Introduction to Analysis Methods for Big Earth Data

Christopher Lynnes[1]

[1]NASA/Goddard Space Flight Center (U.S. Government Civil Servant).

Corresponding author: Christopher Lynnes (christopher.s.lynnes@nasa.gov)

Key Points:

- Big Earth Data are too big to be tractable to simple data inspection and require models to make sense of all the data.
- Useful models for Big Earth Data may be physical, statistical, or machine learning based.
- In many cases, hybrid models combine attributes of two or more of these types.

Abstract

Big Earth Data are too big to be tractable to simple data inspection. Thus, they typically require models to make sense of all the data. Useful models for Big Earth Data may be physical, statistical, or machine learning based. While physical models are ideal for understanding the data, they are not always feasible, particularly when our ability to observe at finer scales exceeds our ability to incorporate the physics. Statistical models are more generalized, but computationally intensive for many Earth Observation datasets. Machine Learning models generally scale well but are sometimes limited in the physical understanding they can offer. Hybrid models combine attributes—and advantages—of two or more of these types.

The National Institute for Standards and Technology (NIST) defines Big Data as "extensive datasets--primarily in the characteristics of volume, velocity, variety, and/or variability--that require a scalable architecture for efficient storage, manipulation, and analysis" (Chang et al., 2018). The Volume aspect of this definition has so far dominated the work in Earth Science Informatics area (though as time goes on and data sources multiply, variety and variability may well surpass the volume challenge). What does this mean for data analysis of Big Earth Observations?

Before the advent of computers, the only practical way to analyze data was through visual inspection of either the individual data points or simple statistical summaries. However, human vision can only scale so far, particularly when limited to static, two-dimensional representations of data, such as science papers. The most common approach to addressing this limitation is to employ some kind of model-based analysis. Model-based analysis can be viewed as three different classes of model: physical, statistical, and machine-learning (Fig. 1), but as we will see, these are not necessarily disjoint.

Physical models are developed based on our understanding of the physical world and its workings. This includes geophysical, biological and ecological models. Observed data may be used in a number of different ways in this type of model: as initial or boundary conditions for the model; directly assimilated into the model itself; or as validation or falsification of the model. As such, particularly in the latter case, analysis based on physical models can be a particularly stringent test of our knowledge of the physical aspects being investigated. However, physical models require a thorough knowledge of underlying principles, which may not be available for studies in new or poorly understood areas. This is particularly the case in Earth Observation as instruments are able to produce data on finer spatial and temporal scales, where the physics may not be well enough known to incorporate into a model. Physical models are also typically complex computer programs which are susceptible to bugs and can be difficult to reproduce by others, unless the model code is both published and portable.

Statistical techniques range from computing simple averages and standard deviations to sophisticated, compute-intensive techniques such as kriging and wavelet analysis. Although not always obvious, these techniques usually rest on underlying assumptions of how the data are likely to be distributed, such as the Gaussian distribution assumed to be the premise for simple averaging. Statistical models can also be powerful tools for exploring the important characteristics of datasets, particularly when we move beyond simple averaging to look at other aspects of the data, such as extreme values of natural phenomena such as air temperature or rainfall (e.g., Palharini et al., 2020). It is likely that extreme-value estimation and prediction will grow in importance for Earth Observations as important environmental measurements such as rainfall and air temperature depart from historical means (Diffenbaugh, 2020). However, if the distribution of the data deviates significantly from the assumed statistical model, statistical analysis can produce misleading results. Also, since most statistical methods compute aggregate quantities to reduce the data volume, the aggregate estimates may fail to find important features in the data. The geospatial nature of Earth Observation statistics represents both a complication and an opportunity, as treated in the following chapter on "Spatial statistics for big data analytics in the ocean and atmospheric realm: techniques, examples, and challenges". The complications arise from the added aspect of spatial proximity as an influence on statistical distributions. The opportunity presented is the ability to identify interesting phenomena such as extreme values within a specified spatial context that might not appear in a larger (e.g., global) context (Liu et al., 2017).

One of the challenges in scaling up statistical analyses is the potential for time-consuming computations. Earth Observations are particularly vulnerable to this because they are collected and processed in thin time slices, such as individual satellite scenes, or hourly or daily grids, and then usually stored by data archives in the form they are received. This arrangement of the data makes time series in particular consuming to compute due to the overhead of opening and closing each thin-time-sliced file. As a result, time series for the globe (or even large regions) can take an exorbitant amount of time to compute. Cloud computing offers the potential to make these calculations more tractable, so long as the data are also reorganized or in some cases partially precomputed, as described in the chapter "Benchmark Comparison of Cloud Analytics Methods Applied to Earth Observations."

Machine learning models are in some ways an extension of statistical models. Indeed, some of the earliest forms of machine learning, such as logistic regression and Naive Bayesian classification, have an obvious basis in statistics. While machine learning algorithms can in theory consume a virtually unlimited amount of data, the usefulness of Big Data varies with the machine learning model and the data characteristics. Adding more data to a machine learning model with high bias (overly simple or underfitting) does little to improve the model. On the other hand, adding more data to model with high variance (underfitting) can be helpful. Even in this case, however, simply acquiring more data is not always useful; class balance and spatial representativeness are essential to mitigating bias in the model (Elmes, 2020).

A challenge with the use of Machine Learning Models in data analysis is that many of the models are purely mathematical, with no obvious connection to the underlying physics. As a result, efforts toward machine learning explainability have sprung up in many fields. A common approach is to examine how sensitive the results are to perturbations in the inputs, say, using Monte Carlo estimates (Sobol, 2001), or by looking at views of the model that directly indicate sensitivity, such as the Jacobian matrix (Maddy and Boukabara, 2021)

Perhaps the most knowledge can be gleaned from Big Earth Data via hybrid models, such as process-guided deep learning, which combine physics-based and machine-learning-based models. For instance, Read et al. (2019) combine process modeling of lake temperature grounded in the conservation of energy with a deep learning model. The physics constraints are used as part of the computation of the deep learning objective function in each training epoch. This approach may produce not only better estimations, but enable scaling out to much larger datasets, thus showing promise for global monitoring.

References

Chang, W.L., 2015. NIST Big Data Interoperability Framework: Volume 1, Definitions (No. Special Publication (NIST SP)-1500-1).

Diffenbaugh, N. S. (2020). Verification of extreme event attribution: Using out-of-sample observations to assess changes in probabilities of unprecedented events. *Science advances*, *6*(12), eaay2368.

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J.R., Fishgold, L., Friedl, M.A., Jain, M., Kohli, D., Laso Bayas, J.C. and Lunga, D., 2020. Accounting for training data error in machine learning applied to Earth observations. *Remote sensing*, *12*(6), p.1034.

Liu, Q., Klucik, R., Chen, C., Grant, G., Gallaher, D., Lv, Q., & Shang, L. (2017). Unsupervised detection of contextual anomaly in remotely sensed data. *Remote Sensing of Environment*, *202*, 75-87.

Maddy, E.S. and Boukabara, S.A., 2021. MIIDAPS-AI: An Explainable Machine-Learning Algorithm for Infrared and Microwave Remote Sensing and Data Assimilation Preprocessing-Application to LEO and GEO Sensors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Palharini R, Vila D, Rodrigues D, Pareja Quispe D, Palharini R, Siqueira R, Afonso J (2020) Assessment of the extreme precipitation by satellite estimates over South America. *Remote Sens.* 12:2085. https://doi.org/10.3390/rs12132085

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, 55, 9173– 9190. https://doi.org/10.1029/2019WR024922

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, *55*(1-3), pp.271-280.

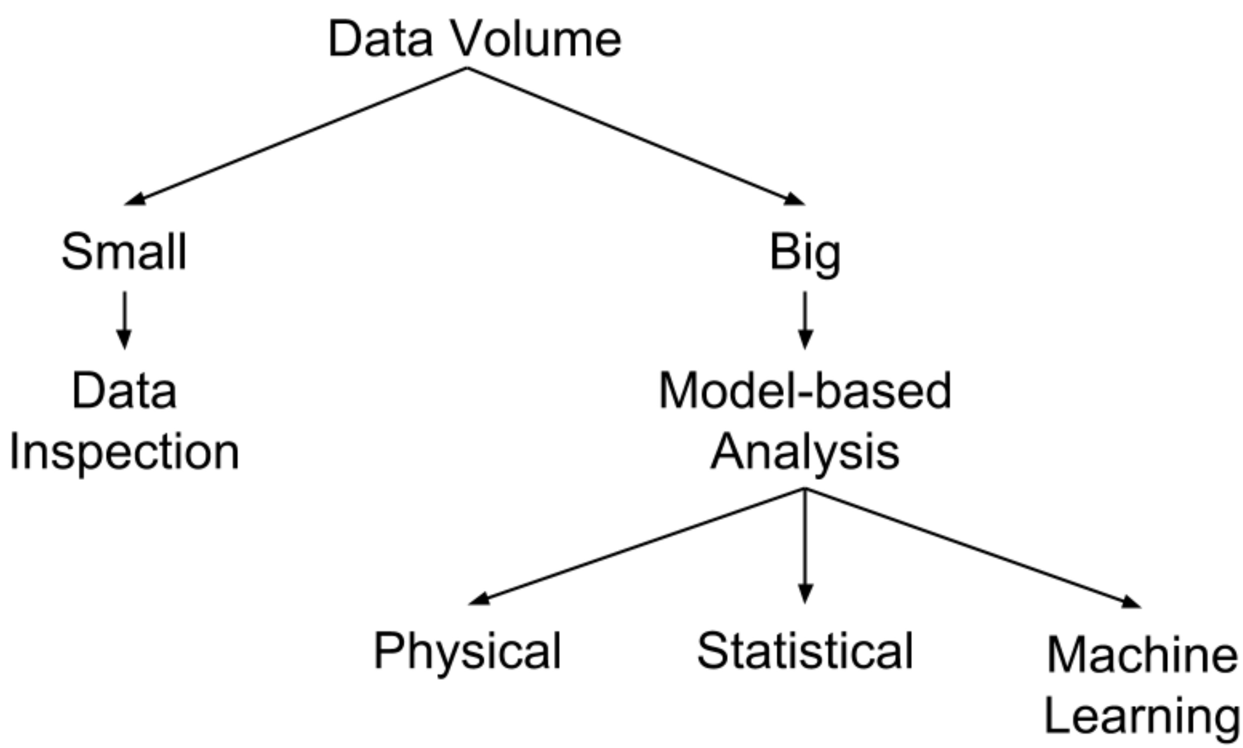

Fig. 1. Big Data volumes are too large to be tractable purely to inspection of the data. Instead, analysis must incorporate some kind of model--physical, statistical or machine-learning--in order to make most sense of the data.