COVIDOWE PIĄTKI SHL
PREZENTACJA Z 20.03.2021

Relacje między zachorowaniami COVID w Polsce – a wynikami z Google Trends (GT)

Badanie ciągłe od 29 września 2020

MARIUSZ KWIATKOWSKI
@altymis

WSTĘP

- Co do zasady – jeśli pojawia się tu odniesienie do danych zewnętrznych – podaję źródło
- Dla danych dotyczących liczby chorych, zajętych respiratorów itp. – czerpię je na użytek tej prezentacji z arkuszy Michała Rogalskiego, które są najpełniejszą i rzetelną bazą takich danych
- https://docs.google.com/spreadsheets/d/1ierEhD6gcq51HAm433knjnVwey4ZE5DCnu1bW7PRG3E/edit#gid=0

Założenie: istnieje związek pomiędzy wyszukiwaniami w Google – a liczbą zachorowań.

- To, co pozwala mi uznać taki związek za istniejący i mierzalny – to 30 lat doświadczeń z informatyką i 20 lat doświadczeń związanych z projektowaniem stron www, marketingiem internetowym, zgodnym z SEO pozycjonowaniem stron internetowych. I duży poziom własnej „ciekawskości” ☺

- Jesienią – słyszałem o badaniach w USA z maja 2020 – ale nie analizowałem ich co do metodyki i wzorów (nie próbowałem ich odtwarzać)
- Ponieważ można się spodziewać, że „doktor Google” przywoływany jest na pierwszym etapie objawów – będzie to wyprzedzało napływ chorych do systemu opieki zdrowotnej o około dwa tygodnie. (Badania amerykańskie sugerowały opóźnienie 16 dni)

link do analizy w Nature -https://www.nature.com/articles/s41598-020-77275-9

- Dodatkowe opóźnienie wynika z odwlekania leczenia przez pacjenta lub trudności w dostępie do testowania i leczenia.

Rozważanie 1: Czy Google Trends zwraca wystarczająco dużo danych liczbowych by podjąć monitoring?

Tak, wartości liczbowe są wysokie, obejmują wszystkie województwa i wykazują się dużą zmiennością dzienną.

Rozważanie 2: Czy wzory z badań w USA będą odpowiadały sytuacji polskiej?

Zapewne – ale matematycznie są zbyt wyrafinowane dla badającego.

Pomiędzy początkiem września 2020 a listopadem 2020 – analizowałem dane GT, przeliczając je z wartościami podawanymi przez Ministerstwo Zdrowia i liczbami dziennych testów – ujawniały się tam pewne korelacje, pozwalające „wyprzedzać” napływ chorych do szpitali – ale od 27 listopada zmieniła się zupełnie struktura danych podawanych przez MZ (zaprzestano publikacji danych z PSSE – Powiatowych Stacji Sanitarno-Epidemiologicznych – i scentralizowano podawanie danych w sposób sugerujący… hmmm… niewiarygodność.)

- Uznałem, że jedyne „lekarstwo” dla chcącego znać rzeczywisty stan epidemii w Polsce - to znalezienie takiej korelacji pomiędzy danymi GT (z punktem wyjścia z wartością wyliczoną poprzednimi wzorami dla 12 października) i dalszym liczeniem na bazie średniej dwutygodniowej (czyli z danymi z Trends pobieranymi od 29 września)
- Metodą prób i błędów przeanalizowałem kilkaset kombinacji wzorów, oddających relacje między wartościami z GT i liczbą raportowanych przez MZ zachorowań, niejako „kalibrując” metodę obliczeń.

- Po uzyskaniu takiej, która wykazywała należyte podobieństwo wartości raportowanych i obliczanych – używam jej do dziś.
- Różnica sum zachorowań wykazanych przez MZ i obliczanych w GT w tamtym czasie (w okresie październik listopad 2020) – wynosi jedynie **3,5%**
- Niestety – z chwilą zmiany przez MZ systemu raportowania – dane zaczęły różnić się dramatycznie.

- Ale – zaprowadziły mnie aż do 2021 roku, kiedy to:
- Pojawiły się badania PZH osób z przeciwciałami (czyli takich, które chorowały) –
- Przeprowadziłem analizę danych GT w relacji z nadzwyczaj wysoką śmiertelnością (dane GUS) w ostatnim kwartale 2020. Zarówno badania PZH (16,5% naprzeciw 16,61% wyliczanych dzięki GT) – jak i analizy relacji zachorowań wg GT i śmiertelności – sugerują mieszczenie się z wynikami GT w granicach wyników spodziewanych.

O metodzie

- Dobrałem dwa zespoły fraz – „wokół testów” i „wokół objawów” – moim zdaniem najbardziej adekwatne do problemu

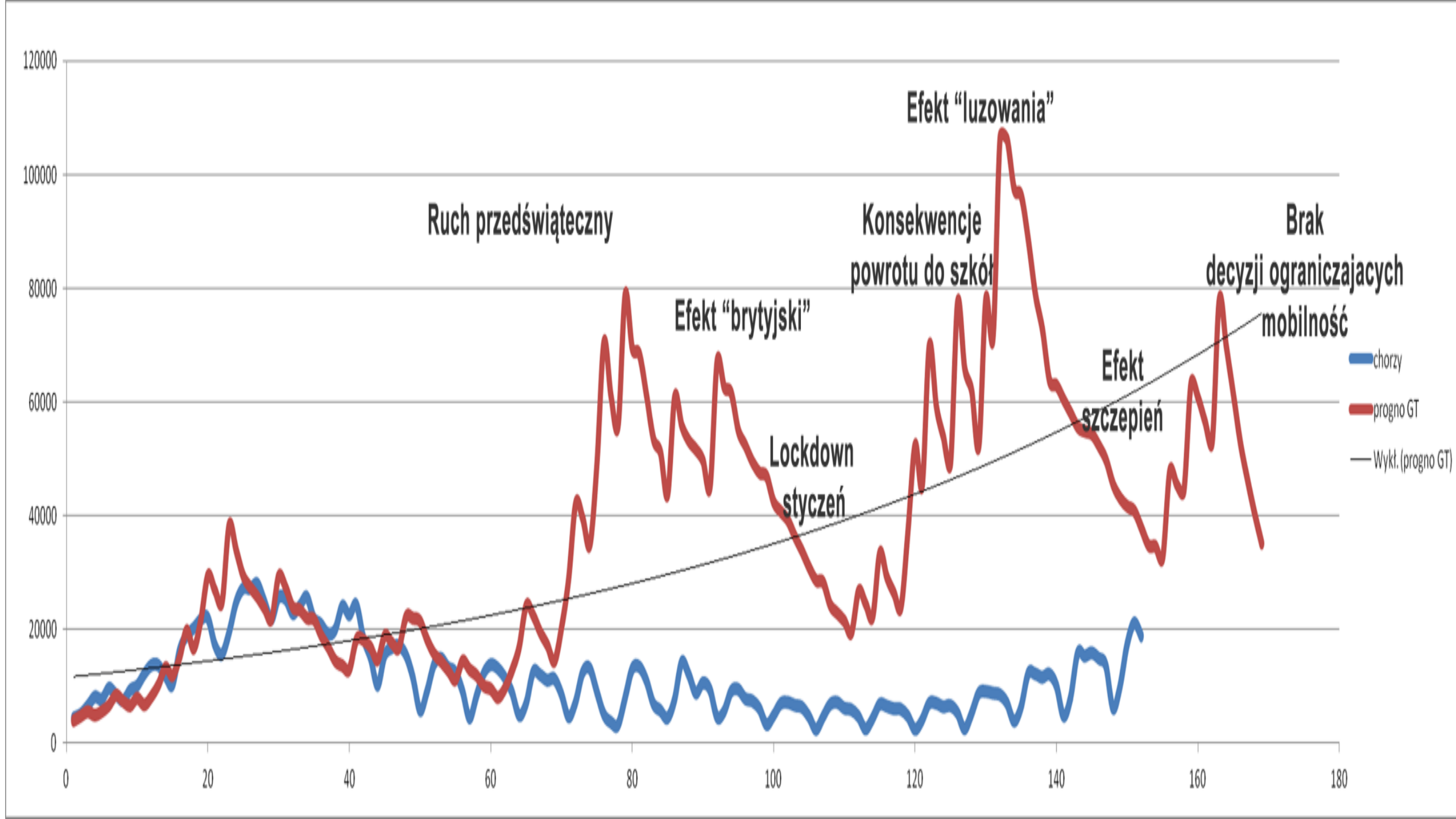
- Omówmy to na przykładzie problemu zepsutego samochodu:
- Frazy „testowe” – awaria samochodu, autoserwis, autonaprawa, naprawa samochodu, diagnostyka itp.
- Frazy „objawowe” – piszczenie w silniku, brak mocy, duże zużycie paliwa, warczący rozrusznik itp.
- Pierwotnie przypisywałem wyniki ręcznie – ale z uwagi na możliwość popełniania błędów i czasochłonność – zautomatyzowałem proces.

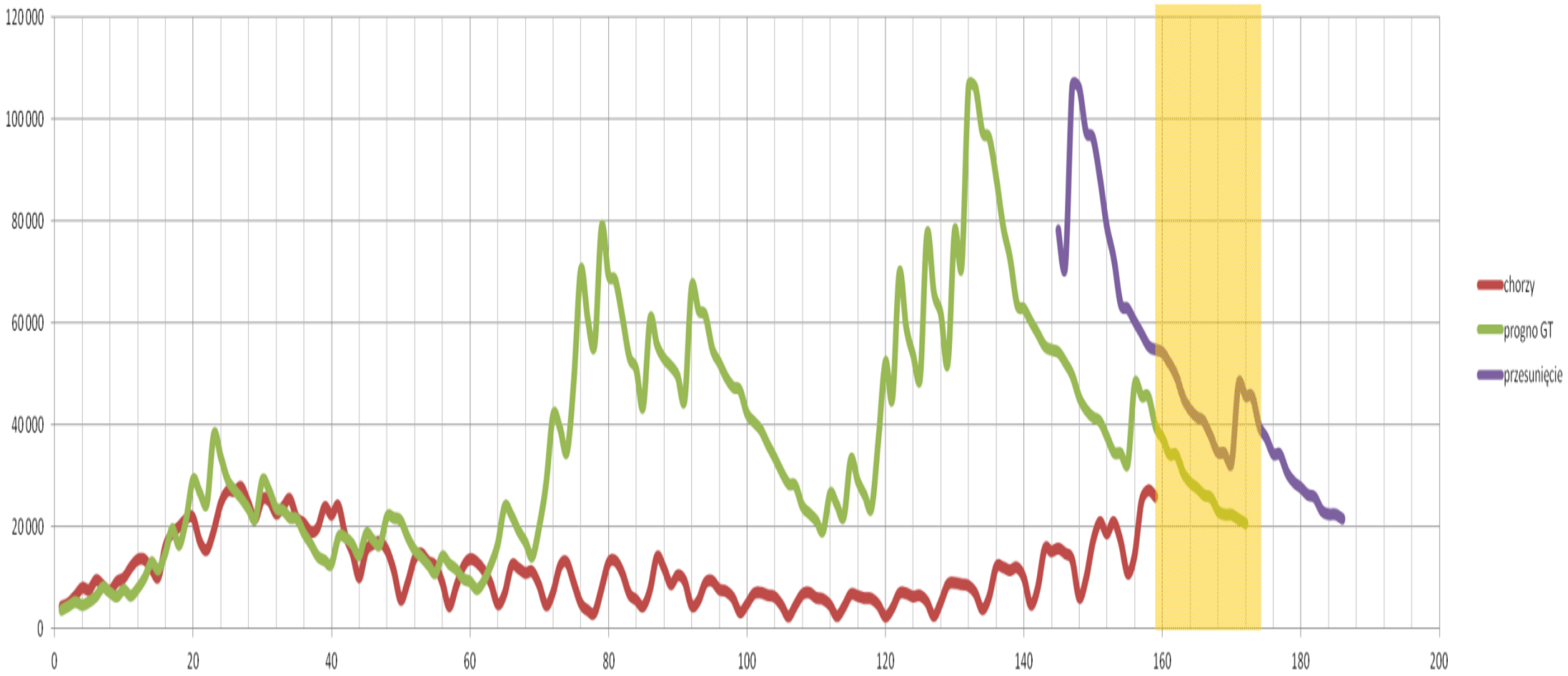
- (Tu uwaga: istnieje prawdopodobieństwo, że różne API Google mogą zbierać wyniki z różnymi wartościami – ja posługuję się oprogramowaniem Open Source o nazwie „R” – to zasadniczo program do modelowania statystycznego [R version 4.0.3 (2020-10-10)] – dlatego do ewentualnego odtwarzania badania warto posłużyć się tym samym oprogramowaniem i tymi samymi skryptami, które przedstawię za chwilę).

- Co otrzymujemy pobierając dane z GT? Wartości liczbowe, podzielone na województwa. Ponieważ nie było zainteresowanych moimi obliczeniami, dane pobieram z podziałem wojewódzkim – ale używam tylko średnich wartości dla grup danych. Wewnątrz grupy danych – używam zestawów fraz, większego dla części objawowej, żeby uniknąć przypadkowych sytuacji, kiedy jakiś objaw zacznie być wyszukiwany bez związku z covid (temu również służą frazy kontrolne do śledzenia anomalii – niestety nadzór nad tym prowadzę metodą oko – komputer ☺)
- Co dzieje się dalej?

Jak widać – na wykresie odzwierciedla się bardzo wyraźnie „tłumienie zakażeń” w pierwszych tygodniach 2021, mimo obecności zakażonych nowym wariantem brytyjski – który rozsiał się, moim zdaniem, głównie przez przedwczesny powrót do szkół.

Na osi X – suma dni badania; na osi Y – liczba zachorowań.


120000
100000
80000
60000
40000
20000
0
0
20
40
60
80
100
120
140
160
180
Ruch przedświąteczny
Efekt "brytyjski"
Lockdown
styczeń
Konsekwencje
powrotu do szkół
Efekt "luzowania"
Efekt
szczepień
Brak
decyzji ograniczajacych
mobilność
chorzy
progno GT
Wykł. (progno GT)



Ten wykres powstał na danych do 19 marca 2021 (z prognozą do 1 kwietnia 2021)
Fioletowa linia to 16-dniowe przesunięcie napływu chorych, dla szpitali

Prognozowanie

- Publikuję w arkuszu dostępnym w sieci (via Google Drive) prognozę rzędu + 10 dni – licząc średnie trzydniowe i porównawczo dwutygodniowe z liczby obliczanych zachorowań, aktualizując codziennie.
- Łącznie daje to perspektywę rzędu miesiąca (bo, jak pamiętamy, obliczone zachorowania „wędrują” do leczenia z opóźnieniem rzędu 14 dni)

Najważniejsze liczby

- Zachorowania w okresie 12.10.2020 – 31.03.2021 = **6 135 953 osoby**
- Szacowana śmiertelność wśród zakażonych (łącznie z danymi MZ, które są niestety tylko cząstkowe) – 1,5% ogółu zakażonych

  = **92.039 osób**
- To oszacowanie wyjaśnię dalej

O wątpliwościach

- Wartości liczbowe, dotyczące zachorowań wydawały mi się tak radykalnie wyższe od zestawu danych serwowanych przez Ministerstwo Zdrowia – że w zasadzie do końca 2020 unikałem publikowania tych wartości.
- Niekiedy – pokazywałem na TT wartości wykresowe, najczęściej przy dużych wzrostach, niewspółgrających z komunikatami oficjalnymi.
- Wydawało mi się absurdalnie niemożliwe, by pewne decyzje, rozluźniające rygory – zapadały przy prawdopodobieństwie tak dużej realnej liczby zachorowań.
- Jednak przypominałem sobie wtedy – jak bardzo zlekceważono doświadczenie izraelskie z maja 2020 – kiedy to powrót dzieci do szkół – zakończył się lawinowym wzrostem zachorowań.
- Analogiczną sytuację mieliśmy w Polsce (tzw. druga fala) – i ponownie, w 2021 roku, po powrocie dzieci do szkół, nawet w wariancie niepełnym (bo nakładało się to na wpływ dużo bardziej zaraźliwego „wariantu brytyjskiego” SARS-Cov-2)

- Nie wiem do dziś – jak moje wzory „zachowywały by” się w wypadku bardzo małej liczby zachorowań – takiej rzędu 300-500 w skali kraju. Da się to oczywiście zbadać, pobierając dane historyczne z GT i używając do przeliczeń moich wzorów. Nie miałem po prostu czasu na taki eksperyment. (korekta poniżej ☺)
- Przygotowując się do tego spotkania pobrałem wyrywkowe dane z kilku lipcowych dni 2020 (kiedy zachorowania były na poziomie kilkuset dziennie – i dane istnieją w wartościach wystarczających do dalszych przeliczeń – nie wprowadzałem ich jednak do „maszyny obliczeniowej)

Inne parametry szczególne

- Czułość – wartości sum uzyskiwanych z Google Trends maleją z upływem czasu (zmniejsza się ogólne zainteresowanie) – ale dane uzyskiwane po przeliczeniu nadal charakteryzują się dużą zmiennością, i podobnymi tempami w zmianach trendów - co sugeruje, moim zdaniem, właściwy dobór relacji między danymi w obliczeniach.
- Dlaczego nie chcę mówić o konkretnych zestawach fraz?
- Bo wyniki badania można zniekształcić, bombardując wyszukiwarkę zestawem tych fraz – nie zmniejszy to wyliczanych zachorowań – ale zniekształci i „zdemoluje” wynik finalny badania.

Poszukiwania potwierdzeń w innych danych i parametrach epidemicznych:

- Trzy główne „idee” kontrolne:
- Zestawienie z badaniami przeciwciał PZH
- Namysł nad relacjami ze śmiertelnością
- Korelacja widocznego obecnie spadku w krzywej zachorowań z najprawdopodobniejszym „nasycaniem” populacji osobami z przeciwciałami

- PZH:
- W artykule https://oko.press/ile-osob-zmarlo-na-covid-ile-sie-zakazilo/ pojawia się info PZH dotyczące przeciwciał u **16,5%** badanych (czyli taki procent populacji chorował)
- Wg danych z GT – otrzymujemy wartość **16,6%** - moim zdaniem to korelacja potwierdzająca prawidłowość wyliczeń.
- Zachorowania MZ do 11.10.2020 = 125.816
- Szacuję, że realnie było ich x 5 = ca. 630.000
- Zachorowanie wg Google Trends 12.10 – 18.03.2021 = 5.677.081
- Razem = 6.607.081
- Populacja Polski 37,97 mln
- Chorych = 16,61%

Obecny spadek krzywej zachorowań:

- Chorych dotąd: Razem = **6.607.081**
- Plus ca 20% bezobjawowych z przeciwciałami = **7.928.000**
- Plus zaszczepieni co najmniej jedną dawką (stan na 17.03; TT @gfkot) = **2.984.642**
- Razem z przeciwciałami = **10.912.642**
- W populacji – **28,74%**

Relacje: zachorowania wg GT a śmiertelność - 1

- Zgony COVID w raportach MZ w tygodniach T42 do T53 (12 października 2020 do 3 stycznia 2021) = 26.114
- Zgony ogółem w tygodniach T42 do T53 (12 października 2020 do 3 stycznia 2021) (GUS) =156.795
- https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/zgony-wedlug-tygodni,39,2.html
- Zgony w okresie jw. za https://mpidr.shinyapps.io/stmortality/ = 158.128
- Zgony znacznie ponad średnią 10-letnią – stanowią o liczbie 66.670
- Zgony „nadspodziewane” minus zgony COVID (MZ) = 66.670 – 26.114 = 40.556
-
- Zachorowania COVID w tygodniach T42 do T53 według MZ = 1.170.999
- Przy współczynniku śmiertelności 3% = zgony spodziewane 35.129
- Zgłoszone zgony COVID (26.114) to 2,23% z zachorowań (1.170.000)
- Zachorowania w T42-T53 według Google Trends – 1.930.387
- Zgony wg współczynnika 2,23% = 43047,63 – zgony (MZ) 26.114 = 16.933,63
- Obliczyliśmy wcześniej różnicę „nadspodziewanych” i „covidowych” zgonów na 40.556
- Uważam, że można uznać, że 16.994 (czyli 42% z „nadspodziewanych”) to zgony powiązane również z COVID (nierozpoznane zachorowania z komplikacjami w chorobach współistniejących, nie zgłoszone w porę opiece medycznej. Zatem nie stanowi to o sprzeczności liczby zgonów wg GUS z wyliczeniem zachorowań wg Google Trends na poziomie 164,85% zachorowań raportowanych przez MZ.
- Zgony w T42-T53 2020 roku to 172,9% ponad średnią 10-letnią.

Relacje: zachorowania wg GT a śmiertelność - 2

- Zatem pewnym istotnym weryfikatorem będzie poziom śmiertelności w Polsce w okresie styczeń-luty oraz statystyki kolejnych miesięcy.
- W pierwszych 6 tygodniach 2021 GUS raportuje łącznie 61.651 zgonów (liczba może być jeszcze w GUS doprecyzowana, to wstępne raporty).
- Zgony COVID (MZ) w tym czasie to 11.674.
- Zachorowania w tym okresie wg Google Trends – to 1.800.114
- Zgony wg współczynnika 2,23% = 40.142
- Zgony (GUS) – zgony (MZ) = 61.651 – 11.674 = 49.977
- Zgony „prawdopodobne” wg GT – zgony (MZ) = 40.142 – 11.674 = 28.468
- Mogłoby to oznaczać, że COVID odpowiada za 56% z 49.977 zgonów niewliczonych do statystyk MZ.
- Mogłoby też oczywiście oznaczać, że statystyka zachorowań wg GT jest w tym okresie nieco zawyżona lub że jeszcze nie odzwierciedliła się we współczynniku śmiertelności.
- Zgony w T1-T6 2021 roku to 124% ponad średnią 10-letnią.

Relacje: zachorowania wg GT a śmiertelność - 3

- Przeanalizujmy 3 okresy co do relacji zachorowań i zgonów
- Średnia dzienna z ostatnich 3 tygodni (dokładnie: 23 dni) dla zachorowań raportowanych przez MZ = 7230
- Średnia dzienna z ostatnich 3 tygodni dla zgonów raportowanych przez MZ = 208
- Zgony do liczby zachorowań = 2,87%
-
- W okresie 28.10 – 20.11
- Średnia dzienna dla zachorowań raportowanych przez MZ = 22.618
- Średnia dzienna dla zgonów raportowanych przez MZ = 342
- Zgony do liczby zachorowań = 1,51%
-
- W okresie 12.12 – 03.01
- Średnia dzienna dla zachorowań raportowanych przez MZ = 8.862
- Średnia dzienna dla zgonów raportowanych przez MZ = 301
- Zgony do liczby zachorowań = 3,4%

Relacje: zachorowania wg GT a śmiertelność - 4

- Czyli: nie budzi zdziwienia relacja zachorowania/zgony na poziomie 1,5%.
- W sytuacji, gdyby przyjąć podobną relację procentową między zachorowaniami liczonymi wg Google Trends – a śmiertelnością w T1-T6 w roku 2021:
- Zachorowania w tym okresie wg Google Trends – to 1.800.114
- Zgony wg współczynnika 1,5% = 27.002
- Zgony (GUS) – zgony (MZ) = 61.651 – 11.674 = 49.977
- Zgony „prawdopodobne” wg GT – zgony (MZ) = 27.002 – 11.674 = 15.328
- Wtedy powiązanie COVID ze zgonami spoza statystyk MZ to 30%.

- Należy zauważyć, że przyjęto tu (dla danych GT i wyliczenia śmiertelności) pewne uproszczenie, nie stosując przesunięcia w czasie – dla precyzyjniejszego ustalenia wartości należy dla wskaźnika śmiertelności w danym dniu – pobierać liczbę zachorowań minus pięć tygodni (trzy tygodnie między zachorowaniem a dotarciem do Opieki Zdrowotnej i kolejne dwa tygodnie intensywnej terapii pod respiratorem

Relacje: zachorowania wg GT a intensywna terapia - 5

- Jeżeli – najprościej w danych Michała Rogalskiego – przyjrzeć się danym o chorych pod respiratorami i danym o zajętych łóżkach szpitalnych, można odnieść istotne wrażenie, że te dane „nie spinają się” z raportowanymi danymi o zakażeniach.
- Można – (na podstawie wiarygodnych danych medycznych w populacji światowej) z pewnym uproszczeniem przyjąć, że:
- 80% chorych na COVID choruje poza szpitalami, nie wymagając leczenia o większej intensywności
- Z chorych trafiających do szpitali 10% wymaga intensywnej terapii pod respiratorem

Relacje: zachorowania wg GT a intensywna terapia - 6

- liczbę chorych pod respiratorami - pomiędzy grudniem a dziś - najniższa liczba to 1267
- czyli - tak uśredniając - mielibyśmy w tym czasie po 13.000 chorych/ dzień

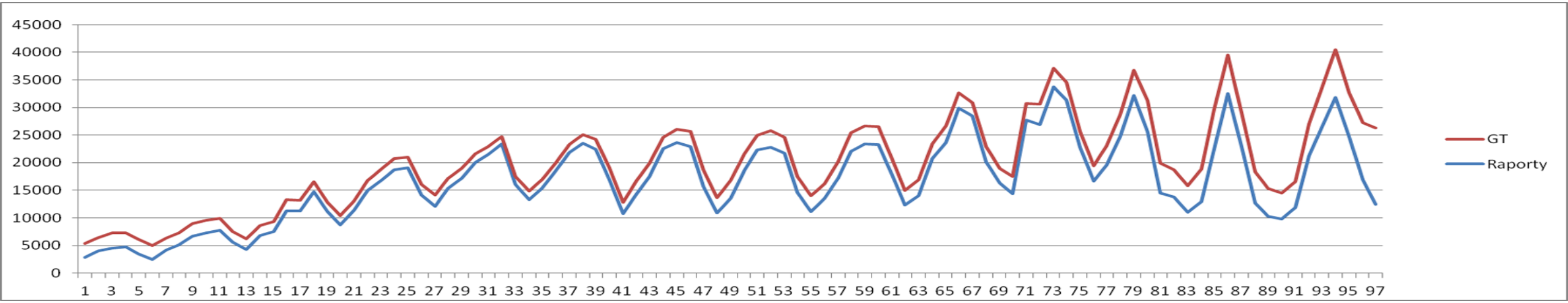
- Tymczasem w danych MZ o zachorowaniach w tym czasie "spadaliśmy" do 2600 przypadków dziennie (25 stycznia, 1 lutego itp); średnio - jest 8400 w zachorowaniach;
- do szpitalnego leczenia trafia ca.10-20 % ogółu chorych;
- 1267 respiratorów x 10 = 12.670 chorych szpitalnych x 5 = 63.500;
- a przy 15% z ogółu w szpitalu - to liczba chorych rzędu 84.000;
- w tym okresie max. liczba chorych z Trends to ok. 78.000;
- Czyli kolejny element analizy porównawczej wskazuje na "mieszczenie" się w wartościach statystycznych.

- Precyzyjniejsze dane medyczne można odnaleźć np. w:
- https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(20)30161-2/fulltext

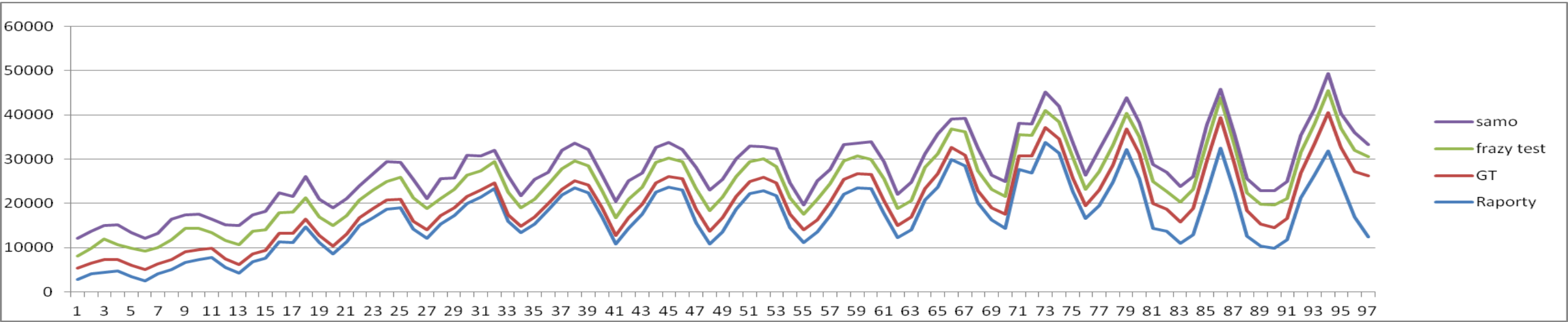
Dodatek – badanie danych z Niemiec

- Z ciekawości – zbadałem dane Google Trends z Niemiec za okres 23.09.2020 – 11.01.2021
- Okazało się, że samo przeniesienie wzoru obliczeniowego nie powoduje uzyskania wartości obliczenia adekwatnych do raportów o zachorowaniach, zgłaszanych przez Niemcy do WHO
- Niemniej jednak – w pełni pokrywają się amplitudy, widoczne w zmianach obydwu grup fraz, wynikach obliczenia i raportach oficjalnych

Zestawienie raportów niemieckich o zachorowaniach dla WHO i przeliczeń zestawów fraz GT



Zestawienie – raporty DE o zachorowaniach, grupy fraz (dane pobrane z Google Trends, obliczenia GT; (**samo** – to opis grupy fraz objawowych)



- Podstawowy wniosek – pełna zgodność amplitud wynika ze stosowania w Niemczech w pełni adekwatnego do rozwoju epidemii poziomu testowania.
- Częstotliwość zmian w grupach fraz i pozostałych wskazuje na nadążanie testowania za ogniskami zakażeń i prowadzenie adekwatnych śledztw epidemiologicznych

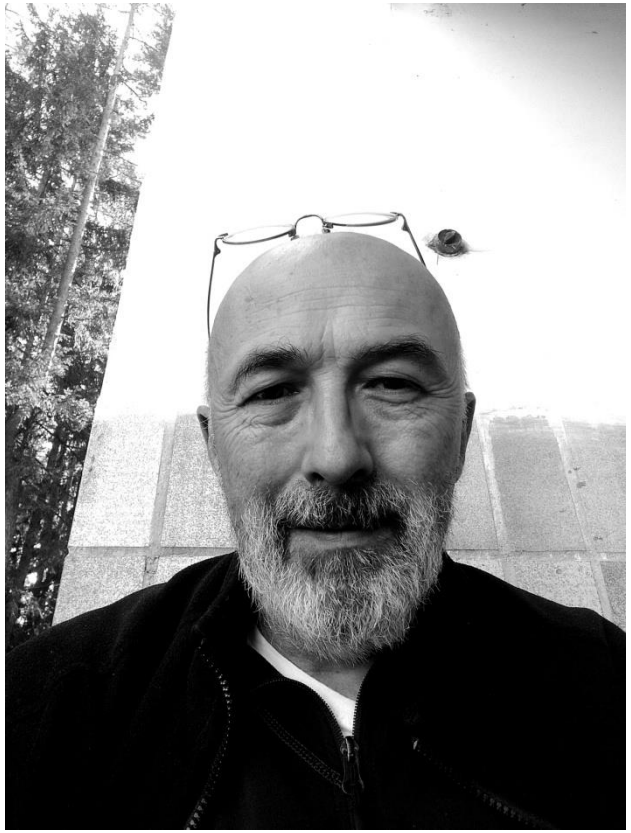
- Ponieważ forma wykresu, uzyskiwanego z obliczenia, jest analogiczna do tych, jakie widać w raportach i frazach – uzyskanie właściwych wyników – to tylko kwestia „kalibracji”
- Jednak z racji wysokiego podobieństwa wykresów dla fraz i raportów – można uznać, że dla Niemiec niepotrzebne jest stosowanie skomplikowanych wzorów – a wystarczy obliczenie stałych proporcji
- Nie rozwijałem analizy „kalibracji” – bowiem najistotniejszą dla mnie informacją było potwierdzenie związku: frazy – obliczenia - raporty

- Mówiąc zupełnie z grubsza – starając się dostosować metodę obliczeniową do warunków niemieckich, wnioskuję, że obliczenie, skonstruowane dla warunków polskich – zawiera rodzaj „wzmacniaczy czułości” – które dla warunków innych obszarów geograficznych trzeba parametryzować
- Zależy to od wolumenu zapytań dla Google z poszczególnych obszarów geograficznych – dla Niemiec jest on wysoki, dla Polski dużo niższy, dlatego bezpośrednie przeniesienie wzorów daje wyniki „nadczułe”, powyżej stanów rzeczywistych

Wartość badania

- Można by tu mówić bardzo wiele – uważam że najistotniejszym czynnikiem jest ten, który pokazuje, że dzięki analizie Google Trends i łączeniu jej z innymi metodami modelowania matematycznego - da się w miarę precyzyjnie i z relatywnie poważnym wyprzedzeniem nadzorować zarządzanie kryzysem – a także, przy przyjmowaniu odpowiednio dociekliwych metod poszukiwawczych – monitorować ewentualne nadciąganie innych zjawisk pandemicznych.
- **Szczególnie istotna „na teraz" może być analogiczna metoda wyszukiwania objawów wskazujących na wzrost grupy chorych z powikłaniami po COVID, zanim zgłoszą się do systemu opieki zdrowotnej**
- Połączenie danych GT z danymi telemetrycznymi operatorów komórkowych (niedostępnymi dla mnie) – w moim przekonaniu pozwala na trzy-czterotygodniowe prognozowanie z precyzją rzędu 5%

- Z pewnym przymrużeniem oka, pełen pokory, przywołam tu cytat z szalonego ale pięknego filmu „Piąty element"

„Chroń życie"

- Arkusz z danymi i wykresami:
- https://docs.google.com/spreadsheets/d/1TfTuLVZNphfEbFgSvAtC_zprdE6ifsr9YTJNjDPTrRY/edit?usp=sharing
- Mariusz Kwiatkowski
- TT - @altymis
- Mail – altymis@gmail.com