Daty podložené omyly

Ondřej Vencálek

1 Úvod – analýza dat v našich životech

V březnu roku 2021, stejně jako o rok dříve – v březnu 2020, prožívali občané České republiky tzv. lockdown, jehož zavedení se v obou případech opíralo o analýzu dat. Těžko si představit přesvědčivější doklad prostého tvrzení, že *analýza dat významně ovlivňuje náš každodenní život.* Jak moc je analýza dat součástí našeho každodenního života snad běžně ani nevnímáme. A přece, spamové filtry v našem emailovém prohlížeči, záhadní permoníci – tzv. recommender systems (`https://en.wikipedia.org/wiki/Recommender_system`) – kteří odkudsi vykutají a zobrazí nám nabídku produktů (třeba knih), která jako by nám byla na míru ušitá, či jen „obyčejná" předpověď počasí, to jsou jen namátkově vybrané příklady výsledků analýzy dat, se kterými se setkáváme takřka denně. Čtenáři, který by si chtěl o pestrosti použití analýzy dat v běžném životě udělat lepší představu lze doporučit populárně-naučné publikace Jeffreyho S. Rosenthala [5] či Natea Silvera [6].

Prostřednictvím analýzy dat hledáme odpovědi na otázky, které nás zajímají. Jaké jsou to otázky ve výše uvedených příkladech využití analýzy dat? Poskytovatele emailových služeb zajímá, jestli příchozí email je obtěžující spam, který by měl být smazán, nebo důležitá zpráva, která naopak smazána být nesmí. Obdchodníka s knihami zajímá, kterou knihu má zákazníkovi nabídnout, aby jej nabídka zaujala. A konečně velké množství lidí zajímá, jaké bude v nejbližší době počasí, aby podle této předpovědi vhodně upravili svůj oděv (aby se tzv. přioblékli, bude-li chladněji) či dokonce program.

Ačkoliv analýza dat hraje v dnešní době důležitou roli, je třeba si neustále připomínat, že proces získávání znalostí na základě analýzy dat má svá úskalí. Je třeba být neustále ve střehu – omylům a chybám se často bohužel nevyhnou ani profesionální datoví analytici, natož poučení laici, kteří dnes mají k dispozici celou řadu softwarových nástrojů pro nejrůznější analýzy.

2 Analýzy, které „zavřely" naši zemi

Připomeňme zde prognózu, kterou tehdejší ministr vnitra ČR Jan Hamáček (`https://cs.wikipedia.org/wiki/Jan_Ham\%C3\%A1\%C4\%8Dek`) zdůvodňoval na konci února 2021 nutnost tzv. tvrdého lockdownu. Tato prognóza poskytnutá dne 24. února 2021 jako podklad pro jednání Národní ekonomické radě vlády (`https://www.vlada.cz/cz/ppov/nerv_2020/narodni-ekonomicka-rada-vlady-182438/`) byla založe-

 DOI: https://doi.org/10.11118/978-80-7509-898-6-0020

na na poměrně sofistikovaném modelu [7]. Podle modelu měl počet nových případů „kulminovat“ přibližně v půlce března, kdy měl dosahovat hodnot cca 20 tisíc nových případů denně (a to za předpokladu zavedení lockdownu na začátku března, jak k němu opravdu došlo). Realita však zdaleka nebyla tak dramatická. Jak se z dostupných dat přesvědčíme (viz níže), počty nových případů dosáhly maxima již 5. března, tedy dříve, než se mohl efekt lockdownu zavedeného 1. března projevit, a to na úrovni přibližně 12 tisíc nových případů denně. Všechna tvrzení o potřebě zavedení přísných opatření tak byla empiricky vyvrácena – k obratu ve vývoji epidemie došlo ještě předtím, než se efekt těchto opatření mohl projevit (viz text Angeliky Bazalové v časopise Reflex [1]). Poznamenejme, že tím ovšem ani není dokázáno, že by opatření neměla žádný efekt.

Na to, že predikce počtu nových případů „nevyšla“, upozornil již 11. 3. reportér Petr Holub [2]. Hlavního autor prognózy – Martin Šmíd, ve své veřejné reakci na Holubův článek [8] správně uvedl, že „Vzhledem k tomu, že jde o model stochastický (zahrnující náhodu), nelze předpokládat, že poskytne přesnou předpověď. Čáry v grafech znázorňují jen jakousi střední předpověď. Tyto bodové předpovědi by správně měly být doplněny mezemi nejistoty (konfidenčními pásy). Tyto hodnoty náš model počítá, v prezentaci pro NERV bohužel nejsou uvedeny [...]“. Selhání predikce počtu nově potvrzených případů však musel uznat: „O selhání stochastického modelu se dá mluvit až v případě, kdy se předpovědi systematicky ocitají mimo tyto pásy, což tady bohužel nastalo u počtu nových případů.“

O podstatně jednodušším modelu, který vedl k lockdownu v březnu 2020, se můžete více dozvědět v knize Pandemie [4].

2.1 Data o COVID – cvičení v R

Chceme-li se přesvědčit, jaký byl skutečný průběh počtu nově diagnostikovaných případů nákazy virem SARS-CoV-2, původcem onemocnění COVID-19, můžeme tak učinit vizualizací dat z některé z veřejně dostupných databází. V České republice počty nově diagnostikovaných případů denně zveřejňuje Ústav zdravotnických informací a statistiky ČR na stránce `https://onemocneni-aktualne.mzcr.cz/covid-19`. Stejné údaje najdeme i v mezinárodní databázi Our World in Data (`https://ourworldindata.org/coronavirus`). Datový soubor `owid-covid-data.csv` můžeme stáhnout např. z repozitáře GitHub: . (`https://github.com/owid/covid-19-data/tree/master/public/data`)

Data načteme pomocí příkazu `read.csv()` a poté vybereme jen data o České republice za období od 24. února do konce dubna 2021.

Pro práci s daty v R-ku lze v současné době doporučit balíček `dplyr`, viz [9], případně balíček `tidyr`, viz [10]. Čtenářům, kteří s těmito balíčky ještě nepracovali, doporučujeme využít některého z „taháků“, tzv. cheat sheets, které na prostoru dvou stran (tedy jednoho listu) formátu A4 přehledně shrnují to nejpodstatnější, co je k dané problematice třeba vědět. Osvědčeným tahákem je Data Wrangling with dplyr and tidyr Cheat Sheet (`https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf`). Novější verze, které jsou zvlášť pro balíčky `dplyr` a `tidyr`, lze nalézt v přehledu taháků . (`https://www.rstudio.com/resources/cheatsheets/`)

```
library(dplyr)

 data    = read.csv("owid-covid-data.csv")
 data.cz = data %>% filter(location=="Czechia")
 data.cz.3_4.2021 = data.cz %>% filter((date>="2021-02-24")&
                                       (date<="2021-04-30"))
```

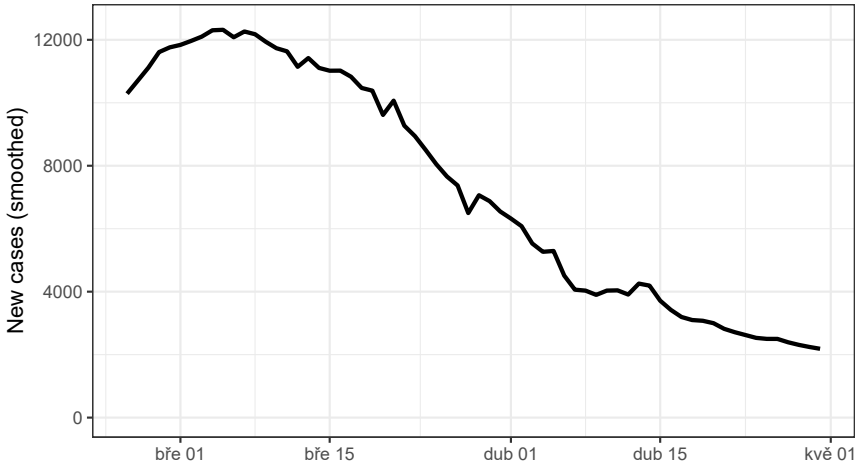
Vývoj počtu nově diagnostikovaných případů můžeme vizualizovat pomocí balíčku `ggplot2`, viz [11]. Opět doporučujeme tahák, tentokrát Data visualization with `ggplot2` cheatsheet. Budeme pracovat se sloupcem `new_cases_smoothed`. Jde o průměrné hodnoty počtu nových případů vždy za posledních 7 dní. Tímto průměrováním jsou z dat odstraněny pravidelné výkyvy s periodou jednoho týdne související s tím, že o víkendu se méně testuje a je tedy i méně zjištěných nových případů. Graf můžeme vykreslit následovně:

```
library(ggplot2)

 ggplot(data.cz.3_4.2021, aes(x=as.Date(date))) +
   geom_line( aes(y=new_cases_smoothed))
```

Obrázek 2.1 byl získán úpravou (rozšířením) tohoto základního příkazu:

```
 ggplot(data.cz.3_4.2021, aes(x=as.Date(date))) +
  geom_line( aes(y=new_cases_smoothed),size=1) +
  ylim(c(0,12500)) +
  labs(y="New cases (smoothed)", x="") +
  theme_bw()
```



Obr. 1: *Průměrné denní počty nově diagnostikovaných případů nákazou virem SARS-CoV-2 v České republice v období 24. 2. 2021 až 30. 4. 2021 (7-denní průměry).*

3 Je lépe chválit nebo kárat?

Je dobré si uvědomit, že proces získávání znalostí z dat není jen něčím, co dělají datoví analytici. To že si utváříme názor na základě svých pozorování (tedy „dat" ukládaných do naší paměti), je přece přirozené. Také v tomto procesu se však naše mysl dopouští mnoha chyb, jak popisuje v knize *Myšlení: rychlé a pomalé* [3] psycholog Daniel Kahneman.

V části nazvané Heuristiky a zkreslení věnoval Kahneman jednu kapitolu jevu nazývanému regrese k průměru (viz [3], str. 189–199). Kapitolu uvádí příhodou z doby, kdy v rámci kuzu psychologie vysvětloval instruktorům izraelského vojenského letectva zásadu, že „odměna za zlepšený výkon funguje lépe než trest za chybný výkon". Jeden ze zkušených instruktorů však byl opačného názoru. Argumentoval takto: „Vždycky jsem si dával záležet, abych lidi za pěkně provedený manévr pochválil, a příště ho vždycky provedli hůř. Ale když jsem je za špatně provedený manévr seřval, většinou se pak zlepšili. Tak mi netvrďte, že chvála pomáhá a trestání ne. Moje zkušenost je přesně opačná." Kahneman uvádí, že byť byla instruktorova pozorování správná, závěr ohledně účinnosti odměny a trestu z těchto pozorování vyvozený byl naprosto špatný. Nezpochybnitelné vysvětlení instruktorovy zkušenosti je známo jako *regrese k průměru*. Tento princip nyní objasníme.

3.1 Fenomén regrese k průměru a jeho ilustrace

Představme si, že dostaneme za úkol se v krátkém časovém intervalu naučit 100 pro nás zatím neznámých slovíček v cizím jazyce. Zvládneme jen 80, zbylých 20 bohužel ne. Jak asi dopadneme v testu, kde máme přeložit 20 slovíček náhodně vybraných ze zadané stovky slov? Umíme 80 ze 100, tedy 80 % všech slovíček, takže bychom očekávali 80% úspěšnost i v testu. Té dosáhneme, když z 20 slovíček správně přeložíme 16. Může se ale stát, že budeme mít štěstí a z vybraných slovíček správně přeložíme více než 16. Jelikož umíme přeložit 80 slovíček, může se dokonce stát, že správně přeložíme všech 20 slovíček. Při hodně velké smůle by se nám však teoreticky mohlo stát, že budeme tázáni právě na těch zbylých 20 slovíček, která neumíme. To by však musela být opravdu neuvěřitelně velká smůla! Jak velká? To umíme spočítat. předpokladu, že všechna slovíčka mají stejnou šanci, že budou vybrána, se

Počet vybraných slovíček, které umíme přeložit, se řídí tzv. *hypergeometrickým rozdělením*. Pravděpodobnost, že jich bude právě x, kde za x můžeme dosadit libovolné celé číslo od nuly po 20, je rovna

$$p(x) = \frac{\binom{80}{x} \cdot \binom{20}{20-x}}{\binom{100}{20}}.$$

Dosadit do tohoto vzorce postupně všechny možné hodnoty x $(0, 1, 2, \ldots, 20)$ můžeme v R-ku jediným příkazem, a to s využitím funkce `dhyper()`:

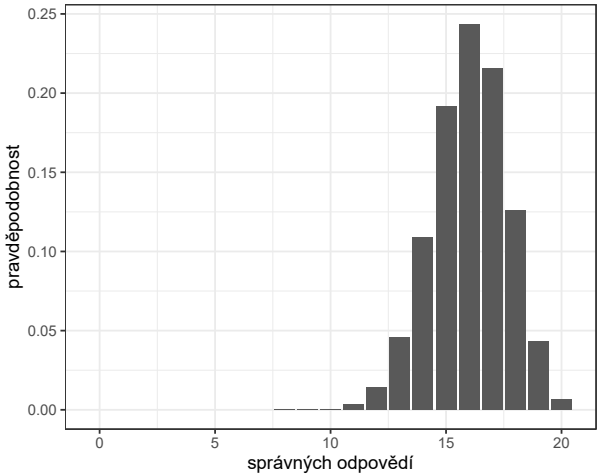
```
dhyper(0:20,80,20,20)
```

Vypočtené pravděpodobnosti můžeme pro přehlednost zobrazit prostřednictvím sloupcového grafu, viz obrázek 3.1. Uvádíme zde kód potřebný k jeho vygenerování v softwaru R (opět používáme balíček `ggplot2`).

```
pocet.spravnych = 0:20
pravdepodobnost = dhyper(0:20,80,20,20)
data = data.frame(pocet.spravnych,pravdepodobnost)

ggplot(data,aes(pocet.spravnych,pravdepodobnost)) +
   geom_col()+
   labs(x = "správných odpovědí", y = "pravděpodobnost") +
   theme_bw()+
```

Na obrázku 3.1 vidíme, že vskutku nejpravděpodobnějším výsledkem testu je zisk 16 bodů z 20. Možná nás však překvapí, že pravděpodobnost tohoto výsledku je „jen“ asi



Obr. 2: *Rozdělení pravděpodobnosti počtu správně přeložených slovíček.*

24 %. Dosti pravděpodobný je také zisk 17 bodů (asi 21 %) nebo 15 bodů (asi 19 %). Více než 10% pravděpodobnost má taky zisk 18, resp. 14 bodů. Výše zmiňovaná možnost, že bychom uměli správně přeložit všech 20 slovíček, má celkem malou pravděpodobnost (0,66 %). Rovněž vidíme, že správné přeložení pouze 10 či dokonce ještě méně vybraných slovíček je velmi nepravděpodobné a muselo by tedy být pokládáno za opravdu velkou nepřízeň štěstěny.

Představme si nyní školní třídu, která má 30 žáků. Každý z těchto žáků se naučil právě 80 ze 100 slovíček[1].

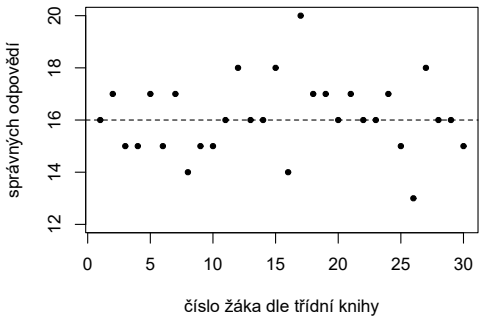
Jak asi dopadnou v testu připravenosti, v němž budou mít za úkol přeložit 20 náhodně vybraných slovíček? Výše vypočtené pravděpodobnosti naznačují, že přibližně čtvrtina žáků (7 až 8) by měla mít 16 správně přeložených slov. Hodně bude i žáků s 15 či 17 správnými překlady, neboť součet $p(15) + p(16) + p(17) = 0{,}19 + 0{,}24 + 0{,}21 = 0{,}64$; čekáme tedy, že kolem dvou třetin žáků (tj. dvacet) by mělo mít výsledek v rozmezí 15 až 17 správných odpovědí.

[1] Snad si můžeme dovolit předpokládat, že slůvka jsou pro zapamatování přibližně stejně obtížná. Debata o tomto předpokladu by jistě byla zajímavá a mohla by přinést vylepšení uvažovaného jednoduchého modelu. K tomu bychom však potřebovali nasbírat skutečná data.

Pomocí generátoru (pseudo)náhodných čísel můžeme pro každého žáka vygenerovat počet získaných bodů. Generujeme z hypergeometrického rozdělení, proto použijeme příkaz `rhyper()`. Data pak vizualizujeme.

```
set.seed(8)
test1 = rhyper(30,80,20,20)

plot(test1,pch=20,
   xlim=c(1,30), ylim=c(16-4,16+4),
     ylab="správných odpovědí", xlab="číslo žáka")
abline(h=16,lty=2)
```



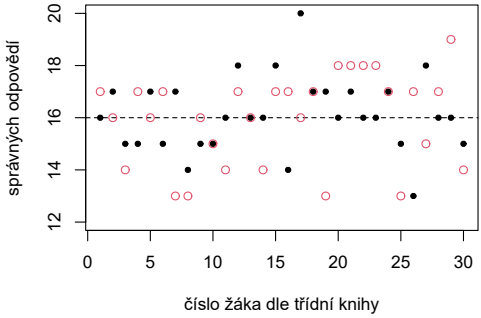
Obr. 3: *Výsledky 30 stejně dobře připravených žáků v testu o 20 otázkách.*

Na obrázku 3.1 můžeme porovnat svá výše popsaná očekávání s „realitou“. Potvrdilo se, že nejčastějším výsledkem je 16 správných odpovědí. Tohoto výsledku dosáhlo 9 žáků. Nejlepší výsledek dosáhl žák, který je v třídní knize veden pod číslem 17. Měl správně všech 20 odpovědí. Naopak nejhůře dopadl žák č. 26, který měl jen 13 správných odpovědí. Uvědomme si, že přitom oba tito žáci (stejně jako i všichni ostatní žáci) byli připraveni stejně dobře (uměli 80 ze 100 slovíček). Jen jeden měl více štěstí (a získal o 4 body více, než by odpovídalo jeho schopnostem) a druhý měl velkou porci smůly (a získal o 3 body méně, než odpovídá jeho schopnostem). Vyučující by si měl být vědom toho, že počet správných odpovědí, kterého žák v testu dosáhl, často není přesným obrazem žákových schopností (znalostí), ale je do značné míry určován náhodou. Kdo z nás by předpokládal, že žák, který dosáhl 20 bodů z 20 možných, byl ve skutečnosti stejně dobře připraven jako ten, který měl jen 13 bodů z 20? Jak správně porozumět dosaženým výsledkům, říká následující „rovnice“:

$$\begin{aligned} \text{výsledek} &= \text{schopnost} + \text{štěstí} \\ 20 \text{ bodů} &= 16 \text{ bodů} + 4 \text{ body} \\ 13 \text{ bodů} &= 16 \text{ bodů} - 3 \text{ body} \end{aligned}$$

Představme si nyní, že se vyučující rozhodne s žáky, z nichž každý stále umí přesně 80 ze 100 slovíček, napsat ještě jeden test o 20 položkách. Opět vybírá náhodně 20 ze 100 slov, přitom nezáleží na tom, jestli už slovo bylo vybráno v prvním testu. Vygenerujme opět výsledky žáků a zobrazme je do grafu spolu s výsledky prvního testu. Můžeme to udělat tak, že do původního grafu přidáme červeně znázorněné symboly odpovídající výsledkům druhého testu.

```
test2 = rhyper(30,80,20,20)
points(test2,col=2,pch=21,cex=1)
```



Obr. 4: *Výsledky 30 stejně dobře připravených žáků v prvním (černě) a druhém (červeně) testu o 20 otázkách.*

Srovnání výsledku prvního a druhého testu vidíme na obrázku 3.1. Z něj je patrné, že ti, kterým v prvním testu štěstí přálo nejvíc (dosáhli extrémně dobrého výsledku), již podruhé tolik štěstí neměli – nejlepší žák si pohoršil z 20 bodů na 16, ani nikdo ze tří žáků, kteří měli 18 bodů, tento výkon při druhém testu nezopakoval. Naproti tomu žák s nejhorším výsledkem v prvním testu si ve druhém polepšil ze 13 bodů na 17. Je dobré si povšimnout i toho, že žák s poměrně dosti špatným výsledkem prvního testu (žák č. 8)

měl ve druhém testu ještě větší smůlu a pohoršil si ze 14 bodů na 13. Tento případ je však spíše výjimečný. Ukazuje, že „návrat k průměru“, který jsme pozorovali v ostatních případech, kdy po extrémně dobrém výsledku došlo k zhoršení a po extrémně špatném došlo ke zlepšení (aniž by se měnily schopnosti žáků), je pravděpodobný, nikoliv však nevyhnutelný.

Všimněme si, že ke skutečnému zlepšení či zhoršení schopností studentů nedošlo. Stále jsme uvažovali situaci, kdy všichni umí přesně 80 ze 100 slovíček. Pozorované zhoršení (výsledek druhého testu byl horší než výsledek prvního) či naopak zlepšení, bylo tedy jen zdánlivé a bylo „způsobeno“ vlivem náhody (štěstí) na výsledek testu. Učitel, který by žáky s nejvyšším počtem bodů v prvním testu pochválil a naopak ty nejhorší by pokáral, a pak pozoroval „negativní efekt“ pochvaly a „pozitivní efekt“ pokárání, se bude v hodnocení ekeftivity svého zásahu (pochvaly, resp. pokárání) mýlit, jako se mýlil instruktor letectva vzpomenutý na začátku této kapitoly.

Čtenáře, který stále ještě není přesvědčen o zdánlivosti efektu pochvaly a pokárání, vybízíme: Vezměte hrací kostku a tu pochvalte vždy, když padne šestka. Uvidíte, že v příštím hodu se nezlepší, častěji se dokonce zhorší. A taky té kostce vynadejte, když padne jednička. Uvidíte, jak se v dalším hodu zlepší (nebo alespoň nezhorší)! Opravdu věříte, že to zhoršení (resp. zlepšení) je kvůli (díky) tomu, že kostku chválíte (káráte)?

Kahneman výše popsaný problém v úsudku ohledně efektivity pochval a kárání komentoval takto: „Zpětná vazba, které nás život vystavuje, je zvrácená. Protože máme tendenci být na jiné lidi milí, když nás potěší, a nepříjemní, když nás nepotěší, ze statistického hlediska jsme trestáni za to, že jsme milí, a odměňováni za to, že jsme nepříjemní.“

4 Závěr

V tomto příspěvku jsme se snažili poukázat na některá úskalí procesu získávání znalostí na základě analýzy dat. Připomněli jsme selhání predikce (podložené daty) sofistikovaného modelu pro předpověď počtu nově diagnostikovaných infekcí virem SARS-CoV-2 z března 2021. Na příkladu hodnocení efektu pochvaly a kárání jsme pak ukázali, jak obtížným úkolem je hodnocení efektu našeho konání a k jakým omylům je naše mysl náchylná. Možná nám to pomůže pochopit obtíže, které máme při vyhodnocování efektivity různých restrikcí, jejichž cílem je různě ovlivňovat rychlost šíření nákazy. Pravdou je, že restrikce se zavádí v situaci, kdy počty nakažených (drameticky) rostou, a naopak „rozvolňuje se“, když epidemie ustupuje. Znamená to však, že opatření fungují? Ne nutně. Tím ovšem netvrdíme, že nefungují. Jen upozorňujeme, že to, co se někomu zdá být zřejmé, zdaleka zřejmé být nemusí.

Snažili jsme se náš výklad doplnit o ukázky práce se softwarem R. Pokud se čtenář dozvěděl něco nového o jeho použití, tím lépe.

Literatura

1. Bazalová, A. Jenom lockdowny virus nezastaví. *Reflex*, č. 12, roč. 2021, str. 9–11.
2. Holub, P. *Hamáčkova předpověď o 20 tisících nakažených nevyšla. Proč?* [online]. 11.3.2021 [cit. 2.11.2021]. Dostupné z:8 . (`https://www.seznamzpravy.cz/clanek/hamackova-predpoved-o-20-tisicich-nakazenych-nevysla-proc-146362`)
3. Kahneman, D. *Myšlení: rychlé a pomalé*. Brno: Jan Melvil, 2012. Pod povrchem. ISBN 978-80-87270-42-4.
4. Kubal, M., Gibiš, V. *Pandemie*. Praha: Kniha Zlín, 2020. ISBN 978-80-7662-047-6.
5. Rosenthal, J. S. *Zasažen bleskem: podivuhodný svět pravděpodobností*. Praha: Academia, 2008. Galileo. ISBN 978-80-200-1645-4.
6. Silver, N. *Signál a šum: mnoho předpovědí selže, některé ne*. Praha: Paseka, 2014. ISBN 978-80-7432-440-6.
7. Šmíd, M. et al. *SEIR Filter: A Stochastic Model of Epidemics*. medRxiv, 2021. Dostupné z doi: 10.1101⊠.02.16.21251834
8. Šmíd, M. *Ještě o Hamáčkově předpovědi...* [online]. 12.3.2021 [cit. 2.11.2021]. Dostupné z: `https://www.bisop.eu/jeste-o-hamackove-predpovedi/`
9. Wickham, H., François, R., Henry, L., Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. `https://CRAN.R-project.org/package=dplyr`
10. Wickham, H. (2021). tidyr: Tidy Messy Data. R package version 1.1.4. `https://CRAN.R-project.org/package=tidyr`
11. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

Autor

Mgr. Ondřej Vencálek, Ph.D., Česká statistická společnost, Na padesátém 81, 100 82 Praha 10