Machine learning algorithms in geostatistical data analysis: formulation and observation

S Pasari¹, P Asudani¹ and A Mehta²
¹ Mathematics Department, Birla Institute of Technology and Science, Pilani, Jhunjhunu – 333031, Rajasthan, India
² Civil Engineering Department, Institute of Infrastructure Technology, Research and Management, Ahmedabad – 380026, Gujarat, India

Abstract. Geostatistical analyses on spatiotemporal datasets have been increasingly popular on diverse disciplines including environmental sciences, soil and earth studies, meteorology, hydrology, and oceanography. They help in documenting and summarizing information to understand the variation of the process and other parameters. Due to high spatial variability and noise in the dataset, traditional geostatistical tools often fail to produce desirable results. In view of this, the present study proposes two machine learning methods, namely artificial neural network (ANN) and k-nearest neighbour (kNN), along with the kriging interpolation to derive temporal distribution of soil parameters. Results in terms of smooth porosity maps demonstrate the effectiveness of the proposed algorithm.

1. Introduction

Geostatistics deals with spatial or spatiotemporal datasets in describing spatial patterns, carrying out interpolation, and thereby analysing the underlying process. Large sized, unstructured and multivariate datasets are often associated with geostatistical analysis. Many areas of science and engineering disciplines directly or indirectly benefit from geostatistical techniques. For example, in mining industry, geostatistics is routinely used to characterize mineral resources and to examine its economic feasibility; in environmental studies, geostatistics is used to estimate and organize pollutant levels to ensure its proper remediation; in the field of soil science, mapping of subsurface topography in terms of acoustic impedance, carbon content, porosity, permeability, temperature and moisture, soil nutrients, soil electrical conductivity and related parameters is performed to prescribe proper fertilizer amount and agricultural practices by agri-consultants; geostatistics is inevitable in weather prediction in meteorology, crime mapping to social science analysis, seismological studies, medical statistics and several such domains. In all of these applications, spatial interpolation using traditional methods like variography and kriging are employed to estimate the value of a variable at new locations. In the present analysis, we implement ANN and kNN models along with kriging to demonstrate the suitability of machine learning models in the estimation of geo-parameters, such as soil maps.

As raw dataset often contains significant amount of measurement errors and associated uncertainty, variogram and other conventional geostatistical tools often fail to produce desirable results, leading to the formulation of machine learning models [1]. In geostatistics, kriging is the predominant method for spatial interpolation, though it assumes that predictions are linearly related to observations [1-2]. On the other hand, deep learning models can easily reveal patterns in high-dimensional and non-linear data related to any complex process. It is observed that neural networks are very useful for modelling patterns in data, though there is a risk of over fitting, whereas geostatistical techniques may overcome such problems using predictions or simulations [1-4].
2. Literature review

Some previous works have demonstrated the application of hybrid models as a mixture of traditional geostatistical tools (e.g., kriging) and machine learning techniques. For instance, Tadic et al. [3] proposed a hybrid model comprising kriging and ANN as spatial interpolator using real airborne carbon-di-oxide mixing-ratio and synthetic data. They compared the performance of the proposed hybrid model with ordinary kriging based on isotropic variogram. Foresti et al. [4] presented a hybrid approach for mapping of precipitation data using geostatistical methods and neural networks. Integrating geostatistics and machine learning algorithms, they obtained both spatial patterns and extreme values, simultaneously. Two case studies were undertaken to demonstrate the model. Wang et al. [5] proposed a kNN-based spatial prediction method that integrates deep learning neural network into a geostatistical model. Based on the implementation on a massive forestry dataset, the authors noted that it is a stochastic process with better performance than available geostatistical models in simulated non-Gaussian data. Kanevski et al. [6] proposed another hybrid approach, namely a feed-forward neural network known as neural network residual kriging (NNRK) for approximating nonlinear trends followed by the application of geostatistical estimator or simulator on residuals. In addition to the above, effects of spatial correlation on hyper-parameter tuning and performance checking have been studied by several machine-learning algorithms such as k-nearest neighbour, boosted regression trees, random forest and support vector machine along with conventional parametric algorithms, such as logistic regression and semi-parametric generalized additive models [1-6].

3. Dataset

An observed non-stationary (non-Gaussian) dataset of porosity and temperature for 294 geographic locations is considered. The porosity values lie between 0.0411 and 0.2069, with their mean and standard deviation 0.1274 and 0.0205, respectively. The temperature data ranges between -1.8790 and 31.0330, with their mean and standard deviation 14.4756 and 11.0929, respectively.

4. Methodology and results

For the studied dataset, first, kriging is implemented to understand the relationship between temperature and porosity (figure 1). It provides a good local estimation of the spatial distribution and porosity’s dependency on temperature. After this, an ANN model is implemented to test its efficacy in modelling spatial changes of subsurface soil porosity. The network functions in two parts, namely training and testing data in about 8:2 ratios. The input parameters are fed to the network to provide porosity values as output. The architecture of the ANN is a feed forward network with four dense layers. The Adam’s optimizer is used with a learning rate of 0.005 and MSE (mean squared error) as the loss function. Results are summarized in figure 2 and figure 3.

It is observed that ANN model is somewhat smooth but not as distributed and accurate as expected, probably due to the high dependency on the neighbouring data. This suggests us to implement a non-parametric kNN model that promises to account for the regression using neighbourhood data exclusively. The kNN captures the idea of similarity (e.g., distance, proximity, or closeness) in its working.

For the implementation of kNN, one needs to choose k number of neighbours, compute distances, and set the hyperparameters (e.g., weights, n-neighbours and distance metric, such as Euclidean and Manhattan distance) iteratively. As kNN uses a nearest training sample search in feature space (like k-means clustering), one may remove the impact feature range from the approach by standardizing predictor features with zero mean and unit variance. Here, the scikit-learn preprocessing library of Python is used to perform this step.

As an illustration of the above process, we provide the kNN results with similarity index as distance with k = 5 and k = 15 (Figure 4 and Figure 5, respectively). Small k results in a local specific prediction model, whereas large k results in a smoother, globally fit prediction model over the predictor features space.
5. Summary
In most of the datasets, the scatter plot of porosity suggests that a simple linear regression model may be inadequate to capture the intrinsic patterns of data. Therefore, nonparametric approaches are often employed to fit the nonlinear response patterns in the predictor feature space.
In this study, we implemented a traditional geostatistical method (kriging) and two machine learning based models (ANN and kNN) to derive temporal distribution of soil parameters, such as porosity. We followed the standard practice of pre-processing (to observe, for example, outliers and data gaps), processing, post-processing (validation) of porosity values. Results in terms of smooth porosity maps demonstrate the effectiveness of the proposed kNN algorithm. The kNN model studied here is extendable to multivariate spatiotemporal datasets.

6. References

[1] Chiles JP and Delfiner P 2012 Geostatistics: Modeling Spatial Uncertainty, Wiley.
[2] Schratz P, Muenchow J, Iturritxa E, Richter J and Brenning A 2019 Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data, *Ecological Modelling* **406**, 109–120.
[3] Tadić JM, Ilić V and Biraud S 2015 Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments, *Atmospheric Environment* **111**, 28–38.
[4] Foresti L, Pozdnoukhov A, Tuia D and Kanevski M 2010 Extreme precipitation modelling using geostatistics and machine learning algorithms, In *Quantitative Geology and Geostatistics*, 41–52.
[5] Wang H, Guan Y and Reich B 2019 Nearest-neighbor neural networks for geostatistics, *International Conference on Data Mining Workshops*.

[6] Kanevski M, Parkin R, Pozdnukhov A, Timonin V, Maignan M, Demyanov V and Canu S 2004 Environmental data mining and modelling based on machine learning algorithms and geostatistics, *Environmental Modelling and Software* **19** (9), 845–855.