Annals of the Academy of Romanian Scientists
Series on Engineering Sciences
ISSN 2066-8570 Volume **13**, Number **1**/**2021**

MANAGEMENT OF BIG DATA IN SCIENTIFIC COMPUTING

Marcel ILIE[1], Augustin SEMENESCU[2]

Rezumat. *Dezvoltarile recente in sistemul de procesare al calculatoarelor au condus la generarea unor cantitati mare de date. Procesarea acestor large baze de data reprezinta adevarate provocari pentru comunintatea stiintifica. O problema particulara este interpretarea datelor, in particular seprarea datelor utile de cele inutile. Mai mult decat atat clasificarea datelor in subgrupuri de date aduce provocari suplimentare. In general este acceptat ca analiza acestor date este foarte dificila. De aceea modele eficiente si acurate de data-mining ar inlesni analiza acestor mari baze de date. In acest studiu propunem un odel computational basat pe algoritmul k-means, identificat ca algoritmul de fuzzy-clustering. Folosirea algoritmului de fuzzy-clustering reduce timpul computational cu 72% comparat cu metodele computationale normale.*

Abstract. *The recent developments in computer processing has led to the generation of significant amount of data. However, the post-processing of this large amount of data poses significant challenges for the scientific community. A particular issues is the data interpretation, particularly regarding the segregation of valuable data from arbitrary data. Moreover, data classification in subgroups poses further challenges. It has been largely accepted that the analysis of these data sets may be cumbersome. Therefore, efficient and accurate data mining models would enable the analysis if large data-sets. In this research we propose a computational model based on the k-means algorithm, identified as fuzzy clustering algorithm. The use of the fuzzy clustering algorithm reduces the computational time by 72%, when compared with regular computational approaches.*

Keywords: data management, numerical modeling, adaptive mesh refinement, k-means algorithm

1. Introduction

The rapid progress in the developments of computing technologies and software such as high-performance computing (HPC) have generated large amounts of data that require further post-processing, interpretation and dissemination [1-5]. Post-processing and visualization of this large amount of data pose significant challenges, particularly when the data needs to be assembled/coupled from various instantaneous time-frames [5, 6]. Some of these cases are encountered in the medical field, biology research, engineering applications, etc. Out of all these

[1]PhD, Assistant Professor, Dept. of Mechanical Engineering, Georgia Southern University, Statesboro, GA 30458, USA, e-mail: milie@georgiasouthern.edu
[2]PhD, Professor, Dept. of Material Sciences, University Politehnica Bucharest, Bucharest, Romania, augustin.semenescu@upb.ro

fields, we focus on engineering applications used in the fluid dynamics analyses, such as particle image velocimetry (PIV). PIV is a flow visualization technique where visual data is collected at different instants in time and post-processed using correlation techniques. It is well know that in both experimental and computational data there is always data uncertainty. The data uncertainty may be reduced by proper data management, which implies, accurate data separation and grouping. Similarly to the experimental data management, the computational research also poses significant challenges with the data allocation. In this research we focus on the computational data resulting from computational fluid dynamics (CFD). In many cases the generated CFD data may or may not represent the accurate solution due to the space and time-discretization schemes employed. The space discretization is defined the size of the grid element, in the sense that the smaller the element grid size the smaller the numerical error. Apparently, a very small element size would minimize the numerical error and ensure the numerical stability. However, there is a drawback associated with approach, namely a larger amount of data than need and thus, post-processing challenges. A method that ensures both numerical accuracy of data and stability would be desirable.

2. Background

In the past decade there have been multiple research studies that concern the development of accurate and efficient computational techniques for data management in CFD [1, 2, and 6]. One of these methods is the adaptive-mesh refinement (AMR). The AMR numerical technique offers the advantage of fast and reliable computations with a minimum number of grid point (data recording points). However, most of these studies focused on a single level of refinement. Thus, we propose a method that is based on multiple levels of refinement with an aposteori data separation technique. The proposed numerical technique represents a novelty in the field of computational sciences.

3. Modeling

Figures 1and 2 show the schematic of the AMR algorithm based on multiple refinements, for both structured, Fig. 1, and unstructured meshes, Fig. 2. The purpose of the refinement is to identify the flow regions that are most dynamic. The criteria for highly dynamic flow region is defined by the values of vorticity vector and it is a user’s option based on the desired values of vorticity field.

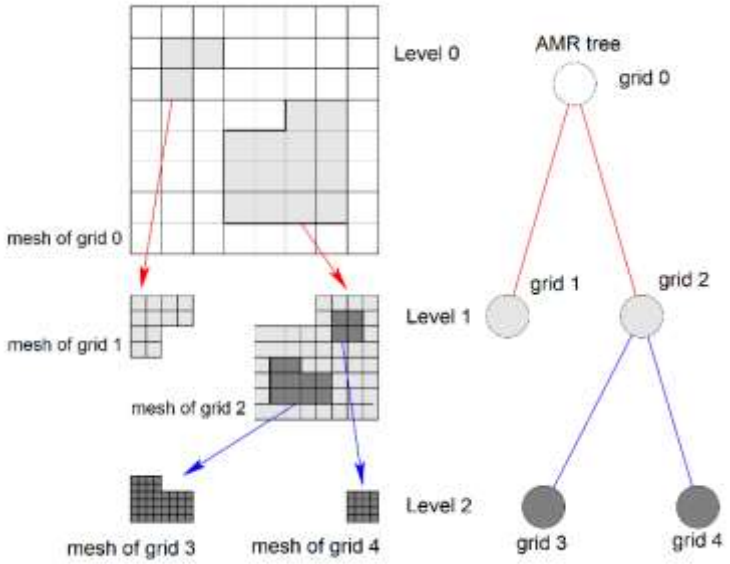


Fig. 1. Adaptive mesh refinement algorithm; structured mesh

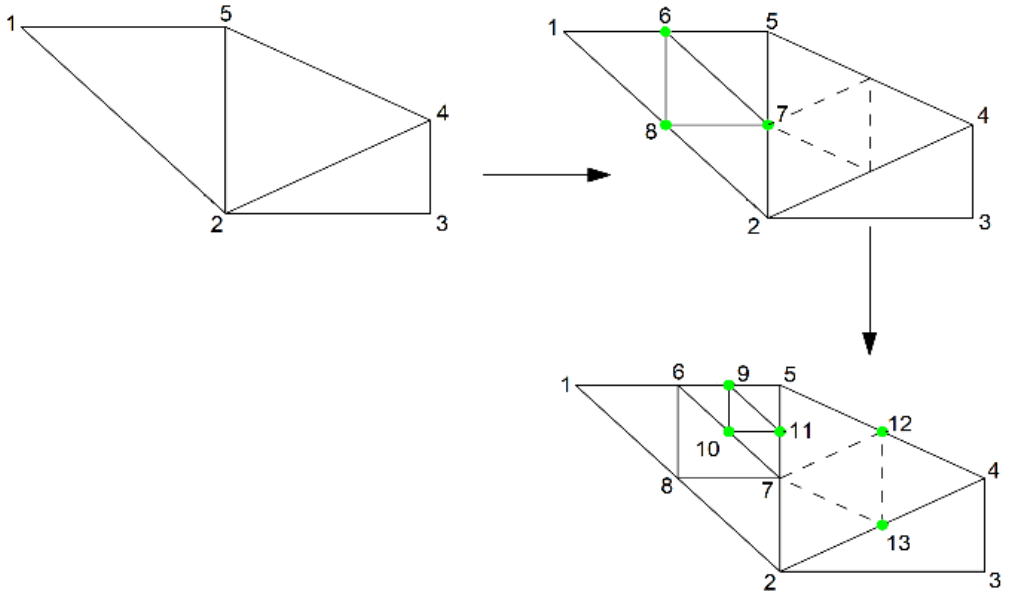


Fig. 2. Adaptive mesh refinement algorithm; unstructured mesh

The vorticity field is given by the determinant of the matrix given by equation 1:

$$\omega = \begin{bmatrix} \vec{i} & \vec{j} & \vec{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ u & v & w \end{bmatrix} \tag{1}$$

where u, v and w are the velocity components, corresponding to the vectors $\vec{i}$, $\vec{j}$, $\vec{k}$.

It is important to mention here that the mesh refinement and implicitly data clustering is performed in the computing stage. The novelty of our approach comes from the fact that we perform the data clustering at the post-processing stage as well. This way we ensure the accuracy of the generated data as well facilitate the data management.

The post-processing data clustering is performed using a fuzzy clustering algorithm, based on the k-means. The mean idea of k-means algorithm is to partition large data points (n-data points) into k clusters. It is worth to mention here that each data belongs to the cluster with the nearest mean.

The k-means algorithm is briefly described in the following. Given a set of data, and assuming 1-D problem, $(x_1, x_2, x_3 \ldots\ldots\ldots x_n)$we can cluster the data into k clusters where ideally $k << n$, resulting into a data set $S = \{S_1, S_2, S_3 \ldots\ldots\ldots S_n\}$. The objective, of the k-means algorithm, is to minimize the variance of the data sets. Thus

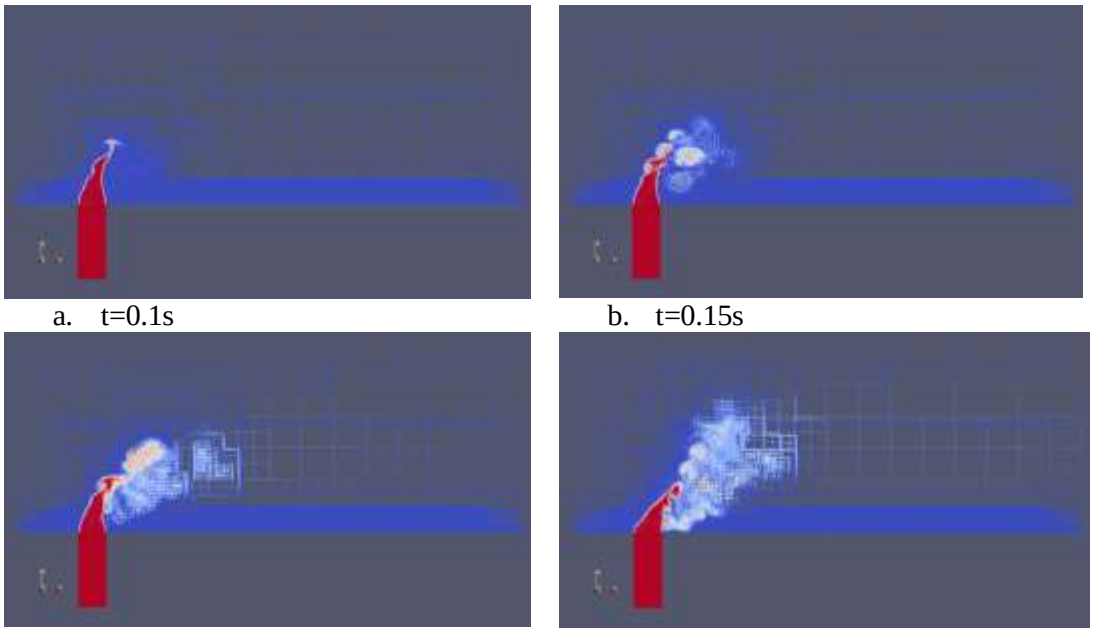
$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2}$$

where μ_i is the mean of the points in the data set S_i . By doing this, the deviations of the points, in the same cluster, are minimized such that

$$\arg\min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} \|x - y\|^2 \tag{3}$$

4. Results and discussion

Figure 3 presents the numerical results using the AMR approach. The data represents CFD results of cross-flow jet at different instants of time. The analysis of the results in Figure 3 reveals the efficiency of the AMR algorithm and illustrates how the grid point are clustered in the regions of highly dynamic vorticity. Regions of low or no vorticity are discretized using a coarser mesh size. This is also observed from the distance between the grid points. The size of the mesh elements, in the case when using the AMR, is about 10^5 .

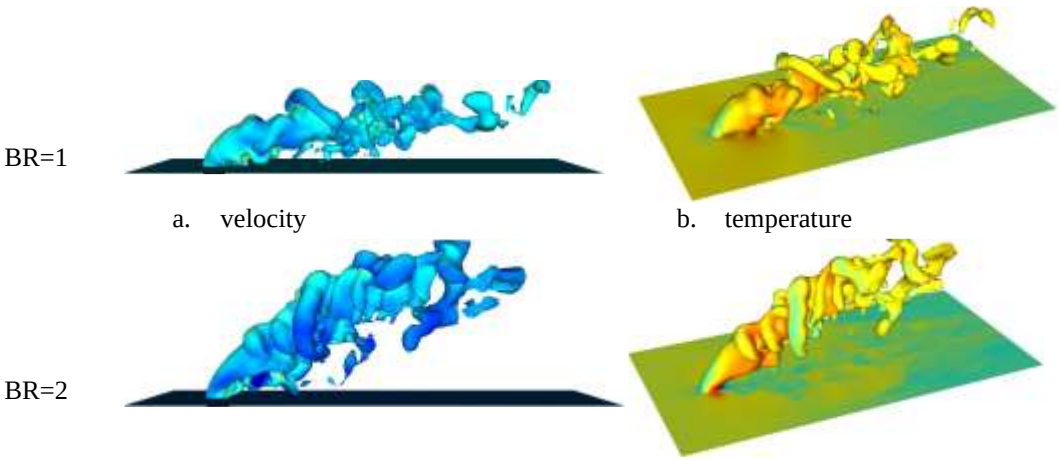


a. t=0.1s b. t=0.15s

c. t=0.2s d. t=0.25s

Fig. 3. Computational results of cross-flow jet using AMR

Figure 4 presents the computational results of the cross-flow fluid dynamics analysis suing the DNS along with the AMR. This study concerns the effect of the velocities' ratio, jet velocity to free-stream velocity, on the development of the fluid flow. The analysis shows that there is a detachment of the flow with the increase of blowing-ratio.



a. velocity b. temperature

c. velocity d. temperature

Fig. 4. Computational results of cross-flow using the AMR

To assess the efficiency of the AMR, we compare the size of the computational domain, and implicitly the number of data points, using AMR versus direct numerical simulation computations. For DNS computations, the size of computational domain is proportional to the Reynolds (R_e) number and proportional with the $R_e^{9/4}$. Thus, for a $R_e \approx 5{,}000$, the number of grid points would be proportional to 210×10^6 grid (data) points. This means a very large number data that is cumbersome to analyze. A simple calculation shows that the use of AMR reduces the number of data points by a factor of 2,100. This is significant data reduction, while capturing all the fluid dynamics of flow field. In the following, the data clustering using the k-means approach is employed for the aerodynamic studies of helicopter blade-vortex interaction (BVI). Thus, Figure 5 presents a schematic of the helicopter BVI phenomenon.

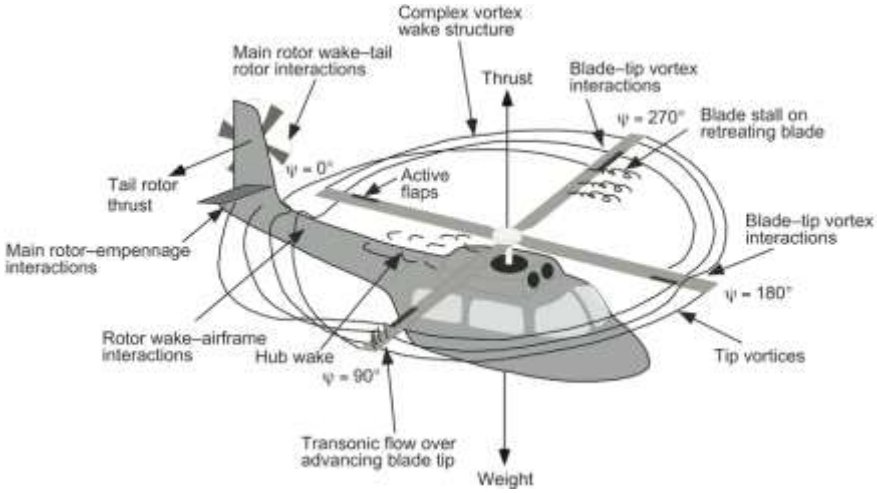
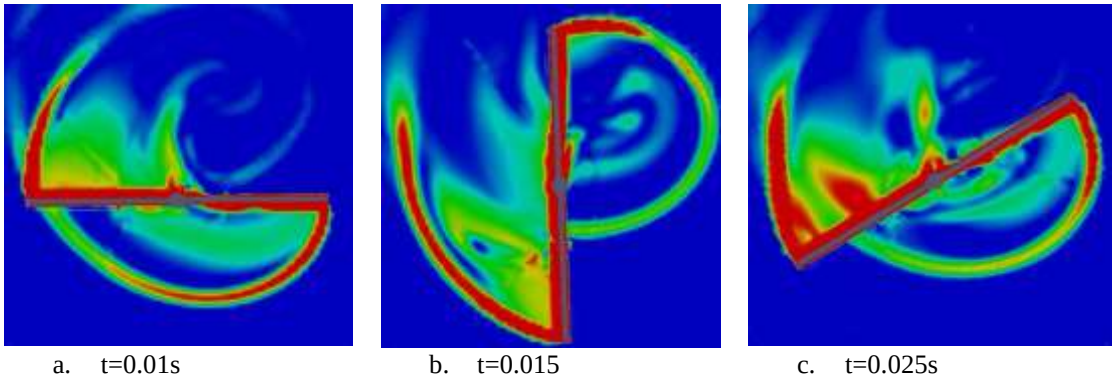
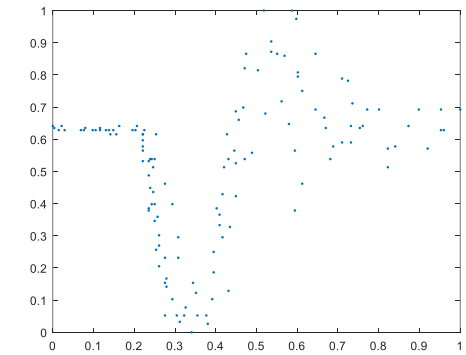


Fig. 5. Schematic of helicopter aerodynamics [7]

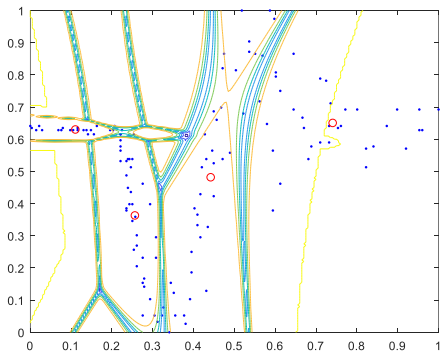


a. t=0.01s b. t=0.015 c. t=0.025s

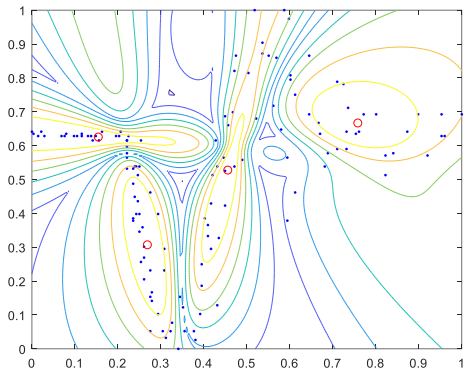
Fig. 6. Computational results of helicopter BVI



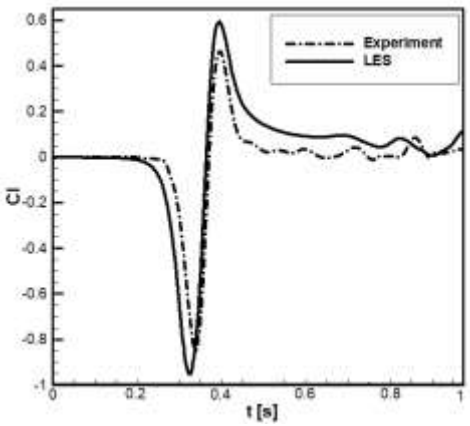
a. data set



b. data clustering



c. refinement of data clustering



d. data comparison

Fig. 7. Data clustering for the problem of BVI

Figure 6 presents the computational results of helicopter BVI for a two-blade configuration. The study reveals the interactions between the advancing blade and the tip-vortex formed at the tip of the blade. These interactions are main factors of noise and vibration in the helicopter dynamics. A set of experimental data (pressure data) is presented in Figure 7a. As it can be seen from Figure 7a, the data contain some noise and we want to separate the good data from the noise data using the k-means algorithm. Performing an initial clustering of data, we are able to cluster the data based on the pressure values, Figure 7b. Further clustering of data reduces even further the variance of the data, Figure 7c. Integrating the data points after the clustering process we obtain the values of the experimental data,

Figure 7d. The comparison between the experimental and computational data shows a very good agreement with the percentage error $\varepsilon \approx 5\%$. It is important to mention here that the AMR algorithm along with the k-means algorithm reduces the computational time by 72%.

Conclusions

An efficient computational approach for data management is developed to facilitate the post-processing of large-set of data. The data management approach comprises two different algorithms, namely the adaptive-mesh refinement and data clustering. The first algorithm reduces the size of the computational domain, while the second algorithm clusters the useful data. The comparison between the experimental and computational data shows very good agreement with a percentage error $\varepsilon \approx 5\%$. The proposed algorithm reduces the computational time by 72% , and this is a significant achievement of the developed method.

R E F E R E N C E S

[1] C. Burstedde, D. Calhoun, K. Mandli, A.R. Terrel, Forest Claw: hybrid forest-of-octrees AMR for hyperbolic conservation laws, in: Parallel Computing: Accelerating Computational Science and Engineering, vol. 25, 2014, pp. 253–262.

[2] X. Chen, V. Yang, Thickness-based adaptive mesh refinement methods for multi-phase flow simulations with thin regions, J. Comput. Phys. 269 (2014) 22–39.

[3] M. Dumbser, O. Zanotti, A. Hidalgo, D.S. Balsara, ADER-WENO finite volume schemes with space-time adaptive mesh refinement, J. Comput. Phys. 248 (2013) 257–286.

[4] N. Favrie, S.L. Gavrilyuk, S. Ndanou, A thermodynamically compatible splitting procedure in hyperelasticity, J. Comput. Phys. 270 (2014) 300–324.

[5] E. Han, M. Hantke, S. Müller, Efficient and robust relaxation procedures for multi-component mixtures including phase transition, J. Comput. Phys. 338 (2017) 217–239.

[6] S. Hank, N. Favrie, J. Massoni, Modeling hyperelasticity in non-equilibrium multiphase flows, J. Comput. Phys. 330 (2017) 65–91.

[7] J. G., Leishman, Principles of helicopter aerodynamics, Cambridge Press., 2016