C. R. Mecanique 341 (2013) 304–313

Contents lists available at SciVerse ScienceDirect

Comptes Rendus Mecanique

www.sciencedirect.com

Error estimate evaluation in numerical approximations of partial differential equations: A pilot study using data mining methods

Estimation d'erreur et approximation d'équations aux dérivées partielles : Une étude pilote fondée sur des méthodes de data mining

Franck Assous [a], Joël Chaskalovic [b,*]

[a] *Ariel University Center & Bar-Ilan University, Ramat Gan, Israel*
[b] *University Pierre-and-Marie-Curie, 4, place Jussieu, 75202 Paris cedex 05, France*

ARTICLE INFO

Article history:
Received 23 September 2012
Accepted 27 December 2012
Available online 5 February 2013

Keywords:
Data mining
Error estimate
Vlasov–Maxwell equations
Asymptotic analysis
Paraxial model

Mots-clés :
Data mining
Estimation d'erreur
Équations de Vlasov–Maxwell
Analyse asymptotique
Modèle paraxial

ABSTRACT

In this Note, we propose a new methodology based on exploratory data mining techniques to evaluate the errors due to the description of a given real system. First, we decompose this description error into four types of sources. Then, we construct databases of the entire information produced by different numerical approximation methods, to assess and compare the significant differences between these methods, using techniques like decision trees, Kohonen's cards, or neural networks. As an example, we characterize specific states of the real system for which we can locally appreciate the accuracy between two kinds of finite elements methods. In this case, this allowed us to precise the classical Bramble–Hilbert theorem that gives a *global* error estimate, whereas our approach gives a *local* error estimate.

RÉSUMÉ

Dans cette Note, on propose une nouvelle méthodologie fondée sur les techniques exploratoires du *data mining* afin d'évaluer les erreurs suscitées par la description d'un système physique donné. Pour ce faire, on identifie quatre type de sources d'erreurs. On constitue alors une base de données regroupant l'ensemble des résultats numériques calculés par différentes méthodes d'approximation, afin d'en comparer les différences significatives, par des techniques telles que les arbres de décision, les cartes de Kohonen, ou encore les réseaux de neurones. À titre d'exemple, nous caractérisons des états spécifiques du système réel pour lesquels on peut *localement* estimer la différence de précision entre deux méthodes d'éléments finis. Il est ainsi possible de préciser les résultats classiques du théorème de Bramble–Hilbert qui procure une estimation *globale*, alors que notre méthode propose une caractérisation *locale* des méthodes d'approximation considérées.

* Corresponding author.
E-mail addresses: franckassous@netscape.net (F. Assous), j.chaskalovic@free.fr (J. Chaskalovic).

1631-0721/$ – see front matter © 2013 Published by Elsevier Masson SAS on behalf of Académie des sciences.
http://dx.doi.org/10.1016/j.crme.2013.01.002

Version française abrégée

Les travaux présentés dans cette Note font suite à ceux publiés dans [1], où des techniques de *data mining* sont appliquées à l'exploitation de résultats numériques de méthodes d'approximation d'équations aux dérivées partielles.

Dans la présente Note, nous étendons le champ d'action de ces techniques au problème de l'évaluation des différentes sources d'erreurs liées au processus de production des résultats numériques. Nous introduisons quatre différentes sources d'erreur liées à ce processus, qui sont susceptibles de dégrader la précision de la description et la compréhension d'un système réel donné (S).

(i) ***L'erreur de modélisation*** **:** considérons un système (S) modélisé par un ensemble d'équations aux dérivées partielles (E). L'erreur de modélisation est définie comme l'écart entre la réalité du système (S) et la manière dont le modèle mathématique (E) l'appréhende. Pour deux modèles d'équations (E_1) et (E_2) modélisant (S), le problème est d'apprécier la qualité relative de ces deux modèles. Idéalement, il faudrait évaluer l'écart respectif entre chacun de ces modèles et le système (S). Dans la mesure où les règles du système (S) sont généralement inconnues, d'où sa modélisation par (E_1) ou (E_2), on se contente alors de définir, puis d'évaluer par des techniques de *data mining*, l'écart entre les modèles (E_1) et (E_2).

Cette approche, complémentaire aux techniques d'analyse phénoménologique, s'appuie sur une analyse *in vivo* et exhaustive des données produites par les deux modèles (E_1) et (E_2), résolus par une même méthode d'approximation, pour un niveau de précision donné. On cherchera alors à isoler par des techniques de *data mining*, dans la base de données constituée de l'ensemble des approximations produites par les deux modèles, les caractéristiques qui distinguent ces deux modèles (cf. [1] pour un premier exemple d'application).

(ii) ***L'erreur d'approximation*** **:** considérons un système (S) modélisé par un modèle d'équations (E), et résolu par deux types de méthodes numériques du même ordre (NUM_1) et (NUM_2), par exemple des différences finies d'ordre 2 et des éléments finis P_2 ou Q_2. Étudier l'erreur d'approximation consiste à determiner d'éventuelles conditions sur le système (S) qui discrimineraient les résultats produits par (NUM_1) et (NUM_2).

Plus précisément, désignons respectivement par u_1 et u_2 les solutions calculées par (NUM_1) et (NUM_2), et par u_{exact} la solution exacte du modèle (E). Soit $\|.\|$ une norme permettant d'apprécier *globalement* l'erreur d'approximation des méthodes numériques (NUM_1) et (NUM_2) telle que (1) et (2) soit satisfaits. Supposons de plus que la constante C_1 soit très petite devant C_2, la méthode (NUM_1) devrait alors être plus précise que (NUM_2). Cependant, il n'est pas exclu qu'il puisse exister des circonstances pour lesquelles (NUM_2) soit plus précis *localement* que (NUM_1), autrement dit que l'inégalité (3) soit vérifiée.

(iii) ***L'erreur de paramétrisation*** **:** c'est l'erreur qui mesure la sensibilité d'une méthode d'approximation par rapport au pas d'espace et de temps. Soient (E_{h_1}) et (E_{h_2}) les résultats numériques produits par une méthode d'approximation donnée – par exemple une méthode de différences finies d'ordre 2 – correspondant à deux pas de discrétisation h_1 et h_2 $(h_1 < h_2)$. D'un point de vue théorique, les résultats numériques (E_{h_2}) sont globalement plus précis que (E_{h_1}), de sorte que (4) et (5) sont satisfaits. Cependant, il peut exister des circonstances locales en (x,t) telles que l'inégalité (6) soit réalisée.

Afin d'identifier et de caractériser de telles situations, on se propose d'explorer, par des méthodes de *data mining*, la base de données constituée par (E_{h_1}) et (E_{h_2}), afin d'isoler localement les conditions qui pourraient expliquer l'inégalité (6).

(iv) ***L'erreur de discrétisation*** **:** il s'agit de l'erreur qui mesure localement la différence de précision entre deux modèles numériques (MN_1) et (MN_2), issus d'une même famille de méthodes d'approximations. Supposons, par exemple, que l'on résout numériquement un modèle mathématique donné (E) par la méthode des éléments finis P_1 et P_2. *A priori*, d'après le théorème de Bramble–Hilbert [2], les résultats obtenus par la méthode P_2 d'ordre 2 sont plus précis que ceux calculés par les éléments finis P_1 d'ordre 1. Cependant, il apparaît dans les estimations d'erreur (7) et (8) des constantes multiplicatives inconnues ou difficilement calculables. Nous proposons de déterminer, par des techniques de *data mining*, s'il existe des circonstances *locales* qui pourraient préciser ce résultat *global*, en recherchant des états (x,t) du système (S) pour lesquels l'inégalité (9) aurait lieu.

Dans cette Note, nous considérons, à titre d'exemple, l'analyse de l'*erreur de discrétisation* par des techniques de *data mining*. Nous introduisons, comme cadre physique, la propagation d'un faisceau de particules, modélisé par les équations instationnaires de Vlasov–Maxwell, dont les solutions seront décrites par des développements asymptotiques. Puis, on approche numériquement par éléments finis de Lagrange P_1 et P_2 le système d'équations aux dérivées partielles $(\mathbf{P})$ défini par (14)–(16).

On constitue une base de données regroupant l'ensemble des résultats numériques obtenus à chaque pas de temps et pour chaque nœud du maillage, par les deux types d'éléments finis. On obtient ainsi une base de donnée comprenant 125 000 enregistrements et 24 variables à analyser en colonne (12 par élément fini).

Afin d'identifer s'il existe des circonstances physiques locales paramétrées par (x,t) telles que l'inégalité (9) ait lieu, ou tout au moins, le cas limite défini par (20) soit réalisé, nous avons analysé la composante radiale E_r du champ électrique. Désignant par $E_r^{(1)}$ (resp. $E_r^{(2)}$) la valeur approchée par éléments finis P_1 (resp. P_2), on introduit la variable binaire « $P_1 vs P_2$ » définie par (22). La catégorie « Same Order » de cette variable nous permet de caractériser les situations correspondant à (20).

Après élimination des données non significatives de la base, nous obtenons une base de données comprenant, après filtrage, 7557 enregistrements. Un premier résultat découlant du tri à plat (cf. Tableau 1) nous permet de constater que 13,7% des éléments de cette base correspondent à la modalité « Same Order » de la variable « $P_1 vs P_2$ ».

Nous avons alors qualifié ces éléments par opposition aux 86,3% autres éléments décrivant des situations où $E_r^{(1)}$ et $E_r^{(2)}$ sont d'un ordre de grandeur significativement différent. Pour ce faire, dans un premier temps, nous avons réalisé, sous le logiciel de data mining IBM SPSS Modeler, une carte de Kohonen [4] afin de constituer des sous-groupes de la base de données dont le comportement serait très homogène vis-à-vis de la variable « $P_1 vs P_2$ ». Le Tableau 2 montre alors l'existence de quatre sous-groupes particulièrement homogènes ; deux par rapport à la valeur « Same Order » et deux autres par rapport à l'autre valeur « Different Order ». On caractérise alors ce qui diffère ces deux catégories de sous-groupes par une méthode de segmentation par un arbre de décision [4].

Le haut de l'arbre ainsi constitué est présenté sur la Fig. 1. La première segmentation de la racine de l'arbre de décision a été réalisée en fonction de la variable « time », qui a été identifiée comme la variable la plus discriminante parmi toutes les variables explicatives disponibles, et ce, avec un seuil optimal de segmentation correspondant au 42$^{\text{e}}$ pas de temps. Ainsi, pour des calculs effectués avant ce pas de temps critique, le nœud de l'arbre à considérer est celui entouré en bleu sur la Fig. 1, qui est composé de 78,9% d'éléments dont la valeur de la variable « $P_1 vs P_2$ » est « Same Order ». Autrement dit, avant le 42$^{\text{e}}$ pas de temps, les approximations $E_r^{(1)}$ et $E_r^{(2)}$ sont du même ordre ordre de grandeur. Il est donc inutile de mettre en œuvre des éléments finis P_2 dont le coût informatique est plus élevé que celui des éléments finis P_1. *A contrario*, après ce pas de temps critique, l'autre noeud produit par la segmentation est composé de 98,4% d'éléments caractéristiques de la valeur « Different Order ».

Cependant, nous ne sommes pas en mesure d'arbitrer pour autant, si cette différence d'ordre de grandeur correspond à une meilleure approximation produite par les éléments finis P_2. De même, nous n'avons pas encore pu mettre en évidence des situations correspondant strictement à l'inégalité (9). Ces deux problématiques constituent les éléments de nos travaux de recherche actuels. Il n'en reste pas moins que le cas limite où les deux types d'éléments finis que nous avons implémentés produisent des résultats équivalents ont pu être identifiés et qualifiés. Au-delà de la méthodologie générale proposée, nous avons illustré sur cet exemple que le coût de calcul des éléments finis P_2 n'est pas justifié pour les pas de temps inférieurs au 42$^{\text{e}}$.

1. Introduction

This Note is a follow-up of [1], in which was introduced a new methodology based on data mining techniques for numerical approximation analysis. Our aim is to extend this methodology to error estimate evaluation in numerical approximations of partial differential equations. It concerns the treatment of the sources of errors involved in any process of approximation, which describes a given real system by a mathematical model, solved by numerical approximation methods, and whose exploitation, *in fine*, will provide the understanding, the control and the forecast of this system.

More precisely, let us consider a real system (S) modeled by a set of partial differential equations (E). Generally, the solution of such a system is carried out by numerical approximations methods, as finite elements, finite volumes, finite differences, or any other appropriate numerical method.

Regarding the production of these approximations, it is important to consider the sources of errors which spoil the accuracy of the description and the understanding of the real system (S). In this Note, we introduce four sources of errors (details will be given in Section 2):

(i) the modeling error,
(ii) the approximation error,
(iii) the parameterization error,
(iv) the discretization error.

Then, we focus our presentation to a test case devoted to the discretization error. For this purpose, we consider numerical solutions to Vlasov–Maxwell equations, computed by an asymptotic model, and numerically discretized by two different finite elements. Hence, we characterize situations in space and in time, in which we can locally appreciate the accuracy between the two kinds of finite elements we have considered. We will see that it may sometimes be possible to clarify the classical Bramble–Hilbert theorem [2], that gives a *global* error estimate, whereas our approach may give a *local* error estimate.

2. Error sources and their evaluation by data mining methods

As it is known, there is always an overall error between a real system and the numerical results produced by its approximation. To evaluate this error with data mining methods, we propose to decompose this error into four possible sources:

(i) ***The modeling error.*** We define it as the gap between the reality of the considered system (S) and how the mathematical model (E) fits with it. Also, if one defines two mathematical models (E_1) and (E_2) which describe the same system (S), the question is to assess the relative quality of these two models. Ideally, one would like to evaluate the gap between each one of these models and the system (S). However, as the actual rules of the system (S) are unknown, which motivates its modeling by (E_1) or (E_2), we focus on circumstances which lead to significant differences between the models (E_1) and (E_2). An alternative could be to get experimental results that would be chosen as the reference for the relative performance evaluation between the models (E_1) and (E_2).

Classical examples illustrating this purpose relate to the comparison of the Navier–Stokes equations solutions versus the Stokes equations, or versus the Euler equations for problems in fluids mechanics, or solutions to the Poisson equation from the Darwin equation in electromagnetism.

The study of this modeling error can be performed by phenomenological analysis techniques [3]. These techniques are based on a fine understanding of the physics of the system (S), on the one hand, and on the meaning of the physical mechanisms described by each term constituting the model equations (E_1) and (E_2) on the other hand.

The method that we suggest is complementary and relies on an *"in vivo"* analysis considering all of the data produced by the two models (E_1) and (E_2), both solved by a same method of approximation, of a given level of accuracy. Then, we will seek to isolate, by the help of data mining techniques, the discriminant features distinguishing the results of the two models (E_1) and (E_2), in the database made up of all approximations produced by the two models. A first attempt can be found in [1].

(ii) ***The approximation error.*** Consider a physical system (S) governed by a given model of equations (E), whose solution is performed by two different numerical methods (NUM_1) and (NUM_2), which theoretically produce approximations of the *same order*. Studying the approximation error is to assess if there exist particular conditions which may significantly distinguish between the results produced by (NUM_1) and (NUM_2).

As an example, consider a finite-difference method of order 2 to be compared with a finite-element method also of order 2, like P_2 or Q_2 finite element. From the theoretical point of view, these two methods of approximation have *globally* the same order of convergence – denoted $O(h^2)$ – where h stands for a reference mesh size. However, the error estimates involve multiplicative constants of h^2, whose magnitude is either unknown or difficult to estimate, C_1h^2 for the numerical method (NUM_1) and C_2h^2 for the numerical method (NUM_2). One can try to identify the circumstances which might locally produce significant differences between the numerical methods (NUM_1) and (NUM_2).

More specifically, let us denote by u_1 and u_2 the solutions computed by the methods (NUM_1) and (NUM_2), and by u_{exact} the exact solution of the model (E). Let us also consider a norm $\|.\|$ to assess the global error of approximation of (NUM_1) and (NUM_2). We have [2]:

$$\left|u_1(x,t) - u_{\mathrm{exact}}(x,t)\right| \leqslant \|u_1 - u_{\mathrm{exact}}\| \leqslant C_1h^2 \tag{1}$$

and

$$\left|u_2(x,t) - u_{\mathrm{exact}}(x,t)\right| \leqslant \|u_2 - u_{\mathrm{exact}}\| \leqslant C_2h^2 \tag{2}$$

where x describes the space variable and t the time dependency.

Assuming that constant C_1 is very small before C_2 $(C_1 \ll C_2)$, the method (NUM_1) should be more efficient that the method (NUM_2). However, it does not prevent the system (S) from eventually producing states such as:

$$\left|u_1(x,t) - u_{\mathrm{exact}}(x,t)\right| \geqslant \left|u_2(x,t) - u_{\mathrm{exact}}(x,t)\right| \tag{3}$$

This corresponds to find circumstances for which the method (NUM_2) would be locally more "accurate" than the method (NUM_1).

We will seek to identify and to qualify such situations by exploratory data mining techniques. Typically, the segmentation by decision trees should be effective in this perspective, or any other supervised data mining technique (see [4]). Therefore, the algorithm of segmentation will identify, if any, and then will characterize, the rows of the database that correspond to the behavior described by expression (3).

Again, without absolute reference – that is the knowledge of the solution u_{exact} of the model (E) – one will focus on a relative comparison between the two methods (NUM_1) and (NUM_2). However, as in the case of the modeling error, if the exact solution u_{exact} is available, one could also proceed to an "absolute" comparison, and not only to a relative one. Moreover, if only experimental data are available, one could evaluate the error between each numerical method (NUM_1) and (NUM_2) and these corresponding data.

(iii) ***The parameterization error.*** This third item deals with the sensitivity of an approximation method with respect to the internal numerical settings, namely the time step or the mesh size. We still consider a real system (S) modeled by a set of equations (E).

Let us denote by (E_{h_1}) and (E_{h_2}) the numerical results produced by a given numerical method (for example, a finite-difference method of a given order), corresponding to two different mesh sizes h_1 and h_2 $(h_1 < h_2)$. As a consequence, (E_{h_2}) are more accurate than (E_{h_1}). Hence, denoting by $u_i(x,t)$ $(i = 1, 2)$, the approximate solution obtained with the mesh size h_i, a numerical method for instance of order 2 will give the following error estimates:

$$\left|u_1(x,t) - u_{\mathrm{exact}}(x,t)\right| \leqslant \|u_1 - u_{\mathrm{exact}}\| \leqslant Ch_1^2 \tag{4}$$

and

$$\left|u_2(x,t) - u_{\text{exact}}(x,t)\right| \leqslant \|u_2 - u_{\text{exact}}\| \leqslant Ch_2^2 \tag{5}$$

where C is the characteristic constant of the approximation error involved in this numerical method.
However, local circumstances in (x,t) such that:

$$\left|u_1(x,t) - u_{\text{exact}}(x,t)\right| \geqslant \left|u_2(x,t) - u_{\text{exact}}(x,t)\right| \tag{6}$$

cannot be excluded. To evaluate this phenomenon, we propose to collect all the numerical results (E_{h_1}) and (E_{h_2}) into an appropriate database, to explore it by data mining techniques, and then to isolate and to qualify the conditions that could significantly distinguish them locally.

(iv) ***The discretization error.*** For a given family of approximations methods – for instance the P_k finite element, $k = 1, 2, \ldots$ – we consider two numerical methods of this family, say (MN_1) and (MN_2), a typical example being the P_1 and P_2 numerical methods. The discretization error is defined as the error due to the difference of order between (MN_1) and (MN_2).
For example, suppose that we want to solve a given mathematical model (E) with finite elements P_1 and P_2. *A priori*, the Bramble–Hilbert theorem claims that, under certain conditions of regularity of the mesh and of the solution, the results obtained by the finite elements P_2 (of order 2) will be more precise than those computed by finite elements P_1 (of order 1).
Once again, the estimations of the approximation error contain multiplicative constants, unknown or difficult to estimate, before h and h^2 – where h is the average diameter of the mesh size. Using obvious notations, the error estimates in this case are written:

$$\left|u_1(x,t) - u_{\text{exact}}(x,t)\right| \leqslant \|u_1 - u_{\text{exact}}\| \leqslant C_1 h \tag{7}$$

and

$$\left|u_2(x,t) - u_{\text{exact}}(x,t)\right| \leqslant \|u_2 - u_{\text{exact}}\| \leqslant C_2 h^2 \tag{8}$$

where C_1, C_2 are two different constants. To deal with this error, we still propose to process data mining techniques to determine if there exist local discriminant conditions which could detail this global result. More precisely, we are looking for states defined by particular values of (x,t) such that:

$$\left|u_1(x,t) - u_{\text{exact}}(x,t)\right| \leqslant \left|u_2(x,t) - u_{\text{exact}}(x,t)\right| \tag{9}$$

Instead of *modeling error*, *approximation error*, *parameterization error* or *discretization error*, to better demonstrate the objectives that we described above, we rather have to retain a mathematical model which will be rich enough to present a consistent number of variables (the columns of the future databases established for each type of error) to provide an appropriate potential of exploration.

For this reason, the mathematical model considered as an example in this Note is the unsteady Vlasov–Maxwell equations whose solutions will be described by asymptotic expansions, see [5].

3. The mathematical model

Charged particle beams problems for non-collisional beams are often modeled by the time-dependent Vlasov–Maxwell system of equations (cf. [6]). However, numerically solving this model requires a large computational effort, so it is worth to derive approximate models leading to cheaper simulations (see [7–10]). We consider here an asymptotic paraxial model introduced in [5] and numerically solved in [11].

3.1. The axisymmetric Vlasov–Maxwell expansion

Let us consider a beam of charged particles with a mass m and a charge q which moves inside a perfectly conducting cylindrical tube, the z-axis being the axis of the tube and the optical axis of the beam. Since the domain under consideration is a bounded axisymmetric three-dimensional domain, we will therefore use the cylindrical coordinates (r,θ,z). We denote by Ω the transverse section of the tube of radius R, by Γ its boundary, so that $\Gamma = \{(r,\theta,z);\ r = R\}$, and by $\boldsymbol{\nu}$ the unit outward normal to Γ. For the sake of simplicity, we assume here that there are no external fields.

Each particle of the beam can be characterized by its position $\mathbf{X} = (r,\theta,z)$ and its velocity $\mathbf{V} = (v_r, v_\theta, v_z)$ in the phase space $(\mathbf{X},\mathbf{V})$. We also introduce the momentum $\mathbf{P} = (p_r, p_\theta, p_z)$ defined by

$$\mathbf{P} = \gamma m \mathbf{V}, \quad \gamma = \left(1 - \frac{\mathbf{V}^2}{c^2}\right)^{-1/2} \tag{10}$$

c denoting the speed of light in the vacuum.

Assuming that the beam is relativistic and non-collisional, the motion of these particles can be described in terms of particle distribution function $f(\mathbf{X},\mathbf{P},t)$, which satisfies the relativistic axisymmetric Vlasov equation:

$$\frac{\partial}{\partial t}(rf)+\frac{\partial}{\partial r}(v_r rf)+\frac{\partial}{\partial z}(v_z rf)+\frac{\partial}{\partial p_r}\left(\left(\frac{1}{r}p_\theta v_\theta+F_r\right)rf\right)+\frac{\partial}{\partial p_\theta}\left(\left(-\frac{1}{r}p_r v_\theta+F_\theta\right)rf\right)+\frac{\partial}{\partial p_z}(F_z rf)=0 \tag{11}$$

In Eq. (11), $\mathbf{F}=(F_r,F_\theta,F_z)$ denotes the electromagnetic Lorentz force given by $\mathbf{F}=q(\mathbf{E}+\mathbf{V}\times\mathbf{B})$, which describes how an electromagnetic field $\mathbf{E}=(E_r,E_\theta,E_z)$ and $\mathbf{B}=(B_r,B_\theta,B_z)$ acts on a particle with a given velocity. This electromagnetic field satisfies the axisymmetric Maxwell equations in the vacuum (see for instance [11]), where the charge and the current densities ρ and $\mathbf{J}=(J_r,J_\theta,J_z)$ – right-hand sides of the Maxwell equations – are obtained as the zero and the first moments of the distribution function f:

$$\rho=q\int f\,\mathrm{d}\mathbf{P},\qquad \mathbf{J}=q\int \mathbf{V}(\mathbf{P})f\,\mathrm{d}\mathbf{P} \tag{12}$$

One then exploits the physical/geometrical properties of the problem to derive paraxial asymptotic models, which approximate the Vlasov–Maxwell system with a known accuracy. Assuming a high-energy short beam, a paraxial relativistic model has been derived (cf. [5,11]) based on the following assumptions:

- the beam is highly relativistic, i.e., satisfies $\gamma\gg 1$,
- the dimensions of the beam are small compared to the longitudinal length of the device,
- the longitudinal particle velocities v_z are close to the light velocity c,
- the transverse particle velocities $(v_r^2+v_\theta^2)^{1/2}$ are small compared to c.

Since $v_z\simeq c$ for any particle in the beam, we rewrite the Vlasov–Maxwell equations in a frame, which moves along the z-axis with the light velocity c. Hence we set

$$\zeta=ct-z,\qquad v_\zeta=c-v_z \tag{13}$$

We denote by $\bar{v}$ the transverse characteristic velocity of the particles. Then, introduce a small parameter η defined by

$$\eta=\frac{\bar{v}}{c}\ll 1$$

The paraxial model described in [5,11] is derived by retaining the terms up to the third order in the asymptotic expansion of the distribution function f. Then, it has been proved in [5] that the third-order asymptotic expansion of f, namely

$$f^{(0)}+\eta f^{(1)}+\eta^2 f^{(2)}+\eta^3 f^{(3)}$$

is entirely determined from the expansion

$$(F_r,F_\theta)^{(0)}+\eta(F_r,F_\theta)^{(1)}+\eta^2(F_r,F_\theta)^{(2)}$$

of the transverse electromagnetic force (F_r,F_θ) up to order 2, and from the expansion

$$F_z^{(0)}+\eta F_z^{(1)}$$

of the longitudinal electromagnetic force F_z up to order 1 only.

Now, to determine these forces, it is sufficient to know:

(i) the principal parts $(E_r^{(0)},E_\theta^{(0)})$ and $(B_r^{(0)},B_\theta^{(0)})$ of the transverse electromagnetic field (zero order);
(ii) the expansions $E_z^{(0)}+\eta E_z^{(1)}$ and $B_z^{(0)}+\eta B_z^{(1)}$ of the longitudinal electromagnetic field up to the order 1;
(iii) the expansion $\boldsymbol{\mathcal{E}}^{(0)}+\eta\boldsymbol{\mathcal{E}}^{(1)}+\eta^2\boldsymbol{\mathcal{E}}^{(2)}$ of the transverse "pseudo-fields" $\boldsymbol{\mathcal{E}}$ up to the order 2, where $\boldsymbol{\mathcal{E}}=(\mathcal{E}_r,\mathcal{E}_\theta)$ denotes the transverse "pseudo-field" defined by $\mathcal{E}_r=E_r-cB_\theta$, $\mathcal{E}_\theta=E_\theta+cB_r$.

Finally, one can show that all the above electromagnetic components are a solution to the following problem (**P**) defined by

$$E_r^{(0)}=cB_\theta^{(0)}=\frac{1}{\varepsilon_0 r}\int_0^r \rho s\,\mathrm{d}s\quad\text{with } E_\theta^{(0)}=B_r^{(0)}=0 \tag{14}$$

$$\begin{cases}\dfrac{\partial E_z^{(1)}}{\partial r}=\dfrac{\partial B_\theta^{(0)}}{\partial t}\\ E_z^{(1)}(r=R)=0\end{cases}\quad\text{and}\quad\begin{cases}\dfrac{\partial B_z^{(1)}}{\partial r}=\mu_0 J_\theta\\ \displaystyle\int_0^R B_z^{(1)}r\,\mathrm{d}r=0\end{cases}\quad\text{with } E_z^{(0)}=B_z^{(0)}=0 \tag{15}$$

$$\begin{cases} \mathcal{E}_r^{(2)} = \dfrac{1}{r} \displaystyle\int_0^r \left(\mu_0 c J_\zeta - \dfrac{1}{c} \dfrac{\partial E_z^{(1)}}{\partial t} \right) s \, \mathrm{d}s \\ \mathcal{E}_\theta^{(2)} = -\dfrac{1}{r} \displaystyle\int_0^r \dfrac{\partial B_z^{(1)}}{\partial t} s \, \mathrm{d}s \end{cases} \quad \text{with } \mathcal{E}_r^{(0)} = \mathcal{E}_\theta^{(0)} = \mathcal{E}_r^{(1)} = \mathcal{E}_\theta^{(1)} = 0 \tag{16}$$

where J_ζ is defined by $J_\zeta = \rho c - J_z = q \int v_\zeta f \, \mathrm{d}\mathbf{V}$. In the rest of the paper, we will only consider the non-vanishing fields. Since there is no ambiguity, we will drop the superscript of the order. For example, $\mathcal{E}_\theta$ instead of $\mathcal{E}_\theta^2$ or E_z instead of E_z^1.

Eqs. (14)–(16) together with the Vlasov equation (11) form the problem (**P**) whose approximations will be explored by data mining techniques in terms of *discretization error* (see Section 4 below). Concerning the numerical discretization, on the one hand, the paraxial Vlasov equation is numerically solved by means of a particle method: it consists in approximating the function $rf(\mathbf{X}, \mathbf{P}, t)$ at any time t by a linear combination of delta distributions in the phase space $(\mathbf{X}, \mathbf{P})$:

$$rf(\mathbf{X}, \mathbf{P}, t) = \sum_k w_k \, \delta\big(\mathbf{X} - \mathbf{X}_k(t)\big) \delta\big(\mathbf{P} - \mathbf{P}_k(t)\big) \tag{17}$$

where w_k denotes the constant weight of the particle k. Its position in the phase space $\mathbf{X}_k = (r, \zeta)$ and $\mathbf{P}_k = (p_r, p_\theta, p_z)$ is a solution to the differential system:

$$\begin{cases} \dfrac{\mathrm{d}r}{\mathrm{d}t} = \dfrac{1}{\gamma m} p_r, \quad \dfrac{\mathrm{d}\zeta}{\mathrm{d}t} = c - \dfrac{1}{\gamma m} p_z \\ \dfrac{\mathrm{d}p_r}{\mathrm{d}t} = \dfrac{1}{\gamma m r} p_\theta^2 + F_r, \quad \dfrac{\mathrm{d}p_\theta}{\mathrm{d}t} = -\dfrac{1}{\gamma m r} p_r p_\theta + F_\theta, \quad \dfrac{\mathrm{d}p_z}{\mathrm{d}t} = F_z \end{cases} \tag{18}$$

together with initial conditions. On the other hand, a variational formulation will be derived from Eqs. (14)–(16) and then discretized by two finite-element methods: a P_1 and a P_2 Lagrange's finite elements that will be compared. Details can be found in [12].

4. Data mining methods for error source analysis

In this Note, we illustrate on an example how to use data mining methods for error source analysis. More precisely, we consider the case of the *discretization error* (iv), that is the error due to the approximation of a given problem by (MN_1) and (MN_2), two numerical methods of the same family. Using the FreeFem++ package [13], we discretize the problem (**P**) described above with a P_1 and a P_2 Lagrange's finite elements to approximate problem (**P**) solutions. We then collect in a suitable database all the numerical results computed by the two families of finite elements.

Hence, the database we consider is composed by data computed at each time step and for each node of the concerned space grid, a given row of the database corresponding to the approximations of all of the physical unknowns, at a given time step and mesh node.

Then, the set of variables which correspond to the columns of the database we considered is listed below:

$$v_r^{(i)}, v_\theta^{(i)}, v_\zeta^{(i)}, E_r^{(i)}, E_z^{(i)}, B_z^{(i)}, \mathcal{E}_r^{(i)}, \mathcal{E}_\theta^{(i)}, J_r^{(i)}, J_\theta^{(i)}, J_\zeta^{(i)}, \rho^{(i)} \quad (i = 1, 2) \tag{19}$$

where the exponent i specifies that the approximations are computed by the P_i Lagrange's finite elements.

Considering all the 100 time steps and the 1250 space nodes, the database we treated was composed by more than 125 000 rows and 24 variables to be analyzed.

4.1. Data mining principles and objectives of exploration

Data mining is an activity of information extraction, whose goal is to discover hidden or *a priori* unknown facts contained in databases.

Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results.

Our objective is to appreciate the difference of accuracy between the P_1 and P_2 approximate problem (**P**) solutions. To this end, we want to determine if there locally exist rows in the database, namely, (r_j, ζ_k, t_n), such that the P_1 method would bring a better approximation or, at least, an equivalent one than the P_2 finite element. For doing this, we focused our attention to identify subgroups in the database such that the numerical approximations computed by the two P_i $(i = 1, 2)$, finite elements are with *"the same order"*.

For the sake of illustration, let us restrict ourselves to the radial component of the electrical field E_r. To identify the approximations of *"the same order"*, we will only keep in the database the rows such that neither $E_r^{(1)}$ nor $E_r^{(2)}$ are too small. This is to eliminate two possible situations:

Table 1
Respective proportions of the "$P_1 vs P_2$" categories.

	Count	Percent
Different Order	6522	86.3
Same Order	1035	13.7

- the case where $E_r^{(1)}$ is small with respect to $E_r^{(2)}$ or reciprocally. This case means that $E_r^{(1)}$ and $E_r^{(2)}$ are not of *"the same order"*;
- the case where $E_r^{(1)}$ and $E_r^{(2)}$ are both small. It corresponds to the situation where we cannot determine if $E_r^{(1)}$ and $E_r^{(2)}$ are or not of *"the same order"*.

In terms of numerical values, each component $E_r^{(1)}$ and $E_r^{(2)}$ is assumed to be small at a given time step t_n and for a given node (r_j, ζ_k), if its value is smaller than 5% of the maximum of all the values of $E_r^{(1)}$ and $E_r^{(2)}$ respectively, available in the database.

By applying this rule, we extract the corresponding rows and we get a dataset made of 7680 rows to be explored, for identifying, if there exist, rows such that:

$$\left|E_r^{(1)}(r_j,\zeta_k,t_n) - E_{r,\text{exact}}(r_j,\zeta_k,t_n)\right| \simeq \left|E_r^{(2)}(r_j,\zeta_k,t_n) - E_{r,\text{exact}}(r_j,\zeta_k,t_n)\right| \tag{20}$$

This situation would precise the Bramble–Hilbert theorem which can be written in our case:

$$\left\|E_r^{(1)} - E_{r,\text{exact}}\right\| \leqslant C_1 h, \quad \text{and} \quad \left\|E_r^{(2)} - E_{r,\text{exact}}\right\| \leqslant C_2 h^2 \tag{21}$$

where C_i $(i = 1, 2)$, are two given but unknown constants.

The situation described by (20) means that, in a certain sense (to be defined), $E_r^{(1)}$ and $E_r^{(2)}$ have the "same numerical order". To identify such situations, let us first define the notion of "same numerical order". For this purpose, we introduce a threshold α and a new binomial variable called "$P_1 vs P_2$" as follows:

$$P_1 vs P_2 \equiv \left| \begin{array}{ll} \textit{Same Order}, & \text{if } |E_r^{(2)} - E_r^{(1)}| \leqslant \alpha \\ \textit{Different Order}, & \text{if not} \end{array} \right. \tag{22}$$

where $\alpha = 0.65$ in our example corresponding to 5% of the maximum of the absolute difference between $E_r^{(1)}$ and $E_r^{(2)}$ found in all the dataset. In such a way, the variable "$P_1 vs P_2$", especially its value "Same Order", will allow us to detect and to characterize situations where relation (20) holds, if any.

The next step is to choose a so-called *target variable* (the variable to be explained) – in our example "$P_1 vs P_2$" – and to process *ad hoc* data mining techniques to qualify the two different categories of this *target* variable.

The first result we found confirmed our suspicion: the dataset contains a non-negligible quantity of rows such that (20) is satisfied. Indeed, as shown in Table 1, almost 14% of the rows in the dataset are in this case. Now, to qualify these 14% of rows, namely the cluster composed by the "Same Order" of the target variable "$P_1 vs P_2$", we process two data mining techniques. The first one is called Kohonen's cards and the second one, the decision trees.

Kohonen's cards [4] belong to the *unsupervised data mining* methods and are based on neural networks [4]. Unsupervised data mining means that no target variable is defined to be modeled or explained but, depending on the given set of the predictor variables (all the available variables except the target ones), the objective of the method is to build clusters inside a given population such that each subgroup has to be homogeneous regarding relations that have to be discovered.

Decision trees [14] belong to the *supervised data mining* methods to process segmentation. The purpose of segmentation is to constitute homogeneous subgroups inside a given population regarding a target variable which is to be explained versus predictor variables. This is processed by an algorithm of segmentation which is basically a minimization of the standard deviation for the concerned target variable.

Then, the idea of our exploration process is to begin with the constitution of homogeneous groups, called clusters, in the dataset, and this, without giving to the algorithm of Kohonen's cards the knowledge of the binomial variable "$P_1 vs P_2$" for each row of the dataset.

Therefore, when the clusters have been discovered by Kohonen's cards, we consider again, for all of the rows which define each cluster, the value of our target variable "$P_1 vs P_2$".

The result we got is very interesting: we found, *a posteriori*, four very homogeneous clusters; two relative to the category "Same Order" and two relative to the category "Different Order". As shown in Table 2, more than 90% of the elements of each cluster correspond to one of the two categories of the target variable "$P_1 vs P_2$".

So, the challenge is now to discover the rules of these two groups, the blue ones corresponding to the "Different Order" and the red ones to the "Same Order".

This was processed by the help of a decision tree, also computed under the software *Modeler* of *IBM SPSS Inc.* Results are shown in Fig. 1 and correspond to the top of the tree. The first segmentation which appears on the decision tree highlights the most discriminated predictor variable, in the set of all the available potential predictors in the dataset. One can observe

Table 2
Kohonen's clusters *versus* the target variable "$P_1 vs P_2$".

	Group 1	Group 2	Group 3	Group 4
Different Order	93.7%	91.5%	6.5%	5.6%
Same Order	6.3%	8.5%	93.5%	94.4%

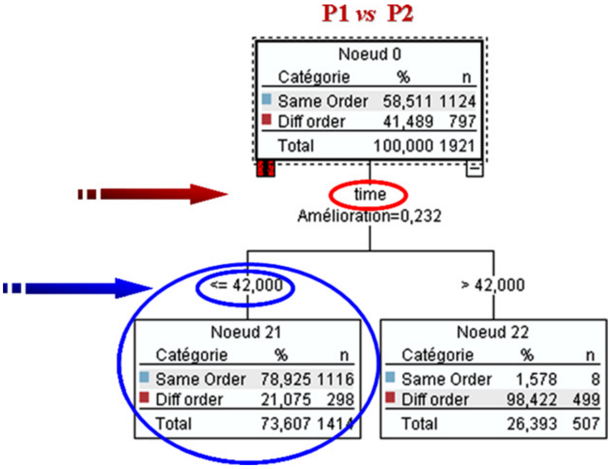

Fig. 1. Decision tree.

that the time is detected as this predictor, with a corresponding computed optimal threshold equal to the 42nd time step over a hundred computed.

This means that when the time is smaller than the 42nd time step, the corresponding blue node in the decision tree is very homogeneous regarding the value "Same Order". More precisely, almost 80% of the elements of this node have this value for the target variable "$P_1 vs P_2$".

We can then conclude that for time steps which are smaller than the 42nd, the implementation of P_2 finite elements was over qualified and then, the cost of the computation could not be justified anymore.

At the opposite, when one considers time steps which are greater than the 42nd, the corresponding cluster, which is the other node in the decision tree (see Fig. 1), is constituted by more than 98% of rows which correspond to the value "Different Order" of the target variable "$P_1 vs P_2$".

Unfortunately, we are not in position to give any conclusion in this case. Indeed, the criteria we choose to define the value "Different Order" do not allow us to make the difference when the finite elements P_1 will lead to a better result in comparison with the P_2 ones.

This is one of the opened questions we have to treat, certainly by the help of above data mining techniques we already proceeded.

5. Conclusion

In this Note, we have proposed original techniques based on data mining methods applied to numerical approximations analysis for PDEs. Our aim was to qualify the numerical approximations computed by P_1 and P_2 finite elements we implemented on Vlasov–Maxwell equations solved by asymptotic expansions which modeled an ultra relativistic beam of particles. We focused our analysis to the radial component of the electrical field E_r and we found, as a complement of the Bramble–Hilbert theorem, that a respectable proportion of a suitable database (almost 14% of the elements) correspond to a numerical evaluation of E_r which has a same order if one compares the two families of finite elements we considered, say, P_1 and P_2. Then, by the help of two data mining techniques (Kohonen's cards and decision tree), we identify the most discriminate predictor which describes the elements such that the evaluation between the two finite elements gave similar numerical results. It is the time that was found with a corresponding threshold which is the 42nd time step. As a consequence, when the time is smaller this time strep, both finite elements P_1 and P_2 have similar results. So, one has not to assume the over cost of the P_2 finite-element method, when the time is smaller than the 42nd time step.

This result is in agreement with standard results describing time discrete approximations. Indeed, it is well known that when the time grows, the accumulated errors with the P_2 finite elements will be smaller than with the P_1 ones. But, what does it mean when time grows? Here, we began to give an answer identified by the decision tree we processed after a first step of data modeling computed by Kohonen's cards. When the time steps are smaller than the 42nd, one has to only consider the P_1 finite elements to decrease the cost of computation, but without diminishing the accuracy of the numerical results.

We do believe to extend and to generalize this kind of results to the other cases of errors we defined in this Note, and we suggest that data mining techniques must be applied more systematically to the numerical approximation analysis as it is in a lot of other applications. This is already the case in marketing and communication [15] and [16], in medicine [17], in biology [18], or in engineering sciences [1].

References

[1] F. Assous, J. Chaskalovic, Data mining techniques for scientific computing: Application to asymptotic paraxial approximations to model ultra-relativistic particles, J. Comput. Phys. 230 (2011) 4811–4827.
[2] J. Chaskalovic, Mathematical and Numerical Methods for Partial Differential Equations, Springer-Verlag, 2013.
[3] E.S. Taylor, Dimensional Analysis for Engineers, Clarendon Press, Oxford, 1974.
[4] R. Lefébure, G. Venturi, Data Mining – Gestion de la relation client, Eyrolles, 2001.
[5] G. Laval, S. Mas-Gallic, P.-A. Raviart, Paraxial approximation of ultrarelativistic intense beams, Numer. Math. 69 (1) (1994) 33–60.
[6] C.K. Birdsall, A.B. Langdon, Plasmas Physics via Computer Simulation, McGraw–Hill, New York, 1985.
[7] M.A. Mostrom, D.I. Mitrovich, D.I.R. Welch, The ARCTIC charged particle beam propagation code, J. Comput. Phys. 128 (2) (1996) 489–497.
[8] S. Slinker, G. Joyce, J. Krall, R.F. Hubbard, ELBA – A three dimensional particle simulation code for high current beams, in: Proc. of the 14th Inter. Conf. Numer. Simul. Plasmas, Annapolis, 1991.
[9] P. Degond, P.-A. Raviart, On the paraxial approximation of the stationary Vlasov–Maxwell, Math. Models Methods Appl. Sci. 3 (4) (1993) 513–562.
[10] P.A. Raviart, E. Sonnendrucker, A hierarchy of approximate models for the Maxwell equations, Numer. Math. 73 (3) (1996) 329–372.
[11] F. Assous, F. Tsipis, Numerical paraxial approximation for highly relativistic beams, Comput. Phys. Commun. 180 (2009) 1086–1097.
[12] F. Assous, J. Chaskalovic, On the error estimate evaluation in PDE's by data mining techniques, 2012, in preparation.
[13] Frédéric Hecht, FreeFem++, Numerical Mathematics and Scientific Computation 3.7, Laboratoire J.L. Lions, Université Pierre et Marie Curie, http://www.freefem.org/ff++/, 2010.
[14] L. Rokach, O. Maimon, Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing Company, 2001.
[15] J. Chaskalovic, A new approach in Media/Marketing Databases explorations for application in E-business, in: National Congress of IREP, Paris, 1999.
[16] J. Chaskalovic, A. Vanheuverzwyn, Innovation in estimations: A reliable approach for radio audience indicators, in: Proc. Esomar, WM^3 2007, Dublin, 3–6 June 2007.
[17] X.L. Nguyên, J. Chaskalovic, et al., Insomnia symptoms and CPAP compliance in OSAS patients: A descriptive study using Data Mining methods, Sleep Med. 11 (8) (2010) 777–784.
[18] O. Kulski, J. Chaskalovic, et al., Explicative factors for prognostics IIU: exploration on 2089 cycles done with statistical and data mining tools, in: 9th Meeting of the French Federation of the Reproduction Studies, Palais des Congrés, Paris, 2004.