Analiza i modelowanie dynamiki chorób zakaźnych w Europie

Bartosz Wolny, Kacper Kortas, Paweł Kamiński, Sebastian Kaca

Cel i zakres projektu

- Cel: Zrozumienie dynamiki czterech chorób (odra, salmonelloza, gruźlica, krztusiec) oraz próba przewidzenia przyszłych zachorowań.
- Dlaczego to ważne: Wsparcie decyzji w zdrowiu publicznym, alokacja zasobów, ocena skuteczności szczepień.
- Zakres:
  - 31 krajów UE/EOG.
  - Lata:
    - Odra: 1999 - 2025.
    - Krztusiec: 1998 - 2023.
    - Salmonelloza: 2007 - 2024.
    - Gruźlica: 1995 - 2022.
  - Dane roczne dla wszystkich chorób.

Źródła danych

- ECDC:
  - Liczba przypadków,
  - Współczynnik zapadalności,
  - Dane roczne i miesięczne dla krajów UE/EOG.
- Eurostat:
  - Populacja całkowita oraz struktura wiekowa,
  - PKB per capita,
  - Wydatki na służbę zdrowia,
  - Poziom wyszczepialności.
- Wyzwania:
  - Liczba przypadków - potwierdzone przypadki, a zgłoszone przypadki,
  - Niejednolite systemy raportowania w poszczególnych krajach,
  - Braki danych w wybranych latach np. COVID-19,
  - Ograniczona dostępność danych o przypadkach importowanych.

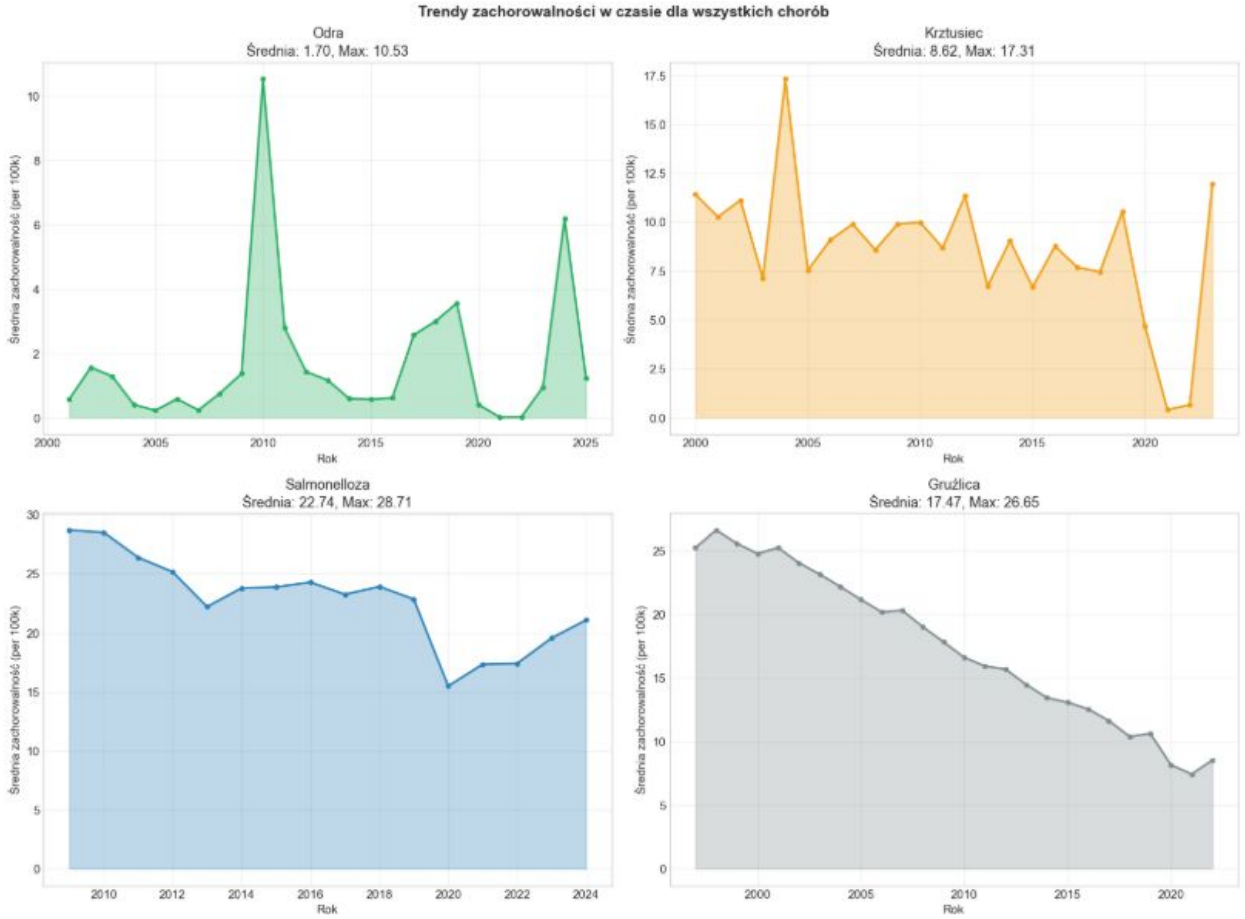
Przygotowanie danych

- Obsługa braków danych
  - Zidentyfikowano braki danych wynikające z różnic w raportowaniu między krajami.
  - Dla pojedynczych braków zastosowano interpolację liniową lub przeniesienie ostatniej dostępnej obserwacji.
  - Kraje z dużą liczbą braków w kluczowych zmiennych wykluczono z dalszego modelowania.
- Standaryzacja
  - Ujednolicono jednostki zapadalności do liczby przypadków na 100’000 mieszkańców.
  - Przeliczono współczynnik zapadalności dla odry z N/1’000’000 na N/100’000, aby była spójna jednostka dla wszystkich chorób.
  - Ujednolicono format daty oraz nazwy krajów (ISO).
- Integracja
  - Dane epidemiologiczne z ECDC połączono z danymi kontekstowymi z Eurostatu.
  - Integracja została wykonana na poziomie kraj-rok (oraz miesiąc dla odry).
  - Ostatecznie uzyskano jeden spójny zbiór danych wejściowych do analizy i modelowania.

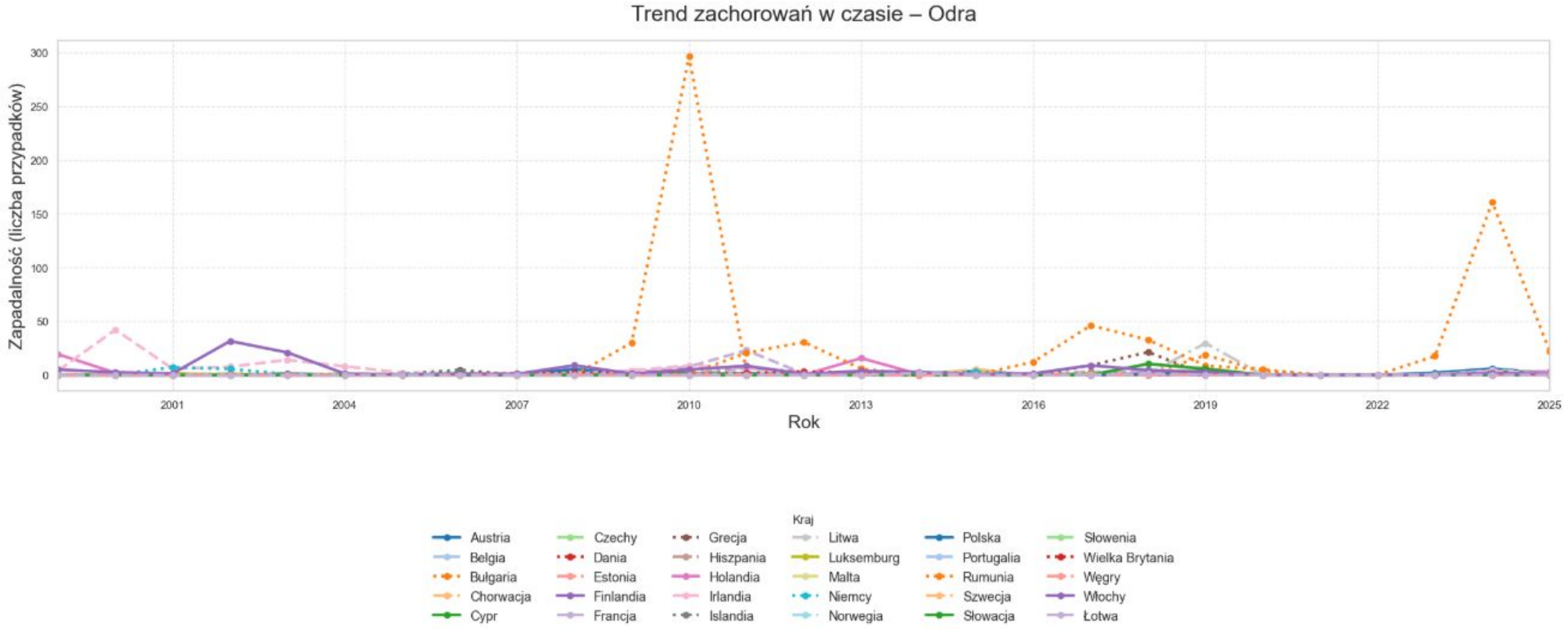
Trendy czasowe

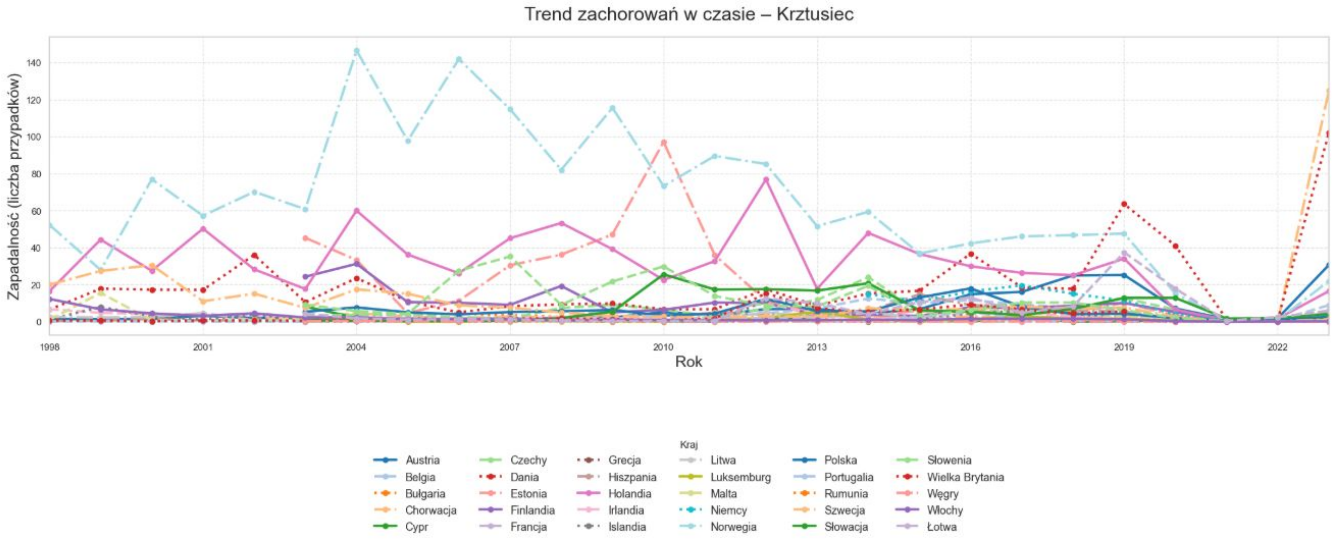
Wnioski:

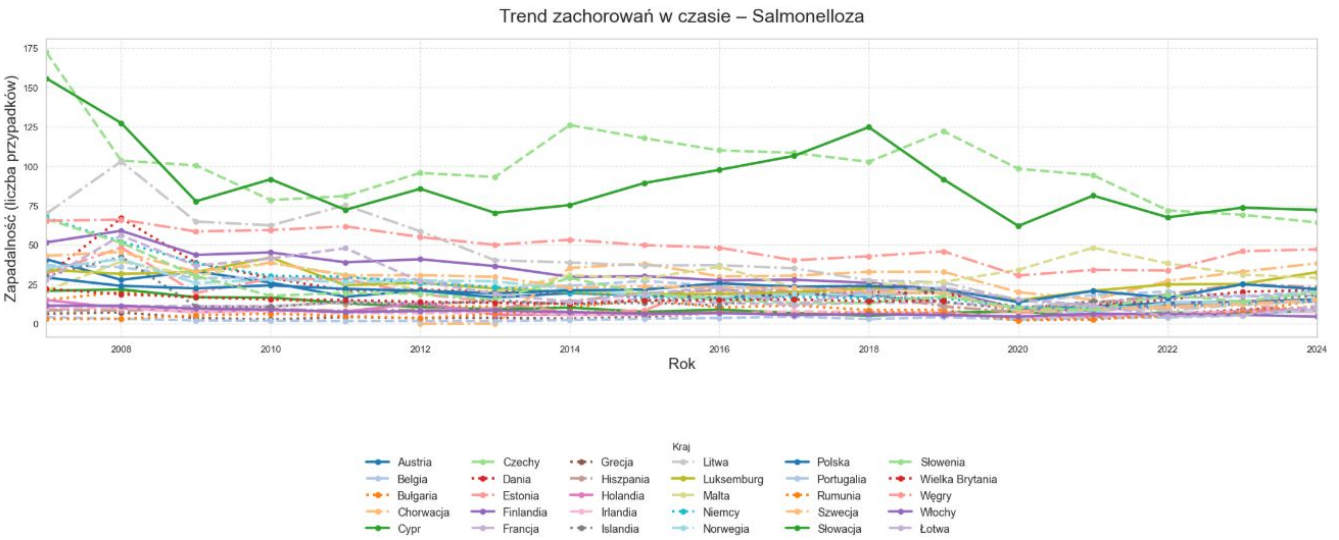
- Wyraźny trend spadkowy dla gruźlicy.
- Odra ma charakterystyczne piki epidemiczne.
- Widać sztuczny spadek w okresie COVID-19, wynikający z ograniczeń oraz braku raportów w tym okresie.

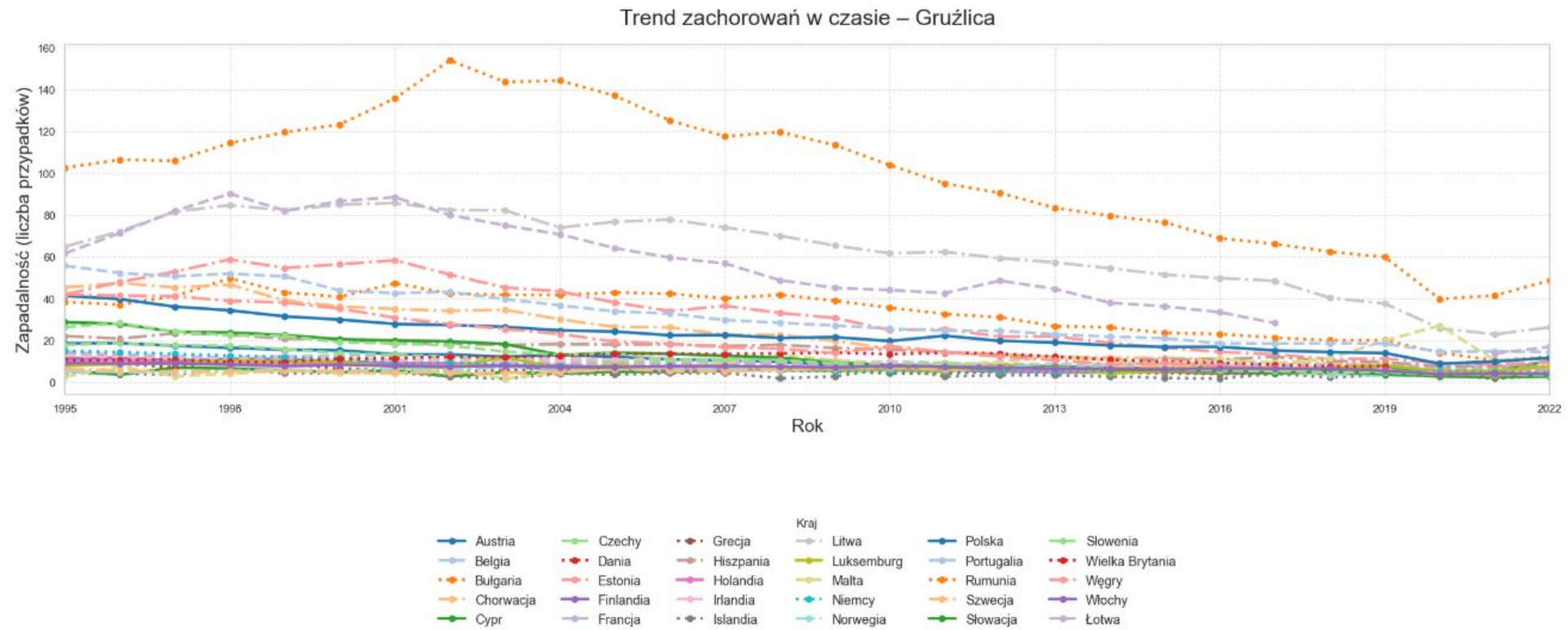


Sezonowość i Geografia




Trend zachorowań w czasie – Krztusiec
Zapadalność (liczba przypadków)
Rok
Kraj
Austria
Belgia
Bułgaria
Chorwacja
Cypr
Czechy
Dania
Estonia
Finlandia
Francja
Grecja
Hiszpania
Holandia
Irlandia
Islandia
Litwa
Luksemburg
Malta
Niemcy
Norwegia
Polska
Portugalia
Rumunia
Szwecja
Słowacja
Słowenia
Wielka Brytania
Węgry
Włochy
Łotwa


Trend zachorowań w czasie – Salmonelloza
Zapadalność (liczba przypadków)
Rok
Kraj
Austria
Belgia
Bułgaria
Chorwacja
Cypr
Czechy
Dania
Estonia
Finlandia
Francja
Grecja
Hiszpania
Holandia
Irlandia
Islandia
Litwa
Luksemburg
Malta
Niemcy
Norwegia
Polska
Portugalia
Rumunia
Szwecja
Słowacja
Słowenia
Wielka Brytania
Węgry
Włochy
Łotwa


Trend zachorowań w czasie – Gruźlica
Zapadalność (liczba przypadków)
0
20
40
60
80
100
120
140
160
1995
1998
2001
2004
2007
2010
2013
2016
2019
2022
Rok
Kraj
Austria
Belgia
Bułgaria
Chorwacja
Cypr
Czechy
Dania
Estonia
Finlandia
Francja
Grecja
Hiszpania
Holandia
Irlandia
Islandia
Litwa
Luksemburg
Malta
Niemcy
Norwegia
Polska
Portugalia
Rumunia
Szwecja
Słowacja
Słowenia
Wielka Brytania
Węgry
Włochy
Łotwa

Wnioski:

- **Odra** - choroba dobrze kontrolowana, ale bardzo wrażliwa na spadek wyszczepialności. Nawet krótkie luki prowadzą do epidemii
- **Krztusiec** - choroba endemiczna, z cyklicznymi falami zachorowań. Jeśli odporność słabnie, to konieczne jest przypomnienie dawki
- **Salmonelloza** - stabilne lub lekko malejące trendy. Brak gwałtownych skoków spowodowany skuteczną kontrolą bezpieczeństwa żywności
- **Gruźlica** - długoterminowy trend spadkowy. Jest to przykład sukcesu zdrowia publicznego, choć choroba nie została wyeliminowana

Wniosek ogólny:
Skuteczność kontroli chorób zależy od szczepień (odra, krztusiec), systemów sanitarnych (salmonelloza) oraz długofalowych programów zdrowotnych (gruźlica).

Korelacje

1. Silne korelacje między odrą a wyszczepialnością, krztusiec - czas od szczepienia, gruźlica - polityka zdrowotna
2. Korelacje międzynarodowe sugerują wspólne czynniki środowiskowe i systemowe
3. COVID-19 działał jak globalny “zakłócacz danych”
4. Brak korelacji między salmonellą a wyszczepialnością

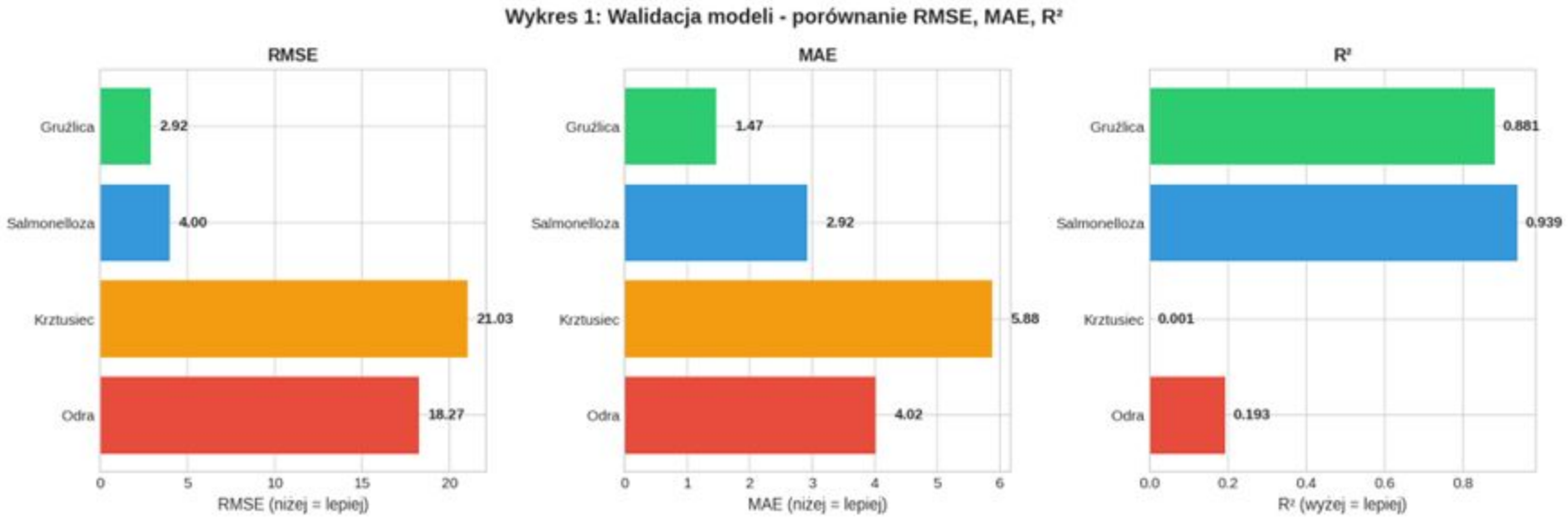
Koncepcja modelu i inżynieria cech

- Zmienna celu: Zapadalność w roku t.
- Zmienne objaśniające:
  - Incidence_Lag1, Lag2 - zachorowalność z poprzednich lat (autokorelacja).
  - Incidence_MA3 - średnia krocząca 3-letnia (wygładzenie trendu).
  - Vaccination_Coverage_Pct - procent wyszczepialności populacji.
  - Pop_Structure_0_14_Pct - udział dzieci (0-14 lat) w populacji.
  - Year_Norm - znormalizowany trend czasowy.
- Podział zbioru:
  - Zbiór treningowy: lata wcześniejsze.
  - Zbiór testowy: ostatnie 2-3 lata.
  - Takie podejście zapobiega wyciekowi informacji z przyszłości i realistycznie ocenia zdolności predykcyjne modeli.

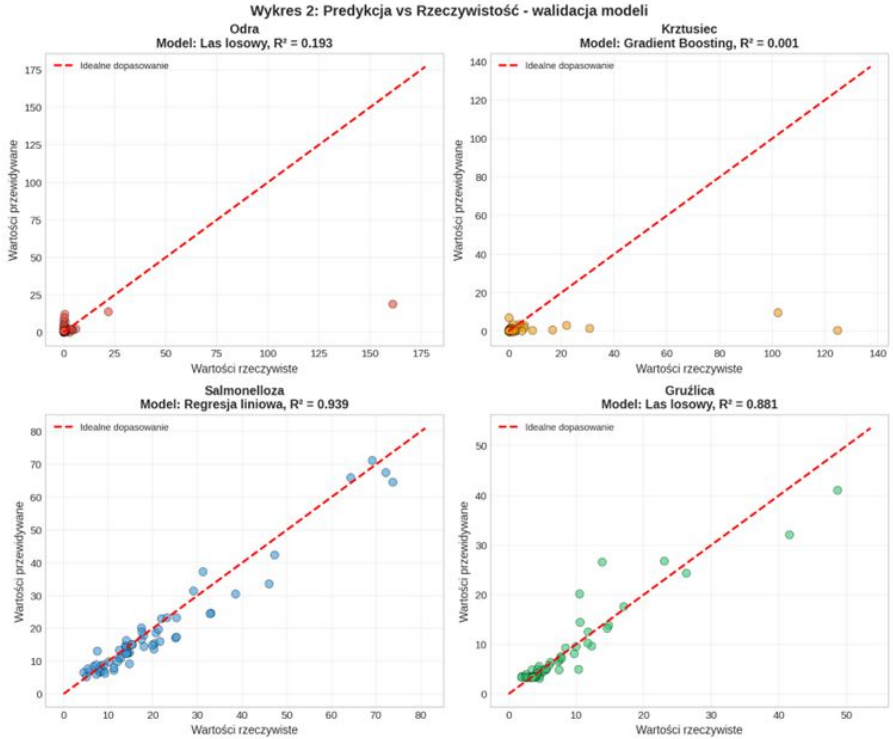
Wybór algorytmu

- Odra: Las losowy
  - Łapie nagłe skoki epidemii lepiej niż modele liniowe
- Krztusiec: Gradient Boosting
  - Model napotkał trudności ze względu na anomalię lat pandemii. Cykliczność choroby została zaburzona przez lockdowny, co sprawiło, że model słabiej przewidział gwałtowny powrót zachorowań w 2023 roku.
- Salmonelloza: Regresja liniowa
  - Zastosowana jako model bazowy dla trendu spadkowego, który w ostatnich latach ulega stabilizacji (wypłaszczeniu).
- Gruźlica: Las losowy
  - Duże różnice między krajami + systematyczny spadek wymagają modelu łapiącego złożone zależności

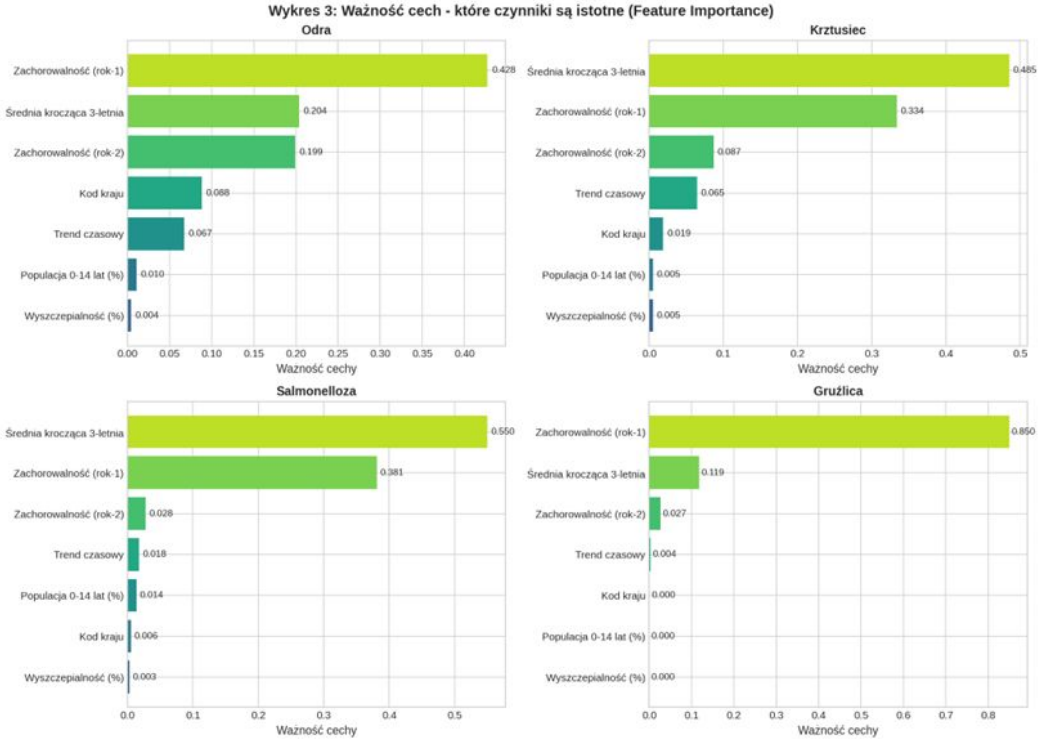
Wyniki i ewaluacja - walidacja modeli



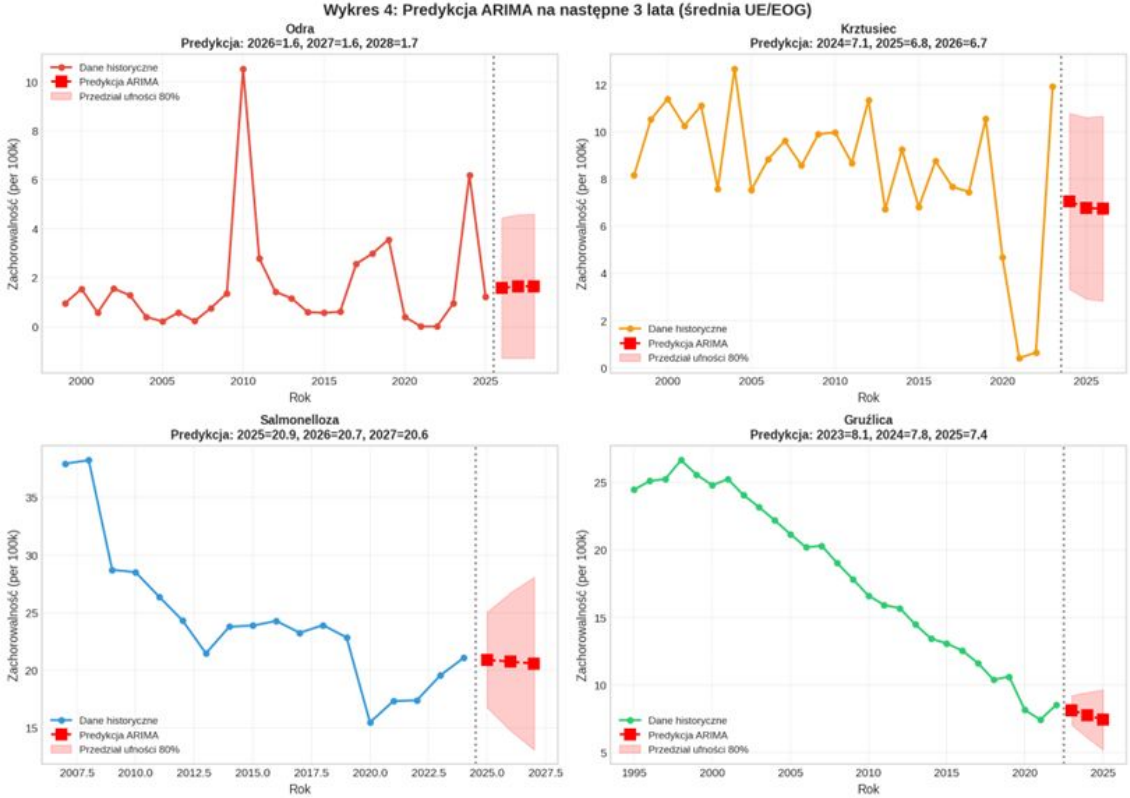
Wyniki i ewaluacja - predykcja vs rzeczywistość



Wyniki i ewaluacja - ważność cech



Predykcja na następne lata



Wnioski merytoryczne

- Ograniczenia projektu
  - Dane roczne: Nie pozwalają na modelowanie sezonowości. Dane miesięczne byłyby lepsze.
  - Brak danych o przypadkach importowanych: Nie można odróżnić transmisji lokalnej od importowanej.
  - Różnice w raportowaniu: Kraje stosują różne definicje przypadków i systemy nadzoru.
  - Brak uwzględniania kontekstu społeczno-ekonomicznego


KONIEC