Interstate infection interaction

Data Preparation

Florian Schweitzer

6.12.2021

CRISP-DM - Data Exploration

Infection data

- Secondary sources wrong scope
  - Timeframe, granularity, ...
- From the source
  - (RKI, Sciensano Epistat, Rivm)

National Institute for Public Health
and the Environment
Ministry of Health,
Welfare and Sport

sciensano

ROBERT KOCH INSTITUT

Factor data

- Border data
  - Mixture of measures
- Holiday & vacation data

Refocussed goals of the Data Mining

Assess the impact of vacation and holidays to infections statewise, crossborder and the EMR as a whole

CRISP-DM - Data Description

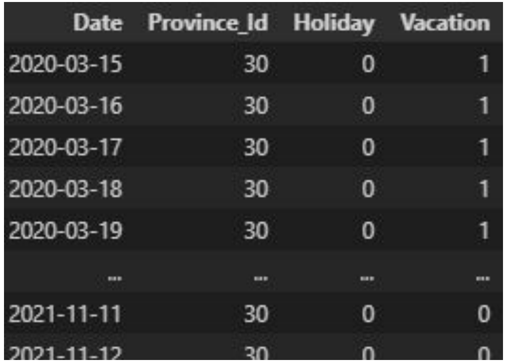
Infection data

Original data from different sources as UTF-8 csv containing the

- reference date
- the number of cases split into subgroups
  - province, age, sex, …
- Other data (hospitalization, death, report date, …)

Day-Off data

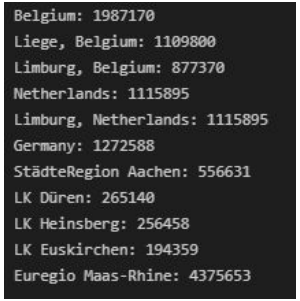
Self-constructed UTF-8 csv format



Date	Province_Id	Holiday	Vacation
2020-03-15	30	0	1
2020-03-16	30	0	1
2020-03-17	30	0	1
2020-03-18	30	0	1
2020-03-19	30	0	1
...	...	...	...
2021-11-11	30	0	0
2021-11-12	30	0	0

CRISP-DM - Data Preparation

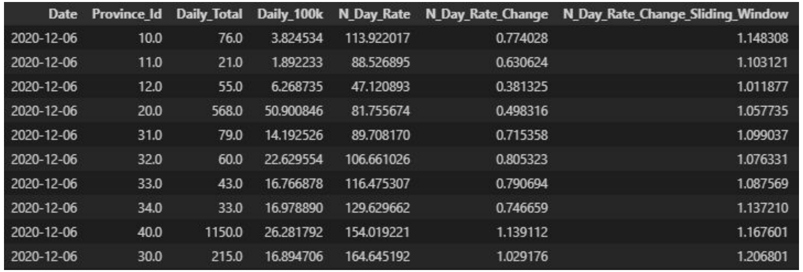
Initial timeframe chosen (2020-03-15 … 2021-11-15)



```
Belgium: 1987170
Liege, Belgium: 1109800
Limburg, Belgium: 877370
Netherlands: 1115895
Limburg, Netherlands: 1115895
Germany: 1272588
StädteRegion Aachen: 556631
LK Düren: 265140
LK Heinsberg: 256458
LK Euskirchen: 194359
Euregio Maas-Rhine: 4375653
```

Original attributes

Date, Province_Id, Daily_Total/Sum_Total



Date	Province_Id	Daily_Total	Daily_100k	N_Day_Rate	N_Day_Rate_Change	N_Day_Rate_Change_Sliding_Window
2020-12-06	10.0	76.0	3.824534	113.922017	0.774028	1.148308
2020-12-06	11.0	21.0	1.892233	88.526895	0.630624	1.103121
2020-12-06	12.0	55.0	6.268735	47.120893	0.381325	1.011877
2020-12-06	20.0	568.0	50.900846	81.755674	0.498316	1.057735
2020-12-06	31.0	79.0	14.192526	89.708170	0.715358	1.099037
2020-12-06	32.0	60.0	22.629554	106.661026	0.805323	1.076331
2020-12-06	33.0	43.0	16.766878	116.475307	0.790694	1.087569
2020-12-06	34.0	33.0	16.978890	129.629662	0.746659	1.137210
2020-12-06	40.0	1150.0	26.281792	154.019221	1.139112	1.167601
2020-12-06	30.0	215.0	16.894706	164.645192	1.029176	1.206801

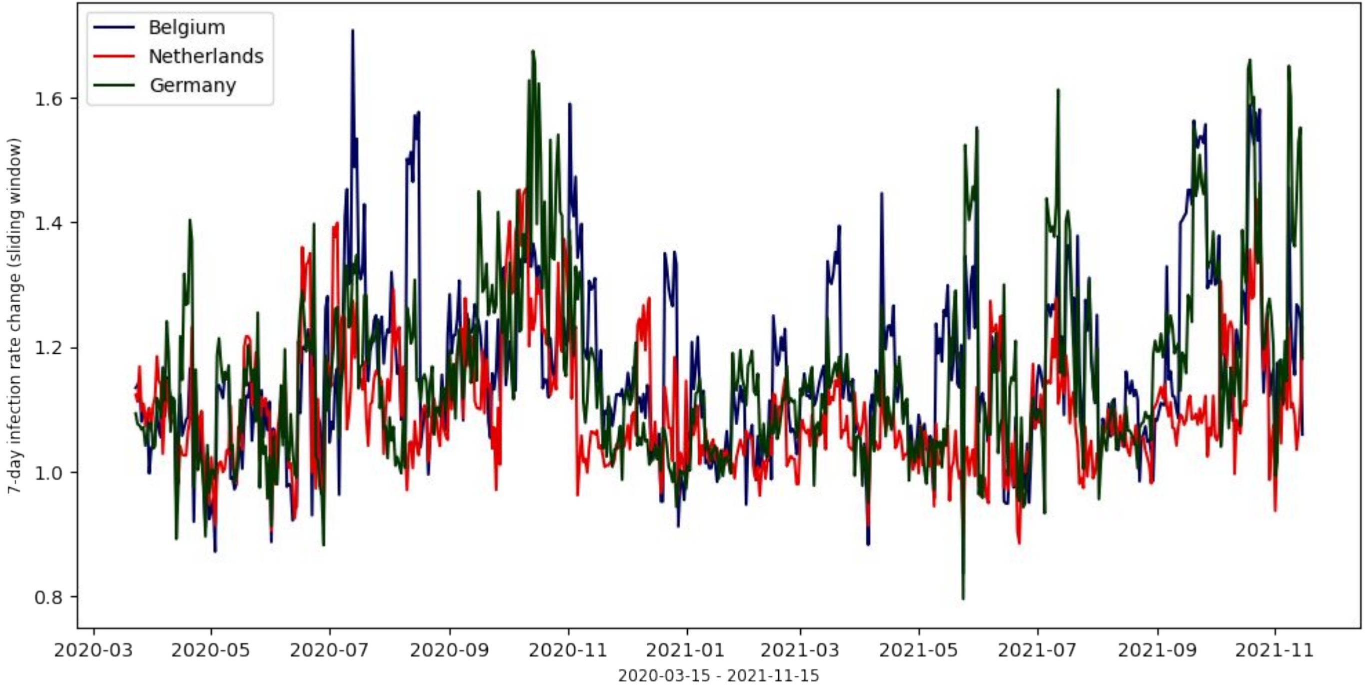
Derived Data

- Daily_100k
- N_Day_Rate
- N_Day_Rate_Change
- N_Day_Rate_Change_Sliding_Window

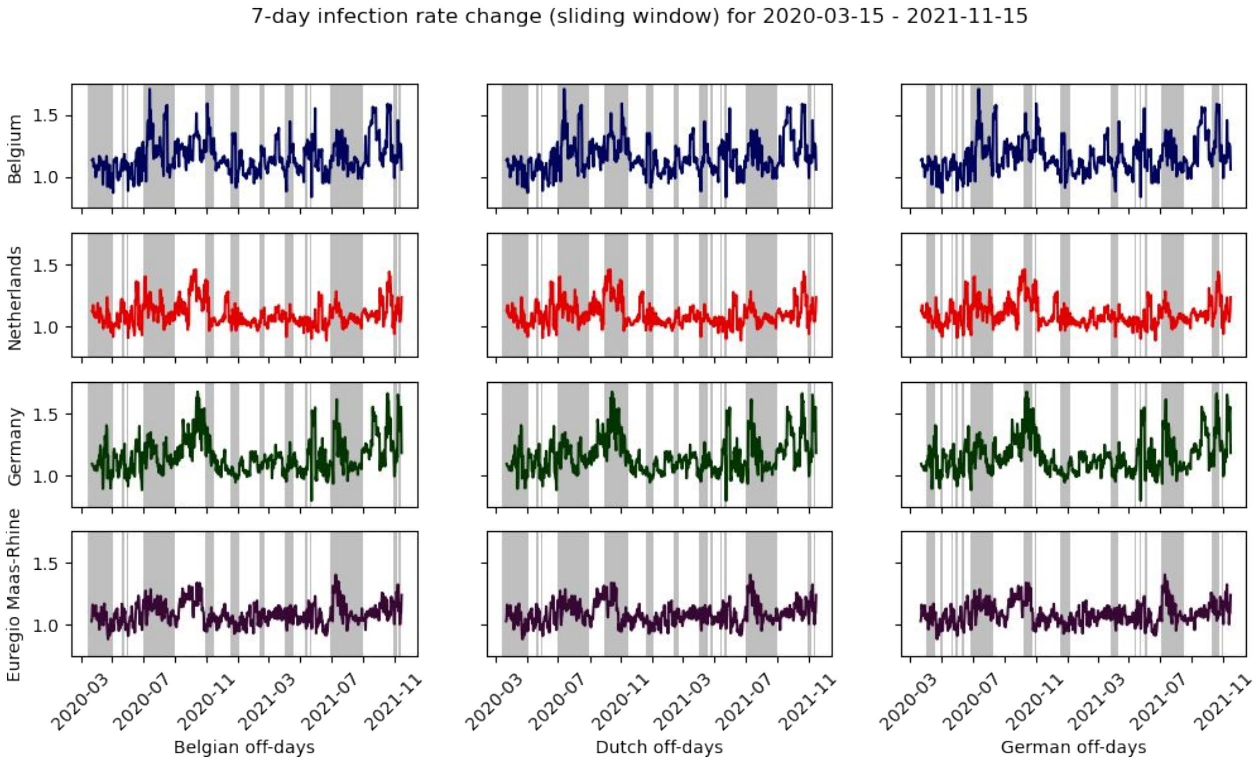
Exported to csv

CRISP-DM - Data Preparation

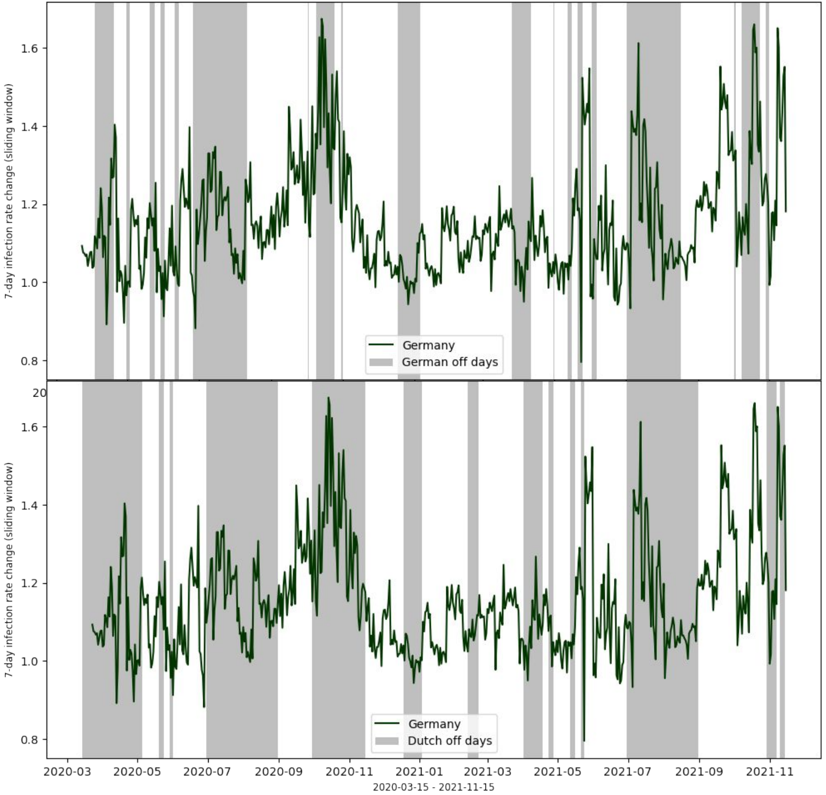
… but before modeling: visualize!



CRISP-DM - Data Preparation



CRISP-DM - Data Preparation

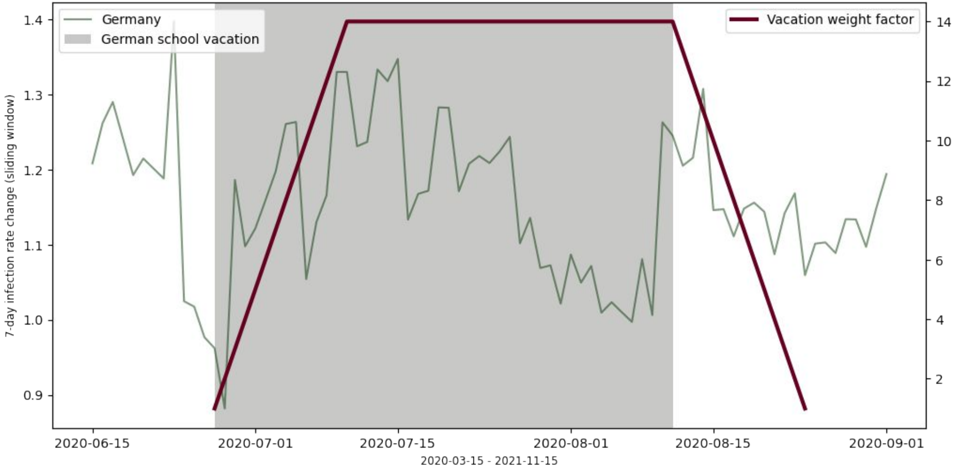


CRISP-DM - Data Preparation

Describing holidays in numbers

Off-Day Streak & Vacation weight factor

- Delayed rise of infection rate
- Longer streaks show change



CRISP-DM - Project Roadmap

Project Plan

Construct infection timelines

- ✔ State based/EMR based
- ✔ Use/Create a measure that is independent from baseline

Inspect infection events

- ✔ Create factor (vacation&holiday) timelines
- ✔ Find/Create a feature for dependency
- Check correlation

Modeling

- Train a prediction-model for change in infections based on focus-factors
- Verify results

CRISP-DM - Modeling and Evaluation

Modeling

Low sample size (~500-600)

Target modeling algorithms

- Boosted Trees
- Support Vector Regression (linear/rbf)
- Ridge Regression

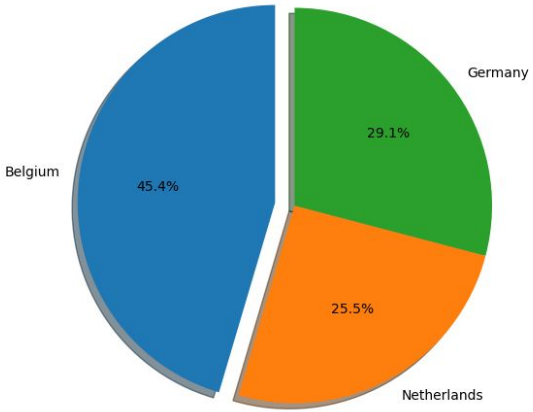
Evaluation

Tendency of change rate

Verification by MSE

Imbalanced population

Differing impacts on neighbours or EMR



Sources

- Border situation/regulations
  https://itemcrossborderportal.maastrichtuniversity.nl
  https://ec.europa.eu/
  https://www.info-coronavirus.be/
- CoViD-19 data
  - RKI CoViD-19 Datahub
    https://npgeo-corona-npgeo-de.hub.arcgis.com/
  - Epistat
    https://data.gov.be
    https://epistat.wiv-isp.be/covid/
  - National Institute for Public Health and the Environment
    https://data.rivm.nl/covid-19/
- Holidays
  https://holidaycalendar.com/
  https://www.feiertagskalender.ch/