Excess Mortality of COVID-19

Prediction Interval of Random Forest

Jinwoo Cho

University of Pittsburgh

Motivation

Definition of Excess Mortality

Excess Mortality

Excess mortality is a term used in epidemiology and public health that refers to the number of deaths from all causes during a crisis above and beyond what we would have expected to see under 'normal' conditions. [1]

- In terms of COVID-19, it is useful measure of the total impact of the pandemic on deaths than the confirmed COVID-19 death count alone.
- There are several methods including comparing average death for 5 years and that of this year (Or t test to see the significance.)
- But this method is too simple in that every year mortality has been increased because of ageing society.

GLM Method

EuroMOMO

EuroMOMO is a European mortality monitoring activity, aiming to detect and measure excess deaths related to seasonal influenza, pandemics and other public health threats.

- Mortality baseline is modelled using a glm poisson corrected for over dispersion.
- EUROMOMO used sine and cosine functions to adjust seasonality.
- Specifically, they estimate influenza attributed excess mortality using FLuMOMO, of which model is a multiplicative Poisson regression time-series model with overdispersion [4].
- FLuMOMO used weekly influenza activity (IA) and temperature data.

Limitation of GLM

- GLM assumed the distribution of data follows poisson distribution with overdispersion.
- Since GLM used sine and consine functions to reflect the seasonality of mortality, and this is quite a strong assumption.
- GLM ignored correlation between features such that influenza activity and extreme temperature.
- GLM ignored serial correlation structure of time series data.

Limitation of Classical methods

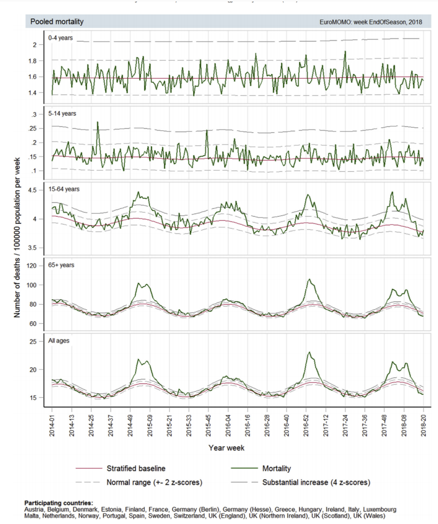


Figure 1: All-cause mortality pooled from 24 European countries based on the EuroMOMO algorithm [4]

Limitation of Classical methods

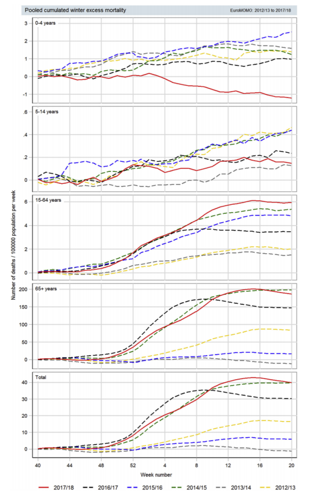


Figure 2: Cumulative excess mortality pooled from 24 European countries based on the EuroMOMO algoritthm [4]

Methodology

What I want to do

When it comes to estimating excess mortality of COVID-19....

- Reflect serial correlation structure of mortality data.
- Instead of considering correlation between features, only focus on mortality data. (There is no total mortality data of 2020, but monthly total mortality data in Korea).
- Use 95% prediction interval to estimate an excess mortality in 2020 (During COVID-19 period).

$\Longrightarrow$ Use three Random Forest method and compare with SARIMA (Times series modeling).

Prediction Interval of Random Forest

1. Quantile regression forest [3]
   - Using estimated conditional distribution, $\hat{H}_n(y|\boldsymbol{x})$, of response variable, Y, given features $\boldsymbol{X} = \boldsymbol{x}$, compute prediction $[\hat{Q}_{\alpha/2}(\boldsymbol{x}), \hat{Q}_{1-\alpha/2}(\boldsymbol{x})]$, where $\hat{Q}_{\alpha}(\boldsymbol{x}) \equiv \inf\{y \in \mathbb{R} : \hat{H}_n(y|\boldsymbol{x}) \geq \alpha\}$.
2. Split Conformal Prediction Intervals [2]
   1) Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{L}_1, \mathcal{L}_2$.
   2) Build a random forest from $\{(\boldsymbol{X}_i, Y_i) : i \in \mathcal{L}_1\}$ (a subset of the full training dataset $\boldsymbol{\mathcal{C}}_n$) to obtain an estimate of the man function $m(\cdot)$ denoted as $\hat{m}_{n/2}(\boldsymbol{X})$.
   3) For each $i \in \mathcal{L}_2$, compute the absolute residual $R_i = |Y_i - \hat{m}_{n/2}(\boldsymbol{X})|$. Let d be the kth smallest value in $\{R_i : i \in \mathcal{L}_2\}$, where $k = \lceil (n/2+1)(1-\alpha) \rceil$.
   4) The SC $100(a - \alpha)\%$ prediction interval for Y is $[\hat{m}_{n/2}(\boldsymbol{X}) - d, \hat{m}_{n/2}(\boldsymbol{X}) + d]$.

Shortage of two methods

Quantile Regression Forest

- Compute a quite wide interval, because of variable estimation in response distribution which depends on only local features.

Split Conformal Prediction Intervals

- The interval is calibrated for measuring the uncertainty of prediction errors from random forests constructed from $n/2$ rather than n observations.
- This fact derives slightly conservative performance.

Out-of-bag(OOB) Prediction Interval

OOB Prediction Intervals [5]

Suppose OOB prediction errors $\{D_i \equiv Y_i - \hat{Y}_{(i)}\}_{i=1}^{n}$, where $\hat{Y}_{(i)}$ is from ith OOB random forest. Then, as the number of samples, n and the number of bootstrapping, B grow large,

$$\begin{aligned} 1-\alpha &\approx \mathbb{P}\left[D_{[n,\alpha/2]} \leq D \leq D_{[n,1-\alpha/2]}\right] \\ &= \mathbb{P}\left[\hat{Y} + D_{[n,\alpha/2]} \leq Y \leq \hat{Y} + D_{[n,1-\alpha/2]}\right], \end{aligned}$$

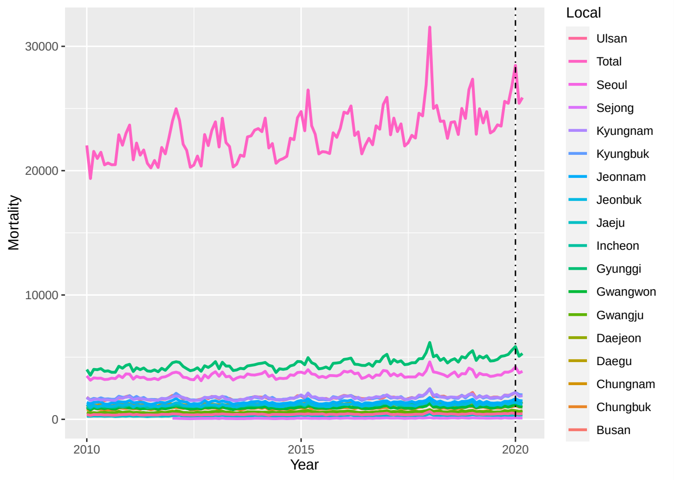
where $D_{[n,\gamma]}$ is the γ quantile of the empirical distribution of $D_1, \ldots, D_n$.

- When D is symmetric, the modified OOB prediction interval given by $\hat{Y} \pm |D|_{[n,\alpha]}$, where $|D|_{[n,\alpha]}$ is the $1-\alpha$ quantile of the empirical distribution of $|D_1|, \ldots, |D_n|$.

Result

Data Description

- Mortality data set is from Jan. 2000 to June. 2020 by month.
- There are mortality data for 17 local region in Korea.

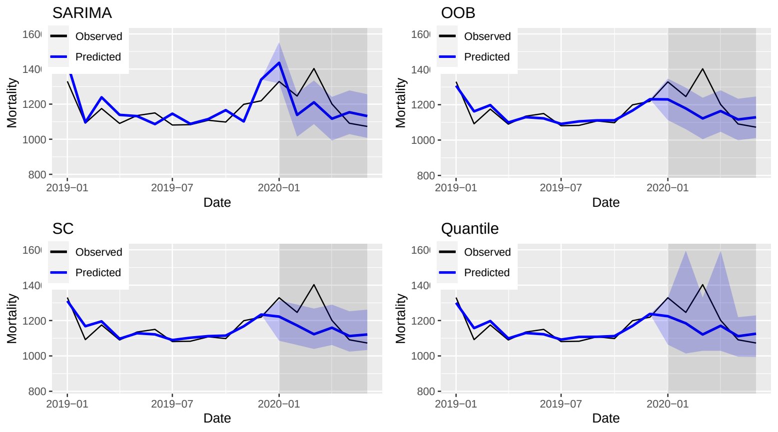


Model setting

- Divide training set (Before Jan/2020) and test set (After Jan/2020)
- SARIMA model was selected by comparing AIC.
- Random forest was conducted by `rfinterval` package in R.
- Fetures of random forest is following
    1. From 1 month lag to 5 months lag
    2. 12 months seasonal lag.

Result

- There is only one local which shows excess mortality (Mar/2020 Daegu).



Discussion

- The result is reasonable in that Daegu had the first regional pandemic in Korea on March. (70.4% of total COVID-19 deaths (114 deaths) were from Daegu until March.)
- Without any feature data, I can estimate excess mortality through prediction intervals.
- Random forest fit well in training data than SARIMA, but not in the test data. This result would be from the unexpected increasing mortality.
- Excess mortality would be more severe in aged group or other countries, so that it needs further study.

References i

F. Checchi and L. Roberts.
Interpreting and using mortality data in humanitarian emergencies.
Humanitarian Practice Network, 52, 2005.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman.
Distribution-free predictive inference for regression.
Journal of the American Statistical Association, 113(523):1094–1111, 2018.

N. Meinshausen.
Quantile regression forests.
Journal of Machine Learning Research, 7(Jun):983–999, 2006.

J. Nielsen, L. S. Vestergaard, L. Richter, D. Schmid, N. Bustos, T. Asikainen, R. Trebbien, G. Denissov, K. Innos, M. J. Virtanen, et al.
European all-cause excess and influenza-attributable mortality in the 2017/18 season: should the burden of influenza b be reconsidered?
Clinical microbiology and infection, 25(10):1266–1276, 2019.

H. Zhang, J. Zimmerman, D. Nettleton, and D. J. Nordman.
Random forest prediction intervals.
The American Statistician, pages 1–15, 2019.

Thanks!