Geostatistical design for optimal sampling of spatially correlated environmental variables using spatial mixture copulas

Mohomed Abraj, You-Gan Wang, and M. Helen Thompson.

Spatial vine copula

A spatial vine copula model can be derived as a conditional function of the spatial mixture copula to interpolate values at the candidate (unsampled) locations. The vine copula consists flexible class of bivariate building blocks of copulas, which is an approximation for I-dimensional multivariate copula. That is, a I-dimensional spatial vine copula can be constructed with all $x_i \in X$, for $i = 1, 2, \ldots, I$. Interpolation at candidate location x'_0 can be achieved using a spatial vine copula and the four nearest neighbouring sampled locations x_1, x_2, x_3 and x_4. Theoretically, all sampled locations can be used as neighbouring locations in the spatial vine copula, however, practical application is limited by computation. Here, four nearby neighbours are used. Figure S.1 shows the spatial vine copula decomposition of the spatial mixture copula C_h.

In Fig. S.1, an edge represents the spatial mixture copula density c_h, and the two nodes linked to each edge represent mixture copulas and conditional mixture copulas. The conditional mixture copula $C_{i|v}(\mathbf{Z}(x_i))$ is given by

$$C_{i|v}(\mathbf{Z}(x_i)) = \frac{\partial C_{ij|v-j}(C_{j|v-j}(\mathbf{Z}(x_i)), C_{j|v-j}(\mathbf{Z}(x_j)))}{\partial C_{j|v-j}(\mathbf{Z}(x_j))},$$

where v is the indices of conditioning variables and $v - j$ is excluding j^{th} index.

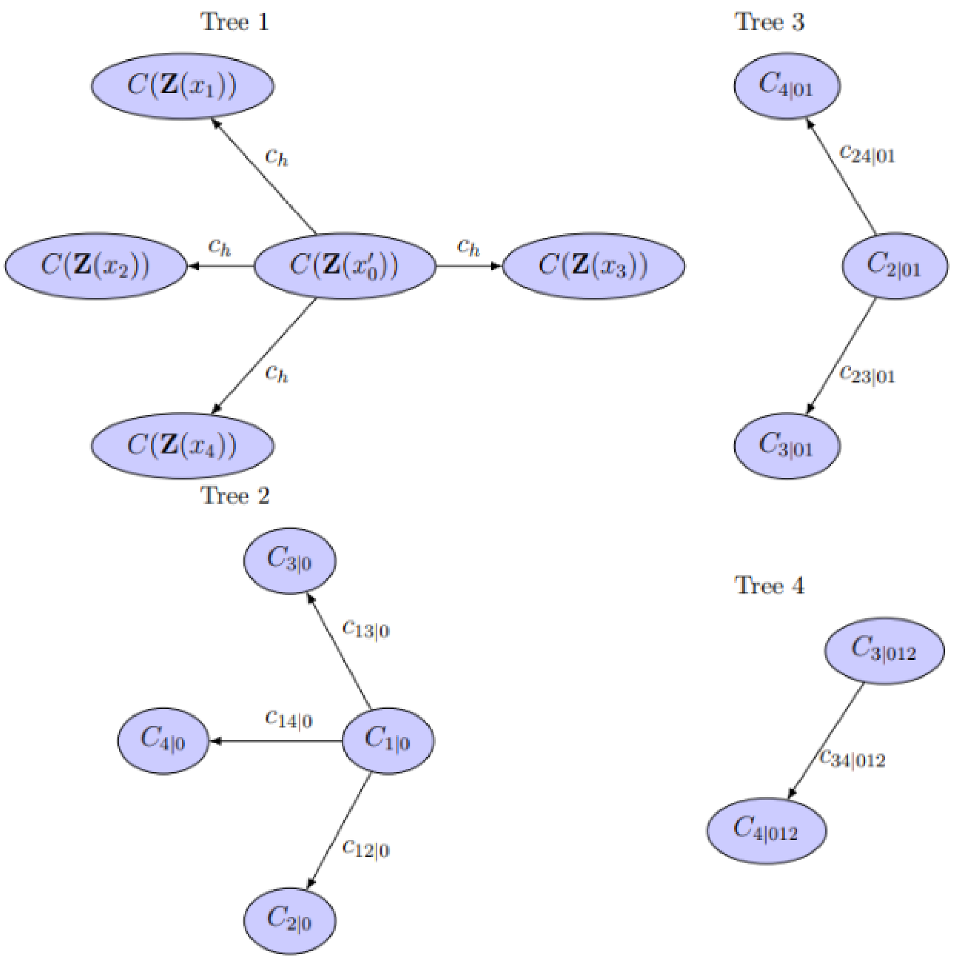

Figure S.1: Five-dimensional spatial vine copula decomposition of C_h.

Let $t_i = C(\mathbf{Z}(x_i))$ and $t'_0 = C(\mathbf{Z}(x'_0))$. Using the decomposition structure in Fig. S.1, the spatial vine copula density is given by

$$c(t'_0, t_1, t_2, t_3, t_4) = \\ c_h[C(\mathbf{Z}(x'_0)), C(\mathbf{Z}(x_1))] \times c_h[C(\mathbf{Z}(x'_0)), C(\mathbf{Z}(x_2))]\times \\ c_h[C(\mathbf{Z}(x'_0)), C(\mathbf{Z}(x_3))] \times c_h[C(\mathbf{Z}(x'_0)), C(\mathbf{Z}(x_4))]... \quad \text{(Tree 1)}$$

$$\times c_{12|0}[C_{1|0}(\mathbf{Z}(x_1)|\mathbf{Z}(x'_0)), C_{2|0}(\mathbf{Z}(x_2)|\mathbf{Z}(x'_0)))]\times \\ c_{13|0}[C_{1|0}(\mathbf{Z}(x_1)|\mathbf{Z}(x_0)), C_{3|0}(\mathbf{Z}(x_3)|\mathbf{Z}(x_0))]\times \\ c_{14|0}[C_{1|0}(\mathbf{Z}(x_1)|\mathbf{Z}(x_0)), C_{4|0}(\mathbf{Z}(x_4)|\mathbf{Z}(x'_0))]... \quad \text{(Tree 2)}$$

$$\times c_{23|01}[C_{2|01}(\mathbf{Z}(x_2)|\mathbf{Z}(x'_0), \mathbf{Z}(x_1)), C_{3|01}(\mathbf{Z}(x_3)|\mathbf{Z}(x'_0), \mathbf{Z}(x_1))]\times \\ c_{24|01}[C_{2|01}(\mathbf{Z}(x_2)|\mathbf{Z}(x'_0), \mathbf{Z}(x_1)), C_{4|01}(\mathbf{Z}(x_4)|\mathbf{Z}(x'_0), \mathbf{Z}(x_1))]... \quad \text{(Tree 3)}$$

$$\times c_{34|012}[C_{3|012}(\mathbf{Z}(x_3)|\mathbf{Z}(x'_0), \mathbf{Z}(x_1), \mathbf{Z}(x_2)), C_{4|012}(\mathbf{Z}(x_4)|\mathbf{Z}(x'_0), \mathbf{Z}(x_1), \mathbf{Z}(x_2))]. \quad \text{(Tree 4)}$$

The conditional density of t'_0 at candidate location x'_0 can then be calculated as

$$c(t'_0|t_1, t_2, t_3, t_4) = \frac{c(t'_0, t_1, t_2, t_3, t_4)}{\int_0^1 c(v, t_1, t_2, t_3, t_4)dv}.$$

The conditional cumulative probability at a candidate (or an unsampled) location is then calculated as

$$\hat{\mathbf{Z}}(x'_0) = C_{h,x'}(t'|t_1, t_2, t_3, t_4) = \int_0^1 c(t'|t_1, t_2, t_3, t_4)dt'.$$

Any desired quantile value can be obtained from above equation for quantile q at the candidate (or unsampled) locations.

Inverse conditional copula approach

In step 4 of the simulation study, in a bivariate context, Z_1 and Z_2 values were obtained by following inverse conditional copula.

Step i: Obtain the joint CDF values of $\mathbf{Z}$ using C_h, and let $\mathbf{T}$ be the vector of the joint CDF values at the removed locations.

Step ii: Obtain the marginal CDF values of Z_2 using F_2, and let R be the vector of the marginal CDF values of Z_2 at the removed locations.

Step iii: Derive the conditional distribution of $\mathbf{T}$, given $R = r$, using the partial derivative of C_h as follows,

$$C_{h,r}(\mathbf{T}|R = r) = P[\mathbf{T} \leq \mathbf{t}|R = r], \\ = \frac{\partial}{\partial r} C_h(\mathbf{t}, r),$$

where $r_1 = (C_{h,r})^{-1}(\mathbf{T}|R = r_1)$ or $r_2 = (C_{h,r})^{-1}(\mathbf{T}|R = r_2)$ if Z_1 is used in step ii.

Step iv: Obtain, $Z_1 = F_1^{-1}(r_1)$, similarly $Z_2 = F_2^{-1}(r_2)$, which are the redesigned simulated values at the removed locations.