Model-free spatial Interpolation and error prediction for survey data acquired by mobile platforms

Maria-João Rendas

Model-free spatial Interpolation and error prediction for survey data acquired by mobile platforms

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

***Abstract*—The paper proposes 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. Numerical results illustrate the impact of randomised cross-validation in real environmental datasets showing that it leads to interpolated fields with smaller error at a much lower computational load. Randomised CV enables a robust parameterisation of interpolation algorithms, in a manner completely driven by the data and free of any modelling assumptions. The new method proposed here resorts to tools and concepts from Computational Geometry, in particular the Yao graph determined by the set of sampled sites. The 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.**

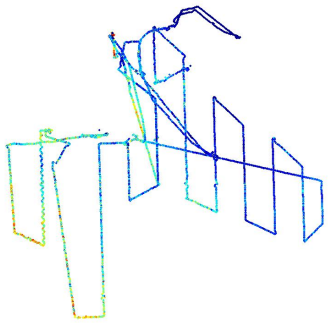

Fig. 1. Design used for a lake observation (courtesy of VITO).

I. Motivation

In environmental sciences, as in monitoring, there is often the need to supplement global earth observation data with on-site observations. The acquired measures should allow a reliable reconstruction of the values of the observed field over the entire region of study (interpolation), and, in many situations, an indication of the uncertainty that affects the reconstructed field is also important (error prediction).

The current trend is to have *in situ* observations collected by sensors carried by mobile platforms, either executing a predefined trajectory – often a series of parallel transects covering the region of interest – or implementing reactive data-driven behaviours that concentrate samples in the most interesting spatial regions, according to some user-defined criterion. In both cases, the set of sampled positions is a discrete subset of the uni-dimensional curve along which the platform traveled (its trajectory).

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.

If not explicitly taken into account, this distinctive *uni-dimensional* characteristic of the datasets acquired by mobile sensors may induce a poor performance of commonly used methods for automatic tuning of interpolators, that may result in a strong degradation of both the quality of the interpolated maps as well as of the associated predicted accuracy. This paper addresses this problem, proposing a model-free Cross Validation technique that performs robustly when applied to datasets collected by mobile sensors.

As the numerical results presented below show, application of common Cross Validation (CV) approaches [1] to *uni-dimensional* datasets may lead to very poor reconstruction of the measured field, since the underlying assumptions on which the (CV) criterion is based, valid when the sampled sites are well distributed over the region of interest, no longer hold. We proposed in [2] a randomised CV method (**rsCV**) that leads to significantly robustness and improved quality of the interpolated maps. Our numerical results show that use of simple model-free local interpolator – like Local Weigthed Regression, [5] – tuned by **rsCV** is more stable able to outperform complex model-based interpolators like Kriging [6], whenever – as it is the case with the majority of real environmental fields – the mathematical assumptions behind Kriging do not hold.

Although providing an efficient tuning of interpolation algorithms, **rsCV** as proposed in [2] is a biased estimator of the prediction error of the reconstructed maps, having the tendency to under-estimate the error for linear transect surveys, i.e. its indication is a *lower bound* of the estimation error. This paper proposes an more complex version of rCV, randomised

[1]Courtesy of VITO.

shape Cross Validation, **rsCV**, that estimates an upper-bound of the prediction error. Use of both criteria establishes thus a confidence interval for the prediction error, giving a more realistic indication to the final users about the quality of the reconstructed map. The price payed by this improved information is numerical complexity: while rCV outperforms standard Cross Validation by using hold-out sets whose *sizes* mimic the distances between points of the region and the sampled sites, **rsCV** improves on **rsCV** by considering the **shape** of the empty (sample-free) regions around each point determined by the design.

The paper summarises the basic principle behind **rCV** and presents the new Cross Validation criterion **rsCV**, resorting to tools and concepts from Computational Geometry, in particular the Yao graph determined by the set of sampled sites. The results are illustrated in real environmental datasets

II. Problem formulation

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

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

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 interpolation operator, depending on some set of user-defined parameters ρ:

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

Choice of ρ is particularly important when Ξ does not sample $\mathcal{A}$ densely.

Ideally, one would choose ρ such that the some functional $C(\cdot)$ of the reconstruction error is minimal: $\rho^\star(Y^{(K)}) = \arg\min_\rho C_{ise}(\rho; Y^{(K)})$. Several criteria can be used. We illustrate the method using the Integrated Square Error (ISE):

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

Obviously, C_{ise} cannot be computed and only estimates of C_{ise} based on the data $Y^{(K)}$ can be used to chose ρ. Model-based estimators of C_{ise} are sensitive to the correctness of the assumed models, as our numerical results below show. A method widely used by practitioners to chose algorithm's parameters is based on Cross Validation (CV) estimation of the prediction error. CV origins go at least as far as the 1930's, see the interesting discussion in [1]. It 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, which is clearly not the case for the in Figure 1.

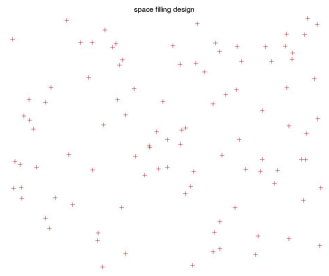

Fig. 2. Space filling design.

III. Cross Validation

Several variants of CV exist [3], but the simple description below 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}

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

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 $Y^{(-i)} = Y^{(K)} \setminus \{y_i\}$.

A necessary condition for $C_{CV}(\rho|Y^{(K)})$ to be a sensible estimate of C_{use} 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 $\mathcal{A}$. As we showed in [2] for uni-dimensional surveys this is never the case for standard CV methods, which led us to propose **rCV**, that we summarise in the next section.

IV. Randomised Cross Validation (rCV)

In a first order approximation the reconstruction error at a point $s \in \mathcal{A}$ is highly (positively) correlated to the distance of s to its closest point in Ξ, see [4]. **rCV** is fundamentally based on the statistical distribution of these distances (induced by the uniform measure in $\mathcal{A}$).

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

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

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 $\mathcal{A}$.

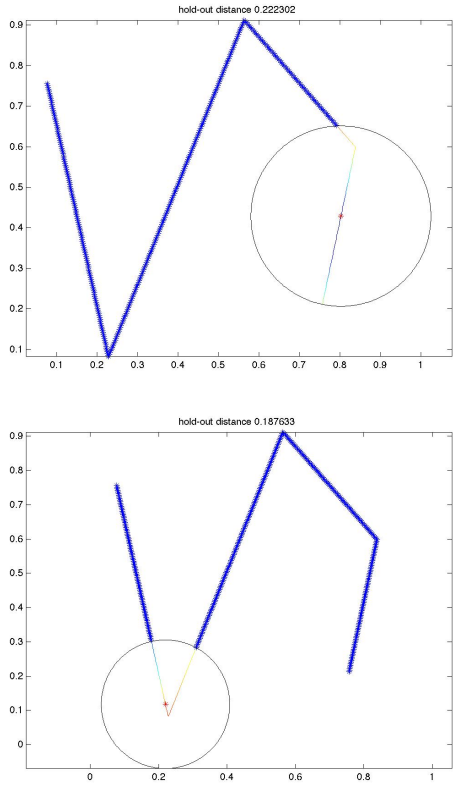

Fig. 3. Example of block-out sets $Y^{-(s,r)}$

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[2]

$$\Xi^{-(s,r)} = \Xi \cap B_r(s)^c \ , \tag{3}$$

and denote by $Y^{-(s,r)}$ the corresponding measures. 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 $\mathcal{A}$ using data $Y^{(K)}$ and design Ξ is

$$C_{rCV}(\rho|Y^{(K)}) = \mathrm{E}_{r,\xi}\left(y(\xi) - \mathbf{F}\left(\xi, Y^{-(\xi,r)}, \rho\right)\right)^2 \ . \tag{4}$$

Above, the average is computed using distributions $r \sim \pi_\Xi$ and $\xi \sim \mathcal{U}(\Xi)$, and the hold-out sets $Y^{-(\xi,r)}$ are given by (3). Obviously, numerical computation of C_{rCV} resorts to stochastic simulation the expected value in (4) being approximated by the corresponding empirical average. Note that C_{rCV} uses randomly chosen hold-out sets that statistically reflect the imbedding of Ξ in region $\mathcal{A}$, and puts CV in the actual conditions under which extrapolation will be done.

The price payed for the robustness of C_{rCV} is its higher numerical complexity, when compared to standard CV techniques, that rely on "homogenous" hold-out sets. However, it can leads to an efficiently and stable tuning of simple interpolators, which may, as the examples below show, significantly outperform more complex interpolators like Krigin at a much lower global (tuning and interpolation) computational cost.

Even if **rCV** provides a preferable alternative to common self-tuning techniques, for uni-dimensional designs **rCV** is a biased estimator of C_{ise}, predicting an interpolation error lower than the true error. This behaviour can be explained by a close analysis of the shapes of the cells of the Voronoi diagram centred an arbitrary point $s \in \mathcal{A}$ (for the set of points $\Xi \cup \{s\}$ obtained by completing Ξ with s) and those corresponding to Ξ alone. The method **rsCV** proposed in the paper overcomes this deficiency by explicitly addressing the geometry of these cells.

Before presenting it, we demonstrate the impact of **rCV** with a numerical study on real datasets.

V. Numerical results

Kriging with range estimated by variogram is presently considered as the preferred interpolation technique in geostatistics. This interpolator is based on strong assumptions about the observed field necessary to be able to estimate the parameters of the model for which Kriging is the optimal predictor. This good performance comes at a significant computational price, as well as complex implementation issues, since the method should ideally simultaneously process the entire observations, which makes it problematic for large environmental datasets.

When Ξ is dense in $\mathcal{A}$, this interpolator produces very good results, and, as importantly, is able to indicate the uncertainty of the estimated map. Unfortunately, the distributional and stationarity assumptions on which it is based are seldom satisfied by real environmental fields, which can lead to very poor behaviour for poor designs. We present below comparison of three estimates: *(a)* Ordinary Kriging (OK) with variogram tuning; *(b)* Locally Weighted Linear Regression (LWLR) with leave-one-out CV; and *(c)* LWLR with **rCV** tuning.

Five different parameters (depth, temperature, CHl_α, turbidity and Ph) were recorded for this survey. Due to space reasons we concentrate on Clh_α and Ph. Figures 4 and 5 show the interpolated Ph. We can see that the map produced by OK is strongly over-smoothed, while the local methods are able to retain the information in the dataset, **rCV** being smoother than LOO-CV while still retaining detailed variation information.

Figures 6 and 7 sow the opposite situation: Kriging produces a good interpolation of the dataset, comparable to the result of **rCV**.

The results above clearly show the robustness of simple local imterpolars combined with **rCV**. We stress that for the 5 parameters in the processed real dataset the behaviour illustrated above was consistent: **rCV** is more stable than OK, producing results of similar quality when OK works well while the total computation load was 6–7 times smaller. It always outperforms LOO-CV as it should be expected.

As we said before, **rCV** produces a negatively biased estimate of the interpolation error C_{ise} (a *lower bound*). We

[2]Notation B^c denotes the complement of set B in $\mathcal{A}$.

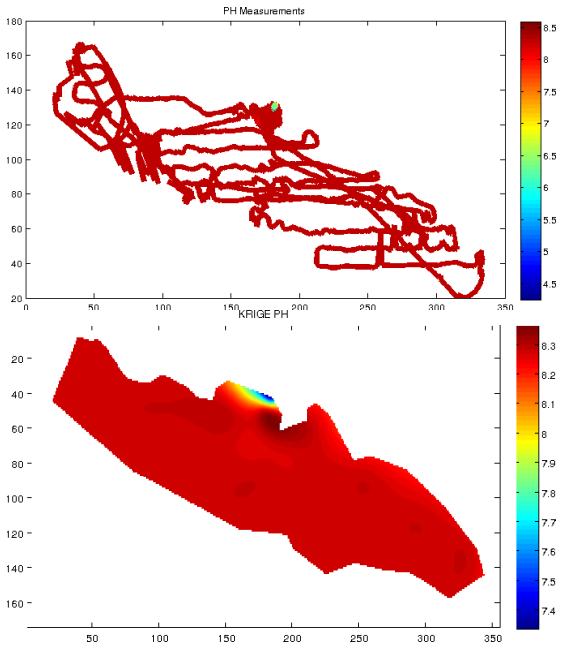

Fig. 4. Ph: measures (top) and OK map.

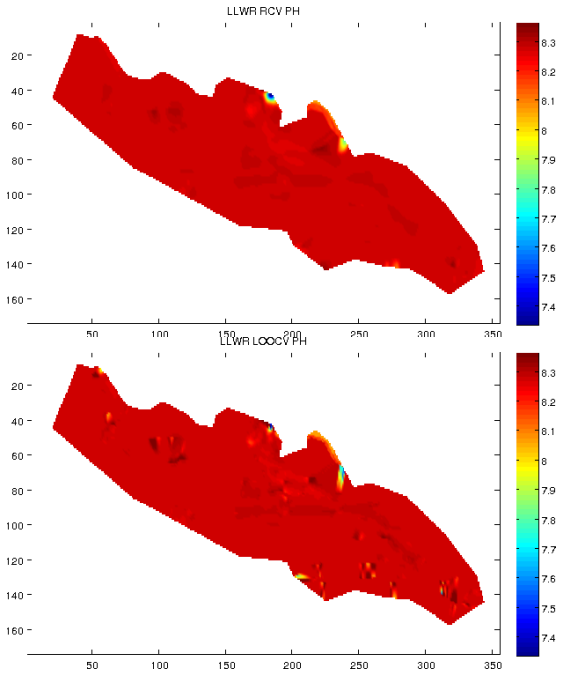

Fig. 5. Ph:**rCV** (top) and LOO-CV maps.

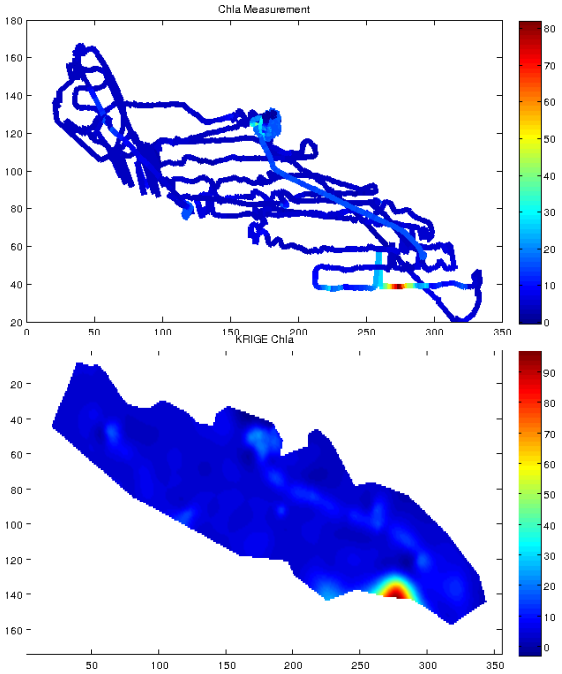

Fig. 6. Chl_α: measures (top) and OK map.

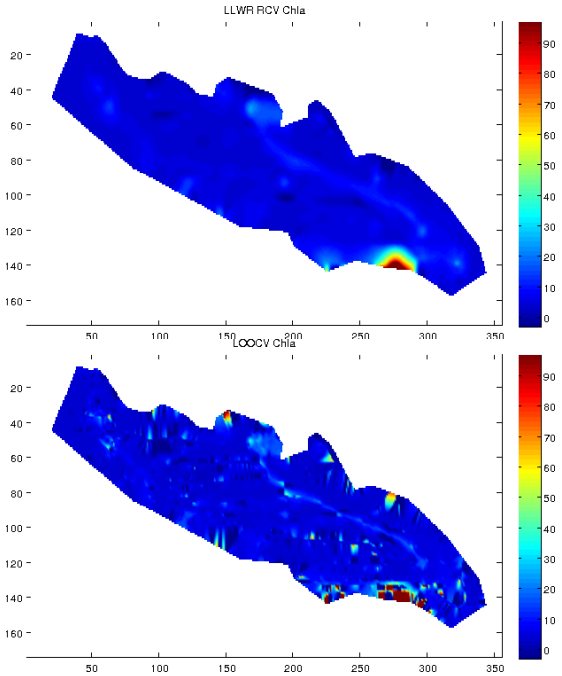

Fig. 7. Chl_α: **rCV** (top) and LOO-CV maps.

outline below a modified version of **rCV** that leads to an *upper bound* of C_{ise}.

VI. Randomised Shape Cross-Validation

The modified rhCV criterion has the same generic expression as **rCV**:

$$rhCV(\rho) = \mathrm{E}_s\mathrm{E}_i \left(y_i - \hat{y}(\xi_i|\Xi^{\xi,\pi_s})\right)^2 \quad .$$

As **rCV**, it can only be computed by Monte Carlo computation of the corresponding excepted values. The main difference resides in the geometry of the hold-out sets. While for **rCV** the hold-out sets have all the same geometry, independently of the point s drawn, depending only on the distance between s and the point in Ξ closest to it. The new **rsCV** draws on a more complex characterisation of the "free space around s. More precisely, we rely on the notion of Yao graph of a set of points $\mathcal{X}$ []. This graph relies on a partition of the space around each point $s \in \mathcal{X}$ in a fixed number (n) of cones centered at s. The graph is obtained by creating an edge between s and its n nearest neighbours in each cone. Figure 8 illustrate the concept and the hold-out sets used by **rsCV**. The green dots are the design points, and the black starts two points s randomly drawn in $\mathcal{A} \setminus \Xi$. The polygons around these points connect the Yao-neighbours of s (elements of Ξ). The red-encircled green dots are the hold-out sets for the two points s shown. Note that this method is able to replicate the error residuals at the boundaries of $\mathcal{A}$.

VII. Conclusion

This paper proposes 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 an improved version of a randomised version of standard CV, where the hold-out sets are randomly chosen in order to statistical replicate the local geometry of the free space around each point in the target interpolation region. We illustrated the advantage of randomised with respect to standard CV techniques (and even parametric moment-matching techniques like use of variogram in the context of kriging) real data. On-going work concerns improvement of the numerical efficiency of the method, which requires the computation of nearest neighbours in several directions.

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 Institute VITO in providing some the dataset used.

In the framework of this project the authors institution implemented an interpolation web service that integrates, amongst other, the cross-validation techniques and interpolators use in the paper. This service is open to interested users outside of the project by request to the authors.

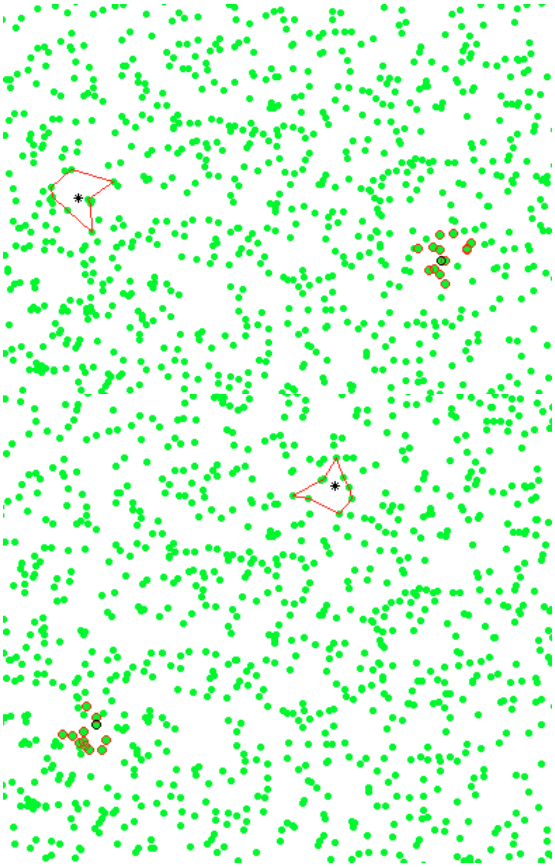

Fig. 8. Hold-out sets based on the Yao graph.

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] Y. Li, M.-J. Rendas, *Tuning interpolation methods for environmental uni-dimensional (transect) surveys*, Proc. Oceans 2015, Washington D.C. USA, October 2015.

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

[4] 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.

[5] 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.

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

[7] Jianqing Fan, Irene Gijbels, *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*, Monographs on Statistics and Applied Probability 66, Chapman & Hall, 1996.