Semiparametric regression GP tool

Konstantin Krivoruchko, July 7, 2015

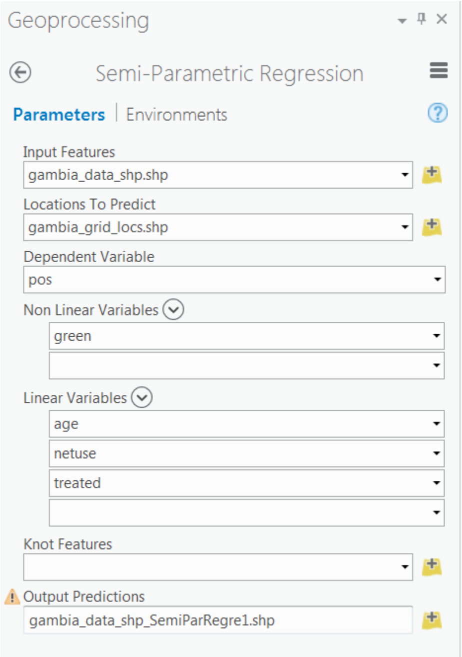
The presence/absence data: interpolation problem and the semiparametric regression solution

Discrete indicator 0/1 data (they are also called categorical or binary variables), where 0 usually indicates absence of the event and 1 indicates its presence, are common in GIS applications such as modeling land use change, epidemiology (e.g. whether or not a certain disease is diagnosed) and forestry (e.g. whether or not a tree is infested). Semiparametric regression is one of the models used to interpolate the binary data, see section “Some theory” below.

How to use this GP tool

The semiparametric regression GP tool models the indicator data Z_i that have one of two possible outcomes:

- 1 (i.e. presence of a particular event) or
- 0 (i.e. absence of a particular event).



Geoprocessing

Semi-Parametric Regression

Parameters | Environments

Input Features: gambia_data_shp.shp

Locations To Predict: gambia_grid_locs.shp

Dependent Variable: pos

Non Linear Variables: green

Linear Variables: age, netuse, treated

Knot Features:

Output Predictions: gambia_data_shp_SemiParRegre1.shp

Figure 1. Semiparametric regression GP tool.

<I suggest the following modifications:

- Rename Non Linear Variables to Spatial Variables (they exist everywhere in the data domain).
- Rename Linear Variables to Non-spatial Variables (individual characteristics).
- Add optional graphics with estimated relationships between dependent variable and independent variables.>

Input features: a table with dependent and all independent variables at the observed point locations.

Locations to predict: table with all independent variables at point locations where prediction is required.

Dependent variable: the variable of interest, which can have one of the following two values: 0 or 1.

Spatial variables: point features with z-values extracted from smoothly changing surfaces such as air pollution, temperature or elevation.

Non-spatial variables: discrete event data such as individual characteristics of people, plants, or houses.

Knot features: from 50 to 200 well-distributed locations for the basis functions selected in such a way that additional knots became unimportant to the final prediction. The default knots configuration is provided.

Output predictions: point feature class with input explanatory variables, predictions and prediction standard errors.

Regression coefficients: the pop up window with graphical relationships between dependent and explanatory variables.

Case study

Working with aggregated epidemiological data is often not appropriate and researchers are interested in individual level inference. For example, the risk of disease depends on age, gender, and bad habits and these individual characteristics are not very helpful when they are averaged over administrative boundaries.

The malaria prevalence samples in children were recorded at villages in Gambia, Africa. There are 2035 observations in 65 villages on the following variables:

- Presence (1) or absence (0) of malaria in a blood sample taken from a child (36 percent of blood samples were tested positive for malaria).
- Satellite-derived measure of the greenness of vegetation in the immediate vicinity of the village (arbitrary units). The greenness of vegetation is a function of the normalized difference vegetation index (NDVI) and it was found in several epidemiological studies that the greenness values are correlated with malaria incidence anomalies, see reference [4] below. In particular, the greenness is an indicator of areas where rainfall runoff may have a significant impact on malaria transmission.
- Age of the child in days (average age is 1080 days)
- Indicator variable denoting whether (1) or not (0) the child regularly sleeps under a bed-net (71 percent of the children used a bed-net).
- Indicator variable denoting the presence (1) or absence (0) of a health center in the village (68 percent of villages have a health center). This non-spatial variable is associated with importance of the delivery of primary health care.

Modeling of the malaria risk is challenging since three variables are binary and two variables are continuous, and there are many measurements in each village. The log odds

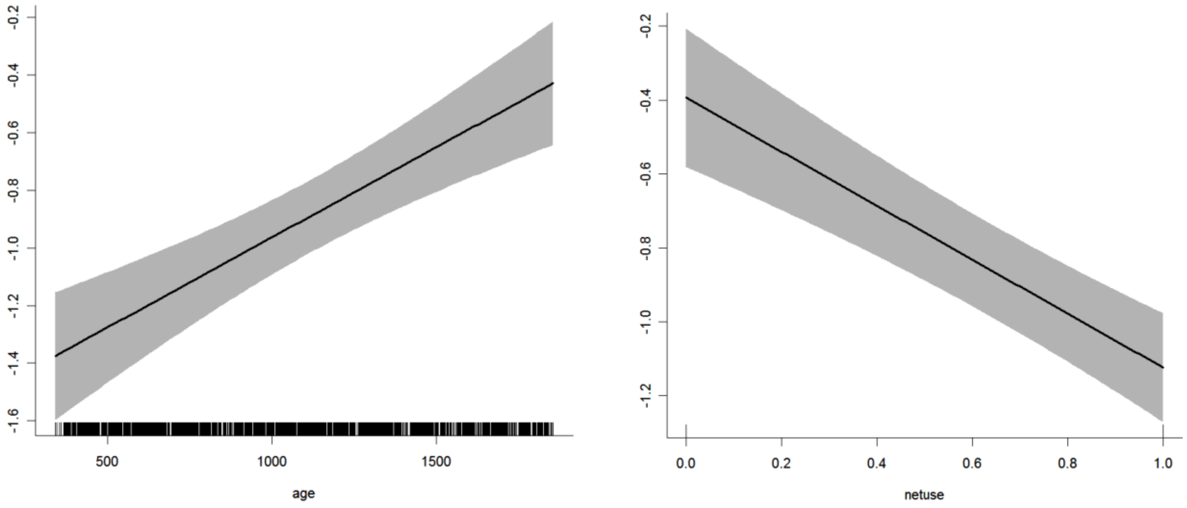
$$\log\left(\frac{p}{1-p}\right),$$

where *p* is a probability of disease presence, of malaria presence at a given location can be modeled by the semiparametric regression model as

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot netuse + \beta_3 \cdot (\text{health center presence}) + f_1(\text{greenness}) + f_2(x, y)\ ,$$

where β_i are coefficients to be estimated, $f_2(x, y)$ is a function of coordinates of the disease presence, and a function of greenness f_1 is modeled using the 2D splines (see detail explanations in the references below).

Figure 2 shows the relationships between the disease outcomes and age, netuse, the health centers presence, and greenness. Gray areas in the graphs in figure 2 show a 95 percent prediction interval around estimated dependency shown as the black lines. Variability bands are obtained by adding and subtracting twice the standard error of the estimated function. Using these relationships, predictions to the locations were the explanatory variables are specified can be obtained.



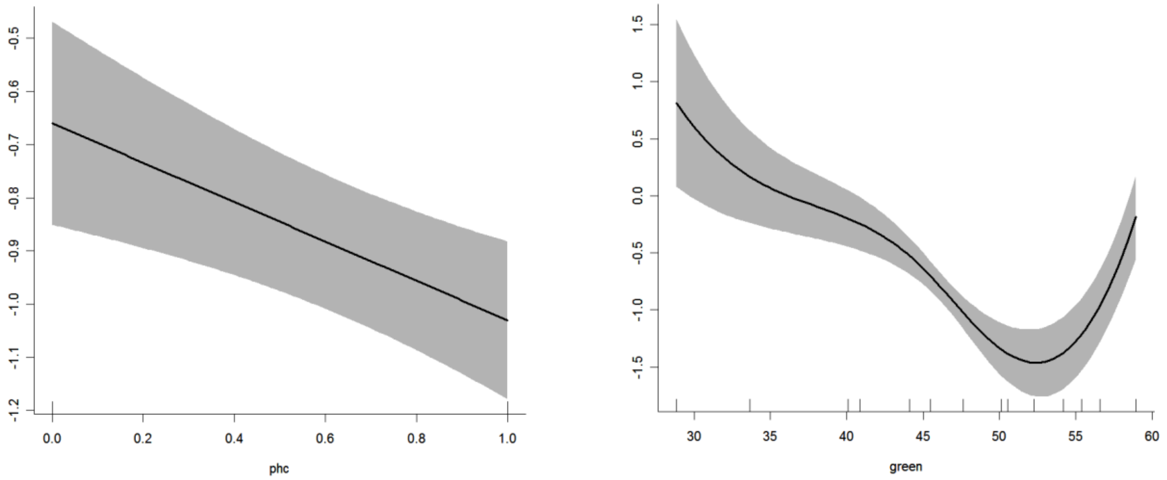


Figure 2. The relationships between the disease outcomes and age, netuse, the health centers presence, and greenness.

Note that the greenness of vegetation (figure 2 at bottom right) is modeled using splines, allowing for a non-linear relationship between the malaria presence and the covariate, while the dependence between the response variable and other explanatory variables is linear.

The semiparametric model can predict values of the disease presence at the locations where explanatory variables are known. Therefore, a map of the malaria risk can be created assuming that the age of children, the use of the mosquito net, the presence of the health center, and the greenness of vegetation are known. The greenness data are available (figure 3 on top; these data can be an output from the Raster to Point GP tool) and we can also assume that the age of children is, for example, three years, that the net is not used and that the health center is present. Using these data as input, the semiparametric model makes the predictions shown in figure 3 at center, and these predictions are accompanied by the estimated prediction standard errors shown in figure 3 at bottom.

It is advisable to try several other combinations of the variables age, netuse, and presence of the health centers to better understand the predicted risk of the disease. In particular, according to the model, removing nets from those children who used them would result in an increase of the malaria risk.

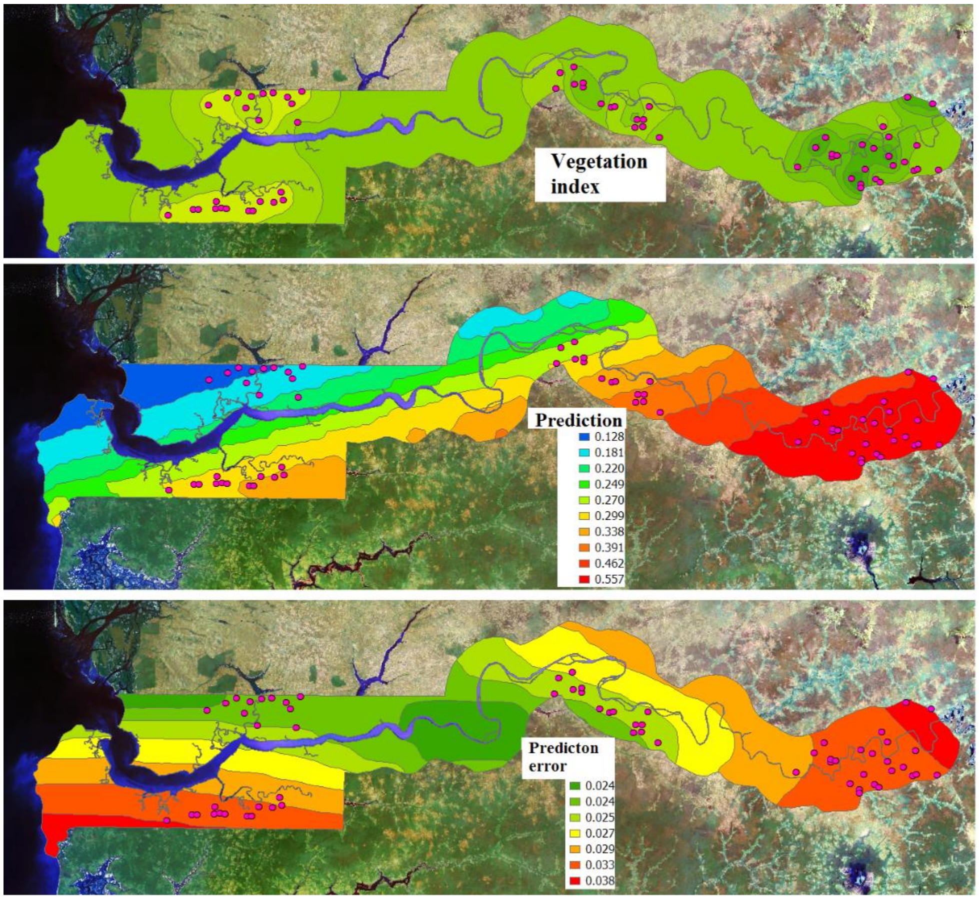


Figure 3. Greenness, malaria probability and its associated prediction standard error.

Some theory

Optimal interpolation of indicator variables is a difficult task, see discussion on indicator kriging in [3], chapters 6 and 9.

In many applications, including environmental monitoring, atmospheric modeling, real estate markets, epidemiology and forestry, several spatially dependent variables are recorded across the region. A proper usage of these additional variables significantly improve spatial predictions. The multivariate geostatistical spatial interpolation model (called cokriging) requires specification of valid and optimal correlation and cross-correlation functions. Fitting reliable cross-correlation functions requires experience in interactive geostatistical modeling. Spatial regression is a popular alternative to the cokriging model.

The difference between cokriging and the spatial regression models is in the way dependence between variables is modeled. The main goal of the spatial regression is to avoid specification of the cross-covariance between variables by modeling the dependence between mean values because it is much easier to estimate regression coefficients and the covariance models, than to assess the cross-covariances, see detailed discussion in [3], chapter 12.

The semiparametric spatial regression is a variant of the generalized additive model. It models the data variation Z_i as a sum of *function* of explanatory variables:

$$Z_i \text{ or } \mathit{function}(Z_i) = s_1(x_1) + s_2(x_2) + \ldots + s_n(x_n) + \varepsilon_i ,$$

where the variable of interest Z_i has a particular statistical distribution and the functions $s_i(x_i)$ are smoothing spline functions for the explanatory variables x_i.

The semiparametric regression can be illustrated using one-dimensional data. Red points in figure 4 display one-hour maximum ozone concentration in Riverside, California, in June 1999. Suppose we want to fit ozone data using a smooth function *$s_1(x)$* as

$$z_i = s_1(x_i) + \varepsilon_i ,$$

where z_i is a response variable (ozone concentration), $s_i(x_i)$ is a smooth function of explanatory variable x_i (day number), and ε_i are independent and identically distributed random errors. This can be done by defining a set of known basis functions *$b_i(x)$* so that *$s_1(x)$* has representation

$$s_1(x) = \sum_i \alpha_i b_i(x) ,$$

where α_i are estimated regression coefficients. A set of four possible basis functions *$b_i(x)$* is shown in figure 1 at bottom. Fitted line is shown in figure 1 on top in blue. According to the linear model diagnostics (R^2 statistics), four basis functions explain 96 percent of the data variation.

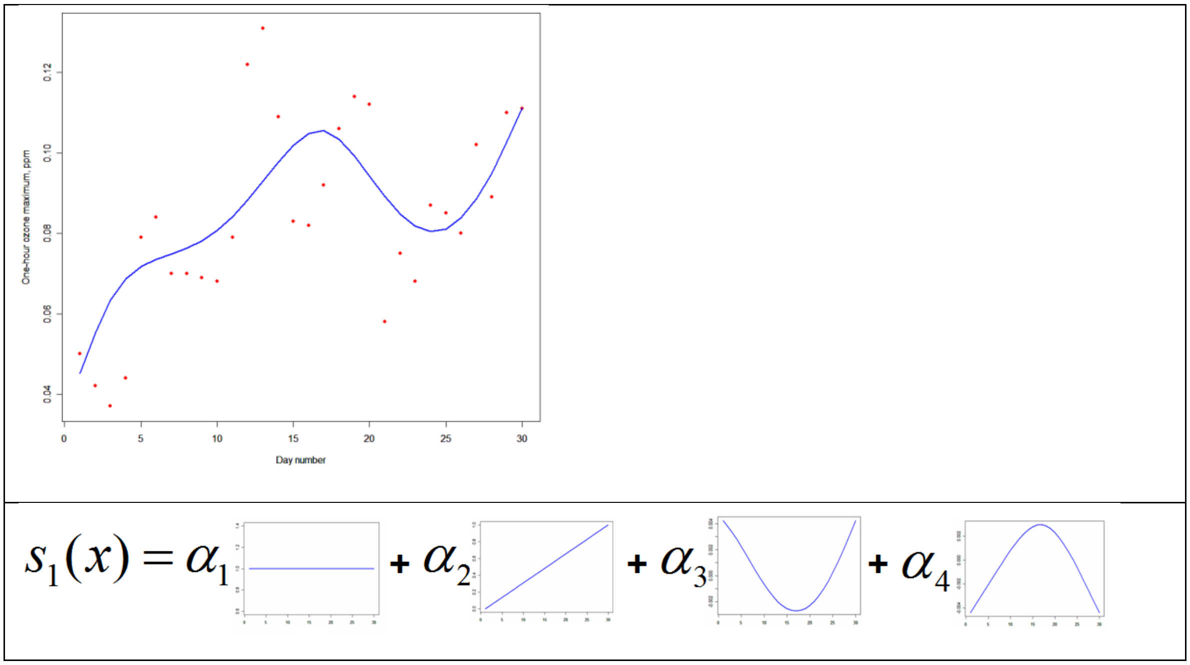


Figure 4. Illustration of the regression approach using splines.

Ozone concentration can be partially explained by the temperature. Therefore, the regression model can be refitted using additional smooth function of the temperature

$$Ozone_i = s_1(day_i) + s_2(temperature_i) + \varepsilon_i$$

and so on. The semiparametric regression allows using both individual-specific and spatial data. In the case of 2D (spatial) explanatory data, smooth function has two arguments: $s_3(x_i, y_i)$ instead of $s_3(x_i)$.

Note that in contrast to cokriging interpolation, the semiparametric regression requires values of the explanatory variables at the predicted locations.

Further reading

[1] Ruppert, D., Wand M. P., and Carroll R. J. (2003) *Semiparametric Regression*. Cambridge University Press, 386 pp.

[2] Wood, S. N. (2006) *Generalized Additive Models: An introduction with R*. CRC Press. 391 pages.

[3] Krivoruchko, K. (2011) *Spatial Statistical Data Analysis for GIS Users*. ESRI Press, 928 pp:

- Semiparametric regression: chapter 12.
- Radial basis functions and kriging: chapter 7.
- Splines and generalized additive model: chapter 6 and appendix 2.
- Radial smoother in SAS: appendix 4.
- Cokriging interpolation: chapters 8 and 9.

[4] Thomson, M., Connor, S., D Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M. & Greenwood, B. (1999). Predicting malaria infection in Gambian children from satellite data and bednet use surveys: the importance of spatial correlation in the interpretation of results. *American Journal of Tropical Medicine and Hygiene* 61: 2–8.