How to Predict Nesting Sites and How to Measure Shoreline Erosion: Fuzzy and Probabilistic Techniques for Environment-Related Spatial Data Processing

Stephen M. Escarzaga, Craig Tweedie,
Olga Kosheleva, and Vladik Kreinovich
University of Texas at El Paso, El Paso, TX 79968, USA
smescarzaga@utep.edu, ctweedie@utep.edu,
olgak@utep.edu, vladik@utep.edu

1. Importance of Environment-Related Spatial Data Processing

- When analyzing the ecological systems, it is important to study:
  - the spatial environment of these systems, and
  - spatial distribution of the corresponding species in this spatial environment.
- In most locations within an ecological zone, the environmental changes are reasonably slow.
- It usually takes decades to see a drastic change.
- However, at the borders between different ecological zones, the changes are much faster.

2. Importance of Studying Shorelines

- In the border between different types of plants the changes are fast but still gradual:
  - new types of plants appear; their proportion grows, and eventually,
  - they take over the area.
- However, there are border areas where the change is the most drastic: namely, the shorelines.
- The shorelines are, in most places, retreating because of the shoreline erosion.
- The overall area of the shorelines is relatively small.
- However, they are a habitat for many species, from birds (like seagulls) to turtles.
- From this viewpoint, it is important to be able to trace and measure shoreline erosion.

3. Studying Spatial Distribution of Different Species

- It is important to trace and measure spatial environments which are important for different species.
- It is also necessary to trace spatial location of these species.
- This problem is especially important for rare birds.
- Birds are most vulnerable when they at their nesting sites.
- It is therefore important to monitor these sites.
- Some species use the same nesting sites year after year.
- Birds from other species vary their sites each year.
- To be able to monitor birds from these species, it is important to be able to predict their nesting sites.

4. Predicting Nesting Sites: Formulation of the Problem

- We observe nesting sites for a certain bird species.
- Our goals are:
  - to analyze which criteria are important for selecting nesting sites, and
  - to come up with formulas that would enable us to predict nesting sites.
- Let $v_1, \ldots, v_n$ be parameters that may influence the selection of a nesting site.
- *Examples:* parameters describing elevation, hydrology, vegetation level, distance form other nesting sites, etc.
- For each geographical location x, we record the values of these parameters $v_1(x), \ldots, v_n(x)$.

5. Formulation of the Problem (cont-d)

- We assume that the birds select a nesting site based on the values of these quantities (at least some of them).
- So, a bird tries to maximize the value of some objective function $F(v_1, \ldots, v_n)$ depending on v_i.
- We do not know the exact form of the dependence $F(v_1, \ldots, v_n)$.
- However, we can expand F in Taylor series and keep the first few terms up in this expansion.
- If we only keep linear terms, we get:

$$F(v_1, \ldots, v_n) = a_0 + \sum_{i=1}^{n} a_i \cdot v_i.$$

- If we also keep quadratic terms, we get:

$$F(v_1, \ldots, v_n) = a_0 + \sum_{i=1}^{n} a_i \cdot v_i + \sum_{i=1}^{n} \sum_{\ell=1}^{n} a_{i\ell} \cdot v_i \cdot v_\ell,$$

6. Formulation of the Problem (cont-d)

- For each of these approximations, the (unknown) objective function has the form

$$F(v_1, \ldots, v_n) = \sum_{j=1}^{N} A_j \cdot V_j(x), \text{ where:}$$

  - $V_j(x)$ are known values (e.g., $v_i(x)$ and $v_i(x) \cdot v_\ell(x)$);
  - A_j are the coefficients that need to be determined.
- We assume that each year, each of the observed nesting sites x_k:
  - has the largest possible value of the objective function
  - in the set C_k of all locations x which are closer to x_k that to any other nesting locations.
- Under this assumption, we want to find $A_1, \ldots, A_N$ that best explain the observations.

7. Analysis of the Problem

- The fact that on the cell C_j, the linear function (2.1) attains its largest value at the site x_j means that

$$\sum_{j=1}^{N} A_j \cdot V_j(x_k) \geq \sum_{j=1}^{N} A_j \cdot V_j(x) \text{ for all } x \in C_k.$$

- In other words, we should have

$$A \cdot \Delta(x) \stackrel{\text{def}}{=} \sum_{j=1}^{N} A_j \cdot \Delta_j(x_k) \geq 0, \text{ where}$$

$$A \stackrel{\text{def}}{=} (A_1, \ldots, A_n), \quad \Delta(x) \stackrel{\text{def}}{=} (\Delta_1(x), \ldots, \Delta_N(x)),$$

$$\text{and } \Delta_j(x) \stackrel{\text{def}}{=} V_j(x_k) - V_j(x).$$

- Similarly, we should have $A \cdot (-\Delta(x)) \leq 0$ for all x.

8. How Can We Solve This Problem?

- From the mathematical viewpoint, this problem is similar to the *linear discriminant analysis*, when:
  - we have two sets $\mathcal{S}$ and $\mathcal{S}'$ and
  - we need to find a hyperplane that separates them, i.e., a vector A such that $A \cdot S \geq 0$ for all $S \in \mathcal{S}$ and $A \cdot S' \leq 0$ for all $S' \in \mathcal{S}'$.
- In our case, S is the set of all vectors $\Delta_j(x)$, and S' is the set of all vectors $-\Delta_j(x)$.
- The standard way of solving this problem is to compute:
  - the mean μ of all M vectors $S \in \mathcal{S}$,
  - the covariance matrix Σ, and
  - then to take $A = \Sigma^{-1}\mu$.

9. How to Solve the Problem

- In our case, we should do the following:
  - compute all M vectors $\Delta(x)$ with components $\Delta_j(x) = V_j(x_k) - V_j(x)$, where $x \in C_k$;
  - compute the average $\mu = \frac{1}{M} \cdot \sum_x \Delta(x)$;
  - compute the corresponding covariance matrix:

$$\Sigma_{ab} = \frac{1}{M} \cdot \sum_x (\Delta_a(x) - \mu_a) \cdot (\Delta_b(x) - \mu_b);$$

  - compute the desired weights as $A = \Sigma^{-1}\mu$.
- We can predict the nesting locations as the points x at which $\sum_{i=1}^{N} A_j \cdot V_j(x)$ is the largest.
- Instead of the above probabilistic clustering, we can use *fuzzy clustering*.

10. How Can We Gauge the Accuracy of the Resulting Estimate

- To gauge the accuracy of this prediction, we can test it against the observed data.
- For each cell C_k, we compute the location c_k at which $F = \sum_{i=1}^{N} A_j \cdot V_j(x)$ is the largest in this cell.
- As a natural measure of prediction accuracy, we can take the mean square distance between:
  - these predicted nesting sites c_k and
  - the actual nesting sites x_k.

11. How to Measure Shoreline Erosion: Formulation of the Problem

- A natural way to measure erosion is to divide:
  - the difference between the observed shoreline locations at two different years by
  - the number of years between the two observations.
- In practice, observers in different years follow slightly different lines when making their measurement.
- Examples: lines at a certain distance from water, or at a certain elevation above water, etc.
- This fact changes the the computed ratio.
- The change can be so large that in the areas with known erosion, the computed ratio becomes negative.
- It is desirable to take this uncertainty into account.

12. How to Take This Uncertainty Into Account: First Approximation

- It is usually assumed that within a few-years period, the rate r of erosion practically does not change, so:

$$x_{t+i} = x_t + i \cdot r.$$

- Due to the observation error ε_t, we have:

$$\widetilde{x}_{t+i} = x_{t+i} + \varepsilon_{t+i} = x_t + i \cdot r + \varepsilon_{t+i}.$$

- We can thus use Least Squares to find r:

$$\sum_{i=0}^{T} (\widetilde{x}_{t+i} - (x_t + i \cdot r))^2 \to \min_{x_t, r}.$$

- Minimization leads to

$$r = \frac{12}{T \cdot (T+1) \cdot (T+2)} \cdot \sum_{i=0}^{T} \left(\left(i - \frac{T}{2} \right) \cdot \widetilde{x}_{t+i} \right).$$

13. How Accurate Is This Estimate?

- From Least Squares, we get $x_t = \frac{1}{T+1} \cdot \sum_{i=0}^{T} \widetilde{x}_{t+i} - r$.
- We can estimate the standard deviation σ of the measurement error ε_{t+i} as:

$$\sigma^2 = \frac{1}{T+1} \cdot \sum_{i=0}^{T} (\widetilde{x}_{t+i} - (x_t + r \cdot i))^2.$$

- In practice, we have several measurements at different spatial locations k, with results $\widetilde{X}_{t,k}$.
- So, to find σ, we should also average over all K locations:

$$\sigma^2 = \frac{1}{L} \cdot \frac{1}{T+1} \cdot \sum_{k=1}^{K} \sum_{i=0}^{T} (\widetilde{x}_{t+i,k} - (x_{t,k} + r_k \cdot i))^2.$$

14. Case of $T = 2$

- In practice, we often have three consequent years of observation x_t, x_{t+1}, and x_{t+2}, i.e., we have $T = 2$.
- In this case:

$$r = \frac{\widetilde{x}_{t+2} - \widetilde{x}_t}{2}.$$

- For $T = 2$, we have:

$$\sigma^2 = \frac{1}{18} \cdot \frac{1}{K} \cdot \sum_{k=1}^{K} (\widetilde{x}_{t,k} - 2\widetilde{x}_{t+1,k} + \widetilde{x}_{t+2,k})^2.$$

15. What If Positive Erosion Values Are Not Always Within 2-Sigma Range?

- The estimated erosion rate r may be negative.
- This is OK if within the 2-sigma interval $[r-2\sigma, r+2\sigma]$, we have a positive value, i.e., if $r + 2\sigma \geq 0$; then:
  - the difference between r and the actual (positive) erosion rate
  - can be explained by the observation uncertainty.
- But what if $r + 2\sigma < 0$?
- This would mean that there is an additional source of error: $\widetilde{x}_{t+i} = x_t + i \cdot r + \varepsilon_{t+i} + \delta_{t+i}$.
- In this case, we still determine our estimates x_t and r from the least squares method.
- However, now, we have an additional source of error, with some standard deviation σ_δ^2.

16. How Accurate Is This Estimate?

- We want to estimate σ_δ^2 and the overall variance $\sigma_t^2 = \sigma^2 + \sigma_\delta^2$.
- For a normal distribution, 95% of the values are within 2 sigma interval.
- So, for 95% of the estimated erosion values r_k, we should have $r_k + 2\sigma_t \geq 0$, i.e., equivalently, $2\sigma_t \geq -r_k$.
- Let us sort the estimates r_k: $r_1 < r_2 < \ldots < r_N$.
- The desired inequality should be satisfied for all $k \geq 0.05 \cdot N$, so $2\sigma_t \geq -r_{0.05\cdot N}$ and $\sigma_t \geq -\frac{1}{2} \cdot r_{0.05\cdot N}$.
- We would like to have the narrowest error bounds.
- So, we choose the smallest $\sigma_t \geq \sigma$ that satisfies this inequality: $\sigma_t = \max\left(\sigma, -\frac{1}{2} \cdot r_{0.05\cdot N}\right)$.

17. Resulting Algorithm: Case of General T

- We start with measurements $\widetilde{x}_{t+i,k}$ make at spatial locations k at years t, $t+1$, …, $t+T$. We compute:

$$r_k = \frac{12}{T\cdot(T+1)\cdot(T+2)}\cdot\sum_{i=0}^{T}\left(\left(i-\frac{T}{2}\right)\cdot\widetilde{x}_{t+i,k}\right);$$

$$x_{t,k} = \frac{1}{T\cdot(T+1)\cdot(T+2)}\cdot\sum_{i=0}^{T}(T\cdot(T+8)-12\cdot i)\cdot\widetilde{x}_{t+i,k};$$

$$\sigma^2 = \frac{1}{L}\cdot\frac{1}{T+1}\cdot\sum_{k=1}^{K}\sum_{i=0}^{T}(\widetilde{x}_{t+i,k}-(x_{t,k}+r_k\cdot i))^2.$$

- We then sort the estimated erosion rates in increasing order: $r_1 < r_2 < \ldots < r_N$.
- The mean square accuracy of the erosion rate estimates r_k is then computed as $\sigma_t = \max\left(\sigma, -\frac{1}{2}\cdot r_{0.05\cdot N}\right)$.

18. Resulting Algorithm: Case $T = 2$

- For the case $T = 2$, we have simplified formulas:

$$r_k = \frac{\widetilde{x}_{t+2,k} - \widetilde{x}_{t,k}}{2} \text{ and}$$

$$\sigma^2 = \frac{1}{18} \cdot \frac{1}{K} \cdot \sum_{k=1}^{K} (\widetilde{x}_{t,k} - 2\widetilde{x}_{t+1,k} + \widetilde{x}_{t+2,k})^2.$$

- We then sort the estimated erosion rates in increasing order: $r_1 < r_2 < \ldots < r_N$.
- The mean square accuracy of the erosion rate estimates r_k is then computed as $\sigma_t = \max\left(\sigma, -\frac{1}{2} \cdot r_{0.05 \cdot N}\right)$.

19. Acknowledgment

This work was supported in part:

- by the National Science Foundation grants:
  - HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
  - DUE-0926721,
- and by an award from Prudential Foundation.