Comparing Kriging, Spline, and MLR in Product Properties Modelization: Application to Cloud Point Prediction

Jean-Jérôme da Costa Soares, Fabien Chainet, Benoît Celse, Marion Lacoue-Nègre, C. Ruckebusch, Didier Espinat

► To cite this version:

Jean-Jérôme da Costa Soares, Fabien Chainet, Benoît Celse, Marion Lacoue-Nègre, C. Ruckebusch, et al.. Comparing Kriging, Spline, and MLR in Product Properties Modelization: Application to Cloud Point Prediction. *Energy & Fuels*, 2018, 32 (4), pp.5623-5634. 10.1021/acs.energyfuels.7b04067 . hal-01810043

HAL Id: hal-01810043

<https://hal.science/hal-01810043>

Submitted on 30 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Comparing kriging, spline and MLR in product properties modelization.

Application to Cloud Point prediction

J.J. Da Costa¹, F. Chainet¹, B. Celse^{1*}, M. Lacoue-Nègre¹, C. Ruckebusch², D. Espinat¹

¹IFPEN, Etablissement de Lyon – Rond-Point de l'Echangeur de Solaize

BP 3 – 69360 Solaize – France

²LASIR CNRS, Université Lille Nord de France, Sciences et Technologies, Cité scientifique, bât C5, 59665

Villeneuve d'Ascq Cedex - France

E-mail address: Benoit.celse@ifpen.fr

Abstract

In the most part of chemometrics application concerning the prediction of physico-chemical properties, regression models are preferred as they are easy to implement and their posterior analysis provide simple interpretations. Interpolation methods are most currently used in geostatistics applications where the dimension of study area is generally limited. In this work, we proposed to develop kriging or splines models for predicting petroleum products properties. Kriging and splines have different foundations as the former is based on stochastic assumptions and the latter is built on deterministic approach. A well-illustrated comparison of both methods was carried out through three suitable examples to point their similarities and their divergences. The advantages of using kriging or / and splines instead of classical regression models were also discussed. Results pointed out the flexibility of interpolation methods as they provide good accuracy for linear and nonlinear cases. They also confirmed previous studies which pointed out equivalence between kriging and spline models performances in some situations. However, kriging approach has more valuable aspect among other interpolation methods since it provides a measure of prediction uncertainties. Kriging modeling were finally compared to multilinear regression for the prediction of diesel cloud point ranging from -39°C to -12°C. Models performances pointed out that kriging enables to improve both accuracy and robustness.

Keywords: Kriging, Multilinear Regression, Splines, Interpolation, Hydrocracking, Diesel, Cloud point

1 Introduction

Interpolation methods were originally developed to solve problem related to geophysics. They are based on local approach since they operate within a small area around the point being estimate and capture local or short-range variation, that contrasts with classical regression (Burrough and McDonnell 1998). Interpolation methods may be divided into two groups (Arnaud and Emery 2000): (1) the deterministic methods which have to map from known points according to a degree of similarity or smoothing; (2) the stochastic methods which estimate the spatial autocorrelation between known points and then consider the spatial configuration of these points around the location to predict.

Kriging is an interpolation method based on stochastic multiGaussian assumptions (Goovaerts 1997). Its goal is to provide the best linear unbiased estimate (B.L.U.E.) and was first used in applied geostatistics for modeling properties related to natural resources (Cressie 1990; Cressie 1993; Goovaerts 1997; Isaaks and Srivastava 1989). Later, kriging models were applied to data of deterministic simulation models (Jones, Schonlau and Welch 1998; Santner, Williams and Notz 2003; Simpson et al. 1998). These applications are particularly different from geostatistics since the dimension of inputs may be greater than 2 (whereas geostatistics considers only two-dimensional coordinates). More recently, other applications of kriging were developed such that metamodeling which is an increasingly need in computing sciences (Kleijnen). Furthermore, the stochastic approach related to kriging is very essential since it provides a measure of the prediction uncertainty that depends on the data configuration. This characteristic was investigated in some works (Baar et al. 2014; Wang).

Splines interpolation has a deterministic approach and was originally developed for ship-building (Ahlberg, Nilson and Walsh 1967; Schumaker 1981). This method is currently used in non-parametric statistical learning for fitting piecewise functions (Hastie, Tibshirani and Friedman 2001) and for solving differential equations (Ascher, Pruess and Russel 1983; Schoenberg). Although some previous studies (Dubrule 1984; Matheron 1975) showed that kriging and spline models generally provide close predictive performances or may be equivalent under some assumptions, their goals and then their foundations are quite different.

Predicting petroleum products properties is an increasingly need for refining industry because of economic considerations. Thus, Machine learning is more and more used to develop predictive models from physicochemical analytical data. The most current models are based on classical (multilinear and nonlinear) regression referring to some equations which describe physicochemical phenomena (kinetic reactions) or derived from empirical observations (Becker et al. 2015; Becker et al. 2016; Celse, Da Costa and Costa 2016; Cookson

1
2
3 et al. 1985; Cookson, Iliopoulos and Smith 1995; Cookson and Smith 1990; Cookson and Smith 1992).
4
5 However, regression models require to identify the adapted analytical expression for the regression function.
6
7 That may be very complex in some situations. Many chemometrics applications in modeling petroleum products
8
9 properties were also proposed using partial least squares (PLS) regression (Abdel-Rahman et al. 2014; Braga,
10
11 Santos and Martins 2014; Santos et al. 2005; Sastry et al. 1998). Although these models provide relatively good
12
13 performances, they use spectroscopic data obtained from analytical techniques of petroleum fractions
14
15 characterization such that nuclear magnetic resonance (NMR) or Infrared spectroscopy. However, this type of
16
17 data are not currently available for refiners. Finally, some authors (Kapur 2004; Marinovic, Bolanca and Ukc
18
19 2012; Santos et al. 2005; Wang, Dong and Sun 2010) proposed to use artificial neural networks (ANNs) for
20
21 modeling of complex properties. The risk in using ANNs is that their optimization requires to perform very
22
23 stringent statistical steps to insure the model robustness. That generally implies a high number of samples which
24
25 is generally not available in petroleum industry. Thus, their application most time result in overlearning.

26
27 The aim of this work is to illustrate first the advantages of using interpolation methods instead of classical
28
29 regression models through basic relevant simulated examples and to point the usefulness kriging compare to
30
31 splines approach. Then, we propose to apply kriging for modeling the cloud point (CP) of diesel cuts produced
32
33 from hydrocracking process by using other physicochemical properties of petroleum cuts.

34
35 This paper is structured as follows: (1) all (simulated and real) databases used in this work are first detailed;
36
37 (2) classical regression, splines interpolation and kriging are reviewed and implemented using simulated data. (3)
38
39 the obtained performances are then discussed before applying kriging for modeling of diesel fuel CP.

40
41 **2 Material and methods**

42
43 All the datasets used to test predictive methods in the present work are described here. Main theoretical aspects
44
45 of the used methods are then recalled.

46
47 **2.1 Parametric regression models**

48
49 In this section, z denotes a variable to model and x a set of explanatory variables. It is assumed that a set of
50
51 measured values $(y_1, \dots, y_n)$ at n sampled data $(x_1, \dots, x_n)$ was observed. In addition, all measured values are
52
53 considered to be submitted to an error due to experimental process such that:

54
55
$$y_i = z_i + \epsilon_i, \quad i = 1, \dots, n \tag{Eq. 1}$$

56
57 where ϵ_i denotes the measurement error related to the i^{th} observation.

The main idea of parametric regression models is to approximate z by a known mathematical function f such that (Hastie et al. 2001):

$$y_i = f(x_i, \theta) + \varepsilon_i \quad (\text{Eq. 2})$$

where θ denotes the set of parameters related to the analytic expression of f and ε_i is the modelling error at the i^{th} sampled data. In what follows, θ is assumed to be a p -dimensional vector ($n > p > 1$).

2.1.1 Multilinear regression

In case of multilinear regression (MLR), f is assumed to have the following form (Azaïs and Bardet 2012):

$$f(x, \theta) = \theta_0 + x^T \theta \quad (\text{Eq. 3})$$

where x^T denotes the transpose vector of x . Let X denotes the design matrix *i.e.* the matrix whose the i^{th} row contains the coordinates of the i^{th} sampled point, and $y = (y_1, \dots, y_n)^T$. The goal of MLR is to estimate “as well as possible” the set of parameters θ . The most current approach consists on introducing Gaussian assumptions for modeling errors. The ε_i are then generally assumed as independent identically normally distributed random variables with mean zero value and of unknown variance σ^2 . Under these assumptions, the least of squares (LS) estimators which consists on minimizing the sum of squared deviation has optimal characteristics according to inferential statistics (Azaïs and Bardet 2012; Hastie et al. 2001). Furthermore, it has been shown that for MLR, the LS-estimator $\hat{\theta}^{LS}$ can be expressed as a linear combination of measured values, (Hastie et al. 2001):

$$\hat{\theta}^{LS} = (X^T X)^{-1} X^T y \quad (\text{Eq. 4})$$

This result may be extend to generalized multilinear regression models (Azaïs and Bardet 2012; Dror and Steinberg 2008). Let g denotes a mathematical differentiable function such that:

$$f(x, \theta) = \theta_0 + g(x)^T \theta \quad (\text{Eq. 5})$$

then, the LS-estimator has the following generalized expression (Hastie et al. 2001):

$$\hat{\theta}^{LS} = (g(X)^T g(X))^{-1} g(X)^T y \quad (\text{Eq. 6})$$

where $g(X)$ denotes the matrix of x by the transformation g .

Although the LS-estimator calculation results in a relatively simple matrix multiplication, it is rigorously based on deterministic approach which does not include the measurement error. Thus, other estimators related to the concept of regularization were developed in statistical learning (Ridge, Lasso, *etc.*) (*Bolasso: Model Consistent Lasso Estimation Through the Bootstrap*; Hastie et al. 2001; Mohri, Rostamizadeh and Talwalkar 2012).

The main advantages of using regression models is that the relation between input variables and the modeled response are obvious to interpret. However, find the adapted form of the regression function f may be very complex, especially in nonlinear case.

2.2 Interpolation methods

All the definitions made in introduction of the previous section are still effective in what follows. The basic idea of interpolation is to estimate the value of a variable at a given point within the study area by computing a weighted average of the observed value (Arnaud and Emery 2000; Cressie 1993). Interpolation methods can be divided into two groups:

- deterministic method where $z(x)$ is considered as a “regionalized variable” (Wackernagel 2003),
- probabilistic method where $z(x)$ is assumed to be a realization of random variable (Goovaerts 1997).

Both types of interpolation methods are illustrated below through spline and kriging approaches.

2.2.1 Spline interpolation

The aim of splines interpolation is to draw contour lines “as smooth as possible”, that is a map which looks like what a draftsman would obtain manually (Dubrule 1984). The splines consist of polynomials of degree m being local. The polynomials describe pieces of a line or surface. For degree $m = 1, 2$ or 3 , a spline is called linear, quadratic or cubic respectively (Ahlberg et al. 1967). In order to simplify the theoretical aspect of spline the study area is assumed to be a subset of the two-dimensional space. Let (u, v) denotes the coordinates of x . The simplest approach of spline so called “Interpolating spline” consists on assigning a local estimate $\hat{z}(u, v)$ which is solution of the following minimization problem:

$$\min_{\phi} J_m(\phi) = \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2 \phi}{\partial u^2} \right)^2 + \left(\frac{\partial^2 \phi}{\partial v^2} \right)^2 + 2 \left(\frac{\partial^2 \phi}{\partial u \partial v} \right)^2 \right] du dv \quad (\text{Eq. 7})$$

among all the functions ϕ honoring the sampled data points. $J_m(\phi)$ can be interpreted, in first approximation as the bending energy of a thin plate represented by ϕ (Dubrule 1984). Thus, the function that minimizes $J_m(\phi)$ takes the shape of a thin plate which would be forced to pass through sampled data points. However, this approach does not consider the measurement error related to the observed data. In practice, a most useful approach called “smoothing spline” is preferred. It consists on adding a term of regularization to the objective function. The function to minimize become (Dubrule 1984; Wahba and Wendelbeger 1980):

$$\sum_{i=1}^n w_i^2 [\phi(x_i) - y_i]^2 + \lambda J_m(\phi) \quad (\text{Eq. 8})$$

λ is called the “smoothing parameter” and the weights w_i^2 are generally chosen inversely proportional to the error variance. Consequently, the added term force the solution to pass “not too far” from the measured values (Dubrule 1984). Duchon *et al.* (Duchon 1975) showed that interpolating and smoothing splines have the same expression. For the two-dimensional case:

$$\hat{z}^{sp}(\mathbf{u}_0, \mathbf{v}_0) = \mathbf{a}_0 + \mathbf{a}_1 \mathbf{u}_0 + \mathbf{a}_2 \mathbf{v}_0 + \sum_{i=1}^n \mathbf{b}_i r_i^2 \log r_i \quad (\text{Eq. 9})$$

where r_i represents the Euclidian distance between $\mathbf{x}_i$ and $\mathbf{x}_0$, *i.e.*:

$$r_i^2 = (\mathbf{u}_0 - \mathbf{u}_i)^2 + (\mathbf{v}_0 - \mathbf{v}_i)^2 \quad (\text{Eq. 10})$$

However, their coefficients $\mathbf{a}_0$, $\mathbf{a}_1$, $\mathbf{a}_2$ and the $\mathbf{b}_i$ are obviously different. In case of smoothing spline they are obtained by solving the following system:

$$\begin{cases} \sum_{i=1}^n \mathbf{b}_i = \mathbf{0} \\ \sum_{i=1}^n \mathbf{b}_i \mathbf{u}_i = \sum_{i=1}^n \mathbf{b}_i \mathbf{v}_i = \mathbf{0} \\ \hat{z}^{sp}(\mathbf{u}_i, \mathbf{v}_i) + \frac{\mathbf{b}_i}{w_i^2} \lambda = \mathbf{y}(\mathbf{u}_i, \mathbf{v}_i), i \in \{1, \dots, n\}, \end{cases} \quad (\text{Eq. 11})$$

These results may be extended to p -dimensional case. The generalized expression of $J_m(\phi)$ is given below (Wahba and Wendelbeger 1980):

$$J_m(\phi) = \sum_{\alpha_1 + \dots + \alpha_p = m} \frac{m!}{\alpha_1! \alpha_2! \dots \alpha_p!} \int \dots \int_{\mathbb{R}^p} \left(\frac{\partial^m \phi}{\partial x_1^{\alpha_1} \dots \partial x_p^{\alpha_p}} \right)^2 d\mathbf{x}_1 \dots d\mathbf{x}_p \quad (\text{Eq. 12})$$

The solution of this generalized minimization problem is discussed in (Wahba and Wendelbeger 1980).

Developing smoothing spline model require to find a suitable value for both smoothing parameter and polynomial order. In practice, a statistical method called generalized cross-validation is usually computed in the learning step (Wahba and Wendelbeger 1980).

2.2.2 Kriging modeling

The goal of kriging is to obtain the best linear unbiased estimate (B.L.U.E.) according to the observed data. Let $\hat{z}^{SK}(\mathbf{x}_0)$ denotes the kriging estimate of the response z at a given point $\mathbf{x}_0$. Thus, it must satisfy the following assumptions:

$$\begin{cases} (1) \hat{z}^{SK}(\mathbf{x}_0) = \sum_{i=1}^n \lambda_i \mathbf{y}(\mathbf{x}_i) \\ (2) E[\hat{z}^{SK}(\mathbf{x}_0)] = \mathbf{z}(\mathbf{x}_0) \\ (3) \hat{z}^{SK}(\mathbf{x}_0) = \arg \min_{\hat{z}^L(\mathbf{x}_0)} \text{Var}[\hat{z}^L(\mathbf{x}_0) - \mathbf{z}(\mathbf{x}_0)] \end{cases} \quad (\text{Eq. 13})$$

where $\hat{z}^L(\mathbf{x}_0)$ denotes any linear estimator of $z(\mathbf{x}_0)$. The definition above calls for a predefined probabilistic model which is generally based on a stochastic approach. Let A denotes the study area. Thus, it is assumed that the set of values $\{z(\mathbf{x}), \mathbf{x} \in A\}$ describes a multiGaussian stochastic process (Stein 1987). According to the process characteristics, two types of kriging may be distinguished. The method is called “simple or ordinary”

1 kriging when the stochastic process is assumed to be stationary of order two, *i.e.* the expected value is invariant
2 within the study area and the covariance between two measured values only depend on the deviation vector of
3 their related locations (Goovaerts 1997). Otherwise, the method is called “universal” and the expected value has
4 the analytical expression of a polynomials (Dubrule 1984).

5 **2.2.2.1 Simple kriging predictor**

6 The multiGaussian assumptions related to simple kriging results in the two following
7 constraints:

8
$$\begin{cases} E[y(x)] = m \\ Cov(z(x_i), z(x_j)) = C(x_j - x_i) \end{cases} \quad (\text{Eq. 14})$$

9 where $E[y(x)]$ is the expected value of $y(x)$. The conditions of optimality verified by the kriging predictor lead
10 to solve the following linear system (Goovaerts 1997):

11
$$\left\{ \sum_{i=1}^n \lambda_i [C(x_j - x_i) + \delta_{ij} \sigma_\epsilon^2(x_j)] = C(x_j - x_0), j \in \{1, \dots, n\} \right. \quad (\text{Eq. 15})$$

12 where

13
$$\delta_{i,j} = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases} \quad (\text{Eq. 16})$$

15 and $\sigma_\epsilon^2(x_j)$ is the variance error related to the j^{th} measured value.

16 This result implies that kriging estimator can be written as a linear combination of the $\{C(x_i - x_0), i = 1, \dots, n\}$,
17 *i.e.* (Matheron 1971):

18
$$\hat{z}^{SK}(x_0) = \sum_{i=1}^n b_i C(x_i - x_0) \quad (\text{Eq. 17})$$

19 where the coefficients b_i are solution of the following system:

20
$$\begin{cases} \sum_{i=1}^n b_i = 0 \\ \hat{z}^{SK}(x_j) + \sigma_\epsilon^2(x_j) b_j = y(x_j), j \in \{1, \dots, n\} \end{cases} \quad (\text{Eq. 18})$$

21 The expression above is analogous to Eq.11 which is related to splines interpolation. Thus, the kriging predictor
22 may be interpreted as an exact interpolator.

23 **2.2.2.2 Universal kriging predictor**

24 A more generalized kind of kriging so called “universal kriging (UK)” or “kriging regression” was also used in
25 the literature (Dubrule 1984; Goovaerts 1997; Matheron 1975). It consists on introducing a polynomial trend
26 depending on the available observations, such that the may be expressed as follows:

27
$$z(x) = \sum_{l=0}^m \beta_l f^l(x) + \alpha(x) \quad (\text{Eq. 19})$$

where m denotes the degree of the polynomial trend and $\alpha(x)$ is the approximation error. The analytical expression of components $f^l(x)$ are given in Appendix C. It has been shown that the UK weights predictor are solution of the following system:

$$\begin{cases} \sum_{i=1}^n b_i f^l(x_i) = 0, \quad l = 0, \dots, m \\ \hat{z}^{SK}(x_j) + \sigma_\epsilon^2(x_j) b_j = y(x_j), \quad j \in \{1, \dots, n\} \end{cases} \quad (\text{Eq. 20})$$

All the system related to the calculation of kriging weights predictor depends obviously on the chosen covariance model C . Some examples of covariance function are given in Table 1.

Table 1 : Example of covariance models

Type of covariance	Expression
Exponential	$\sigma_0^2 e^{-\sum_{j=1}^p \theta_j s_j - w_j }$
Gaussian	$\sigma_0^2 e^{-\sum_{j=1}^p \theta_j (s_j - w_j)^2}$
Linear	$\max\{0, 1 - \theta_j s_j - w_j \}$
Spherical	$1 - 1.5\xi_j + 0.5\xi_j^3, \xi_j = \min\{1, \theta_j s_j - w_j \}$

In the simplest case, the measurement errors variance is assumed to be homogeneous around the study area. In case of a heterogeneous errors variance the method is then called “cokriging” (Dubrule 1984; Goovaerts 1997). Kriging also requires previous analysis to choose the most adapted covariance structure. Geostatisticians use a graphical recognition tool called “semi-variogram” (Goovaerts 1997; Isaaks and Srivastava 1989). However, there is no method that enables to compute this step automatically. In this work we only used a Gaussian model (see **Error! Reference source not found.**) for all developed kriging models. Thus, the learning step consists on estimating intrinsic parameters of Gaussian covariance (including a possible nugget effect, see Appendix E). This step is usually computed by maximizing the likelihood function (see Appendix C) (Isaaks and Srivastava 1989). In what follows, only simple kriging was considered.

2.3 Prediction uncertainty

Most of the predictive methods are deterministic as they have no assessment of errors with predicted values. The traditional approach for modeling prediction uncertainty at an unsampled location x_0 consists on computing a

minimum error variance $\hat{y}(x_0)$ of the unknown value $y(x_0)$ and the associated variance $\sigma^2(x_0) = \text{Var}[\hat{y}(x_0) - y(x_0)]$. The estimate and error variance are then typically combined to derive a Gaussian-type confidence interval centered on the estimate value (Goovaerts 1997). For example, the 95% confidence interval is taken as:

$$\text{Prob}\{y(x_0) \in [\hat{y}(x_0) - 2\sigma(x_0), \hat{y}(x_0) + 2\sigma(x_0)]\} = 0.95 \quad (\text{Eq. 21})$$

For example, it has been shown that an estimate of the minimum error variance related to MLR model is expressed as follows (Azaïs and Bardet 2012):

$$\hat{\sigma}^2(x_0) = \hat{\sigma}^2(\mathbf{1} + x_0'(X^T X)^{-1}x_0) \quad (\text{Eq. 22})$$

where $\hat{\sigma}^2$ is the commonly used estimate of the modeling errors variance which is defined as (Azaïs and Bardet 2012):

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{Eq. 23})$$

A more rigorous approach consists on assessing first the uncertainty about the unknown, then deduce an estimate optimal in some appropriate sense (Journel and Rossi 1989; Srivastava 1987). Let $u(x_0)$ denotes the random variable modeling the uncertainty about $y(x_0)$. The distribution function $F(x_0, z) = \text{Prob}\{u(x_0) \leq z | (I)\}$ made conditional to the available information (I) fully models that uncertainty in the sense that probability intervals can be derived, such as:

$$\text{Prob}\{u(x_0) \in [a; b] | (I)\} = F(x_0, b) - F(x_0, a) \quad (\text{Eq. 24})$$

where a and b are scalars. Note that these probability intervals are independent of any particular estimate $\hat{y}(x_0)$ of the unknown value $y(x_0)$, but only depend on the available information (I) and x_0 . Each conditional probability distribution function provides a measure of local uncertainty in that it relates to a specific location x_0 . Under the multiGaussian assumptions related to simple kriging (paragraph 2.2.2.1), it has been shown that the mean and the variance of the conditional distribution function at x_0 are equal to the simple kriging estimate $\hat{z}^{SK}(x_0)$ and simple kriging variance obtained from the available sampled data (Journel and Huijbregts 1978).

2.4 Data sets

2.4.1 Simulated data

Dataset 1 and dataset 2 are represented in Fig. 1a and Fig. 1b respectively. Each dataset contains 50 points uniformly distributed around $[-3; 3]^2$. For the dataset 1, the function to model is affine. For the dataset 2, the function to model is rational fraction. Their analytical expression is given in Table 2.

Fig. 1 : a) 2D plots of dataset 1 in the corresponding study area; b) 2D plots of dataset 2 in the corresponding study area

Dataset 3 is represented in Fig. 2. In this case, the function to model is a sum of inverse polynomials of order 2 which may be interpreted as potential functions (Table 2).

Fig. 2 : 2D plots of dataset 3 in the corresponding study area

Three sets of three-dimensional data were then simulated using the same following process:

1. A study area D was defined as a subset of $\mathbb{R}^2$.
2. 2D points were then randomly generated around the study area as sampled data
3. A predefined bivariate function f was applied to each sampled data and a “white noise” (*i.e.* normally distributed random variable with mean zero and a fixed variance) was added to the result to obtain the corresponding observed responses.

Thus, each data set consists on n triplet (x, y, z) such as

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

$$z = f(x, y) + \varepsilon \tag{Eq. 25}$$

Where ε were chosen as a normally distribute random variable with mean zero and a fixed variance σ^2 . Details are given in Table 2.

In the third situation, we considered a function f which may be interpreted as the sum of two potential functions centered in $P_1 = (-0.5; -0.5)$ and $P_2 = (0.5; 0.5)$ respectively. The sampled points were generated such that they describe two well-identified clusters. Then, the coordinates of points were assumed to follow a two-dimensional Gaussian distribution (see Appendix B). Half of coordinates were distributed such that:

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

$$(x_i, y_i) \sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}; \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right) \tag{Eq. 26}$$

The second part were distributed such that:

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

$$(x_i, y_i) \sim \mathcal{N}\left(\begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}; \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right) \tag{Eq. 27}$$

Table 2 : Information about simulated data

Type of function	Analytical expression	Study area	Noise variance
Affine	$z = f(x, y) = 2x - 3y + 4$	$[-3; 3] \times [-3; 3]$	1
Rational	$z = f(x, y) = \frac{(x^3 - 3x)}{1 + y^2}$	$[-3; 3] \times [-3; 3]$	1
Sum of potential	$z = f(x, y) = \frac{0.3}{0.3 + (x + 0.5)^2 + (y + 0.5)^2} + \frac{0.7}{0.7 + (x - 0.5)^2 + (y - 0.5)^2}$	$[-1; 1] \times [-1; 1]$	0.025

2.4.2 Real data

Diesel samples were produced using hydrocracking (HCK) process. A block flow diagram for data acquisition is given Fig. 3. Vacuum Gas Oil (VGO) was analyzed before being hydrotreated and hydrocracked in hydrotreatment (HDT) and HCK reactor successively. Total effluent obtained from HCK reactor provide is then distilled to obtain diesel cuts. 56 diesel samples with CP varying from -12°C to -39°C were produced from 7 different VGOs using different catalysts and under various operating conditions. Their properties are given in

Table 3. In refinery, petroleum cuts are usually characterized using some basic properties such as density (d), refractive index (n) or simulated distillation (Riazi 2005). The details of properties measured to characterize diesel as well as their corresponding of standard measurement methods are given in Table 4. Boiling ranges are represented by a series of values ($T_i, i = 5, 10, 15 \dots, 95$) such that T_i refers to the temperature on simulated distillation following ASTM D2887 method (ASTM D2887 2016). $C_{370^+, reactor\ input}$ and $C_{370^+, reactor\ output}$ are the part of petroleum cuts that correspond to a boiling range up to 370°C measured before (on VGO) and after (on the unconverted oil or UCO) the given process respectively. Thus, X_{370} is the conversion of 370^+ petroleum cut (Becker et al. 2016).

Fig. 3 : Block flow diagram for data acquisition

Table 3: Properties of the seven different VGO used to produce diesel samples

VGO	Origin	Density at 15°C (g/cm ³)	Nitrogen (wt ppm)	Sulphur (wt%)	IBF – FBF (°C)	Kinematic viscosity at 70°C	Kinematic viscosity at 100°C
B	Gansu (Chine)	0,8974	1190,00	1,0409	348,00 – 584,00	82,25	12,55
C	Iranian Saniya	0,9375	1300,00	2,8743	321,90 – 615,10	124,74	11,11
G	Husky	0,9580	1080,00	3,0300	215,00 – 565,00	126,70	25,20
K	SR	0,9346	1755,00	2,2375	337,70 – 632,50	25,24	9,73
M	Hoil+Oural	0,9314	2160,00	1,0777	292,00 – 606,20	81,87	8,06
N	Etats-Unis	0,9208	1160,00	0,2974	158,20 – 581,40	32,07	4,69
P	Mix of Arabian Light/Basrah (Irak) 85/15	0,9284	1395,00	1,8921	316,00 – 624,00	133,24	11,93

Table 4 : Measured and evaluated properties of feedstock and diesel samples

Property	Petroleum cut	Standard Methods	References
Cloud point (CP)	diesel	NF ISO 23015	(NF EN ISO 23015 1994)
Simulated distillation	diesel	ASTM D2887	(ASTM D2887 2016)
$(T_i, i = 5, 10, 15, \dots, 90, 95)$			

For the prediction of diesel CP, samples were divided into training (40 samples) and prediction (16 samples) sets. Training samples were selected using a space filling algorithm (Santiago, Claeys-Bruno and Sergent 2012) in order to ensure that most predictive situations may be rigorously considered as interpolation cases. The projection of training samples in Conversion- T_{95} space is plotted in Fig. 4. The color scale is related to measured CP (using NF ISO 23015 (NF EN ISO 23015 1994)). Samples were numbered and labeled from 1 to 40. A gradient of CP may be globally observed since CP tends to increase with the T_{95} value. In contrast, the higher the conversion rate, the lower the CP is. These observations are in accordance with previous study which

pointed that diesel CP and T_{95} values are both strongly dependent on its heavy n-alkanes contents. Furthermore, hydrocracking enables to convert molecules combining cracking and isomerization reactions. Then, n-alkanes are mostly converted into branched alkanes when the conversion rate is high. The third variable used to model CP is related to Feedstock which may also significantly affect the diesel quality.

Fig. 4 : Projection of training samples in conversion rate-T95 space

2.5 Assessment of models quality

For each simulated function, models were developed using simulated data (paragraph 2.4.1) as training set.

Models performances were then evaluated using the following process:

1. the study area was discretized as a 100x100 grid points regularly spaced to use them as validation set
2. both theoretical and predicted values were calculated at each grid point.
3. the root mean square error (*RMSE*) of prediction and the mean relative deviation (*MRD*) were calculated on validation set as detailed in Table 5.

For real data, models were compared in two ways:

1. a leave-one-out (LOO) cross-validation was first applied to training data and a *RMSE* of cross validation and the percentage of predicted points with a precision less than confidence interval (CI) of the standard measurement method were then deduced (see formulae in Table 5) (Wahba and Wendelbeger 1980)
2. Same statistics were evaluated on prediction set

Table 5 : Definition of statistics for models performance

Type of data	Statistics	Formula
Simulated data	Mean Relative Deviation	$MRD = \frac{1}{n} \sum_{i=1}^n \frac{ y_i^{predicted} - y_i^{real} }{y_i^{real}}$
	Root Mean Square Error of Prediction	$RMSEP = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{predicted} - y_i^{real})^2}$
Real data	Mean Absolute Deviation	$MAD = \frac{1}{n} \sum_{i=1}^n y_i^{predicted} - y_i^{measured} $
	Root Mean Square Error of Prediction	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{predicted} - y_i^{measured})^2}$
	Percentage of predicted points with a precision under $\pm\gamma$ (only for real data)	$\tau_{\pm\gamma} = \frac{\sum_{i=1}^n 1_{ y_i^{predicted} - y_i^{measured} \leq \gamma}}{n}$

2.6 Used software

All models have been implemented using statistical software R (version 3.2.2, Copyright (C) 2015 The R Foundation for Statistical Computing). R is a free software and comes with absolutely no warranty. It is a collaborative project with many contributors. It compiles and runs a wide variety of UNIX platforms, Windows and MacOS. Various performing packages are available and allow to implement models from personal data. The package “stats” contains function that allow fitting linear regression models and statistical calculations. The package “DiceKriging” was used to implement kriging models (Olivier Roustant 2015). The package “fields” contains a function called “Tps” that enables to develop smoothing spline models.

3 Results and Discussion

Spline and kriging models were developed for each simulated database introduced in paragraph 2.4. Their performances were evaluated over the study area and were compared with those of classical regression model. All models performances are discussed in the following paragraphs.

3.1 Models comparison on simulated database

In each case of simulated data, classical regression, kriging and spline models were developed. The intrinsic parameters estimates related to spline and kriging model (introduced in paragraph **Error! Reference source not found.**) are specified in Table 6. The models performances are discussed in three ways:

- First by overlapping the response surface that the model provided with the original graph of the function to fit.
- Secondly by plotting a 2D-graphics of the prediction errors distribution within the study area.
- Thirdly by comparing their statistics given in Table 7.

Table 6 : Intrinsic parameters estimates in learning step for interpolation methods

Interpolation method	Intrinsic parameters	Data set 1	Data set 2	Data set 3
Splines	Order (m)	2	2	3
	Smoothing (λ)	0.17	4e-4	3e-6
Kriging	Nugget	1.12	0.70	2e-3
	θ_1	11.50	1.24	0.45
	θ_2	10.43	1.21	0.50

3.1.1 Data set 1 (affine function)

Statistics for models performance in simulated situations are given in Table 7. For data set 1, MLR obviously provides very good performances as all the assumptions of gaussian model are verified. Results are illustrated in Fig. 5. Consider first the results for MLR model. Fig. 5a illustrates the overlapping of the MLR response surface (plane colored in blue) to the original graph (plane colored in black); green points represent the training data. The corresponding prediction errors are plotted in Fig. 5b; the study area is delimited by the two-dimensional space and the absolute deviation related to each point is represented by a color scale; white circles localize training points. It may be observed in Fig. 5a that MLR response surface and original graph are obviously very close. Note also that blue regions which refers to low prediction errors are predominant in Fig. 5b. Both above observations show the good quality of the fit, in accordance with the good properties of the LS estimator in case of Gaussian linear model. However, MLR response surface and original graph are not entirely

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 equivalent. Indeed, the visual contrast between blue and black color that appears in Fig. 5a combined with the
2 symmetric distribution of prediction errors observed in Fig. 5b suggest a slight tilt of MLR response surface
3 relative to the original plane. Note also that the symmetrical aspect of the prediction errors distribution does not
4 depend on the training set configuration.

5 Results for kriging and splines models are illustrated in Fig. 5c and Fig. 5d respectively where the
6 corresponding prediction errors are plotted. The proximity between kriging predictions and real values is also
7 effective. However, some differences may be noted relative to MLR model: Firstly, it may be observed in Fig. 5c
8 that the points where prediction error is close to 0 describe a nonlinear curve which is the intersection line
9 between kriged surface and original map. Then, kriged surface is curved and the quality of the fit is less effective
10 for kriging than for MLR. Indeed, red regions that refers to absolute deviation around 1.5 may be observed in
11 Fig. 5c whereas the maximum value of absolute deviation for MLR model is less than 1 (see Fig. 5b).

12 The curvy characteristic of the response surface may be extended to the thin plate spline (Fig. 5d).
13 However, the prediction error distribution that spline model provides is closer to MLR than kriging. Globally, all
14 obtained models provided good performances since blue regions are predominant in absolute deviation plots
15 (Fig. 5b, c and d). Models statistics given in Table 7 are in accordance with these observations above. Indeed, a
16 MRD value of 0.07 was obtained for MLR against 0.10 and 0.13 for spline and kriging respectively.

Fig. 5 : a)- Overlapping of MLR response surface (blue) and original map (black) for affine function (data set 1) – green points -> training samples ; b)-Absolute deviation for all points within the study area – white circles (training samples locations); c) Absolute deviation for all points within the study area – white circles (training samples locations); d) Absolute deviation for all points within the study area – white circles (training samples locations)

3.1.2 Data set 2 (rational function)

For simulated data set 2, the function to fit is a rational fraction. Therefore, MLR is obviously low performing here since it is clearly impossible to correctly fit the sampled data by a plane surface. Moreover, find the adapted analytical expression to compute classical regression (linear or nonlinear) appears unlikely. Thus, only kriging and spline response surface are discussed below. Prediction errors distribution is represented in Fig. 6a for kriging model. The model appears to provide good accuracy in the most part of the study area. The dark blue regions that refer to the lowest prediction errors are localized around the training samples locations (white circles). A low quality of fit is particularly observed in a well-localized region that correspond to both low

abundance of training samples and a high variability of the function to fit. These observations may be extended to spline model (Fig. 6b). This relative equivalence between both models is supported by statistics since RMSE value of 1.69 was obtained for spline model against 1.88 for kriging (see Table 7). Note also that the high values of MRD (3.83 and 3.60) related to both models reflect the very low quality of the fit in low samples abundance and high function variability regions within the study area.

Fig. 6 a)- Overlapping of kriged response surface (blue) and original map (black) for rational function (data set 2) – green points -> training samples; b)-Absolute deviation for all points within the study area – white circles (training samples locations)

3.1.3 Data set 3 (sum of potential)

For this third example, the same process was used to compare models. As in previous case, we obviously considered that MLR cannot be appropriate to the simulated data observed in Fig. 2. Then, only kriging and spline models performances were illustrated in Fig. 7a and Fig. 7b respectively. As in the previous example, kriging and spline predicted values are close to the real value in the region localized around clusters and in the intermediate region. In contrast, significant difference of surface behavior may be observed in regions of low samples abundance. Thus, kriging model seems to overestimate the response values in these regions whereas spline model underestimates them. Furthermore, thin plate spline significantly diverges from the original graph (white zone in Fig. 7b) whereas for kriging the estimate values appears to aim at an expected value. Then, kriging provides more accurate prediction than spline model for these points. Models statistics

given in Table 7 are in accordance with this last point since MRD value obtained from kriging (0.10) significantly better than those of spline model (0.22).

Fig. 7 : a)- Overlapping of spline response surface (blue) and original map (black) for sum of potential (data set 3) – green points -> training samples; b)-Absolute deviation for all points within the study area – white circles (training samples locations)

Table 7 : Statistics for Models performances in simulated situations

Data	Model	MRD	RMSEP
Data set 1	Regression	0.13	0.24
	Kriging	0.07	0.40
	Spline	0.10	0.23
Data set 2	Regression	7.80	2.52
	Kriging	3.83	1.88
	Spline	3.60	1.69
Data set 3	Regression	0.40	0.32
	Kriging	0.10	0.18
	Spline	0.22	0.29

3.1.4 Uncertainties of prediction

Uncertainties of predicted values were computed by estimating their 95% level confidence interval amplitudes.

Results are illustrated for datasets 2 and 3 in Fig. 8 for stochastic approach. Classical intervals have obviously

ellipsoidal form due to their analytical expression. The ellipsoid size increases regularly with the distance from global barycenter of training data.

Interval amplitude that stochastic approach provided are illustrated in Fig. 8 for dataset 2 and 3. Amplitudes level are given through a color scale (blue regions correspond to lowest interval amplitude and red regions indicate highest amplitudes). It may be observed in Fig. 8a and Fig. 8b that blue regions are well-localized around training samples. Bluest regions are those that have a high abundance of sampled points whereas red regions have clearly less abundance of samples.

Fig. 8 : Amplitude of 95% level confidence interval within the study area; a) for dataset 2; b) for dataset 3

3.1.5 Statement

MLR, kriging and spline models was tested in three different case of simulated data. Two distinct situations may be noted: the first one concerns rigorous interpolation case where predicted point is surrounded by training samples (or not too far from them). In this situation kriging and models obviously provide good accuracy. The other one consists on predicting a point relatively far from the training set, that rigorously corresponds to an extrapolation case. Then, various behaviors of interpolation methods are observed according to those of the function to fit. When the response varies slightly the models accuracy remain consistent. When there are significant variation of the function, kriging and spline surface have distinct behaviors. This divergence is mainly due to the choice of covariance structure for kriging model. Indeed, Dubrule (Dubrule 1984) pointed that kriging and spline are strictly equivalent when a polynomial covariance function with a degree equal to the spline order is used.

However, previous examples showed that this polynomial structure is not always appropriate, especially in case of potential function. An advantage in using Gaussian type is that the predicted value always aims at an expected value.

Globally, the results showed the adaptability of both kriging and spline models to linear or nonlinear situations whereas classical regression require to find the suitable model structure according to the observed data. Except for classical functions (linear, affine, second degree polynomial, *etc.*), this may be very complex.

Kriging and spline provided very close performances in the tested situations. In addition, the stochastic approach of kriging enables to define local uncertainties that are related to the training points configuration as shown in the previous paragraph. When an unsampled location is surrounded by training points (interpolation) the prediction uncertainty is low. When it is relatively far from sampled data (extrapolation) a high uncertainty value is observed.

3.2 Application to the modeling of cloud point in diesel fuel.

MLR and kriging models were developed using training samples. The 95% level confidence interval related to each predictive method were also computed for all predicted values. Resulting confidence interval amplitudes are illustrated by error bars (colored in blue). Models performances were evaluated on both training (using leave-one-out method) and prediction set. Parity graph are plotted in

Fig. 9 for kriging model. Statistics of Models performances are available in **Error! Reference source not found.** On training set, MAD and RMSE values are clearly better for kriging (2.1°C and 2.9°C respectively) than for MLR (3.0°C and 4.0°C respectively). Moreover, the percentage of predicted values at a precision level

under the confidence interval of the standard measure is quite higher for kriging (70%) than for MLR (60%).

Fig. 9a shows that training samples which have the highest absolute prediction errors by leave-one -out (numbered 8, 10, 22 and 32) correspond to the largest confidence intervals. Furthermore, it may be observed in Fig. 4 that all these samples are located relatively far from the rest of training data. That points the most valuable characteristic of stochastic interpolation.

On prediction set, models performance are also quite better for kriging model than for MLR. Actually, kriging model provides a RMSE value of 2.0°C and predicts 94% of test samples at a precision level under the confidence interval of the standard measure, whereas the corresponding values are respectively 2.5°C and 75% for MLR model.

Table 8 : Statistics for Models performances in prediction of diesel cloud point

Type of Validation	Model	MAD (°C)	RMSEP (°C)	% points predicted at ± CI	% points predicted at ± 2CI
Leave-one-out	MLR	3.0	4.0	60	82
	Kriging	2.1	2.9	70	90
Prediction	MLR	2.0	2.5	75	94
	Kriging	1.8	2.0	94	94

Fig. 9: Parity plots of kriging model for prediction of diesel cloud point; a) on training set; b) on prediction set

4 Conclusion

In this work, kriging and splines models were first developed and compared to MLR for modeling simulated functions. Results pointed that interpolation methods have a high capacity of adaptability to linear or nonlinear situations without requiring previous analysis. Actually, intrinsic parameters of splines may be automatically selected and some basic covariance such as Gaussian structure related to kriging are always intuitively workable in any situation. Although, kriging and splines provide very close performance in some situations, kriging is more valuable since it provides a measure of uncertainties related to the predicted value based on stochastic assumptions. That enables to discuss on how far a given unsampled location is from the training samples and then rigorously specify what kind of (interpolation or extrapolation) situation has to be considered. This point is very important in modeling physicochemical properties of petroleum products. It is even particularly essential when the number of descriptors increases significantly (higher than 3).

Kriging and MLR models were developed and compared in modeling of diesel cloud points. Although MLR provides good performances that are in accordance with the uncertainties of the standard measure, kriging enables to improve accuracy. In our knowledge, this is the first model for the prediction of diesel CP using kriging.

Globally, interpolation methods are well adapted to the modeling of petroleum products properties since they provide good performances when the number of samples is limited. Use of kriging methods to high

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 dimensional study remains always challenging. The real difficulty consists on recognizing the adapted
2 covariance structure. On this point, Gaussian structure appears as a good default alternative.
3

Appendix A : Estimation of regression parameters

In statistics, an estimator is a rule to calculate an estimate of a given quantity based on observed data (Azaïs and Bardet 2012). Let $\hat{\theta}$ denotes an estimator of θ . It is generally obtained by solving an optimization problem such as:

$$\hat{\theta} = \arg \min_{\theta} J(\theta) \quad (\text{Eq. A 1})$$

where J is an objective function that depends on the available observed data. There are various types of estimator according to the analytic expression of the objective function. The most currently used is the least squares (LS) estimator which is defined as follows:

$$\hat{\theta}^{LS} = \arg \min_{\theta} \{ J(\theta) = \sum_{i=1}^n (y_i - f(x_i, \theta))^2 \} \quad (\text{Eq. A 2})$$

The quality of an estimator is evaluated by two statistical characteristics:

- Its bias

$$B[\hat{\theta}] = E(\hat{\theta}) - \theta \quad (\text{Eq. A 3})$$

- Its variance

$$V[\hat{\theta} - \theta] = E[(\hat{\theta} - \theta)^2] - B(\hat{\theta})^2 \quad (\text{Eq. A 4})$$

where $E[X]$ and $V[X]$ denotes respectively the expected value and the variance of the random variable X . An estimator is considered as an optimal one whether it is an unbiased with a minimal variance (Azaïs and Bardet 2012). Under the assumptions that modelling errors are normally independently identically distributed (i. i. d.), it has been shown that the LS-estimator presents optimal characteristics (Antoniadis, Berruyer and Carmona 1992; Azaïs and Bardet 2012).

Regression models may be divided into two groups: the generalized linear models if f can be expressed as a linear combination of its parameters (affine, polynomial, quadratic, etc.); the nonlinear models (rational fraction, exponential, logarithmic, etc.)

Appendix B : multivariate normal distribution

A gaussian vector is a numeric vector such that any linear combination of his components is a normally distributed random variable. A multivariate normal distribution is entirely characterized by its mean vector m and its variance-covariance matrix Σ . Let $(X_1, \dots, X_d)$ denotes a Gaussian vector. The probability density is defined as follows:

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

$$f(x) = \frac{1}{\sqrt{2\pi}^n |\det \Sigma|} e^{-(x-m)' \Sigma^{-1} (x-m)} \text{ pour tout } x \in \mathbb{R}^d \quad (\text{Eq. A 5})$$

Appendix C : General trend for universal kriging

The f^l that appear in the definition of universal kriging predictor (paragraph **Error! Reference source not found.**) are polynomials of degree lower or equal to m . If x is a d – dimensional vector, they are denies as follows (Journel and Rossi 1989):

$$f^0(x) = 1, f^1(x) = x_1, \dots, f^p(x) = x_p, f^{p+1}(x) = x_1^2, \dots, f^{2p}(x) = x_1 x_d, \dots, f^m(x) = x_d^m \quad (\text{Eq. A 6})$$

Appendix D : Maximum likelihood estimate

In statistics, a likelihood function is a function of a statistical model parameters given data. Likelihood functions play a key role in statistics inference. The basic idea of likelihood function consists on considering models parameters as a set of random variable. It is then defined as the probability density function of parameters given the observed data. The maximum likelihood estimate is the set of values that maximizes the likelihood function (Myung 2003).

Appendix E : Nugget effect

In the early development of geostatistics, the term ‘nugget effect’ was coined for the apparent discontinuity at the beginning of many semi-variogram graphs. This name was chosen to reflect the large differences found between neighboring samples in ‘nuggety’ mineralization such as Wits gold reefs (Goovaerts 1997). It may be defined as follows:

$$\lim_{h \rightarrow 0} C(h) = C_{nugget} + \sigma_\epsilon^2 \quad (\text{Eq. A 7})$$

where C is the recognized covariance model by plotting semi-variogram.

5 References

- Abdel-Rahman, Elfatih M., Onesimo Mutanga, John Odindi, Elhadi Adam, Alfred Odindo, and Riyad Ismail. 2014. "A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data." *Computers and Electronics in Agriculture* 106(0):11–19.
- Ahlberg, J. H., E. N. Nilson, and J. L. Walsh. 1967. *The Theory of Splines and Their Applications*. New York: Academic Press.
- Antoniadis, Anestis, Jacques Berruyer, and R. Carmona. 1992. *Régression non linéaire et applications*. Paris: Economica.
- Arnaud, Michel, and Xavier Emery. 2000. *Estimation et interpolation spatiale: Méthodes déterministes et méthodes géostatistiques*. Paris: Hermes Science.
- Ascher, U., S. Pruess, and R. D. Russel. 1983. "On spline basis selection for solving differential equations." *SIAM Journal on numerical analysis* 20(1):121–42.
- ASTM D2887. 2016. *Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography*: ASTM International.
- Azaïs, Jean-Marc, and Jean-Marc Bardet. 2012. *Le modèle linéaire par l'exemple: Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS*. 2nd ed. Paris: Dunod.
- Baar, Jouke H. S. de, Mustafa Percin, Richard P. Dwight, Bas W. van Oudheusden, and Hester Bijl. 2014. "Kriging regression of PIV data using a local error estimate." *Experiments in Fluids* 55(1):159.
- Becker, Per J., Benoît Celse, Denis Guillaume, Hugues Dulot, and Victor Costa. 2015. "Hydrotreatment modeling for a variety of VGO feedstocks: A continuous lumping approach." *Fuel* 139:133–43.
- Becker, Per J., Benoît Celse, Denis Guillaume, Victor Costa, Luc Bertier, Emmanuelle Guillon, and Gerhard Pirngruber. 2016. "A continuous lumping model for hydrocracking on a zeolite catalysts: Model development and parameter identification." *Fuel* 164:73–82.
- Bolasso: Model Consistent Lasso Estimation Through the Bootstrap*.
- Braga, J. W., A. A. Santos, and I. S. Martins. 2014. "Determination of viscosity index in lubricant oils by infrared spectroscopy and PLSR." *Fuel* 120:171–78.
- Burrough, P. A., and Rachael McDonnell. 1998. *Principles of geographical information systems*. Oxford, New York: Oxford University Press.
- Celse, Benoît, Jean-Jérôme Da Costa, and Victor Costa. 2016. "Experimental Design in Nonlinear Case Applied to Hydrocracking Model: How Many Points Do We Need and Which Ones?" *International Journal of Chemical Kinetics* 48(11):660–70.
- Cookson, David J., Jozef L. Latten, Ian M. Shaw, and Brian E. Smith. 1985. "Property-composition relationships for diesel and kerosene fuels." *Fuel* 64(4):509–19.
- Cookson, David J., and Brian E. Smith. 1990. "Calculation of Jet and Diesel Fuel Properties Using ¹³C NMR Spectroscopy." *Energy and Fuels* 4(6):152–56.
- Cookson, David J., and Brian E. Smith. 1992. "Observed and Predicted Properties of Jet and Diesel Fuels Formulated from Coal Liquefaction and Fischer-Tropsch Feedstocks." *Energy and Fuels* 6:581–85.
- Cookson, David J., Peter Iliopoulos, and Brian E. Smith. 1995. "Composition-property relations for jet and diesel fuels of variable boiling range." *Fuel* 74(1):70–78.

- Cressie, Noel A. C. 1990. "The origins of kriging." *Mathematical geology* 22(3):239–52.
- Cressie, Noel A. C. 1993. *Statistics for spatial data*. New York, Chichester: Wiley.
- Dror, Hovav A., and David M. Steinberg. 2008. "Sequential Experimental Designs for Generalized Linear Models." *Journal of the American Statistical Association* 103(481):288–98.
- Dubrule, Olivier. 1984. "Comparing splines and kriging." *Computers and Geosciences* 10(2-3):327–38.
- Duchon, J. 1975. "Spline associated to n observations of a random function." *Comptes Rendus Hebdomadaires Des Seances De L'academie Des Sciences Série A* 280(14):949–51.
- Goovaerts, Pierre. 1997. *Geostatistics for natural Resources evaluation*. New York: Oxford University Press.
- Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2001. *The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman*. New York: Springer.
- Isaaks, Edward H., and R. M. Srivastava. 1989. *Applied geostatistics*. New York: Oxford University Press.
- Jones, Donald R., Matthias Schonlau, and William J. Welch. 1998. "Efficient global optimization of expensive black-box functions." *JOURNAL OF GLOBAL OPTIMIZATION* 13(4):455–92.
- Journel, A. G., and Ch J. Huijbregts. 1978. *Mining geostatistics*. London, New York: Academic Press.
- Journel, A. G., and M. E. Rossi. 1989. "When do you need a trend model in Kriging ?" *Mathematical geology* 21(7):715–39.
- Kapur, G. 2004. "Establishing structure–property correlations and classification of base oils using statistical techniques and artificial neural networks." *Analytica chimica acta* 506(1):57–69.
- Kleijnen, Jack P. C. *Kriging metamodeling in simulation: A review*. Vol. 2007,13. Tilburg: Center for Economic Research.
- Marinovic, Slavica, Tomislav Bolanca, and Sime Ukic. 2012. "Prediction of Diesel Fuel Cold Properties Using Artificial." 48(1):47–51.
- Matheron, G. 1971. *The theory of regionalized variables and its applications*. Vol. 5. Paris: École Nationale Supérieure des Mines.
- Matheron, G. 1975. *Splines and Kriging: Their Formal Equivalence*: Fontainebleau.
- Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of machine learning*. Cambridge, MA: The MIT Press.
- Myung, In J. 2003. "Tutorial on maximum likelihood estimation." *Journal of Mathematical Psychology* 47(1):90–100.
- NF EN ISO 23015. 1994. *Produits Pétroliers - Détermination du Point de Trouble*.
- Olivier Roustant, David G. Y. D. 2015. "Package DicKriging: Kriging Methods for Computer Experiments." (<http://dice.emse.fr/>).
- Produits pétroliers lourds, solides ou pâteux et Catalyseurs dosage azote TOTAL par la méthode de Dumas*.
- Riazi, M. R. 2005. *Characterization and properties of petroleum fractions*. 1st ed. West Conshohocken, PA: ASTM International.
- Santiago, J., M. Claeys-Bruno, and M. Sergent. 2012. "Construction of space-filling designs using WSP algorithm for high dimensional spaces." *Chemometrics and Intelligent Laboratory Systems* 113:26–31.
- Santner, Thomas J., Brian J. Williams, and William I. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York, NY: Springer New York.

- Santos, Vianney O., Flavia C. C. Oliveira, Daniella G. Lima, Andrea C. Petry, Edgardo Garcia, Paulo a. Z. Suarez, and Joel C. Rubim. 2005. "A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis." *Analytica chimica acta* 547(2):188–96.
- Sastry, M. I., Anju Chopra, A. S. Sarpal, S. K. Jain, S. P. Srivastava, and A. K. Bhatnagar. 1998. "Determination of Physicochemical Properties and Carbon-Type Analysis of Base Oils Using Mid-IR Spectroscopy and Partial Least-Squares Regression Analysis." *Energy and Fuels* 12(5):304–11.
- Schoenberg, I. J. *Cardinal Spline Interpolation*. Vol. 12. Philadelphia: Soc. for Industrial and Applied Mathematics.
- Schumaker, Larry L. 1981. *Spline functions, basic theory*.
- Simpson, Timothy, Timothy M Mauery, John J Korte, and Farrokh Mistree. 1998. "Comparison Of Response Surface And Kriging Models For Multidisciplinary Design Optimization." *American Institute of Aeronautics and Astronautics* 98.
- Srivastava, R. M. 1987. "Maximum variance of profitability." *Cim Bulletin* 80(901):63–68.
- Stein, Michael. 1987. "Gaussian approximations to conditional distributions for multi-Gaussian processes." *Mathematical geology* 19(5):387–405.
- Wackernagel, Hans. 2003. *Multivariate geostatistics: An introduction with applications*. 3rd ed. New York: Springer.
- Wahba, G., and J. Wendelbeger. 1980. "Some New mathematical -Methods for Variational Objective Analysis Using Splines and Cross Validation." *Monthly Weather Review* 108(8):1122–43.
- Wang, Dezhi. "Kriging regression in digital image correlation for error reduction and uncertainty quantification.", University of Liverpool, degree granting institution.
- Wang, Shouchun, Xiucheng Dong, and Renjin Sun. 2010. "Predicting saturates of sour vacuum gas oil using artificial neural networks and genetic algorithms." *Expert Systems with Applications* 37(7):4768–71.