A Methodology for Automatic Analysis and Modeling of Spatial Environmental Data

Mikhail Kanevski
University of Lausanne, Centre for Research on Terrestrial Environment, Geopolis building
1015 Lausanne, Switzerland
Mikhail.Kanevski@unil.ch

Abstract—**The research paper deals with a step-by-step methodology for the automatic modeling of geospatial environmental data. The methodology proposed is based on general regression neural networks (GRNN) and probabilistic neural networks (PNN) as modeling tools. GRNN and PNN are nonparametric nonlinear models suitable for the automatic analysis, modeling, and spatial predictions of complex environmental data. The simulated and real data case studies illustrating the methodology are considered and discussed.**

Keywords-environmental geospatial data; automatic data modelling; machine learning algorithms.

I. INTRODUCTION

The problem of automatic environmental data modeling becomes more and more important taking into account the volume of data available from different sources: measurements, automatic monitoring networks, remote sensing, GIS (Geographical Information Systems), etc. These data are widely used to calibrate science-based models (e.g., in meteorology, climate, pollution dispersion), to estimate environmental risks and natural hazards (landslides, avalanches, forest fires etc.), and to estimate renewable resources. Most of environmental data bases contain extremes and outliers and data are highly variable at several spatial scales. Moreover, the environmental phenomena are nonlinear and in many cases should be considered in a high dimensional feature spaces composed of 3d geographical coordinates and additional characteristics derived, for example, from digital elevation models [1][2][3].

A classical approach to analyze environmental geospatial data is based on geostatistical models [4][5]. Most of geostatistical models explicitly take into account the anisotropic spatial correlations analyzed and modeled by an application of variography. In general, geostatistics is a powerful and well established data modeling tool in a low dimensional space.

Recently, an intercomparison of models suitable for automatic two dimensional interpolations in a geographical space was carried out and the results are presented in a report [6]. A wide variety of methods was used – from traditional geostatistical kriging models to advanced neural networks. The general regression neural network produced very good results in terms of testing error and other global statistics usually used to quantify the quality of modeling.

The present study generalizes the ideas proposed in [2][7][8][9] for modeling and predictions of high dimensional complex spatial environmental data. In the following section, a description of a general problem is given and an operational and efficient methodology for automatic geospatial data analysis and modeling is considered step-by-step with a short discussion of real and simulated data case studies. The paper is completed with a brief discussion and some conclusions for the future research.

II. THE METHODOLOGY AND CASE STUDIES

In general, the problem considered is the following: having environmental data embedded in a high dimensional space, develop a model that 1) explains the phenomena under study without overfitting of data and 2) is good enough (the criteria should be defined) for the generalization and spatial predictions, often called mapping. Moreover, it is necessary to justify the quality of the results obtained by using some general criteria and to quantify the uncertainties of the predictions.

Environmental phenomena are nonlinear and their data modeling should be considered in a high dimensional space (dimension d>3) that is composed of geographical coordinates and some additional variables (features), for example, produced from digital elevation models (e.g., slopes, curvatures, variability at different scales, etc.). Moreover, only some of features could be relevant for the analysis and predictions, and some of them can be just a noise. Therefore, the problem of relevant automatic features selection or features extraction during an automatic analysis can be important.

Taking into account the comments and demands given above, the following generic methodology can be proposed for the spatial environmental data analysis and modeling (from exploratory analysis to spatial predictions and decision-oriented mapping):

1. Preparing of an input/feature space (a collection of independent variables). In principle, the library of input features should be quite general to cover a wide range of possible scenarios.
2. Analysis of monitoring networks and data clustering taking into account the validity domains of raw data and a prediction grid. Monitoring network analysis also helps to understand the representativeness of data and their spatial topology. It improves the decision on data splitting and declustering procedures, and, if necessary, in a monitoring network optimization (redesign) [5][10].

 ISBN: 978-1-61208-251-6

3. Exploratory spatial analysis of data (ESDA) using visualization (geo-visual analytics), (geo)statistical tools [4][5], and machine learning algorithms (MLA) [8]. At this important stage of the analysis, MLA can help to detect potential patterns in a high dimensional feature space. Data pre-processing and data transformations.
4. Splitting of data into training (development of the model), validation (fitting or calibration of the user-defined hyper-parameters), and testing (assessment of the generalization) subsets. Different criteria can be used: random splitting, spatial declustering, etc.
5. Detection of the available patterns/structures in a high dimensional data. The discrimination between white non-structured noise and spatially structured information. In geostatistics usually the variography is used. In a more general case, MLA can be efficient as well.
6. Training (optimization, calibration) of the models. Modeling of the observed structured information.
7. Iterative application of the feature selection algorithms - either the features are weighted according to their importance or the group of the most relevant features for the prediction is selected [1][2][3][8]. These techniques can also be applied at the steps 1, 5 and 6.
8. Analysis of the training residuals using visualization, (geo)statistical and machine learning tools. Analysis of the residual patterns. The same procedure as in 5 is applied: the residuals should have no spatial structure and should be normally distributed. Moreover, an overfitting of the training data should be avoided. One of the possibilities is to estimate a noise level in data, for example, using an estimate of a nugget in a high dimensional feature space. Then, the variance of a noise can be used as a stopping criterion.
9. Testing of the models. Application of the developed models to testing data subset. Analysis of the testing residuals and their spatial structure. Again, retrained GRNN can be used to perform an exploratory spatial analysis of the testing residuals.
10. Application of the validated and tested models for the spatial predictions (mapping in high dimensional spaces).
11. Quantification of the modeling quality: confidence and prediction intervals.
12. Decision–oriented mapping. At this phase GIS can be widely applied.

Current version of the methodology does not include the recommendations on the monitoring network optimization (MNO). This is a separate but closely related problem. The contemporary reviews on MNO approaches along with space-time environmental data case studies are given in [10].

Recently, the application of general regression neural networks (GRNN) for the regression and the probabilistic neural networks (PNN) for the classification were reconsidered taking into account their properties of patterns detection and adaptivity in the feature selection problems [2][6][7][8]. For example, GRNN is an efficient tool to discriminate noise from structured information. This property can be used both for the original data and for the analysis of the residuals to estimate the quality of modeling. As it was mentioned above, “good” residuals should be white spatially non-structured noise. Basic GRNN model has no hyper-parameters and is easy to train. Therefore, GRNN is attractive model for the automatic data processing.

Anisotropic GRNN can automatically neglect highly noisy features and takes into account only the relevant ones [2][7]. Anisotropic GRNN (when Gaussian kernel is used) means that different kernel bandwidths are used for different features (independent variables). This version of GRNN is sometimes called an adaptive GRNN [2].

It is important to note, that non-parametric statistics is a solid theoretical background both for GRNN and PNN. Therefore, these models can produce also extended results including the characterization of the uncertainties. This is extremely important in a real decision–oriented mapping process when the uncertainties can be even more important than the predictions themselves.

The GRNN/PNN training procedure applied in this research (training = selection of the optimal kernel bandwidths by applying optimization algorithms) is based on a cross-validation error cost function. Either a leave-one-out or a leave-k-out error functions are considered, depending on the number of available training data. In case of too many data a validation data set can be used to train the model. It accelerates the training procedure and reduces the computational time.

The test data set is used only to estimate the generalization properties of the models, i.e., their abilities to predict independent data never seen during the training.

In the present research GRNN was used with an anisotropic Gaussian kernel. In a more general setting a complete Mahalanobis distance can be applied. More theoretical details about the models and their implementation can be found in [8].

The PNN has the same kind of properties and can be used for the classification problems when working with categorical data – discrete classes.

In the present research, the methodology is illustrated using the simulated and real data case studies. Simulated data were produced by adding to the real data several noisy artificial features in order to test the ability of GRNN to neglect the non-relevant information. Artificial additional features were generated using a shuffling procedure, i.e., by randomizing the raw variables. In this case, the original global distributions are preserved but the spatial structures, even if present, are destroyed. The following case studies were considered:

- real data case study. Topo-climatic modeling of the monthly temperature and precipitation in mountainous regions. These are typically three dimensional problems.
- simulated data case study. Three new artificial features were generated either by shuffling of X, Y and Z geographical coordinates or by noise injection

 ISBN: 978-1-61208-251-6

with different variance. Finally, the problem was considered in a six dimensional space [7].

During the case studies all phases of the methodology were applied. The homogeneity of monitoring network was studied using topological, statistical and fractal measures. Exploratory analysis was carried out using statistical and geostatistical (variography) tools. Measurements were split into training and testing subsets using spatial declustering procedure. Validation subset was not necessary because of the cross-validation (leave-one-out) training technique was applied.

Below the modeling results are discussed briefly. In real data case studies the kernel bandwidths for geographical coordinates reflect spatial 3d anisotropy of the phenomena: for longitude X and latitude Y they are of order 10 km and for altitude Z few hundred meters. The results were compared with a geostatistical model – kriging with external drift [4][8].

In the simulated data case study, noise features after training have very large kernel bandwidths, exceeding the variability of these features. In this case the corresponding part of the Gaussian kernel equals almost to one and these features do not influence the solution. Then, according to the methodology, adaptive GRNN was applied for the exploratory spatial analysis of the training residuals. No spatial structures were detected. At the end, both 3d and 6d (3d+noise) solutions were very similar.

Finally, the models developed were evaluated using testing data subset and good generalization errors were obtained.

More powerful and much more computationally intensive approach is based on a complete analysis of all possible models, i.e. on all possible combinations of features. In a d dimensional input space the number of possible models is (2^d-1). In this case GRNN is applied both as a modeling and as a feature selection tool. Using a cross-validation error, all models can be sorted and the best one with a minimum error can be selected for the predictions. Such approach was also applied for both case studies. Important result is that the best selected models did not include noise features.

III. DISCUSSION AND CONCLUSIONS

A basic methodology for spatial data processing was proposed. The methodology includes the analysis of input space structure (monitoring networks), comprehensive exploratory analysis of data and the residuals, detection and modeling of structured information using nonlinear nonparametric models. As an efficient and operational tool adaptive GRNN for the regression problems and adaptive PNN for the classification problems were proposed. Training of models was based on cross-validation procedures. The modeling results were evaluated by using independent testing subsets and by analyzing the testing residuals.

One of the important and useful conclusion from the study is that an application of machine learning algorithms at all phases of the data analysis and modeling is strongly recommended [8][9]. In many cases, it helps to reveal complex hidden patterns and structures in data that improves the selection and calibration of models, even if other modeling approaches finally are applied.

The potential extension of the methodology and models can be guided into the following directions: scaling of models with the dimension of space and number of measurements, robustness of the approach, more elaborated assessment of the uncertainties, extension of MLA modeling tools and adaptive kernels [8]. An important future research will be in developing multi-scale multivariate models. For the clustered monitoring networks, kernels can be not only feature-adapted but space-adapted as well.

Only two basic problems of learning from data were considered – classification and regression. The third one and the most difficult – modeling of spatially distributed probability density functions, is still an open question.

Finally, the same kind of methodology can/should be generalized and adapted for the modeling and predictions of spatial-temporal environmental data.

ACKNOWLEDGMENTS

The author acknowledges useful discussions on the topics presented with Dr. A. Pozdnoukhov, Dr. V. Timonin, Dr. L. Foresti and S. Robert.

The research was partly supported by the Swiss NSF grant #140658.

REFERENCES

[1] L. Foresti, D. Tuia, M. Kanevski, and A. Pozdnoukhov, "Learning wind fields with multiple kernels", Stochastic Environmental Research and Risk Assessment. Volume 25, Number 1, pp. 51-66, 2011.

[2] S. Robert, L. Foresti, and M. Kanevski, "Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks", International Journal of Climatology, **32**, DOI: 10.1002/joc.3550, 2012.

[3] A. Pozdnoukhov, G. Matasci, M. Kanevski, and R. Purves, "Spatio-temporal avalanche forecasting with support vector machines", Natural Hazards and Earth System Sciences. Vol. 11, pp. 367-382, 2011.

[4] J-P. Chiles and P. Delfiner, "Geostatistics: Modelling Spatial Uncertainty". John Wiley & Sons, Hoboken, NJ, 2012.

[5] M. Kanevski and M. Maignan, "Analysis and Modelling of Spatial Environmental Data", EPFL Press, Lausanne, 2004.

[6] G. Dubois (Editor), "Automatic Mapping Algorithms for Routine and Emergency Monitoring Data", EU Report No. EUR 21595EN, 152 pp, 2005.

[7] V. Timonin and M. Kanevski, "On Automatic Mapping of Environmental Data Using Adaptive General Regression Neural Network". GIS Research UK annual conference, London, pp. 423-427, 2010.

[8] M. Kanevski, A. Pozdnoukhov, and V. Timonin, " Machine Learning for Spatial Environmental Data. Theory, Applications and Software", EPFL Press, Lausanne, 2009.

[9] M. Kanevski, V. Timonin, and A. Pozdnoukhov, "Automatic Mapping and Classification of Spatial Environmental Data", in Geocomputation, Sustainability and Environmental Planning, B. Murgante, Burusso G., and A. Lapucci, Eds., Springer, pp. 205-223, 2011.

[10] J. Mateu and W. Mueller (Editors), "Spatio-temporal design. Advances in efficient data acquisition", Wiley, 2012.

 ISBN: 978-1-61208-251-6