Tuning interpolation methods for environmental uni-dimensional (transect) surveys

You Li, Maria-João Rendas

Tuning interpolation methods for environmental uni-dimensional (transect) surveys

You Li and Maria-João Rendas
Laboratoire I3S, CNRS
Sophia Antipolis, FRANCE
Email: {yli}{rendas}@unice.fr

***Abstract*—The paper proposes rCV, a new randomised Cross Validation (CV) criterion specially designed for use with data acquired over non-uniformly scattered designs, like the linear transect surveys typical in environmental observation. The new criterion enables a robust parameterisation of interpolation algorithms, in a manner completely driven by the data and free of any modelling assumptions. The new CV method randomly chooses the hold-out sets such that they reflect, statistically, the geometry of the design with respect to the unobserved points of the area where the observations are to be extrapolated, minimising biases due to the particular geometry of the designs. Numerical results on both simulated and realistic datasets show its robustness and superiority, leading to interpolated fields with smaller error.**

I. Introduction

Environmental observation of an extended region $\mathcal{A}$ is often accomplished by using a motorised platform equipped of sensing equipment that performs a trajectory $\mathbf{p}(\cdot) \subset \mathcal{A}$ that "covers" $\mathcal{A}$ with a series of nearly paralell line transects. The trajectory followed by a surface boat during observation of a lake in Belgium shown in Figure 1[1] is a representative example. Sensor acquisition rate and carrier speed, along with limitations in power and time, result in a much higher sampling rate along the trajectory of the carrier than the average point density over $\mathcal{A}$. This apparent in Figure 1, that actually plots (color coded) the individual sampled values.

The ultimate goal of spatial surveys is to produce a map of some observed field $f(\cdot)$ over the entire region $\mathcal{A}$, i.e., to extrapolate (the term "predict" is also used) the point samples of $f(\cdot)$ taken along the carrier path to the entire surface.

In this paper we address the problem of tuning the parameters of the algorithm used to estimate the value of $f(\cdot)$ at the unobserved points of $\mathcal{A}$, taking into account that the set of sampled points (the design) may have an arbitrary geometry, in particular the intrinsically one-dimensional geometry described above. A method widely used by practitioners to chose algorithm's parameters is Cross Validation (CV), whose origins go at least as far as the 1930's, see the interesting discussion in [1]. As we show below, CV works well in the geo-statistical context when the data points are "space filling", i.e., uniformly scattered in $\mathcal{A}$, as in Figure 2. This is clearly not the case for the uni-dimensional design of Figure 1. This paper presents rCV, a novel randomised version of CV specially tailored to be robust with respect to strongly "non space filling" design geometries.

[1]Courtesy of VITO.

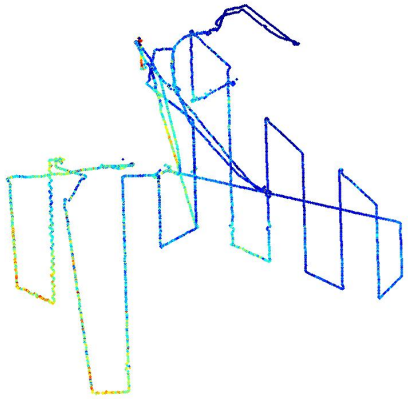

Fig. 1. Design used for a lake observation (courtesy of VITO).

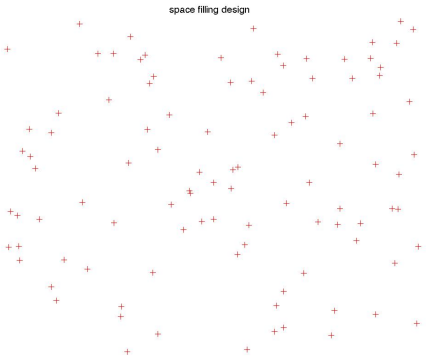

Fig. 2. Space filling design.

We start by introducing notation and formulating the algorithm tuning problem (section II) and subsequently present CV methodologies (section III) pointing out the specific difficulties that arise for uni-dimensional designs. Section IV presents rCV, the new randomised CV method that we propose. Finally, we demonstrate (section V) the advantages of rCV in both

real and simulated datasets, and section VI summarises our contribution and lists some directions for future work.

II. Problem formulation

Let $Y^{(K)}$ denote the complete set of measures acquired during a one-dimensional survey done along path $\mathbf{p}(\cdot) \subset \mathcal{A}$

$$Y^{(K)} = \{y_k = f(\mathbf{p}(\ell_k)), k = 1, \ldots, K\}$$

Let Ξ be the design, i.e., the set of points at which observations are made

$$\Xi = \{\mathbf{p}(\ell_k), k = 1, \ldots, K\} \ ,$$

and denote by $\xi_k = \mathbf{p}(\ell_k)$ a generic point of Ξ.

Let $\mathbf{F}$ denote the operator that when applied to the set of measures $Y^{(K)}$ generates the field predictions (the extrapolation algorithm):

$$\hat{f}(s|Y^{(K)}; \rho) = \mathbf{F}\left(s, Y^{(K)}; \rho\right), \qquad s \in \mathcal{A} \ .$$

As indicated above, $\mathbf{F}$ usually depends on a series of parameters $\rho \in \mathbb{R}^q$ (for instance, the kernel scale and trend model for kriging methods, or the number of neighbours or neighbourhood size in local regression methods) that must be defined by the user. Choice of ρ is particularly important when, as it is often the case in environmental applications, Ξ does not sample $\mathcal{A}$ densely.

Several criteria can be used to measure the quality of spatial extrapolation. We consider here the most common one, the Integrated Square Error (ISE), even if the method proposed is independent of this particular choice:

$$C_{ise}(\rho; Y^{(K)}) = \frac{1}{|\mathcal{A}|} \int_{\mathcal{A}} e^2_\rho(s; Y^{(K)})\, ds, \tag{1}$$
$$e_\rho(s; Y^{(K)}) = f(s) - \hat{f}(s|Y^{(K)}; \rho) \ .$$

Ideally, one would choose ρ such that the reconstruction error is minimal:

$$\rho^\star(Y^{(K)}) = \arg\min_\rho C_{ise}(\rho; Y^{(K)}) \ . \tag{2}$$

Note that C_{ise} cannot be computed, as it depends on the field's values outside the design Ξ, and only estimates of C_{ise} based on the available data $Y^{(K)}$ can be used to chose ρ. Two different frameworks are envisageable: *(a)* a parametric *stochastic* model $\mathcal{M}(\gamma)$ is known to capture the characteristics of the field $f(\cdot)$, enabling determination of the expected value of C_{ise}; *(b)* C_{ise} must be estimated using only $Y^{(K)}$, no additional knowledge about $f(\cdot)$ being available. Kriging methods fall under *(a)*, the interpolation algorithm being intrinsincally tied to a stochastic model of the observed field: data $Y^{(K)}$ can be used to estimate the model parameters $\hat{\gamma}(Y^{(K)})$, which determine the optimal (in an expected value sense) estimator of $f(\cdot)$ at unobserved points of $\mathcal{A}$. Model-based frameworks are sensitive to the correctness of the assumed models, and for this reason *(b)* is often the preferred practitioners' choice. In this paper we propose a new estimator of the reconstruction errror for subsequent selection of the value of ρ using Cross Validation, a popular methodology belonging to framework *(b)*, see the section below.

III. Cross Validation

Several variants of CV exist [2], but the simple description below captures the main principle at work behind them all and is sufficient for the purpose of this paper.

Let ξ_i denote a generic point of Ξ, $\Xi^{(-i)}$ a subset of Ξ that *does not contain* ξ_i and $Y^{-(i)}$ the corresponding measures. A realisation of the interpolation error for algorithm $\mathbf{F}$ with parameters ρ is obtained by using $Y^{(-i)}$ to estimate the field at ξ_i:

$$\epsilon_\rho(\xi_i; Y^{(-i)}) = y_i - \hat{f}(\xi_i|Y^{(-i)}; \rho), \qquad \xi_i \in \Xi \ .$$

Averaging these residuals over $\xi_i \in \Xi$ yields a CV estimate of C_{ise}

$$C_{CV}(\rho|Y^{(K)}) = \widehat{C_{ise}}(\rho|Y^{(K)}) = \frac{1}{|\Xi|} \sum_{\xi_i \in \Xi} \epsilon^2_\rho(\xi_i; Y^{(-i)}) \ , \tag{3}$$

that can be used to select ρ by $\rho^\star = \arg\min_\rho C_{CV}(\rho|Y^{(K)})$. Different choices for the sets $Y^{(-i)}$ give rise to different variants of CV, the most common being "leave-one-out," where $Y^{(-i)} = Y^{(K)} \setminus \{y_i\}$.

As it is obvious from the presentation above, one necessary condition for $C_{CV}(\rho|Y^{(K)})$ to be a sensible estimate of the true error is that the set of "cross-validation residuals" $\epsilon_\rho(\xi_i; Y^{(-i)})$ be a representative sample of the prediction error process at unobserved points of $\mathcal{A}$. This is never exactly the case, and in fact one can show that C_{CV} is in general a conservative estimate of the error, i.e., it predicts a performance poorer than the expected one, and several corrections have been proposed to eliminate this negative bias. However, it is customarily accepted that this bias it does not affect the location of the minimum of C_{CV} as a function of ρ, i.e., the uncorrected CV criterion can safely be used to select the best interpolating model.

Let us now analyse the impact of the one-dimensional design geometry on the quality of C_{CV} as an estimator of the true prediction error.

Spatial interpolation is known to be particularly sensitive to the geometry of Ξ around each reconstructed point. An additional condition for the validity of C_{CV} as a proxy of the interpolation error C_{ise} is thus that, around each design point ξ_i, the geometry of the designs $\Xi^{(-i)}$ be representative of the spatial distribution of the points of Ξ around generic unobserved points of $\mathcal{A}$. This is true for "space filling designs"[2], where the density of the design points is nearly homogeneous inside $\mathcal{A}$, but is violated in the case of uni-dimensional surveys as the ones considered in this paper.

Consider the extreme case of the leave-one-out CV, where $Y^{(-i)} = Y^{(K)} \setminus \{y_i\}$, that clearly reveals the biasing effect of uni-dimensional designs Ξ. Points of Ξ are naturally numbered along the curve $\mathbf{p}(\cdot)$. There are two nearest neighboors in the sets $\Xi^{(-i)}$, for each ξ_i: ξ_{i-1} and ξ_{i+1}, at distances $d_- = \|\xi_i - \xi_{i-1}\| \simeq d_+ = \|\xi_i - \xi_{i+1}\|$. The estimates at ξ_i produced

[2]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.

by most interpolation algorithms will be strongly determined by these two points, with an error that will be small if these distances are small, i.e., when $\mathbf{p}(\cdot)$ is densely sampled, as we consider in this paper.

To facilitate analysis, consider the simple case of a regular survey with transects of length L spaced D apart. The distribution of distances of a generic points $s \in \mathcal{A}$ to Ξ can be approximated by

$$\pi_\Xi(d) \simeq \left(1 + \frac{D}{2L} - \frac{2}{L}d\right), \qquad d \in \left[0, \frac{D}{2}\right] \ ,$$

which, for $D \ll L$ is close to the uniform distribution in $[0, \frac{D}{2}]$. With probability $\simeq 1 - 2\delta_+/D$ the distance of points in $\mathcal{A}$ to Ξ will be larger than $d_- \simeq d_+$, the distance to the closest points in the leave-one-out hold sets $Y^{(-i)}$, leading to largely optimistic CV estimates of the reconstruction error, i.e., the CV criterion (3) will be much smaller than the actual reconstruction error (1).

For model-based estimators, if the model assumed is correct, the minimum of $C_{CV}(\cdot)$ can still occur, in average, near the "good value" of parameter ρ, as our results on simulated data in section V show. However, in the most likely situation that the observations do not follow the model for which $\mathbf{F}$ has been designed, standard CV criteria will be unable to capture the behaviour of the prediction error over the entire range of possible situations (in terms of distances to the closest data points). The randomised CV method presented in the next section overcomes this limitation.

IV. Randomised Cross Validation (rCV)

As mentioned in the previous section the reconstruction error at a point $s \in \mathcal{A}$ is highly dependent on the local geometry of the points of Ξ in the neighborhood of s. In a first order approach this error is dominated by the distance of s to its closest point in Ξ, see [3]. This justifies use of the common *minimax* design criterion that minimises $C_{mM}(\Xi)$, the maximum distance between points of $\mathcal{A}$ and the design Ξ:

$$C_{mM}(\Xi) = \max_{s \in \mathcal{A}} \min_{\xi \in \Xi} \|s - \xi\| \ .$$

This criterion is difficult to compute, requiring a search over the entire region $\mathcal{A}$, and the *maximim* criterion C_{Mm} is often used instead. It evaluates a design by the distance between the closest points of Ξ, which should be as large as possible:

$$C_{Mm}(\Xi) = \min_{\xi_1, \xi_2 \in \Xi} \|\xi_1 - \xi_2\| \ .$$

Maximising this criterion leads to maximally spreading all points inside $\mathcal{A}$, and is known to push some of the design points to its boundary, which is certainly not optimal from the point of view of the C_{mM} criterion. For designs with good space filling properties the set of distances $\{d(\xi_i, \Xi^{(-i)}), \xi \in \Xi\}$ is expected to be a typical sample from the distribution of distances $\{d(s, \Xi), s \in \mathcal{A}\}$, which is a condition for (3) to yield a valid indication of (1).

It is obvious that dense one-dimensional transect designs like the ones considered in this paper are not space filling under any purely geometric criterion. When Ξ is a regular sampling of a one-dimensional curve $\mathbf{p}(\ell)$, C_{Mm} will be trivially equal to the distance between adjacent points, while C_{mM} will most likely be equal to $D/2$ as the approximate expression for π_Ξ given above shows. That the difference between C_{Mn} and C_{mM} is large is an indication that standard CV techniques will fail to provide a valid indication of the expected performance when extrapolating over the entire $\mathcal{A}$.

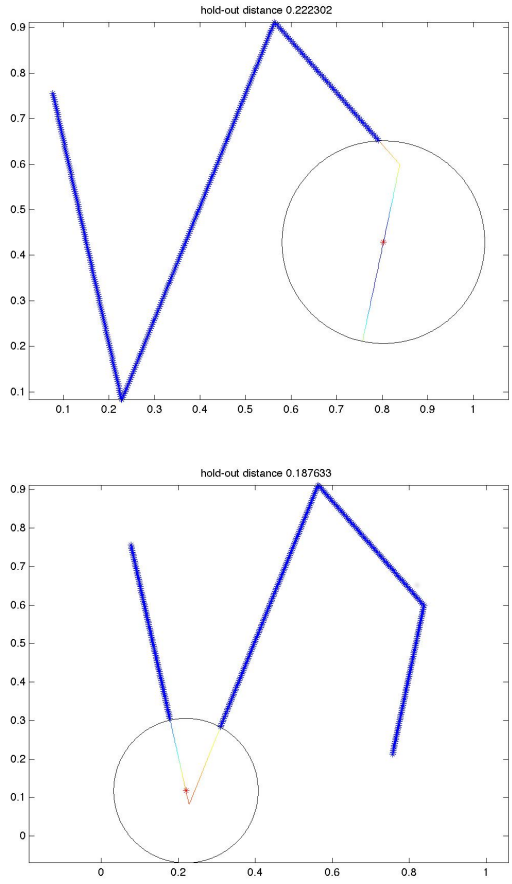

Fig. 3. Example of block-out sets $Y^{-(s,r)}$

Let $d(s, \Xi)$ be the distance between point $s \in \mathcal{A}$ and Ξ:

$$d(s, \Xi) = \min_{\xi \in \Xi} \|s - \xi\| \ ,$$

and, as before, let $\pi_\Xi(d)$ the probability distribution of $d(s, \Xi)$ when $s \sim \mathcal{U}(\mathcal{A})$, i.e., when s is uniformly distributed in $\mathcal{A}$.

For any $s \in \mathcal{A}$ and $r \geq 0$ let $B_r(s)$ be the ball of radius r centred at s, and denote by $\Xi^{-(s,r)}$ be the sets

$$\Xi^{-(s,r)} = \Xi \cap B_r(s)^c \ , \tag{4}$$

and denote by $Y^{-(s,r)}$ the corresponding measures. Notation A^c denotes complement of set A in $\mathcal{A}$. Figure 3 illustrates the definition of the sets $Y^{-(s,r)}$, the hold-out sets used in our randomised CV.

Randomised CV criterion

The randomised CV criterion for the prediction error over region $\mathcal{A}$ using observations $Y^{(K)}$ taken over design Ξ is

$$C_{rCV}(\rho|Y^{(K)}) = \mathrm{E}_{r,\xi}\left(y(\xi) - \mathbf{F}\left(\xi, Y^{-(\xi,r)}, \rho\right)\right)^2 \quad , \quad (5)$$

where the statistical average is computed using distributions $r \sim \pi_\Xi$ and $\xi \sim \mathcal{U}(\Xi)$, and the hold-out sets $Y^{-(\xi,r)}$ are defined by (4).

Criterion C_{rCV}, as equation (5) shows, uses randomly chosen hold-out sets, that statistically reflect the geometry of design Ξ with respect to the entire region $\mathcal{A}$: it puts CV in the geometrical conditions under which extrapolation will actually be done.

The numerical computation of C_{rCV} ressorts to stochastic simulation, the expected value in the definition (5) being approximated by the empirical average

$$C_{rCV}(\rho|Y^{(K)}) \simeq \frac{1}{M}\sum_{i=1}^{M}\left(y_i - \mathbf{F}\left(\xi_i, Y^{-(\xi_i,r_i)}, \rho\right)\right)^2 \quad ,$$

where $y_i = y(\xi_i)$, $M \propto |\mathcal{A}|$ is a large number and the pairs (ξ_i, r_i) are independently drawn according to

1) $\xi_i \sim \mathcal{U}(\Xi)$
2) $r_i = d(s_i, \Xi)$, $s_i \sim \mathcal{U}(\mathcal{A})$.

The price payed for the robustness of C_{rCV} is its higher numerical complexity: the number of criterion evaluations (predictions) is now of the order of the size of $\mathcal{A}$, usually much larger than the size of Ξ, required for instance by leave-one-out CV. Also, the determination of the distances $d(s_i, \Xi)$, requiring a minimisation over Ξ, are computationally expensive. Note that for designs based on paths $\mathbf{p}(\cdot)$ ans sets $\mathcal{A}$ of simple geometry, for which π_Ξ can be approximated analytically, computation of these distances can be avoided.

V. Numerical Results

We compare the performance of several sptatial interpolation methods tuned by the randomised CV proposed in the paper to the following methods.

1) *Usual leave-one-out* (LOOCV). This corresponds to using all other points to estimate each observation, i.e., to hold-out sets of the form

$$\Xi^{(-i)} = \Xi \setminus \{\xi_i\} \quad .$$

Comparison with this criterion will show the importance of holding out points close to the interpolation point.

2) *Fixed distance hold-out* (Fix). This is a simplified version of rCV, and corresponds to considering hold-out sets which are of the form

$$\Xi^{(-i)} = \Xi^{-(\xi_i, r^\star)} \quad .$$

This is a standard approach used for correlated data, where the value of $r^\star$ is specified by the user, based on its expectation with respect to the field's correlation range, which in general is unknown. We propose instead to adjust the hold-out distance $r^\star$ based on the distribution π_Ξ, and set $r^\star$ to a fixed quantile α of π_Ξ, i.e., such that

$$\mathrm{Prob}\left\{d(s,\Xi) \le r^\star\right\} = \alpha \quad .$$

We used $\alpha = .75$ in the numerical results below. Comparison with this criterion will enable us to assess the importance of using hold-out sets that reflect $\pi_\Xi(d)$, even if in a simplified manner.

3) *Variogram* (Var). To stress the robustness of the proposed criterion, we also show numerical results for Kriging, comparing the observed prediction errors of the optimal Kriging estimator using a range parameter ρ chosen by C_{rCV} to the standard use of variogram considered by most geo-statistics packages.

A. Simulated Data

This section considers simulated datasets which are realisations of a Gaussian Process (GP) of zero mean and Matérn correlation function with parameter $\nu = 3/2$

$$R(s,s') = \sigma^2\left(1 + \frac{\sqrt{3}\|s-s'\|}{\rho}\right)\exp\left(-\frac{\sqrt{3}\|s-s'\|}{\rho}\right) \quad ,$$

with $\rho = 10$. The field is simulated inside a square region $\mathcal{A} = [0,60] \times [0,60]$. Figure 6 shows one of the simulated fields and the one-dimensional design used in the numerical results shown below.

Interpolation by both non-parametric (Local Weighted regression, LWR) and parametric (Kriging) methods are studied.

1) LWR: is a common interpolation method that fits simple local parametric models $f(s) = F(s;\theta(s,\rho))$ to the dataset. At each point $s \in \mathcal{A}$, the parameters $\theta(s,\rho)$ minimise the weighted average

$$\theta(s,\rho) = \arg\min_\theta \sum_{y_i \in Y^{(K)}} w(s,\xi_i;\rho)\left(y_i - F(s;\theta)\right)^2 \quad ,$$

where the weights $w(s,\xi_i;\rho)$ are a decreasing function of $\|s-\xi_i\|$ such that

$$\sum_i w(s,i;\rho) = 1 \quad .$$

The local models that are fit are most commonly of polynomial type, θ being the coefficients of these polynomials. The original method [4] considers weighting functions of compact support, the parameter ρ being the size of this support, making at each point a truly local (weighted) fit to a strict subset of $Y^{(K)}$. Alternatively, we may consider functions of infinite support, like exponential or Gauss functions, and in this case ρ controls the speed with each the $w(s,\cdot;\rho)$ tends to zero. The numerical results below use a Gauss weight function.

Figures 4 and 5 illustrate the performances observed in 100 distinct simulations of the model above for quadratic LWR. The three tuning methods studied here are LOOCV (green), rCV (red) and the fixed-distance hold-out set (blue). We also plot (black curve) the ISE values for the best ρ parameter that can be adjusted to each of the 100 simulated datasets.

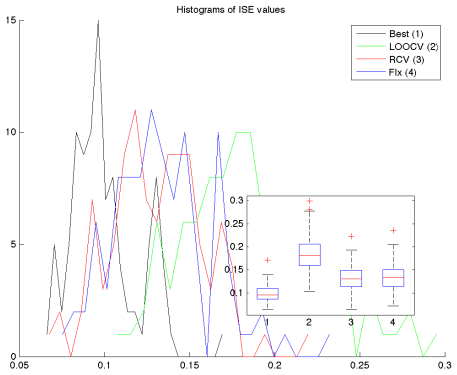

Fig. 4. ISE for LWR using different tuning methods (one-dimensional design).

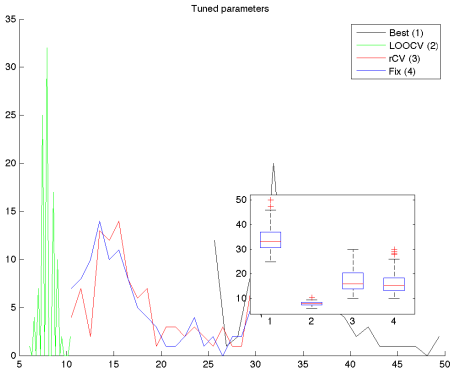

Fig. 5. Tuned LWR parameters (uni-dimensional design).

We can see that rCV consistently leads to better interpolation performance (lowest ISE), see Figure 4. Average values of ISE are equal to 0.186 for LOOCV, 0.131 for rCV and 0.135 for fixed hold-out. The actual minimum of ISE (computed using knowledge of the entire simulated field) is equal to 0.101. As we see, even the simple hold-out sets using a fixed distance already lead to a good improvement over LOOCV, showing how important it is that the sets $\Xi^{(-i)}$ reflect the actual geometric conditions under which the dataset will be interpolated. Figure 5 shows histograms and corresponding box-plots of the tuned parameters ρ (in this case the rate of the Gaussian weight function) for the 100 datasets. As it should be expected, the rate chosen by LOOCV is systematically lower than the rates chosen by the other two methods, leading to a worst performance when interpolating over distant points. The "best" ρ parameters are also shown for comparison.

Figure 7 shows the interpolation results for one of the sampled fields using the parameters tuned by LOOCV (top) and rCV (bottom).

2) Kriging: Kriging [5] is a popular geo-statistical interpolation method that implements the optimal mean-square error estimate of a spatial field based on the assumption that it is a realisation of a stationary Gaussian Process. Several variants

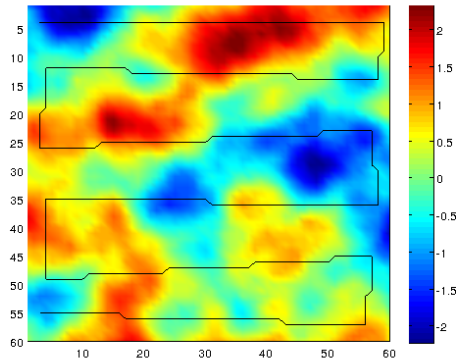

Fig. 6. Simulated Gaussian field and one-dimensional design.

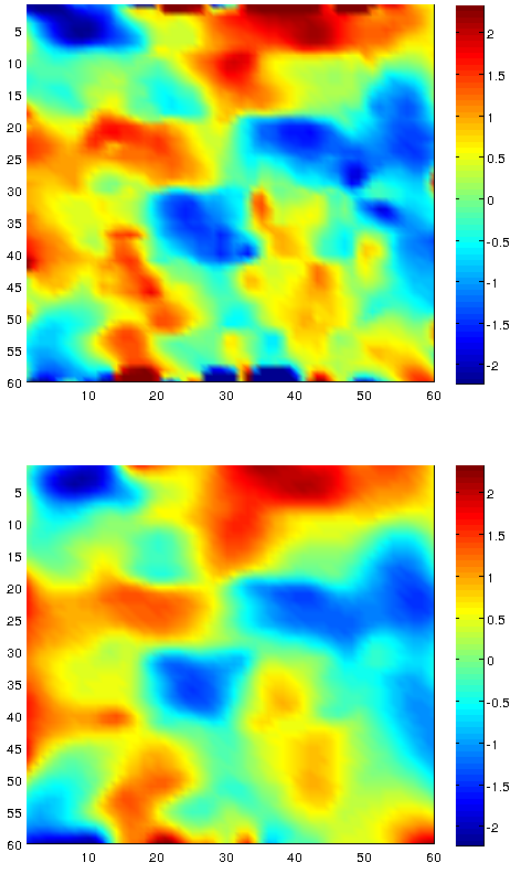

Fig. 7. LWR interpolated fields with weight rate chosen by LOOCV (top) and rCV (bottom). Compare with Figure 6.

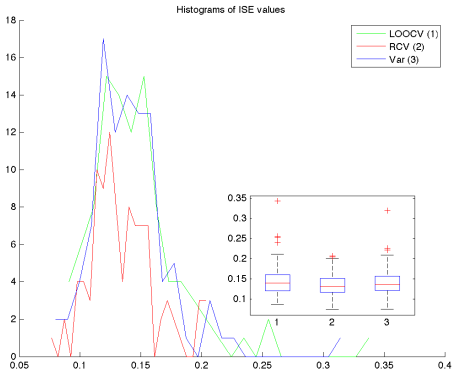

Fig. 8. ISE values for Kriging, one-dimensional design.

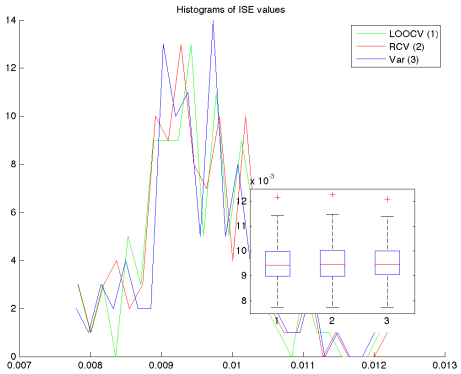

Fig. 9. Values of ISE for Kriging using, regular design.

of Kriging exist, considering different modelling approaches for the mean function of the process. Simple Kriging assumes a known mean, ordinary kriging a constant unknown mean, and universal Kriging assumes that the mean function of the process belongs to the linear span of a finite set of known functions, most commonly polynomials. The estimator is linear in the observations, and is fundamentally determined by the correlation function of the GP. In this paper we studied ordinary Kriging, considering tuning of the range parameter ρ of the Matérn covariance using the variogram, LOOCV and rCV.

Figure 8 plots the ISE observed for Kriging. It can be seen that the new criterion leads to slightly better tuning of the interpolation method, although impact is less in this case, since the observations are realisation of the assumed stochastic model: the average ISE over theses 100 realisation is 0.1454 for LOOCV, 0.1412 for the Variogram and 0.1351 for rCV, such that, again rCV leads to the best tuning.

Figure 9 shows the ISE for observations over a regular grid with the same number of points as the previous designs, showing that randomisation does not lead to ISE increase, even when its higher numerical complexity is not justified. In fact, although all three methods lead approximately to the same performance, as it should be expected since in this case the inter-design distances reflect well what happens for generic points of $\mathcal{A}$. Although the designs have the same size, remark the strong impact of design geometry on the absolute value of the prediction error.

Figures 10 and 11 show the values of ρ that were chosen by the different criteria. For the one-dimensional design (Figure reffig:line:rg) rCV and Var lead to the values of ρ closer to the true one, LOOCV seriously underestimates ρ, while for the regular design (Figure 11) CV is best able to identify the true model (although this does not map into a better ISE, as Figure 9 shows). The parameter estimates for the regular design shows similar performance of rCv and LOOC, underestimating the value of ρ, while the variogram is not significantly affected by the change of design geometry.

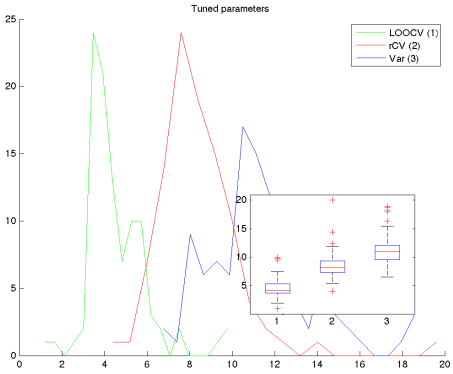

Fig. 10. Values of $\hat{\rho}$, Kriging, one-dimensional design.

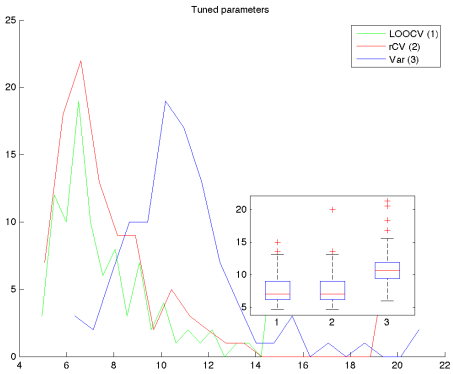

Fig. 11. Values of $\hat{\rho}$, Kriging, regular design.

TABLE I
ISE, INTERPOLATION OF FIELD IN FIG. 12.

	LWR (linear)	LWR (quad.)	Kriging
CV	0.007536	0.083414	0.003984
rCV	0.002778	0.003551	0.002887
fix	0.002947	0.008507	0.003746
Var	$\star$	$\star$	0.002961

TABLE II
ρ, FIELD IN FIG. 12.

	LWR (linear)	LWR (quad.)	Kriging
CV	2.5	10	3.3
rCV	10	26	8.6
fix	7.5	18	16.3
Var	$\star$	$\star$	9.43

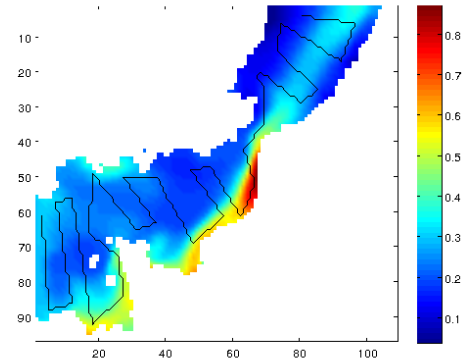

Fig. 12. Chlorophyll$_\alpha$ field produced by model MIRO&CO (courtesy of MUMM).

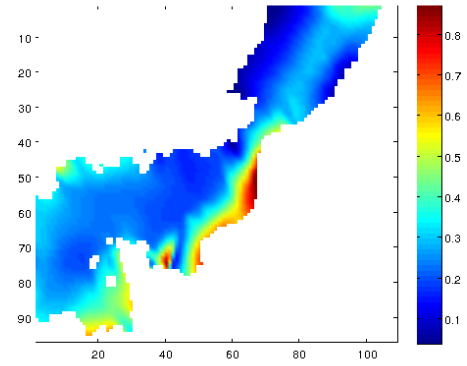

Fig. 13. Interpolated field (LWR with linear model, rCV tuning).

B. Real Data

We also compare rCV to the other methods on a realistic dataset, produced by the MIRO&CO oceanographic model [6], that predicts the annual cycle of inorganic and organic carbon and nutrients, phytoplankton, bacteria and zooplankton under realistic forcing conditions. The model covers the entire water column of the Southern Bight of the North Sea. The scalar field used, see Figure 12, is Chlorophyll$_\alpha$. The uni-dimensional survey trajectory that has been simulated is the thin black line in the Figure, that samples the field rather coarsely along a series of nearly parallel transects. Tables I and II summarise our results.

The tables show the ISE values and the identified parameters ρ for both linear and quadratic LWR and Kriging (using a Matérn covariance with $\nu = 3.2$). For the Kriging estimator, use of the variogram has also been considered. We can see that rCV consistently leads to better interpolation performance for all interpolation methods. The global minimum being achieved for linear LWR, with an ISE value of 9.7 (the interpolated field is shown in 13, remark the large errors outside the convex hull of the design), and it is only slightly degraded for the other interpolation methods, being highest for quadratic LWR (ISE = 12.4). Compare with use of standard CV, that may lead to ISE values up to 291 (quadratic LWR). The fixed CV using π_Ξ is also consistently better than CV, but is less robust than rCV, leading to an important degradation for quadratic LWR. The method that is the less sensitive to the tuning method is Kriging for which the ISE of all methods is in the interval $[10.08, 13.91]$.

Analysis of the parameters identified, Table II, confirms that standard CV always leads to more local interpolation for this type of designs (smaller values of ρ).

VI. CONCLUSION

This paper presents a new Cross Validation Criterion intended to be used on datasets acquired along linear/transect surveys, a common practice in environmental observation. The new criterion is a randomised version of standard CV, where the hold-out sets are randomly chosen in order to statistical replicate the local geometry of the design around each point in the target interpolation region. We formally motivated the new method, and illustrated its advantage with respect to standard CV (and even parametric moment-matching techniques like use of variogram in the context of kriging) in both simulated and realistic oceanographic data. Future work along the same line concerns two issues: *(i)* $C_{rCV}(\cdot)$ makes sense only in the context of prediction of spatially stationary fields, and we are studying how we can apply the same idea to enable local adjustment of the interpolator parameters, see [7]; *(ii)* the numerical complexity of $C_{rev}(\cdot)$ is much higher than standard CV, and two directions will be explored to improve its efficiency: *(a)* consider analytical approximations to $\pi_\Xi(d)$, as outlined in section III, eliminating the need to estimate π_Ξ; *(b)* consider "fast CV" approaches, that exploit the recursive version of some estimators.

ACKNOWLEDGMENTS

This work has been partially supported by the European Union through project DRONIC (FP7-ICT 611428, "Application of an unmanned surface vessel with ultrasonic, environmentally friendly system to (map and) control blue-green algae (Cyanobacteria)," http://dronicproject.com/), and by ANR (France) through project DESIRE (https://www.i3s.

unice.fr/desire/). The authors acknowledge the collaboration of the Institutes VITO and MUMM (Royal Belgian Institute of Natural Sciences) in providing some of the datasets used in this work.

REFERENCES

[1] M. Stone, *Cross-validatory Choice and Assessment of Statistical Prediction*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 36, No. 2, pp. 111-147, 1974.

[2] S. Arlot, *A Survey of Cross Validation Procedures for Model Selection*, Statistics Surveys, Vol. 4, pp 40-79, 2010.

[3] L. Pronzato and W. Muller, *Design of computer experiments: space filling and beyond*, Statistics and Computing, Springer Verlag (Germany), 22 (3), pp.681-701, 2012.

[4] W. S. Cleveland and S. J. Devlin, *Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting*, Journal of the American Statistical Association, vol. 83, no 403, p. 596610, 1988.

[5] C. E. Rassmussen, *Gaussian processes for machine learning*, MIT Press, 2006.

[6] G. Lacroix, K. Ruddick, Y. Park, N. Gypens, and C. Lancelot, C., *Validation of the 3D biogeochemical model MIRO&CO with field nutrient and phytoplankton data and MERIS-derived surface chlorophyll α images.*, Journal of Marine Systems, Vol 64, p. 66-88, 2007.

[7] G. Lacroix, K. Ruddick, Y. Park, N. Gypens, and C. Lancelot, C., *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*, Monographs on Statistics and Applied Probability 66, Chapman & Hall, 1996.