

Statystyczne modele pandemii i metody predykcyjne w prognozowaniu rozprzestrzeniania się wirusa SARS-CoV-2 w różnych krajach z uwzględnianiem szczepień

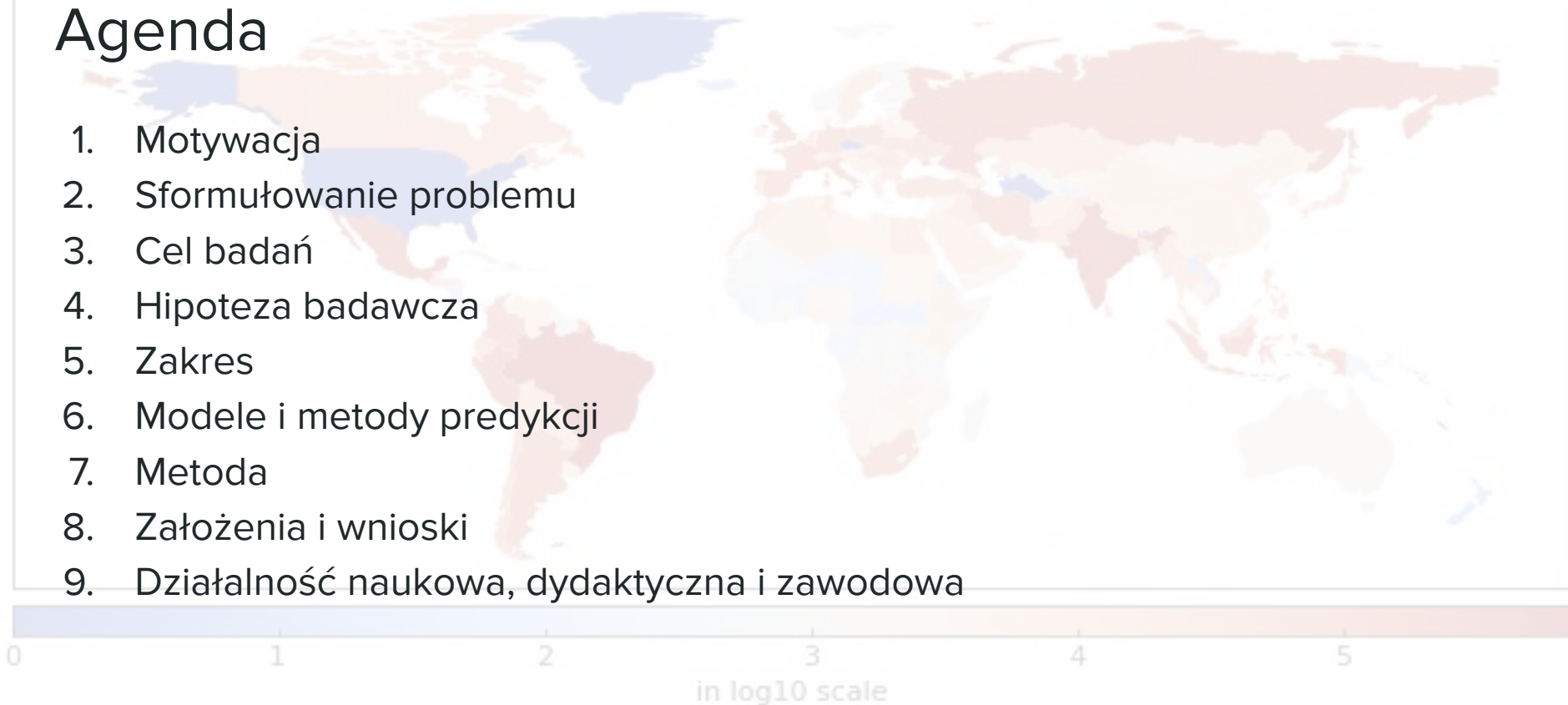
Autor: mgr inż. Łukasz Kupracz

Koszalin 2021

Opiekun naukowy: prof. dr hab. inż. Antoni Wiliński

Agenda

1. Motywacja
2. Sformułowanie problemu
3. Cel badań
4. Hipoteza badawcza
5. Zakres
6. Modele i metody predykcji
7. Metoda
8. Założenia i wnioski
9. Działalność naukowa, dydaktyczna i zawodowa



Motywacja

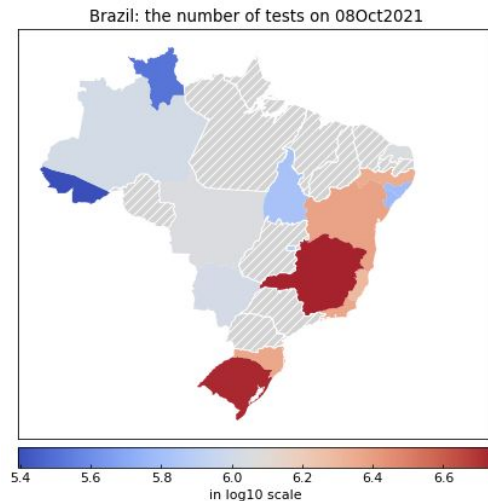
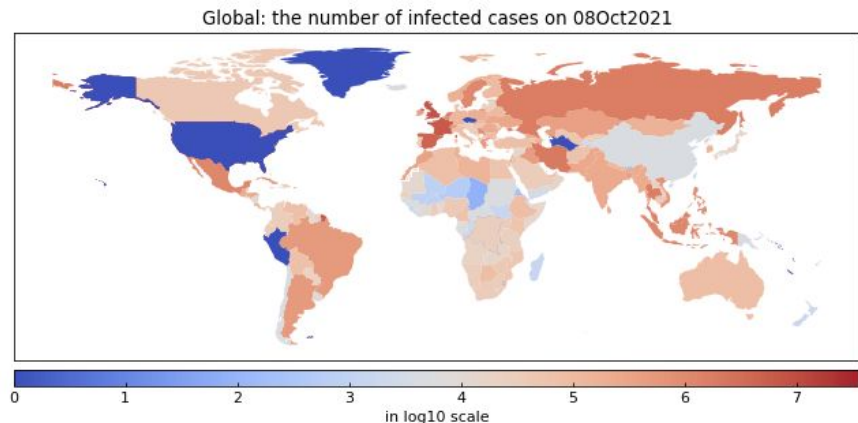
- COVID-19 największym kryzysem globalnym od czasów II wojny światowej.
- Przewyższa śmiertelnością pandemie **globalne** (hiszpanka, HIV) i **lokalne** SARS, MERS i Ebola).
- Rozwój pandemii w krajach jest dotychczas **nieprzewidywalny**.
- Dotychczasowe modele przebiegu choroby i szacunki rozprzestrzeniania się wirusa **nie są spójne**.

Sformułowanie problemu

- **Model SIR** i pochodne: sprawdzają się dla krótkotrwałych interwałów, nieskuteczne dla długoterminowych prognoz.
- **Modele sztucznej inteligencji** zależą od wielu etapów uczenia się, ich ważność jest ograniczona wystarczającą liczbą zestawów danych treningowych.
- Prognozowanie na poziomie dni bazujące na **szeregach czasowych** są kolejnym podejściem, które łatwo powiela poprzednie wzorce i nie może przewidzieć trendu zmian.
- **Modelowanie oparte na agentach** symuluje współtowarzyszy osobników (agentów) w celu obliczenia rozprzestrzeniania się choroby w społeczności. Niemniej jednak takie modele opierają się na parametrach na poziomie populacji które **nie są jeszcze znane**.
- **Istotnym problemem jest mała ilość informacji i brak wiarygodnych danych dla nowopowstałej choroby.**
- **Poszukuje się przydatnego podejścia do przewidywania dynamiki choroby w celu podejmowania najlepszych decyzji.**

Cel badań

Cel: Celem pracy jest opracowanie i przetestowanie poprzez **symulacje** komputerowe narzędzia prognostycznego funkcjonującego w obszarze danych statystycznych pandemii COVID-19. Narzędzie to może realizować zadania **prognostyczne** umożliwiające wczytanie danych dla dowolnego państwa (także on-line) i wykonanie prognoz wraz z ich **wizualizacją**.



Hipoteza badawcza

Hipoteza: Istnieje wiele metod opartych na różnych założeniach, które umożliwiają prognozę zmian w szeregach czasowych dotyczących pandemii COVID-19. Stawia się hipotezę, że dla wybranego **interwału** czasowego, dla wybranego **państwa** daje się wybrać **model** zmian statystycznych w pandemii i opartą na nim najskuteczniejszą **metodę prognozowania** w powszechnie uznawanych za ważne **szeregach czasowych** – liczby nowych zakażeń, zgonów i ozdowieńców z uwzględnieniem testów i szczepień.



Zakres

1. Ogólna charakterystyka pandemii
2. Przegląd stosowanych w dostępnych źródłach modeli pandemii i metod predykcyjnych np. SIR i pochodne, ARIMA i inne ekonometryczne, oparte na sieciach neuronowych, regresyjne, klasyfikacyjne, kNN i inne.
3. Autorska hybrydowa metoda prognozowania rozprzestrzeniania się wirusa (liczby nowych zakażeń i zgonów) oparta na regresji wielorakiej z pseudoinwersją Moore'a – Penrose'a.
4. Porównanie z metodą GMDH i siecią neuronową ze wsteczną propagacją błędów
5. Symulacje komputerowe na danych historycznych z zastosowaniem uczenia maszynowego
6. Opracowanie zasad i metod akwizycji danych
7. Prezentacja badań dla różnych interwałów czasowych i różnych państw
8. Dyskusja wyników badań
9. Konkluzje, dalsze kierunki badań

Model SIR i modele pochodne

Modele i metody predykcji SIR - wprowadzenie

- Model zaproponowany przez Kermack i McKendrick w **1927** r. w celu symulacji **rozprzestrzeniania się chorób zakaźnych** takich jak odra i różyczka.
- **Zastosowanie: przewidywanie przebiegu choroby w krótkich odstępach czasu.**
- Wymaga obecności **niewielkich ilości** wiarygodnych danych nowopowstałej choroby:
 - wstępne oszacowanie **współczynników śmiertelności i wyzdrowień** dla populacji Wuhan
 - **predykcję polityk rządowych** epidemii w Chinach
 - **predykcja efektów różnych strategii lockdown'u** we Włoszech

Skonstruowany model SIR nie był w stanie przewidzieć rzeczywistego rozprzestrzeniania się i wzorca epidemii COVID-19 w długim okresie

Modele i metody predykcji SIR - opis

Modele SIR zakładają frakcje **podatne** (S), **zakaźne** (I) i **ozdrowiałe** (R) w bliskiej populacji i obliczają tempo zmian w każdej z frakcji za pomocą zwykłego równania różniczkowego (ODE)

Dla sum frakcji musi być spełniony warunek: $\forall t \geq 0 \quad s(t) + i(t) + r(t) = 1$

Warunek ten można interpretować jako skumulowaną liczbę nowych przypadków w każdym punkcie czasowym, które mogą być **ozdrowiałe**, **zmarłe** lub są nadal **zakaźne** z powodu przebiegu choroby.

Suma $r(t) + i(t)$ określa wszystkie osobniki **nie podatne** już na infekcję.

Modele i metody predykcji SIR - zapis różniczkowy

Różniczkowy zapis modelu:

$$\frac{ds}{dt} = -\lambda si$$

$$\frac{di}{dt} = \lambda si - \gamma i$$

$$\frac{dr}{dt} = \gamma i$$

$$R_0 = \frac{\lambda}{\gamma}$$

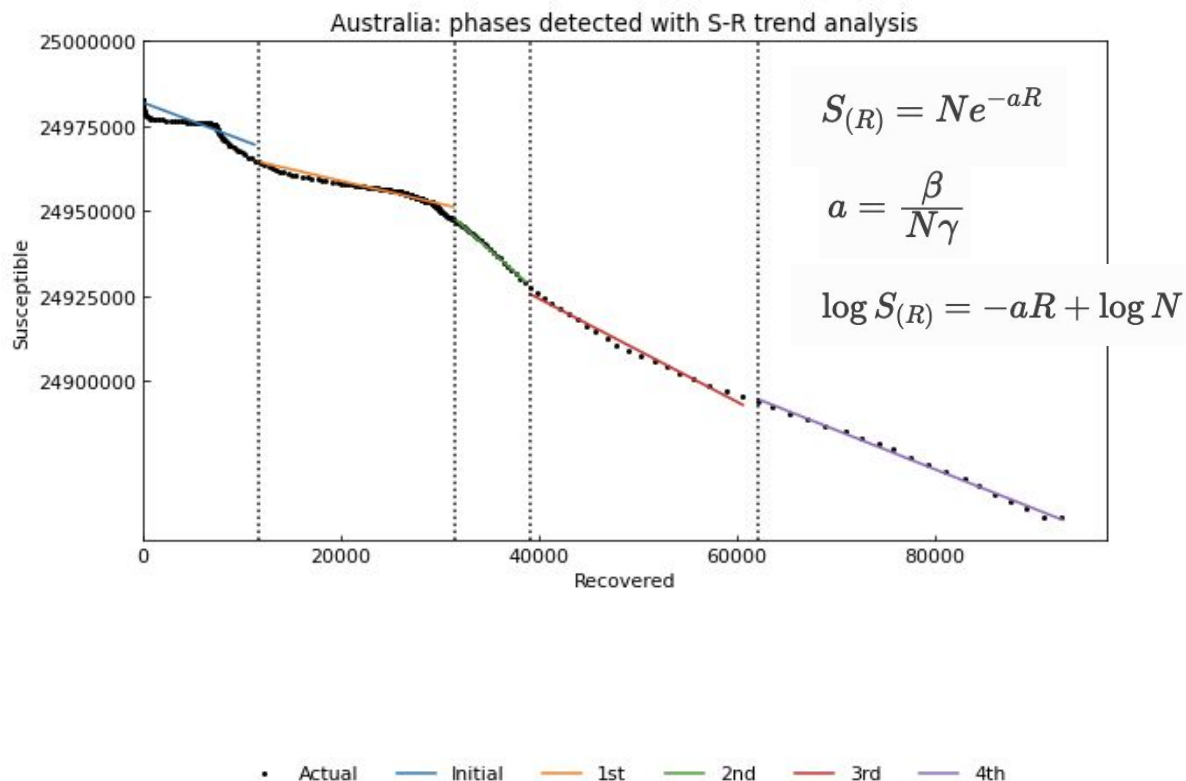
R_0 jest podstawowym wyznacznikiem epidemii i może być bezpośrednio interpretowane jako **oczekiwana liczba nowych przypadków** spowodowane przez osobę zakaźną przed wyzdrowieniem

λ – współczynnik infekcji

γ – współczynnik ozdrowień

R_0 – podstawowa liczba reprodukcji

SIR - analiza trendów - fazy pandemii (Australia)



S - podatni
R - ozdrowieńcy
N - liczba osobników w populacji
Algorytm detekcji zmiany trendu: “Binseg-normal”

Zmiana trendu następuje wraz ze zmianą wartości:
 β (współczynnik kontaktu) i γ (współczynnik ozdrowień + śmiertelności)

Model matematyczny oparty na węzłach

Metoda Manava Bhatnagara

COVID-19 model matematyczny oparty na agentach

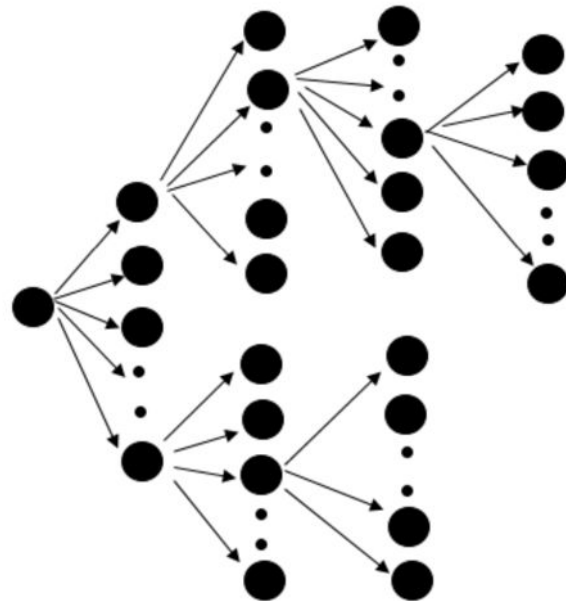
Model matematyczny oparty na agentach:

- Osoba **A** zaraża kolejne **a** osoby

$$T_{cases} = 1 + a + (1 + a)a + (1 + a + (1 + a)a)a + \dots$$

- Po przekształceniu, gdzie **N** - przedział czasu, GP - progresja geometryczna.
- W każdym przedziale czasu N wzrost zakażeń odzwierciedla GP ze wskaźnikiem (1+a)

$$T_{cases} = 1 + a + \underbrace{(1 + a)a + (1 + a)^2a + (1 + a)^3a + (1 + a)^4a + \dots}_{\text{GP of } (N-2) \text{ terms}}$$



COVID-19 model matematyczny: predykcja

W wyniku przekształceń otrzymujemy finalną zależność:

$$T_{cases} = 1 + a + \frac{(1+a)a((1+a)^{N-2}-1)}{a}$$

$$T_{cases} = (1+a)^{N-1}$$

$$T_{cases} = (1+a)^{N-1}$$

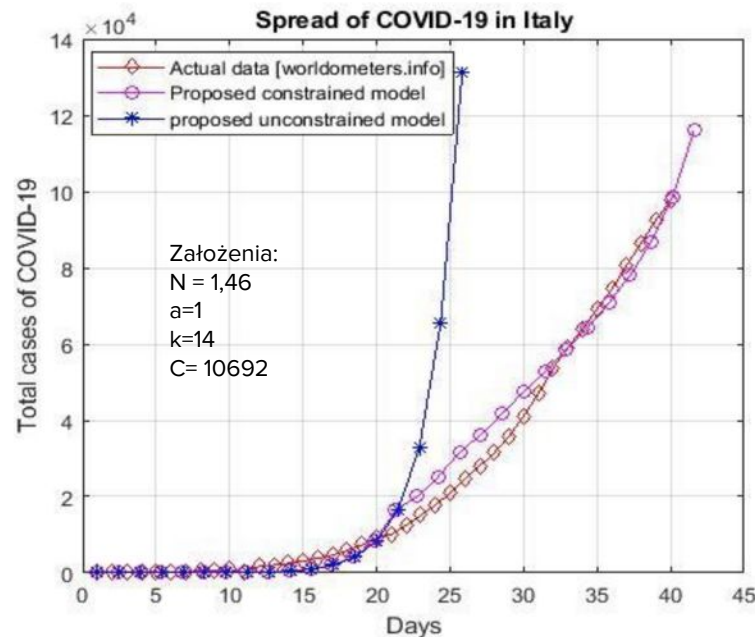
Zakładając, że **jedna osoba zaraża jedną osobę** ($a=1$), cała populacja **P** zostanie zarażona kiedy zostanie spełniona nierówność:

$$P \leq 2^{N_p-1}$$

$$\begin{aligned} T_{cases} &= X_n - X_{n-k} - X_{n-k+1} \\ &= a(1+a)^{n-2} - a(1+a)^{n-k-2} - a(1+a)^{n-k-1}, \quad n > k, \end{aligned}$$

N_p - najmniejsza wartość N spełniająca nierówność
 K - czas inkubacji (aktywności) wirusa w dniach

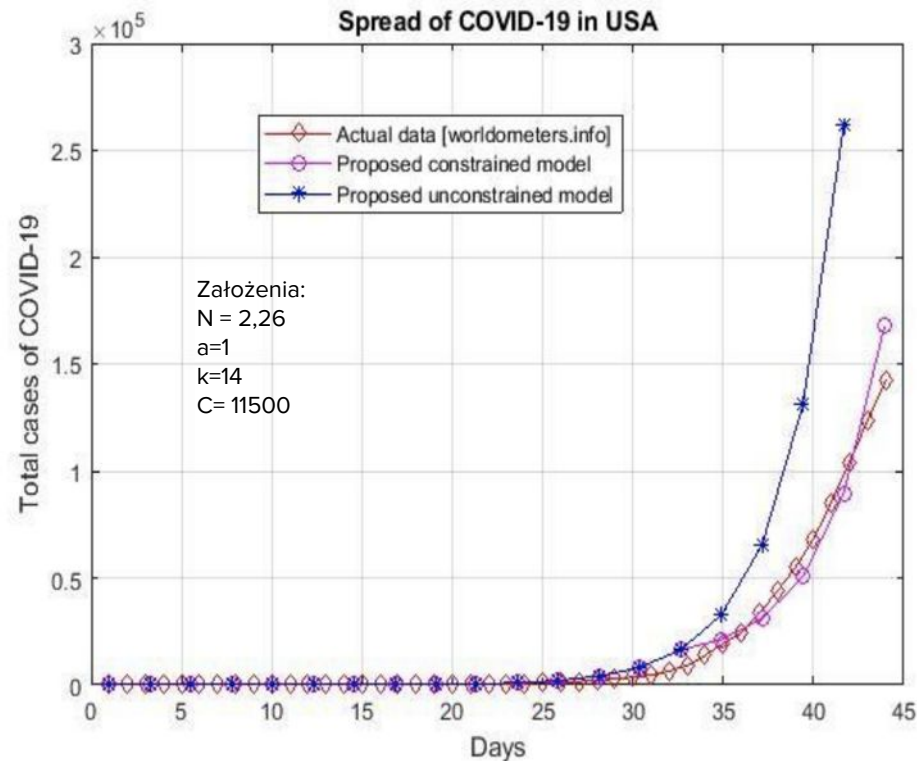
Uwzględnienie izolacji społecznej oraz ozdowień:
C - wskaźnik korekcyjny



COVID-19 model matematyczny: wpływ izolacji i kwarantanny

Model z ograniczeniami:

$$T_{cases} = \begin{cases} X_n, & \text{for } n \leq k, \\ X_n - X_{n-k+1} - C, & \text{for } n > k. \end{cases}$$



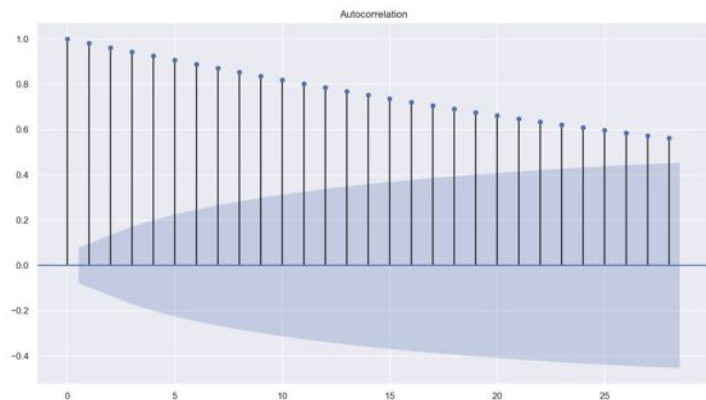
Model ARIMA

ARIMA - stacjonarność danych

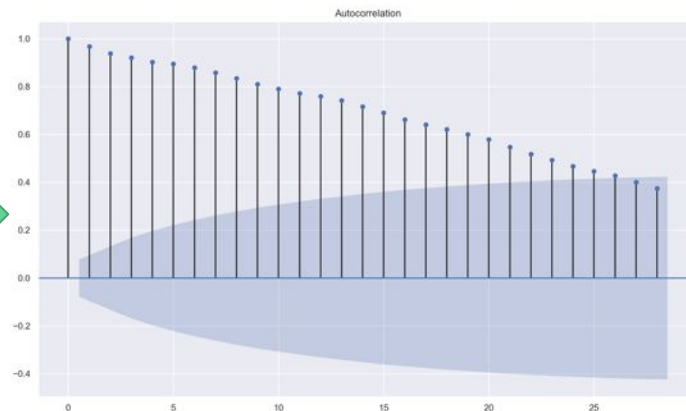
Etapy:

1. Identyfikacja - wykresy szeregów, autokorelacja, stacjonarność szeregów
2. Estymacja
3. Diagnoza

Różnicowanie szeregu - stacjonarność

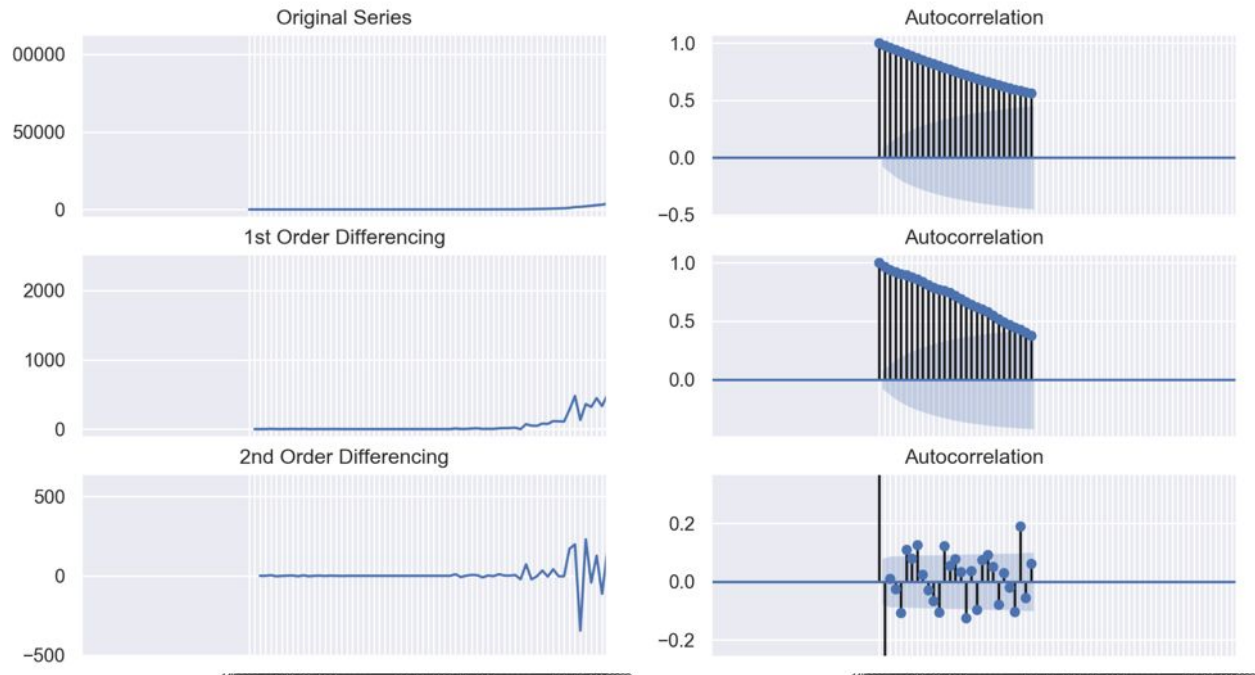


Różnicowanie - 1 poziom



ARIMA - stacjonarność danych (zoom)

Różnicowanie szeregu - stacjonarność



ARIMA - permutacja parametrów

Parametry modelu ARIMA:

- p - parameter autoregresji
- d - rząd różnicowania
- q - parametr średniej ruchomej

AIC - Kryterium informacyjne Akaikego

Tworzenie zbioru - permutacji parametrów {p,d,q} z zadanego zakresu 0..n (np. n=12)

Predykcja oparta na 70% próbek.

Zbadanie parametrów: min(AIC), min(błąd średniokwadratowy)

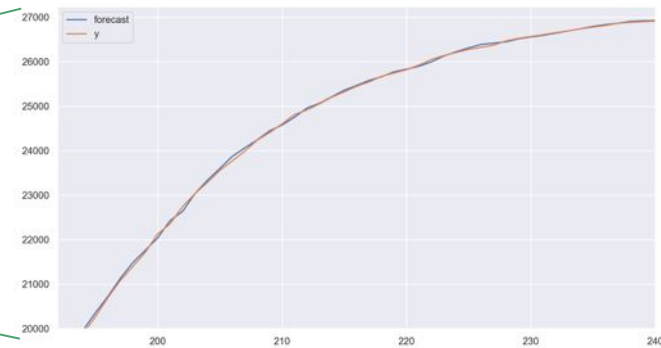
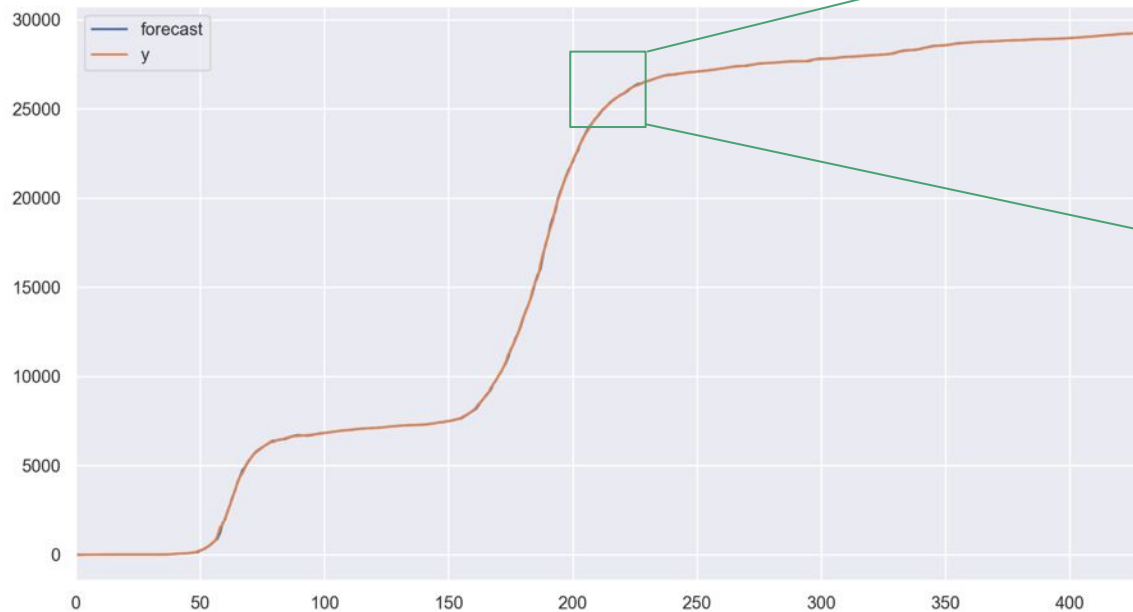
$$AIC = -2 \sum_j \ln(\hat{\pi}_j) + 2q,$$

gdzie:

$\hat{\pi}_j$ – estymowane prawdopodobieństwo, przy założeniach danego modelu, uzyskania takiej właśnie wartości obserwacji j jaka była naprawdę uzyskana,

q – liczba parametrów modelu.

ARIMA - wyniki (6, 2, 6)



ARIMA Model Results						
Dep. Variable:	D2.y	No. Observations:	430			
Model:	ARIMA(2, 2, 2)	Log Likelihood	-2206.139			
Method:	css-mle	S.D. of innovations	40.800			
Date:	Sun, 03 Oct 2021	AIC	4424.278			
Time:	12:32:07	BIC	4448.661			
Sample:	2	HQIC	4433.906			
	coef	std err	z	P> z	[0.025	0.975]
const	0.0508	1.356	0.037	0.970	-2.607	2.709
ar.L1.D2.y	0.6731	0.055	12.210	0.000	0.565	0.781
ar.L2.D2.y	-0.1744	0.054	-3.200	0.001	-0.281	-0.068
ma.L1.D2.y	-1.5454	0.027	-57.293	0.000	-1.598	-1.492
ma.L2.D2.y	0.8919	0.024	36.433	0.000	0.844	0.940
Roots						
	Real	Imaginary	Modulus	Frequency		
AR.1	1.9303	-1.4175j	2.3949	-0.1008		
AR.2	1.9303	+1.4175j	2.3949	0.1008		
MA.1	0.8663	-0.6088j	1.0589	-0.0975		
MA.2	0.8663	+0.6088j	1.0589	0.0975		

Model oparty na regresji wielorakiej GMDH

Metoda GMDH

Grupowa metoda przetwarzania danych (GMDH): rodzina algorytmów indukcyjnych do komputerowego modelowania matematycznego **wieloparametrycznych zbiorów danych**, która umożliwia w pełni automatyczną optymalizację strukturalną i parametryczną modeli.

Metoda GMDH – Group Method of Data Handling

Zastosowanie GMDH:

- poszukiwanie modelu aproksymującego wyniki **obserwowanej zmiennej wyjściowej**
- próba osiągnięcia celu predykcyjnego. Szczególnie wtedy, gdy wyniki pomiarów mają związek z chwilą przeprowadzenia pomiaru a ich wzajemne następstwo ma znaczenie, czyli wówczas, gdy wyniki te **tworzą szereg czasowy**

Metoda GMDH – Group Method of Data Handling

y - zmienna obserwowana

$X = \{x_1, x_2, \dots, x_i, \dots, x_M\}$ - zbiór argumentów, mogących wpływać na wartość zmiennej obserwowanej

Każda zmienna x_i tworzy szereg czasowy:

$X_{ik} = x_{i, t-K}, x_{i, t-K+1}, \dots, x_{i, t-1}, x_{it}$ dla $i=1, 2, \dots, M$ oraz $k=1, 2, \dots, K$

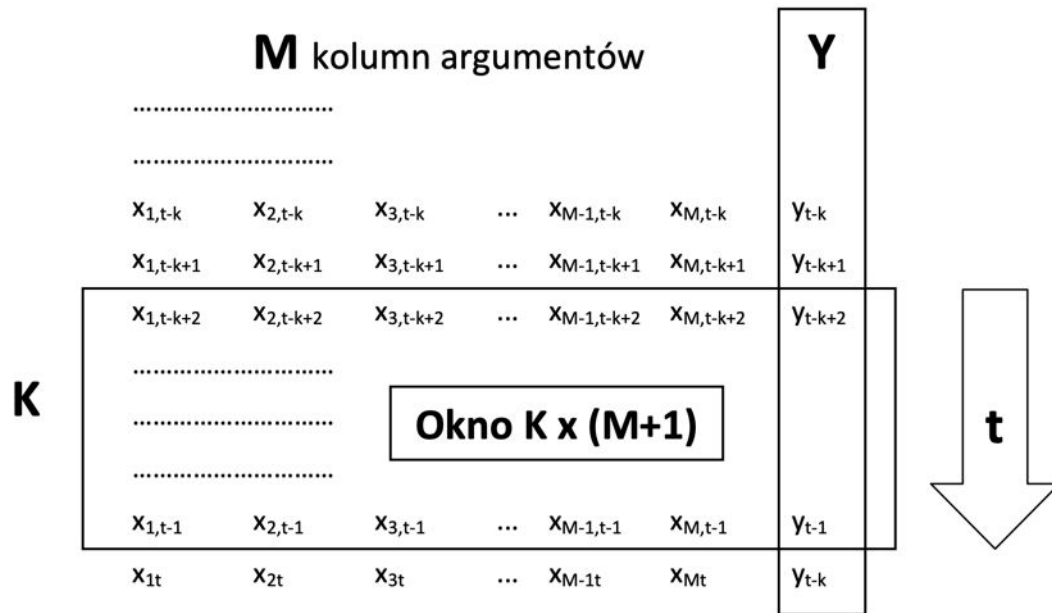
$t=K+1, K+2, \dots$ (indeks chwili bieżącej)

$(t-k)$ - indeks chwili odległej o k okresów

Podobnie:

$y_k = y_{t-K}, y_{t-K+1}, \dots, y_{t-1}, y_t$ dla $k=1, 2, \dots, K$ $t=K+1, K+2, \dots$

K - skończona liczba pomiarów



Modele i metody predykcyjne - regresja wieloraka

Algorytm predykcji:

1. Dane początkowe
 - początek uczenia w szeregach czasowych – beg
 - okno uczenia window - $X_{M \times N}$
 - horyzont predykcyjny hp
 - koniec uczenia fin=beg+window;Uczenia współczynników A przeprowadza się na macierzy X

$X_{M \times N}$ - kroczące okno czasowe o wymiarach $M \times N$, gdzie:

- X macierz danych (skumulowana liczba zakażeń)
- M - liczba wierszy podmacierzy zawierające dane z wybranych państw
- N - liczba dni tworzących okno czasowe

Dane wejściowe (JHU):

```
C=[1750659 1766490 1781345 1794914 1801083
    88668 89414 90009 90331 90525
    1284288 1299002 1312164 1321331 1325291
    793407 796465 799727 801575 805116
    10793732 10869227 10938836 11019344 11051665
    15621 15683 15724 15782 15806];
```



Znormalizowane dane:

```
X = [1.0000 0.7420 0.7108 0.8420 0.7970 0.8790 0.9119
     1.0000 0.7501 0.7161 0.8507 0.8033 0.8807 0.9151
     1.0000 0.7603 0.7232 0.8580 0.8083 0.8846 0.9183
     1.0000 0.7689 0.7297 0.8641 0.8128 0.8881 0.9218
     1.0000 0.7772 0.7359 0.8689 0.8172 0.8909 0.9252
     1.0000 0.7852 0.7420 0.8726 0.8209 0.8942 0.9281
     1.0000 0.7896 0.7455 0.8757 0.8229 0.8967 0.9303];
```

Modele i metody predykcyjne - regresja wieloraka

2. Określa się przesunięty o hp wektor zmiennej objaśnianej Y

$$Y = C(beg+hp : fin+hp)$$

3. Przygotowanie macierzy danych X poszerzonej o wektor jedynek (wyrazów wolnych w wielomianach)

$$X = [\mathbf{1} \quad X_1 \quad X_2 \quad X_3 \quad \dots \quad X_N]$$

4. Obliczenie współczynników równania multiregresji poprzez zastosowanie pseudoinwersji Moore'a-Penrose'a

$$A = X \backslash Y'$$

5. Obliczenie modelu zmiennej objaśnianej Y wykorzystując obliczone współczynniki równań A

$$Y_m = X * A$$

6. Określenie wektora X_{last} ostatnich danych w macierzy X jako wektora z jedynką na pierwszym miejscu (na podobieństwo pkt.3)

$$X_{last} = [1 \quad C1(fin+hp, 2) \quad C1(fin+hp, 3) \quad C1(fin+hp, 4) \quad C1(fin+hp, 5) \quad C1(fin+hp, 6) \quad C1(fin+hp, 7)];$$

7. Obliczenie prognozowanej wartości Y znajdującej się hp pozycji za X_{last}

$$y_{next}(i) = X_{last} * A$$

Modele i metody predykcyjne - regresja wieloraka

8. Obliczenie błędu bezwzględnego prognozy

```
Error(i)=abs(ynext(i)-C1(fin+hp+hp,nr_panstwa));
```

9. Obliczenie bezwzględnego błędu procentowego prognozy

$APE = \text{Error}(i) / C1(\text{fin}+\text{hp}+\text{hp}, \text{nr_panstwa})$

Średni błąd prognozy $MAPE = \text{mean}(APE)$



Modele klasyfikacyjne

Modele i metody predykcyjne - klasyfikacyjne

Metoda bazuje na **podobieństwie wzrostu zachorowań** w Indiach do zmian liczby zakażeń w innych krajach o **podobnych trendach** zmian.

- Zastosowanie **wskaźnika relatywnych zmian** zależnego od szeregów czasowych potwierdzonych przypadków zakażeń (dane JHU).
- Predykcja w horyzoncie 12-dniowym.
- APE ("Absolute Percentage Error") jako **miara dopasowania modelu**.

$$Y_j^m = \max y_j \quad j = 1, 2, \dots, i \text{ dla } i = 1, 2, \dots, N$$

N - liczba dni - obserwacji wartości Conf

$$\text{Conf}(t) = y(t)$$

$$y(i), i = 1, 2, \dots, N$$

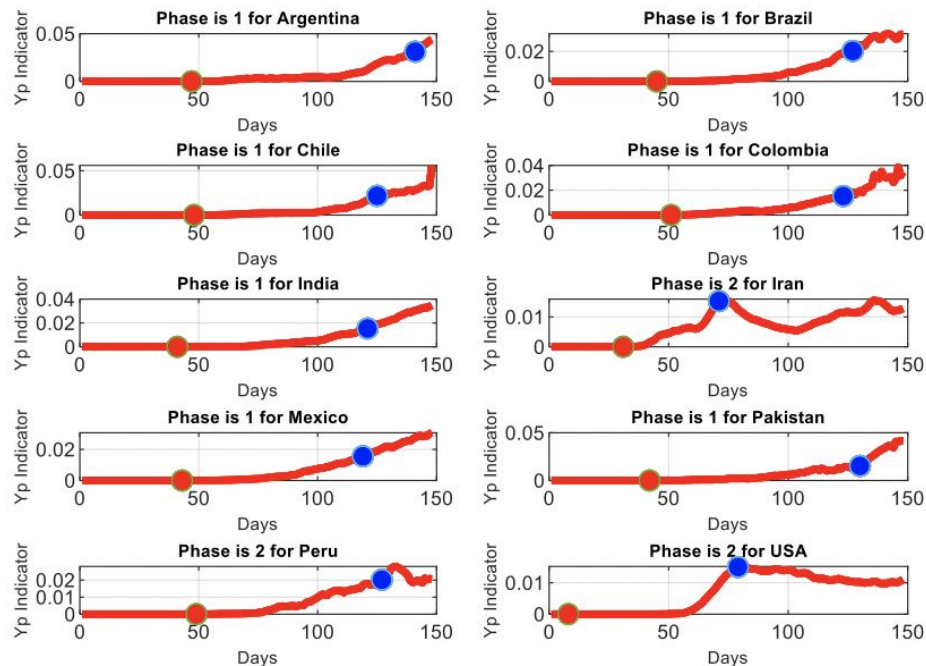
Dzienny przyrost zakażeń: $yp(i) = y(i) - y(i - 1)$ dla $i = 2, 3, \dots, N$

$$\text{Odniesienie do maksymalnej wartości } Y_j^m : Yp(i) = \frac{yp(i)}{Y_i^m}$$

Modele i metody predykcyjne - klasyfikacyjne

- Wskaźnik $Yp(i)$ służy do klasyfikacji faz pandemii w różnych krajach.
- Zaletą wskaźnika jest relatywnie mały zakres zmian w porównaniu z zakresem zmian bezwzględnych w poszczególnych krajach.
- Dla rozróżnienia zmian w mniej lub bardziej dotkniętych krajach wprowadzono progi wartości Yp
- Przekroczenie progu przez ciągłe zmiany Yp oznacza zmianę fazy pandemii w danym kraju.
- Wykres: lista krajów z największą dynamiką wskaźnika Yp
- Iran: wyraźnie widoczny ponowny wzrost

Phases of the Epidemics for the different countries



Modele i metody predykcyjne - klasyfikacyjne

Zmiany wskaźnika Y_p (czerwona linia) dla krajów, w których widoczne jest spowolnienie wzrostu zakażeń oraz kraje referencyjne (stan na 01.06.2021)

Zielone punkty - kraje, w których pandemia wygasa.

Polska, Rumunia, Rosja - utrzymanie trendu zakażeń.

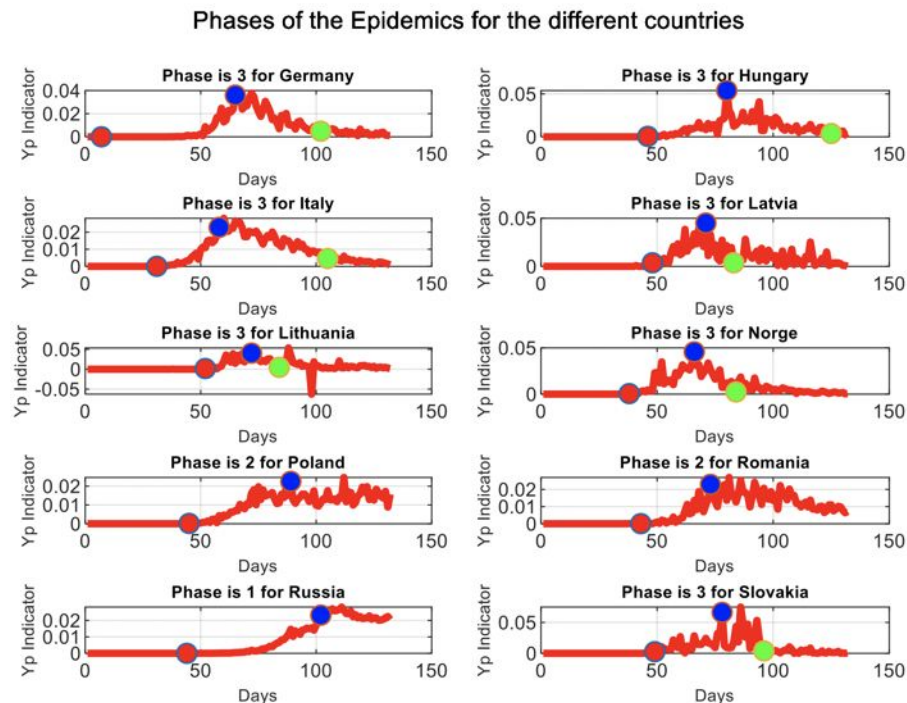
Dla rozróżnienia faz wprowadza się progi:

- górny th_u i dolny th_d , przy czym $th_u > th_d$
- empirycznie określone wartości na podstawie dostępnych danych:

$$th_u = 0.04$$

$$th_d = 0.005$$

Wskaźnik th_u powinien się zmieniać wraz ze zmianą danych i być powiązany ze wzrostem Y_p



Źródło: Antoni Wilinski, "COVID-19 - Forecasts of virus spread in India based on similarity to other countries - June 2020"

Modele predykcyjne zakażeń bazująca na MA

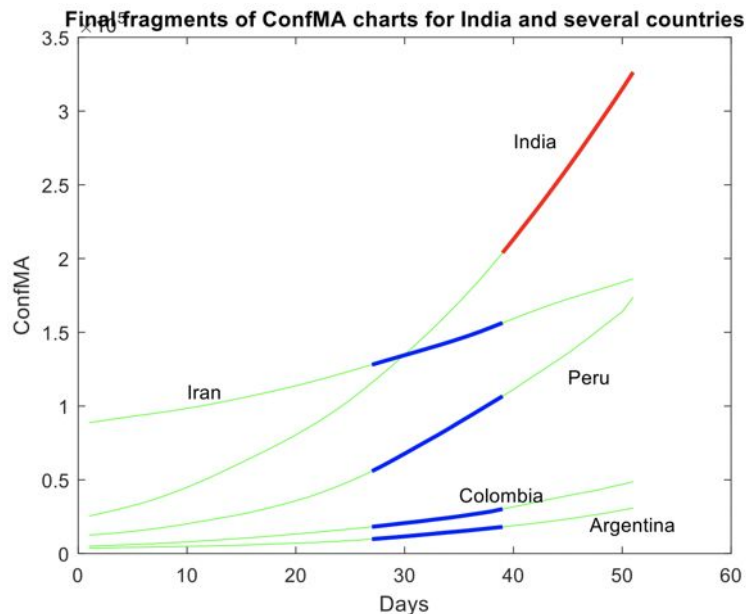
Model zakłada hipotezę poszukiwania podobieństw w przeszłości zmian krzywej Conf wśród innych krajów, nawet tych, które są niżej w rankingu przypadków potwierdzonych, ale mają podobny dynamiczny wzrost w przeszłości. Model bazuje na średniej kroczącej (MA - moving average)

$$ConfMA(i) = \frac{\sum_{i-MA}^i Conf(i)}{MA}$$

Empirycznie ustalona wartość parametru MA=7 (dni)

Wykres: krzywe ConfMA w wybranych krajach i Indiach w wartościach bezwzględnych za ostatnie dni badanego okresu.

- Segmenty ciągłych to fragmenty szeregów czasowych poszczególnych krajów pojawiające się po odcinkach zaznaczonych na niebiesko.



Modele predykcyjne zakażeń bazująca na MA

Czerwony segment przebiegu dla Indii będzie bazowy dla szukania **krajów o podobnych przebiegach** oraz predykcji w oknie czasowym $w = 12$ dni.

W celu porównania zmian ConfMA różnych krajów przeprowadza się **normalizację danych** w zakresie $<0;1>$

Wprowadza się RCI - indeks zastępczy ConfMA:

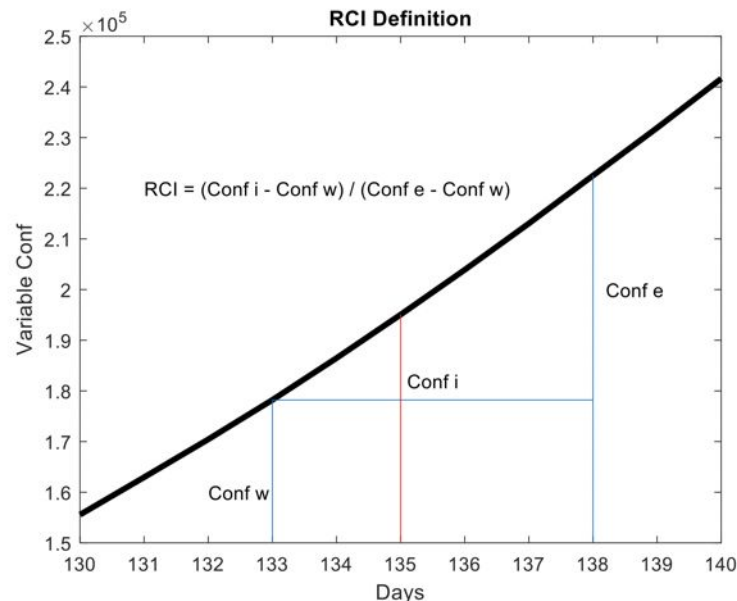
$$RCI_i = (ConfMA_i - ConfMA_w) / (ConfMA_e - ConfMA_w)$$

ConfMA i - średnia krocząca wartość ConfMA z i-tego dnia po w-tym dniu wstecz od końca (bieżącego dnia) w szeregu czasowego ConfMA

ConfMA w - średnia krocząca wartość ConfMA w w-tym dniu wstecz od końca (bieżącego dnia) w szeregu czasowego ConfMA

ConfMA e - średnia krocząca wartość ConfMA na końcu szeregu czasowego ConfMA

Wykres: graficzne wyjaśnienie znormalizowanego wskaźnika RCI

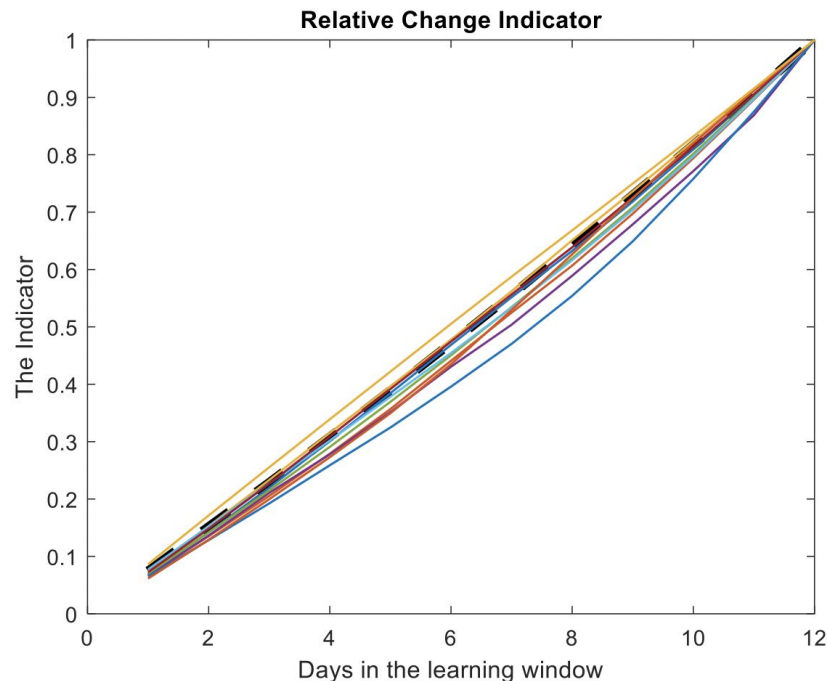


Modele predykcyjne zakażeń bazująca na MA

Wykres RCIc wykonano z uwzględnieniem γp
następujących ograniczeń:

- RCI $c = 0$ na początku okna czasowego w , dla $c = 1, 2, \dots, 9$
- RCI $c = 1$ na końcu okna czasowego w , dla $c = 1, 2, \dots, 9$

Krzywe RCI dla Indii (linia przerywana) i innych krajów
wybranych według kryterium podobnego rozwoju



Model kNN

Modele i metody predykcyjne - kNN i inne

Metoda umożliwia znalezienie państw o **podobnym przebiegu szeregu czasowego**.

Podobieństwo przebiegu krzywej (tutaj: Conf Cases - potwierdzone przypadki zakażeń) względem **zadanego bazowego kraju** - punktu odniesienia.

Dane są **znormalizowane** i dzielone na **segmenty** czasowe: 12 dniowe wstecz oraz 12 dni wprzód.

Państwa wyszukiwane są metodą kNN (np. K=10, K=3, K=5, K=7)

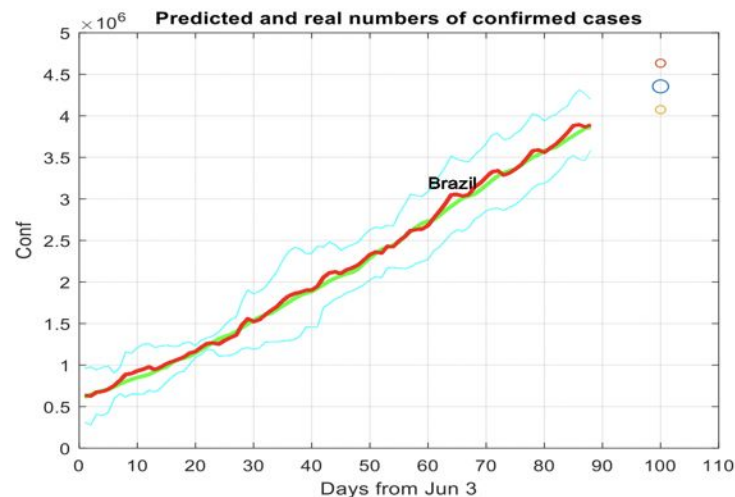
Day	rConf	pConf	APEc	BestCountry	SecondCountry	ThirdCountry
185	2.8591e+06	2.9458e+06	0.03	"Iran"	"Nigeria"	"Poland"
186	2.9122e+06	3.0511e+06	0.048	"Pakistan"	"Mexico"	"Chile"
187	2.9624e+06	3.0531e+06	0.031	"Iran"	"Mexico"	"Peru"
188	3.0124e+06	3.0342e+06	0.007	"Nigeria"	"Indonesia"	"Peru"
189	3.0354e+06	3.0509e+06	0.005	"Peru"	"Nigeria"	"Germany"
190	3.0575e+06	3.1417e+06	0.028	"Italy"	"Ethiopia"	"Mexico"
191	3.1096e+06	3.1922e+06	0.027	"Egypt"	"Ethiopia"	"Indonesia"
192	3.1648e+06	3.2622e+06	0.031	"Egypt"	"Indonesia"	"Chile"
193	3.2249e+06	3.3263e+06	0.031	"Egypt"	"Chile"	"Indonesia"
194	3.2755e+06	3.339e+06	0.019	"Italy"	"Senegal"	"Chile"
195	3.3171e+06	3.2885e+06	0.009	"Philippines"	"Peru"	"India"
196	3.3402e+06	3.3136e+06	0.008	"Kenya"	"Argentina"	"Germany"
197	3.3596e+06	3.3587e+06	0	"Germany"	"Senegal"	"Kenya"
198	3.4074e+06	3.4096e+06	0.001	"India"	"Italy"	"Senegal"
199	3.4567e+06	3.4901e+06	0.01	"Italy"	"Argentina"	"Senegal"
200	3.502e+06	3.5789e+06	0.022	"Senegal"	"Germany"	"Italy"
201	3.5323e+06	3.5893e+06	0.016	"Israel"	"Germany"	"Kenya"
202	3.5824e+06	3.5612e+06	0.006	"Poland"	"Peru"	"Germany"
203	3.6058e+06	3.6106e+06	0.001	"Philippines"	"Colombia"	"Italy"
204	3.6229e+06	3.6595e+06	0.01	"Ethiopia"	"Argentina"	"Kenya"
205	3.67e+06	3.7276e+06	0.016	"USA"	"Senegal"	"Chile"
206	3.7172e+06	3.7996e+06	0.022	"Nigeria"	"USA"	"Israel"
207	3.7614e+06	3.8827e+06	0.032	"Pakistan"	"Israel"	"Nigeria"
208	3.8048e+06	3.8946e+06	0.024	"Israel"	"USA"	"Italy"
209	3.8462e+06	3.8668e+06	0.005	"Italy"	"Kenya"	"Iran"
210	3.8623e+06	3.8928e+06	0.008	"Peru"	"Italy"	"Colombia"

Modele i metody predykcyjne - kNN i inne

12 dniowa prognoza dla Brazylii:

- Prognoza jest sprawdzana po osiągnięciu horyzontu prognozowania
- Różnica pomiędzy **aktualną** a **prognozowaną** wartością oznaczono jako błąd APE (Absolute Percentage Error)

$$APE = \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

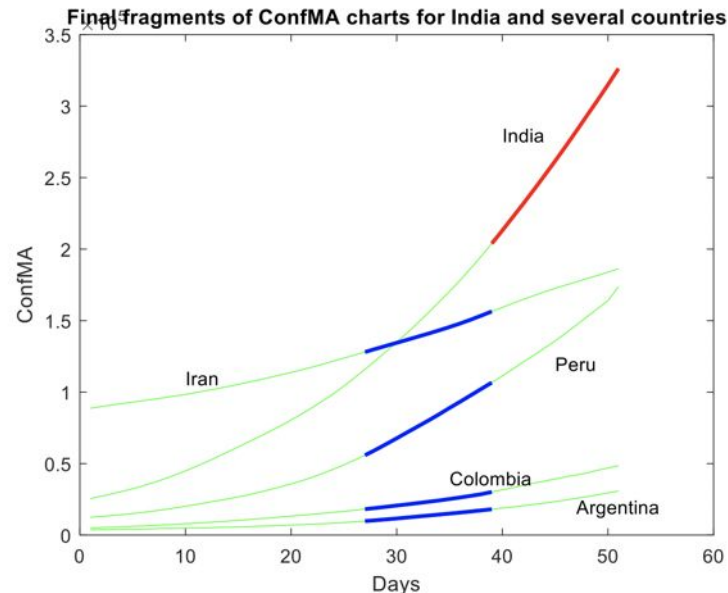
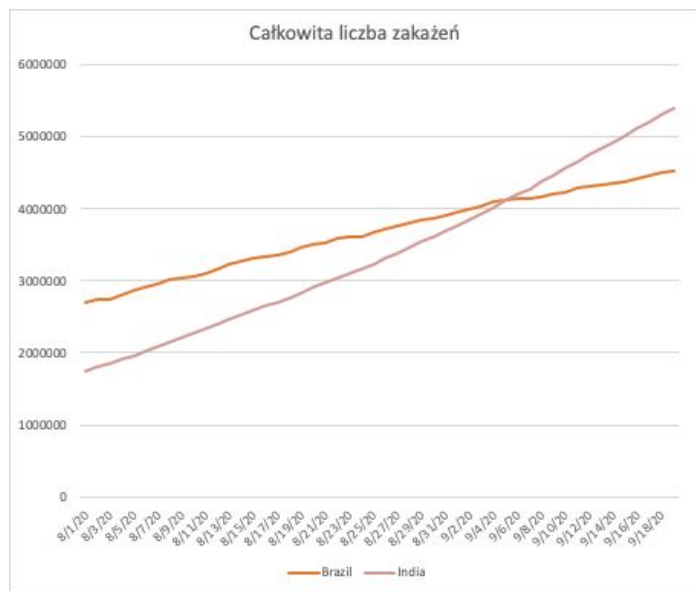


Actual (green) and predicted 12 days earlier (red) time series Conf (t). The larger marker on day 100 is the Confirmed Cases forecast for Brazil. The hundredth day from Jun 3 is Sep 11.

Źródło: Antoni Wilinski, "COVID-19 - When will India catch up with Brazil?"

Modele i metody predykcyjne - kNN i inne

Porównanie segmentów dla kraju bazowego - Indii



The essence of the method - comparing the previous segments with the last base segment (red). In this example, the forecasts are made for India.

Źródło: Antoni Wilinski, "COVID-19 - When will India catch up with Brazil?"

Model oparty na sztucznej sieci neuronowej

Model oparty na sztucznej sieci neuronowej Hopfieldda ze wsteczną propagacją błędów

Sieć ma dość prostą strukturę, ale wystarczającą, jak się okazało w wyniku przeprowadzonych eksperymentów, do stosunkowo **dokładnej krótkoterminowej predykcji**.

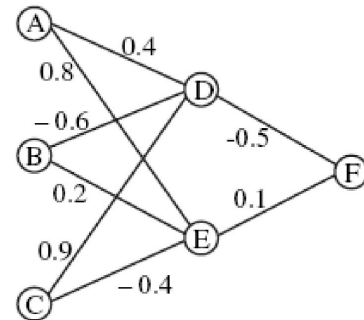
Sieć ma trzy neurony w warstwie wejściowej – A, B i C, dwa D i E w warstwie ukrytej i jeden neuron wyjściowy F, po którym następuje porównanie wyjścia OutputF z wynikiem oczekiwanym ActualF (w fazie uczenia wag), po czym obliczany jest błąd ErrorF i „cofany” by korygować wagi sieci.

Celem była prognoza liczby zakażeń w Polsce (Confirmed Cases) i zmienna ta była wprowadzana jako WejścieA. Dodatkowo na wejście wprowadzano liczby zakażeń w dwóch innych krajach – po kilku próbach wybrano Niemcy i Chile.

Struktura takiej sieci z
przykładowymi wejściami:

Step 1

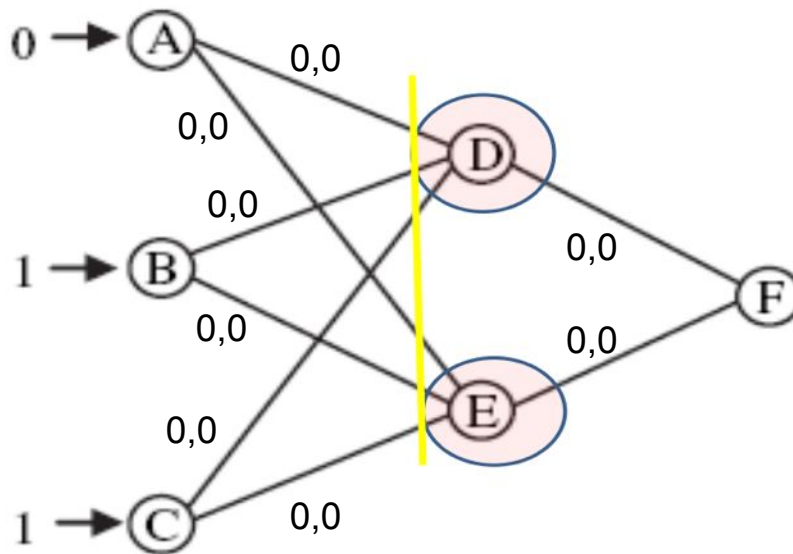
	I_1	I_2	I_3	O
i	1	1	1	1
ii	0	0	0	0
iii	1	0	0	1
iv	0	0	1	0
v	0	1	1	0



Model oparty na sztucznej sieci neuronowej Hopfielda ze wsteczną propagacją błędów
Przykład jednego ze stanów obliczeniowych

Step 2

	I₁	I₂	I₃	O
i	1	1	1	1
ii	0	0	0	0
iii	1	0	0	1
iv	0	0	1	0
v	0	1	1	0



$$Input_D = I_1 \times w_{AD} + I_2 \times w_{BD} + I_3 \times w_{CD}$$

$$Input_D = (0 \times 0,0) + (1 \times 0,0) + (1 \times 0,0) = \text{wynik}$$

$$Input_E = I_1 \times w_{AE} + I_2 \times w_{BE} + I_3 \times w_{CE}$$

$$Input_E = (0 \times 0,0) + (1 \times 0,0) + (1 \times 0,0) = \text{wynik}$$

Model oparty na sztucznej sieci neuronowej Hopfieldda ze wsteczną propagacją błędów

Fragment kodu w środowisku Matlab (funkcja do przeprowadzania obliczeń w sieci)

cz.1.

```
function [W InputF ErrorD ErrorF OutputF]=annLA(Welok, Wp, L, ActualF);
```

```
W=Wp;  
%step2  
%wagi dla drugiego kroku  
Wagi1=W(1:3,1); %wagi pierwszej kolumny, warstwy pierwszej  
Wagi2=W(1:3,2);
```

```
InputD=Welok*Wagi1;  
InputE=Welok*Wagi2;
```

```
%Step3 - wyjście z warstwy ukrytej  
e=2.7182;
```

```
OutputD=1/(1+e^(-InputD));  
OutputE=1/(1+e^(-InputE));
```

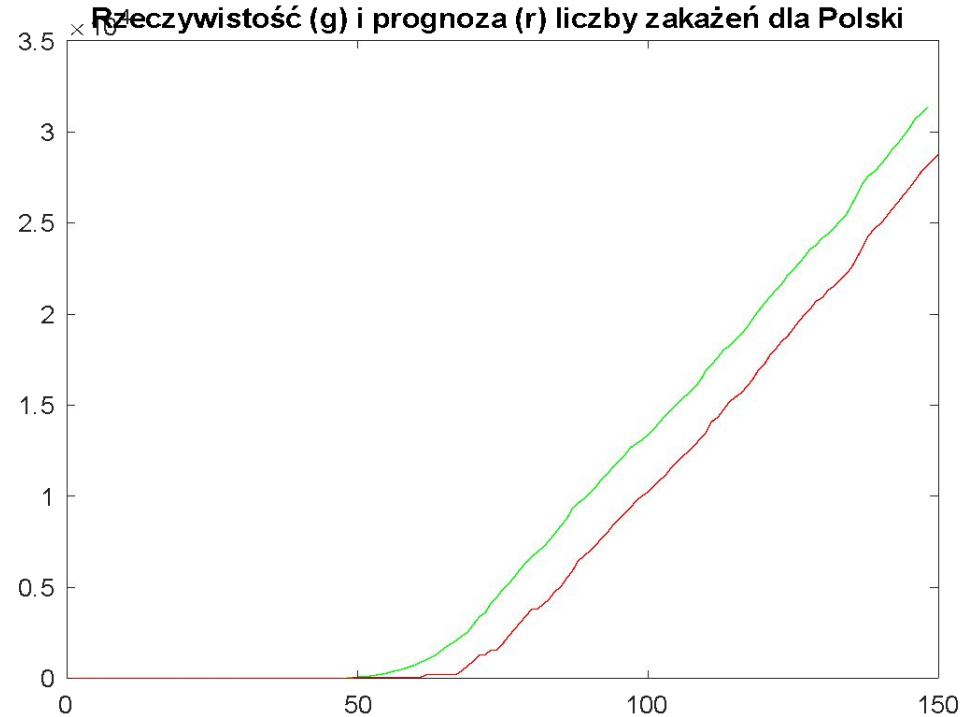
```
%Step 4  
InputF=OutputD*W(4,1)+OutputE*W(4,2);  
OutputF=1/(1+e^(-InputF));
```

```
%Step 5  
%ActualF=0;  
ErrorF=OutputF*(1-OutputF)*(ActualF-OutputF);
```

```
%Step 6 back propagation  
ErrorD=OutputD*(1-OutputD)*ErrorF*W(4,1);  
ErrorE=OutputE*(1-OutputE)*ErrorF*W(4,2);
```

Model oparty na sztucznej sieci neuronowej Hopfielda ze wsteczną propagacją błędów

Przebieg rzeczywisty liczby zakażeń dla Polski dla pierwszych 150 dni pandemii i przebieg jednodniowej prognozy tych zakażeń.



Model oparty na sztucznej sieci neuronowej Hopfielda ze wsteczną propagacją błędów

Symulację wykonano dla wektora wejść neuron A – Polska; neuron B – Niemcy; neuron C – Chile. Na wyjściu notowano wartość liczby zakażeń dla Polski w dniu następnym.

Wnioski:

- symulacja prognoz jest **łatwa**, obserwowane odchylenie jest **proste do zredukowania** wskutek jego systematyczności i przewidywalności.
- nie jest oczywiste jakie szeregi czasowe (jakie państwa) należy podać na wejściu należy je dobrać np. **technika kNN**
- symulacja prognoz w późniejszym okresie pandemii komplikuje się wskutek większej wariancji wartości szeregu czasowego, ale również daje precyzyjne wyniki w horyzoncie jednodniowym

Metoda hybrydowa autorska

Metoda

- Autorska hybrydowa metoda prognozowania rozprzestrzeniania się wirusa (liczby nowych zakażeń i zgonów) oparta na regresji wielorakiej z pseudoinwersją Moore'a – Penrose'a z wierszami (państwami podobnymi) dobieranymi metodą podobną do kNN.
- Porównanie z metodą GMDH i siecią neuronową ze wsteczną propagacją błędów
- Symulacje komputerowe na danych historycznych z zastosowaniem uczenia maszynowego
- Opracowanie zasad i metod akwizycji danych (Dane pobierane w sposób ciągły (raz na dobę) z macierzy instytutu CSSE Uniwersytetu Johnsa Hopkinsa z Baltimore)
- Prezentacja badań dla różnych interwałów czasowych i różnych państw
- Dyskusja wyników badań
- Konkluzje, dalsze kierunki badań

Działalność naukowa, dydaktyczna i zawodowa

Działalność naukowa i edukacja - artykuły

Artykuł	Punktacja
Antoni Wilinski, Łukasz Kupracz, "A Comparative Study of the Multiple Intelligence Profiles of First-Year IT Students and Employed Graduates" (2020), Informatics in Education	100

Edukacja:



POLITECHNIKA KOSZALIŃSKA



ALaRI

Advanced Learning and
Research Institute
Education and innovation in Cyber-Physical
and Embedded Systems



Universitat
de les Illes Balears

Działalność naukowa - artykuły w recenzji

Artykuł	Punktacja
Antoni Wilinski, Łukasz Kupracz, Aneta Senejko, Grzegorz Chrząstek, "COVID-19: Average time from infection to death on the example of Poland, USA, India and Germany", (2021) Quality & Quantity	100
Antoni Wilinski, Arti MK, Lukasz Kupracz, "COVID-19: About the inequality of the territorial distribution of the number of deaths in the world and the indicator of the quality of epidemic management", (2021) IEEE Transaction on Computational Social System	100
Antoni Wilinski, Lukasz Kupracz, Aneta Senejko, Grzegorz Chrząstek, "COVID-19: O jaki interwał czasowy przesunięta jest krzywa zgonów w stosunku do krzywej zakażeń - na przykładzie statystyk Polski", Research Gate (preprint),	

Działalność dydaktyczna

- Zaprojektowanie specjalności “Inżynieria testów oprogramowania”
- Konsultacje odnośnie przedmiotów specjalistycznych na Wydziale Elektroniki i Informatyki Politechniki Koszalińskiej.
- Współautorstwo programów stażowych.
- Prowadzenie zajęć i opiekun przedmiotu: “Bazy danych NoSQL - Wykład”, 2 rok, studia 2 stopnia.
- Konsultant odnośnie utworzenia specjalności Internetu Rzeczy (IoT):
Politechnika Koszalińska, Akademia Pomorska

Działalność zawodowa



Dyrektor programu z zakresu Connected Car



Konsultant ds. jakości

Odbiorcy usług, m. in.:



Dziękuję za uwagę

mgr inż. Łukasz Kupracz