Data mining techniques for scientific computing: Application to asymptotic paraxial approximations to model ultrarelativistic particles

Franck Assous\textsuperscript{a,*}, Joël Chaskalovic\textsuperscript{b}

\textsuperscript{a} Ariel University Center, 40700 Ariel, Israel & Bar-Ilan University, 52900 Ramat-Gan, Israel
\textsuperscript{b} University Pierre and Marie Curie, Paris, France & Ariel University Center, 40700 Ariel, Israel

\textbf{A R T I C L E  I N F O}

\textbf{Article history:}
Received 27 August 2010
Received in revised form 12 January 2011
Accepted 2 March 2011
Available online 9 March 2011

\textbf{Keywords:}
Data mining
Asymptotic methods
Paraxial approximation
Vlasov–Maxwell equations

\textbf{A B S T R A C T}

We propose a new approach that consists in using data mining techniques for scientific computing. Indeed, data mining has proved to be efficient in other contexts which deal with huge data like in biology, medicine, marketing, advertising and communications. Our aim, here, is to deal with the important problem of the exploitation of the results produced by any numerical method. Indeed, more and more data are created today by numerical simulations. Thus, it seems necessary to look at efficient tools to analyze them. In this work, we focus our presentation to a test case dedicated to an asymptotic paraxial approximation to model ultrarelativistic particles. Our method directly deals with numerical results of simulations and try to understand what each order of the asymptotic expansion brings to the simulation results over what could be obtained by other lower-order or less accurate means. This new heuristic approach offers new potential applications to treat numerical solutions to mathematical models.

\textcopyright 2011 Elsevier Inc. All rights reserved.

\textbf{1. Introduction}

With the development of super-computers, numerical methods produce today a huge quantity of numerical results. This is especially true with the availability of massively parallel computers where, more and more data are computed by numerical simulations. Moreover, this problem is certainly the most critical for unsteady and three-dimensional problems, in which are computed, at each node of a given mesh, approximations of scalar fields, vector or tensor components. It seems thus necessary to look at new powerful tools to analyze the corresponding volumetric simulated data.

In addition, the numerical results are generally only partially exploited. Typically, one exploits the time evolution of a variable, a snapshot of a given quantity at a fixed time, or some specific diagnostics. For instance for the Vlasov Maxwell simulation that we will consider here, one often depicts the position-velocity phase space at a given time.

The purpose of this paper is to present a new approach based on data mining techniques applied to scientific computing. Why data mining techniques could help scientific computing? Such techniques have proved to be efficient in other contexts which deal with huge data, like in biology [5], medicine [8,9], marketing [18], advertising and communications [6,7]. Indeed, no one has to prove that these methods are relevant and well adapted to analyze such data.

In this paper, we focus our presentation to a test case dedicated to an asymptotic paraxial approximation to model ultrarelativistic particles. Indeed, solving the time-dependent Vlasov Maxwell equations, which is one of the most complete mathematical model for collisionless plasma or non-collisional beams, can lead to very expensive computations especially in a three-dimensional domain. Therefore, whenever possible, it is worthwhile to take into account the particularities of the physical problem to derive approximate asymptotic models leading to cheaper simulations.

\* Corresponding author.
\textit{E-mail address:} francaassous@netscape.net (F. Assous).

0021-9991/\$ - see front matter \textcopyright 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2011.03.005
However, despite some theoretical convergence results, it is not always easy to determine which terms to retain in the asymptotic expansion to get a sufficiently precise but not too expensive model. For instance, in the case of the paraxial approximation we consider here, despite a convergence result proved in [12], a numerical study, see [3] shows that the comparison of the different orders of approximations is not obvious. In other terms, the asymptotic models are often difficult to compare directly one to the other.

In this paper, we propose a methodology that we hope helpful for many domains of applications in scientific computing. Basically, this approach can be decomposed into the following steps:

- **Mathematical model elaboration.** In this work, an asymptotic approximation of Vlasov–Maxwell equations.
- **Numerical approximation processed.** Here a discretization by finite differences and particle-in-cell methods.
- **Collection of all the numerical data in a well structured and adapted database.**
- **Definition of relevant target variables to be explained and potential predictors.** In this study, the “native” numerical variables, namely corresponding to the computed approximations, together with those deduced form the formal analysis.
- **Implementation of pertinent data mining tools (supervised or not) in relation with the objectives of the exploration.** In this paper, the use of decision trees.
- **Analysis of the results produced by the data mining due to appropriate statistical techniques.** Here, statistical tests and spearman’s rho correlation, to appreciate the main differences between the nodes identified in the decision tree.

In this study, we will deal with the comparison of asymptotic models. Hence, data mining techniques will be used to perform a sensitivity as a continuation of [1], analysis of approximate models. Our method directly processes numerical results of simulations and try to understand what each order of the asymptotic expansion brings to the simulation results over what could be obtained by other lower-order or less accurate means.

For this purpose in the present study, we consider a second order accurate formulation (denoted by $M_2$) and a first order model (denoted by $M_1$) derived from the same asymptotic expansion, see [12,3]. Beyond the mathematical analysis or theoretical results of approximation ($M_2$ is obviously more “accurate” than $M_1$!), we propose here to check a given order accuracy on the numerical results themselves. Of course, one could assess this accuracy with a series of numerical test cases where the parameters of the asymptotic expansion vary. But this can lead to tedious computations. First, since there are often several parameters, so that one of them varies whereas other aspects are fixed. Secondly, since we often get a huge quantity of results to compare. Moreover, contrary to the data mining techniques, such sensitivity studies are not always supported by a theory. For these reasons, we propose an alternative based on those techniques. In this approach, the numerical results considered as a “database” are entirely used and data mining techniques allows us to exploit all the time steps at all the nodes of the mesh.

This paper is organized as follows. In Section 2, we first recall the time-dependent Vlasov–Maxwell system of equations, written in axisymmetric geometry, from which the ultrarelativistic paraxial model is derived. We then expose the approximate models $M_1$ and $M_2$, that will be compared in section 3. After recalling some necessary material about data mining methodology, we present our specific approach based on decision trees. Section 4 is devoted to numerical results, in which the model $M_1$ and $M_2$ are compared. Finally, a conclusion is drawn. Let us emphasize that, beyond this particular study, we think that this new heuristic approach can operate in a lot of other domains of applications, where numerical simulations take a central part to understand the features of a given system.

## 2. The paraxial model

To solve charged particle beams or plasma physics problems for collisionless plasma or non-collisional beams, one of the most complete mathematical models is the time-dependent Vlasov–Maxwell system of equations, cf. [4]. However, the numerical solution of such a model requires a large computational effort. Therefore, whenever possible, it is worth to take into account the particularities of the physical problem to derive approximate models leading to cheaper simulations, see [13,16,10,14]. Note that this requirement of “cheaper computations” is often the main motivation of most of the asymptotic models. This remark would give to the approach exposed here the possibility of a lot of other applications. In this paper, we will concentrate on an asymptotic paraxial model that has been introduced in [12] and numerically solved in [3] for high energy short beams. We recall it here by rewriting the equations differently than in [3], in a way that is well adapted to data mining analysis.

### 2.1. 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 $(r, \theta, z)$. We denote by

1 In the transverse direction.
the transverse section of the tube of radius $R$, by $\Gamma$ its boundary, so that $\Gamma = \{(r, \theta, z); r = R\}$, and by $v$ the unit outward normal to $\Gamma$. For the sake of simplicity, we assume here that there is no external fields.

Each particle of the beam can be characterized by its position $X = (r, \theta, z)$ and its velocity $V = (v_r, v_\theta, v_z)$ in the phase space $(X, V)$. Assume that the beam is relativistic and non collisional, we introduce the momentum $P = (p_r, p_\theta, p_z)$ (defined below). Hence, the motion of these particles can be described in terms of particle distribution function $f(X, P, t)$ by the relativistic Vlasov equation. In addition, if the particle distribution function and the data are independent of $\theta$, we obtain that the function $rf$ – usually used instead of $f$ in the axisymmetric case – satisfies the axisymmetric Vlasov equation

\[
\frac{\partial}{\partial t}(rf) + \nabla \cdot (v_r rf) + \frac{1}{c^2} \left( \frac{1}{r} p_r v_0 + F_r \right) rf + \frac{\partial}{\partial p_r} \left( - \frac{1}{c^2} p_r v_0 + F_0 \right) rf + \frac{\partial}{\partial p_z} (F_z rf) = 0.
\]

where

\[
P = \gamma m V, \quad \gamma = \left( 1 - \frac{V^2}{c^2} \right)^{-1/2}.
\]

c denoting the speed-of-light in the vacuum.

In Eq. (1), $F = (F_r, F_\theta, F_z)$ denotes the electromagnetic Lorentz force given by $F = q(E + V \times B)$, that describes how an electromagnetic field $E = (E_r, E_\theta, E_z)$ and $B = (B_r, B_\theta, B_z)$ acts on a particle with a given velocity. This electromagnetic field satisfies the axisymmetric Maxwell equations in the vacuum.

In axisymmetric geometry, these equations are usually split into two independent systems of unknowns $(E_r, E_\theta, E_z)$ for the first system, and $(B_r, B_\theta, B_z)$ for the second one. If we denote $\bar{E} = (E_r, E_\theta, E_z)$ and $\bar{B} = (B_r, B_\theta, B_z)$, Maxwell’s equations are written

\[
\begin{cases}
\frac{1}{c^2} \frac{\partial E_r}{\partial t} - \text{curl} B_\theta = -\mu_0 J_r, \\
\frac{\partial B_\theta}{\partial z} + \text{curl} E_r = 0, \\
\text{div} E = \frac{1}{\epsilon_0} \rho,
\end{cases}
\quad \text{and} \quad
\begin{cases}
\frac{1}{c^2} \frac{\partial E_\theta}{\partial t} - \text{curl} B_r = -\mu_0 J_\theta, \\
\frac{\partial B_r}{\partial z} + \text{curl} E_\theta = 0, \\
\text{div} B = 0.
\end{cases}
\]

where $\epsilon_0$ and $\mu_0$ are respectively the dielectric permittivity and the magnetic permeability that verify $\epsilon_0 \mu_0 c^2 = 1$.

The axisymmetric operators are defined by

\[
\text{curl} B_\theta = \left( -\frac{\partial B_r}{\partial z} \frac{1}{r} \frac{\partial}{\partial r} (r B_\theta) \right), \quad \text{curl} E_r = \frac{\partial E_\theta}{\partial z} - \frac{\partial E_z}{\partial \theta}, \quad \text{div} E = \frac{1}{r} \frac{\partial}{\partial r} (r E_r) + \frac{\partial E_z}{\partial z}.
\]

The perfectly conducting boundary condition gives here $E_z = E_\theta = 0$ on the surface of the tube $\Gamma$. On the $z$-axis, symmetry considerations imply $E_r = B_\theta = 0$ and $B_\theta = E_\theta = 0$. Artificial boundary conditions are generally required to close the domain. However, they have no influence on this paraxial model, cf. [12].

The charge and the current densities $\rho$ and $J = (J_r, J_\theta, J_z) = (J_r, J_\theta)$ are obtained as the zero and the first moments of the distribution function $f$

\[
\rho = q \int f d\mathbf{P}, \quad J = q \int V \mathbf{P} f d\mathbf{P}.
\]

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. For high energy short beams, a paraxial relativistic model has been derived, cf. [12,3] 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 $v_\theta$ 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 $v_\theta \approx c$ for any particle in the beam, we rewrite the Vlasov–Maxwell equations in the beam frame, which moves along the $z$-axis with the light velocity $c$. Hence we set

\[
\zeta = ct - z, \quad v_c = c - v_\theta.
\]

As a consequence, the bunch of particles is evolving slowly in this frame. We denote by $v$ the transverse characteristic velocity of the particles. Then, introduce a small parameter $\eta$ defined by

\[
\eta = \frac{v}{c} \ll 1.
\]

The paraxial model described in [12,3] is derived by retaining the first four terms in the asymptotic expansion of the distribution function and the electromagnetic fields with respect to $\eta$. Following [12], one can show that the second order asymptotic expansion of
\[ f = f^{(0)} + \eta f^{(1)} + \eta^2 f^{(2)} \]
is entirely determined from the expansion of the electromagnetic force \((F_r, F_\theta, F_z)\) up to the order one:
\[ (F_r, F_\theta, F_z) = (F_r, F_\theta, F_z)^{(0)} + \eta (F_r, F_\theta, F_z)^{(1)}. \]

Now, to determine this force in the axisymmetric case, it is sufficient to know:

1. The principal part \((B_r^{(1)}, B_\theta^{(1)})\) and \((E_r^{(1)}, E_\theta^{(1)})\) of the transverse electromagnetic field.
2. The expansions \(E_z^{(1)} + \eta E_z^{(2)}\) and \(B_z^{(1)} + \eta B_z^{(2)}\) of the longitudinal electromagnetic field up to the order 2.

In [3], a particle-in-cell method is constructed from this asymptotic model, and numerical results illustrate the accuracy of the method. Nevertheless, one can ask to ourselves in which way such a model is more precise than a simpler first order model. In other words, which terms to retain in the asymptotic expansion to get a sufficiently precise, but not too expensive model? To deal with this question, we propose using data mining explorations in asymptotic models, derived in the next section.

2.2. The approximate models \(M_1\) and \(M_2\)

Using the asymptotic expansion described above, one can derive several “nested” approximate models of the Vlasov–Maxwell equations. Indeed, a paraxial model is derived by retaining the first terms in the asymptotic expansion of the distribution function and the electromagnetic fields with respect to \(\eta\). Hence, one can consider the model denoted \(M_1\) in which the asymptotic function \(f\) is approximated by the \(i\)th order expansion \(f^{(0)} + \eta f^{(1)} + \cdots + \eta^i f^{(i)}\).

In this paper, our aim is to illustrate the possibility of data mining techniques applied on scientific computing. Hence we will only consider and compare the 2 first models \(M_1\) and \(M_2\). Let us now expose them.

Following [3,12] one can show that the \(i\)th order asymptotic expansion of \(f\) (here \(i = 1, 2\)) is entirely determined from the knowledge of the \((i – 1)\)th order expansion of the electromagnetic Lorentz force \((F_r^{(i-1)}, F_\theta^{(i-1)}, F_z^{(i-1)})\). One thus obtain the following two models.

1. The model \(M_1\):
   In this model, the asymptotic expansion \(f^{(0)} + \eta f^{(1)}\) is entirely determined from the zero order expansion \((F_r^{(0)}, F_\theta^{(0)}, F_z^{(0)})\) of the electromagnetic force. To compute them, it is sufficient to know the principal part of the transverse electromagnetic fields, which satisfies following [12,3]:
   \[
   \begin{align*}
   E_r^{(1)} &= c B_\theta^{(1)} = \frac{1}{\varepsilon_0 \mu_0} \int_0^r \rho^{(1)} s ds, \\
   E_\theta^{(1)} &= B_z^{(1)}, \\
   E_z^{(1)} &= 0.
   \end{align*}
   \]  
   (6)
   whereas the corresponding forces have the following expression
   \[
   F_r^{(0)} = q v_r^{(1)} B_\theta^{(1)}, \quad F_\theta^{(0)} = 0, \quad F_z^{(0)} = q v_\theta^{(1)} B_z^{(1)}.
   \]  
   (7)
   Note that in this model, the longitudinal fields \(E_z^{(1)}, B_z^{(1)}\) are identically zero.

2. The model \(M_2\):
   We also consider the model \(M_2\), in which the expansion \(f^{(0)} + \eta f^{(1)} + \eta^2 f^{(2)}\) is entirely determined from the first order expansion \((F_r^{(1)}, F_\theta^{(1)}, F_z^{(1)})\) of the electromagnetic force. To characterize them, it is proved [12,3] that the transverse electromagnetic fields have to verified the same equations as (6) for the transverse fields, but at the order 2, namely
   \[
   \begin{align*}
   E_r^{(2)} &= c B_\theta^{(2)} = \frac{1}{\varepsilon_0 \mu_0} \int_0^r \rho^{(2)} s ds, \\
   E_\theta^{(2)} &= B_z^{(2)}, \\
   E_z^{(2)} &= 0.
   \end{align*}
   \]  
   (8)
   supplemented with, for the longitudinal fields:
   \[
   \begin{align*}
   \frac{\partial E_z^{(2)}}{\partial r} &= \frac{\partial B_z^{(2)}}{\partial r}, \quad \text{and} \quad \frac{\partial B_z^{(2)}}{\partial r} = \mu_0 j_z^{(2)}, \\
   E_z^{(2)} (r = R) &= 0.
   \end{align*}
   \]  
   (9)
   Finally, the corresponding forces are expressed:
   \[
   \begin{align*}
   F_r^{(1)} &= q \left( v_r^{(2)} B_z^{(2)} + v_\theta^{(2)} B_\theta^{(2)} \right), \quad F_\theta^{(1)} = -q v_\theta^{(2)} B_z^{(2)}, \quad F_z^{(1)} = q \left( E_z^{(2)} + v_\theta^{(2)} B_\theta^{(2)} \right).
   \end{align*}
   \]  
   (10)
   Since the model is written in a frame which moves along the optical axis at the speed of light, the bunch of particles is evolving slowly in that frame. As a consequence, the computational domain is defined as a simple rectangular domain in variables \((r, \zeta)\), and the above equations are easily solved by a finite-difference method written on a uniform rectangular
Concerning the modeling of the particle dynamic, recall that the paraxial Vlasov equation is numerically solved by means of a particle method: it consists in approximating the function \( f(\mathbf{X}, \mathbf{P}, t) \) at any time \( t \) by a linear combination of delta distributions in the phase space \( (\mathbf{X}, \mathbf{P}) \):

\[
rf(\mathbf{X}, \mathbf{P}, t) = \sum_k w_k \delta(\mathbf{X} - \mathbf{X}_k(t)) \delta(\mathbf{P} - \mathbf{P}_k(t)),
\]

where \( w_k \) denotes the constant weight of the particle \( k \). Its position in the phase space \( \mathbf{X} = (r, \zeta) \) and \( \mathbf{P} = (p_r, p_\zeta, p_z) \), where \( p_r = \gamma v_r, p_\zeta = \gamma v_\zeta, p_z = \gamma (c - v_z) \) is solution to the differential system:

\[
\begin{aligned}
\frac{dr}{dt} &= \frac{1}{\gamma m} p_r, \\
\frac{d\zeta}{dt} &= c - \frac{1}{\gamma m} p_\zeta, \\
\frac{dp_r}{dt} &= \frac{1}{\gamma m} p_r^2 + F_r, \\
\frac{dp_\zeta}{dt} &= -\frac{1}{\gamma m} p_r p_\zeta + F_\zeta, \\
\frac{dp_z}{dt} &= F_z,
\end{aligned}
\]

and initial conditions.

In the next section, our aim is to perform a sensitivity analysis of these two models via data mining techniques; For instance to understand what the second order in the model \( M_2 \) practically brings to the simulation results over what could be obtained by the model \( M_1 \). In such Vlasov Maxwell simulations, one is often interested in the particle motion. For this reason, we will use the particle velocities as significant variables in the data mining analysis. Note that the choice of these variables cannot be automatic: it will always depend on the human expertise that will decide what to be explored in the data. Following [3], we introduce for each model \( M_i \), \( i = 1, 2 \), the variables

\[
\delta v^i_r := |v^i_r - \bar{v}^i_r|,
\]

for the radial velocity and

\[
\delta v^i_z := |v^i_z - \bar{v}^i_z|,
\]

for the longitudinal one, where the index \( \bar{v} \) denotes in each case the average velocity.

3. Data mining and decision trees

3.1. Segmentation by decision tree

Data mining goal is to discover hidden or \textit{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 [11,18].

Decision trees [15] belong to the supervised data mining tools to process the so-called segmentation, whose aim is to constitute homogeneous subgroups inside a given population. For this purpose, we select in a given database, with the aid of the expert of the domain (physician, physicist, economist, etc.) a variable \( y \) to be explained, named the \textit{target variable}. Then, we assume a formal unknown relation \( y = f(x_1, x_2, \ldots, x_n) \) between the target \( y \) and \( n \) other variables \( x_1, x_2, \ldots, x_n \) of the database, called the \textit{predictors}. Basically based on the minimization of the standard deviation of the target variable \( y \), an algorithm of segmentation determines the resulting optimized homogeneous subgroups. This results to a decision tree.

In the case of the segmentation we considered in our study, the target variable is a \textit{categorial} one, \textit{i.e.}, is defined by a finite number of values, that can be numeric or not, like “low” and “high”.

A decision tree is then a tree composed by different subgroups (called \textit{nodes}) of the initial population (called \textit{root node}). At each level of the tree, these \textit{nodes} are obtained by the segmentation algorithm, by identifying among the predictor variables \( (x_1, x_2, \ldots, x_n) \), the most discriminating one, regarding the \textit{homogeneity degree} of the resulting \textit{nodes}. The homogeneity degree of a node (or a subgroup) is usually measured by the standard deviation of the target variable \( y \). For instance for a categorial target variable \( y \) defined by two categories “low” and “high”, It simply represents the percents of individuals which belong to one category.

Each split of the segmentation divides a given node into several nodes, (here, in our study into two nodes which is the specific case of a binary decision trees), based on the most discriminating predictor variable, let us say \( x_j \) for a given \( j \), \( 1 \leq j \leq n \), such that the left resulting node verifies the inequality \( x_j \leq \tau \) and the right one \( x_j > \tau \) (\( \tau \) being a threshold optimally computed by the algorithm of segmentation).

This process stops when the splitting is not feasible: either any new subgroup cannot be found to be more homogeneous than the previous one or the resulting segmentation is composed by insignificant subgroups, typically composed by a two low number of individuals. Hence, in the decision tree, the path from the root node to each terminal node, called a \textit{leaf}, defines a succession of inequalities on the predictor variables that characterize the solutions belonging to the leaf with a certain accuracy, which depends on the percentage of misclassified solutions in the leaf.

By choosing the most inequalities leaves, one is able to characterize the different categories of the predictor variable with a set of rules, composed by a succession of inequalities, with the best accuracy.
3.2. A basic example of data mining

This subsection is devoted to a basic application of data mining to an everyday life problem, to help non-specialists in data mining. We consider a fictive case based on a medical application. The collected data are a set of patients suffered from the same illness. During their course of treatment, each patient responded to one of five medications denoted Drug A, Drug B, Drug C, Drug X and Drug Y. Our purpose is to show how data mining methods, especially decision trees, can help us to find out which drug might be appropriate for a future patient with the same illness.

Let us list the available variables in the database: Age, Sex, Blood pressure (BP), Cholesterol, Blood sodium concentration (Na), Blood potassium concentration (K), Prescription drug to which a patient responded (Drug).

Clearly, the target variable here is the prescribed drug, as one is interested to determine the efficiency of each drug category on each patient. Let us have a look at what factors might influence drug. Following the expert’s advise, here the physician, the concentrations of sodium and potassium in the blood are important factors. So, we can create a scatterplot of sodium versus potassium, using the drug categories as a color overlay, (see Fig. 1).

This plot clearly shows a threshold above which the correct drug is always drug Y and below which the correct drug is never drug Y. This threshold is a ratio, namely between the sodium (Na) and the potassium (K). Since the ratio of sodium to potassium seems to predict when to use drug Y, we introduce it as a new variable. This variable might be useful later when one builds a model to predict when to use each of the five drugs.

On the other hand, the ability to fully explain how the ratio of sodium to potassium in the blood affects the choice of drug is not possible at this stage of our exploration. Indeed, one cannot determine for which level of this ratio such drug is more efficient than another one. This is where data modeling will likely provide some answers. To achieve this objective, we will try to fit the data using a segmentation model based on a decision tree (see Fig. 2).

To conclude this basic example, one has now to choose the most homogeneous leaves in the decision tree, to characterize the different categories of the predictor variable with a succession of inequalities. From Fig. 2, one readily obtains that people with an Na-to-K ratio less than 14.64 and high blood pressure (described by node 3), age determines the choice of drug (nodes 6 and 7). For people with low blood pressure (node 4), cholesterol level seems to be the best predictor (nodes 8 and 9).

3.3. Application to our data

Based on the methodology illustrated in the previous section, we will derive now decision tree analysis for numerical computing applications.

In the database we considered, the data are computed by the help of finite differences method and described numerical approximations of problems (6)–(10) solutions. Then, at each time step and for each node of the concerned space grid, we get a set of variables which are:

\[
v^{(i)}(t), v_z^{(i)}(t), v_{\phi}^{(i)}(t), E_r^{(i)}, E_z^{(i)}, B_{\phi}^{(i)}, B_z^{(i)}, \rho^{(i)}, f_{\phi z}^{(i)}, f_{\phi z}^{(i-1)}, F_r^{(i)}, F_z^{(i)}, F_{\phi z}^{(i)}, \delta^2 v_r^{(i)}, \delta^2 v_z^{(i)} \quad (i = 1, 2).
\]

Therefore, we organize the data such that each row of the database (or “individual”, the devoted terminology in database language) contains the information of the above variables for a given time step and for a space node. In the simulation case we have run, the number of time steps \(N_t = 100\), whereas the number of grid points \(N_r \times N_z\) was equal to 1250.
Considering all the 100 time steps and the 1250 space nodes, the database we treated was composed by 125,000 rows and 26 variables given by (13) to be analyzed.

Because our objectives are to appreciate the improvement of the results depending on the order of the asymptotic development of problems (6)–(10) solutions, we introduce the two following variables to define an appropriate target variable:

Let us denote by $X$ a given variable to be computed by the two asymptotic models $M_1$, $M_2$. We set $X^{(1)}$ its value computed by the model $M_1$ and $X^{(2)}$ its corresponding value from the model $M_2$. As we are basically interested with the relative order of magnitude of $X^{(1)}$ w.r.t. $X^{(2)}$, it is natural to consider the ratio of the variable $X^{(1)}$ by the variable $X^{(2)}$. Moreover, without any a priori on the data, there is no reason to use another transformed target variable.

Then, the first variable we consider here, $x_{1,2}$, is defined by:

$$x_{1,2} = \frac{X^{(1)}}{X^{(2)}}.$$  \hspace{1cm} (14)

It measures the weight of the model $M_1$ in the model $M_2$, regarding the variable $X$.

More precisely, one has to consider $\omega_{1,2}$ in the two following main cases: The first one is when $\omega_{1,2}$ is around 1 which corresponds to an equivalence of numerical results obtained between the two models $M_1$ and $M_2$ for the calculation of $X$. The second case describes the situation when the numerical results between $M_1$ and $M_2$ are significantly different. This case is available when $\omega_{1,2}$ is either very small or very great compared to 1.

Without any a priori on the meaning of low or high contributions of the model $M_1$ in the model $M_2$, it is usual to define the categorial variables $\omega_{1,2}$ as follows: the three classes of individuals are determined based on an equal number of individuals for each category, (low, medium and high).

Without any a priori on the meaning of low or high contributions of the model $M_1$ in the model $M_2$, it is usual to define the categorial variables $\omega_{1,2}$ as follows: the three classes of individuals are determined based on an equal number of individuals for each category, (low, medium and high).

Then, because our objectives are to identify which variables could be predictors of a significant difference in the evaluation of the variable $X^{(1)}$ and $X^{(2)}$, we only keep in our analysis the extreme groups defined by the “low” class and the “high” class of the categorial target variable to be explained, namely, $\omega_{1,2}^{-}$. 
Otherwise, the purpose of our analysis being to point out the role of the electromagnetic fields in the sensitivity between the models $M_1$ and $M_2$, the dependent variables – that is to say the predictors – we kept to explain the above two classes are the non-vanishing electromagnetic components, the charge and current densities and the particles velocities computed by the model $M_2$, namely

$$\nu_i^{(2)}, \phi_i^{(2)}, E_i^{(2)}, B_i^{(2)}, \rho_i^{(2)}, j_i^{(2)}.$$ \hfill (15)

As a complement to take into account the coupling with the Vlasov equation, we also add to the above list of predictors the components of the Lorentz force involved in the model $M_2$, that is

$$F_r^{(1)}, \ F_\phi^{(1)}, \ F_z^{(1)}.$$ \hfill (16)

Note that the other available variables of the database could be considered as predictors or targets in further developments, depending on the objectives of the exploration.

4. Result: comparison between the model $M_1$ and $M_2$

The question is to understand, at least in a formal way from the equations, what are the variables which bring a significant contribution from the model $M_1$ to the model $M_2$, regarding first the variable $\delta \nu_r$, then the variable $\delta \nu_z$.

4.1. The case of $\delta \nu_r$

4.1.1. The formal approach

In this paragraph the general variable $X$ defined in section 3 is chosen equal to $\delta \nu_r$. Then, first of all, we plot the distribution of $\omega_{1,2}$ defined here by: $\omega_{1,2} = \delta \nu_i^{(1)}/\delta \nu_i^{(2)}$.

As one can see (Fig. 3), this distribution is very gathered around the value 1 of $\delta \nu_i^{(1)}/\delta \nu_i^{(2)}$.

To understand this phenomena, we will formally analyze the relation between $\delta \nu_i^{(1)}$ and $\delta \nu_i^{(2)}$ with respect to all the potential predictors. Then, we will confirm by statistical and data mining tools the resulting features we will get.

From Eq. (12), one readily sees that the radial velocity $\nu_r$ basically depends on the radial force $F_r$ and on the square of the azimuthal velocity $\nu_{z,\text{ave}}^2$. Recall also the definition of $\delta \nu_i^{(0)}$ for each model $M_i$, that is:

$$\delta \nu_i^{(0)} = \gamma (\nu_i^{(0)} - \nu_{i,\text{ave}}^{(0)}),$$

where the index $\text{ave}$ denotes the average velocity. Hence we can write in a formal way that the variable $\delta \nu_i^{(0)}$ is a function $\varphi$ of $F_i^{(1)}$ and $(\nu_i^{(0)})^2$ (the shift $-1$ of the superscript is due to the asymptotic expansion, see Section 2.1), that we denote by:

$$\delta \nu_i^{(0)} = \varphi (F_i^{(1)}, (\nu_i^{(0)})^2) \quad (i = 1, 2).$$

Let us use now the expressions of $F_i^{(1)}$, for $i = 1, 2$. We have:

\footnote{In our case, the factor $\gamma m$ is quasi-constant. Anyway, if it not the case, one has to introduce $\delta \rho_i^{(0)}$ instead of $\delta \nu_i^{(0)}$. Moreover, $\nu_{i,\text{ave}}^{(0)}$ as an average velocity is a constant for each time step of the simulation.}
Recall that each model $M_i$ was derived by retaining the $i$ first terms in the asymptotic expansion of the distribution function, the electromagnetic fields and forces, with respect to the small parameter $r$. As a consequence, one can consider that

$$
\begin{align*}
F_r^{(0)} &= q v_r^{(1)} B_{r,0}, \\
F_r^{(1)} &= q v_r^{(2)} B_{r,0} + q v_r^{(2)} B_{r,0}.
\end{align*}
$$

where $q v_r^{(0)}$ denotes the difference of evaluation for $B_s$ (resp: $B_t$), between the models $M_1$ and $M_2$.

Substituting the above expressions into $F_r^{(1)}$ and retaining only the “zero order” terms, one can approximate the force $F_r^{(1)}$ by

$$
F_r^{(1)} \approx F_r^{(0)} + q v_r^{(2)} B_{r,2},
$$

and we get that

$$
\delta F_r^{(2)} := \varphi \left(F_r^{(0)}, (v_r^{(2)})_1\right) \text{ can be approximated by}
$$

$$
\delta F_r^{(2)} \approx \varphi \left(F_r^{(0)} + q v_r^{(2)} B_{r,2}, (v_r^{(2)})_1\right).
$$

One can carry on these approximations in a formal way. Assuming that we only consider the linear part of the functional $\varphi$ as a first level of formal approximation, this yields

$$
\delta v_r^{(2)} \approx \varphi \left(F_r^{(0)}, (v_r^{(2)})_1\right) + \varphi \left(q v_r^{(2)} B_{r,2}, (v_r^{(2)})_1\right).
$$

In addition, one uses now the same kind of approximation for the velocity $v_r$,

$$
v_r^{(1)} = v_r^{(0)} + \Delta v_r^{(1)}.
$$

By substituting it in the first term of the above expression, and retaining only the “zero order” term, we finally obtain

$$
\delta v_r^{(2)} \approx \varphi \left(F_r^{(0)}, (v_r^{(2)})_1\right) + \varphi \left(q v_r^{(2)} B_{r,2}, (v_r^{(2)})_1\right) \approx \delta v_r^{(1)} + \varphi \left(q v_r^{(2)} B_{r,2}, (v_r^{(2)})_1\right).
$$

This shows that if the variables $v_r^{(0)} B_{r,2}$ and $(v_r^{(2)})_1$ do not bring a significant contribution, there is almost no difference between $\delta v_r^{(1)}$ and $\delta v_r^{(2)}$. As a consequence it will also explains why the distribution $\delta v_r^{(1)}/\delta v_r^{(2)}$ is gathered around 1.

### 4.1.2. Data mining explorations by decision tree and validation by statistical tests

The purpose of this section is to check the formal result of the above section by the help of data mining and statistical tools. Basically, we use decision trees to identify the main features of two kinds of individuals in our population, the “low” and the “high” classes we considered. To that aim, we take into account the results of the formal analysis presented in Section 4.1.1., that helps us to complete the list of potential predictors. For this reason, we also considered the quantity $v_r^{(2)} B_{r,2}$ as a predictor variable, together with the other variables described in (15) and (16). Without any complementary information one might only considered the native numerical variables corresponding to the computed approximations, as the potential predictors.

To identify the role, if any, of the elements such that $\omega_{1,2}$ is significantly far from 1, we will focus our analysis on the individuals which belong to the low class and the high class defined by $\omega_{1,2}^{(3,13)}$. In addition, having eliminated the medium group of $\omega_{1,2}^{(3,13)}$, this will lead to a better accuracy level for the resulting segmentation.

Our interest is to evaluate the differences, if any, between the distributions of the dependent variables which are suspected to be different between the models $M_1$ and $M_2$. For this purpose, we performed a decision tree on the target variable $\omega_{1,2}^{(3,13)}$, whose purpose is to identify subgroups in the database which are homogenous to the category “low” and “high”.

First of all, let us consider the decision tree (see Fig. 4) we got under IBM SPSS Modeler to model the two classes of the variable $\omega_{1,2}^{(3,13)}$, says the “low” and the “high” classes, in relation with the potential predictors listed above. Recall that the exploration was processed on the whole available time steps and for all the space nodes of the considered grid.

The precision of the decision tree, that evaluates the tree accuracy, is computed by the risk estimate which is equal to 6.05 percents; the risk estimate tells the chances of misclassification by the decision tree. Here, for a binomial modeling prediction, the risk estimate describes the proportion of cases incorrectly classified by the tree. So, if the risk estimate is equal to 6.05 percents, it means that 93.95 percents of data are correctly classified by the model of segmentation computed by the decision tree. Accordingly, the quality level of the decision tree is very high and its reading is as follows.

As one can see (Fig. 4), the first segmentation which appears on the decision tree shows the most discriminating predictor variable in the set of all the available potential predictors. Hence, $E_2^{(2)}$ appears as this predictor with a corresponding computed threshold equal to ~0.006. This means that the group of the “low” $\omega_{1,2}$ is mainly different from the group of the “high” $\omega_{1,2}$, if one splits the whole involved population of the database up to the found threshold of $E_2^{(2)}$.

Because of the meaning of the two classes of the variable $\omega_{1,2}^{(3,13)}$, we considered, a first practical conclusion is that the computation of $\delta v_r$ is significantly different between the two models $M_1$ and $M_2$, primarily due to the presence of the variable...
$E^{(2)}_z$ in the asymptotic model $M_2$. This result is expected since the $E_z$ component is not present in the model $M_1$, we mean $E^{(1)}_z = 0$.

On the other hand, at the first level of the root of the decision tree, i.e., at the top of the tree, a relevant tool is a whole classification of the predictors (see Table 1) from the most discriminant to the less one, to distinguish the low class to the high one. In this classified list of predictors appears after the first variable $E^{(2)}_z$, the component $E^{(2)}_z$ as the second more significant predictor to explain the numerical difference of the asymptotic development computed by the model $M_2$ versus $M_1$.

This feature is unexpected since the corresponding component $E^{(1)}_z$ is non-zero in the model $M_1$. Because of the strong nonlinearity of the partial differential system, this was not a result that could be expected before this exploration.

Furthermore, and at the same stage of the decision tree, says the root, the classification of the predictors also pointed out that the variable $B^{(2)}_z$ does not appear as a predictor which explains a significant difference between the two asymptotic modelings.

---

**Table 1**

Hierarchy of the $E^{(2)}_{1,2}$ predictors for $\partial v_r$.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Improvement scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E^{(2)}_z$</td>
<td>0.144</td>
</tr>
<tr>
<td>$F^{(1)}_z$</td>
<td>0.144</td>
</tr>
<tr>
<td>$(v^{(2)}_h)^2/(v^{(1)}_h)^2$</td>
<td>0.112</td>
</tr>
<tr>
<td>$E^{(2)}_z$</td>
<td>0.092</td>
</tr>
<tr>
<td>$F^{(1)}_z$</td>
<td>0.092</td>
</tr>
<tr>
<td>$\rho^{(2)}$</td>
<td>0.063</td>
</tr>
<tr>
<td>$f^{(2)}_z$</td>
<td>0.055</td>
</tr>
<tr>
<td>$B^{(2)}_z$</td>
<td>0.028</td>
</tr>
<tr>
<td>$F^{(1)}_z$</td>
<td>0.028</td>
</tr>
<tr>
<td>$v^{(2)}_h B^{(2)}_z$</td>
<td>0.027</td>
</tr>
</tbody>
</table>

---

Fig. 4. Decision tree related to $\partial v_r$. 
In contrast to the component $E^{(2)}$ mentioned above, we recall here that $B^{(1)}$ was null in the model $M_1$. Nevertheless, it does not bring a significant contribution in the second order model $M_2$, while this might be intuitively expected.

After these first conclusions, let us identify the nodes corresponding to the most homogeneous one relatively to the low and high classes defined by $\omega^{(3|2)}_{1,2}$. The exploration of the decision tree from the root to the end nodes, namely the leaves, shows the following results.

It appears (see Fig. 4) that node 5 corresponds to the most homogenous regarding the “low class” and that node 10 is the most characteristic of the “high class”. Therefore, these two groups are defined as the most likely ones to characterize the “individuals” in the database which correspond, on the first hand, to those such that $\Delta \nu^{(2)}_1$ is significantly greater than $\Delta \nu^{(2)}_2$, (we mean the “low class”), and on the other hand, at the opposite, those such that $\Delta \nu^{(2)}_2$ is significantly less than $\Delta \nu^{(2)}_1$, (we mean the “high class”).

Let us describe now these two groups by the help of standard statistical tools, to highlight and to classify their main features. We processed two statistical tests to identify significant differences for each predictor computed on the two nodes 5 and 10, mentioned above.

The first test we considered is the $t$-test, whose purpose is to identify significant differences of the mean of a given variable computed in different groups. This test is standard for predictors which follows normal distributions. But, it is also available for other more general categories of distributions, regarding the IBM SPSS recommendation: “This test assumes that the data are normally distributed; however, this test is fairly robust to departures from normality.” [17].

The second test we processed is the Mann–Whitney test, whose purpose is approximately the same as the $t$-test, but which is a non-parametric one. Consequently, it does not assume any hypothesis of normality for the variables to be tested. As one can see on Tables 2 and 3, both tests lead to the same conclusions; significant differences between the two groups composed by the nodes 5 and 10 are detected for the following predictors: $\nu^{(2)}_1$, $\nu^{(2)}_2$, $E^{(2)}_1$, $E^{(2)}_2$, $r^{(2)}$, $p^{(2)}$, $J^{(2)}$, $F^{(2)}_1$, $F^{(2)}_2$.

On the other hand, the variables $\nu^{(2)}_v$ and $B^{(2)}_v$ do not play a discriminant role to distinguish the low class to the high one. Then, accordingly to formula (17) in section 4.1.1, $\Delta \nu^{(2)}_v$ does not significantly differs from $\Delta \nu^{(1)}_r$ because neither $\nu^{(2)}_v B^{(2)}_v$ nor $(\nu^{(2)}_v)^2$ appears as a major predictor to explain the two relevant classes of the target variable $\omega^{(3|2)}_{1,2}$.

This result is confirmed by a complementary analysis which consists to classify all the predictors, from the most significant one to the least one, regarding the target variable $\omega^{(3|2)}_{1,2}$ limited to the two groups, the low class and the high class. This has been done by computing the non-parametric Spearman's correlation coefficients.

Then, one can observe (see Table 4) the following rating between the potential predictors: The most correlated one with the target variable $\omega^{(3|2)}_{1,2}$ is the component $E^{(2)}_2$, (or equivalently $F^{(2)}_2$) due the explicit relation between $E^{(2)}_2$ and $F^{(2)}_2$, and the fact that $\nu^{(2)}_v B^{(2)}_v$ is not significant, see Eq. (10) and Table 1), with a correlation coefficient $\rho (-1 \leq \rho \leq 1)$ equal to 0.858 associated to a $p$-value less than $10^{-5}$. Here, the meaning of the $p$-value corresponds to the degree of confidence that a relationship exists between 2 variables. In our case, $p$-value less than $10^{-5}$ means that we are more than 99.99% confident that a

| Table 2 |
|------------------|------------------|
| $t$-test applied to nodes 5 and 10 of decision tree related to $\nu_v$. |
| Levene's test for equality of variances | $t$-test for equality of means |
| $F$ | Sig. | $t$ | df | Sig. (2-tailed) |
| $\nu^{(2)}_v$ | Equal variances | 0.186 | 0.666 | 17.804 | 1340 | 0.000 |
| | Different variances | 17.748 | 1229.7 | 0.000 |
| $\nu^{(2)}_v$ | Equal variances | 14.584 | 0.000 | -0.279 | 1340 | 0.781 |
| | Different variances | -0.285 | 1327.273 | 0.775 |
| $\nu^{(2)}_v$ | Equal variances | 724.708 | 0.000 | 34.184 | 1340 | 0.000 |
| | Different variances | 38.833 | 839.97 | 0.000 |
| $E^{(2)}_1$ | Equal variances | 7785.76 | 0.000 | -23.457 | 1340 | 0.000 |
| | Different variances | -26.817 | 792.520 | 0.000 |
| $E^{(2)}_2$ | Equal variances | 1525.12 | 0.000 | 34.472 | 1340 | 0.000 |
| | Different variances | 39.575 | 762.226 | 0.000 |
| $B^{(2)}_v$ | Equal variances | 3.180 | 0.075 | -0.972 | 1340 | 0.331 |
| | Different variances | -0.982 | 1288.971 | 0.326 |
| $\rho^{(2)}$ | Equal variances | 47.217 | 0.000 | -4.911 | 1340 | 0.000 |
| | Different variances | -5.053 | 1334.256 | 0.000 |
| $J^{(2)}$ | Equal variances | 41.160 | 0.000 | -4.085 | 1340 | 0.000 |
| | Different variances | -4.203 | 1334.466 | 0.000 |
| $F^{(1)}$ | Equal variances | 7785.759 | 0.000 | 23.457 | 1340 | 0.000 |
| | Different variances | 26.817 | 792.520 | 0.000 |
| $F^{(1)}$ | Equal variances | 3.18 | 0.075 | -0.972 | 1340 | 0.331 |
| | Different variances | -0.982 | 1288.971 | 0.326 |
| $F^{(1)}$ | Equal variances | 1525.794 | 0.000 | -34.432 | 1340 | 0.000 |
| | Different variances | -39.528 | 762.226 | 0.000 |
| $\nu^{(2)}_v B^{(2)}_v$ | Equal variances | 3.331 | 0.068 | 0.972 | 1340 | 0.331 |
| | Different variances | 0.982 | 1289.374 | 0.326 |
| $(\nu^{(2)}_v)^2$ | Equal variances | 14.632 | 0.000 | 0.260 | 1340 | 0.795 |
| | Different variances | 0.266 | 1327.119 | 0.790 |
strong relationship exists between $E^{(2)}$ and $\omega^{(3)\text{CLS}}$. Let us recall that this result was also identified by the decision tree algorithm (the predictor which appears at the first segmentation in the decision tree).

The second predictor in the rating list is $E^{(2)}$, (or $F^{(1)}$), also with a reliable computation in relation with the $p$-value which is also less than $10^{-5}$. This was another result we found by the decision tree.

Finally, let us mention that both predictors $B^{(2)}_2$ and $\nu^{(2)}_2$ are not correlated with $\omega^{(3)\text{CLS}}$. This can be observed from the respective Spearman’s correlation coefficient which is equal to 0.032 for the predictor $B^{(2)}_2$ and −0.042 for $\nu^{(2)}_2$.

Let us also remark that the two corresponding $p$-values are equal to 0.23 for $B^{(2)}_2$ and 0.12 for $\nu^{(2)}_2$. This means that we cannot be 95% confident that a relationship exists between $B^{(2)}_2$ and $\omega^{(3)\text{CLS}}$ on the first hand, and $B^{(2)}_2$ and $\nu^{(2)}_2$ on the other hand.

So, regarding this complement of exploration, we conclude again that neither $B^{(2)}_2$ nor $\nu^{(2)}_2$ plays a significant role within the computation of $\delta \nu$, either from the models $\mathcal{M}_2$ or $\mathcal{M}_1$. Once more, due to formula (17), $\delta \nu$, has approximatively the same value if one computes it either with the model $\mathcal{M}_1$ or $\mathcal{M}_2$.

This makes sense and explains why the distribution of $\frac{\delta \nu^{(1)}_i}{\delta \nu^{(2)}_i}$ is concentrated around 1, (see Fig. 3), as it was suspected by the formal analysis we presented above.
4.2. The case of $dv_z$

4.2.1. The formal approach

As we proceeded for the analysis of $dv_r$, we set now the variable $X$ to be equal to $dv_z$ and we plot the resulting distribution of $\delta v_z^{(1)} / \delta v_z^{(2)}$, (see Fig. 5). Here, unlike the distribution of $\delta v_r^{(1)} / \delta v_r^{(2)}$, it is gathered around the 0-value of $\delta v_z^{(1)} / \delta v_z^{(2)}$.

Again from Eq. (12), it is clear that the longitudinal velocity $v_z$ depends only on the longitudinal force $F_z$. Using the definition of $\delta v_z^{(i)}$ for each model $M_i$:

$$\delta v_z^{(i)} = \frac{c v_z^{(i)}}{C_0} / \Delta v_z^{(i)};$$

we can still express the variable $\delta v_z^{(i)}$ is a function $\psi$ of $F_z^{(i-1)}$:

$$\delta v_z^{(i)} = \psi(F_z^{(i-1)}) \quad (i = 1, 2).$$

We exploit now the expressions of $F_z^{(i-1)}$ for $i = 1, 2$. We have

$$\begin{align*}
F_z^{(0)} &= q v_z^{(1)} B_0^{(1)}, \\
F_z^{(1)} &= q v_z^{(2)} B_0^{(2)} + q E_z^{(2)}.
\end{align*}$$

Again, each model $M_i$ being derived by retaining the $i$ first terms in the asymptotic expansion, one can consider that

$$\begin{align*}
B_0^{(2)} &= B_0^{(1)} + \Delta B_0^{(1)}, \\
v_z^{(2)} &= v_z^{(1)} + \Delta v_z^{(1)}.
\end{align*}$$

Substituting these expressions into $F_z^{(1)}$ and retaining only the “zero order” terms, one can approximate the force $F_z^{(1)}$ by

$$F_z^{(1)} \simeq F_z^{(0)} + q E_z^{(2)}$$

and we get that $\delta v_z^{(2)} := \psi(F_z^{(1)})$ can be approximated by

$$\delta v_z^{(2)} \simeq \psi(F_z^{(0)} + q E_z^{(2)}).$$

Let us still can carry on this approximation. Again, we only consider the linear part of the functional $\psi$ as a first level of formal approximation to obtain:

$$\delta v_z^{(2)} \simeq \psi(F_z^{(0)} + q E_z^{(2)}) + \psi(q E_z^{(2)}) \simeq \delta v_z^{(1)} + \psi(q E_z^{(2)}).$$

(18)

This would explains why, if the variable $E_z^{(2)}$ brings a significant contribution, there is a real difference between $\delta v_z^{(1)}$ and $\delta v_z^{(2)}$, or, in other words, why the distribution of $\frac{\delta v_z^{(1)}}{\delta v_z^{(2)}}$ is mainly gathered around 0.

\footnote{The factor $\gamma m$ is quasi-constant.}
Table 5

$t$-test applied to nodes 2 and 11 of decision tree related to $\delta v_z$.

<table>
<thead>
<tr>
<th></th>
<th>Levene's test for equality of variances</th>
<th>$t$-test for equality of means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>Sig</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>0.852</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>35.161</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>389.041</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>3362.368</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>29.002</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>239.052</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>26.953</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>66.891</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>3362.368</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>239.052</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>28.993</td>
</tr>
<tr>
<td>$E_z^{(2)}$</td>
<td>Equal variances</td>
<td>3362.364</td>
</tr>
</tbody>
</table>
4.2.2. Data mining exploration by decision tree and validation by statistical tests

Here, the precision of the decision tree given by the risk estimate is equal to 5.2 percent. So, 94.8 percent of data are correctly classified by the model of segmentation computed by the corresponding decision tree. One more time, the quality level of the decision tree is very high and it let us to use it with a high level of confidence.

The first segmentation which appears on the decision tree (Fig. 6) highlights the most discriminated predictor variable in the set of all the available potential predictors. More precisely, \( F^{(1)}_x \) is detected as this predictor with a corresponding computed optimal threshold equal to 37.24.

Based on the information available within the decision tree dedicated to \( \hat{\nu}_x \), we get similar conclusions as we mentioned for the decision tree relatively for \( \hat{\nu}_y \), namely:

- \( F^{(1)}_x \) is the most discriminate variable. This was an expected result, since \( E^{(1)}_x = 0 \) in the model \( M_1 \), which implies that the main difference between \( E^{(1)}_x \) and \( E^{(0)}_x \) is essentially due to \( E^{(2)}_x \) (see Eq. (10)).
- The second most important predictor identified at the root of the decision tree is \( E^{(2)}_x \) which an unexpected feature since the corresponding component \( E^{(1)}_x \) is non-zero.
- On the contrary, \( B^{(2)}_x \) appears as a non-significant predictor even if \( B^{(1)}_x \) was null in the model \( M_1 \).

Let us now identify in the decision tree the nodes which corresponds to the most homogeneous relatively to the low and high classes defined by \( \omega^{(3x12)}_y \), with respect to \( \hat{\nu}^{(1)}_y \) and \( \hat{\nu}^{(2)}_y \). The exploration of the decision tree (see Fig. 6) allows us to isolate the two nodes 2 and 11 as the most homogenous ones, regarding the “low class” and the “high class” respectively.
As in the above analysis dedicated to the computation of \( \Delta v_z \) by each model, these two groups characterized the most likely ones either for the individuals such that \( \Delta v_z^{(2)} \) is significantly less than \( \Delta v_z^{(1)} \) or otherwise.

Then, we are in position to describe these two groups by identifying the predictors which corresponds to significant differences between them. Tables 5 and 6 together lead to the same conclusions: The sought predictors are given by the whole list of potential predictors. Thereby, \( \Delta v_z^{(2)} \) significatively differs from \( \Delta v_z^{(1)} \) because, according to formula (18) in section 4.2.1, \( E_z^{(2)} \) will produce a relevant contribution. This explains why the distribution of \( \frac{\Delta v_z^{(1)}}{\Delta v_z^{(2)}} \) is grouped around the 0-value as it was predicted by the formal analysis we detailed above.

Another time, this result is confirmed by the computation of the non parametric Spearman’s correlation coefficients.

Indeed, one can observe (see Table 7) the following rating list between the potential predictors: The most correlated one with the target variable \( \omega_{1,2}^{(2)} \) is \( E_z^{(2)} \), with a correlation strength estimated to 0.86 and associated with a p-value less than \( 10^{-3} \), which leads us to consider the correlation evaluated by Spearman’s \( \rho \) as a reliable one.

The above results both confirm and strengthen our formal approach described in subsection 4.1.1. Then, we conclude that the computation of \( \Delta v_z \) significantly differs in the two models \( M_1 \) and \( M_2 \). For this reason, the distribution of \( \frac{\Delta v_z^{(1)}}{\Delta v_z^{(2)}} \) is gathered around its 0-value. This means that the model \( M_2 \) increases the value of \( \Delta v_z \) computed within the model \( M_1 \), since we considered absolute values in our analysis of \( \Delta v_z \).

5. Conclusion

In this paper, we have presented a new approach based on data mining techniques and statistical tools applied to scientific computing. We focused our study to the specific case of an asymptotic paraxial approximation to model ultrarelativistic particles. Our aim was to determine the role of the different powers in the asymptotic expansion, restricted to the models \( M_1 \) and \( M_2 \). As we have considered an approximate model of the Vlasov–Maxwell equations, we have chosen \( \Delta v_z^{(0)} \) and \( \Delta v_z^{(1)} \) (i = 1, 2), as the two main physical variables.

For \( \Delta v_z \), the analysis of the decision tree shows that the most important predictor is \( E_z^{(2)} \). On the other hand, \( E_z^{(2)} \) does not bring any significant contribution to distinguish the value of \( \Delta v_z \) between the two models \( M_1 \) and \( M_2 \). The two most homogeneous nodes in the decision tree allows us to confirm why the distribution \( \omega_{1,2} \) is gathered around one, as suspected by the formal analysis. Namely, neither \( E_z^{(2)} \) nor \( v_z^{(1)} \) does not play a significant role for the computation of \( \Delta v_z \) in the two models \( M_1 \) and \( M_2 \).

Concerning \( \Delta v_z \), the analysis of the decision tree shows here that the most important predictor is \( F_z^{(3)} \). Similarly to the case of \( \Delta v_z \), \( B_z^{(2)} \) is not a relevant predictor for \( \Delta v_z \). Again, the identification of the two most homogeneous nodes in the decision tree confirms that the distribution \( \omega_{1,2} \), in that case, is gathered around its zero value, as expected by the formal analysis. Indeed, \( F_z^{(3)} \) has a strong correlation with \( \omega_{1,2}^{(2)} \).

Despite the mathematical analysis which states that an asymptotic model of high order (here \( M_2 \)) is globally more accurate than a lower order one (here \( M_1 \)), our take-home message is first to show that this improvement is not true, uniformly for each variable of the model. Then, we show how data mining tools allow us to identify which variables are actually improved from one model to the other.

Beyond the particular case we treated in this paper, we suggest that data mining techniques can be applied to the analysis of any scientific computations as it is applied in a lot of other domains. Indeed, data mining would help to investigate the relevance and/or the quality of numerical simulations, particularly when a large quantity of data – as for instance with massively parallel computers – is available.

Nevertheless, several problems remain open: up to now, this approach remains heuristic and mathematical analysis to assess the method would be useful. In addition, data mining tools will always involve the human expertise to guide the analysis and to explore the data.

References