COVID-19 EPIDEMIC DATA MODELING IN SPACE-TIME USING INNOVATION DIFFUSION KRIGING

Konstantin Krivoruchko & Jorge Mateu

June 2020

A Swedish geographer Torsten Hägerstrand is known for his work on migration, cultural diffusion and time geography. In [1], he suggested that a large number of natural, cultural, and economic processes can be described as “innovation diffusion”. Although his general model of spatial interaction was not successful when fitted to the data in comparison with geostatistical and lattice data approaches, it is qualitatively attractive when the data are complex and non-stationary. Spatial empirical Bayesian kriging (EBK) model (for continuous data) which is in some extent qualitatively similar to the Hägerstrand’s innovation diffusion was recently discussed in [2-4]. It consists of the following internal steps:

1. Parameters of the spatial process Θ, including the semivariogram model, are estimated from the data.
2. Using Θ, new values are unconditionally simulated at each of the input data locations Ksim times.
3. New parameters Θi, i = 1 . . . Ksim, are estimated from the simulated data. A histogram of Θi is an approximation of prior distribution.
4. Call Θi, i = 1 . . . Ksim, the empirical prior distribution. It is assumed that the model parameters can take only Θi values.
5. A weight for each simulated model is calculated using Bayes’ rule.
6. Predictions and prediction standard errors are produced at the specified locations using the formulas [2,4].

The default EBK model is “intrinsic random function (IRF)” process with power spatial correlation model. This correlation model is qualitatively similar to the innovation diffusion concept because it corresponds to fractional Brownian motion process. Simplistic illustration of IRFK can be found in [5]. When the dataset is large, the software creates subsets with a specified number of samples. A user defined data subset option is also provided. In that case, model fitting in the is performed for data subsets, and predictions are made using the weighted sum of the models from the possibly overlapping or disjoint nearby subsets. Numerous results of the models’ comparison show that EBK outperforms other predictors, and increasingly so with data complexity.

IRF with non-stationary semivariogram models described in [2-5] requires the following major additional options to be used with epidemic data:

- Generalization from space to space time.
- Count data transformation (assuming Binomial or negative binomial distribution of the original data) to Gaussian distribution.
- Use of covariates (explanatory data). This option is essential for better epidemic forecast (extrapolation in time direction).
- Conditional simulation from the mixture of the models

We have developed research version of the innovation diffusion kriging (IDK) software which combines ideas described in [2-4]. Spatial correlation in the IDF regression model for data y(s, t) is described by IRF kriging with power semivariogram model with time scaled to be consistent with metric space, similar to the approach discussed in [4, section 4.2.1]. The data transformation parameters Θ in p(s,t)=transformation(y(s, t)|Θ), where p(s, t) is probability of the individual to be sick, are estimated simultaneously with other model parameters similar to [4, section 3]. We use chordal distance metric, as justified in [3], which allows model fitting and predictions both for small and very large territories, including the entire Earth surface. We support measurement error methodology [6, 7], which allows both accurate predictions and correct model validation (comparing predictions with noisy observation data).

The output from the model is a realization from the binomial distribution Binomial(K|N, p(s,t)), where K is the number of new cases for the population of N, assuming the same risk factor p(s, t) for each individual in the data subset.

IDK can be used starting from time when sufficiently large number of infected people is available, in practice, two-three weeks after the first registered case of infected local people in the region. When testing IDK with covid-19 data available in the Internet, we use time steps from 15-20 to 50 days and about 50 locations in space in each data subset.

For each day, the output from the IDK is the prediction, prediction standard error, estimated number of new infected cases and its credible 90 percent interval, and *locally* estimated semivariogram model parameters.

The research IDK software is written in Python, while all pre- and post-processing steps are done in R. Typical IDK model fitting and prediction when the number of space-time points is 1000 or larger requires a few hours. Because of long computation time, we don't use empirical Bayesian methodology described in steps 2-4 above. Substantial software optimization which allows much faster calculations and Bayesian inference is possible, but it requires substantial development time.

References

[1] T. Hägerstrand, Innovationsforloppet ur Korologisk Synspunkt, Gleerup, Lund, 1953.

[2] Krivoruchko, K., Gribov, A., 2019. Evaluation of empirical Bayesian kriging. Spatial Stat. 32, 100368. https://doi.org/10.1016/j.spasta.2019.100368.

[3] Krivoruchko, K., Gribov, A., 2020. Distance metrics for data interpolation over large areas on Earth’ s surface. Spatial Stat. 35, 100396. https://doi.org/10.1016/j.spasta.2019.100396.

[4] Gribov, A. and Krivoruchko, K., 2020. Empirical Bayesian kriging implementation and usage. Science of the Total Environment Volume 722, 20 June 2020, 137290. https://doi.org/10.1016/j.scitotenv.2020.137290.

[5] Krivoruchko, K., Fraczek, W., 2015. Interpolation of data collected along lines. https://apl.maps.arcgis.com/apps/MapJournal/index.html?appid=e7bd9a788b584f21ae738363b9b55d41.

[6] Krivoruchko K. (2011) Spatial Statistical Data Analysis for GIS Users. ESRI Press, 928 pp. Freely available at https://community.esri.com/thread/201550-spatial-statistical-data-analysis-for-gis-users-available-free-for-download since September 11, 2017.

[7] Krivoruchko, K., Gribov, A., and Ver Hoef, J.M. (2006) A new method for handling the nugget effect in kriging. In T.C. Coburn, J.M. Yarus, and R.L. Chambers, eds., Stochastic modeling and geostatistics: Principles, methods, and Case Studies, Volume II: AAPG Computer Applications in Geology 5, p. 81 – 89.