ARTICLE IN PRESS

Big data sampling and spatial analysis: "which of the two ladles, of fig-wood or gold, is appropriate to the soup and the pot?"

Roger Bivand [a,*], Konstantin Krivoruchko [b]

[a] *Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway*
[b] *Environmental Systems Research Institute, 380 New York Street, Redlands, CA 92373, United States*

A B S T R A C T

Following from Krivoruchko and Bivand (2009), we consider some general points related to challenges to the usefulness of big data in spatial statistical applications when data collection is compromised or one or more model assumptions are violated. We look further at the desirability of comparison of new methods intended to handle large spatial and spatio-temporal datasets.

© 2018 Published by Elsevier B.V.

1. Changes in data collection

Although in some application areas, data continue to be collected by carefully planned and costed surveys (for example mining and petroleum), supplemented by secondary use of remote sensing and ancillary information, in others the data are often harvested rather than acquired. Costs associated with the use of such data are related to its capture and storage, rather than its planned collection. This suggests that there is a major difference in data acquisition strategies between key sectors of the economy and government for whom geographical position matters, and those who hold user-generated data and try to retro-fit geographical position to this information.

For example, collecting a series of dynamic measurements taken from a car, boat, or other means of transportation using a GPS-enabled sensor has become a common practice (Krivoruchko and Frączek, 2016). A strong effort was made to estimate the magnitude of the radioactively polluted air over the urban area of Fukushima six months after the accident at the NPP. Radiation meters installed on 16 survey cars recorded the intensity of radiation between September 13th and 29th, 2011. Overall, more than 112,000 recordings of radiation intensity were taken. The main effort of those who were sampling radiation was to cover the largest possible length of transects. Some of the streets were driven twice, as sometimes the car had to drive in different directions along the same street. On some wide streets, possibly with separated directional lanes, the measured values of radiation varied significantly. The difference in radiation on the two sides of the street can be explained not by the spatial distance, but by temporal separation between the times at which the observations were collected. Such a sampling design is problematic even for estimating pollution on the road where the air is changing every minute due to the movement of the vehicle (Krivoruchko and Frączek, 2016). The post-Fukushima data collection from cars resulted in much more data than it was available post-Chernobyl. However, in the latter case much attention was given to sample design and careful data collection, based on understanding the physical and chemical processes. We feel strongly that careful design of data collection is much more important for modeling and handling uncertainty than the rapid generation of large volumes of data.

---

* Corresponding author.
    E-mail addresses: roger.bivand@nhh.no (R. Bivand), kkrivoruchko@esri.com (K. Krivoruchko).

Among spatial data collection issues, spatial and temporal support are central, with ensuing change of support problems when data from different sources need to be integrated (Gotway and Young, 2002; Krivoruchko and Gotway Crawford, 2005; Gelfand, 2010). Naturally, if the observations are collected by design, these problems can be mitigated at least in some measure, so the design of spatial samples should be considered carefully, especially in connection with the prior understanding of the spatial "reach" of underlying processes (Wang et al., 2012). Contemporary applications of "big data" often have no secure positional information, and proxy by informed guesses. For example, the overwhelming majority of tweets do not provide GPS-based point positions. Having millions of observations with large unknown and varying measurement/locational error is not better than having hundreds of poor measurements.

One useful option for dealing with the uncertainty of individual observations is allowing for the specification of the individual measurement errors in the format of one standard deviation, or weights so that data with higher weights will have more impact on the interpolation model (as provided, for example, in empirical Bayesian kriging and local polynomial interpolation in Geostatistical Analyst). Locational error can be treated as measurement error. Although the error in the data coordinates may change the covariance model significantly (see for example Fig. 1 in Cressie and Kornak, 2003 p. 443), the suggested or similar models are not used in practice. However, the locational error has the same effect on the kriging model as the measurement error: increasing the nugget effect parameter.

Positional error is also discussed by Fanshawe and Diggle (2011) and Chakraborty and Gelfand (2010), but without the concerns raised there being followed up in software implementations. Another article from the same year, Diggle et al. (2010), points to the additional serious issue of inference when the data come from preferential samples that violate the basic assumptions of the methods used. A recent article using these results questions the usefulness of using big data acquired for precision farming for soil mapping; farmers will collect more samples at locations of greater interest for them (Rawlins et al., 2017).

What is the threshold for big data? For statisticians, it is simply when there are numerical problems in linear algebra on large dense matrices. But for GIS user, it is when there is a problem with data storage and data querying. When spatial statisticians report advances in solving big data problems, the numbers involved seldom exceed around one hundred thousand points, so there is still plenty to do, not least because big datasets appear to force the use of simple models with unrealistic assumptions.

2. Some problems with modeling big spatial data

Although big spatial data are collected in the ionosphere, atmosphere, ocean, on the land and subsurface, only a small number of organizations are using recently developed statistical methodologies for the big data analysis. One example is data permanently collected by NOAA (at the moment there are about 50 million multivariate observations). Each "measurement" typically consists of the values of temperature, salinity, oxygen, nitrates, phosphates and silicates at the particular depth collected at a particular time (NOAA, 2013). Note that the number of observations collected at different locations varies dramatically making data interpolation challenging, see Fig. 1. It seems that for many large organizations, the time of big data arrived totally unexpectedly. NOAA is still using old trusted functionality, interpolating the ocean data using three nearest points (Reiniger and Ross, 1968). The World Ocean Atlas documentation states:

> If data coverage allows, we use the 4-point Reiniger–Ross interpolation method directly, no changes to their algorithm. . . . We also use three point Lagrangian or, as a last resort, linear interpolation, when there are not enough valid points for Reiniger–Ross (https://www.nodc.noaa.gov/OC5/wod-woa-faqs.html).

We think that it is a problem that such an indefensible approach is used, but perhaps they are apprehensive about having to re-estimate everything backwards if they tried better alternatives? Many organizations seem to be stuck in maintaining backward compatibility, but should move to sounder methods that take modern computational capacity into account. In particular, Lagrangian interpolation (Reiniger and Ross, 1968) has too many shortcomings to be used today.

We illustrate another common problem with real data interpolation using average annual rainfall values in South Africa. 8397 measurements are available for the territory covered by mountains, desert, and jungles. The data are provided by Krivoruchko (2011).

Fig. 2 (top left) shows the data locations and their estimated density and (top right) the annual rainfall standard deviation in areas with approximately 200 observations. Note that standard deviation values in areas colored in red is about eight times larger than the values in green polygons. We expect that the estimated prediction uncertainty should be larger in the areas with larger data variation. When the data variation is about the same, the prediction error should be larger in the areas with lower observation density.

Fig. 2 (bottom) shows the annual precipitation prediction standard errors produced by the stochastic partial differential equation (SPDE: left Lindgren et al., 2011) and empirical Bayesian kriging (EBK: right Krivoruchko and Gribov, 2014) models. There is a big difference in the spatial structures of the prediction errors: SPDE largely reflects the density of the data locations, while the EBK prediction standard errors pattern is qualitatively similar to the precipitation data variance map. Additionally, the EBK prediction error is a function of the data density.

Note that in the meteorological literature, precipitation values are typically modeled using the gamma distribution and the mean and variance of that distribution are related. Therefore, the precipitation prediction standard error should be a function of the prediction. Accordingly, large prediction standard errors in the northwestern part of South Africa near the