Statistics and Probability Letters 136 (2018) 87–91

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

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

ARTICLE INFO

Article history:
Available online 16 February 2018

Keywords:
Change of support
Sampling design
Data transformation
Prediction standard error

ABSTRACT

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.

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).

https://doi.org/10.1016/j.spl.2018.02.012
0167-7152/© 2018 Elsevier B.V. All rights reserved.

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

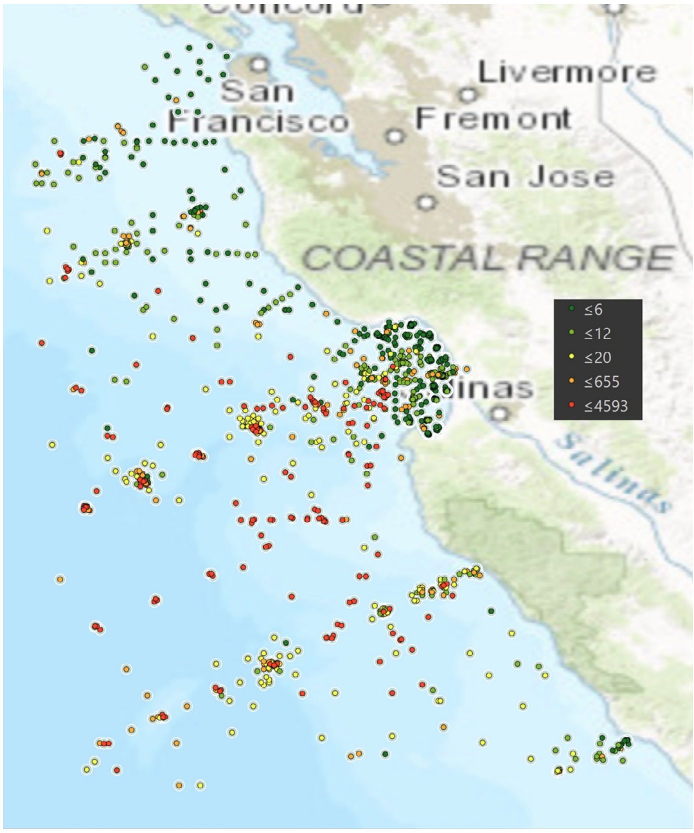

Fig. 1. The number of dissolved oxygen observations, taken from a small subset of the global spatio-temporal data in the World Ocean Atlas 2013 version 2 (NOAA, 2013).

Kalahari Desert with low annual rain and small prediction standard errors in the rainy tropical part of the country do not appear realistic. It should be noted that prediction standard error maps created by fixed rank kriging (Cressie and Johannesson, 2008), Gaussian predictive process (Banerjee et al., 2008) and lattice kriging (Nychka et al., 2015) models are very similar to the SPDE map; we think that the only difference is in the degree of smoothness. This is a serious problem if statistics is really a science of uncertainty because large and complex data variation is never the same in space and space–time.

Although interpolating up to 100,000 points is sufficient for many applications, there are always larger datasets. Then one obvious solution to the problem is divide and conquer algorithm. One successful implementation of this approach is EBK, see the case study in Krivoruchko and Gribov (2014) with the interpolation of 1.4 billion points of Lidar data.

Another typical problem is the Gaussian process assumption. Real data are never Gaussian and the usage of Gaussian models introduces unknown error to the predictions and especially prediction standard errors. One solution to the problem is using flexible data transformation globally (Gribov and Krivoruchko, 2012) and locally (Krivoruchko and Gribov, 2014).

Back to the support problem. Suppose that the end user of statistical software collects the data taking into account the change of support problem. This end user then meets the barrier of the absence of reliable software for big data interpolation and spatial regression handling the change of support problem. Available implementations are fitting one covariance model to the entire dataset (i.e. Krivoruchko et al., 2011) and, as shown by Oliveira (2013, 2014), all currently used models for Poisson data effectively describe only a narrow class of spatial correlation.

As a final example, one useful approximation for big data modeling over large areas is the usage of chordal distances instead of more complex approaches based on arcs and kernel convolution on sphere or ellipsoid (i.e. Krivoruchko and Gribov, 2014; Gribov and Krivoruchko, 2017). There are many GIS applications where accurate distance calculation using a geodesic distance metric is essential. Big global data interpolation is an interesting exception when the predictions made using chordal distance metrics give practically the same results as more theoretically sound models based on spherical and geodesic distances. For example, interpolation using EBK with chordal distance is not only accurate but also very fast. In addition, chordal distance metrics can be used successfully when an optimal data projection for a particular region on the Earth or another planet is not available.

3. Discussion

Poor data sampling and violation of statistical model assumptions always were problems but have become much more important in the case of big spatial and space–time data, because data quality is getting worse and the number of violated

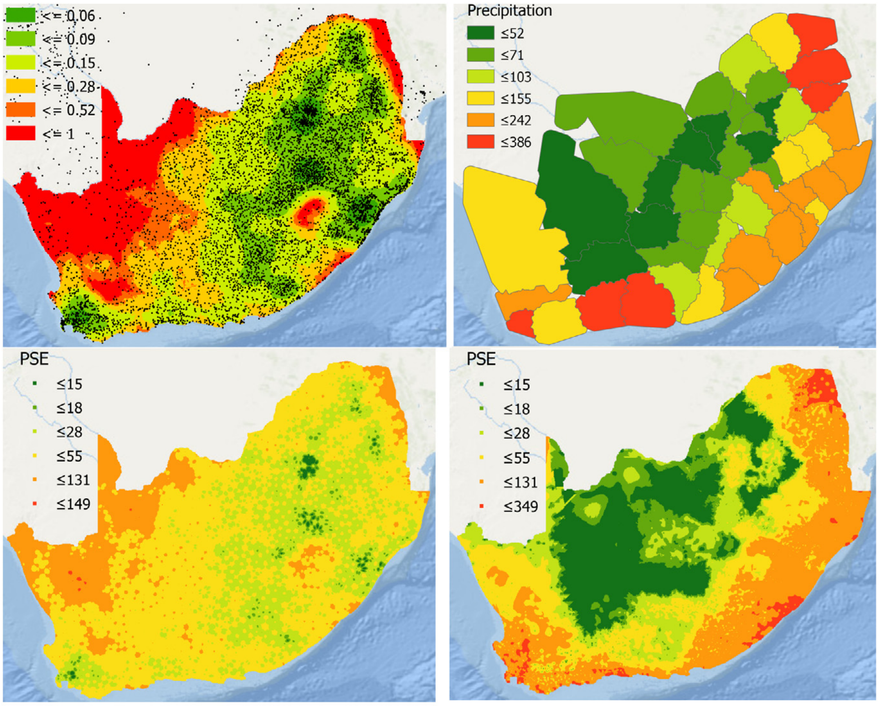

Fig. 2. Rainfall in South Africa: top left—data locations and their density; top right—annual rainfall standard deviations in areas with about 200 observations; lower row: annual precipitation prediction standard errors (PSE): left—SPDE; right—EBK.

assumptions is increasing. Use of data "to hand" often disturbs the link between what we are trying to measure, and the available data (Höhle, 2017). Therefore, much time is required to clean up the data, and this in turn requires substantial skills.

Natural hazards such as heat waves, extreme rainfall or windstorms, arise due to spatial physical processes and the amount of such data, both model output and empirical observations, is permanently increasing. Spatial statistical extreme values models are computationally demanding and it seems that currently there is no good solution for fast and efficient mapping of high distribution quantiles. For example, in order to model the complex non-stationary dependence structure of precipitation extremes over the continental U.S. using data collected at 1218 stations and make predictions at 2200 grid points, it took 25 thousand core-hours on a cluster with 39 nodes of 20 cores each (Castro Camilo and Huser, 2017). In practice, such a long computation time is unacceptable for both commercial and freeware software users, limiting the very important topic of spatial extreme values data analysis to theoretical researchers. It would be exciting to develop much faster approximation models, which produce reasonably accurate results at most in an hour.

Fitting models to big data based on gold standards for small and medium size data is problematic—there is a risk of breaking the pot. Maybe a fig-wood ladle is more like a combination of carefully selected approximations, such as the divide and conquer algorithm, flexible data transformation, and providing individual measurement errors?

References

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848. https://doi.org/10.1111/j.1467-9868.2008.00663.x.

Castro Camilo, D., Huser, R., 2017. Local likelihood estimation of complex tail dependence structures in high dimensions, applied to U.S. precipitation extremes. ArXiv e-Prints, URL: https://arxiv.org/abs/1710.00875.

Chakraborty, A., Gelfand, A.E., 2010. Analyzing spatial point patterns subject to measurement error. Bayesian Anal. 5 (1), 97–122. https://doi.org/10.1214/10-BA504.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 209–226. https://doi.org/10.1111/j.1467-9868.2007.00633.x.

Cressie, N., Kornak, J., 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. Statist. Sci. 18 (4), 436–456 URL: http://www.jstor.org/stable/3182834.

Diggle, P.J., Menezes, R., Su, T.-l., 2010. Geostatistical inference under preferential sampling. J. Roy. Statist. Soc. Ser. C 59 (2), 191–232. https://doi.org/10.1111/j.1467-9876.2009.00701.x.

Fanshawe, T.R., Diggle, P.J., 2011. Spatial prediction in the presence of positional error. Environmetrics 22 (2), 109–122. https://doi.org/10.1002/env.1062.

Gelfand, A.E., 2010. Misaligned spatial data: The change of support problem. In: Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes, M. (Eds.), Handbook of Spatial Statistics. Chapman & Hall/CRC, Boca Raton, pp. 517–539.

Gotway, C.A., Young, L.J., 2002. Combining incompatible spatial data. J. Amer. Statist. Assoc. 97, 632–648.

Gribov, A., Krivoruchko, K., 2012. New flexible non-parametric data transformation for trans-gaussian kriging. In: Abrahamsen, P., Hauge, R., Kolbjørnsen, O. (Eds.), Geostatistics Oslo 2012. Springer Netherlands, Dordrecht, pp. 51–65. https://doi.org/10.1007/978-94-007-4153-9_5.

Gribov, A., Krivoruchko, K., 2017. New flexible compact covariance model on a sphere. ArXiv e-prints, arXiv:1701.03405.

Höhle, M., 2017. A statistician's perspective on digital epidemiology. Life Sci. Soc. Policy 13 (1), 17. https://doi.org/10.1186/s40504-017-0063-9.

Krivoruchko, K., 2011. Spatial Statistical Data Analysis for GIS Users. ESRI Press, Redlands, CA, URL: https://community.esri.com/thread/201550-spatial-statistical-data-analysis-for-gis-users-available-free-for-download.

Krivoruchko, K., Bivand, R., 2009. GIS, users, developers, and spatial statistics: on monarchs and their clothing. In: Pilz, J. (Ed.), Interfacing Geostatistics and GIS. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 203–222. https://doi.org/10.1007/978-3-540-33236-7_16.

Krivoruchko, K., Frączek, W., 2016. Interpolation of Data Collected along Lines. ESRI, URL: https://apl.maps.arcgis.com/apps/MapJournal/index.html?appid=e7bd9a788b584f21ae738363b9b55d41.

Krivoruchko, K., Gotway Crawford, C., 2005. Assessing the uncertainty resulting from geoprocessing operations. In: Maguire, D.J., Batty, M., Goodchild, M.F. (Eds.), GIS, Spatial Analysis, and Modeling. ESRI Press, pp. 67–92.

Krivoruchko, K., Gribov, A., 2014. Pragmatic bayesian kriging for non-stationary and moderately non-gaussian data. In: Pardo-Igúzquiza, E., Guardiola-Albert, C., Heredia, J., Moreno-Merino, L., Durán, J.J., Vargas-Guzmán, J.A. (Eds.), Mathematics of Planet Earth: Proceedings of the 15th Annual Conference of the International Association for Mathematical Geosciences. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 61–64. https://doi.org/10.1007/978-3-642-32408-6_15.

Krivoruchko, K., Gribov, A., Krause, E., 2011. Multivariate areal interpolation for continuous and count data. Proc. Environ. Sci. 3 (Supplement C), 14–19. https://doi.org/10.1016/j.proenv.2011.02.004. 1st Conference on Spatial Statistics 2011 Mapping Global Change.

Lindgren, F., Rue, H., Lindstrm, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x.

NOAA, 2013, 2013. World ocean atlas 2013 version 2. URL: https://www.nodc.noaa.gov/OC5/woa13/.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2015. A multiresolution gaussian process model for the analysis of large spatial datasets. J. Comput. Graph. Statist. 24 (2), 579–599. https://doi.org/10.1080/10618600.2014.914946.

Oliveira, V.D., 2013. Hierarchical poisson models for spatial count data. J. Multivariate Anal. 122 (Supplement C), 393–408. https://doi.org/10.1016/j.jmva.2013.08.015.

Oliveira, V.D., 2014. Poisson kriging: A closer investigation. Spat. Stat. 7 (Supplement C), 1–20. https://doi.org/10.1016/j.spasta.2013.11.001.

Rawlins, B.G., Marchant, B., Stevenson, S., Wilmer, W., 2017. Are data collected to support farm management suitable for monitoring soil indicators at the national scale?. Eur. J. Soil Sci. 68 (2), 235–248. https://doi.org/10.1111/ejss.12417.

Reiniger, R., Ross, C., 1968. A method of interpolation with application to oceanographic data. Deep Sea Res. Oceanogr. Abstr. 15 (2), 185–193. https://doi.org/10.1016/0011-7471(68)90040-5.

Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling. Spat. Stat. 2 (Supplement C), 1–14. https://doi.org/10.1016/j.spasta.2012.08.001.