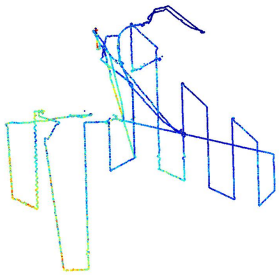
Tuning interpolation methods for environmental line/transect surveys

You Li and Maria-João Rendas
Laboratoire I3S, CNRS, Sophia Antipolis, FRANCE

May 22, 2015

Abstract

Environmental observation of extended regions is often accomplished by using motorised platforms equipped of sensing equipment that perform a trajectory that "covers" the region of interest $\mathcal{A}$ with a series of line transects, as in the figure below. Ultimate goal is to map a field $f(\cdot)$ over $\mathcal{A}$. Sensor acquisition rate and carrier speed, along with limitations in the available power and time, result on very dense sampling along the trajectory of the carrier compared to the average point density over the region.



Let $Z = \{z(\ell), \ell \in \mathcal{L} \subset \mathcal{A}\}$ be the set of acquired measures. Design $\mathcal{L}$, the set of observation sites, is a discrete subset of $\mathcal{A}$. The interpolated map $\hat{f}(s), s \in \mathcal{A}$ is obtained by applying some algorithm $G_\theta[\cdot]$ to the acquired data:

$$\hat{f}(s; G_\theta, Z) = G_\theta\left[s; Z\right], \qquad s \in \mathcal{A} \ .$$

Above, θ represents a set of user-defined algorithm parameters (size of neighbourhoods of local methods, scale parameters of Kriging, etc). These parameters influence the quality of the map, that is most often assessed by the Integrated Mean Square Error:

$$IMSE_{G,Z}(\theta) = \frac{1}{|\mathcal{A}|}\int_{\mathcal{A}} \left(f(s) - \hat{f}(s; G_\theta, Z)\right)^2 \, ds \ .$$

Choice of θ is particularly important when the $\mathcal{L}$ does not sample $\mathcal{A}$ densely, and can be done in two different contexts: *(a)* a parametric *stochastic* model $\mathcal{M}(\gamma)$ exists that captures knowledge about the field; *(b)* no further facts about $f(\cdot)$ besides Z are known.

Kriging, for instance, falls under case *(a)*. Using the data, estimates of the model parameters $\hat{\gamma}_{\mathcal{M})}(Z)$ can be produced, allowing optimisation of the expected IMSE value, $\mathrm{E}_{\mathcal{M}(\hat{\gamma}(Z))}\left[IMSE_{G_\theta,Z}\right]$. This approach is intrinsically sensitive to the correctness of the assumed model.

Approach *(b)*, generically known as *cross-validation* (CV), is free of any prior assumptions about the field, being often the preferred practitioners' choice. Several variants of CV exist, but the following captures their generic form. Let s_i denote a generic point of $\mathcal{L}$, z_i the corresponding measure $(y_i = f(s_i))$ and $Z^{(i)}$ a subset of Z that does *not contain* y_i. An estimate of the interpolation error for G with parameters θ is obtained by using $Z^{(i)}$ to estimate the field at s_i:

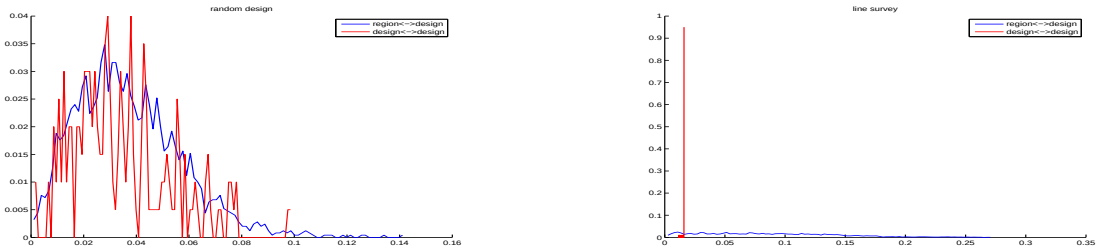
$$\epsilon_{s_i;Z^{(i)}}(G, \theta) = y_i - \hat{f}(s_i; G, Z^{(i)}, \theta), \qquad s_i \in \mathcal{L} \ .$$

Averaging these estimates over $\mathcal{L}$ yields an estimate of $IMSE_{G,Z}(\theta)$ that can be used to choose θ:

$$\widehat{IMSE}_{G,Z}(\theta) = \frac{1}{|\mathcal{L}|} \sum_{s_i \in \mathcal{L}} \epsilon^2_{s_i;Z^{(i)}}(G,\theta) \ .$$

Different choices for the sets $Z^{(i)}$ give rise to different variants of CV, the most common being "leave-one-out," where $Z^{(i)} = Z \setminus \{y_i\}$. Existing literature on CV implicitly assume that the design $\mathcal{L}$ is a "space filling design". The exact definition of "space filling" varies in the literature, but here it is sufficient to say that points of these designs are as far away from each other as possible.

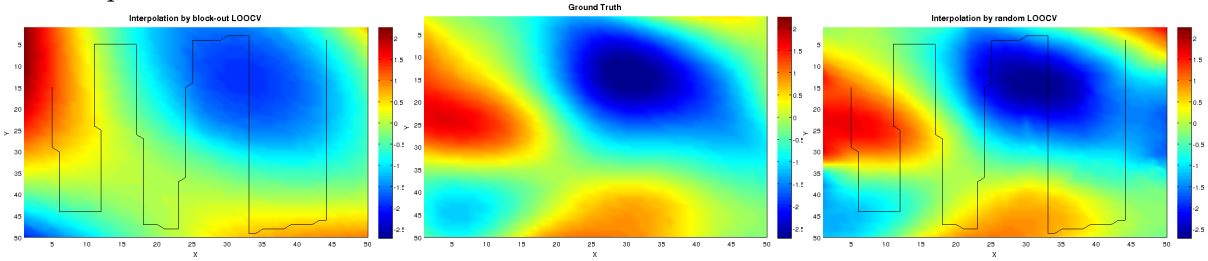
It is generally accepted that the interpolation error at $s \in \mathcal{A}$ depends strongly on the distance from s to its closest point in $\mathcal{L}$: $d(s) = \min_{s_i \in \mathcal{L}} \|s - s_i\|$. For "space filling" designs, the distribution of $d_{cv}(s_i) = \min_{s_j \in \mathcal{L} \setminus \{s_i\}} \|s_i - s_j\|, s_i \in \mathcal{L}$, the distance between each s_i and the other design points is similar to the distribution of $\{d(s), s \in \mathcal{A}\}$ (see the red and blue curves on left of the figure below) yielding sensible estimates of the overall IMSE.



Distribution of $d_{cv}(s_i)$ (red) and $d(s)$ (blue). Left: random design; right: line survey.

This is clearly not the case in line/transect surveys, where the distance between consecutive points is small, as shown on the right of the figure. This induces biases on estimation of $IMSE(\theta)$ that prevent optimal tuning of the algorithm. Our results on real and simulated data confirm this.

We propose a new randomised CV method able to cope with non-uniformly scattered designs, like line surveys, enabling a robust parameterisation of the interpolation algorithms free of model assumptions. The new method selects the sets $Z^{(i)}$ randomly, such that the induced distribution of $d_{cv}(s_i)$ matches the distribution of $d(s)$ regardless of the geometry of $\mathcal{L}$, preventing biases due to its particular intrinsic geometry. Below we plot the interpolated maps using the parameters of locally weighted regression selected by standard CV (SCV, left) and by our method (RCV, right) that clearly show its advantage. The black line shows the observation path. The integrated residuals are 0.326 for SCV, 0.195 for RCV, and attains a minimum value of 0.190 (parameters that minimise the residuals' sum over the entire simulation grid). Note the small loss of RCV compared to the actual optimal value.



SCV $f(\cdot)$ RCV

We justify theoretically the new method, and illustrate the gains that can be expected on real data related to monitoring of toxic algae blooms.