Using R with ArcGIS to model presence/absence data using semiparametric regression with the *SemiPar* package

Before proceeding to the example, you must have the following installed on your computer:

Prerequisites

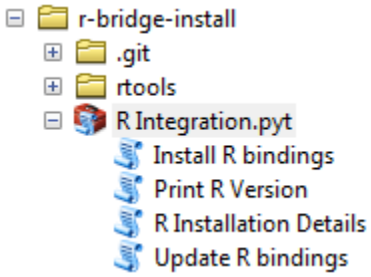
ArcGIS 10.3.1 or ArcGIS Pro 1.1 (don't have it? try a 60 day trial)

1. R Statistical Computing Software, 3.1.0 or later
    - 32-bit version required for ArcMap, 64-bit version required for ArcGIS Pro (Note: the installer installs both by default).
    - 64-bit version can be used with ArcMap by installing Background Geoprocessing and configuring scripts to run in the background.
2. R ArcGIS Bridge

Setup Instructions

ArcGIS 10.3.1

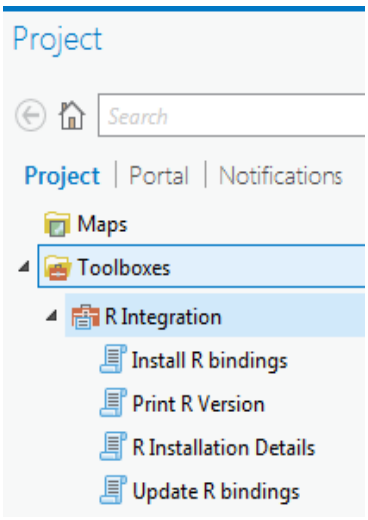
- In the Catalog window, navigate to the folder containing the Python Toolbox, `R Integration.pyt`. *Note*: You may have to first add a folder connection to the location that you extracted the files or downloaded via GitHub.
- Open the toolbox, which should look like this:



- Run the `Install R bindings` script. You can then test that the bridge is able to see your R installation by running the `Print R Version` and `R Installation Details` tools.

ArcGIS Pro 1.1

- In the Project pane, either navigate to a folder connection containing the Python toolbox, or right click on *Toolboxes* > *Add Toolbox* and navigate to the location of the Python toolbox.
- Open the toolbox, which should look like this:



- Run the `Install R bindings` script. You can then test that the bridge is able to see your R installation by running the `Print R Version` and `R Installation Details` tools.

Background information

A common type of data both within and outside the GIS sector is presence/absence data. Most commonly, this type of data is stored as a binary variable that takes only the values 0 and 1. The presence of an event is indicated by a 1, and the absence of an event is indicated by a 0. Common sources of presence/absence data include epidemiology (e.g. whether or not a certain disease is diagnosed), forestry (e.g. whether or not a tree is infested with a particular parasite), and geology (e.g. the presence/absence of mineral deposits at a particular location). Semiparametric regression is one of the methods that can be used to model the presence/absence data, and the model can then be used to make recommendations about how to increase or decrease the prevalence of the event (such as reducing the prevalence of a disease).

The *Semiparametric Regression* tool, which uses the *SemiPar* package [5], models the presence/absence variable using a set of explanatory variables that are known to be related to the probability of presence. It also allows you to separate explanatory variables into two general classes:

1. Explanatory variables that have a linear relationship with the presence/absence data. For these variables, increasing the value of the explanatory variable will either increase or decrease the probability of presence.
2. Explanatory variables that have a nonlinear relationship with the presence/absence data. For these variables, increasing the value of the explanatory variable will elicit a nonlinear response in the presence/absence data. For example, the probability of presence might initially increase, then decrease, then increase again.

The linear explanatory variables are modeled parametrically, and the nonlinear explanatory variables are modeled nonparametrically. Because the regression model uses both parametric and nonparametric components, it is called semiparametric regression.

Explanation of Parameters

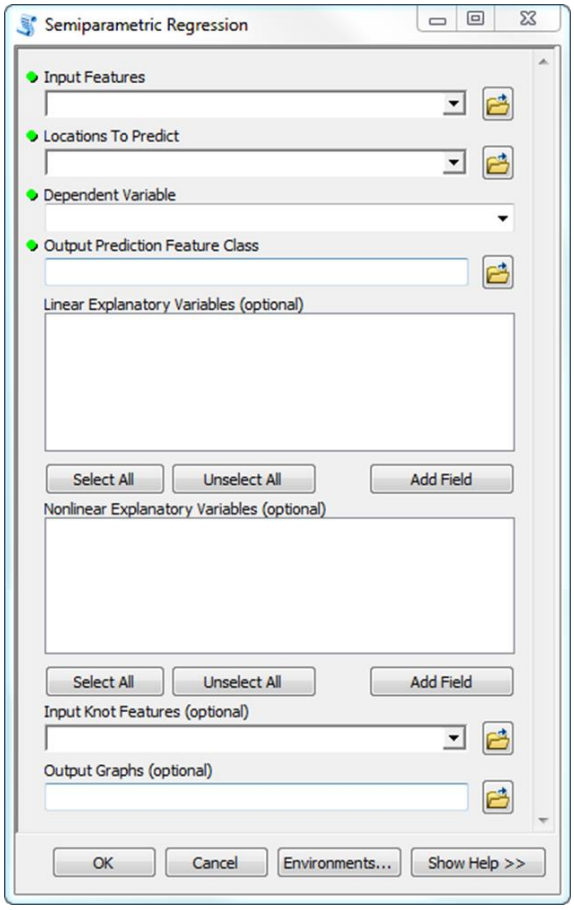


Figure 1. *Semiparametric Regression* tool.

Input Features: Input point features containing fields of the dependent binary variable and all explanatory variables.

Locations To Predict: Input point features representing locations where you would like to predict the probability of presence. These point features must have all explanatory variables stored as fields. The names of these fields must match the names of the explanatory variables in the *Input Features*.

Dependent Variable: Field from the *Input Features* containing the presence/absence variable, which can only take the values 0 and 1. A value of 0 indicates absence, and a value of 1 indicates presence.

Output Prediction Feature Class: Output point feature class that contains the predictions and lower/upper bounds of the 95% confidence interval of the probability of presence. These values are calculated at the positions in the *Locations To Predict* features. In addition to the fields contained in the *Input Features*, the output feature class will also contain the:

1. *prediction* – The predicted probability of presence.
2. *LCL_95* – The lower limit of a 95% confidence interval for the probability of presence.
3. *UCL_95* – The upper limit of a 95% confidence interval for the probability of presence.

Linear Explanatory Variables: Fields from the *Input Features* containing explanatory variables that are linearly related to the *Dependent Variable*.

Nonlinear Explanatory Variables: Fields from the *Input Features* containing explanatory variables that are nonlinearly related to the *Dependent Variable*.

Input Knot Features: Input point features containing 50 to 200 well-distributed locations to connect the piecewise polynomials of the spline basis functions. These points should be selected in such a way that additional knots became unimportant to the final prediction. These knots will be used to interpolate the spatial effect of the semiparametric model. If no knot features are provided, a default knot configuration will be used.

Output Graphs: Creates a PDF containing graphs generated from the *plot* function in the *SemiPar* package [5]. These graphs show both the linear and nonlinear relationships between the explanatory variables and the dependent variable. For nonlinear explanatory variables, a graph of the fitted knots will be displayed. A map of the spatial effect will also be graphed. For more information about these graphs, consult the documentation for the *SemiPar* package [5].

The tool will also display messages that give information about the individual explanatory variables. For linear explanatory variables, the tool will list the estimated coefficient ("coef"), standard error ("se"), likelihood ratio ("ratio"), and p-value. For nonlinear explanatory variables, the messages will list the degrees of freedom ("df"), the smoothing parameter used in the spline ("spar"), and the number of knots that were used ("knots").

Case study

The sample data for this tool involves blood tests for malaria among children in the Republic of The Gambia in Africa [4]. Samples were taken from 2035 children in 65 villages, as shown in Figure 2.

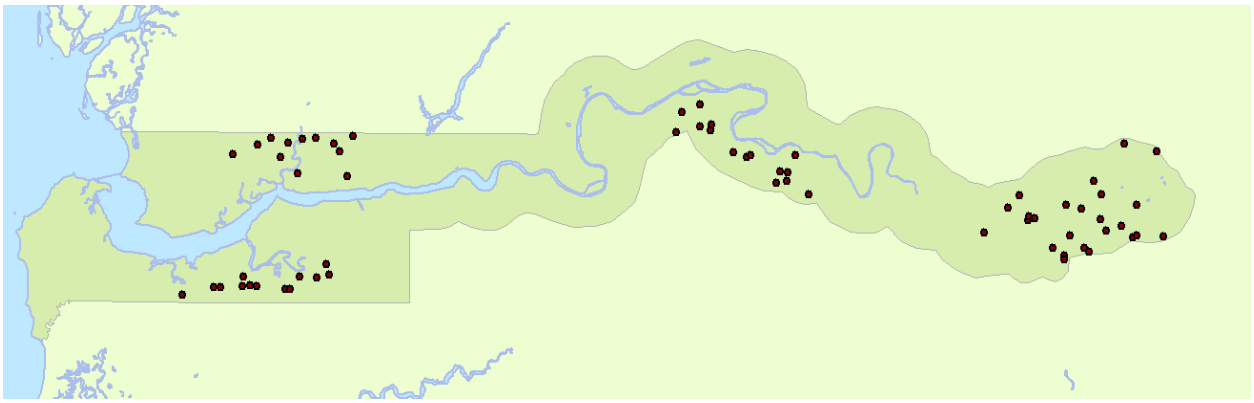


Figure 2. Locations of the 65 villages in The Gambia where children were tested for malaria. Each point represents a village, and many children were sampled from each village.

For the study, the following variables were recorded for each child:

- Presence (1) or absence (0) of malaria in a blood sample taken from a child (36 percent of blood samples tested positive for malaria). In the *gambia_data* feature class provided with this example, this field is named *malaria*.
- Satellite-derived measure of the greenness of vegetation in the immediate vicinity of the village. The greenness of vegetation is calculated from the normalized difference vegetation index (NDVI), and previous epidemiological studies [4] have found that the level of greenness is an important predictor of the number of mosquitos, the transmitter of malaria, and the greenness of vegetation has a

nonlinear relationship with malaria risk. In the *gambia_data* feature class provided with this example, this field is named *green.*

- Age of the child in days (average age is 1080 days, approximately 3 years). In general, older children have higher malaria rates than younger children [4] because older children have had more potential exposures to the malaria virus. In the *gambia_data* feature class provided with this example, this field is named *age.*
- 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). Bed nets are designed to prevent insect bites during sleep, so the use of a bed net is important to limit the number of mosquito bites and, hence, reduce the prevalence of malaria. In the *gambia_data* feature class provided with this example, this field is named *netuse.*
- The presence (1) or absence (0) of a health center in the village. Malaria is a curable disease, so access to health facilities is an important predictor of the prevalence of malaria. In the *gambia_data* feature class provided with this example, this field is named *healthcenter.*

Modeling of the malaria risk is challenging because 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 malaria presence, can be modeled by the semiparametric regression model as

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot netuse + \beta_3 \cdot healthcenter + f_1(green) + f_2(x,y)$$

where β_i are linear coefficients to be estimated, $f_2(x, y)$ is a spatial effect (estimated with 2D splines), and f_1 is a function of greenness that is modeled using 1D splines [1], [2]. Because greenness is modeled using splines, this allows for a nonlinear relationship between the malaria presence and the greenness. To fit this model to the sample data, provide the *gambia_data* feature class as the *Input Features*. The fields *age*, *netuse*, and *healthcenter* should be provided in the *Linear Explanatory Variables*, and *green* should be provided in the *Nonlinear Explanatory Variables*.

The semiparametric model can predict probabilities of malaria presence at any locations where explanatory variables are specified. Therefore, a map of the malaria risk can be created assuming that the age of children, the use of a bed net, the presence of a health center, and the greenness of vegetation are specified. This allows us to answer hypothetical questions such as:

- What would a map of malaria risk look like if every 3 year old (age = 1095 days) did not use a bed net and did not have access to a healthcare facility?

Using this hypothetical data (*gambia_locations*) as the *Locations To Predict*, the semiparametric model predicts the probability of malaria presence in children for this above scenario (Figure 3, top).

By specifying other values of the explanatory variables *age*, *netuse*, and *healthcenter*, other hypothetical scenarios can be tested, such as seeing predictions for 5 year olds that use a bed net but do not have access to a health center. Using these hypothetical situations, we can see, for example, that removing nets from those children who used them would result in an increase of the risk of malaria.

Because all explanatory variables are set to constant values in the *Locations To Predict* in this example (age three years, does not use a bed net, and does not have access to a health center), these point predictions can be interpolated to create a continuous map of the probability of malaria. In other hypothetical situations where the explanatory variables are not set to constant values, it is not valid to interpolate the predicted malaria probabilities. Figure 3 shows the predictions to points (top) and the result of interpolating the predicted probabilities using the Kernel Interpolation With Barriers geoprocessing tool in the Geostatistical Analyst toolbox (bottom).

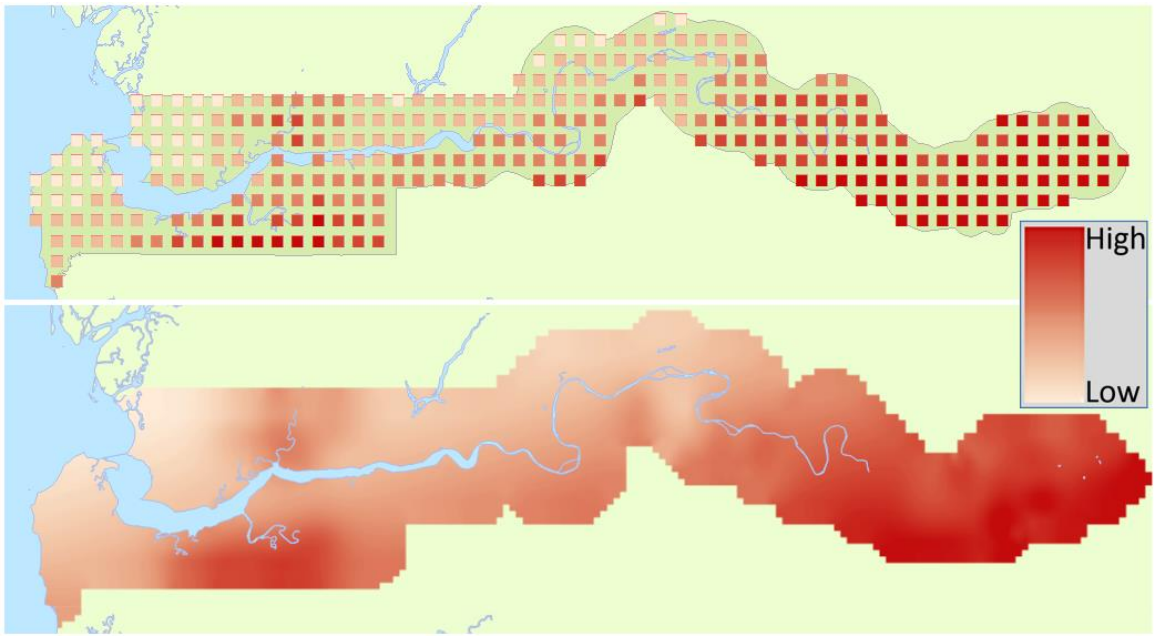


Figure 3. Predicted malaria probability at point locations (top) and as a raster (bottom)

References and 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.

[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.

[5] Matt Wand (2014). SemiPar: Semiparametic Regression. R package version 1.0-4.1. http://CRAN.R-project.org/package=SemiPar