Estimating the number of cases of podoconiosis in Ethiopia using geostatistical methods
Supplementary material - Geostatistical analysis

December 11, 2017

Covariates

Seven covariates were used to model the spatial variation in podoconiosis prevalence: elevation and derived slope, long-term average of precipitation, enhanced vegetation index (EVI), clay and silt content at the top soil (0-15 cm) and night light-emissivity. Each of these is displayed in Figure 1.

Model formulation

Let $Y_i$ denote the random variable associated with the number of podoconiosis cases, out of $n_i$ individuals, sampled at a village location $x_i$ in Ethiopia. We assume that conditionally on spatial random effects, $S(x_i)$, and Gaussian noise, $Z(x_i)$, the $Y_i$ are mutually independent binomial with probability of a positive test for podoconiosis, $p_i$. The linear predictor is

$$
\log \left\{ \frac{p_i}{1 - p_i} \right\} = d(x_i) \top \beta + S(x_i) + Z(x_i),
$$

where $d(x_i)$ is a vector of explanatory variables with associated vectors of regression coefficients $\beta$. A plot of the logit-transformed prevalence against elevation, shows a positive association up to about 2000 meters and negative thereafter (Figure 2). For this reason elevation was included in the model using a quadratic polynomial. The spatial random effects $S(x)$ can be interpreted as the cumulative effect of unmeasured risk factors for podoconiosis. Finally, the unstructured random effects $Z(x)$ represents extra-binomial variation within villages.

We assume that $S(x)$ is a stationary and isotropic Guassian process with mean zero and covariance function given by

$$
cov\{S(x), S(x')\} = \sigma^2 \exp\{-\|x - x'\|/\phi\},
$$
where $\|x - x'\|$ is the Euclidean distance between any two locations $x$ and $x'$.

We use $\tau^2$ to denote the variance of the Gaussian noise $Z(x)$.

We use Monte Carlo maximum likelihood (MCML) (Geyer & Thompson, 1992; Geyer, 1994, 1996, 1999) to obtain point estimates of $\beta, \sigma^2, \phi$ and $\tau^2$.

Table 1 reports the (MCML) and the corresponding 95% confidence intervals.

Table 1: Monte Carlo maximum likelihood estimates with association 95% confidence intervals (CI).

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-13.931</td>
<td>(-14.944, -12.917)</td>
</tr>
<tr>
<td>Precipitation $\times 10^3$</td>
<td>1.268</td>
<td>(0.475, 2.060)</td>
</tr>
<tr>
<td>Night emissivity lights</td>
<td>-0.006</td>
<td>(-0.019, 0.006)</td>
</tr>
<tr>
<td>Slope</td>
<td>-0.024</td>
<td>(-0.051, 0.002)</td>
</tr>
<tr>
<td>EVI</td>
<td>0.871</td>
<td>(-0.504, 2.245)</td>
</tr>
<tr>
<td>Elevation $\times 10^3$ (linear term)</td>
<td>7.376</td>
<td>(6.271, 8.481)</td>
</tr>
<tr>
<td>Elevation $\times 10^3$ (quadratic term)</td>
<td>-1.999</td>
<td>(-2.261, -1.738)</td>
</tr>
<tr>
<td>Silt</td>
<td>0.035</td>
<td>(0.017, 0.052)</td>
</tr>
<tr>
<td>Clay</td>
<td>0.003</td>
<td>(-0.003, 0.009)</td>
</tr>
<tr>
<td>$\log(\sigma^2)$</td>
<td>0.979</td>
<td>(0.774, 1.185)</td>
</tr>
<tr>
<td>$\log(\phi)$</td>
<td>3.928</td>
<td>(3.657, 4.199)</td>
</tr>
<tr>
<td>$\log(\tau^2)$</td>
<td>-0.857</td>
<td>(-1.325, -0.390)</td>
</tr>
</tbody>
</table>

**Model validation**

We check the validity of the assumed covariance model for the spatial correlation using the following Monte Carlo algorithm.

1. Simulate $S(x_i)$ and $Z(x_i)$ under the fitted model at each of the sampled village locations $x_i$.

2. Simulate binomial data $y_i$ based on (1).

3. Fit a standard logistic regression (i.e. $S(x_i) = Z(x_i) = 0$, for all $x_i$) to the simulated data $y_i$ using explanatory variables $d(x_i)$.

4. Obtain the Pearson’s residuals from the standard logistic regression of the previous step and compute the empirical semi-variogram.

5. Repeat steps 1 to 4 for 10,000 times.

6. Use the resulting 10,000 empirical semi-variograms to compute 95% tolerance intervals at each distance bin.
7. Compute the empirical semi-variogram using the residuals of a standard logistic regression as in step 3, for the observed data.

8. If the empirical semi-variogram from step 7 falls inside the 95% tolerance intervals, we conclude that the adopted covariance function is compatible with data. If, instead, the empirical semi-variogram from step 7 falls outside the 95% tolerance intervals, we conclude that the assumed covariance function is not compatible with the data.

Figure 3 shows the results of the outlined validation procedure. Since the empirical semi-variogram (solid line) falls within the 95% tolerance intervals (dashed lines), we then conclude that the adopted covariance model is compatible with the data.

References


Figure 1: Spatial covariates used to model podocniosis prevalence.
Figure 2: Plot of the logit-transformed prevalence against elevation. The red line is obtained using locally weighted smoothing.
Figure 3: Results from the model validation. The dashed lines represent the 95% tolerance intervals from step 6. The solid line corresponds to the observed semi-variogram of step 7.