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

Stephen M. Escarzaga[1,2], Craig Tweedie[1,2], Olga Kosheleva[3], and Vladik Kreinovich[2]
[1]Environmental Science Program, [2]Cyber-ShARE Center
[3]Department of Electrical and Computer Engineering
University of Texas at El Paso, El Paso, TX 79968, USA
smescarzaga@utep.edu, ctweedie@utep.edu, olgak@utep.edu, vladik@utep.edu

Abstract—**In this paper, we show how fuzzy and probabilistic techniques can be used in environment-related data processing. Specifically, we will show that these methods help in solving two environment-related problems: how to predict the birds' nesting sites and how to measure shoreline erosion.**

I. Formulation of the Problem: Importance of Environment-Related Spatial Data Processing

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; see, e.g., [1], [3], [7].

Studying spatial environment: the importance of studying shorelines. 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. 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.

While the overall area of the shorelines is reasonably small in comparison with the areas of the land and the sea areas, shorelines play a large role in ecological systems, since they are a habitat for many species, from birds (like seagulls) to turtles to numerous other creatures.

From this viewpoint, it is important to be able to trace and measure shoreline erosion.

Studying spatial distribution of different species. In addition to tracing and measuring 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, but birds from other species vary their sites each year. To be able to monitor birds from these species, it is therefore important to be able to predict their nesting sites.

What we do in this paper. In this paper, we show that fuzzy and probabilistic techniques can help in solving these two environment-related spatial data processing problems.

II. How to Predict Nesting Sites?

Formulation of the environmental 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.

Reformulating this problem in precise terms. Let $v_1, \ldots, v_n$ be parameters that may influence the selection of a nesting site: e.g., parameters describing elevation, hydrology, vegetation level, distance form other nesting sites, etc. For each geographical location x, we record the values of these parameters $v_1(x), \ldots, v_n(x)$.

We assume that the birds select a nesting site based on the values of these quantities (at least some of them). In general, this means that a bird tries to maximize the value of some objective function $F(v_1, \ldots, v_n)$ depending on these values v_i.

We do not know the exact form of the dependence $F(v_1, \ldots, v_n)$. However, we can always expand this dependence in Taylor series and keep only terms up to a certain order in this expansion. For example, if we only keep linear terms, this means that we consider objective functions of the type

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

for some to-be-determined coefficients a_i. If we also keep quadratic terms, this means that we consider objective functions of the type

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

etc. The more terms we keep, the more accurately we describe the objective function and thus, the more accurately we predict the nesting sites.

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

$$F(v_1,\ldots,v_n)=\sum_{j=1}^{N} A_j\cdot V_j(x), \tag{2.1}$$

where $V_j(x)$ are known values (e.g., $v_i(x)$ and $v_i(x)\cdot v_\ell(x)$) and 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 among all locations within the corresponding *Voronoi cell* C_k – i.e., among all locations x which are closer to x_k that to any other nesting locations. Under this assumption, we would like to find the weights $A_1,\ldots,A_N$ that best explain the observed nesting sites.

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

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

In other words, we should have

$$A\cdot\Delta(x)\stackrel{\text{def}}{=}\sum_{j=1}^{N} A_j\cdot\Delta_j(x_k)\ge 0 \tag{2.2}$$

where we denoted

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

and $\Delta_j(x)\stackrel{\text{def}}{=}V_j(x_k)-V_j(x)$. Similarly, we should have $A\cdot(-\Delta(x))\le 0$ for all x.

How can we solve this problem? From the mathematical viewpoint, this problem is similar to the *linear discriminant analysis* (see, e.g., [2]), 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 $A\cdot S\ge 0$ for all $S\in\mathcal{S}$ and $A\cdot S'\le 0$ 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 the vectors $S\in\mathcal{S}$, the covariance matrix Σ, and then to take $A=\Sigma^{-1}\mu$. So, in our case, we should do the following:

- compute all the vectors $\Delta(x)$ with components $\Delta_j(x)=V_j(x_k)-V_j(x)$, where $x\in C_k$; let M be the total number of such vectors;
- compute the average $\mu=\dfrac{1}{M}\cdot\sum\limits_x V(x)$ of these vectors;
- compute the corresponding covariance matrix Σ with components

$$\Sigma_{ab}=\frac{1}{M}\cdot\sum_x(V_a(x)-\mu_a)\cdot(V_b(x)-\mu_b); \tag{2.3}$$

- compute the desired weights as $A=\Sigma^{-1}\mu$, i.e., as a solution to a linear system $\Sigma A=\mu$.

The above procedure is equivalent to using probabilistic clustering of the vectors $V_j(x)$ and $-V_j(x)$, i.e., clustering based on probabilistic ideas (see, e.g., [6]). Alternatively, we can use *fuzzy clustering* techniques, i.e., clustering based on using fuzzy ideas (see, e.g., [4], [5], [8]).

Once we know the coefficients A_j, we can use the objective function (2.1) to predict the nesting locations as the points x at which the objective function $\sum\limits_{i=1}^{N} A_j\cdot V_j(x)$ attains a local maximum.

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. Specifically, for each cell C_k, we compute the location c_k at which the weighted combination $\sum\limits_{i=1}^{N} A_j\cdot V_j(x)$ attains its maximum on this cell. The mean square distance between these predicted nesting sites c_k and the actual nesting sites x_k can serve as a natural measure of prediction accuracy.

III. How to Measure Shoreline Erosion?

Formulation of the problem. Many coastal areas are affected by erosion, the sea expands and the shore retreats. A natural way to measure erosion is to observe the shoreline year after year.

In principle, the rate of erosion in each location can be determined as follows: we compute the difference between the observed shoreline locations at two different years, and divide this difference by the number of years between the two observations. In practice, however, observers in different years follow slightly different lines when making their measurement: e.g., lines at a certain distance from water, or at a certain elevation above water, etc. This fact changes the difference between observations and thus, the computed ratio is, in general, different from the actual erosion rate. The difference can be so large that in the areas with known erosion, the computed ratio becomes negative – erroneously indicating the sea retreat.

In short, we have an additional measurement uncertainty. It is desirable to take this uncertainty into account.

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, if we perform observations at years t, $t+1$, ..., $t+T$, then we expect the observed coordinates x_t, x_{y+1}, ..., of the shoreline take the form

$$x_{t+i}=x_t+i\cdot r. \tag{3.1}$$

Due to the presence of the above-mentioned observation error ε, the observed coordinate $\widetilde{x}_{t+i}$ has the form

$$\widetilde{x}_{t+i} = x_{t+i} + \varepsilon_{t+i} = x_t + i \cdot r + \varepsilon_{t+i}. \tag{3.2}$$

It is therefore reasonable to use the use the usual Least Squares techniques to estimate the erosion rate r, i.e., to find r as the value corresponding to the following optimization problem:

$$\sum_{i=0}^{T} (\widetilde{x}_{t+i} - (x_t + i \cdot r))^2 \to \min_{x_t, r}. \tag{3.3}$$

How can we solve the corresponding minimization problem? Differentiating the expression (3.3) with respect to both unknowns r and x_t, equating the derivatives to 0, dividing both sides of the resulting equation by $T + 1$, and taking into account that

$$\sum_{i=0}^{T} i = \frac{T \cdot (T+1)}{2}$$

and

$$\sum_{i=0}^{T} i^2 = \frac{T \cdot (T+1) \cdot (2T+1)}{6},$$

we get the following system of two equations:

$$\overline{x} = x_t + r \cdot \frac{T}{2}; \tag{3.4}$$

$$\frac{1}{T+1} \cdot \sum_{i=0}^{T} (i \cdot \widetilde{x}_{t+i}) = x_t \cdot \frac{T}{2} + r \cdot \frac{T \cdot (2T+1)}{6}, \tag{3.5}$$

where

$$\overline{x} \stackrel{\text{def}}{=} \frac{1}{T+1} \cdot \sum_{i=0}^{q} \widetilde{x}_{t+i} \tag{3.6}$$

is the arithmetic average of all the observed values $\widetilde{x}_{t+i}$.

From these two equations, we can find the estimates r and x_t. Let us first find the estimate r. Multiplying the equation (3.4) by $\dfrac{T}{2}$, we get

$$\frac{1}{T+1} \cdot \sum_{i=0}^{q} \left(\frac{T}{2} \cdot \widetilde{x}_{t+i} \right) = x_t \cdot \frac{T}{2} + r \cdot \frac{T^2}{2}. \tag{3.7}$$

Subtracting (3.7) from (3.5), we get

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

$$r \cdot \frac{T}{2} \cdot \frac{2T+1}{3} - r \cdot \frac{T}{2} \cdot \frac{T}{2} =$$

$$r \cdot \frac{T}{2} \cdot \left(\frac{2T+1}{3} - \frac{T}{2} \right) = r \cdot \frac{T}{2} \cdot \frac{T+2}{6}. \tag{3.8}$$

Thus,

$$r = \frac{12}{T \cdot (T+1) \cdot (T+2)} \cdot \sum_{i=0}^{q} \left(\left(i - \frac{T}{2} \right) \cdot \widetilde{x}_{t+i} \right). \tag{3.9}$$

Substituting this expression into the formula (3.4), we get

$$x_t = \overline{x} - r =$$

$$\frac{1}{T+1} \cdot \sum_{i=0}^{q} \widetilde{x}_{t+i} -$$

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

$$\frac{1}{T \cdot (T+1) \cdot (T+2)} \cdot \sum_{i=0}^{q} (T \cdot (T+8) - 12 \cdot i) \cdot \widetilde{x}_{t+i}. \tag{3.10}$$

Once we have computed r and x_t from these equations, then, based on a single measurement, we can then estimate the standard deviation σ of the measurement error ε_{t+i} as the mean square difference between the observed and predicted values:

$$\sigma^2 = \frac{1}{T+1} \cdot \sum_{i=0}^{T} (\widetilde{x}_{t+i} - (x_t + r \cdot i))^2. \tag{3.11}$$

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 these locations:

$$\sigma^2 = \frac{1}{L} \cdot \frac{1}{T+1} \cdot \sum_{k=1}^{K} \sum_{i=0}^{T} (\widetilde{x}_{t+i,\ell} - (x_{t,k} + r_k \cdot i))^2, \tag{3.11a}$$

where K is the overall number of spatial locations.

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, the formulas (3.9) and (3.10) take the following form:

$$r = \frac{\widetilde{x}_{t+2} - \widetilde{x}_t}{2}, \tag{3.12}$$

and

$$x_t = \frac{5\widetilde{x}_t + 2\widetilde{x}_{t+1} - \widetilde{x}_{t+2}}{6}. \tag{3.13}$$

Here,

$$x_t + r = \frac{\widetilde{x}_t + \widetilde{x}_{t+1} + \widetilde{x}_{t+2}}{3} \tag{3.14}$$

and

$$x_t + 2r = \frac{\widetilde{x}_t + \widetilde{x}_{t+1} + \widetilde{x}_{t+2}}{3} + r =$$

$$\frac{-\widetilde{x}_t + 2\widetilde{x}_{t+1} + 5\widetilde{x}_{t+2}}{6}. \tag{3.15}$$

Thus,

$$\widetilde{x}_t - x_t = \frac{\widetilde{x}_t - 2\widetilde{x}_{t+1} + \widetilde{x}_{t+2}}{6}, \tag{3.16}$$

$$\widetilde{x}_{t+1} - x_{t+1} = \frac{\widetilde{x}_t - 2\widetilde{x}_{t+1} + \widetilde{x}_{t+2}}{3}, \tag{3.17}$$

and

$$\widetilde{x}_{t+2} - x_{t+2} = \frac{\widetilde{x}_t - 2\widetilde{x}_{t+1} + \widetilde{x}_{t+2}}{6}. \tag{3.18}$$

Therefore, in this case,

$$\sigma^2 =$$

$$\frac{1}{3} \cdot ((\widetilde{x}_t - x_t)^2 + \widetilde{x}_{t+1} - x_{t+1})^2 + (\widetilde{x}_{t+2} - x_{t+2})^2) =$$

$$\frac{1}{3} \cdot \left(\frac{1}{6^2} + \frac{1}{3^2} + \frac{1}{6^2} \right) \cdot (\widetilde{x}_t - 2\widetilde{x}_{t+1} + \widetilde{x}_{t+2})^2 =$$

$$\frac{1}{18}\cdot(\widetilde{x}_t-2\widetilde{x}_{t+1}+\widetilde{x}_{t+2})^2. \tag{3.19}$$

By taking the average over all spatial locations, we get

$$\sigma^2=\frac{1}{18}\cdot\frac{1}{K}\cdot\sum_{k=1}^{K}(\widetilde{x}_{t,k}-2\widetilde{x}_{t+1,k}+\widetilde{x}_{t+2,k})^2. \tag{3.20}$$

What if positive erosion values are not always within 2-sigma range? The estimated erosion rate r may be negative, but it is OK if within the corresponding 2-sigma interval

$$[r-2\sigma,r+2\sigma],$$

we have a positive value, i.e., if we have $r+2\sigma>0$. This means that the difference between the actual (positive) erosion rate and our (negative) estimate r can be explained by the observation uncertainty.

But what is $r+2\sigma<0$? This would mean that we need an additional source of error, i.e., that instead of the formula (3.2), we will have

$$\widetilde{x}_{t+i}=x_{t+i}+\varepsilon_{t+i}+\delta_{t+i}=x_t+i\cdot r+\varepsilon_{t+i}+\delta_{t+i}. \tag{3.21}$$

In this case, we still determine our estimates x_t and r from the least squares method (3.3). However, now, in addition to the error component (3.11), we have an additional source of error, with some standard deviation σ_δ^2, so the overall variance σ_t^2 now has the form

$$\sigma_t^2=\sigma^2+\sigma_\delta^2. \tag{3.22}$$

How can we determine σ_t^2 and σ_δ^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 $r_k+2\sigma_t\ge 0$, i.e., equivalently, $2\sigma_t\ge -r_k$. If we sort the estimated erosion rates in increasing order, as

$$r_1<r_2<\ldots<r_N, \tag{3.23}$$

then this means that the desired inequality should be satisfied for all $k\ge 0.05\cdot N$, i.e., that we should have $2\sigma_t\ge -r_{0.05\cdot N}$, $2\sigma_t\ge -r_{0.05\cdot N+1}$, etc. Since the sequence r_k is sorted in increasing order, the first inequality implies all the others, so it is sufficient to satisfy the first inequality $2\sigma_t\ge -r_{0.05\cdot N}$, i.e., $\sigma_t\ge -\frac{1}{2}\cdot r_{0.05\cdot N}$.

We would like to have the narrowest error bounds, so we choose the smallest $\sigma_t\ge\sigma$ that satisfies this inequality, i.e., we take $\sigma_t=\max\left(\sigma,-\frac{1}{2}\cdot r_{0.05\cdot N}\right)$.

Resulting algorithm: case of general T. We start with measurements $\widetilde{x}_{t+i,k}$ make at different spatial locations k at years t, $t+1$, …, $t+T$.

For each location k, we compute the estimated erosion rate

$$r_k=$$

$$\frac{12}{T\cdot(T+1)\cdot(T+2)}\cdot\sum_{i=0}^{q}\left(\left(i-\frac{T}{2}\right)\cdot\widetilde{x}_{t+i,k}\right) \tag{3.24}$$

and the estimated initial erosion

$$x_{t,k}=$$

$$\frac{1}{T\cdot(T+1)\cdot(T+2)}\cdot\sum_{i=0}^{q}(T\cdot(T+8)-12\cdot i)\cdot\widetilde{x}_{t+i,k}. \tag{3.25}$$

Then, we estimated the first approximation σ to the corresponding uncertainty as

$$\sigma^2=\frac{1}{L}\cdot\frac{1}{T+1}\cdot\sum_{k=1}^{K}\sum_{i=0}^{T}(\widetilde{x}_{t+i,\ell}-(x_{t,k}+r_k\cdot i))^2. \tag{3.26}$$

We then sort the estimated erosion rates in increasing order:

$$r_1<r_2<\ldots<r_N, \tag{3.27}$$

and take

$$\sigma_t=\max\left(\sigma,-\frac{1}{2}\cdot r_{0.05\cdot N}\right). \tag{3.28}$$

This σ_t is the mean square accuracy of the erosion rate estimates r_k.

Resulting algorithm: case $T=2$. For the case $T=2$, when we have three consecutive years of measurement, we have simplified formulas

$$r_k=\frac{\widetilde{x}_{t+2}-\widetilde{x}_{t,k}}{2}, \tag{3.29}$$

and

$$\sigma^2=\frac{1}{18}\cdot\frac{1}{K}\cdot\sum_{k=1}^{K}(\widetilde{x}_{t,k}-2\widetilde{x}_{t+1,k}+\widetilde{x}_{t+2,k})^2. \tag{3.30}$$

Then, we take σ_t as determined by the formula (3.28).

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 "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

[1] S. H. Ackers, R. J. Davis, K. A. Olsen, and K. M. Dugger, "The volution of mapping habitat for northern spotted owls (*Srix occientalis caurina*): A comparison of photo-interpreted, Landsat-based, and lidar-based habitat maps", *Remote Sensing of Environment*, 2015, Vol. 156, pp. 361–373.

[2] A. Afifi and S. May, *Practical Multivariate Analysis*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.

[3] R. Early, B. Anderson, and C. D. Thomas, "Using habitat distribution models to evaluate large-scale landscape priorities for spatially dynamic species", *Journal of Applied Ecology*, 2008, Vol. 45, pp. 228–238.

[4] G. Klir and B. Yuan, "Fuzzy Sets and Fuzzy Logic", Prentice Hall, Upper Saddle River, New Jersey, 1995.

[5] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[6] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[7] G. Singh, A. Velmurugan, and M. P. Dakhate, "Geospatial approach for tiget habitat evaluation and distribution in Corbett Tiger Reserve", *Journal of Indian Society of Remore Sensing*, 2009, Vol. 37, pp. 573–585.

[8] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.