



A practical method of global sensitivity analysis under constraints

Jean-Pierre GAUCHI

Technical Report 2015-1, february 2015

UR1404 MaIAGE

**Mathématiques et Informatique Appliquées
du Génome à l'Environnement**

INRA

Domaine de Vilvert

78352 Jouy-en-Josas Cedex

France

<http://maiage.jouy.inra.fr>

© 2015 INRA

Jean-Pierre GAUCHI
Jean-Pierre.Gauchi@jouy.inra.fr

Abstract

A novel practical method to conduct a Global Sensitivity Analysis (GSA) for computer models is proposed in this article. It is based on Partial Least Squares polynomial regression metamodeling, used in conjunction with a D-optimal computer experiment design. This practical method is particularly well adapted to the following four situations: (i) when the cost of a single computer simulation run is very high (long computing time), limiting computation to just hundreds of simulation runs, whereas the usual GSA methods (typically, those based on Latin Hypercube Samplings or Sobol sequences) require thousands or tens of thousands of simulation runs; (ii) when the inputs are not independent because some of them are stochastically linked (correlated) or deterministically (functionally) linked; (iii) when some inputs are of a qualitative (categorical) nature; and (iv) when the outputs are multivariate. This practical method, proposed in a deterministic simulation framework, is useful for moderately nonlinear computer models. It makes it possible to define and compute novel sensitivity indices, referred to here under the general term of SIVIP, not to be confused with Sobol indices defined for independent inputs. We have called our new method the SIVIP method. It is applied here to the aeronautics field, in particular, for the estimation of detection performances of infrared sensors and, more precisely, for the computation of aircraft InfraRed Signatures (IRS).

Keywords: sensitivity indices; PLS regression; D-optimal design; aeronautics; InfraRed Signature.

1 Introduction

Global Sensitivity Analysis (GSA) is defined as the study of how the output uncertainty of a computer model can be allocated to different sources of uncertainty in the model inputs [1]. GSA leads to a quantification and a ranking of the input influences. A major final objective is the simplification of the model, enabling us to obtain a parsimonious model in the end. GSA based on variance methods has undergone considerable development since the 1990s, with many applications in several scientific areas such as applied physics and, more recently, in the biological sciences [2]. Several books provide a good overview of the techniques used in GSA [3,4]. Sobol' Sensitivity Indices (*SSI*) [5] are intended to represent the sensitivities for general (nonlinear) models. Almost all of the published methods are based on Monte Carlo simulations [6,7] and require thousands or even tens of thousands of simulation runs for computing accurate estimations of the *SSI*. This is typically the case of the methods based on space-filling designs [8]. Moreover, these classical methods only work correctly if the random inputs of the computer model are continuous and independent because the *SSI* are rigorously defined only in this situation [5]. Thus, the need for a large number of simulation runs and the necessity for the inputs to be continuous and independent constitute three major constraints for *SSI* estimation.

To address the first of these constraints, i.e., the large number of simulations to be carried out with Monte Carlo methods, some authors [9,10] have proposed a completely different approach for computing the *SSI*, mainly based on sparse polynomial chaos expansions. This latter approach leads to a large decrease in the number of simulation runs to be performed. This major advantage makes it possible to obtain the *SSI*, even for huge computer models where a single simulation run requires several minutes or hours and where it is obvious that none of the classical methods based on Monte Carlo simulations can be used. However, even the methods [9,10] cannot take account of the correlations among some (or all) of the quantitative inputs, or of linear functional relationships between them (typically, bounded linear combinations of inputs), and the presence of qualitative (categorical) inputs. In [11-14], the authors focus on correlated inputs, but these inputs must be quantitative and cannot be either strictly functionally linked or qualitative.

In this article, we consider dependent inputs: either correlated quantitative inputs or inputs linked by some linear combinations of the general form $\alpha_{\min} \leq \alpha_1 X_1 + \dots + \alpha_K X_K \leq \alpha_{\max}$, where $K \leq p$, p the total input number, and α_{\min} , α_1 , \dots , α_K , α_{\max} are real numbers. Qualitative inputs can also be present in the form of their (0/1)-indicator variables. Hence, if we are in a situation where we have to simultaneously deal with the following three specific constraints - (1) Constraint A: a number of simulation runs that must be small or moderate (relative to the

number in Monte Carlo simulations); (2) Constraint B: quantitative inputs that are correlated or deterministically linked by relationships of the previously mentioned general form; and (3) Constraint C: presence of qualitative inputs - we then propose a novel practical approach in this paper for conducting a GSA. Moreover, this approach also enables us to deal with multivariate outputs. Our approach is jointly based on a Partial Least Squares Regression (PLSR) for a polynomial meta-modeling and on a D-Optimal Computer Experiment Design (DOCED) strategy. Consequently, the new sensitivity indices we propose here, referred to under the general term of *SIVIP* (not to be confused with the *SSI* that are only correctly defined for independent inputs), are based on a metamodeling approach. Concerning the DOCED, it should be stressed that we are in a deterministic simulation framework, i.e., the inputs are controlled (as in standard factorial designs) and take the values determined by the DOCED. Note that some types of *SIVIP* proposed in this paper were briefly presented and partially defined in [2,15-17], but were not statistically and mathematically justified.

Other computer experiment modeling strategies obviously already exist, typically, those based on Gaussian Process (GP) models, beginning with the work of Sacks et al. [18] (for detailed discussions of such models, see [19,8]), but they only perform for continuous inputs and univariate outputs. A relevant article [20], however, does exist, which also proposes taking the qualitative inputs into account with the construction of correlation functions. Lastly, the approach of [21,22] for multivariate outputs should also be mentioned, but no optimal computer experiment strategy is given.

Our approach is simpler than [20]. It can be seen as a practical alternative method, very different from [20-22], particularly well adapted if the computer model nonlinearity is moderate, and useful with multivariate outputs as well.

The rest of the article is organized as follows. Section 2 presents the novel approach. Section 3 gives statistical justifications of this approach. Section 4 illustrates this novel approach with a case study in aeronautics, in particular, for the estimation of detection performances of infrared sensors and, more precisely, for the computation of aircraft InfraRed Signatures (IRS), revealing the effectiveness of this methodology. Section 5 is the conclusion. The Appendix provides some background on PLSR.

2 The novel approach proposed

2.1 Postulation of a polynomial metamodel and its estimation

Let \mathcal{M} be the computer model of the "black box" type. Assume that for $\mathbf{x}_i = (X_{i1}, \dots, X_{ip})$, a point in the p -dimensional space of the p X_j inputs, $j = 1, \dots, p$, the simulation run result of the univariate Y output is $y_i = \mathcal{M}(\mathbf{x}_i)$. Note that p can be large in practical applications (frequently as high as 30).

We assume that the Y output can be approximated by a (full or not) polynomial metamodel (surrogate model) of degree d , referred to as \mathcal{M}_d , built from the p inputs (the categorical inputs are coded with their (0/1)-indicator variables). For one simulation run, we can write $y_i = \mathcal{M}(\mathbf{x}_i) = \mathcal{M}_d(\mathbf{x}_i) + \eta_i$ and $\mathcal{M}_d(\mathbf{x}_i) = \hat{\mathcal{M}}(\mathbf{x}_i) + \varepsilon_i$ where η_i is a deterministic model error because \mathcal{M}_d is a deterministic polynomial approximation of \mathcal{M} , and all of the \mathbf{x}_i form a deterministic design. Moreover, ε_i is a residual due to the estimation method chosen.

With several y_i obtained from \mathcal{M} at several \mathbf{x}_i , the estimation of the β vector of the \mathcal{M}_d coefficients is computed by PLSR (see Appendix for details on this regression method), leading to the estimated metamodel, $\hat{\mathcal{M}}_{d,PLS}$, that will be simply referred to as $\hat{\mathcal{M}}$ in the text. Therefore, we obtain $\hat{\mathcal{M}} = \mathbf{X}\hat{\beta}_{PLS}$ as the estimation of \mathcal{M} , where \mathbf{X} is the design model matrix of \mathcal{M}_d . The value of d and the monomials present in \mathcal{M}_d are chosen beforehand by the user. For practical applications of moderate nonlinearity, a value of 2 or 3 for d is generally used.

2.2 Building strategy of an efficient DOCED

Our basic idea for determining an optimal computer experiment design is inspired by the theory of optimal design of real (non-computer) experiments for linear models (see the book by Atkinson and Donev [23] for a very clear introduction to this theory).

We now explain how to determine a small-size computer experiment design formed by computer experiments selected from a large candidate computer experiment set that we will refer to as the X_C set. This selection is made according to the D-optimality criterion [24]. A DOCED is associated with a model, \mathcal{M}_d in this case. The following two steps are therefore required to determine a DOCED*. This DOCED* will be considered as a deterministic design.

Step 1: Construction of the X_C candidate set

This construction depends on the (quantitative or qualitative) nature of the p inputs, as well as on the constraint types (correlations or linear combinations).

Let p_1 be the quantitative input number and p_2 be the qualitative input number, with $p = p_1 + p_2$.

If we want to correlate the p_1 inputs, a convenient and simple method is to start from a space-filling design, SFD, based on a large N_0 row number (typically $N_0 = 1000 \times p$), e.g., a Latin Hypercube Sampling (LHS) [25], and then use the Iman and Conover method [26] to (approximately) obtain the desired correlations among the concerned inputs. Another possible method is to use copulas [27-28] for correlating the p_1 inputs, which is the case for the application detailed in Section 4. In another situation, if the p_1 inputs are linked by a linear combination (see Introduction section), then only the rows among the N_0 original rows that satisfy this combination are selected to form a new (smaller) constrained SFD. Hence, the number of rows of this latter SFD can be considerably decreased. The resulting SFD remains acceptable if its row number is greater than or equal to $100 \times p$. Note that this problem of functional (or structural) constraints is encountered in the constrained optimization field as well. The reader may refer to the relevant approach [29] to deal with this situation.

Let N_C be the row number of this constrained SFD. Moreover, if p_2 qualitative inputs are present, then their levels are repeated and randomized until N_C values are obtained for each of the p_2 inputs. Finally, we obtain a constrained SFD of N_C rows and p columns, which will be referred to as SFD* in the text.

Step 2: Construction of a sequence of several DOCED and choice of a DOCED*

Firstly, using SFD*, a design model matrix is built depending on d and the chosen monomials. The modalities of the p_2 inputs are replaced by their (0/1)-indicator variables, and the construction of eventual interactions are then performed with some quantitative inputs. Let \mathbf{X} be this design model matrix of N_C and P columns, where P corresponds to the number of chosen monomials in \mathcal{M}_d , i.e., main effects of the quantitative inputs, indicator variables of the qualitative inputs, interaction terms, quadratic terms, and cubic terms, if necessary.

The DOCED sequence to be constructed is referred to as $(\boldsymbol{\xi}_{n \times P}^D, n = P(= n_0), P + 1, \dots, N_C - 1)$, where each DOCED, $\boldsymbol{\xi}_{n \times P}^D$, is a $(n \times P)$ matrix. Every $\boldsymbol{\xi}_{n \times P}^D$ is optimal in the sense of the D-optimality criterion [24], i.e., among all n size designs drawn from SFD* – forming the so-called Ξ_n set – it is the one with the largest normalized determinant $\Delta_n = [\det(\mathbf{X}_n^T \mathbf{X}_n)] / n^P$ where $\mathbf{X}_n^T \mathbf{X}_n$ is the so-called information matrix. It was generally found that n is very small compared to N_C , which is the major advantage of the method, therefore satisfying constraint A. Several specific (discrete) algorithms exist for finding a $\boldsymbol{\xi}_n^D$. The well-known Fedorov [24] and Mitchell [30] exchange algorithms can be used, and they can be easily and efficiently put into practice by means of [31].

An important remark must be made here about finding a D-optimal design

with algorithms of [24]. These algorithms require the inversion of the information matrices. However, strong correlations between quantitative inputs and/or the presence of all of the indicator variables (resulting from the qualitative inputs) lead to the non-invertibility of these information matrices. Before using these algorithms, we propose, in this case, to replace the \mathbf{X} matrix by the orthogonal matrix formed by its principal components (obtained from a Principal Component Analysis), by retaining a number of components corresponding to at least 99% of the explained inertia. This prior transformation was achieved for the application in Section 4.

Finally, on the graph of the normalized determinants (or their logarithms) versus n , we can detect a good compromise between a high level of the D-optimality criterion and a reasonable value of n (see Figure 1 in the Application section). Let n^* be this optimal value of n . The DOCED^* , $\boldsymbol{\xi}_{n^* \times P}^D$, is therefore found.

2.3 The novel Sensitivity Indices

These new sensitivity indices, which we refer to under the general term of *SIVIP* in this article, are based on the *VIP* statistics obtained in PLSR outputs. The *VIP* statistics are defined by formula (14) given in the Appendix. We propose two types of *SIVIP*:

- the **Individual Sensitivity Index** defined for each k monomial term among the P monomial terms of \mathcal{M}_d , for univariate and multivariate output, by:

$$ISIVIP_k = VIP_k^2/P \quad (1)$$

If the output is multivariate, we refer to it as the **Generalized Individual Sensitivity Index** in this case, or $GISIVIP_k$.

- the **Total d -order Sensitivity Index** defined for each of the p original inputs, for univariate and multivariate output, by:

$$T_dSIVIP_j = \sum_{u=1}^J ISIVIP_{\Omega_{ju}} \quad (2)$$

where Ω_{ju} is the u th index set where a j index appears, J is the total number of these Ω_{ju} , and $ISIVIP_{\Omega_{ju}}$ is the corresponding monomial individual sensitivity index. For example, with three inputs, X_1, X_2, X_3 , and $d = 2$, the Total 2-order Sensitivity Index for X_1 is expressed as $T_2SIVIP(X_1) = ISIVIP(X_1) + ISIVIP(X_1^2) + ISIVIP(X_1X_2) + ISIVIP(X_1X_3)$. If the output is multivariate, we refer to it as the **Generalized Total d -order Sensitivity Index** in this case, or GT_dSIVIP_j .

We emphasize here that the $GISIVIP_k$ and the GT_dSIVIP_j are particularly useful from a practical point of view. Indeed, they can reveal several strong or weak inputs for all the outputs studied simultaneously, which is especially useful when the multivariate output is a spectrum, e.g., infra-red spectrum.

3 Justifications of the novel approach proposed

3.1 Justification of the PLSR choice

Firstly, we can observe that the PLSR metamodel deals with **partial covariances** for constructing its t_h components, $h = 1, \dots, H$. This aspect is clear if we examine the developed vectorial structure of a t_h component given in formula (9) of the Appendix. Indeed, the term $cov(\mathbf{F}_{h-1}, \mathbf{E}_{h-1,j})$ is a covariance between the \mathbf{F}_{h-1} residual matrix and the $\mathbf{E}_{h-1,j}$ residual vector and is therefore a partial covariance by definition. Formula (9) thus explains why the multicollinearity among the monomials can be correctly taken into account. Moreover, PLSR leads to the approximation $\hat{\mathbf{F}}_0 = \sum_{h=1}^H \mathbf{t}_h \mathbf{r}_h^T$, and we therefore obtain for the Y_l univariate output:

$$var(\hat{Y}_l) = var(Y_l) \sum_{h=1}^H r_{lh}^2 var(t_h), \quad l = 1, \dots, L \quad (3)$$

which is an empirical **orthogonal** decomposition of the total variance of \hat{Y}_l . Moreover, since $Y_l \approx \hat{Y}_l$, formula (3) gives an approximation of an empirical orthogonal decomposition of the Y_l output. We can then see with formula (9) that the decomposition (3) is linked to the partial covariances between the Y_l output and the monomial inputs.

3.2 Justification of the D-optimality choice

If we use Ordinary Least Squares Regression (OLSR) for estimating the $\beta_{P \times 1}$ coefficient vector of \mathcal{M}_d with n real data, the variance of the OLS estimator, $\hat{\beta}_{OLS}$, has the following well-known form:

$$var(\hat{\beta}_{OLS}) = \sigma^2 (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \quad (4)$$

where σ^2 is a constant, unknown, experimental variance. On the basis of (4), we can see that in order to minimize the variances and covariances of the $\hat{\beta}_{OLS}$ vector components, which we always want to do, a D-optimal design, $\xi_{n \times P}^D$, is a good choice because

$$\boldsymbol{\xi}_{n \times P}^D = \text{Arg} \left\{ \max_{\boldsymbol{\xi}_{n \times P} \in \Xi_n} [\det(\mathbf{X}_n^T \mathbf{X}_n)] \right\} = \text{Arg} \left\{ \min_{\boldsymbol{\xi}_{n \times P} \in \Xi_n} [\det(\mathbf{X}_n^T \mathbf{X}_n)^{-1}] \right\} \quad (5)$$

regardless of the value of the unknown σ^2 .

In the computer experiment framework, if the metamodel is fixed, then σ^2 is based on the ε_i residuals (see Section 2.1), and formula (4) can then makes sense as well.

Now, if we want to prove that it is also relevant to use D-optimal designs in the PLSR context— strictly defined in the framework of OLSR — then it might be proven that the determinant of an approximation (no exact formula exists) of the $\hat{\boldsymbol{\beta}}_{PLS}$ variance will be decreased if \mathbf{X}_n was based on a D-optimal design instead of on just any design. This is an open problem at that time. Only some numerical studies show an advantage of using D-optimal designs in the PLS context.

Lastly, it should be emphasized here that the choice of the D-optimality criterion is appropriate only if the postulated model is correct. It is for this reason that we assumed a moderate nonlinearity for the computer model and, therefore, a polynomial approximation with $d \leq 3$.

3.3 Justification of the VIP-based Sensitivity Indices choice

A first justification of the *VIP*-based Sensitivity Indices, instead of a choice based on $\hat{\boldsymbol{\beta}}_{PLS}$, for example, can be warranted by the fact that it has been observed for many years and in many studies that the *VIP* remains much more stable than $\hat{\boldsymbol{\beta}}_{PLS}$ when h increases. Moreover, the following two important arguments can be given in favor of this choice.

Deriving the *SSI*

It is possible to directly derive the *SSI* in the case of independent inputs in some particular situations. If the computer design is orthogonal, $Q_G^2 = 1$, and $L = 1$ (one univariate output Y_l is considered), then only one PLS component will be extracted. Its corresponding vectorial form will then be $\mathbf{t}_1 = \mathbf{E}_{01}w_{11} + \dots + \mathbf{E}_{0P}w_{1P}$, and formula (8) becomes $\mathbf{F}_{0l} = w_{11} + \dots + w_{1P}$ because the \mathbf{r} vector becomes the scalar $r = 1$ and $\text{var}(\mathbf{E}_{0j}) = 1$, $j = 1, \dots, P$. These w_{1j}^2 are equal to $\text{cor}^2(F_{0l}, E_{0j}) = \text{cor}^2(Y_l, Z_j)$. In this case, we can prove that w_{1j}^2 is the *SSI* for the input Z_j . For an input Z_j , we have:

$$\begin{aligned} ISIVIP_j &= VIP_j^2/P = P^{-1} \left(P (\text{cor}^2(Y_l, t_1))^{-1} \text{cor}^2(Y_l, t_1) w_{1j}^2 \right) = w_{1j}^2 \quad (6) \\ &= \text{cor}^2(Y_l, Z_j) = V(E(Y_l|Z_j))/V(Y_l) = SSI_j \end{aligned}$$

Summation of the $ISIVIP_j$

As a result of the property given in the Appendix, in the general case (univariate or multivariate outputs), for dependent or independent inputs, we have $\sum_{j=1}^P ISIVIP_j = 1$, which is a very convenient formula from a practical point of view if we wish to compare the $ISIVIP_j$ to each other. It is analog to the Sobol' theoretical summation [7] for independent inputs $\sum_{j=1}^Q SSI_j = 1$, where $Q = 2^p - 1$.

Illustration of the SIVIP method for the Ishigami function

In GSA, this function is a very well known test function with three independent inputs [9]. Even if the SIVIP method is specifically designed for dependent inputs, it can be relevant to apply it in this situation. Therefore, with a DOCED* of only 32 computer experiments (sampled in a 64-Legendre orthogonal array) and several (not full) \mathcal{M}_3 polynomial metamodels, we obtained accurate estimations of the three analytical total sensitivity indices (ATSI). For example, with a polynomial metamodel formed by the 11 monomials $X_1, X_3, X_1X_2, X_1X_3, X_2X_3, X_1^3, X_1^2X_2, X_1X_2^2, X_2^3, X_2X_3^2$ and X_3^3 , we obtained the following estimations (in %) $\widehat{ATSI}_1 = 44.91$ (44.83), $\widehat{ATSI}_2 = 35.62$ (35.58), and $\widehat{ATSI}_3 = 19.47$ (19.59); the analytical values are given in brackets.

However, we do not have any efficient algorithm for finding these correct monomials at this time. Only an exhaustive analysis enabled us to find them here. In any case, it is not the aim of the SIVIP method to compute sensitivity indices in the situation of independent inputs.

4 A case study in aeronautics

4.1 Presentation of the problem

This methodology has been applied to aircraft InfraRed Signature (IRS) analysis, within the framework of detection performances of infrared sensor estimation [32]. For an aircraft in a given atmospheric environment, the first order effects on the IRS are related to the spectral range, the presentation geometry, the aircraft speed and the engine power setting. Several sources of variability lead to a dispersion of the values likely to be observed: weather, aircraft aspect angles, aircraft type and optical properties. These uncertainties in the input data are propagated through the computer model to the output data. Therefore, the simulated result is an interval of possible IRS that should include the IRS measured at a given moment. Note that a single run with the computer model takes approximately 3 minutes (on a 64 bit Sun Fire workstation with four Intel Xeon processors).

Since many of the input factors of our IRS simulation were uncertain, we used a GSA to identify the most important ones. We considered a daylight air-to-ground full-frontal approach in France by a generic aircraft flying at low altitude (800-1200 ft). The infrared sensor was located on the ground, at a distance of 20 kilometers from the aircraft, in the flight direction. The aircraft was assumed to be spatially unresolved by the sensor.

4.2 Definition of inputs and outputs

Sixteen uncertain original inputs were analyzed. Eight were related to flight conditions and IR optical properties of aircraft surfaces: altitude (ALTI), Mach number (MACH), engine power setting (POWERS), emissivity of air intake (EAI), and to aspect angles of the aircraft: cap (CAP), yaw (YAW), roll (ROLL) and pitch angles (PITCH). Eight were related to atmospheric conditions: visibility (VIS), relative humidity (RH), ground air temperature (TA), atmospheric model (MODEL), aerosol model (IHAZE), cloud presence (CLOUDS), base altitude of the cloud layer (HBASE), and hour to compute solar position (HOUR). The three inputs, MODEL, IHAZE and CLOUDS, were qualitative, with two, three and two levels, respectively. The three aerosol models are the modalities of the IHAZE qualitative input: urban (IHAZE1), maritime (IHAZE2) and rural (IHAZE3). By using the three indicator variables of IHAZE, we had a total of 18 inputs. The modalities of MODEL and CLOUDS were coded as 0/1.

The outputs analyzed were the sensor differential irradiances between the target and the background, and we focused both on: (i) the 3-5 μm spectrally integrated target intensity, which corresponds to typical infrared sensor spectral ranges, designated as IRStot hereafter; and (ii) a multispectral intensity, integrated on 14 spectral bands equally spaced in the wave-number domain. The usefulness of multispectral or hyperspectral sensors for remote sensing assignments has been proven [33-35], and some studies [36-38] emphasize their potential for target detection.

We are currently working on the specification of a multispectral sensor for different missions such as aircraft detection or classification, and being able to perform a sensitivity analysis on a multispectral infrared signature is therefore of the utmost importance.

4.3 Construction of the dependent candidate set

We started by building a type of SFD by sampling 20,000 candidate experiments using the estimated distribution of variables related to atmospheric conditions, except for HOUR, and a Sobol low discrepancy sequence for the nine other variables: ALTI, MACH, POWERS, EAI, CAP, YAW, ROLL, PITCH and HOUR, as explained in [3,8].

RH, TA and HBASE were correlated, and their distribution was estimated on the basis of 3200 measured data using a non-parametric kernel reconstruction of marginal distributions, coupled with a dependence modeling based on a Normal copula [39,40]. This modeling was performed using the scientific library OpenTURNS (<http://www.openturns.org/>).

We also obtained a non-parametric reconstruction for the distribution of VIS and CLOUDS. A discrete uniform distribution was associated with the qualitative MODEL. The three aerosol models, IHAZE1, IHAZE2 and IHAZE3, are not equally probable in France, and we considered a $\{0.45; 0.1; 0.45\}$ probability law.

At this stage, the SFD* was obtained. With these 18 inputs, we built the design model matrix, \mathbf{X} , from this SFD*, by postulating an incomplete \mathcal{M}_d of 158 monomials: the 13 original continuous inputs, their 13 squared terms, their 13 cubic terms, and the 114 two-input and three-input interaction terms specified by the infrared signature expert. Therefore, the dimensions of \mathbf{X} were 20,000 rows and 158 ($= P$) columns.

4.4 Construction of the DOCED*

Firstly, since the information matrices, $\mathbf{X}_n^T \mathbf{X}_n$, built with \mathbf{X}_n selected in the preceding \mathbf{X} , were generally not invertible, this \mathbf{X} matrix was replaced by the \mathbf{Z} matrix formed by the first 50 principal components of a principal component analysis (these 50 components represented 99% of inertia). A DOCED sequence was extracted from the SFD*, from $n = n_0 = 50$ to $n = 1000$. A good compromise was obtained in 180 experiments between the maximization of the (log) normalized determinant (see Step 2 in Section 2.2) and the number of runs, if we observe Figure 1. The DOCED*, $\xi_{n^*=180, P=158}^D$, was then formed with $n^* = 180$.

4.5 Results and discussion

We focused here on the total indices, the T_3SIVIP and the GT_3SIVIP , since it was quite difficult to analyze all of the 158 monomial $ISIVIP$. However, it should be noted that the 20 largest $ISIVIP$ and $GISIVIP$ were associated with terms involving VIS or IHAZE2, and that the largest one-degree monomials were also associated with VIS and IHAZE2. The bar plot of Figure 2 presents the T_3SIVIP (in %) associated with the 18 variables.

The total IHAZE input contribution can be computed by summing the T_3SIVIP corresponding to the IHAZE1, IHAZE2 and IHAZE3 indicator variables. The T_3SIVIP associated with IHAZE is then 10.65, corresponding to the second highest T_3SIVIP in comparison with the others. Thus, two variables are particularly predominant: VIS and IHAZE and, more precisely, the maritime model of aerosol

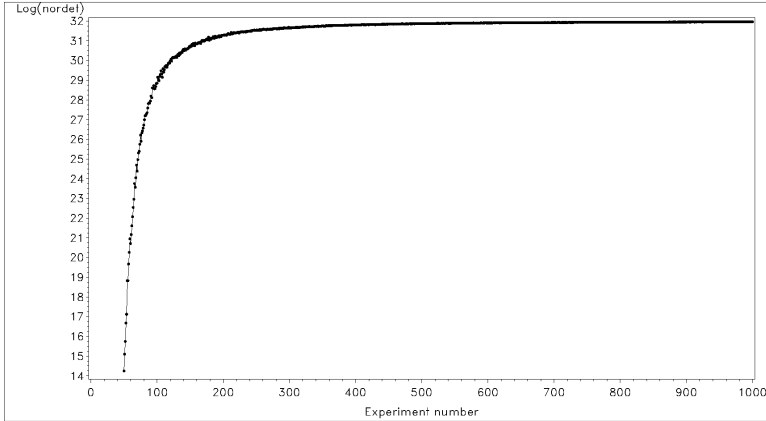


Figure 1: Evolution of the normalized determinant logarithm versus the number of experiments n .

IHAZE2, since indices associated with the two other models, IHAZE1 and IHAZE3 (urban and rural), are the lowest ones. This can be explained by the fact that urban and rural aerosols are quite similar and very different from the maritime one. Three additional variables also have an important impact on IRS variability: RH, TA and HOUR. It is then difficult to discriminate among the following eight variables in terms of total indices: MODEL, CAP, HBASE, EAI, MACH, POWERS, ALTI and PITCH. CLOUDS, ROLL and YAW have relatively negligible contributions to IRS variability since their total indices represent less than 10% of sum of all of the indices.

If we now focus on the multispectral infrared signature, the bar plot of Figure 3 presents the GT_3SIVIP (in %) associated with the 18 variables. As previously mentioned, the GT_3SIVIP associated with the total IHAZE contribution is computed by summing the GT_3SIVIP corresponding to the IHAZE1, IHAZE2 and IHAZE3 indicator variables. Their sum is 10.8 and corresponds to the second highest GT_3SIVIP in comparison with the others. Hence, as in the IRStot case, the two predominant variables are VIS and IHAZE. The three additional variables that also have an important impact on IRS variability are unchanged: RH, TA and HOUR. The main difference with IRStot stems from the higher indices corresponding to the three angles: PITCH, which extends from rank 13 up to rank 8, and, to a lesser extent, YAW and ROLL, which remain without influence. A possible explanation is that some of the 14 bands significantly emphasize the impact of phenomena that vary with aspect angles, such as the plume emission contribution or the airframe reflected light from the surrounding background.

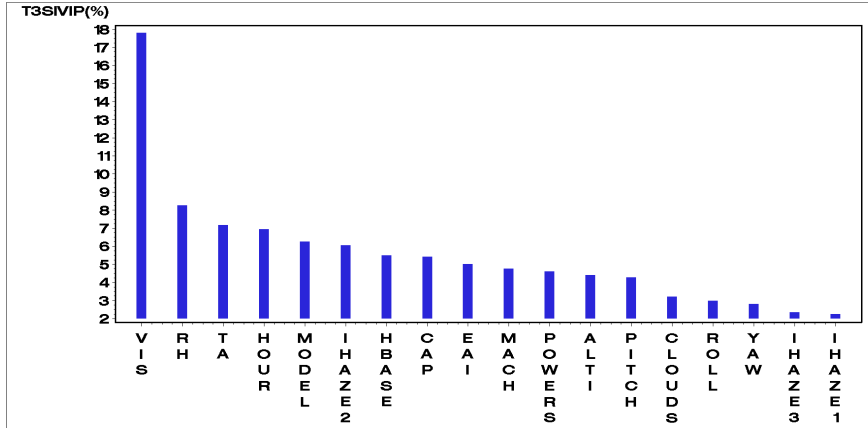


Figure 2: The 18 T_3SIVIP (in %) corresponding to the 18 inputs for the IRStot univariate output.

Finally, only five variables among the whole set actually influence the spectrally integrated and the multispectral infrared signature variability for the scenario considered, and they are all atmosphere-related factors. If we want to reduce the IRS uncertainty, we can thus combine the optics sensor with some detectors that can measure these atmospheric data. By virtue of the scenario chosen, a full frontal approach, large variations in IRS due to changes in aircraft aspect angle are excluded. The variations in aspect may dominate the rest of the variables. Hence, by essentially fixing the aspect to focus on, the variability of the meteorological phenomena may be given more importance in the overall IRS than if viewed from a different angle. Secondly, the low aircraft altitude and the long path between the sensor and the aircraft lead to an optical path that crosses the lowest atmospheric layers.

Consequently, the infrared transmission between sensor and aircraft and the atmospheric radiance, which both have a strong impact on the sensor differential irradiance, are themselves influenced by the quantity and type of low altitude atmospheric aerosols. Even if the sensitivity analyses of the multispectral and the spectrally integrated IRS lead to rather similar results for our scenario, it is especially interesting to be able to perform a sensitivity analysis in dozens of spectral bands simultaneously. We can thereby consider different selections and mergings of bands in order to perform the specification of a multispectral sensor.

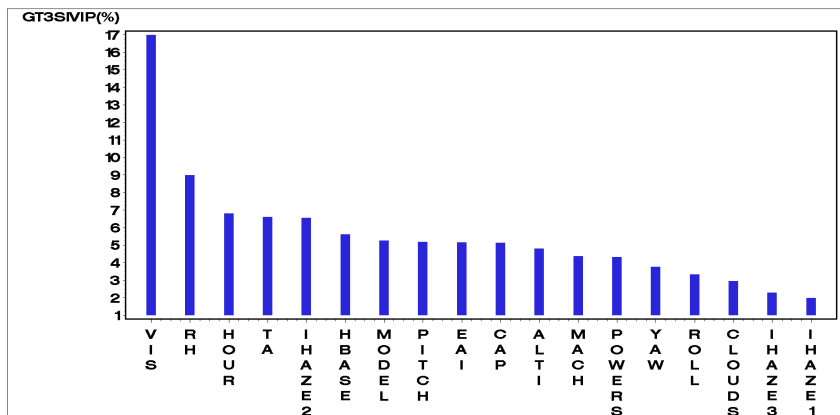


Figure 3: The 18 GT_3SIVIP (in %) corresponding to the 18 inputs for the multivariate output.

5 Conclusions

5.1 Application aspects

Thanks to the methodology proposed in this paper, we were able to identify five variables that have a strong impact on integrated and multispectral IRS variability for the chosen scenario with only 180 simulations, and we can now proceed to the next step, the establishment of an IRS simulation metamodel, in order to be able to estimate IRS dispersion. Since these first results are quite promising, it would be interesting to extend this model to other scenarios with more uncertain variables, and to other military objects.

5.2 Methodological aspects

Our method, designated the SIVIP method, is based on two very well-established theories (PLS regression and D-optimal designs), and it is very simple and very easy to put into practice by means of an SAS/IML procedure or an R-language procedure, both available upon request to the corresponding author (the SIVIP R-package will soon be available on the CRAN). To our knowledge, no other practical method exists to simultaneously manage the A, B and C constraints, notably with multivariate outputs and, therefore, no realistic comparison of our method can be made today with an alternative method. Only scientific experts in the field are qualified to validate the results at this time. This has been done in the real application problem presented in Section 4, as well as in the application presented in [15] with success. We hope that new competitive methods will be developed

in the near future, which will then make it possible for the researchers involved to test our dataset using their own methods. It should also be noted that even with independent inputs, this method can be useful in the frequently encountered situation of costly simulations, even if the *SIVIP* will generally be different from the *SSI* (even in this particular situation of independent inputs).

The *SIVIP* method works correctly in the situation where the user can postulate that the nonlinearity of the computer model is reasonable, i.e., if $d \leq 3$. Many real applications fall within this framework. However, if no prior knowledge is available about the nonlinearity, then neither the value of d nor the structure of the polynomial metamodel can be assumed (no previously chosen monomials are possible). In this situation, giving a large value to d can be a dangerous choice because of the well-known poor behavior of high-degree polynomials in terms of approximations. To address this difficulty, we propose an adaptive - and therefore parsimonious, in terms of the total simulation run number - version of the method for strong nonlinear computer models, which is currently being developed.

Acknowledgments: We would like to thank Gail Wagman, a professional English language translator, who corrected the English, and Régis Lebrun, from EADS Innovation Works, who performed the dependence modeling based on copulas.

References

- [1] A. Saltelli, *Sensitivity Analysis for Importance Assessment*, Risk Analysis 22 (2002), pp. 579-590.
- [2] M. Ellouze, J.P. Gauchi, and J.C. Augustin, *Sensitivity analyses applied to an exposure model of Listeria monocytogenes in smoked cold salmon*, Risk Analysis 30(5) (2010), pp. 841-852.
- [3] A. Saltelli, K. Chan, and E.M Scott (Editors), *Sensitivity analysis*, Wiley, New York, 2000.
- [4] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity analysis in practice - A guide to assessing scientific models*, Wiley, New York, 2004.
- [5] I.M. Sobol', *Sensitivity estimates for nonlinear mathematical models*, Mathematical Modeling Computational Experiments 1 (1993), pp. 407-414.
- [6] A. Saltelli, *Making best use of model evaluations to compute sensitivity indices*, Computational Physics Communications 145 (2002), pp. 280-297.
- [7] I.M. Sobol', *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*, Mathematics and Computers in Simulation 55 (1-3) (2001), pp. 271-280.
- [8] K.T. Fang, R. Li, and A. Sudjianto, *Design and modelling for computer experiments*, Chapman and Hall / CRC, London, 2006.
- [9] B. Sudret, *Global sensitivity analysis using polynomial chaos expansions*, Reliability Engineering & System Safety 93 (2008), pp. 964-979.

- [10] G. Blatman and B. Sudret, *Adaptive sparse polynomial chaos expansion based on Least Angle Regression*, Journal of Computational Physics 230 (6) (2011), pp. 2345-2367.
- [11] J. Jacques, C. Lavergne, N. Devictor, *Sensitivity analysis in presence of model uncertainty and correlated inputs*, Reliability Engineering & System Safety 91 (2006), pp. 1126-1134.
- [12] H. Rabitz, *Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs*, Procedia Social and Behavioral Sciences 2 (2010), pp. 7587-7589.
- [13] T.A. Mara and S. Tarantola, *Variance-based sensitivity indices for models with dependent inputs*, Reliability Engineering & System Safety 107 (2012), pp. 115-121.
- [14] S. Kucherenko, S. Tarantola, and P. Annoni, *Estimation of global sensitivity indices for models with dependent variables*, Computer Physics Communications 183 (2012), pp. 937-946.
- [15] J.P. Gauchi, S. Lehuta, and S. Mahévas, *Optimal Sensitivity Analysis under Constraints: Application to Fisheries*, Procedia Social and Behavioral Sciences 2 (2010), pp. 7658-7659.
- [16] J.P. Gauchi and S. Lefebvre, *PLS-based global sensitivity analysis for numerical models: an application to aircraft infrared signatures*, In: Proceedings of PLS2012, the 7th International Conference on PLS and Related Methods, (2012), pp. 43-45, Houston, Texas, USA.
- [17] S. Lefebvre and J.P. Gauchi, *Multidimensional Global Sensitivity Analysis for Aircraft Infrared Signature Models with Dependent Inputs*, 7th International Conference on Sensitivity Analysis of Model Output, July 1-4 2013, Nice, France.
- [18] J. Sacks, S.B. Schiller, and W.J. Welch, *Design for computer experiments*, Technometrics 31(1) (1989), pp. 41-47.
- [19] T.J. Santner, B.J. Williams, and W.I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.
- [20] P.Z.J. Qian, H. Wu, and C.F.J. Wu, *Gaussian process models for computer experiments with qualitative and quantitative factors*, Technometrics 50(3) (2008), pp. 383-396.
- [21] M. Lamboni, H. Monod, and D. Makowski, *Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models*, Reliability Engineering & System Safety 96(4) (2011), pp. 450-459.
- [22] A. Marrel, B. Iooss, M. Jullien, B. Laurent, and E. Volkova, *Global sensitivity analysis for models with spatially dependent outputs*, Environmetrics 22(3) (2011), pp. 383-397.
- [23] A.C. Atkinson and A.N. Donev, *Optimum experimental designs*, Clarendon Press, Oxford, UK, 1992.

- [24] V.V. Fedorov, *Theory of optimal experiments*, Academic Press, New York, 1972.
- [25] M.D. McKay, R.J. Beckman, and W.J. Conover, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, *Technometrics* 21 (1979), pp. 239-245.
- [26] R.L. Iman, W.J. Conover, *A distribution-free approach to inducing rank correlation among input variables*, *Communications in Statistics-Simulation and Computation* 11 (1982), pp. 311-334.
- [27] R.B. Nelsen, *An Introduction to Copula 2nd Ed.*, Springer, New York, 2006.
- [28] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, New York, 1997.
- [29] J.P.C. Kleijnen, W. van Beers, and I. van Nieuwenhuyse, *Constrained optimization in expensive simulation: Novel approach*, *European Journal of Operational Research* 202 (2010), pp. 164-174.
- [30] T.J. Mitchell, *An algorithm for the construction of D-optimum experimental designs*, *Technometrics* 16 (1974), pp. 203-210.
- [31] SAS Software, QC module, *Optex procedure*, Version 9.2, SAS Institute, North Carolina, USA, 2011.
- [32] S. Lefebvre, A. Roblin, S. Varet, and G. Durand, *A methodological approach for statistical evaluation of aircraft infrared signature*, *Reliability Engineering & System Safety* 95 (2010), pp. 484-493.
- [33] L.E Hoff, A.M. Chen, X. Yu, and E.M. Winter, *Enhanced Classification Performance from Multiband Infrared Imagery*, In: *Proceedings of the 30th IEEE Asilomar Conference on Signals, Systems and Computers* Asilomar, 29 (1996), pp. 837-841.
- [34] C.I. Chang, S.S. Chiang, *Anomaly detection and Classification for Hyperspectral Imagery*, *IEEE Trans. Geosci. Remote Sensing* 40 (2002), pp. 1314-1325.
- [35] U. Amato, A. Antoniadis, V. Cuomo, L. Cuttillo, M. Franzese, L. Murino, and C. Serio, *Statistical cloud detection from SEVIRI multispectral images*, *Remote Sensing of Environment* 112 (2008), pp. 750-766.
- [36] J. Karlholm and I. Renhorn, *Wavelength band selection method for multispectral target detection*, *Applied Optics* 41 (32) (2002), pp. 6786-6795.
- [37] J. Ahlberg and I. Renhorn, *Multi and Hyperspectral Target and Anomaly Detection*, Scientific Report of the Swedish Defence Research Agency (FOI), FOI-R-1526-SE, 2004.
- [38] D. Manolakis, G. Shaw, *Detection algorithms for hyperspectral imaging applications*, *IEEE Signal Process. Mag.* 19 (2002), pp. 29-43.
- [39] R.B. Nelsen, *An Introduction to Copula 2nd Ed.*, Springer, New York, 2006.
- [40] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, New York, 1997.

- [41] H. Martens and T. Naes, *Multivariate calibration*, Wiley, New York, 1992.
- [42] M. Tenenhaus, J.-P. Gauchi, and C. Ménardo, *Régression PLS et Applications*. *Revue de Statistique Appliquée* XLIII(1) (1995), pp. 7-63.
- [43] M. Tenenhaus, *Régression PLS - théorie et pratique*, Technip, Paris, 1998.
- [44] S. Wold, M. Sjöström, and L. Frikkson, *PLS Regression: a basic tool of chemometrics*, *Chemometrics and Intelligent Laboratory Systems* 58 (2001), pp. 109-130.
- [45] N. Kettaneh-Wold, *Analysis of mixture data with partial least squares*, *Chem Intel Lab Sys* 14 (1992), pp. 57-69.
- [46] H. Wold, *Estimation of principal components and related models by iterative least squares*, in *Multivariate Analysis*, Krishnaiah PR (Ed.). New York, Academic Press (1966), pp. 391-420.
- [47] SIMCA-P9 software, *User Guide and Tutorial*, Umetrics AB, Umea, Sweden, 2001.
- [48] A. Lazraq, R. Cléroux, and J.-P. Gauchi, *Selecting both latent and explanatory variables in the PLS1 regression model*, *Chemometrics and Intelligent Laboratory Systems* 66 (2003), pp. 117-126.

Appendix: PLSR background

A brief overview of PLSR

PLSR is a bilinear method for relating inputs to outputs, and that is very popular in the chemometrics field [41-44]. It is very different (particularly from the algorithm point of view) from the familiar Ordinary Least Squares Regression (OLSR) in that no matrix inversion is needed. For the sake of clarity, we use input notations in this Appendix that are different from those used in the other sections of this paper. Bold characters are used for vectors and matrices, but standard characters are used when a general meaning is assumed, e.g., a PLS component is designated as t_h , whereas in its vectorial form, it is designated as t_h .

The general goal of PLSR is to model, via linear combinations, the link between R explanatory (input) variables Φ_j , $j = 1, \dots, R$, and L (≥ 1) responses (outputs) Y_l , $l = 1, \dots, L$, both observed on the same N objects. We can represent the PLSR model by using the multivariate model form:

$$\mathbf{Y}_{N \times L} = \mathbf{\Phi}_{N \times R} \boldsymbol{\beta}_{R \times L} + \boldsymbol{\Sigma}_{N \times L} \quad (7)$$

where $\mathbf{Y}_{N \times L}$ is the observed multivariate output matrix, $\mathbf{\Phi}_{N \times R}$ is the input matrix, $\boldsymbol{\beta}_{R \times L}$ is a $(R \times L)$ -matrix of coefficients to be estimated, and $\boldsymbol{\Sigma}_{N \times L}$ is a $(N \times L)$ -matrix of error terms. PLSR presents some very strong advantages over OLSR: (i) the input variables - monomials built with these input variables can also be taken into account, including quadratic terms, interactions terms, cubic terms, etc. - can be highly correlated or even functionally linked, whereas the $\hat{\beta}_{PLS}$ estimates of $\beta_{R \times L}$ remain interpretable (e.g., see

the very well illustrated example in [45] where the inputs add up to one at each row of the design matrix); (ii) R and/or L can be larger than N ; (iii) information about the probability distributions of the Σ elements is not needed (it is considered to be a so-called distribution-free method). The first two cases, (i) and (ii), make it impossible to use OLSR. The PLSR algorithm is based on a very specific algorithm, the NIPALS algorithm [43,46,47]. Following are some formal elements.

Let \mathbf{E}_0 be the centred and scaled Φ matrix (i.e., mean = 0 and variance = 1 for each column); \mathbf{F}_0 , the centered and scaled \mathbf{Y} matrix; \mathbf{E}_h , the residual matrix from the decomposition of \mathbf{E}_0 by using h PLS components, t_h ; \mathbf{F}_h , the residual matrix from the decomposition of \mathbf{F}_0 by using h PLS components; \mathbf{F}_{lh} , the l th column of \mathbf{F}_h ; and H , the PLS component number retained. Therefore, the PLSR objective is to construct a linear combination, $\mathbf{u}_1 = \mathbf{F}_0 \mathbf{c}_1$, with columns of \mathbf{F}_0 , and a linear combination, $\mathbf{t}_1 = \mathbf{E}_0 \mathbf{w}_1$, with columns of \mathbf{E}_0 , by the maximization of the covariance between the t_1 and u_1 components, subject to the constraints $\|\mathbf{w}_1\|_2 = \|\mathbf{c}_1\|_2 = 1$. Hence, we obtain two variables, u_1 and t_1 , as correlated as possible, and best summing up \mathbf{E}_0 and \mathbf{F}_0 . The following regressions:

$$\begin{aligned}\mathbf{E}_0 &= \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{E}_1 \\ \mathbf{F}_0 &= \mathbf{t}_1 \mathbf{r}_1^T + \mathbf{F}_1\end{aligned}$$

are then performed. The \mathbf{E}_0 and \mathbf{F}_0 are deflated by $\mathbf{t}_1 \mathbf{p}_1^T$ and $\mathbf{t}_1 \mathbf{r}_1^T$, and second linear combinations, \mathbf{u}_2 and \mathbf{t}_2 , are computed, and so on. We finally obtain:

$$\mathbf{F}_0 = \mathbf{t}_1 \mathbf{r}_1^T + \dots + \mathbf{t}_h \mathbf{r}_h^T + \mathbf{F}_h \quad (8)$$

where the t_h are orthogonal between them. We can develop the vectorial form of \mathbf{t}_h as:

$$\mathbf{t}_h = \left[\sum_{j=1}^P \text{cov}^2(\mathbf{F}_{h-1}, \mathbf{E}_{h-1,j}) \right]^{-1/2} \sum_{j=1}^P [\text{cov}(\mathbf{F}_{h-1}, \mathbf{E}_{h-1,j})] \mathbf{E}_{h-1,j} \quad (9)$$

On the basis of (8), we can deduce the following norm decomposition, i.e., an empirical decomposition of the output variance (because the \mathbf{F}_{l0} in \mathbf{F}_0 are centered and scaled):

$$\|\mathbf{F}_0\|^2 = \|\mathbf{r}_1\|^2 \|\mathbf{t}_1\|^2 + \dots + \|\mathbf{r}_H\|^2 \|\mathbf{t}_H\|^2 + \|\mathbf{F}_H\|^2 \quad (10)$$

For one (univariate) centered-scaled output, \mathbf{F}_{l0} , we can then write:

$$\|\mathbf{F}_{l0}\|^2 = \text{var}(\mathbf{F}_{l0}) = r_1^2 \|\mathbf{t}_1\|^2 + \dots + r_H^2 \|\mathbf{t}_H\|^2 + \|\mathbf{F}_{lH}\|^2 \quad (11)$$

The significant number H of PLS components is given by means of a specific cross-validation test (see the end of this Appendix).

On the basis of (8) we can obtain the PLS regression equations:

$$\hat{Y}_l = \hat{\beta}_{l0} + \hat{\beta}_{l1} \Phi_1 + \dots + \hat{\beta}_{lM} \Phi_M \quad , l = 1, \dots, L$$

Definition of the VIP statistics in PLSR

Let the redundancy statistics be as follows:

$$Rd(Y; t_h) = \frac{1}{L} \sum_{l=1}^L cor^2(Y_l, t_h) \quad , \quad L \geq 1 \quad (12)$$

and

$$Rd(Y; t_1, \dots, t_H) = \sum_{h=1}^H \frac{1}{L} \sum_{l=1}^L cor^2(Y_l, t_h) \quad , \quad L \geq 1 \quad (13)$$

The *VIP* (Variable Importance in the Projection) statistics [43,47] for an Φ_j input in the presence of a univariate or multivariate Y output is then defined as:

$$VIP_{Hj} = \left[\frac{P}{Rd(Y; t_1, \dots, t_H)} \sum_{h=1}^H Rd(Y; t_h) w_{hj}^2 \right]^{1/2} \quad (14)$$

where the w_{hj} are the components of the w_h vectors. Note the important property $\sum_{j=1}^P VIP_{Hj}^2 = P$.

Obviously, the *VIP* statistics make sense if the value of the fitting-prediction criterion, Q_G^2 , (see below) is close enough to one, implying that $\hat{\mathcal{M}} = \Phi_{N \times R} \hat{\beta}_{R \times L}$ is a reasonable estimated approximation of the unknown underlying model for both fitting and prediction. In Section 2.3, we omitted the index H for the sake of brevity because non-ambiguity is possible once $\hat{\mathcal{M}}$ has been cross-validated with H components by means of the Q_G^2 .

Validation of the PLSR model

The PLSR model is validated by a cross-validation procedure based on the Q_{Gh}^2 and Q_G^2 statistics (bounded between 0 and 1) computed for each univariate output, detailed in [48]. The PLSR model is thus iteratively built, and these iterations are stopped by means of a procedure on the t_h components. A commonly used procedure is the one inspired by the rule proposed in [47]. This rule is the following: a t_h component (from $h = 1$) is retained if $Q_{Gh}^2 \geq 1 - (0.95)^2$. Finally, with the H retained components, the Q_G^2 cumulated index is computed [47]. Thus, this Q_G^2 index is a fitting-prediction type criterion typically designed for a PLSR model. Its major advantage, in comparison to a classical R^2 , is that it avoids overfitting. From a practical point of view, we have to choose a Q_G^2 threshold to be overtaken of at least 0.80.