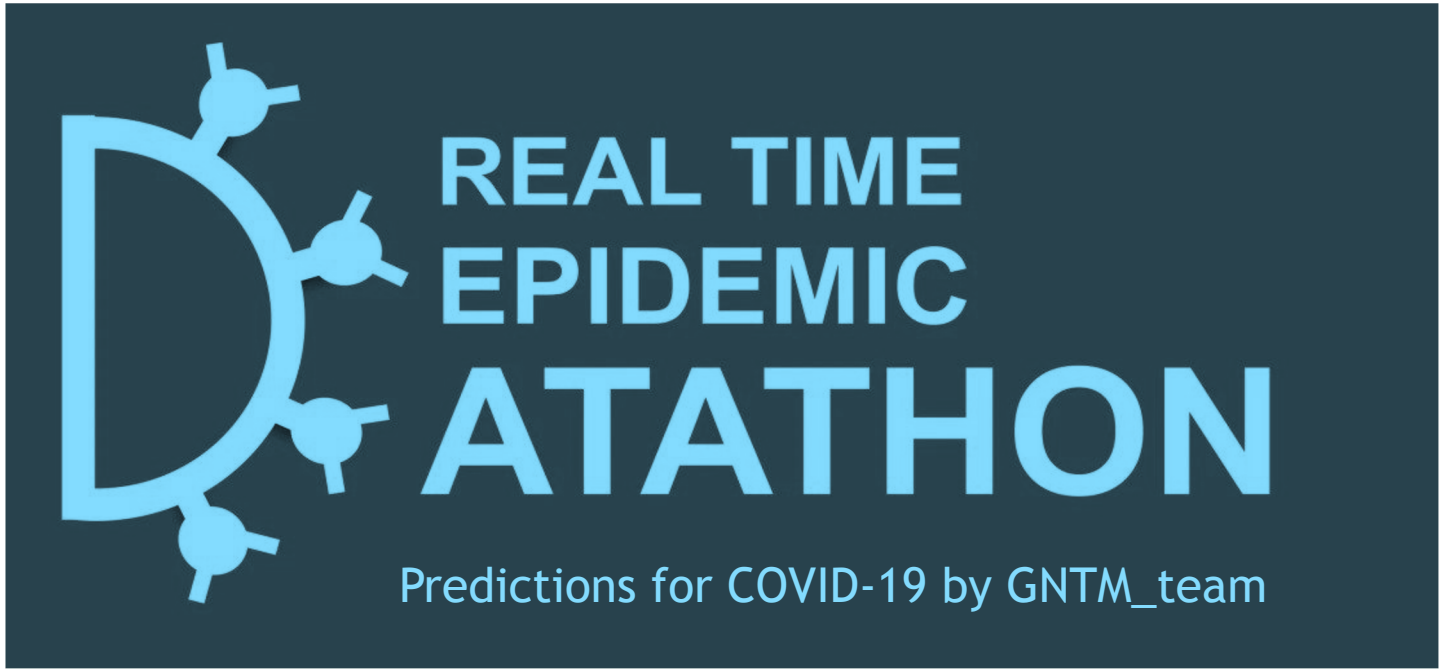


Data Science in Techno-Socio-Economic Systems, ETH Zürich, 2020

Dimitrios Gkouletsos
Athina Nisioti
Marina Panteli
Chrysa Papadopoulou
Iliana Papadopoulou

Presentation Outline

- Motivation for taking part in the Epidemic Datathon
- Biological assumptions for Covid-19 & Focal groups
- Data pre-processing
- Characteristics of chosen models
- Conclusions

Participation Motivation

- The pandemic has undoubtedly been the defining characteristic of 2020:
  - Afflicted the entire planet
  - Ignited interest in the scientific community
- Passion for Data Science (using prior experience and continuing to learn)
- Chance to contribute in the scientific sphere during these seemingly ‘unpredictable’ times

Covid-19: Assumptions

Nature of Covid-19

- No chance of reinfection
- Symptoms can be expressed within 14 days of contracting infection
- Critical condition patients: 5% of cases

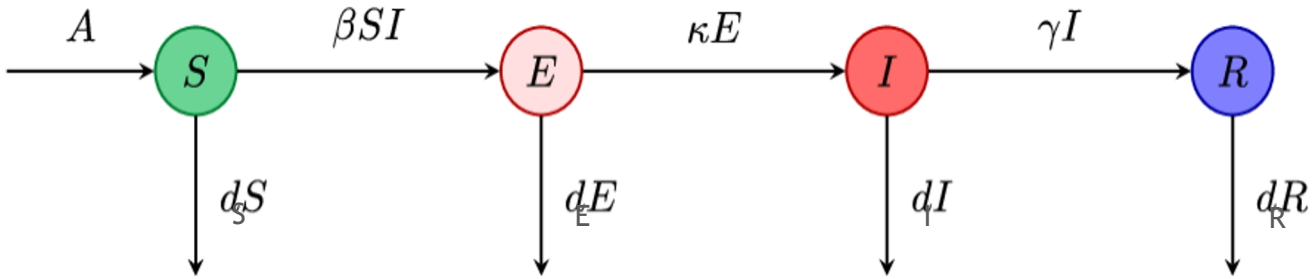
Subpopulations (cumulative)

- Infected
- Recovered
- Deceased
- Critically ill

All countries (provinces/states)

Rejected models for predictions

- Epidemiological model e.g. SEIR



  - Reason for rejection: estimation of parameters incredibly challenging, especially because of insufficient information (e.g. initial stochasticity)

Data Preprocessing

- Data used from the “John Hopkins Repository”
- Cleaning of the dataset (removal of irrelevant features and countries with non-reasonable values)
- Aggregation of all the cities/ sub-regions to acquire total cases for states
- Concatenation of the global dataset with the US dataset

Complete training set

ARIMA: Characteristics

- ARIMA (p,d,q): Autoregressive Integrated Moving Average model

p: Number of lag observations included in the model
d: Number of times that the raw observations are differenced
q: Size of the moving average window

- After differencing:

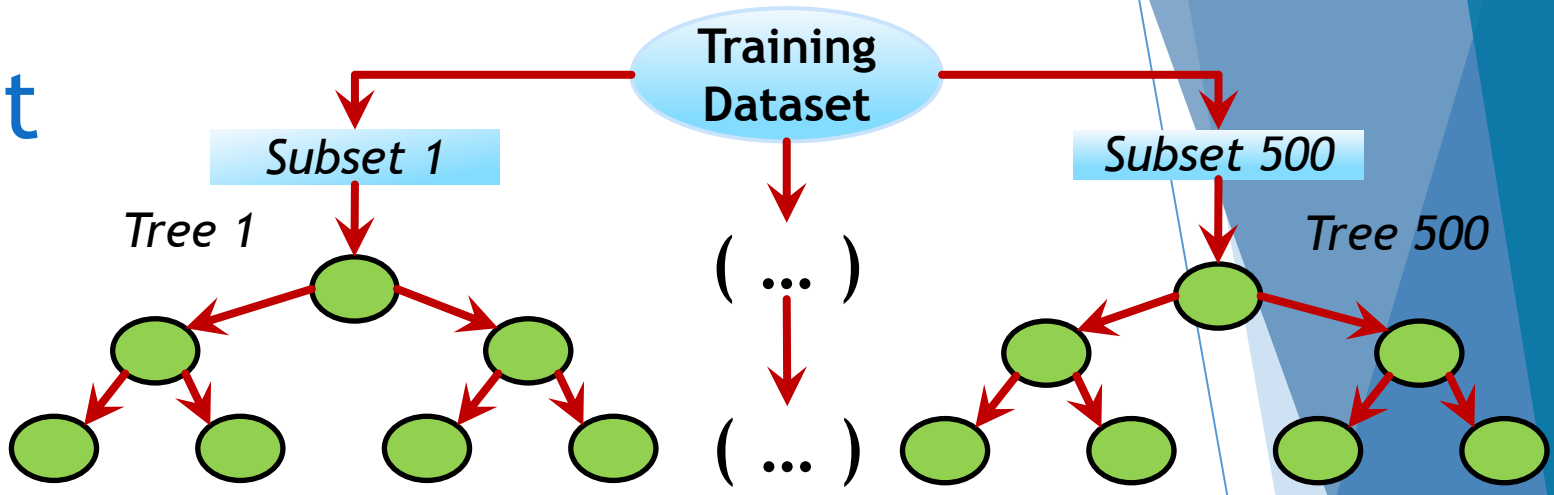
$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + Z_t - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q}$$

$$Z_t \sim WN(0, \sigma_Z^2)$$

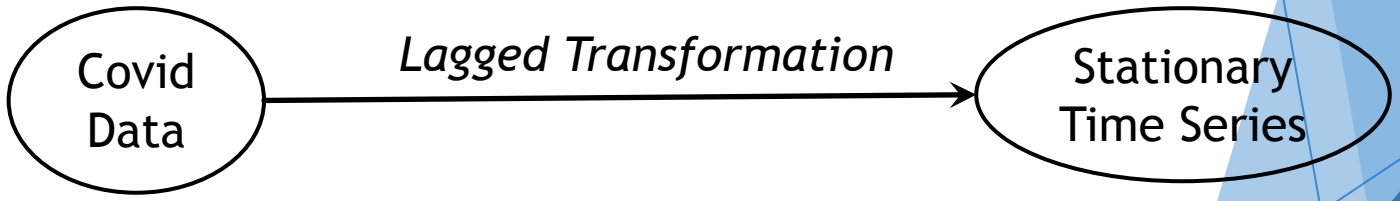
ARIMA: Parameter estimation

- Our dataset: cumulative cases → Non-stationary timeseries
- Augmented Dickey - Fuller statistical test in $\alpha=0.1$ to estimate d
- Estimation of remaining ARIMA parameters → Akaike information criterion (**AIC**)

Random Forest

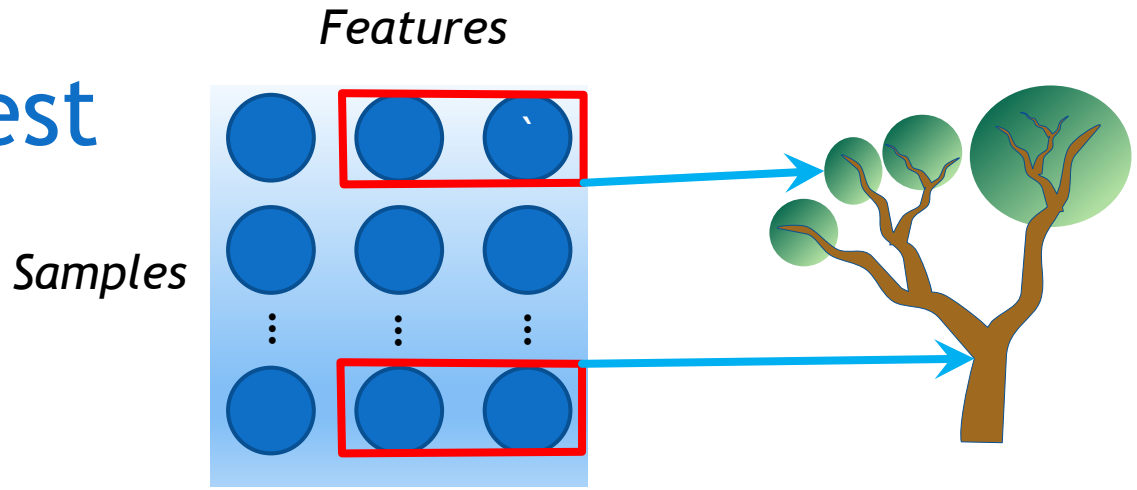


- **Main Structure:**
  - Consists of Regression Trees. Each tree spits out a prediction.
  - Builds independent parallel model
- **Data Preprocessing:**



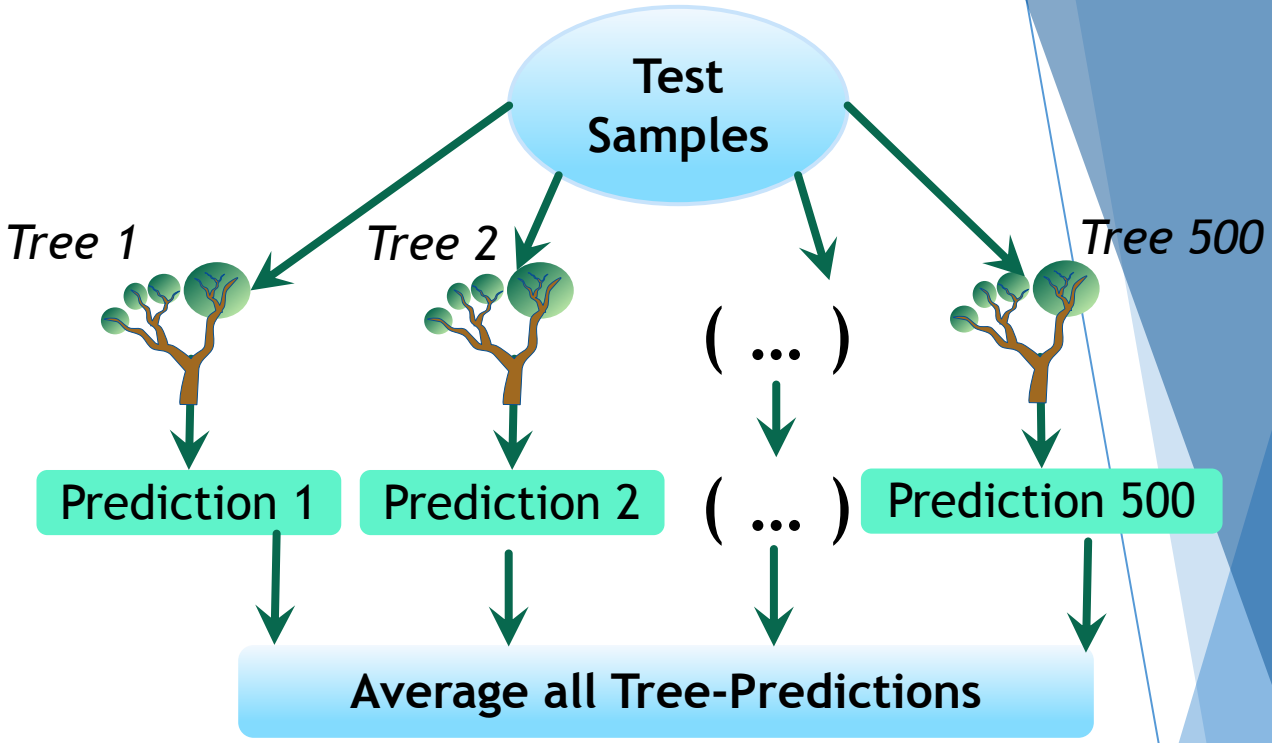
- **Primary Advantages:**
  - Desired properties: Low-variance predictions and avoids overfitting

Random Forest



- **Hyper-Parameters:**
  - Number of trees = **500** → Tradeoff between performance - complexity
- **Bootstrapping:**
  - Utilize a subset of samples for each tree (with replacement)
- **Random Subspace Method:**
  - Select randomly a portion of features for each tree

Random Forest



- Feed future days (test samples) into Random Forest model
- Each tree produces an independent prediction
- **Random Forest Final Prediction:** Average of all tree predictions

Exponential Smoothing

Forecast future values using a weighted average of all previous values in series

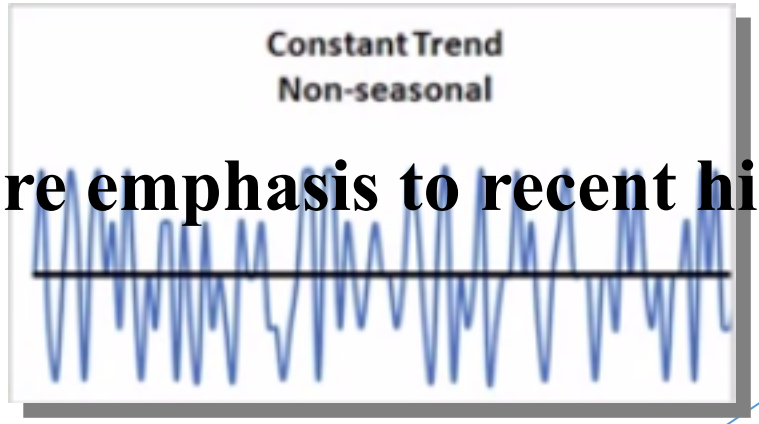
Advantages

- Popular
- Cheap to compute
- Easy to apply

Assumptions

- Adaptivity
- Robustness
- Continuity

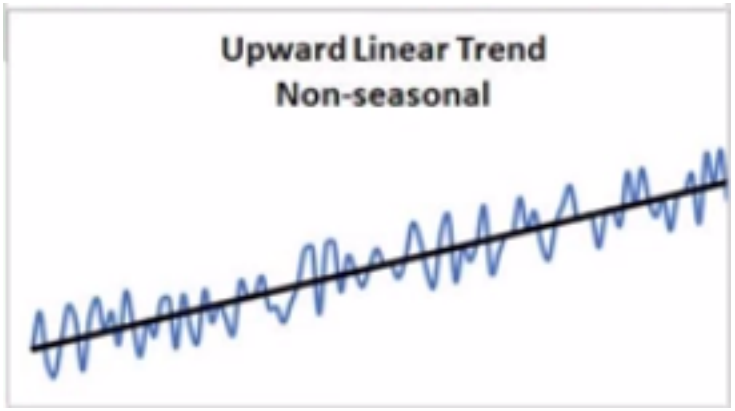
α



More emphasis to recent history

Holt's Exponential Smoothing

Different α,β for each country



steps in future

$$F_{t+k} = L_t + kT_t$$

recent trend

most recent estimated level

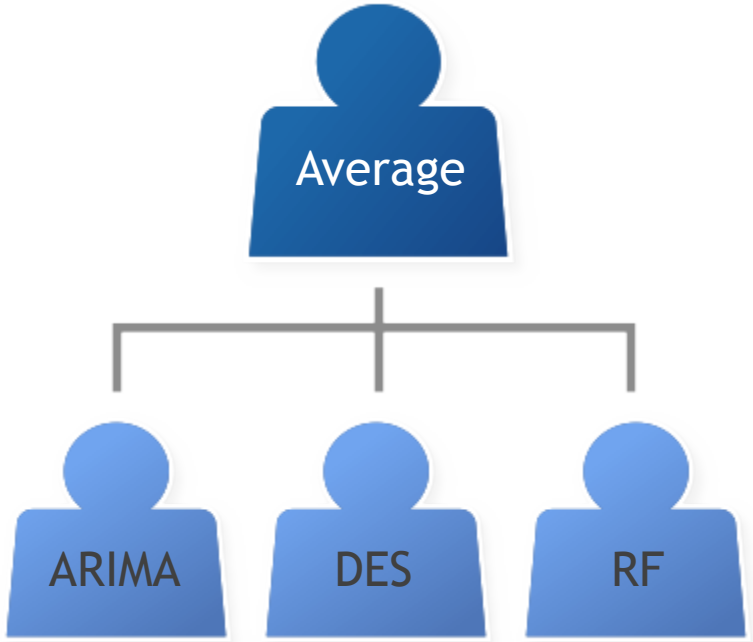
Update level

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$$

Update trend

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

Model Averaging



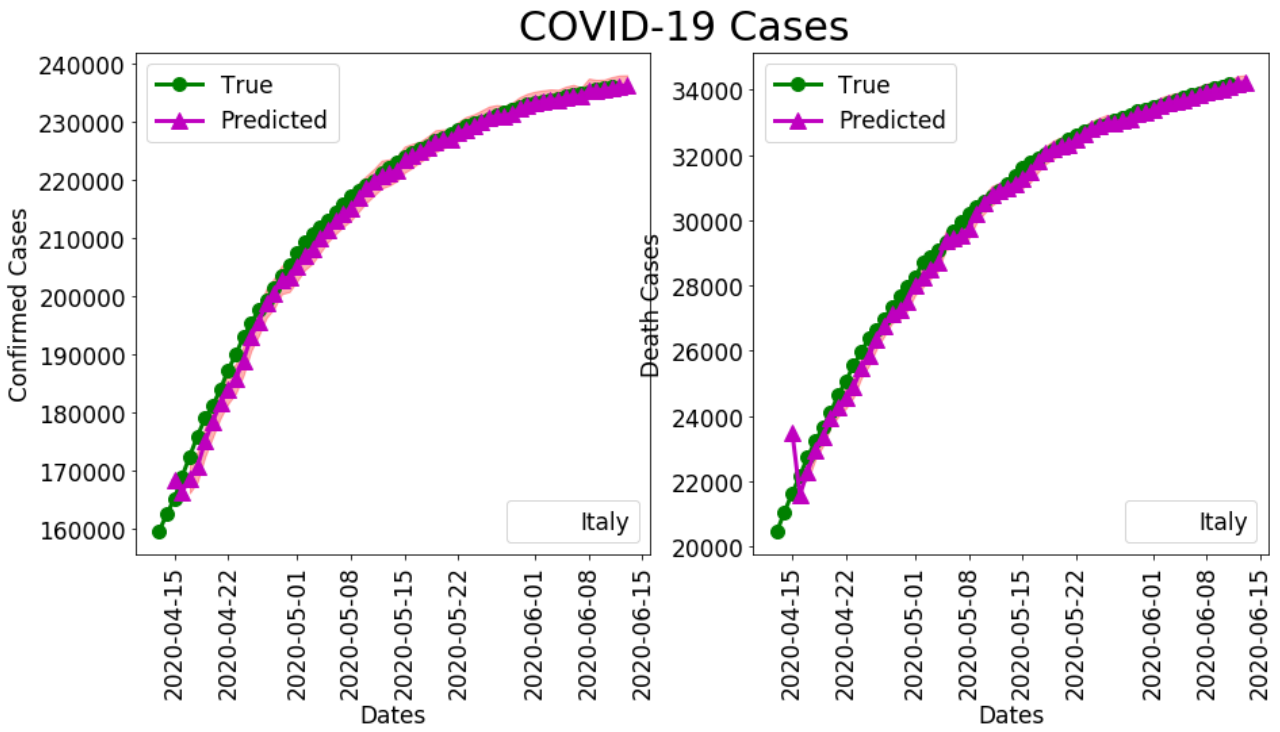
Not a single forecast

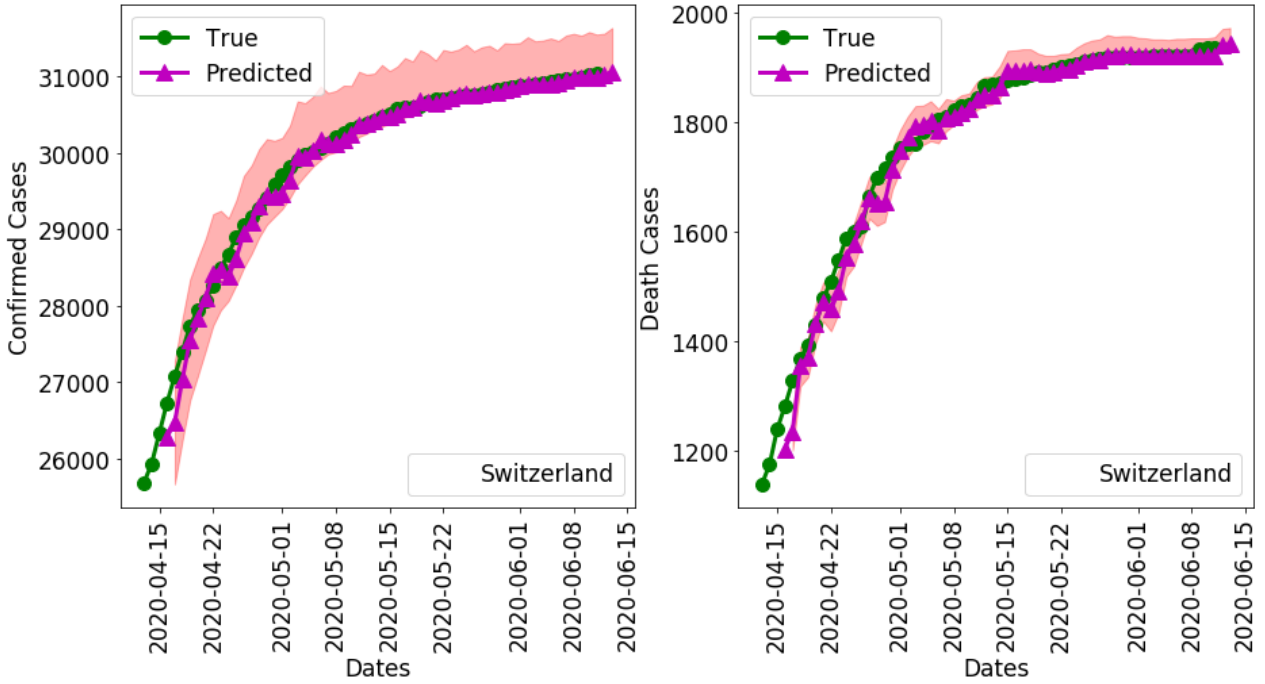
- Reduction of risk of choosing the wrong hypothesis
- Reduction of error variance of forecast

"The accuracy of the combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods."
Makridakis and Hibon (2000)

2-day Prediction

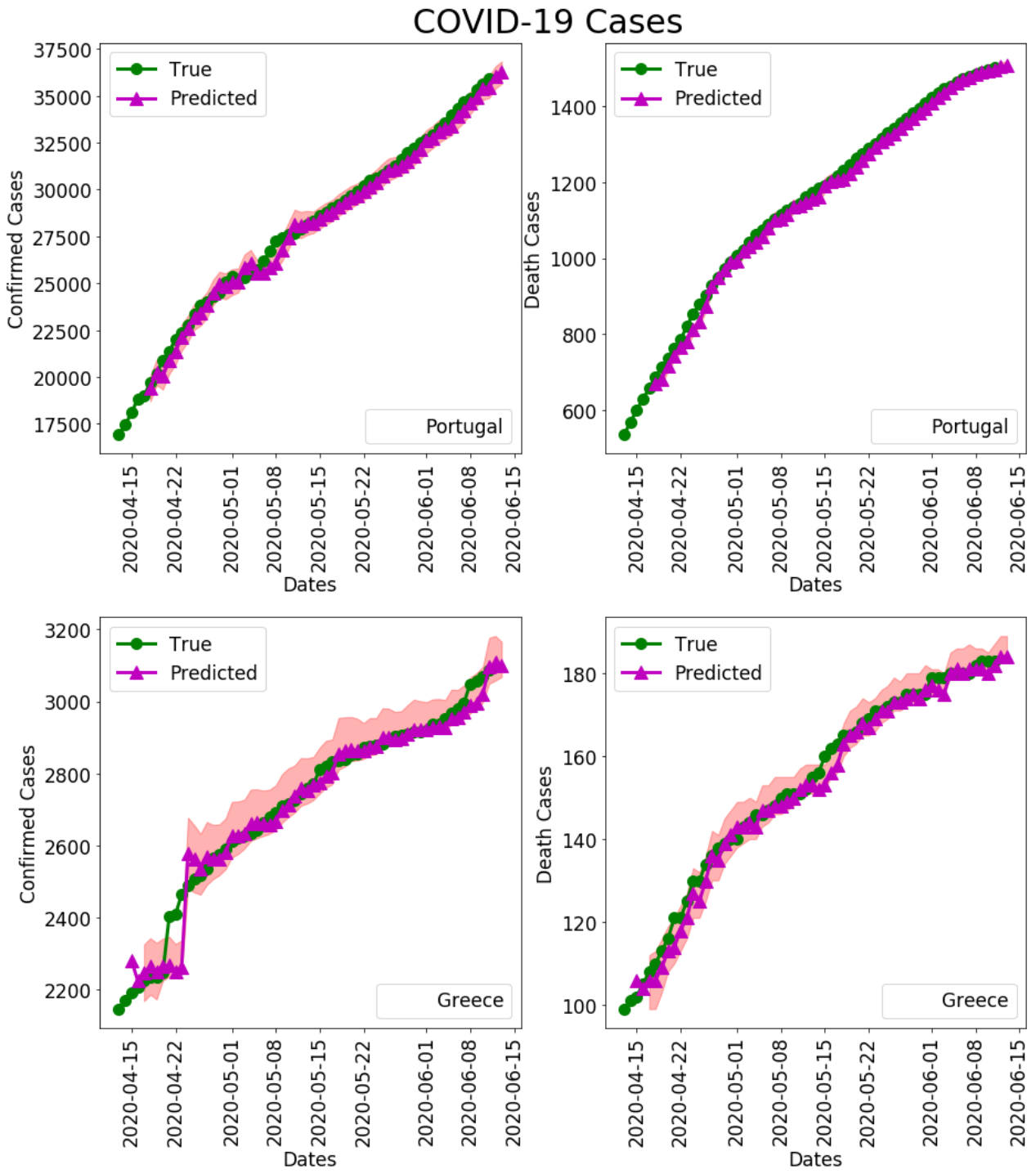
- Italy
- Switzerland





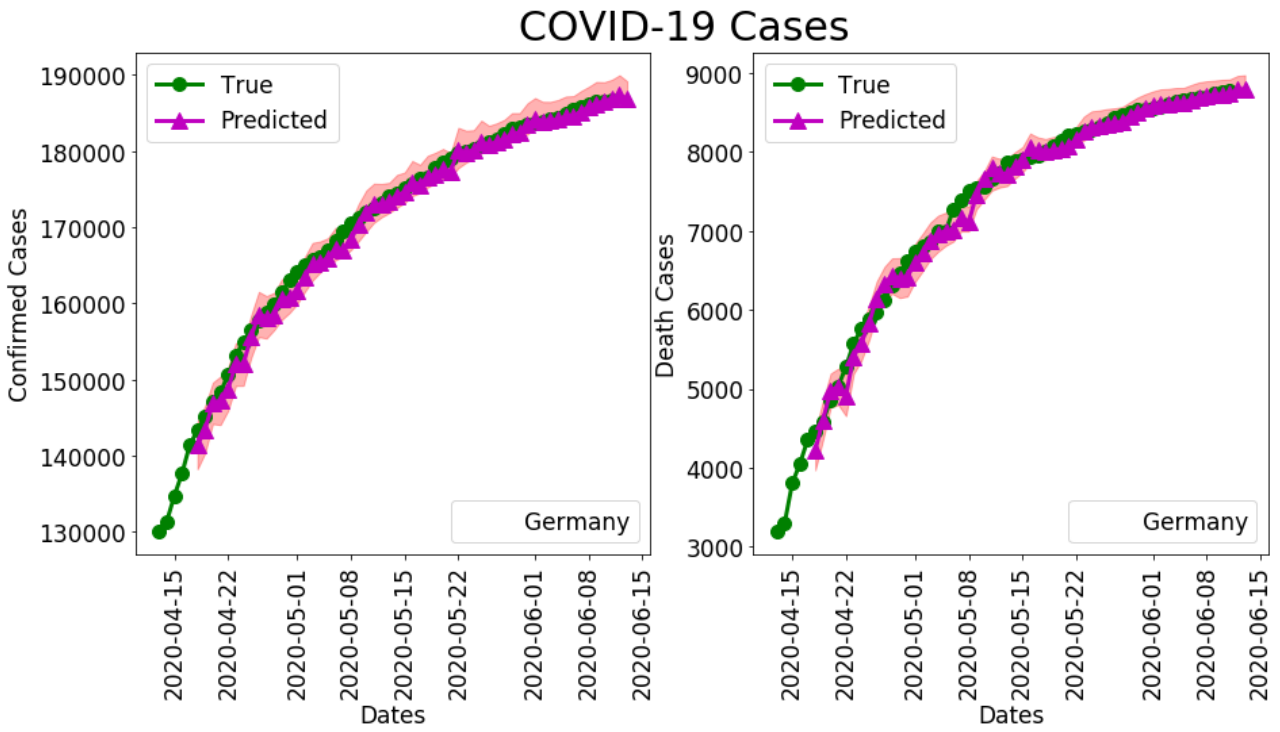
2-day Prediction

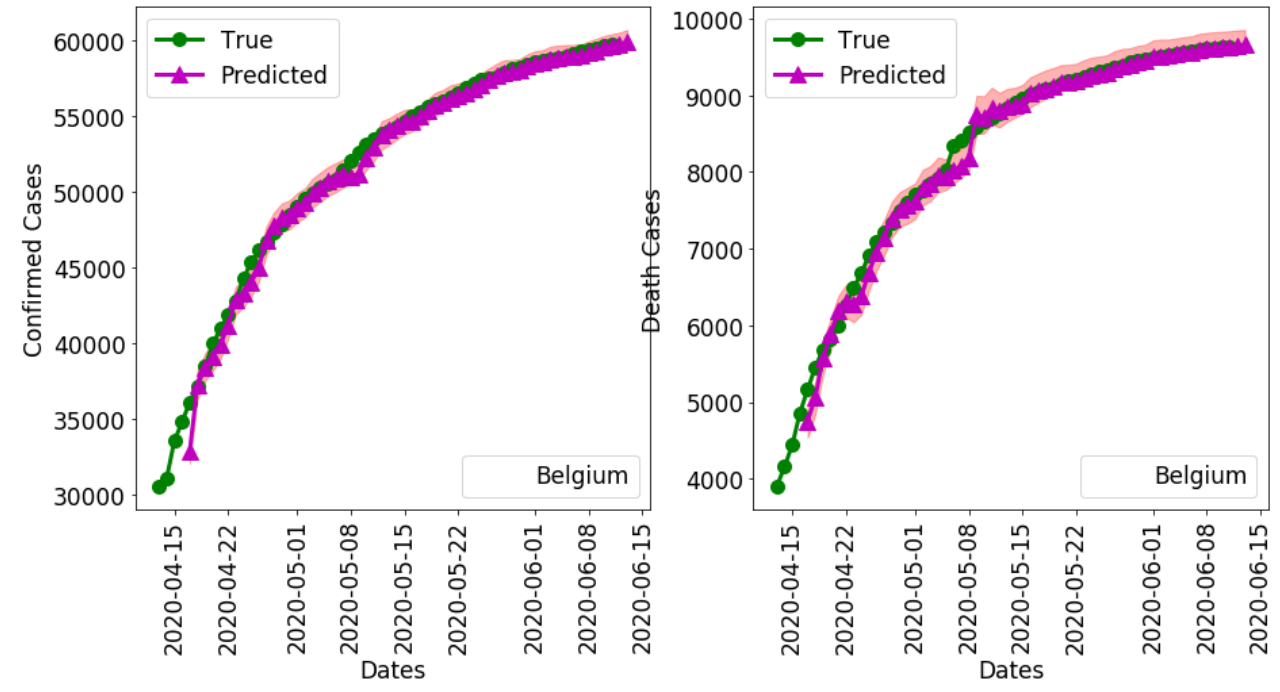
- Portugal
- Greece



2-day Prediction

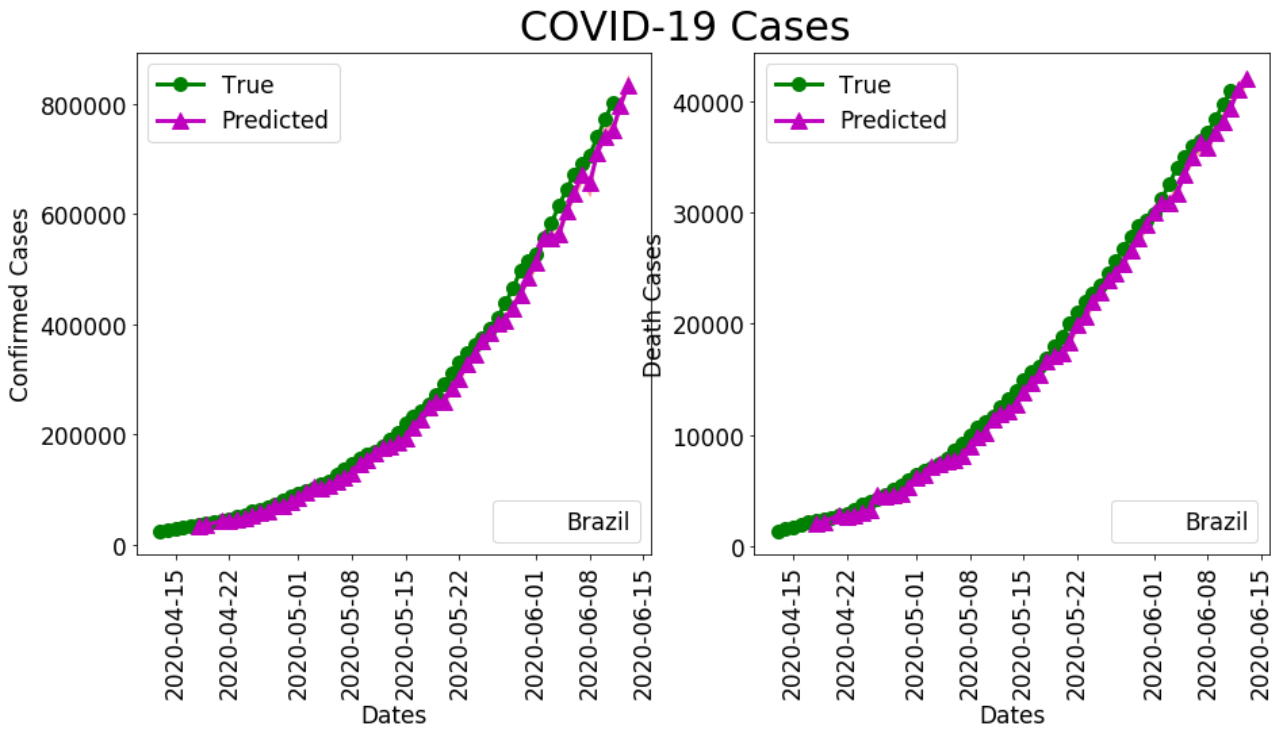
- Germany
- Belgium

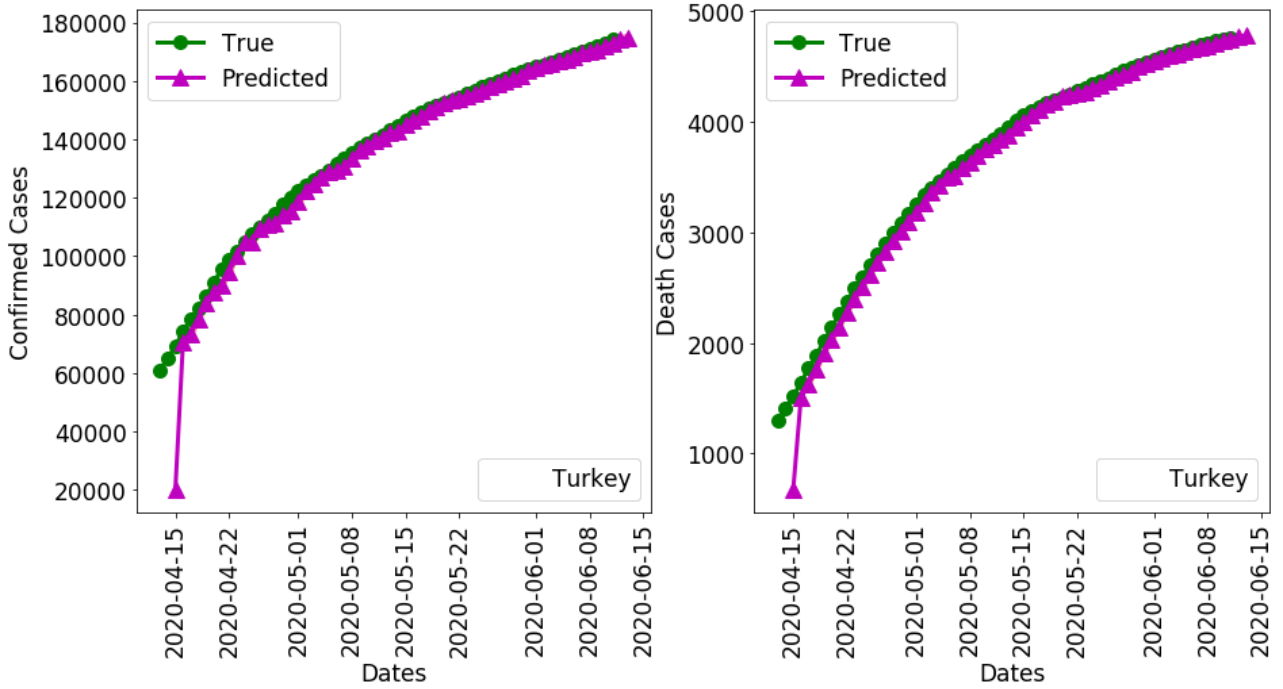




2-day Prediction

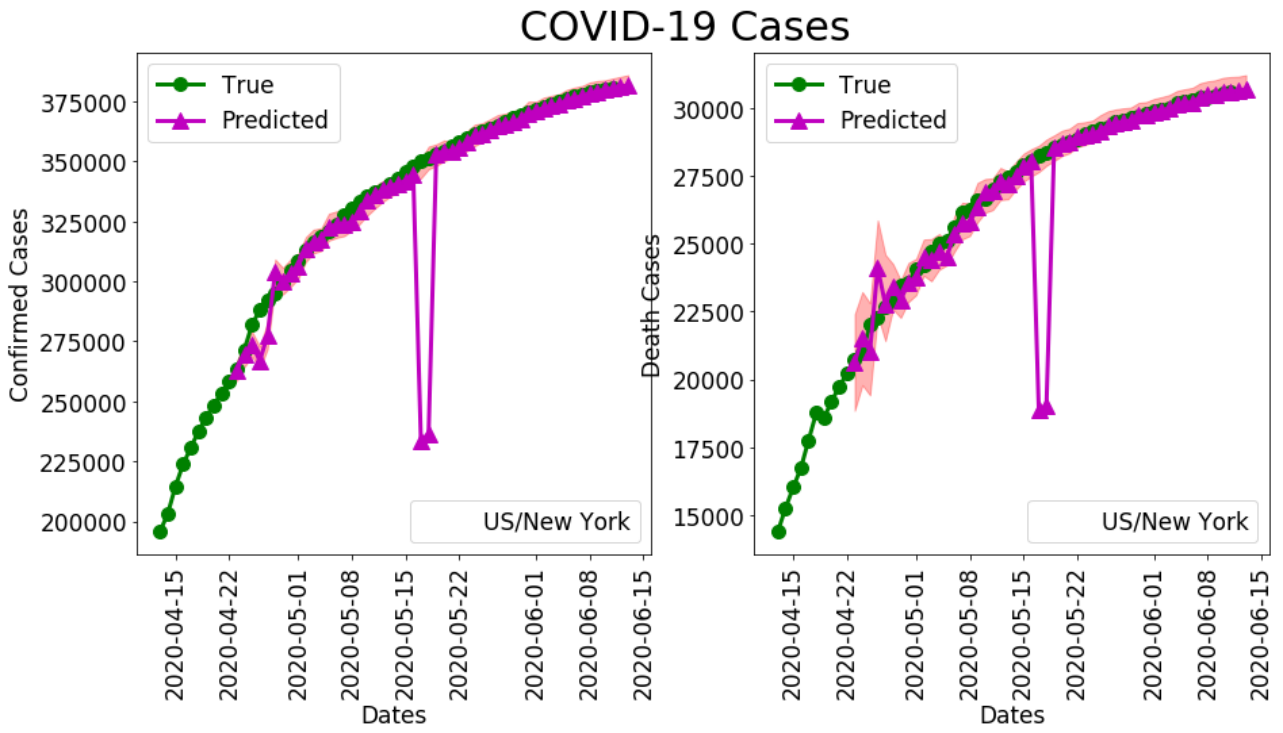
- Brazil
- Turkey

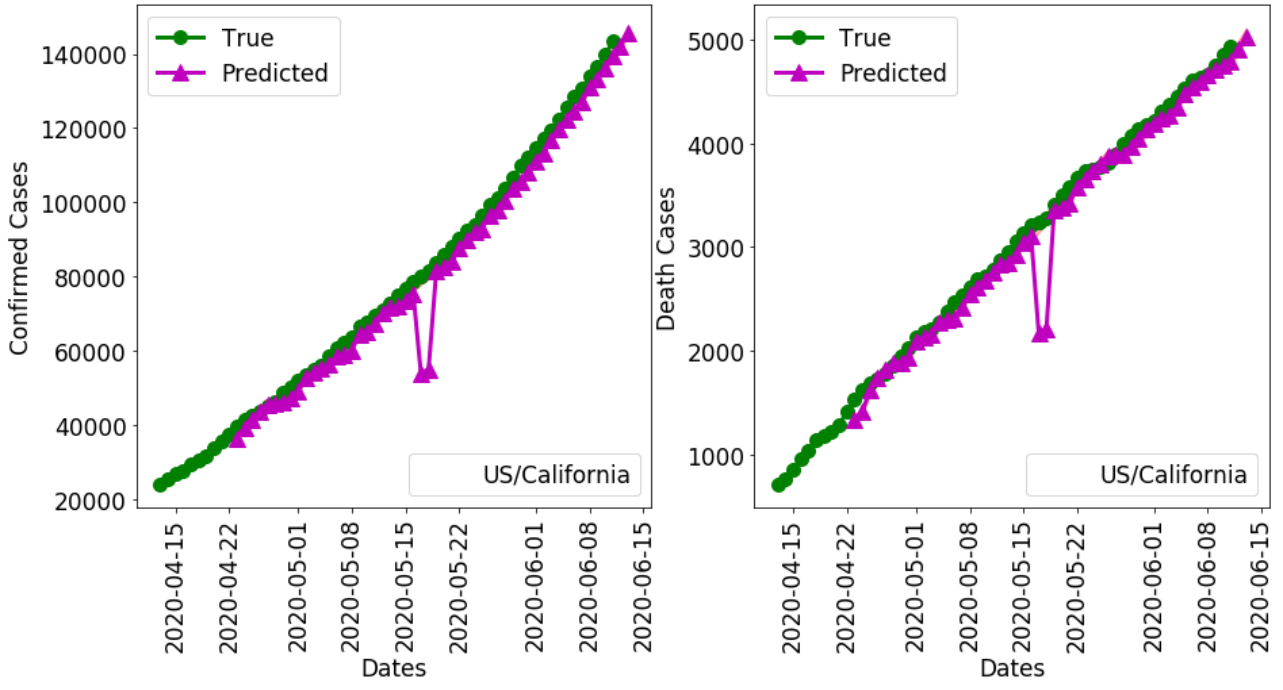




2-day Prediction

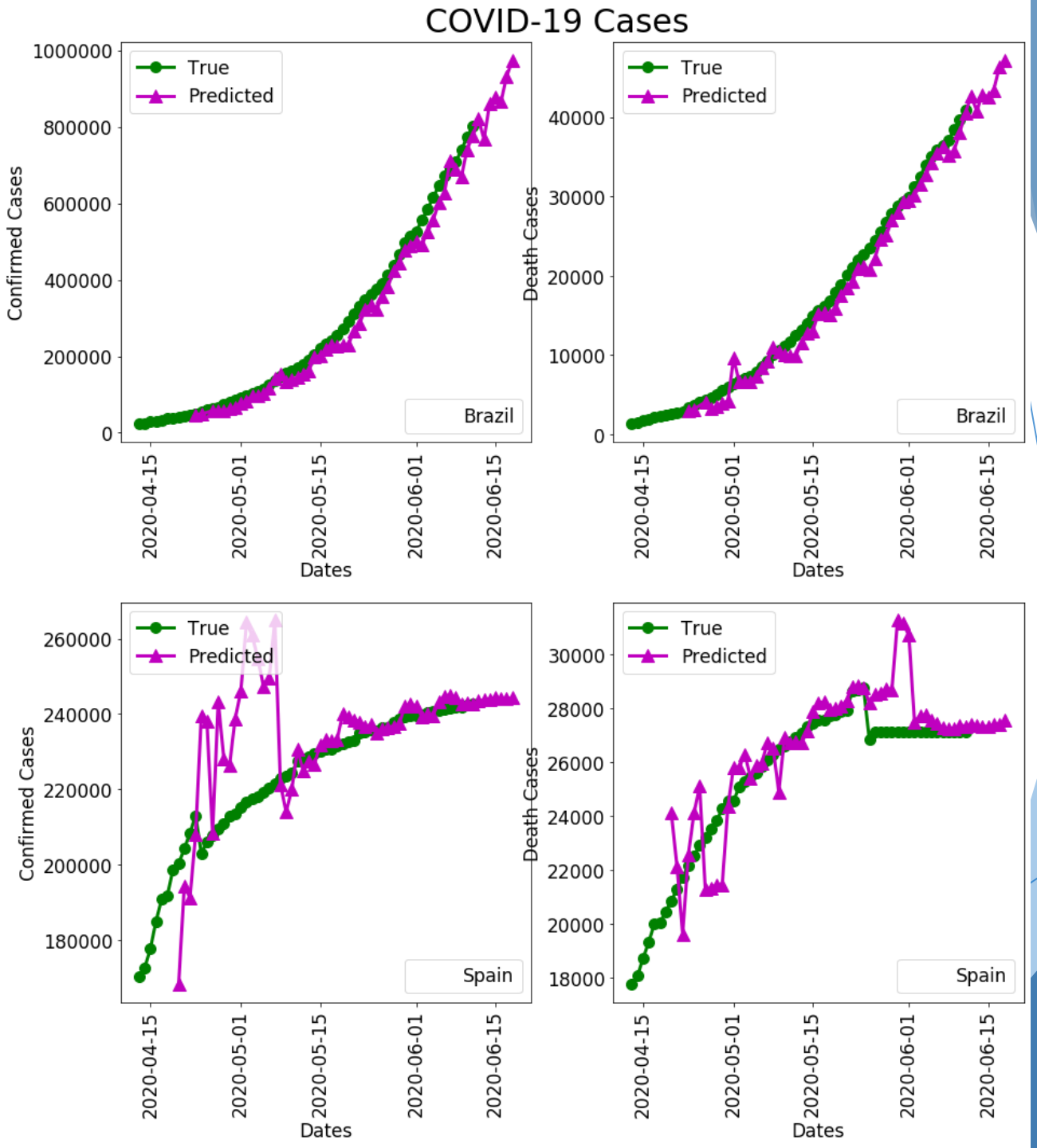
- New York
- California





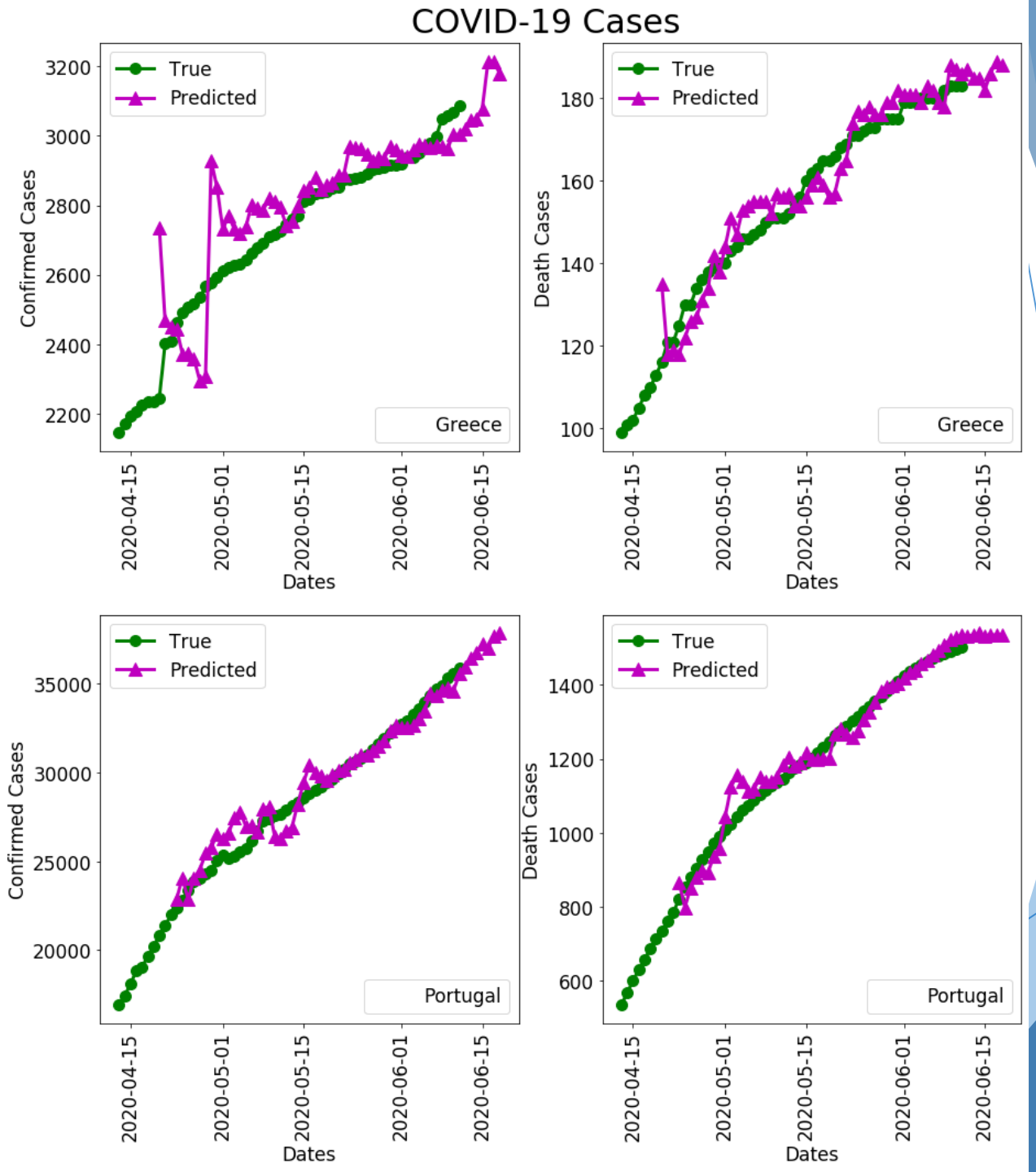
7-day Prediction

- Brazil
- Spain



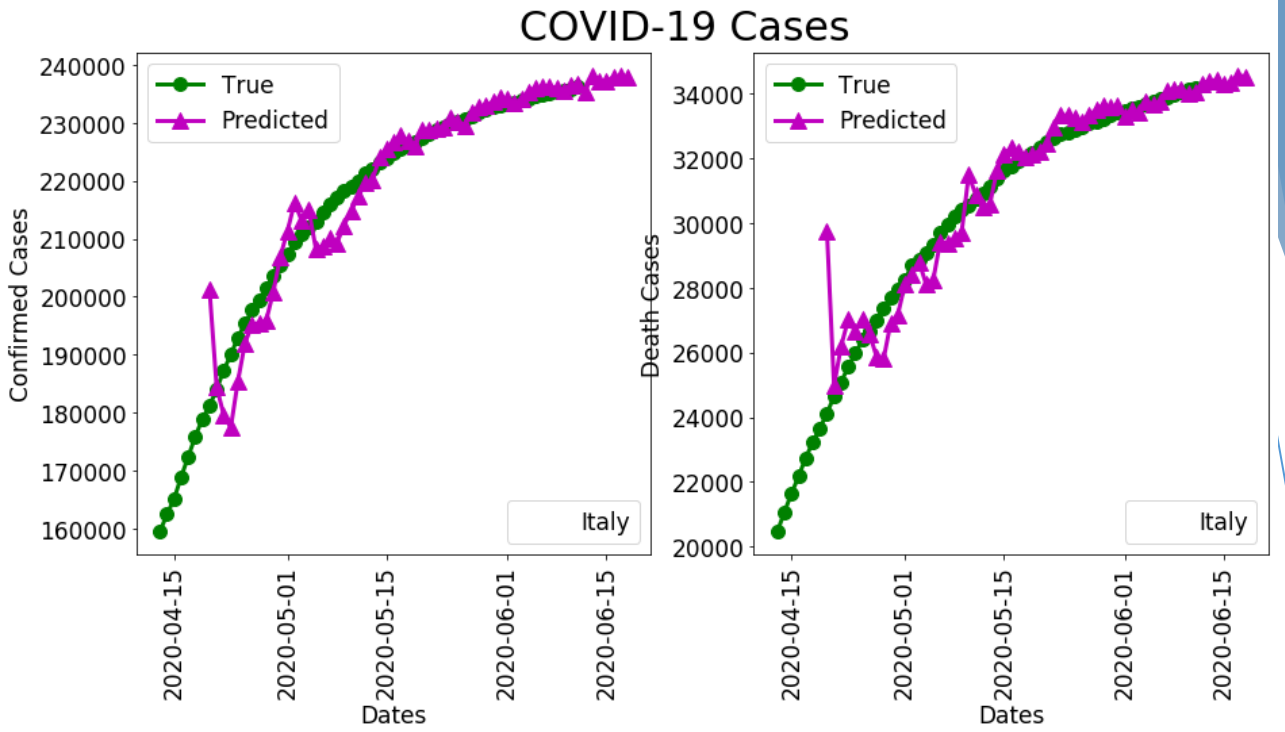
7-day Prediction

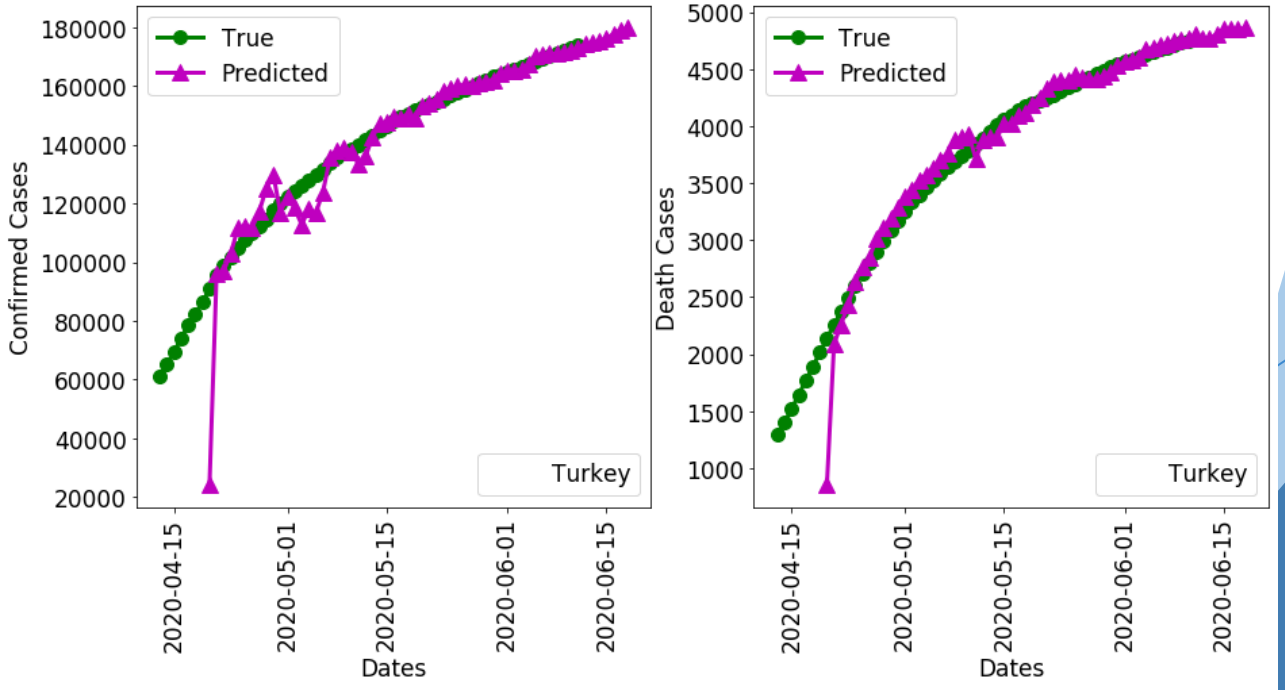
- Greece
- Portugal



7-day Prediction

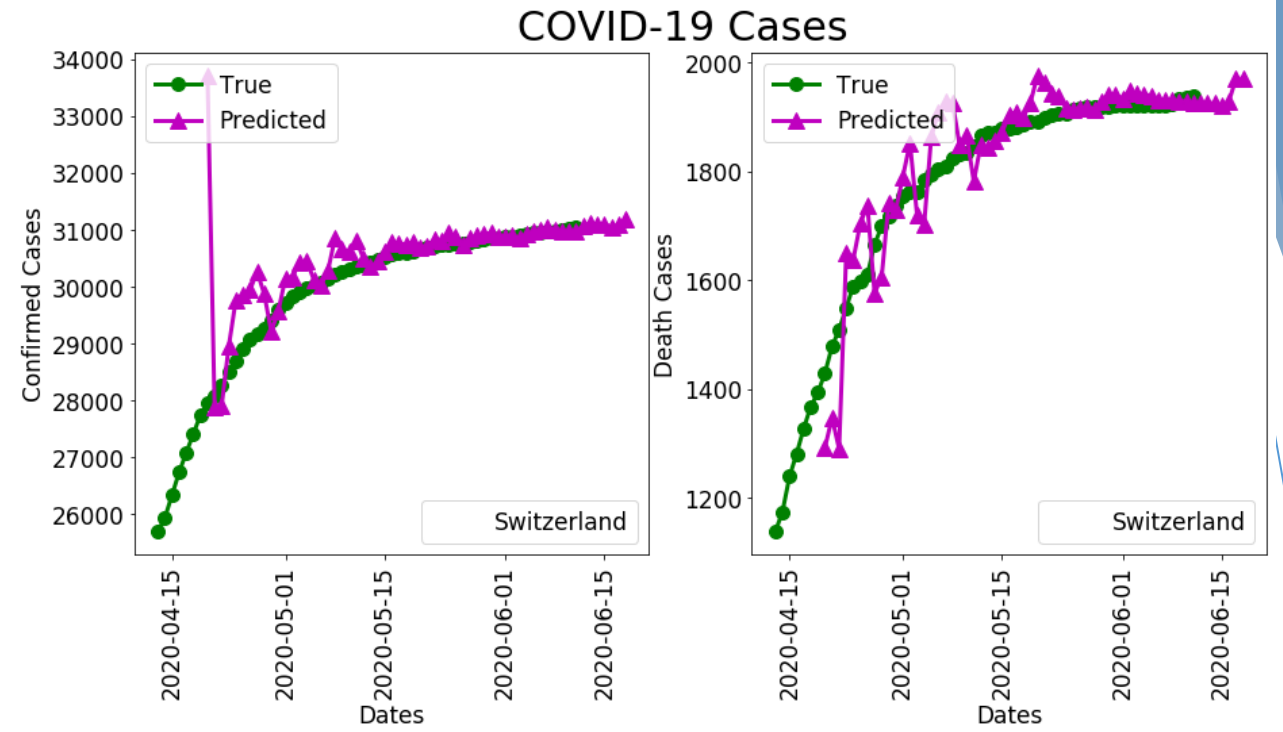
- Italy
- Turkey

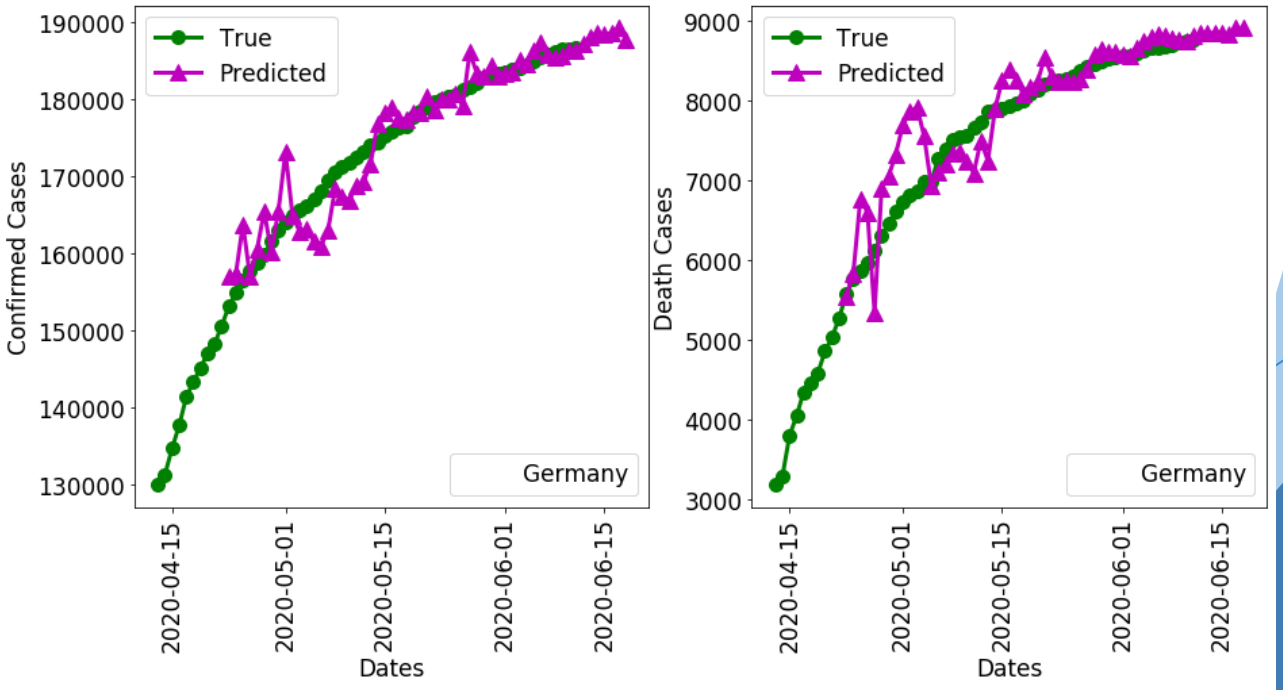




7-day Prediction

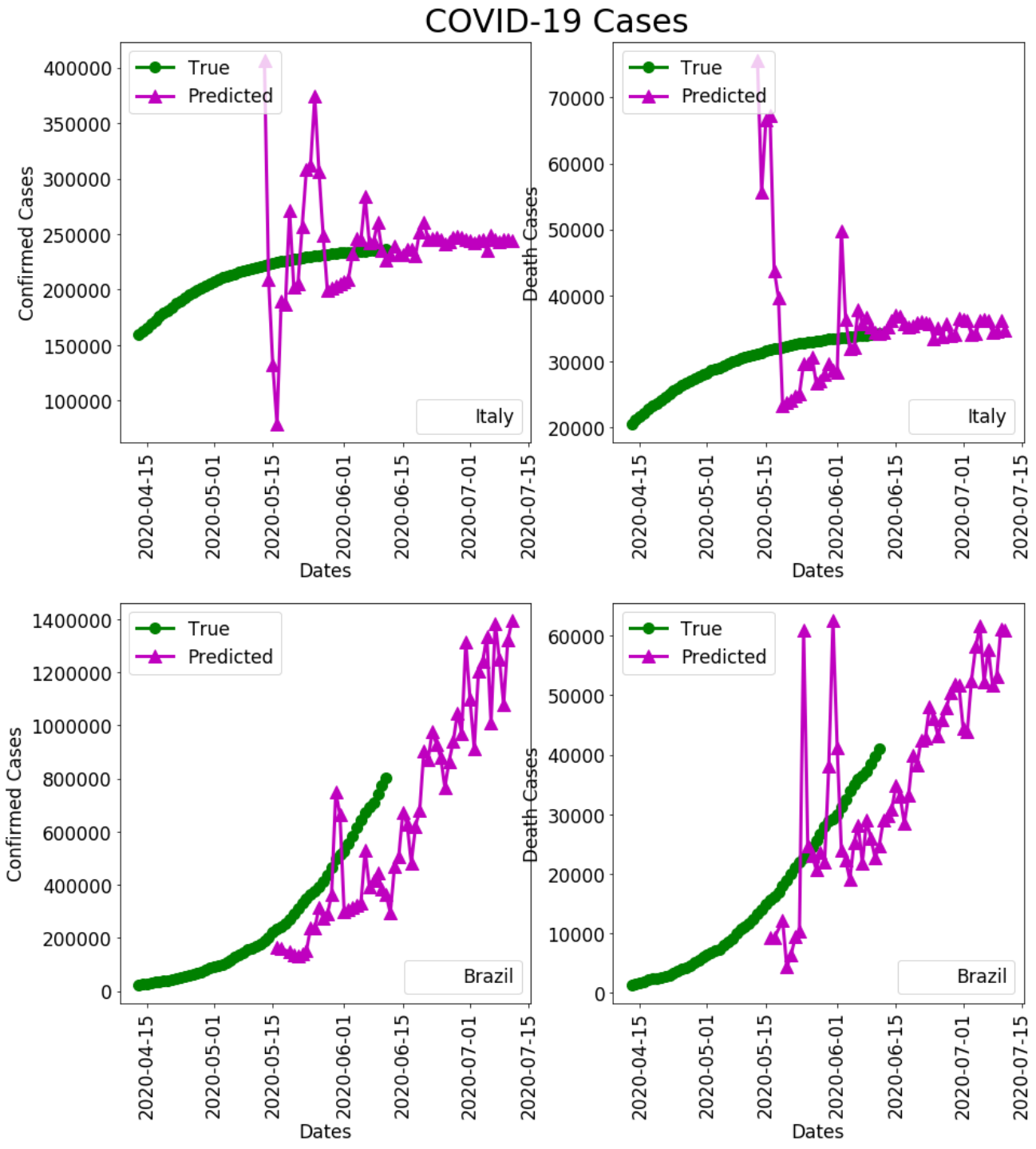
- Switzerland
- Germany





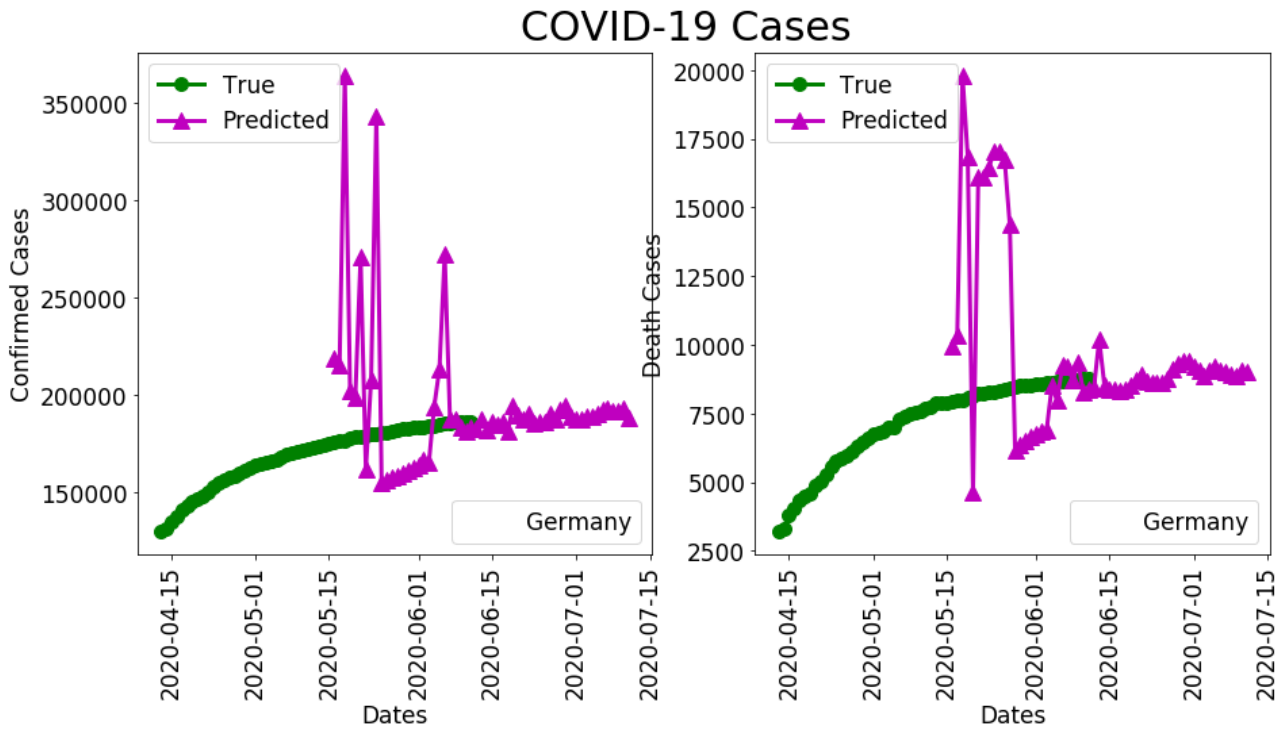
30-day Prediction

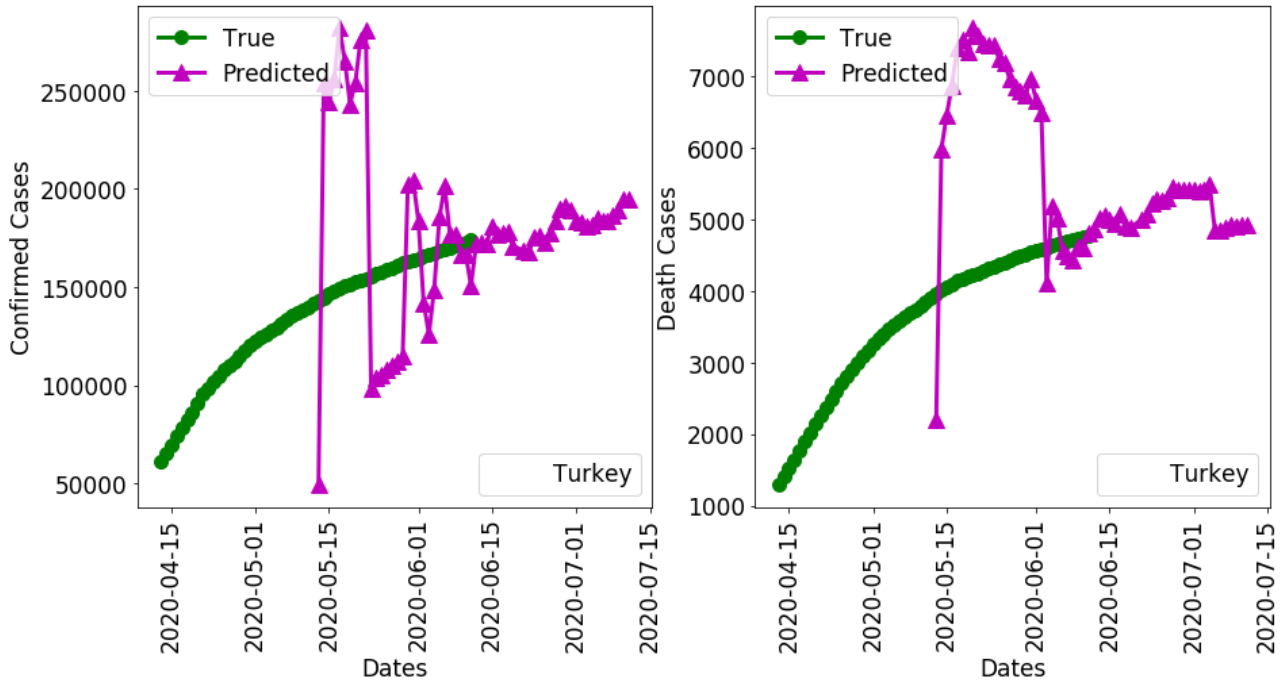
- Italy
- Brazil



30-day Prediction

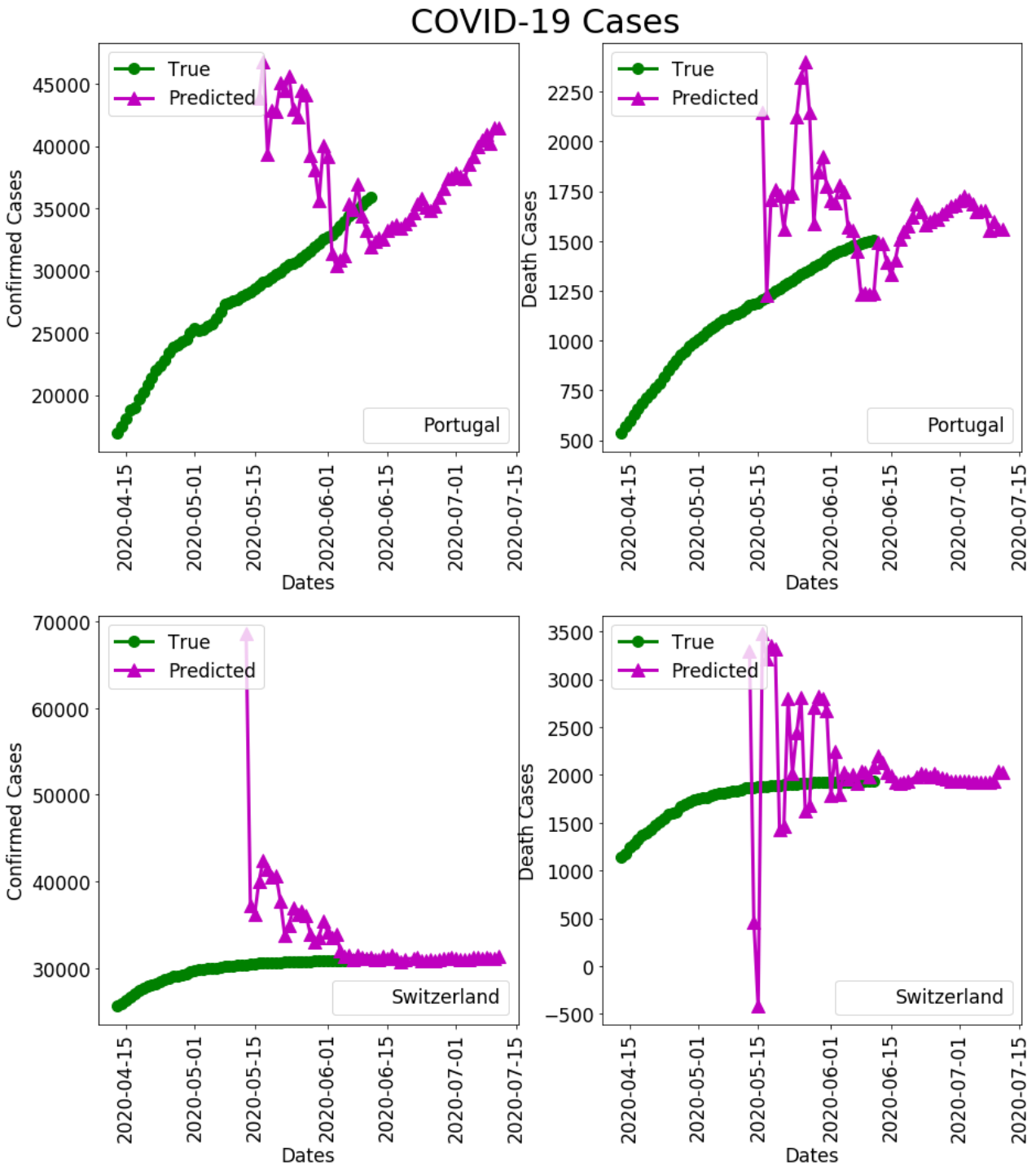
- Germany
- Turkey



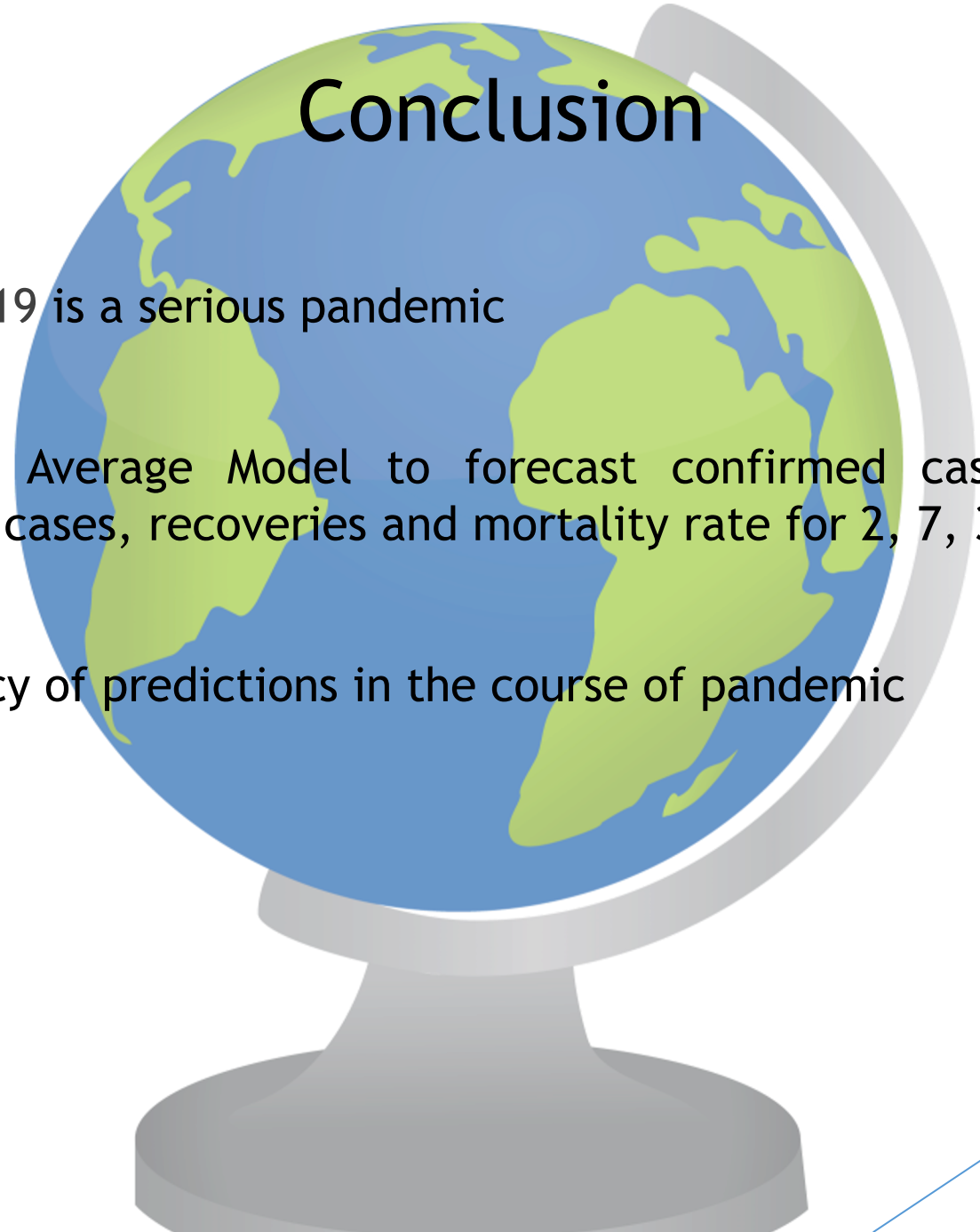


30-day Prediction

- Portugal
- Switzerland



Conclusion



- COVID-19 is a serious pandemic
- Use of Average Model to forecast confirmed cases, deaths, critical cases, recoveries and mortality rate for 2, 7, 30 days
- Accuracy of predictions in the course of pandemic

References

- E. Blair, M. Leonard and B. Elsheimer (2012) Combined Forecasts: What to Do When One Model Isn’t Good Enough. SAS Global Forum 2012 Paper 341–2012
- Benvenuto, D., et al. (2020). "Application of the ARIMA model on the COVID-2019 epidemic dataset." Data in Brief **29**: 105340.

Thank you very much for your attention!