

Zanim dopadnie nas grypa

Piotr KRZYŻANOWSKI*

* Uniwersytet Warszawski, Wydział
Matematyki, Informatyki i Mechaniki

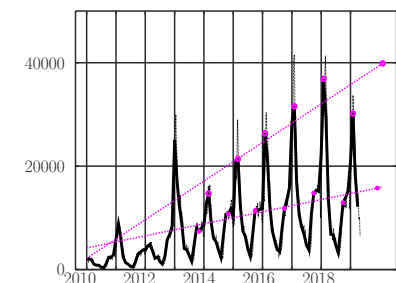
Czy lubicie długoterminowe prognozy pogody? Odbija się w nich głęboko zakorzeniona ludzka wiara, że odległą przyszłość można przewidzieć – na przekór *efektowi motyla*. No cóż, sam muszę przyznać, że cieszę się, gdy grudniowe prognozy przewidują piękną, słoneczną i mroźną aurę na zimowe ferie; ale jeśli zapowiadają ponury mokry standard, wtedy ratuję się myślą, że to przecież tylko prognoza...

Oczywiście można przewidywać nie tylko pogodę: na przykład zimą byłoby całkiem ciekawie przepowiedzieć rozwój corocznej epidemii grypy. I właśnie teraz, mimo braku wiedzy medycznej, uzbrojony jedynie w serię liczb, prosty model i kilka linii kodu (wybiorę moje ulubione środowisko obliczeniowe Octave, www.octave.org) spróbuję przewidzieć, jak będzie przebiegać w tym sezonie.

Szczerze mówiąc, zdziwiłbym się, gdyby przyszłość posłuchała się moich prognoz: użyte przeze mnie narzędzia są raczej mało wyrafinowane. Ale... gdyby przewidywania się sprawdziły... czyż nie byłoby zabawnie?

Skorzystamy z publicznie dostępnych danych NIZP-PZH na temat raportowanej przez lekarzy w całej Polsce liczby zachorowań na grypę (jak je zdobyć samemu, pisaliśmy w Δ_{19}^{10}). Zaczniemy od naszkicowania wykresu zachorowań w ostatnich latach, zobacz rysunek obok. (Dane zbierane są cztery razy w miesiącu).

Wykres jest mało regularny, niemniej możemy zauważyć, że w pierwszym kwartale każdego roku ma silny pik, a w sierpniu – wyraźne minimum. Tabela obok pokazuje daty, gdy zostało osiągnięte ekstremum, a dane do niej można wygenerować kodem:



Średnia miesięczna liczba zachorowań na dzień (uśrednienie trochę wygładza wykres). Górna kolorowa prosta została dopasowana do pików z lat 2014–2018 metodą najmniejszych kwadratów. Analogicznie dopasowano dolną do chwil listopadowej stagnacji

minimum	maksimum
08.08.2012	23.01.2013
08.08.2013	08.03.2014
08.08.2014	01.03.2015
08.08.2015	16.02.2016
08.07.2016	23.01.2017
08.08.2017	08.02.2018
08.08.2018	08.02.2019

Daty rocznych ekstremów raportowanej zachorowalności na grypę (podana jest data początku pomiaru).

Dla prostoty w powyższej tabelce porównywaliśmy liczby z kolejnych raportów, które obejmują okresy *nierównej* długości. W rzeczywistości powinniśmy działać na średniej dziennej zachorowalności – i tak właśnie robimy w późniejszych obliczeniach

Wyjątkowa była ubiegła zima, kiedy grypa przebiegała nad wyraz łagodnie.

```

1 pkg load signal;
2 [~, loc] = findpeaks(infected, 'MinPeakDistance', 24);
3 maxinfected = infected(loc); maxdays = days(loc);
4 disp( datestr(maxdays) );

```

W programie przyjęliśmy, że w zmiennej *infected* znajdują się liczby zachorowań, które wcześniej ściągnęliśmy z NIZP-PZH, a w *days* – daty odpowiadających pomiarów. Funkcja *findpeaks* z pakietu *signal* znajduje położenia *loc* lokalnych maksimum danych zawartych w wektorze *infected*. Dodatkowo wymagamy, by odległość między maksimum wynosiła co najmniej 24 pomiary, czyli 6 miesięcy (jasne, są przecież odległe z grubsza o rok). W trzeciej linii wydobywamy z danych tylko te odpowiadające znalezionym maksimum, które w kolejnej linii wypisujemy na ekranie.

Minimum zawsze przypada na drugi tydzień sierpnia. Maksimum nie zachowuje się aż tak regularnie, ale jak możemy obliczyć, średnia odległość między maksimum na przestrzeni ostatnich 6 lat wynosi około 368 dni – czyli z grubsza też jeden rok (co w końcu nie powinno nas aż tak dziwić...).

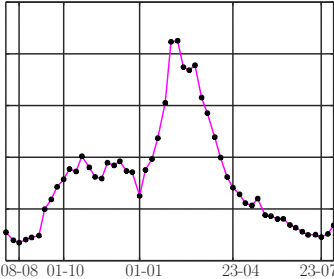
Rzuca się w oczy, że od kilku lat piki zgłaszanej liczby zachorowań systematycznie rosną (niemal liniowo). Nie będąc lekarzem, ale za to trochę wyobrażając sobie, jak działa *sprawozdawczość*, przypuszczałbym, że za tym wzrostem nie stoi realny wzrost liczby wszystkich zachorowań na grypę w Polsce – ale że zapewne jest on efektem coraz lepiej działającego systemu zbierania danych (coraz więcej lekarzy odsyła stosowne formularze?). Gdyby więc liniowy trend miał utrzymać się jeszcze w tym sezonie, łatwo będzie przewidzieć orientacyjną szczytową liczbę zgłoszeń:

```

1 P = polyfit(maxdays, maxinfected, 1);
2 disp( polyval(P, maxdays(end)+365) );

```

Całą robotę – dopasowanie wykresu prostej do maksimum – wykonuje pierwsza linia powyższego kodu: funkcja **polyfit** wyznacza *metodą najmniejszych kwadratów* wielomian P stopnia pierwszego, najmniej oddalony od punktów pomiarowych (x_k, y_k) , gdzie $x_k = \text{maxdays}(k)$, a $y_k = \text{maxinfected}(k)$. W drugiej linii obliczamy wartość tego wielomianu w dniu następnego maksimum zachorowań (przyjeliśmy, że wypadnie dokładnie za rok od ostatniego).



Typowy przebieg sezonu grypowego (tu dla sezonu 2015/2016)

Naturalnie w dłuższej perspektywie liniowy wzrost nie może się utrzymać, co zrozumiemy, przedłużając kolorową prostą *w lewo*: „przewiduje” ona, że już w 2008 roku lekarze powinni byli zgłosić... ujemną liczbę zachorowań.

Modelowanie w wybranym sezonie

Powyższe miało charakter czysto rozrywkowy, więc czas na coś poważniejszego. Prawie każdy sezon chorobowy ma podobny przebieg (rysunek obok):

- dosyć szybki wzrost zachorowań od połowy sierpnia,
- z początkiem października następuje stabilizacja liczby zachorowań, a nawet – w pierwszym tygodniu stycznia – jej chwilowy spadek,
- od początku stycznia – kilkutygodniowa faza szybkiego wzrostu liczby zachorowań,
- osiągnięcie szczytowej liczby zachorowań (najczęściej w lutym),
- szybkie zmniejszanie się liczby zachorowań gdzieś tak do końca kwietnia,
- powolny spadek do kolejnego minimum.

Czy więc dałoby się – dysponując *bieżącymi* danymi ze strony internetowej NIZP-PZH – przewidzieć rozwój tegorocznej grypy? Do tego wykorzystamy następujący, bardzo prosty, model zachorowań. Niech S_t oznacza liczbę osób podatnych na wirusa (ale na razie zdrowych) w dniu t , z kolei I_t – liczbę osób zarażonych (i przy tym zarażających), a R_t – liczbę osób, które już nie mogą być zarażone ani zarażać, bo np. wyzdrowiały i nabrały odporności. W modelu przyjmujemy następującą zależność S, I, R w dniu następnym od tychże w dniu obecnym:

$$\begin{aligned} S_{t+1} &= S_t - b \cdot I_t \cdot S_t, \\ I_{t+1} &= I_t + b \cdot I_t \cdot S_t - a \cdot I_t, \\ R_{t+1} &= R_t + a \cdot I_t. \end{aligned}$$

Gdybyśmy znali współczynniki modelu: a, b , to moglibyśmy go użyć do wyznaczenia przyszłych wartości S, I, R tak, jak w poniższej funkcji:

```
function [S,I,R] = KMCK(a,b,S1,I1,R1,n)
% wyznacza S = [S1, S2, ..., Sn], I = [I1, I2, ..., In], R = [R1, R2, ..., Rn].
S(1) = S1; I(1) = I1; R(1) = R1;
for t = 1:n-1
    S(t+1) = S(t) - b*I(t)*S(t);
    I(t+1) = I(t) + b*I(t)*S(t) - a*I(t);
    R(t+1) = R(t) + a*I(t);
end
end
```

Opisany model to tzw. model Kermacka–McKendricka, o którym pisaliśmy w Δ_{16}^4 – u nas w wersji z czasem dyskretnym.

Między innymi wiadomo, że istotnym czynnikiem wpływającym na skuteczność rozprzestrzeniania się grypy jest pogoda – co powoduje, że *w rzeczywistości* wartość współczynnika b zmienia się w czasie, podczas gdy my radośnie przyjmujemy, że jest stała.

To bardzo zgrubny model, ale mamy nadzieję, że okaże się wystarczający, jeśli ograniczymy zakres jego stosowalności. Nie możemy liczyć na to, że uda się nam w modelu dopuszczającym tylko jeden pik zachorowań odtworzyć wieloletni przebieg choroby. Ale wymodelowanie okresu największej zachorowalności w sezonie, powiedzmy: od początku stycznia do końca kwietnia, wydaje się celem jak najbardziej realnym.

Ponieważ $1/a$ odpowiada średniej długości okresu, gdy zainfekowany zaraża, to możemy z dobrym skutkiem przyjąć, że $1/a = 2,5$ dnia (potem bierzemy zwolnienie i kładziemy się do łóżka), więc $a = 0,4$ (na dzień).

Nie znając dobrej wartości współczynnika b , dopasujemy ją tak, by wartości I wyznaczone przez model jak najlepiej przybliżyły te, które już znamy z raportów

Pomijamy istotny szczegół, że raporty NIZP-PZH nie podają liczby wszystkich zachorowań, co więcej – podawane przypadki nie zawsze muszą być prawdziwą gripą. Taki urok medycyny.

NIZP-PZH. W tym celu definiujemy funkcję

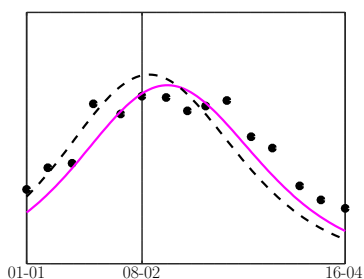
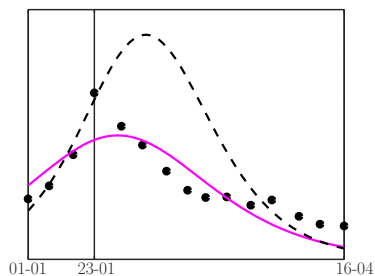
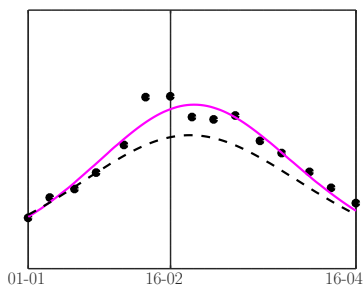
$$b \mapsto F(b) := [I_1(b), I_2(b), \dots, I_n(b)],$$

która dla zadanego b wyznacza (za pomocą powyższego modelu) wartości I_t w punktach pomiarów w zadanym przez nas okresie. Następnie, korzystając z funkcji `leasqr` z pakietu `optim`, znajdziemy wartość b , która zminimalizuje sumę kwadratów odchyłeń wartości przewidywanych od wartości zmierzonych:

$$(I_1(b) - \text{infected}(1))^2 + \dots + (I_n(b) - \text{infected}(n))^2 \rightarrow \min!$$

Listing 1. (Octave)

```
1 [infected, days] = przytnijdane(infected, days, '01-01-2016', '23-04-2016');
2 pkg load optim;
3 [~, b] = leasqr(days, infected, 0.4, F);
```



Wymodelowany przebieg grypy dla sezonów 14/15, 15/16 i 16/17. Pionowa kreska wskazuje realne maksimum zachorowań. Linia kolorem: wynik działania modelu dla b wyznaczonego na podstawie wszystkich danych. Czarna, przerywana linia: dla b wyznaczonego na podstawie tylko pierwszych czterech pomiarów

Najpierw ograniczamy zbiór danych `infected`, `days` do interesującego nas okresu (implementację funkcji `przytnijdane` można sobie z grubsza wyobrazić), a po załadowaniu pakietu `optim` funkcja `leasqr` wyznacza b minimalizujące sumę kwadratów odchyłeń (czwarty argument, F , jest potrzebny, bo ta funkcja wyznacza wartości $I_1(b), I_2(b), \dots$ potrzebne do obliczenia odchyłeń) – omówienie jej implementacji, wykorzystującej KMCK, pominiemy.

Jak widać na wykresach obok, dla tak wyznaczonego b dostaliśmy kolorową krzywą, całkiem dobrze wpasowującą się w punkty pomiarowe (sezon 2016/2017 jest – jak widzieliśmy wcześniej – pod wieloma względami wyjątkowy i model nie przewiduje prawidłowo wielkości piku). Okazuje się, że dla większości sezonów dostajemy podobną wartość optymalnego b . Aby więc przewidywać przyszłe epidemie, możemy po prostu skorzystać ze „średniej” wartości $b \approx 0,434$, co zamyka sprawę (po prostu startujemy symulację z wartości znanych w konkretnym dniu). Z drugiej strony wybór uśrednionego b powoduje, że nie możemy reagować na ewentualne *atypowości* bieżącego sezonu (np. niesprzyjającą aurę lub szczególnie zjadliwy szczep wirusa, co w każdym przypadku powoduje zmianę współczynnika b). Wówczas lepiej dopasować nowe b do bieżących danych. Przykładowy wynik dopasowania b na podstawie jedynie czterech pomiarów ze stycznia (linia przerywana na wykresach obok) pokazuje, że możemy się wtedy sporo pomylić; jednak im więcej pomiarów wykorzystamy, tym lepsze będą przewidywania.

A zatem... zaglądając do mojej szklanej/krzemowej kuli...

P R Z E P O W I A D A M, Ż E:

1. Następny sezon grypy... już się zaczął: w połowie sierpnia!
2. Od października do końca grudnia raportowana liczba zachorowań nie przekroczy... 140 tysięcy w tygodniu! *[Odczytany liniowy trend z pierwszego wykresu]*.
3. Szczyt zachorowań przypadnie... w drugim tygodniu lutego! *[Po prostu zakładam, że stanie się to za rok od ostatniego]*. Będzie wtedy zgłaszanych aż 400 tysięcy zachorowań tygodniowo! *[Z liniowego trendu na pierwszym wykresie; jednak nie wierzę w to zbyt i mam nadzieję, że naprawdę będzie ich znacznie mniej... Chociaż, z drugiej strony, grypa jest chorobą zdradziecką, więc po poprzednim roku ciszy teraz może zaatakować ze zdwojoną siłą]*.
4. Przebieg zachorowalności w I kwartale 2020 roku... będę na żywo monitorować i prognozować na swojej stronie internetowej (www.mimuw.edu.pl/~przykry) już od teraz, co tydzień uwzględniając nowe dane! *[Będę dopasowywać na bieżąco parametry modelu, minimalizując błąd średniokwadratowy, podobnie jak opisano powyżej]*.

No dobra, czas kończyć pisanie. Już w trakcie poczułem się – jakoś tak – niewyraźnie: boli mnie głowa, drapie w gardle i... tak jakby... łamie w kościach? Czyżby –

Aaa... apsik!...