LAZY MAPPING OF ENVIRONMENTAL AND POLLUTION DATA

M. F. Kanevski
Institute of Nuclear Safety (IBRAE)
B. Tulskaya 52, 113191 Moscow, Russia, m_kanevski@ibrae.ac.ru

SUMMARY. The work deals with the application of modern lazy learning algorithms to environmental and pollution decision-oriented mapping. The term ‘lazy’ comes from the recent developments in machine learning algorithms terminology. The ‘learning’ here means understanding of original data and development of the corresponding algorithm capable to respond/predict/interpolate to new requests. Lazy mapping algorithms presented in the report have close relation to general regression (GRNN) and probabilistic (PNN) neural networks and to nonparametric function estimations based on locally weighted regression [1-7]. The case study is based on mapping of radioactively contaminated territories after the Chernobyl accident.

1. INTRODUCTION

The paper discusses application of locally weighted regression, a form of lazy learning algorithm, to environmental and pollution spatial mapping. Synonyms for lazy learning are memory-based, instance-based, k-nearest neighbour algorithms. Lazy methods often use local approaches, which yield highly adaptive behaviour. Recently, mapping with spatially localised basis functions has become a popular approach. Lazy learning algorithms usually have no training phase, and perform computation only at prediction time. In general, lazy learning algorithms exhibit three characteristics that distinguish them from other (eager) learning algorithms [8]. First, they defer processing of their inputs until they receive requests for information. They simply store their inputs for future use. As a result, they typically require greater memory storage. Second, they reply to information requests by combining their stored (original) data: take a query, search the database for similar data points and build an on-line local model with which to predict an output value. Third, they discard the constructed answer at any intermediate result. As a result, the learning is simple - one-shot learning, but they have higher computational costs when answering requests. For typical environmental problems in 2D and 3D with moderate number of samples it is not a problem.

Lazy learning algorithms can be modified to return a confidence interval along with its spatial prediction, which is especially important in environmental decision-oriented mapping. Of course, lazy algorithms can be adapted for local conditional probability density function estimations and corresponding probabilistic environmental and pollution mapping. Of course, very rarely data are homogeneously distributed in space. Local learning algorithms attempt to adjust locally the capacity of the training system to the properties of the data in each area of the space. Adding new data to lazy models is very cheap and does not require model restructuring. Therefore, lazy models can be easily adapted for automatic decision-oriented environmental mapping [5]. Finally, local outlier detection methods do not label points as outliers for all queries, as do global methods. Points can be outliers for some queries and not outliers for others. Local robust regression models (e.g., LOESS, LOWESS) can be used in case of data contaminated by outliers [2,4]. Some extensions of the nonparametric kernel regression approach were made so that it too displayed kriging features (both estimations and estimation variances) were carried out in (Yakowitz and Szidarovszky 1985). In particular, a data-driven estimator of the expected square error was derived.
2. LOCAL REGRESSION MODELS

In the present study, the main attention is paid to the local linear models applied to spatial predictions (approximations) of smooth functions. A smooth function is: 1) a continuous function with a limited number of discontinuities; 2) close points in the space have close values of function (spatial continuity). The basic idea is to decompose difficult global problems into simpler local problems.

Let us consider in short the basic equations and formulas. If the observed data are denoted by \( Z(X_i, Y_i) \) then for estimation at the point \((X, Y)\) the weighted least squares formulation is

\[
\min_{\alpha, \beta, \gamma} \sum_{i=1}^{n} \left\{ z(x_i, y_i) - \alpha - \beta (x_i - x) - \gamma (y_i - y) \right\}^2 
K \{d(x_i, y_i; x, y); h_x, h_y\}
\]

where \( K\{d(); h_x, h_y\} \) is a two-dimensional kernel function depending on distance functions \( d \) and smoothing parameters (bandwidth) \( h_x, h_y \). When the kernel is separable

\[
K \{d(x_i, y_i; x, y); h_x, h_y\} = K_x \{d(x_i - x) / h_x\} \cdot K_y \{d(y_i - y) / h_y\}
\]

The resulting estimator, corresponding to the quantity \( \alpha \) can be compactly defined in matrix notation [2,7]. If \( R \) denotes an \( n \times 3 \) design matrix whose \( i \)-th row consists of the elements \( \{1, (x_i - x), (y_i - y)\} \), and \( W \) denotes a matrix of 0s with the weights \( K\{d(); h_x, h_y\} \) for each observation down the diagonal, then the local linear estimator can be written as the first element of the least squares solution \( (R^TWR)^{-1}R^TWZ \) where \( Z \) denotes the vector of responses for each observation [2,4,5].

There are several basic kernel functions frequently used in the kernel density estimates and regressions [2]. The requirements on a kernel function are straightforward. Usually kernels are symmetric and smoothly decreasing functions from maximum value at zero. Discontinuities in kernel functions lead to discontinuities in the predictions. The smoother the kernel function, the smoother the estimated function. There is a variety of distance functions to apply [1]. The most frequently used is the unweighted Euclidean distance:

\[
d_E(u_i, u) = \left[ (u_i - u)^T (u_i - u) \right]^{1/2}
\]

where \( u_i^T = (x_i, y_i), \ u^T = (x, y) \). In general, diagonally weighted Euclidean distance is used

\[
d_m(u_i, u) = \left[ (u_i - u)^T M^T M (u_i - u) \right]^{1/2}
\]

where matrix \( M \) can be arbitrary. This is also known as the Mahalanobis distance. Commonly used kernels include the symmetric Beta family \((t \sim d/h) [2]\):

\[
K(t) = \left\{ Beta(1/2, \gamma + 1) \right\}^{-1} \left( 1 - t^T t \right)^\gamma_+ , \ \gamma = 0,1,\ldots
\]

where the subscript + denotes the positive part, which is assumed to be taken before the exponentiation. The choices \( \gamma = 0,1,2, \) and 3 lead to respectively the uniform, the Epanechnikov, the biweight and the triweight kernel function. Gaussian kernel is in the limit \( \gamma \rightarrow \infty \). The widely used Epanechnikov kernel is \( K(t) = (3/4)(1-t^2)_+ \).
The only unknown parameters are bandwidths \( h \). In GRNN formulation this is a parameter to be obtained in training. Smoothing parameters \( h \) influence upon the type of solution. When \( h \) is small (\( h \to 0 \)) the solution converges to interpolation (i.e., \( Z_m \to Z_i \) if \( (x,y) \to (x_i,y_i) \)). When \( h \) is large, the smoothing is applied and the solution converges to approximation. If \( h \to \infty \) \( Z_m \to \Sigma Z_i / n \) - the sample mean of the observations. By changing the smoothing parameter the quantity and quality of extracted spatial information can be controlled.

In case of local constant regression the resulting estimator is the Nadaraya-Watson kernel regression estimator:

\[
Z_{NW} (x, y; h_x, h_y) = \frac{\sum_{i=1}^{n} K \left\{ d \left( ; h_x, h_y \right) \right\} Z_i}{\sum_{i=1}^{n} K \left\{ d \left( ; h_x, h_y \right) \right\}}
\]

This estimator was reformulated as a General Regression neural Network (GRNN) by Specht [9]. There are several ways how to tune bandwidths [2,4]:

- **Global tuning**: the fit parameters are set globally by an optimisation process that typically minimises cross-validation error over the data, and therefore constant size and shape volumes of data are used to answer queries.
- **Query-based local tuning**: the fit parameters are set on each query based on local information
- **Point-base local tuning**: the weighted training criteria uses different fit parameters for each point \( X_i \): a bandwidth \( h_i \), a distance metric \( d_i \), a weighting function \( K() \), and possibly a weight \( w_i \):
- **Plug-in-approach**: the fit parameters can be set by a direct calculation.
- **Optimisation approaches**: the fit parameters can be set by an optimisation process that either: minimises the training set error; minimises the test or validation set error; minimises the cross-validation error; minimises generalised cross-validation error; minimises Akaike’s information criterion; or adjusts Mallow’s \( C_p \).

### 3. CASE STUDY

The case study is based on regional [approximately \((110\times70)\) sq. km] irregularly sampled data on soil contamination in Western part of Briansk region (Russia) by Sr90 radionuclide. The original input space \((X,Y)\) was projected into the region \([-1,+1]\). This is the most contaminated region in Russia due to the Chernobyl accident and has been studied by using both traditional and geostatistical models. The problem, that complicates the study and mapping, is high variability of fallout at different scales. As usually, when dealing with environmental data, non-linear large scale trends in the region are important. Previously the same data have been studied with geostatistical models and artificial neural networks [11,12]. Below the main results of using Nadaraya-Watson (GRNN) spatial predictor are presented. Gaussian kernel is used, which is a standard for the GRNN.

\[
K \left\{ d \left( x_i, y_i ; x, y \right); h_x, h_y \right\} \sim \left[ 1 / h_x h_y \right] \exp \left\{ -\frac{\left( x_i - x \right)^2 \left( y_i - y \right)^2}{h_x^2 h_y^2} \right\}
\]

One of the most important problems is the selection of bandwidth. In the present study cross-validation (leave-one-out) method has been used. For the present data it was founded, that using two dimensional anisotropic bandwidth does not essentially influence upon final
results. This is the reason why isotropic bandwidth parameter was used. Moreover, because the monitoring network (design) is not highly clustered (there are no very sparse regions), applied bandwidth parameter was a constant and did not depend on the position in space. Cross-validation error curve is shown in Figure 1. The easily detected minimum corresponds to the optimal bandwidth parameter.

![Fig.1. Cross-validation error function.](image)

The regression residuals were analysed with the help of exploratory data analysis and variography. Examples of omnidirectional variograms of the original data and the residuals are presented in Figure 2. Residuals are normally distributed and are not spatially correlated - pure nugget effect. The most important information, described by univariate statistics and measures of spatial continuity (variograms) has been extracted. It should be noted that nugget values for data variogram and residual variogram are the same. By changing bandwidth it is possible to control the model and residuals. At higher bandwidth values it is possible to model only large scale trends; residuals under this conditions are correlated. The bandwidth parameter is not optimal. At present, local models when residuals are correlated are under development.

The optimal bandwidth parameter was used for spatial prediction mapping (Fig.3.). The results were compared with the outputs of geostatistical models and feedforward neural networks (multilayer perceptrons, MLP). For the comparison original data were split into training and validation data sets. Training data set was used in order to develop model, validation data set was used as an independent additional measurements for testing and comparison of models. In terms of mean square error for validation data set local model was as good as geostatistical models and better than feedforward neural network. It should be noted that multilayer perceptrons construct global approximations to nonlinear input-output mapping. They are capable of interpolations in regions of the input space where little data are available. Local models construct local interpolations/approximations to nonlinear input-output mapping. Thus, different approaches can have advantages, depending on the monitoring network design and the quantity and quality of data. The possibilities to model large scale nonlinear trends with the help of MLP were studied in detail in [11,12].
The preliminary results obtained with local regression models are promising. The case study presented is based on the application of kernel smoothers (GRNN) for mapping of radioactively contaminated territories. Kernel smoothers can be efficiently used for the estimation of local probability distribution functions (probabilistic mapping) [2-4]. An application of the kernel smoother to the direct interpolation of indicator variables is presented in [9].

Kernel smoothers (particularly GRNN) have some disadvantages which are common to any nonparametric methods [2-4,7]. In the nonparametric setting all approximations $F$ are biased estimators of the unknown function $G$, because there does not typically exist a function $F^*$ with finite number of parameters able to approximate $G$. Especially substantial bias can be observed near the boundary where the neighbourhood is asymmetric. Bias can also be a problem in the interior region if the predictors are non-uniform or if the regression function has
substantial curvature [5]. Some of these problems can be solved by development multivariate (co-estimations of several variables carrying different quantity and quality of information) kernel smoothers (MKS). Additional variables can reduce boundary effect, effect of clustering and improve predictions in sparse regions. Another solution is to use local polynomial regression models, which also provide natural estimations for the derivatives of unknown function. Local polynomial estimators have nice minimax efficiency properties: the asymptotic minimax efficiency for commonly used orders is 100% among all linear estimators and only a small loss has to be tolerated beyond this class [2]. The local polynomial approximation methods are appealing on general scientific grounds: the least squares principle to be applied opens the way to a wealth of statistical knowledge and thus easy generalizations. Local methods are highly adaptive at different spatial scales. These are open issues for further research.

ACKNOWLEDGEMENTS. The work was supported in part by INTAS grant 96-1957.

REFERENCES