena (tj. pokud v první vlně byly otázky číslovány 1 až 50, první otázka druhé vlny dostala číselný základ kódu 51 atd.).

Řazení přehledu otázek ilustruje následující příklad:

* vlna 1: ot1, ot2, ot3, ot4, ot5
* vlna 2: ot1, ot6, ot3, ot4, ot5
* vlna 3: ot1, ot2, ot7, ot5
* kombinace: ot1, ot6, ot2, ot7, ot3, ot4, ot5

Přehled otázek zahrnuje:

* kódy proměnných odpovídající kódům v datovém souboru (s výjimkou kódů pro filtr = 66666 a kódů pro chybové hodnoty = 99998)
* znění otázky a možností odpovědí (u některých proměnných se znění otázky a/nebo odpovědí lišilo mezi vlnami – v takových případech jsou uvedena všechna využívaná znění se značením příslušných vln; v datových souborech jsou pak v popiscích proměnných a hodnot použita nejnovější znění)
* filtrační podmínku (pokud byla formulace filtru mezi vlnami měněna, opět jsou uvedeny všechny využívané varianty se značením příslušných vln)
* technická specifikace:
    - rotace položek (ROTACE)
    - využití posuvníku (SLIDER)
    - v otázkách s více možnostmi odpovědí značení odpovědi, která vynuluje ostatní případné odpovědi (NONE)
    - minimální a maximální povolené hodnoty numerických proměnných (pokud byly vyžity)
    - limity pro měkké kontroly a text souvisejícího upozornění (pokud bylo využito)
    - popis případných transformací proměnných (prováděno výjimečně)

Pro snazší orientaci ve vlnách dotazování jednotlivých otázek byl připraven rovněž **seznam proměnných** se značením vln, v nichž byly zjišťovány. Proměnné jsou v tomto seznamu řazeny stejným způsobem jako v přehledu otázek a datových souborech. Hodnota 1 ve sloupci dané vlny značí, že proměnná byla v této vlně zařazena do dotazníku.

Další dostupné zdroje

/ 1 / **Přehled otázek** zařazených do veřejného datového souboru (kombinace dotazníků z jednotlivých vln)

/ 2 / **Seznam proměnných** zařazených do veřejného datového souboru se značením vln, ve kterých byly dotazovány

Zjišťování sociodemografických charakteristik a stavu před epidemií

Základní sociodemografické charakteristiky respondentů byly obvykle zjišťovány jen jednou (velká část z nich v první vlně, tj. pohlaví, vzdělání, kraj, velikost místa bydliště). Věk byl od druhé vlny dotazován vždy, protože v první vlně nebyla obsažena otázka na věk v letech, ale pouze věkové

kategorie. Protože účast ve vlně nebyla vázána na účast ve všech předchozích, mohl respondent poprvé uvést svůj věk v letech ve kterékoli z následujících vln.

Do veřejného datového souboru byl **věk v letech** zpracován jako proměnná *age_start*, která představuje hodnotu věku respondenta v první vlně po vlně 2, které se respondent zúčastnil. Pokud se respondent žádné další vlny neúčastnil, je doplněn věk získaný z jeho profilu zadaného při registraci v Českém národním panelu. Proměnná *age_source* označuje vlnu, z níž byla hodnota čerpána (pokud hodnota vychází z profilu v ČNP, kódováno 0).

Opakovaně bylo v jednotlivých vlnách dotazováno také **složení domácnosti** kvůli potřebě filtrování a postupně se zvyšující míře podrobnosti dotazování na počty jednotlivých členů domácnosti.

Ve vybraných vlnách byly zjišťovány **charakteristiky respondenta či jeho domácnosti před nástupem epidemie v České republice** (tj. před březnem 2020), zejména: příjem domácnosti před epidemií (dotazováno ve vlně 3), počet hodin odpracovaných týdně v lednu 2020 (dotazováno ve vlně 2), vybrané zdroje příjmů před epidemií (dotazováno ve vlně 8) či omezení pracovní aktivity před epidemií (vlna 18). Hodnoty těchto ukazatelů jsou proto dostupné jen pro respondenty, kteří se účastnili příslušné vlny zjišťování.

Zkoumané týdny

Klíčové aktivity respondentů byly zjišťovány samostatně za každý ze dvou předcházejících týdnů sběru dat. Jedná se o počty osobních kontaktů, sociální aktivity, docházku na pracoviště, počet odpracovaných hodin a čerpání dovolené. Tyto referenční týdny zjišťování shrnuje *tabulka 5*.

V obdobích, kdy byl sběr dat realizován s dvoutýdenní periodicitou, byl tak pokryt celý časový úsek od minulého dotazování (první týden referenčního období představuje týden sběru dat předchozí vlny). Pokud činil odstup mezi vlnami tři týdny, není pokryt týden odpovídající týdnu sběru dat předchozí vlny.

Ve vybraných otázkách zjišťujících stav v předchozích týdnech (testování q027, změna v práci q278) bylo v textu otázky upravováno referenční období dle periodicity sběru (tj. pokud se sběr opakoval po dvou týdnech, otázka se ptala na minulé dva týdny a analogicky u tří týdnů).

Týdny, za které bylo dotazováno absolvování testu na koronavirus (ve formě týdnů zjišťováno od 21. vlny, dříve respondenti sami uváděli přesné datum), tak vždy pokrývaly období od posledního sběru dat. Pozor, při delší periodicitě sběru dat než dvoutýdenní neodpovídá týden 1 a týden 2 absolvování testů týdnům 1 a 2 kontaktů, sociálních aktivit, práce z domova atd. U testování týden 1 vždy značí nejstarší týden (počáteční v referenčním období), u ostatních otázek je to předminulý týden vzhledem k týdnu dotazování.

Tabulka 5: Týdny zkoumané v jednotlivých vlnách

Vlna	Zkoumaný týden 1 (od–do)		Zkoumaný týden 2 (od–do)		Referenční období pro týdny testování (od–do)		Počet dotazovaných týdnů testování
1	2.3.2020	8.3.2020	9.3.2020	15.3.2020	-		-
2	16.3.2020	22.3.2020	23.3.2020	29.3.2020	-		-
3	30.3.2020	5.4.2020	6.4.2020	12.4.2020	-		-
4	13.4.2020	19.4.2020	20.4.2020	26.4.2020	-		-
5	27.4.2020	3.5.2020	4.5.2020	10.5.2020	-		-
6	11.5.2020	17.5.2020	18.5.2020	24.5.2020	-		-
7	25.5.2020	31.5.2020	1.6.2020	7.6.2020	-		-
8	15.6.2020	21.6.2020	22.6.2020	28.6.2020	-		-
9	6.7.2020	12.7.2020	13.7.2020	19.7.2020	-		-
10	27.7.2020	2.8.2020	3.8.2020	9.8.2020	-		-
11	17.8.2020	23.8.2020	24.8.2020	30.8.2020	-		-
12	31.8.2020	6.9.2020	7.9.2020	13.9.2020	-		-
13	14.9.2020	20.9.2020	21.9.2020	27.9.2020	-		-
14	28.9.2020	4.10.2020	5.10.2020	11.10.2020	-		-
15	12.10.2020	18.10.2020	19.10.2020	25.10.2020	-		-
16	26.10.2020	1.11.2020	2.11.2020	8.11.2020	-		-
17	9.11.2020	15.11.2020	16.11.2020	22.11.2020	-		-
18	23.11.2020	29.11.2020	30.11.2020	6.12.2020	-		-
19	7.12.2020	13.12.2020	14.12.2020	20.12.2020	-		-
20	21.12.2020	27.12.2020	28.12.2020	3.1.2021	-		-
21	11.1.2021	17.1.2021	18.1.2021	24.1.2021	4.1.2021	24.1.2021	3
22	1.2.2021	7.2.2021	8.2.2021	14.2.2021	25.1.2021	14.2.2021	3
23	15.2.2021	21.2.2021	22.2.2021	28.2.2021	15.2.2021	28.2.2021	2
24	1.3.2021	7.3.2021	8.3.2021	14.3.2021	1.3.2021	14.3.2021	2
25	15.3.2021	21.3.2021	22.3.2021	28.3.2021	15.3.2021	28.3.2021	2
26	29.3.2021	4.4.2021	5.4.2021	11.4.2021	29.3.2021	11.4.2021	2

Kontrola a zpracování dat

Kontroly během sběru

Od druhé vlny byla na počátek dotazníku zařazena otázka zjišťující, zda se respondent účastnil minulé vlny výzkumu. Pokud odpověděl záporně, objevil se text vyzývající, aby dotazování bylo přenecháno jinému členu domácnosti („Přenechte prosím dotazování členovi Vaší domácnosti, který vyplňoval dotazník minule.“).

Část numerických proměnných měla nastavenou minimální a maximální povolenou hodnotu a/nebo limity, po jejichž překročení se objevilo upozornění, aby respondent ověřil, že zadaná odpověď je opravdu správná (tzv. „soft checks“).

Od druhé vlny byly do dotazníku zařazeny kontrolní otázky (testy pozornosti), v nichž byl respondent vyzván, aby z nabídky vybral odpověď uvedenou v pokynu. Počet zařazených testů pozornosti se mezi vlnami mírně lišil, obvykle však dotazník obsahoval čtyři. Pokud respondent vybral v testu pozornosti chybnou odpověď, byl mu zobrazen text upozorňující na důležitost pozorného vyplňování (včetně skutečnosti, že výherce tabletu se losuje jen mezi respondenty, kteří dotazník vyplňují pozorně).

Kontroly dat a čištění dat

Na souborech z jednotlivých vln šetření byla provedena **kontrola povoleného rozsahu** hodnot kategorických a numerických proměnných (dle kódovacího schématu obsaženého v dotazníku) a kontrola **konzistence dat vůči filtračním podmínkám**.

Pokud nebyla v důsledku technických problémů při ukládání odpovědi, chybné specifikace filtru či zařazení položky až po začátku dotazování zaznamenána platná odpověď, je pozorování kódováno hodnotou 99998. Při zpracování dat byl kód 999998 doplněn rovněž do proměnných, u nichž mezi vlnami docházelo k úpravám filtračních podmínek a v části vln byla použita podmínka více omezující dotázanou množinu respondentů. Kód 99998 je aplikován u těch respondentů, kteří by podmínce vyhovovali, kdyby byla formulována širším způsobem jako v jiných vlnách. Nakonec je kód 99998 využit pro případy, kdy byla začištěna (např. v rámci rekódování textových odpovědí v polootevřených otázkách) proměnná vstupující do filtrační podmínky a tato podmínka se pak na respondenta v důsledku vztahuje, ačkoli při vyplňování dotazníku tomu tak nebylo. Kód 99998 tak v obecném smyslu zachycuje situace, kdy měla být získána platná hodnota, ale nedošlo k tomu.

Proměnné, u nichž je kód 99998 použit z jiných důvodů než technických problémů při záznamu odpovědi, obsahuje *tabulka 6*.

Hodnoty numerických proměnných, které přesahovaly kontrolní limity nastavené v dotazníku (např. více než 100 hodin u proměnných q049 a q050) byly ponechány beze změny. U vybraných numerických proměnných byly však extrémně vysoké hodnoty nahrazeny chybovým kódem 99998. Tyto proměnné jsou uvedeny v *tabulce 6*.

Tabulka 6: Přehled vybraných proměnných, v nichž je využit chybový kód 99998

Kód proměnné	Popis situace
q028, q029	užší filtrační podmínka v 1. vlně než v ostatních
q017b, q021b	užší filtračních podmínka v 3. a 4. vlně než v ostatních
q049, q050	užší filtračních podmínka ve 2. vlně než v ostatních
q118	ve vlnách 10 až 14 nebyla filtrační podmínka nascriptována v souladu se zadáním v dotazníku
tQ172	užší filtračních podmínka v 11. vlně než v ostatních
q211a_18, q211a_19, q211b_18, q211b_19	proměnné zařazeny v průběhu sběru dat
q016, q020	hodnoty > 200
q018, q022, q266, q224	hodnoty > 2 000
q213–16, q265, q165	hodnoty > 10 000 000
q164	hodnoty > 200 000

U menšího počtu proměnných byly prováděny logické kontroly a související čištění. Především hodnoty proměnných q027_01b a q027_02b (zda byl respondent/člen domácnosti testován v posledních týdnech) byly začištěny podle odpovědí v navazujících otázkách q038a: pokud v nich bylo uvedeno, že respondent/člen domácnosti testován ve skutečnosti nebyl, byly v souladu s tím upraveny hodnoty proměnných q027. Podobně byly dle odpovědí v polootevřených otázkách na navštívené země a uplatňované slevy v případě potřeby upraveny nadřazené proměnné (např. respondent uvedl, že v některém předchozím týdnu cestoval do zahraničí, a poté v otevřené otázce upřesnil, že se jednalo jen o cestu po ČR; či na otázku po cestování od února 2020 uvedl starší datum).

Od 2. do 20. vlny byly respondenti dotazováni, ve kterých dnech absolvovali testy na koronavirus. V některých případech uváděná data nespadají do týdnů, za které měl respondent v této sérii otázek vypovídat. Původní hodnoty zde byly ponechány.

V datovém souboru ve formátu .sav byly jako user-missing hodnoty nastaveny kódy 66666 (filtr) a 99998 (chybějící či chybná hodnota). Odpovědi ve smyslu „nevím“ či „odmítám odpovědět“ nebyly jako user-missing nastaveny (obvykle kód 99, 99999).

Průběžně byla prováděna kontrola struktury dotázaných souborů dle výše příjmu domácnosti, sociálních aktivit a dalších charakteristik (pomocí srovnání distribucí odpovědí z první vlny u vzorku účastníků první vlny se vzorkem účastníků aktuální následné vlny). Původní příjmová struktura ani další původní parametry respondentů se během vln významně nemění.

Kódování otevřených odpovědí

V textových proměnných v polootevřených otázkách na navštívené země (q011, q169) a uplatňované slevy na dani (q264) byly opraveny jazykové chyby a odpovědi byly upraveny tak, aby se obsahově unikátní výpovědi vyskytovaly jen jednou. Pokud respondent uvedl více navštívených zemí, jsou od sebe odděleny čárkou a řazeny v abecedním pořadí (podobně u typů slev na dani). Sjednocen byl rovněž formát uváděných dat návštěv zemí (většina respondentů uvedla jedno datum, a pokud se vyskytoval interval, bylo vloženo jen koncové datum). Pokud textová odpověď v polootevřených otázkách odpovídala nabízené kategorii, byla odpověď rekódována do uzavřené možnosti.

Odpovědi ve formátu dat (date)

Proměnné obsahující informaci ve formě data (q038g, q011b) jsou formátovány jako stringové proměnné, aby bylo možné zachovat kódy neplatných hodnot. V SPSS je možné k jejich jednoduchému převodu na formáty DATE použít např. příkaz ALTER TYPE.

Vyřazování případů

Z datových souborů za jednotlivé vlny byli vyřazováni respondenti s mimořádně krátkým časem průchodu dotazníkem a/nebo chybně zodpovězenými kontrolními otázkami. V kontrolních otázkách (test pozornosti) byl respondent vyzván, aby z nabídky vybral odpověď uvedenou v pokynu. Vyřazování probíhalo dle následujících pravidel:

* vyplnění za méně než 4 minuty NEBO
* vyplnění za méně než 5 minut a více než jeden test pozornosti špatně NEBO
* všechny zařazené testy pozornosti špatně.

Tato pravidla byla uplatňována od druhé vlny (dotazník první vlny byl výrazně kratší, testy pozornosti byly zařazeny až ve druhé vlně). Počet zařazených testů pozornosti se mezi vlnami mírně lišil, obvykle však byly čtyři.

V jednotlivých vlnách byly takto obvykle vyřazeny nižší jednotky pozorování. Účastníci, jejichž odpovědi byly z datového souboru vyřazeny více než třikrát v posledních pěti vlnách z důvodu nekvalitního vyplnění, byli z panelu trvale vyloučeni a nebyli oslovováni k další účasti. Před vlnou 26 bylo takto vyřazeno trvale sedm panelistů.

Průběžně byla také kontrolována stabilita sociodemografických ukazatelů mezi vlnami. Během prvních deseti vln bylo vyloučeno sedm respondentů, u nichž docházelo k opakovaným změnám v základních sociodemografikách (věk, struktura domácnosti).

Anonymizace

Textové proměnné zařazené do veřejného souboru byly zkontrolovány, zda neobsahují detailnější informace o respondentovi, které by mohly samostatně nebo v kombinaci s dalšími údaji vypovídat o jeho identitě.

Identifikační čísla respondentů Českého národního panelu byla změněna na náhodná identifikační čísla.

Nejpodrobnější úroveň geografické identifikace respondentů v datech představuje kraj.

Paradata

Datový soubor obsahuje proměnné *start* a *end*, které odpovídající začátku a konci vyplňování dotazníku, tedy okamžikům, kde respondent do elektronického dotazníku vstoupil a kdy jej kompletně dokončil (do datového souboru zařazeny pouze ty kompletně dokončené). Proměnná *filling_time* je vypočtena jako rozdíl začátku a konce dotazování. Respondent ale nemusel vyplňování strávit celý tento čas, mohl mít například dotazník otevřený, zatímco se věnoval jiným činnostem. Proměnnou *filling_time* tedy není možné interpretovat jako čas, po který respondent s dotazníkem přímo interagoval.

Proměnné *start* a *end* jsou dostupné od třetí vlny výzkumu, proměnná *filling_time* se nachází v každé vlně.

Vážení

Vzhledem k nadvýběru větších obcí a dále nestejné pravděpodobnosti opakované účasti různých sociodemografických skupin, která je typická pro výběrová šetření (vyšší šance účasti u osob s vyšším vzděláním, starších atd.), se sociodemografické složení respondentů dotázaných v jednotlivých vlnách odchyluje od populačního a je vhodné jej korigovat post-stratifikačním vážením. Po použití vah lze výsledky šetření považovat za kvótně reprezentativní pro dospělou populaci ČR.

Data byla vážena vzhledem k populačním distribucím (dle dat Českého statistického úřadu) dle níže uvedených charakteristik:

Tabulka 7: Přehled proměnných pro post-stratifikační vážení

Proměnná	Vlna zjišťování	Počet kategorií
Vážení dle distribuce údajů ČSÚ		
Pohlaví	vlna 1	2
Věk	aktuální vlna	6
Vzdělání	aktuální informace z profilu respondenta v ČNP	4
Velikost místa bydliště	vlna 1	7
Kraj	vlna 1	14
Ekonomický status v první vlně šetření, březen 2020	vlna 1	6
křížení věku a vzdělání	viz výše	3 x 3
křížení věku a pohlaví	viz výše	3 x 2
Vážení dle distribuce z 1. vlny šetření		
Náplň práce v první vlně šetření, březen 2020	vlna 1	6

Pro zajištění srovnatelnosti mezi vlnami šetření byl vzorek ve druhé a dalších vlnách vážen také podle distribuce typu práce zjištěné ve vzorku z první vlny šetření (proměnná q014, 9 kategorií). Tedy struktura pracovní náplně respondentů vztahující se k první vlně je v druhé a dalších vlnách vážena na proporce odpovídající celému vzorku první vlny. Kategorie této proměnné se vztahují k rizikovým skupinám zaměstnání z pohledu šíření nákazy a fixování této pracovní struktury by proto mělo podpořit srovnatelnost ukazatelů epidemiologicky významného chování mezi vlnami.

Vážení bylo provedeno v software R pomocí algoritmu **kvadratického programování**, který umožňuje nastavit povolený rozsah vah (nastaveno na 0,3 až 3) a maximální povolenou odchylku populační a

pozorované četnosti (nastaveno na 1 p. b.). K apriornímu omezování rozsahu vah a kvadratickému programování viz např.:

Isaki, C. T., Tsay, J. H., Fuller, W. A. 2004. Weighting Sample Data Subject to Independent Controls. Survey Methodology 30(1): 35–44.

Valliant, R., Dever, J. A. Kreuter, F. 2018. Practical Tools for Designing and Weighting Survey Samples. Springer.

Tabulka 8: Populační distribuce využité při vážení

Kvótní charakteristika 1	Kvótní charakteristika 2	Zastoupení v populaci (%)
Pohlaví		
Muž		48,8
Žena		51,2
Věk		
18–24 let		7,9
25–34 let		16,1
35–44 let		19,9
45–54 let		17
55–64 let		15,1
65 a více let		24,1
Nejvyšší dosažené vzdělání		
Základní		10,8
Střední bez maturity		34,3
Střední s maturitou		35,1
Vysokoškolské		19,8
Kraj		
Praha		12,2
Středočeský		12,5
Jihočeský		6
Plzeňský		5,5
Karlovarský		2,8
Ústecký		7,7
Liberecký		4,1
Královéhradecký		5,2
Pardubický		4,9
Vysočina		4,8
Jihomoravský		11,2
Olomoucký		6
Zlínský		5,5
Moravskoslezský		11,5
Velikost místa bydliště		
Do 999 obyvatel		17
1000 až 1999 obyvatel		9,6

2000 až 4999 obyvatel		11,4
5000 až 19 999 obyvatel		18,4
20 000 až 49 999 obyvatel		11,6
50 000 až 99 999 obyvatel		9,9
100 000 obyvatel a více		22,1
Věk	**Vzdělání**	
18–34 let	Bez maturity	8,3
18–34 let	S maturitou	9,8
18–34 let	Vysokoškolské	5,9
35–54 let	Bez maturity	15,1
35–54 let	S maturitou	13,6
35–54 let	Vysokoškolské	8,2
55 a více let	Bez maturity	21,9
55 a více let	S maturitou	11,6
55 a více let	Vysokoškolské	5,6
Ekonomický status		
Zaměstnanci		47,5
Podnikatelé		9,9
Nezaměstnaní		2,7
Nepracující důchodci		29,6
Žáci, studenti, učni		5,7
Mateřská a jiné		4,6
Věk	**Pohlaví**	
18–34 let	Muž	12,3
18–34 let	Žena	11,6
35–54 let	Muž	19
35–54 let	Žena	18
55 a více let	Muž	17,5
55 a více let	Žena	21,6
Pracovní náplň v březnu 2020		**Zastoupení ve vzorku 1. vlny**
Zdravotnictví		2,4
Pečovatelská profese		0,6
Školství/škola		3,8
Ve službách, pohostinství		4,5
Řidič/ka hromadné dopravy, taxi		0,9
Úřad		3,4
Zaměstnání, kde přicházím do kontaktu s velkým množstvím lidí		14,1
Nic z výše uvedeného		27,8
Nepracuje		42,6

Struktura datového souboru

Pro usnadnění práce s panelovým rozměrem dat byly připraveny dva formáty datového souboru, který spojuje data za všechny dosud zveřejněné vlny šetření. Tyto formáty odpovídají dvěma obvyklým formátům užívaným pro panelová data: široký („wide“) a dlouhý („long“).

Ve spojených souborech nebylo prováděno žádné doplňování dat z jedné vlny do jiné. Hodnoty proměnných figurují v datech tedy jen v těch vlnách, v nichž byly dané proměnné dotázány.

V obou typech datových souborů jsou na počátku uvedeny základní identifikace:

* identifikátor respondenta *(caseid)*
* proměnné *ucastXY* značící účast respondenta v dané vlně (1 = Ano, Ne)
* dlouhý soubor obsahuje na počátku také indikátor vlny *(vlna)*

Široký formát („wide“)

Jeden řádek datového souboru představuje jednoho respondenta a jeho odpovědi ve všech vlnách. Soubor tedy zahrnuje 3101 pozorování odpovídající tomuto počtu respondentů, kteří se zúčastnili první vlny šetření (a nebyli vyřazeni z důvodu výraznějších nesrovnalostí v sociodemografických proměnných v prvních deseti vlnách).

Prázdné hodnoty *(system missing)* značí, že respondent se neúčastnil dané vlny dotazování.

Každá proměnná je obsažena tolikrát, v kolika vlnách byla dotázána. Respondent je identifikovaný pomocí identifikátoru *caseid*, vlnu dotazování označuje předpona proměnné ve tvaru v01, v02, v10 atd. Vlna dotazování je rovněž uvedena na konci popisku příslušné proměnné *(variable label).*

Proměnné jsou řazeny po jednotlivých vlnách, tedy nejprve jsou uvedeny všechny proměnné vztahující se k první vlně (včetně vah), dále všechny proměnné z druhé vlny atd. Proměnné jsou uvnitř vln řazeny podle kombinovaného pořadí ve všech vlnách, které je užíváno v přehledu otázek a seznamu proměnných (s vyloučením prázdných sloupců za proměnné, které v dané vlně nefigurovaly).

Dlouhý formát („long“)

Jeden řádek datového souboru představuje odpovědi respondenta v jedné vlně. Respondent je identifikovaný pomocí identifikátoru *caseid*, vlnu označuje proměnná *vlna*. Každé pozorování v datech

tedy přestavuje unikátní kombinaci respondenta a vlny. Datový soubor z prvních 26 vln šetření obsahuje 59 741 pozorování.

Prázdné hodnoty *(system missing)* značí, že daná proměnná nebyla v dané vlně dotazována.

Proměnné jsou v datovém souboru uspořádány v tematických blocích, které respektují řazení bloků v dotaznících jednotlivých vln a rovněž pořadí proměnných v tematických blocích odráží umístnění proměnné v dotazníku (k řazení proměnných více viz v části Evidence otázek)

Syntaxe pro vytváření souborů

Pro převod mezi širokým a dlouhým formátem a jejich rozdělení na jednotlivé vlny byl připraven .sps script využívající Python3 plug-in, který je dostupný ve verzích SPSS 24 a vyšších.

/ 3 / blok 1: převod wide na long

/ 4 / blok 2: převod long na wide

/ 5 / blok 3: rozdělení wide na jednotlivé vlny

/ 6 / blok 3: rozdělení long na jednotlivé vlny