Multimedia Tools and Applications (2024) 83:10115–10146
https://doi.org/10.1007/s11042-023-15686-2

Parallel and distributed processing for high resolution agricultural tomography based on big data

Gabriel M. Alves[1,2,3] · Paulo E. Cruvinel[1,2]

Received: 1 November 2021 / Revised: 16 July 2022 / Accepted: 23 April 2023 /
Published online: 20 June 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In the field of high-resolution tomography, there is currently a notable increase in the volume of tomographic projections and data produced. Such a context has been demanding new computational approaches to the process of reconstruction and processing of the resulting digital images. This paper presents a new approach to meet such a demand, such as optimizing the set of tomographic projections for the reconstruction process, parallelizing algorithm reconstruction, and processing the data in a distributed manner. In this context, a customized method for the high-resolution tomographic reconstruction of agricultural samples has been validated. Hence, tomographic projections with greater amounts of information based on measurements of the spectral density of the projections can be prioritized, and the reconstructive process parallelization using the known filtered back-projection can be considered (i.e., distributed data flow and the use of the Apache Spark environment). For the operation, such an approach based on the big data environment has been organized, that is considering a cluster installed on the Amazon Web Services platform, whose configuration has been defined after the evaluation of the speedup and efficiency metrics. The developed method proved to be useful for carrying out high-resolution tomography analyses of large quantities of agricultural samples, based on the paradigms of precision agriculture for gains in sustainability and competitiveness of the production process.

Keywords Tomographic image reconstruction · Tomographic selection projections · Big data · Image processing · Precision agriculture

Paulo E. Cruvinel contributed equally to this work.

✉ Gabriel M. Alves
gabriel.marcelino@ifsp.edu.br

Paulo E. Cruvinel
paulo.cruvinel@embrapa.br

[1] Post-Graduation Program in Computer Science, Federal University of São Carlos (UFSCar), Rod. Washington Luís, Km 235, 13565-905 São Carlos, São Paulo, Brazil

[2] Embrapa Instrumentation, 15 de Novembro, 1452, 13560-970 São Carlos, São Paulo, Brazil

[3] Federal Institute of Education, Science and Technology, Av. Marginal, 585, 13871-298 São João da Boa Vista, São Paulo, Brazil

1 Introduction

The study of computed tomography (CT) applied to agriculture, which began in the early 1980s, focuses on soil science and included investigating the processes of water infiltration and the properties of density, moisture, and porosity [24]. In Brazil, the first X-ray and γ-ray minitomograph scanners for soil science applications were built in 1987. This makes it is possible to measure samples in the laboratory and constitutes an important step in the development and advancement of the tomography technique in the country. Subsequently, other tomographs were developed on at millimeter scale, such as portable γ-ray tomography and Compton scattering tomography [4, 7, 20, 27, 31].

Moreover, other types of agricultural research started using CT to develop their studies and perform analyses [1, 2, 13, 23]. CT enables non-invasive analysis of the interior of a body or object and, therefore, an alternative method for evaluating the internal morphology of agricultural samples. The non-invasive analysis of the interior of the agricultural samples is possible because CT produces an image of the interior of a body by reconstructing the projections obtained from X-ray beams that go through a body without damaging it. Therefore, reconstruction from projections is considered a fundamental step and demands high computational capacity, in addition to managing a large amount of data [9, 11, 22].

In this context, the term *big data* can be applied to a large volume of tomographic data because it represents a new method of handling available data nowadays, which is often unstructured. Big data can be applied to the increased demand for analyses, as the number of species and varieties of seeds continues to increase. In addition, it should be noted that big data techniques have already been used in various agricultural applications, such as in the process of tomographic reconstruction, in the treatment of information to be reconstructed three-dimensionally and in the development of new algorithms [3, 6, 12, 17, 19, 26, 34–36]. Therefore, the opportunity to integrate these three areas (e.g., CT, agriculture, and big data) is intended to allow the reconstruction of tomographic images in a big data environment to enable a greater number of agricultural analyses. Thus, good quality image reconstruction should be initially considered using smaller sets of tomographic projections to reduce the time involved in the reconstruction and allow, a significant increase in the number of analyses in the same frame. Consequently, new solutions are of interest in the parallelization of the algorithms involved in reconstruction and in the use of architectures that allow hardware processing [8, 28, 32]. However, it should be noted that the use of cloud computing clusters for tomographic reconstruction has not yet been explored.

This study aims to develop a method for two-dimensional (2D) and three-dimensional (3D) (volumetric) high-tomographic image reconstruction in a parallel and distributed big data environment that will allow the selection of the most relevant projections to reconstruct good quality images to allow a greater number of agricultural analyses to be completed in the same time frame. The main contribution of this study is the distinguished reduction in the time requested for a high resolution CT image reconstruction based on the projections selection by its spectrum of energy. In addition, for both 2D and 3D image reconstruction methods it has been considered the inclusion of parallelization and a framework operating in a distributed environment. The remainder of this paper proceeds as follows. Section 2 presents the fundamentals of CT and the power spectral density (PSD) used for the selection of the projections. Section 3 presents the organization of the method for tomographic reconstruction of agricultural samples in a big data environment. Section 4 presents the results and discussion of this work. Finally, the conclusions are presented in Section 5.

2 Fundamentals of computed tomography and power spectral density

2.1 Computed tomography and methods of reconstruction

The main problem of CT is obtaining an image of the object under study from the reconstruction of projections that were obtained based on transmission. The solution is to reconstruct an image by obtaining line integrals along straight lines that pass through the object.

The physical model of X-ray attenuation in transmission CT is illustrated in Fig. 1. A narrow beam represented by a straight L with intensity $I(x)$ comes from the source and passes through the object, which has a certain attenuation coefficient μ. The detector registers the remaining intensity of the beam, and this information is used to reconstruct the 2D image of the object [14, 16, 25].

From the physical model, the following (1) can be obtained, known as the *Lambert-Beer* equation, which expresses the amount of exponentially attenuated X-rays along straight L.

$$I = I_0 \exp\left(-\int_L \mu(x)dx\right) \tag{1}$$

For tomographic reconstruction purposes, the variation of this attenuation should be measured along the straight L, which can be obtained using the following (2).

$$p(L) = \int_L \mu(x)dx = -\ln\left(\frac{I}{I_0}\right) \tag{2}$$

From this equation, a reconstruction method is obtained by the radon transform, to discover a function $f : \mathbb{R}^2 \to \mathbb{R}$ from all line integrals in a previously determined domain. In CT, it is used to determine the distribution of attenuation $\mu(x)$ which corresponds to the density of the object under study. Therefore, the problem is considered an inverse problem because it seeks to find the attenuation coefficient from the available data, that is, from I and I_0.

One approach to understanding the process of tomographic reconstruction is to consider an X-ray beam as a straight line from the *source* to the *detector*. This set (i.e., source and detector) is rotated by an angle $\theta \in [0, 2\pi)$ so that the entire object is scanned in the plane at one fixed position z. Figure 2 presents a schematic diagram of a parallel projection, highlighting the distance t. Evidently, a projection is a set of line integrals, represented by $P_\theta(t)$, considering the same cross-section or position in the z-axis, as well as the same angle θ of the set (i.e., source and detector) in relation to the fixed coordinates (x, y).

In practice, it is ideal if the linear attenuation coefficient values are given as a result of fixed coordinates (x, y), which is possible using the relation between the polar and Cartesian coordinates. Therefore, the perpendicular distance from the origin to line L (X-ray beam) can be determined using the following (3):

$$t = x\cos(\theta) + y\sin(\theta). \tag{3}$$

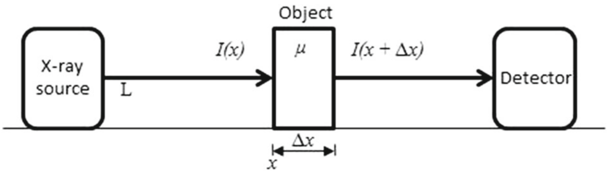


Fig. 1 Physical model of X-ray attenuation

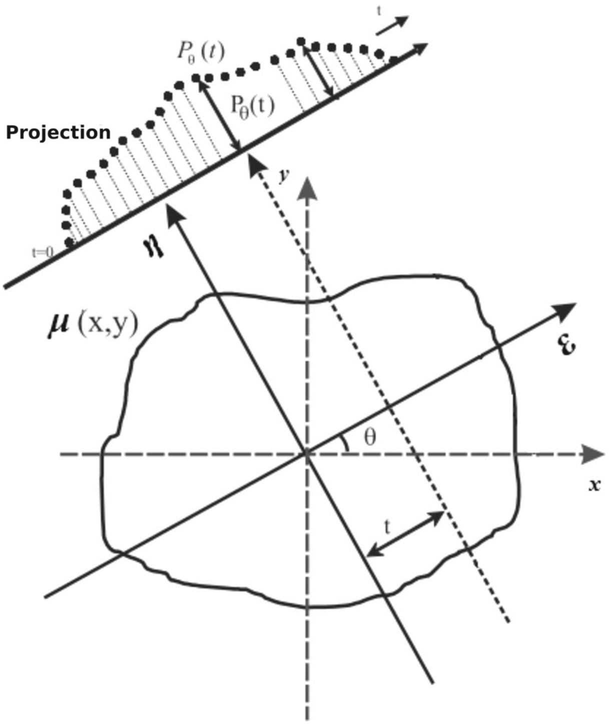


Fig. 2 Schematic diagram of a parallel projection

Thus, $P_\theta(t)$ is based on ε in which angle θ determines the inclination of the ε-axis concerning the horizontal line, and the integral of the function is made on a straight line perpendicular to this axis. Further, scanning the entire interval $\theta \in [0, 2\pi)$ is unnecessary, but only the interval $\theta \in [0, \pi)$ to avoid data redundancy.

Because, computationally, infinite line integrals cannot be obtained, the cross-section, can be represented at a certain angle, making use of the Dirac delta function, which has the sampling property. Equation (3) in the 2D case, can be rewritten as

$$P_\theta(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mu(x, y)\delta(x \cos\theta + y \sin\theta - t)dxdy. \tag{4}$$

Equation (4) is known as the *Radon transform*, $\mathcal{R}_\theta \mu(t) = P_\theta(t)$. Therefore, the problem of reconstructing an image consists of determining $\mu(x, y)$ from $\mathcal{R}_\theta \mu(t)$. The Radon transform maps the space domain (x, y) in the domain (t, θ), where each point in space (t, θ) corresponds to a line in space (x, y).

The Radon inverse transform, $\mathcal{R}^{-1}$, is used to reconstruct μ and can be obtained through the Fourier slice theorem, or the central slice theorem, which relates the projections of the Radon transformation to the Fourier transform.

2.2 Power spectral density

The PSD of a signal is often solved by estimating the autocorrelation function with the available data, which is applied after the Fourier transform to obtain the desired spectral description. However, different approaches are available to perform spectral estimation that can be classified as parametric or non-parametric methods.

The first type is generally simpler to calculate, but it requires a *priori* knowledge signal, whereas the second type assumes no particular structure behind the available data [10, 21].

Given a random signal in the time domain, $\mathcal{X}(t)$, it is assumed that it is sampled over a finite time interval $(-T/2,\ T/2)$ and is denoted by $X_T(t)$. When applying the Fourier transform, we obtain:

$$\tilde{X}_T(f) = F\{X_T(t)\} = \int_{-\infty}^{\infty} X_T(t)e^{-2\pi jft}dt = \int_{-T/2}^{T/2} \mathcal{X}(t)e^{-2\pi jft}dt \tag{5}$$

From (5), we obtain the module and argument of $\tilde{X}_T$ the *amplitude spectrum* and *phase spectrum*, respectively. The *spectral energy density*, is calculated from $\tilde{X}_T$ using the expected value of the square of the amplitude spectrum, as indicated in the (6):

$$E(f) = \mathcal{E}\{|\tilde{X}_T(f)|^2\} \tag{6}$$

It is observed that $E(f)$ tends to infinity when T tends to infinity. Therefore, dividing (6) by the interval of T limits the growth and provides the *density of the power spectrum* expressed by (7), which is real and not negative. This definition is valid and exists for all stationary processes with zero mean and finite variance. For agricultural tomography, the samples to be tested are moved to the tomographic table. They remain stationary during the projection acquisition process so that this theory can be used [5]. Additionally, as the Poisson noise is a priority in the tomographic process, it is also considered that stationary behavior will be exhibited throughout the tomographic process.

$$S(f) = \lim_{T\to\infty} \mathcal{E}\left\{\frac{1}{T}\left|\int_{-T/2}^{T/2} \mathcal{X}(t)e^{-2\pi jft}dt\right|^2\right\} \tag{7}$$

In the discrete case, considering the sequence $x[n]$, we obtain (8), where $\hat{S}$ is the estimator per periodogram. This is equivalent to applying a rectangular window over to interval $0 \leq n \leq (T-1)$ of sequence $x[n]$ to square the Fourier transform module of the truncated sequence and normalizes the result by a factor T to obtain a measure of PSDs.

$$\hat{S}(e^{jf}) = \frac{1}{T}\left|\sum_{n=0}^{T-1} x[n]e^{-jfn}\right|^2 \tag{8}$$

Based on the spectral density of each tomographic projection present in a considered sinogram, the energy of each projection was evaluated to try to identify those that have a more relevant set of information to obtain the tomographic reconstruction in two dimensions. In this context, information on the spectral density of each tomographic projection can be obtained by considering the power spectrum related to it.

When considering signal $s = s(t)$, continuous in time, as a function that represents a random signal and $S = S(\omega)$, a function representing the periodogram of this signal, it is possible to decompose S as follows:

$$S = S_r + jS_i, \tag{9}$$

where S_r and S_i are the real and imaginary parts, respectively, and $j = \sqrt{-1}$. This equation can still be written in polar form as follows:

$$S = |S|e^{j\theta(p)}. \tag{10}$$

Therefore, the amplitudes in the spectrum translated by (10) can be given by:

$$|S| = \sqrt{S_r^2 + S_i^2}. \tag{11}$$

Hence, (8) can be considered when working with an X-ray tomographic projection, which is represented as a sequence $s[n]$, as the signal is discrete; that is,

$$\hat{S}(e^{jf}) = \frac{1}{M}\left|\sum_{n=0}^{M-1} s[n]e^{-jfn}\right|^2, \tag{12}$$

where n is in the interval $0 \leq n \leq (M-1)$ and represents the number of samples in the sequence $s[n]$.

3 Materials and methods

Figure 3 presents the block diagram that illustrates the overview of the method developed for the 2D and 3D (volumetric) reconstruction of tomographic images of agricultural samples in big data environment.

The samples obtained using agricultural tomographs were projections that were inserted and stored in a big data environment, which is represented by the dashed line. The process consisted of selecting the projections that used spectral density to evaluate the associated energy in each tomographic projection to select those that had more relevant information. Subsequently, 2D and 3D parallel reconstruction steps were performed, and finally, the images were made available for viewing.

3.1 Big data environment

The organization of the big data environment was considered from two perspectives: infrastructure and application. These two perspectives are built using a technology stack. Figure 4

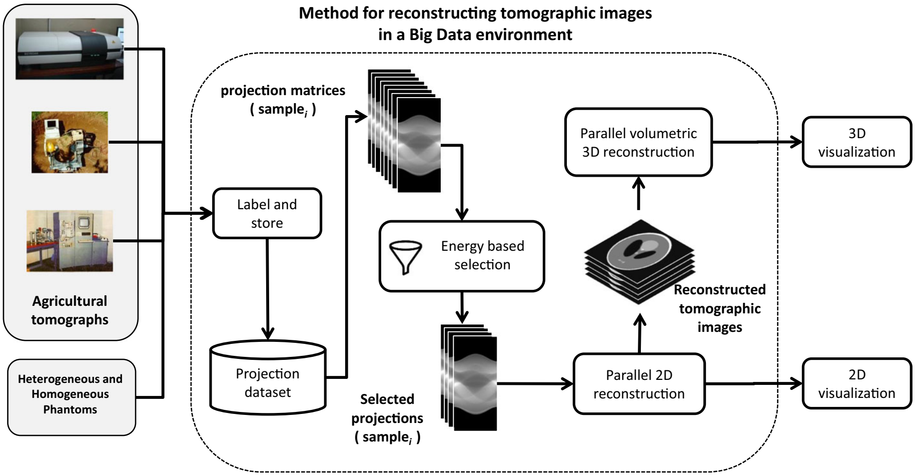

Method for reconstructing tomographic images
in a Big Data environment
Agricultural tomographs
Heterogeneous and Homogeneous Phantoms
projection matrices (sample$_i$)
Label and store
Projection dataset
Energy based selection
Selected projections (sample$_i$)
Parallel volumetric 3D reconstruction
Reconstructed tomographic images
Parallel 2D reconstruction
3D visualization
2D visualization

Fig. 3 Block diagram of the method of reconstruction of tomographic images in big data environment for the analysis of agricultural samples

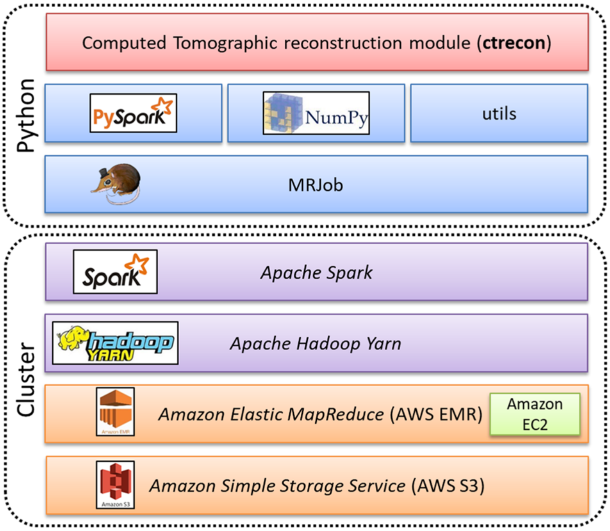

Python
Computed Tomographic reconstruction module (ctrecon)
PySpark
NumPy
utils
MRJob
Cluster
Spark
Apache Spark
hadoop YARN
Apache Hadoop Yarn
Amazon Elastic MapReduce (AWS EMR)
Amazon EC2
Amazon Simple Storage Service (AWS S3)

Fig. 4 Technology stack used for the organization of the big data environment

illustrates the technology stack used in the organization of the big data environment for this work and emphasizes, through the dashed lines, the cluster representing the infrastructure, and the method of reconstruction of the tomographic images, developed using Python language.

Table 1 presents the technologies used for the organization of the big data environment and the respective versions installed and configured to compose the environment.

When observing the technology stack, the first layer refers to data storage using Amazon's distributed file system technology, `S3`. The platform *Amazon Elastic MapReduce* (`EMR`) made it possible to structure the cluster where the computers, or cluster nodes, are instances of the Amazon *Elastic Compute Cloud* (`EC2`). Table 2 presents the instance types that were used to evaluate the prepared environment for the correct operation of the developed method.

The use of homogeneous clusters, which are organized using the same machine configuration, should be highlighted. Thus, for evaluation purposes, the number of instances from each developed model can be used.

Table 1 Versions of the technologies used in the organization of the big data environment

Technology	Version
PySpark	2.4.2
Numpy	1.16.4
MRJob	0.6.9
Apache Spark	2.4.2
AWS EMR	emr-5.24.0
Apache Hadoop YARN	2.8.5
Java (OpenJDK)	1.8.0_201
Python	3.7.6

Table 2 Types of the instances used in the composition of the cluster

Model	vCPU	Memory (GiB)
m5.xlarge	4	16
m5.2xlarge	8	32
m5.4xlarge	16	64

From the perspective of the application, the layer of the library `MRJob`[1] was responsible for integrating applications written in Python with various cloud computing services, such as those offered by Amazon.

The reconstruction method of 2D and 3D tomographic images (volumetric) represented by the last layer in Fig. 4 was written using Python language, with a module called `ctrecon`, and several libraries that are represented in the penultimate layer, such as PySpark[2] and, Numpy[3], as well as several auxiliary libraries represented by the block *utils*.

3.2 Tomographic projection selection model

The model for the selection of the tomographic projections, developed in this work, is based on (12), where the energies of the tomographic projections contained in a sinogram can be calculated from the PSD.

From the calculation of tomographic projections and the number of projections are contained in the sinogram in question, (13) is used to determine the number of classes contained in this set of energies.

$$k = \lfloor \sqrt{N} \rfloor, \tag{13}$$

where N represents the number of tomographic projections contained in a sinogram. The floor function, denoted by $\lfloor x \rfloor$, converts a real number x in the higher whole number less than or equal to x, which in this case refers to the number of classes that will be defined for the sinogram considered.

In this context, the set of tomographic projections that compose a sinogram is understood to be a set of energies $E = \{\hat{S}_0, \hat{S}_1, \hat{S}_2, \ldots, \hat{S}_{N-1}\}$, in which each energy $\hat{S}_i$, where $i = 0, 1, 2, \ldots, N-1$, represents a projection.

From the set of energies, the classes are defined to make the set of classes $C_T = \{C_0, C_1, C_2, \ldots, C_{k-1}\}$, where a particular C_j, where $j = 0, 1, \ldots, k-1$, represents a subset of the energies contained in the whole E.

The interval Δ, of energies associated with each class, is expressed using (14), which considers the greater and lesser energy found in the set of energies E, as well as the number of classes defined using (13).

$$\Delta = \frac{\max E - \min E}{k} \tag{14}$$

[1] https://github.com/Yelp/mrjob

[2] https://spark.apache.org/docs/latest/api/python/index.html

[3] https://numpy.org/

Therefore, each class has an initial energy ei, and a final energy, ef, so that $C_j = [ei, ef[$. The initial energy of a class is given by (15).

$$ei_j = \begin{cases} \min E & \text{se } j = 0, \\ ef_{j-1} & \text{c.c.} \end{cases} \tag{15}$$

The final energy of a class is given by (16)

$$ef_j = \begin{cases} ei_j + \Delta & \text{se } j < (k-1), \\ \max E & \text{c.c.} \end{cases} \tag{16}$$

After defining the energy classes and intervals, the tomographic projections were classified according to their energy values. A Gaussian distribution was considered during the development of the model. In this context, the classes are following this distribution model. Therefore, after the classification of the projections, the averages $(\mu_0, \mu_1, \ldots, \mu_{k-1})$ and standard deviation (σ) of each class were calculated.

Figure 5 illustrates a conceptual representation of the energy classes, as well. Classes C_j and the initial and final energies and the Gaussian distribution associated with each class. The hatched region indicates the region where the most significant projections of each class were identified.

Therefore, the selection criterion consisted of choosing the most significant probabilities within each energy class, which translated to the tomographic projections that presented the largest amount of information. Shannon showed that information can be quantified and, the amount of information is related to probability [29, 30, 33]. Therefore, projections that contained energy within the range of a standard deviation $([-\sigma, \sigma])$ in each class were considered significant, leading to the formation of the sets C_j^{sel}, to $j = 0, 1, \ldots, k-1$, which contained the selected projections for each energy range associated with the classes.

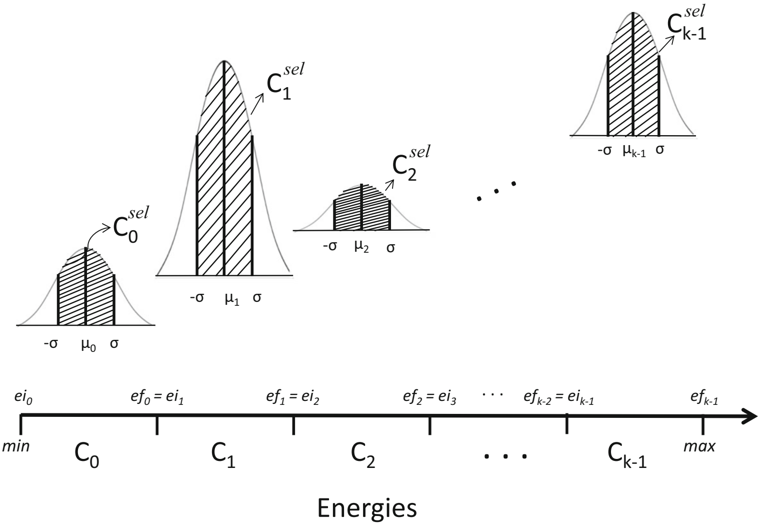


Fig. 5 Conceptual representation of energy classes considering the Gaussian distribution in each class. The hatched region corresponds to the region of each class where the most significant projections are found

Next, the tomographic projections identified as more significant in each class, based on the energy information, were grouped to form a new sinogram composed of a smaller number of projections in relation to the original sinogram.

The new sinogram refers to $F = \{C_0^{sel}, C_1^{sel}, \ldots, C_{k-1}^{sel}\}$. In addition, the tomographic projections contained in the new sinogram were organized according to the angle at which they were acquired to prepare for the reconstruction stage in two dimensions.

3.3 Reconstruction algorithm from the selected projections

Algorithm 1, implemented in Python language, was executed in the big data environment prepared in this study, to perform tomographic reconstruction from the previously selected projections.

The algorithm was structured in three main steps: (i) selection of tomographic projections, (ii) 2D reconstruction, and (iii) 3D reconstruction (volumetric). After reading the projection matrices (set M), the distribution of the matrices to the nodes of the cluster was determined. Subsequently, the projections are selected in a distributed manner, in which each node is responsible for processing a subset of projection matrices. Figure 6 shows the tomographic projection selection process.

The second stage of the algorithm consists of 2D tomographic reconstructions of the matrices with the selected projections of the previous stage. The reconstruction was performed using the FBP algorithm, which is based on the Fourier slice theorem. After filtering the projections that compose the matrix s_i, the process of reconstruction was initiated, and interpolation in the domain of space was conducted. Subsequently, the back-projection stage

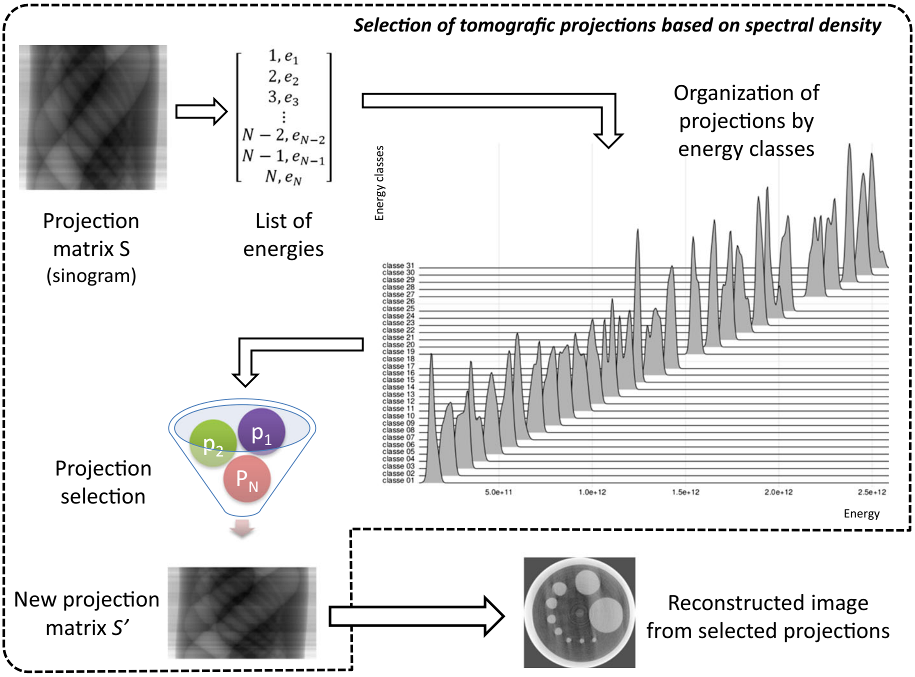


Fig. 6 Selection of tomographic projections applied to a *phantom* sinogram

Algorithm 1: Tomographic Reconstruction Method in a Big Data environment

```
Input  : Set of projection matrices $M = \{m_1, m_2, \ldots, m_N\}$
Output : Reconstructed volume $V$

distribute $M$ to cluster nodes

selects projections in the matrices of the $M$ set and produces the $S$ set
$S \leftarrow \varnothing$
foreach $m_i \in M$ that is in a cluster node, do
    $list_e \leftarrow$ evaluteEnergy($m_i$)
    $F_i \leftarrow \varnothing$
    $k \leftarrow \lfloor \sqrt{N} \rfloor$
    $\Delta \leftarrow \lfloor (\max(list_e) - \min(list_e)) \div k \rfloor$
    $C_T \leftarrow$ generateClasses($list_e$, $\Delta$, $k$)
    for each $C_j \in C_T$ do
        $\mu_j \leftarrow$ compute class average $C_j$
        $\sigma \leftarrow$ compute class standard desviation $C_j$
        $C_j^{sel} \leftarrow$ class projections $C_j$ with $\sigma \leq 1$
        $F_i \leftarrow$ add $C_j^{sel}$
    end
    $S \leftarrow F_i$ // adds the new sinogram to set $S$
end
performs 2D reconstruction
foreach $F_i \in S$ that is in a cluster node, do
    $I_i \leftarrow \varnothing$ // reconstructed image
    for each $P_\theta \in F_i$ do
        $P_{filtrada} \leftarrow$ filter($P_\theta$)
        $tmp \leftarrow$ interpolate($P_{filtered}$)
        $I_i \leftarrow I_i + tmp$
    end
end
performs 3D reconstruction (volumetric)
foreach $I_i$ that is in a cluster node, do
    split in ( id-region, (section, tile) )
end
collect tiles por id-region forming blocks
foreach block that is in a cluster node, do
    $v_i \leftarrow$ bspline(block)
    $V \leftarrow v_i$
end
```

was responsible for the sum of the filtered and interpolated projections to compute the contribution of each projection in the pixel of the reconstructed image I_i, which can also be referred to as a slice.

The volumetric reconstruction stage developed in this study used the reconstructed slices (I) and added virtual slices, generated using B-spline interpolation, to produce the final volume. The slices were stacked, and for each position, a set of points was formed, which were interpolated to generate a set of voxels that composed the final volume. The parallelization strategy adopted in this study consisted of dividing the slices into small regions (*tiles*) and applying the interpolation process to a subset of points. Figure 7 presents a block diagram of parallel volumetric reconstruction.

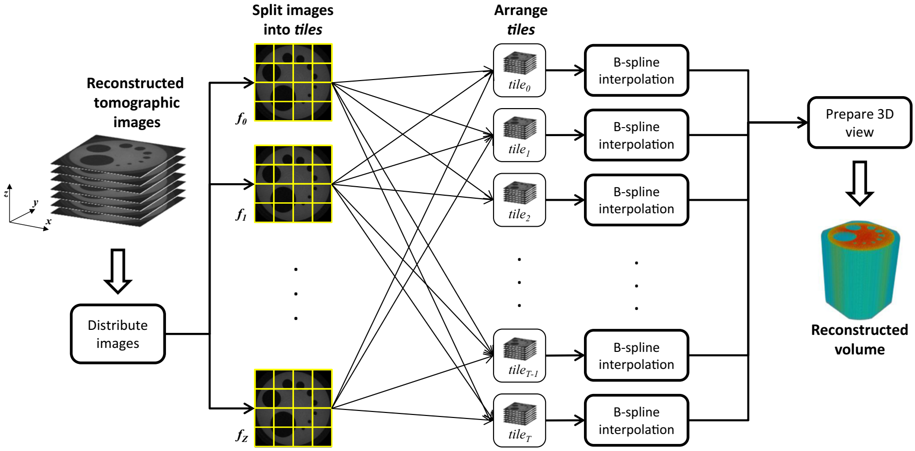


Fig. 7 Block diagram of parallel volumetric reconstruction considering an example of the processing of a *phantom*

The first phase consisted of dividing each slice into regions or *tiles*. One region received an identification necessary for the final volume reconstruction. Each slice was divided into rows and columns, where identifiers were used to identify a certain region, as illustrated in Fig. 8.

The identifier (`id-region`) was formed by eight characters, where the first four indicated the row and the others indicated the column. In addition to the identifier, it was necessary to include the section (position of z-axis) and the set of pixels associated with the region. Therefore, a region was registered in the big data environment in the following format: (`<id-region>`, (`<section>`, `<tile>`)). The format is a tuple, where the first value is the region identifier and the second value is a new tuple containing the section and set of pixels. The next phase of volumetric reconstruction consisted of an intermediate reduction operation responsible for grouping the regions by an identifier so that each subset contained all the slices of a given region. Subsequently, B-spline interpolation was performed for each

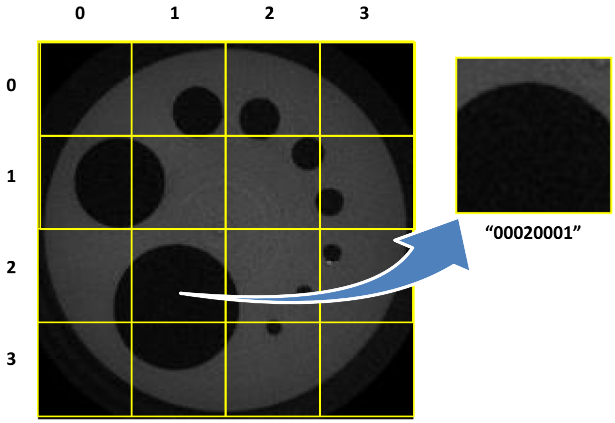


Fig. 8 Identification of a region (*tile*) in a slice. The region and its identifier are composed of eight characters

Table 3 SkyScan 1172 tomograph parameter values were adjusted for acquisition of *phantom* projections and seed samples

Parameter	Value
Number of files	976 (`.tif`)
Number of sections	2096
Rotation step (degrees)	0.20
Time of exposure	790 ms
Rotation interval	0° a 180°
Rotation step	0.500
Image pixel size	$8.54\mu m$
Source voltage	100kV
Source current	$100\mu A$
First section	68
Last section	1968
Reconstruction angular range (degrees)	195.20

subset. Finally, in the last phase, data were made available for 3D viewing. Considering that in this study, visualization was planned to be performed outside of the big data environment, the final data of the volumetric reconstruction were saved in files for later use.

3.4 Experimental evaluation

For the evaluation of the developed method, a heterogeneous plexiglass *phantom* sample was prepared, each with nine holes[4] and $100mm$ diameter and $150mm$ in height.

Additionally, in this study, samples of seven types of seeds were used: peanut (*Arachis hypogaea*), cowpea (*Vigna unguiculata*), sunflower (*Helianthus annuus*), chickpea (*Cicer arietinum*), wheat (*Triticum*), pumpkin (*Cucurbita*) and soybean (*Glycine max*).

The acquisition of the seed sample projection matrices and *phantom* projection matrices was performed using the SkyScan 1172 tomograph. To achieve this, equipment should be adjusted to use the same configuration for both the samples and *phantom*. Table 3 presents the values of the parameters adjusted in the tomograph for the projection acquisition.

Five samples were prepared for each type of seed, except for cowpea and peanut, where four samples were prepared, totaling 33 samples. The samples were scanned in the range of $0° - 195.2°$ with an angular pitch of 0.2°. Therefore, a matrix of projections of a slice contained 976 projections and 2000 points per projection. Each projection point had 2 bytes, so a sample corresponded to $1960 \times 2000 \times 976 \times 2 \approx 7.13$ GB. In addition to the 33 samples, a *phantom* was prepared with the same settings as the seeds; therefore 34 samples were used, totaling 66, 640 tomographic projections. Each sample was 7.13 GB, making a total of approximately 242 GB of analyzed tomographic data.

The analysis of the selection of projections consisted of calculating the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR). For each calculated metric, the maximum, minimum, and median measurements were observed. The reference image (*ground truth*) for each analyzed slice was prepared from 2D tomographic reconstruction considering the 976 projections. In addition, a region of interest (ROI) was defined with dimensions of 1200×1200 pixels, positioned at the center of the reconstructed images, which provided the data for the analysis of the measurements.

[4] The diameters of the holes were: 3.00, 4.00, 4.60, 7.40, 8.60, 11.13, 13.70, 24.00 and $34.55mm$

Each sample contained 1960 slices; for the calculation of the metrics, a subset of the total slices was selected, considering a confidence interval of 99%, margin of error of 5%, and proportion $p = 0.50$, as no a priori information on the slices was observed. Therefore, 498 projection matrices on each seed sample were used to analyze the selection of tomographic projections.

The SSIM index, given by the (17), considers image degradation as a perceived change in structural information while incorporating phenomena such as luminance and contrast. Structural information consists of the idea that pixels have a strong interdependence, especially when they are spatially close. These dependencies provide important information about the structures of the objects in the scene.

$$SSIM(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{17}$$

where x and y are the original and resulting images, respectively; μ_x is the average of x; μ_y is the average of y; σ_x^2 is the variance of x; σ_y^2 is the variance of y; σ_{xy} is the covariance of x and y; $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are variables that stabilize the division, with L being the range of pixel values (often $2^{bits} - 1$), and $k_1 = 0.01$ and $k_2 = 0.03$, respectively.

The PSNR measure, given by (18), is based on the signal-to-noise ratio, which is an estimate of the reconstructed image compared to the original image.

$$PSNR(x, y) = 10 \log \frac{s^2}{MSE(x, y)}, \tag{18}$$

where $s = 255$ for images with 256 gray levels and the MSE measure (*mean squared error*), given by the (19), is computed as the average of the squared intensity between the original and resulting images, both of size MXN.

$$MSE(x, y) = \frac{1}{MN} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} e(n, m)^2, \tag{19}$$

where $e(m, n)$ is the difference between the original and the resulting image.

To evaluate the parallel environment, the speedup metric was used, that is, calculated by the (20), which determines the increase in speed by the execution of a program when using p processors, in relation to its sequential execution using a single processor. In the equation, T_{seq} and T_{par} are the sequential and parallel times, respectively, to execute the same program.

$$S_p = \frac{T_{seq}}{T_{par}} \tag{20}$$

Another measure was used to evaluate the big data environment namely efficiency, whose value is obtained using (21).

$$E_p = \frac{S_p}{p} \tag{21}$$

The efficiency measure evaluates how much parallelism is explored in an algorithm, and it quantifies the processor utilization. Generally, the measurement value is in the interval [0, 1) and the closer to 1 the value of E_p the greater the efficiency. Generally, the measurement value is in the interval [0,1), because a superlinear speedup may occur during parallel processing [15, 18].

4 Results and discussions

This section presents the results obtained in the process of selecting projections based on spectral density, as well as the 3D (volumetric) visualization of agricultural seed samples.

4.1 Infrastructure analysis for the tomographic reconstruction method based on big data

We sought to evaluate the most appropriate cluster configuration for the analysis of big data. In fact, we have considered the evaluation not only for 2D but also for volumetric (3D) tomographic reconstruction methods. In this sense, the following aspects were considered:

1. **Number of cluster nodes**: Four types of clusters were organized, each containing a certain number of nodes. The aim was to evaluate the operation of the method in the cluster as a function of the number of machines used. It is noteworthy that out of all of the nodes, for all types, one node has been configured as a master and the others as workers. In such an analysis, four types of clusters were considered (i.e., each of them having 4, 6, 8, and 10 nodes, respectively).
2. **Cluster nodes' model**: Three machine configurations were selected to compose the cluster, as presented in Table 2. Table 4 presents the cluster capacities as a function of the number of vCPUs and RAM memory capacity. In addition, to make it easier to understand, the individual configurations of each node model are presented.

Table 4 shows the cluster configuration, where 12 different configurations were evaluated, that is, from the smallest capacity containing 16 vCPU and 64 GB of RAM memory to the largest capacity configuration, with 180 vCPU and 640 GB of RAM memory. In this study, note that the strategy adopted for the parallelization of the two-dimensional reconstruction considers the distribution of the tomographic projection matrices by the nodes of the cluster so that the granularity was considered in terms of average levels. This is because the algorithm has only divided the total volume of the matrices related to the tomographic projections.

The results obtained from the measurements of both speedup and efficiency are presented below. They have allowed the definition of the most appropriate cluster configuration for high-resolution agricultural tomographic reconstruction based on the big data method.

4.2 Speedup assessment

As discussed previously, the Speedup calculation is based on the sequential operation time to the parallel operation time ratio. Therefore, for the evaluation and analysis, a set of tomographic projection matrices, with dimensions equal to 976 × 2000 (2k), was submitted to

Table 4 Cluster's capacity is a function of the number of nodes, and the individual configuration of each node model

	Node configuration		Cluster capacity (vCPU/Memory)			
Model	vCPU	Memory	4 nodes	6 nodes	8 nodes	10 nodes
m5.xlarge	4	16 GB	16/64	24/96	32/128	40/160
m5.2xlarge	8	32 GB	32/128	48/192	64/256	80/320
m5.4xlarge	16	64 GB	64/256	96/384	128/512	160/640

Table 5 Sequential time evaluation for tomographic reconstruction - (2D) and volumetric (3D)

Model	2h Reconstruction	Volumetric Reconstruction 3D	Total
m5.xlarge	31h 41min	8h 52min	40h 33min
m5.2xlarge	31h 23min	8h 32min	39h 55min
m5.4xlarge	31h 05min	8h 29min	39h 34min

a single machine, considering each of the models indicated in Table 4. The results of the sequential time obtained for both 2D and volumetric (3D) tomographic reconstruction are shown in Table 5.

Figures 9 and 10 are shown respectively the evaluation of the speedup measurements for a set of 2D and 3D reconstructions, both calculated from the results presented in Table 5.

The clusters that used m5.4xlarge machines delivered higher speedup values owing to the larger number of processors (vCPU) and the amount of available RAM memory. Clearly, the Speedup for clusters that used m5.xlarge and m5.2xlarge machines could produce better results, mainly for volumetric reconstruction (3D). The reason for this is associated with the fact that, after rebuilding the volume, the algorithm developed saves the data on disk directly from the worker nodes, rather than sending it to the master node, therefore reducing the communication time.

From the perspective of analyzing the big data infrastructure, it is noteworthy that results have illustrated fast reconstruction from tomographic sinograms when considering a total of 1960 high-resolution tomographic projections and even when considering complete matrices from tomographs. In fact, the described method has proved to be adequate, as it has become possible to consider the assessment of a large number of analyses. This allows for the greatest demand for tomographic evaluation based on less time consumption. Next, the efficiency measurements are discussed to analyze the operation of the method as a function of the cluster configuration.

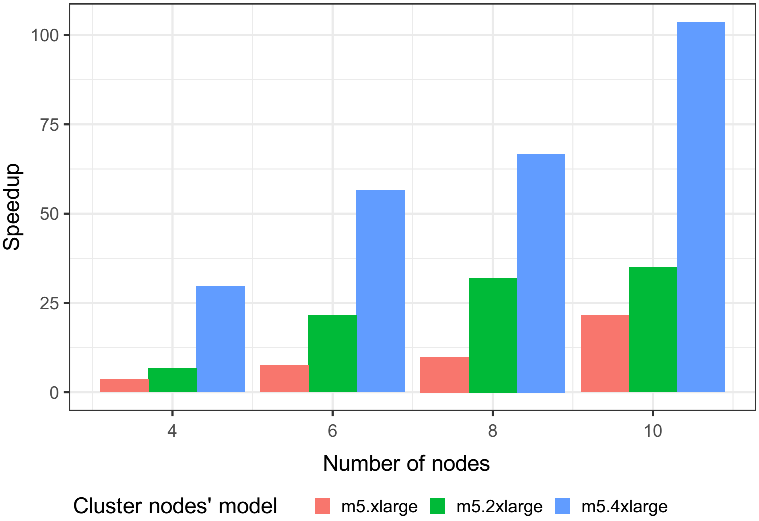


Fig. 9 Speedup evaluation for a set of 2D reconstructions

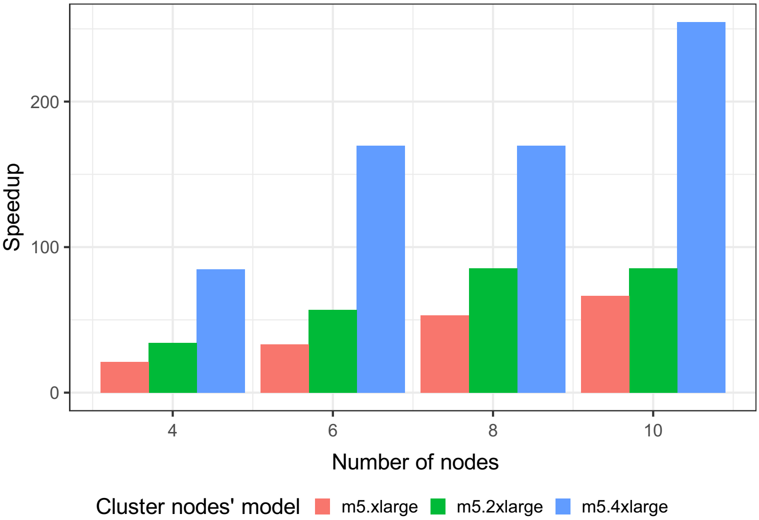


Fig. 10 Speedup evaluation for a set of volumetric (3D) reconstructions

4.3 Efficiency assessment

As mentioned previously, the measurement of efficiency consists of the speedup to the number of processors ratio, as well as a result that allows the evaluation related to how much parallelism can be explored in an algorithm, as well as for quantifying the use of each processor.

Figure 11 shows the resulting evaluation of the efficiency related to the 2D tomographic reconstruction. In such a plot, clusters with m5.4xlarge machines and 10 nodes showed a better efficiency, as expected.

Figure 12 shows the resulting evaluation of the efficiency related to the volumetric (3D) tomographic reconstruction. In such a plot, the cluster with six m5.4xlarge machines and six nodes has the highest efficiency when compared to the other evaluated configurations. This also includes the clusters with machines that operate based on the same model, but with a larger number of nodes. Such a result may also affect the final processing cost.

Furthermore, by considering the processing cost of each machine in the cluster, it was possible to observe the sequential processing of the parallel processing cost ratio. Table 6 presents the cost in US dollars per hour as a function of the machine model configuration used for processing.

It can be observed in Table 7 that the cluster configuration with six nodes, that is, based on the m5.4xlarge model, presented better sequential processing to parallel processing ratio in terms of cost. In fact, in the 2D tomographic reconstruction, the parallel processing was

Table 6 Processing cost in US dollars (USD)

Model	Cost/hour (USD)
m5.xlarge	0.192
m5.2xlarge	0.384
m5.4xlarge	0.768

Table 7 Cost ratio between sequential and parallel processing for both 2D and volumetric (3D) tomographic reconstructions

	4 nodes		6 nodes		8 nodes		10 nodes	
Model	2D	3D	2D	3D	2D	3D	2D	3D
m5.xlarge	1.25	7.09	1.51	6.65	1.40	7.60	2.40	7.39
m5.2xlarge	2.27	11.37	4.33	11.37	4.56	12.19	3.87	9.48
m5.4xlarge	9.87	28.29	***11.31***	***33.95***	9.52	24.25	***11.52***	***28.29***

11 times more than in sequential processing. Similarly, for volumetric (3D) tomographic reconstruction, parallel processing has been proven superior, that is, in this case, in the order of 33 times, when compared to sequential processing. For this reason, it was decided the cluster with six nodes based on the m5.4xlarge model (one master node and five workers nodes), was to be used not only to process the experimental samples but also to evaluate the selection process of their tomographic projections.

In order to have a comparative analysis of the developed method in relation to a commercially available algorithm for 2D tomographic reconstruction, it has been considered projection matrices containing 976 projections with 2,000 points per projection, i.e., from one heterogeneous phantom, as presented in Section 3.4. For such a comparison all the tomographic projections have been considered and then tomographic reconstructions based on both the SkyScan 1172's software that uses the FDK algorithm (Feldkamp-Davis-Kress) and the developed method were carried out. Figure 13 presents the result for the comparative study in which the developed method took an average of 1.07s to reconstruct a slice, while the SkyScan 1172's software took an average of 4.00s to reconstruct the same slice, i.e., having the same dimension and number of tomographic projections.

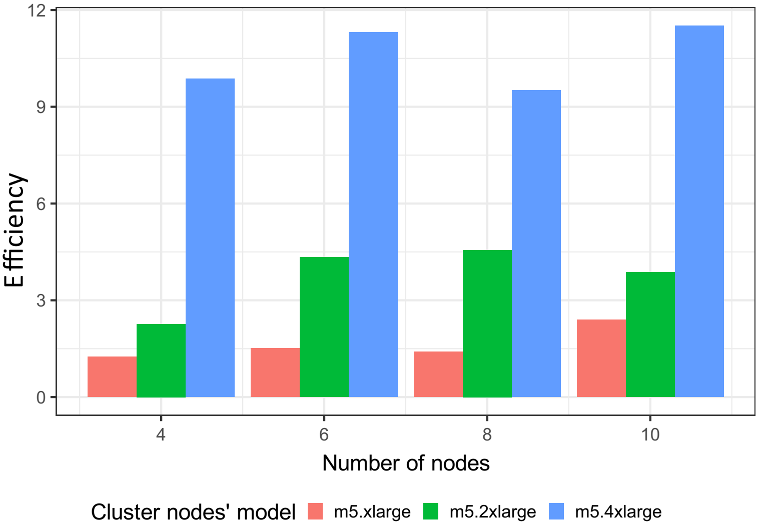


Fig. 11 Efficiency evaluation for 2D tomographic reconstruction

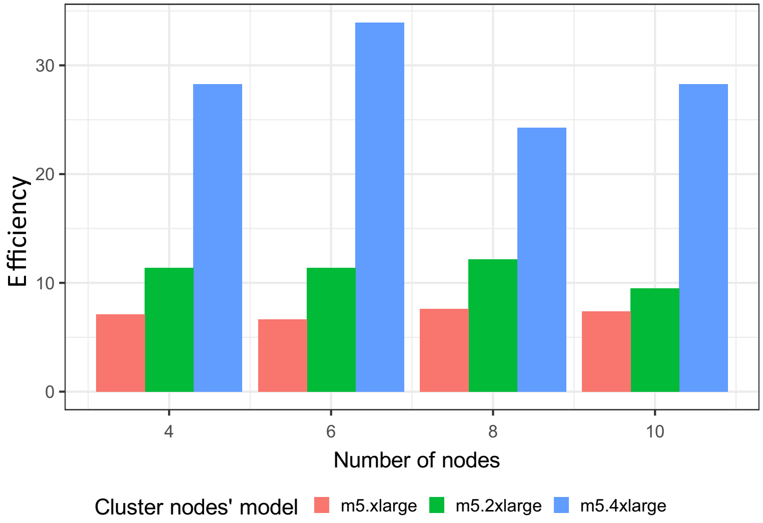


Fig. 12 Efficiency evaluation for 3D tomographic reconstruction (volumetric)

After the analyses, it was observed that the processing time for tomographic reconstruction in a big data environment was approximately 35 min, that is when considering the 1960 sinograms. This corresponds to a total of 1,912,960 projections or 7.13 GB of data. This processing time includes the loading of the projections in the environment, the selection of the projections in the sinograms, and 2D and volumetric (3D) reconstructions.

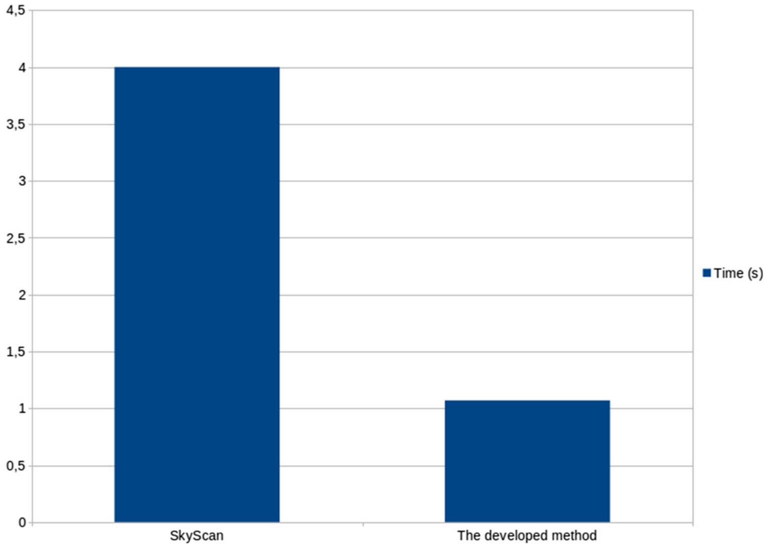


Fig. 13 Time comparison between the 2D tomographic reconstruction of the SkyScan 1172 software and the method proposed in this study

Table 8 Minimum, median, and maximum SSIM values calculated for the *phantom*

Value	Slice	SSIM	Selection
Minimum	728	0.774	61.78%
Median	1276	0.813	62.50%
Maximum	988	0.931	63.83%

4.4 Structural similarity assessment

The SSIM was observed, which allowed the structural information of the reconstructed images to be evaluated from a subset of projections. This allowed the quality of the 2D reconstruction to be verified by reducing the amount of data. The closer one gets to 1 in the value of the SSIM, the more identical the reconstructed image, and the fewer the projections in relation to the reconstructed image with all available projections.

Initially, the analysis of the SSIM for the *phantom* was performed considering 498 slices, chosen according to the above strategy. Table 8 presents the minimum, median, and maximum values of the SSIM and indicates the slice and selection rate of the projections.

The selection rate was 62.50% for an SSIM of 0.813. With a reduction of 37.50% from the initial data set, it was still possible to obtain an SSIM value of approximately 0.800.

Figure 14 presents the reconstructed images of the slices mentioned in Table 8. It can be observed that the main information of the slices was preserved even with the reduction of the number of projections.

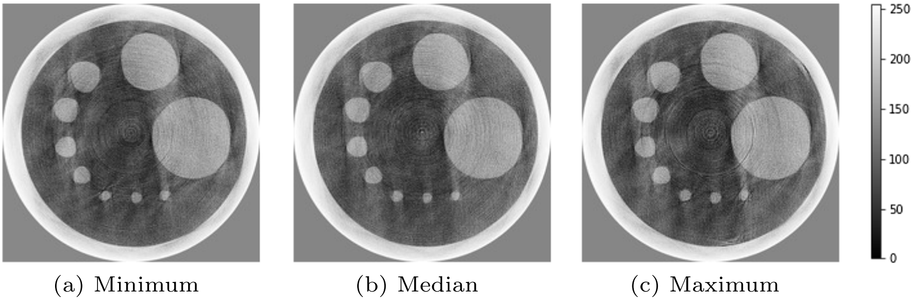
The first analyzed data set refers to cowpea seeds, whose SSIM values are reported in Table 9.

Figure 15 shows reconstructed images of the slices of an analyzed sample of cowpeas.

In the cowpea samples, the highest selection rate (82.27%) resulted in the highest SSIM value (0.947), whereas the lowest selection rate (49.59%) did not produce the lowest SSIM value. In this case, it is observed that the slice that obtained the lowest SSIM value (0.687) presented a selection rate of projections of 61.27%, which is identical to the selection rates of the slices whose SSIM values correspond to the median of each set.

Table 10 presents the values obtained from the SSIM, as well as the selection rate of the projections of the second set of sunflower seed samples.

In the sunflower seed samples, it was observed that the reconstructed image of slice 1368 obtained an SSIM value of 0.910, with a projection selection rate of 65.98%. It is interesting

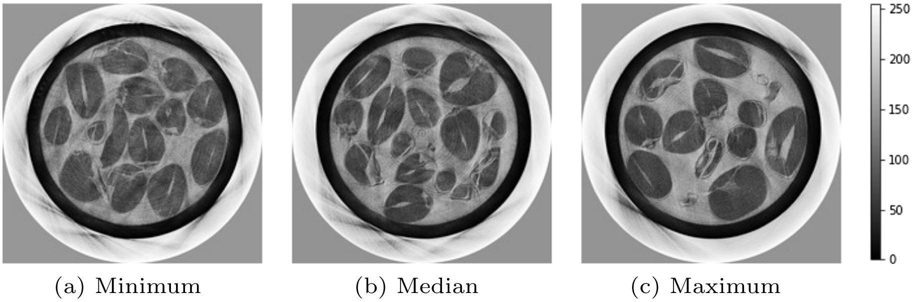


(a) Minimum (b) Median (c) Maximum

Fig. 14 Slices of *phantom* that represent the minimum, median and maximum values of the SSIM measure. All images are represented by the gray scale ranging from 0 to 255

Table 9 Evaluation of tomographic images of cowpea samples: minimum, median, maximum SSIM values, and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
06	1696	0.687	61.27%	1152	0.854	69.47%	700	0.947	82.27%
07	1924	0.753	57.89%	1256	0.828	66.70%	484	0.922	71.31%
08	444	0.702	49.59%	1080	0.809	61.68%	832	0.893	68.24%
09	360	0.789	62.19%	876	0.846	69.26%	1580	0.928	72.23%

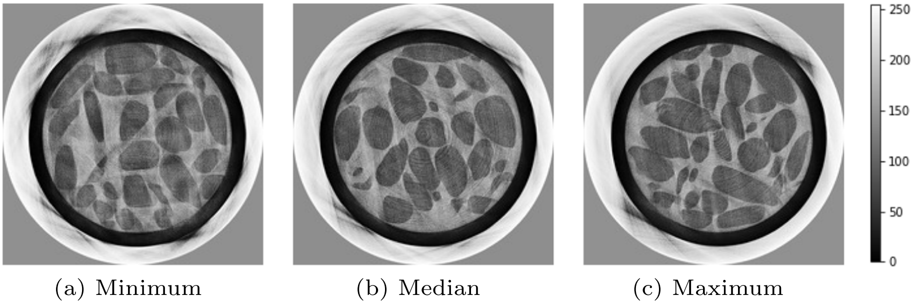


(a) Minimum (b) Median (c) Maximum

Fig. 15 Tomographic images of a sample of cowpeas corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

Table 10 Results for the evaluation of tomographic images of sunflower samples. Minimum, median, maximum SSIM values, and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
10	268	0.745	58.20%	620	0.854	71.62%	1236	0.928	77.05%
11	1656	0.773	71.72%	704	0.854	68.65%	1708	0.922	75.72%
12	1856	0.734	73.98%	500	0.840	58.71%	1020	0.933	83.20%
13	1284	0.746	59.12%	680	0.848	65.06%	1328	0.923	77.25%
14	520	0.735	53.38%	1156	0.822	66.60%	1368	0.910	65.98%



(a) Minimum (b) Median (c) Maximum

Fig. 16 Tomographic images of a sample of sunflower corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

Table 11 Evaluation of tomographic images of chickpea samples: minimum, median, maximum SSIM values, and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
15	1296	0.760	56.97%	968	0.832	65.27%	652	0.914	69.26%
16	1868	0.744	54.30%	700	0.809	61.58%	1324	0.902	71.31%
17	188	0.752	58.50%	140	0.817	59.12%	488	0.876	69.77%
18	1464	0.744	54.82%	112	0.831	64.45%	1056	0.922	70.90%
19	320	0.746	59.22%	176	0.828	64.86%	1836	0.897	69.47%

to note that slice 1020 obtained the highest SSIM value from the sample set, but selected 21% of slice 1368 to obtain a higher SSIM value of 2.5%. In contrast, the reconstructed image that obtained the lowest SSIM value (0.734) selected more projections (73.98%). It is also worth noting that slice 500 selected less than 60% and obtained an SSIM value greater than 0.800.

Figure 16 presents the reconstructed images of the slices of an analyzed sample of sunflower.

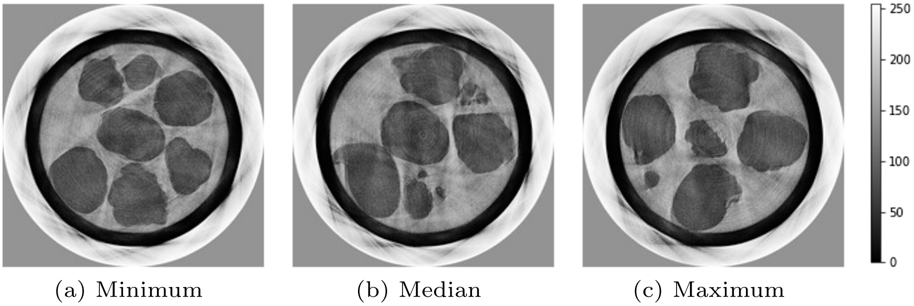
Table 11 presents the SSIM and the rate of selection of the projections of the set of chickpea seed samples.

In the chickpea sample group, the SSIM values were higher than 0.700 with selection rates between 54.30% and 71.31%. In the case of the lowest selection rate (54.30%), 530 projections which produced an SSIM value of 0.744 were selected. In contrast, the highest selection rate in the group (71.31%) obtained an SSIM of 0.902, that is, an SSIM value of 21.23% higher than the lowest selection rate using 31.32% more projections.

In this group, by increasing the selection rate even higher than 30%, a gain in SSIM of more than 20% was observed when the selection rate was increased to more than 30%, which can be understood as an expressive result. Figure 17 presents reconstructed images of the slices of an analyzed chickpea sample.

Table 12 presents the values obtained from the SSIM and, the selection rate of the projections from the set of wheat seed samples.

The wheat seed sample set showed higher selection rates than the chickpea samples, with the highest selection rate being 74.90%. In contrast, an SSIM value of 0.912 was observed



(a) Minimum (b) Median (c) Maximum

Fig. 17 Tomographic images of a sample of chickpeas corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

Table 12 Results of the evaluation of tomographic images of wheat samples: minimum, median, maximum SSIM values, and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
20	340	0.757	63.63%	636	0.859	71.31%	856	0.926	74.08%
21	1356	0.754	51.84%	148	0.821	62.81%	468	0.897	74.90%
22	324	0.751	57.07%	664	0.818	63.63%	1492	0.912	63.11%
23	1336	0.743	51.84%	264	0.819	57.17%	1784	0.910	68.14%
24	328	0.775	58.09%	552	0.832	65.98%	1784	0.906	71.31%

with 63.11% of the selected projections. In the median column, it was possible to observe that slice 264 obtained an SSIM value of 0.819 with a selection rate lower than 60%.

Figure 18 presents reconstructed images of the slices of an analyzed wheat sample.

The next set consisted of pumpkin seed samples. Table 13 presents the values obtained from the SSIM, and the selection rate of the projections.

Figure 19 presents reconstructed images of the slices of an analyzed pumpkin sample.

Table 13 shows that all slices obtained an SSIM value higher than 0.750. Table 14 presents the values obtained from the SSIM and the selection rate of projections from the set of samples of soybeans.

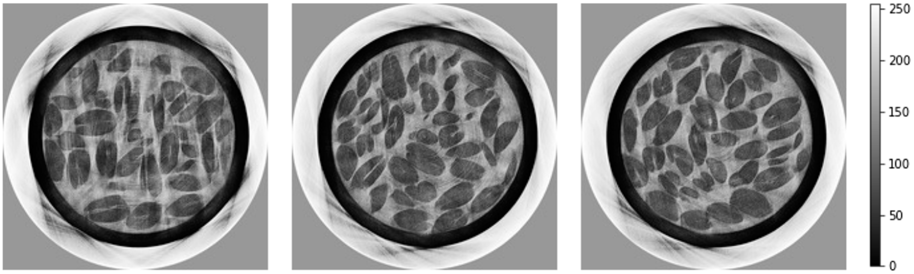
Table 14 shows that slice 948 obtained a higher SSIM value than slice 268, although both slices have similar selection rates. Figure 20 presents reconstructed images of the slices of an analyzed soybean sample.

The last set refers to the peanut seed samples. Table 15 presents the values obtained from the SSIM and, the selection rate of the projections.

Table 15 shows that slice 992 obtained a higher SSIM value (0.959) with selection rate of 74.08%, whereas the projections that obtained SSIM values of approximately 0.80 had selection rate around 63%.

Figure 21 presents reconstructed images of the slices of an analyzed peanut sample.

Figure 22, in the boxplot graphic, presents the calculation of the SSIM measurement for the slices of the seed samples, which total the reconstruction and analysis of slices, and provides a complete view of the SSIM for the image base used in this evaluation, after the analysis of each set of seed samples.



(a) Minimum (b) Median (c) Maximum

Fig. 18 Tomographic images of a sample of wheat corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

Table 13 Results of the evaluation of tomographic images of pumpkin samples: minimum, median, maximum SSIM values and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
25	1104	0.751	59.43%	388	0.821	63.01%	988	0.935	71.52%
26	1768	0.758	59.73%	668	0.834	67.93%	1144	0.964	80.74%
27	368	0.760	67.62%	444	0.847	67.83%	196	0.935	76.84%
28	1356	0.758	53.69%	172	0.839	66.80%	992	0.926	71.21%
29	1084	0.739	55.23%	72	0.831	67.21%	920	0.925	83.40%

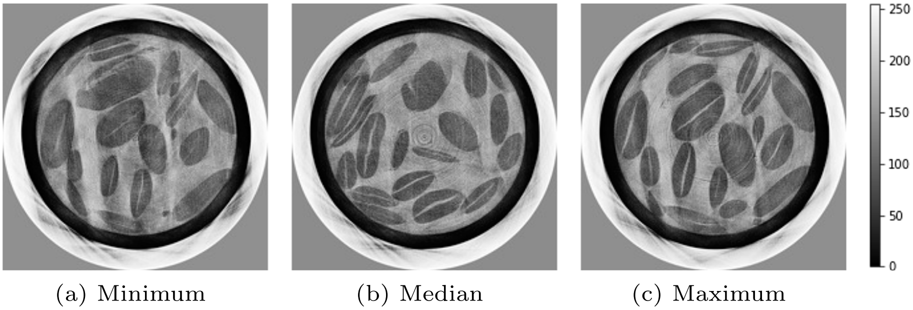
As mentioned in Section 3.4, for each considered sample having different agricultural seeds, we considered the CT image reconstruction, i.e., resulting in 498 slices for each of them. Such CT reconstructed slices were arranged in relation to the z-axis. In fact, the Fig. 23 presents the result of a statistical analysis to calculate the coefficient of linear regression (R^2) considering the results of SSIM as a function of the number of selected projections, i.e., for instance, selected from the central slice of each different seed sample that was evaluated. Further, the R^2 obtained was equal to 0.87. In the chart, each point is related to the central slice of each sample that has been considered for such analysis.

The next section presents the results obtained from the analysis of the PSNR measure for the set of 33 samples and their respective tomographic projection matrices.

4.5 Peak signal to noise ratio assessment

The objective of the PSNR measurement analysis was to observe the signal-to-noise ratio of the reconstructed image in comparison to the reference image to verify whether the selection of projections implied an increase in noise, compromising the 2D slice reconstruction stage.

Therefore, such analysis was motivated by the fact that, when projections were selected, the amount of data was less than expected, so the selection may have degraded the image by generating artifacts during the tomographic reconstruction. Table 16 presents the calculated PSNR values for the phantom, and the calculations for the samples are presented later.

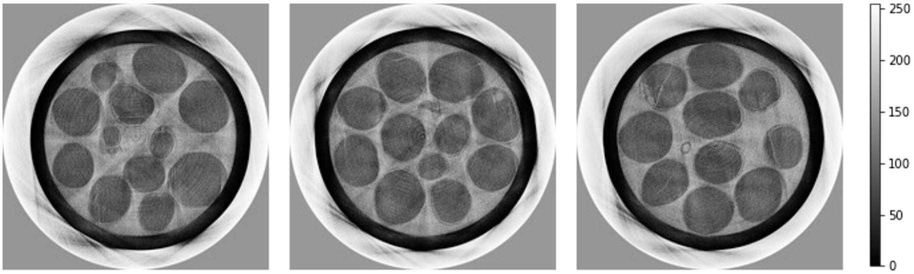


(a) Minimum (b) Median (c) Maximum

Fig. 19 Tomographic images of a sample of pumpkin corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

Table 14 Results of the evaluation of tomographic images of soybean samples: minimum, median, maximum SSIM values, and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
30	1664	0.751	57.48%	324	0.827	68.85%	848	0.910	76.43%
31	100	0.718	55.33%	180	0.830	63.01%	1252	0.907	78.18%
32	1804	0.701	66.29%	356	0.821	69.98%	1992	0.905	72.03%
33	980	0.625	62.40%	268	0.855	69.67%	1108	0.922	72.85%
34	140	0.768	60.96%	228	0.837	67.11%	948	0.912	69.88%

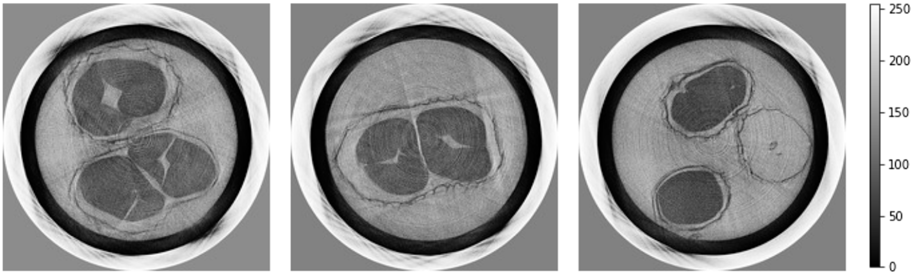


(a) Minimum (b) Median (c) Maximum

Fig. 20 Tomographic images of a sample of soybean corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

Table 15 Evaluation of tomographic images of peanut samples: minimum, median, maximum SSIM values, and rate of projection selection (Sel.) of the analyzed samples

	Minimum			Median			Maximum		
Sample	Slice	SSIM	Sel.	Slice	SSIM	Sel.	Slice	SSIM	Sel.
35	1640	0.762	60.86%	116	0.829	64.55%	992	0.959	74.08%
36	244	0.760	59.22%	124	0.817	61.48%	492	0.911	73.57%
37	1560	0.781	58.81%	104	0.829	64.45%	988	0.916	67.83%
38	1988	0.745	56.97%	316	0.826	65.37%	1436	0.905	66.39%



(a) Minimum (b) Median (c) Maximum

Fig. 21 Tomographic images of a peanut sample corresponding to the reconstructed slices, where the SSIM values observed were minimum, median, and maximum. All images are represented by a grayscale ranging from 0 to 255

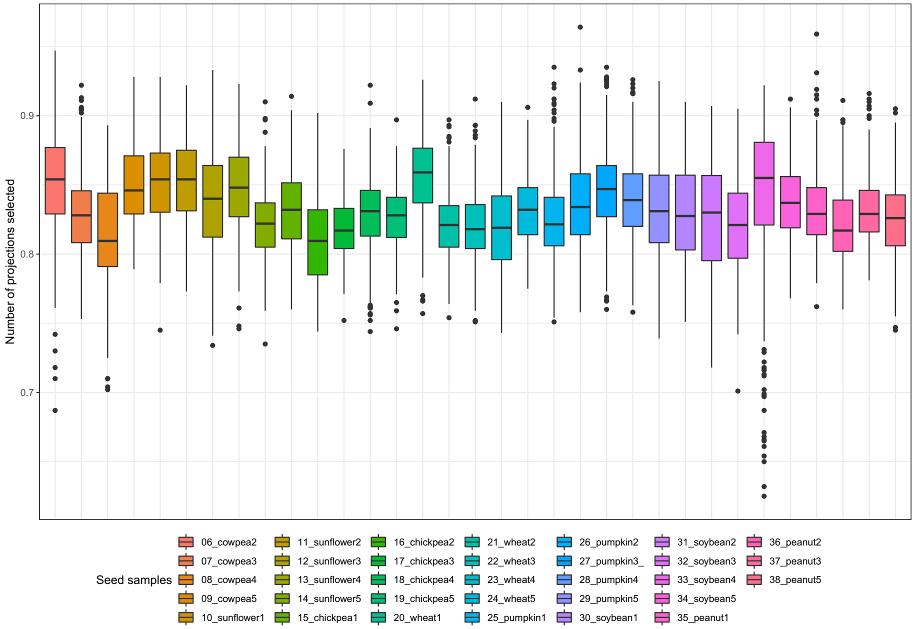


Fig. 22 SSIM analysis of the reconstructed images of the seeds set containing 33 samples

Figure 24 shows slice 1066, which refers to a minimum value of PSNR; 1928, which refers to the median value of PSNR; and 988, which refers to the maximum value of the PSNR.

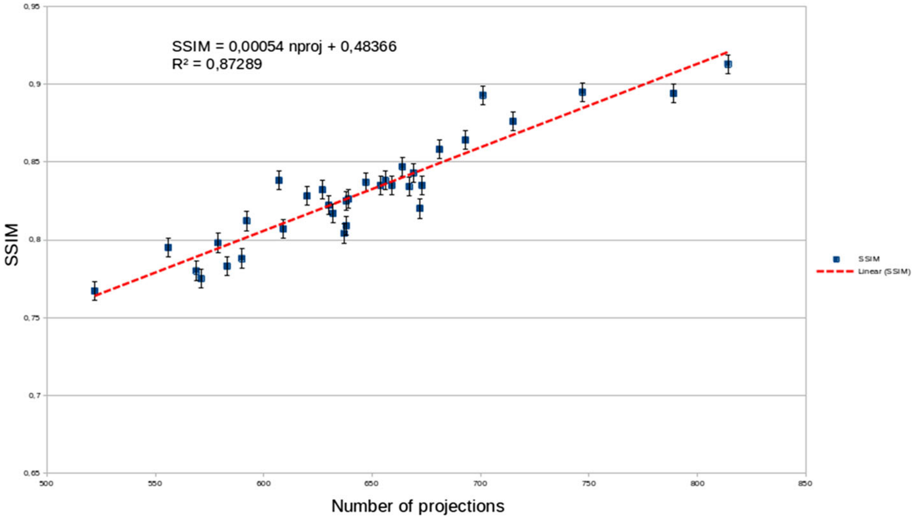


Fig. 23 Linear regression considering the results of SSIM as a function of the number of selected projections for each central slice from the 33 tomographic assay carried out for the samples having different agricultural seeds

Table 16 Minimum, median, and maximum PSNR values calculated for the phantom

Value	Slice	PSNR (dB)	Selection
Minimum	1066	16.953	64.14%
Median	1928	27.027	63.63%
Maximum	988	39.873	63.83%

Figure 25 presents the PSNR calculation for the analysis of the database used in this evaluation, which totaled 16.434 slices, distributed among 33 samples of agricultural seeds.

As shown in Fig. 25, the median of the PSNR for all samples was above 25 dB, which indicates that the process of projection selection did not compromise the quality of the 2D tomographic reconstruction.

4.6 Volumetric visualization of tomographic images

In this section, the results of the 3D visualization (volumetric) for the samples reconstructed in the big data environment are presented. As discussed, slices of the samples were divided into regions and organized into blocks that were interpolated to generate a portion of the volume or subvolume, of the sample. The subvolumes were then archived in Amazon's environment (AWS S3). Therefore, the process of viewing a sample consisted of two steps: (i) grouping the subvolumes to generate the complete volume of the sample; and (ii) viewing the complete volume using the `itkwidgets`[5] tool.

Both steps occurred outside the big data environment that was structured in this study and the following resources were used: Python language and the Jupyter notebook development environment[6] integrated with Kitware's VTK visualization library, through the itkwidgets plugin. The visualization was performed on a computer with 32 GB of RAM and Intel core i9 processor.

In the first stage, a Python script was prepared to recover the subvolumes stored in AWS S3. Each subvolume featured the identification of its position, so it was possible to organize the complete volume with this information. The second stage was to load the complete volume into memory, using the itkwidgets plugin features, and view it. Figure 26 shows a volumetric view of the phantom.

The *phantom* was reconstructed with 996 slices with size of 2000 × 2000 pixels. Of all the slices, 498 were real and 498 were virtual, generated by the interpolation process. Thus, entire volume produced 2000 × 2000 × 996 voxels, totaling 15 GB of data. Figure 27 shows the volumetric view of a sample of cowpeas.

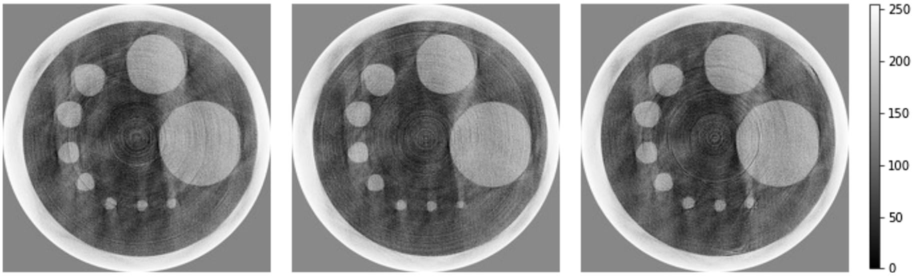
Similar to the *phantom*, the volume of a sample of cowpeas was reconstructed with a size 2000 × 2000 × 996, resulting in 15 GB of data.

5 Conclusion

This work presented a method of 2D and 3D (volumetric) reconstruction of high tomographic images of agricultural samples using big data techniques. An important aspect of the developed method was the parallelization strategy adopted for 2D reconstruction, which consisted

[5] Disponível em: https://github.com/InsightSoftwareConsortium/itkwidgets

[6] Available at: https://jupyter.org/



(a) Minimum (b) Median (c) Maximum

Fig. 24 Slices of phantom that represent the minimum, median and maximum values of the PSNR measure. All images are represented by the gray scale ranging from 0 to 255

of parallelizing the projection matrices, rather than the individual projections. This strategy allowed the reconstruction of agricultural samples, such as seeds, in a distributed big data environment.

For the execution of this customized developed method, it was necessary to structure a cluster of computers. The infrastructure used for this purpose was provided through the Amazon AWS service. In this context, 12 different cluster configurations are evaluated. In addition, the configuration that allowed not only the use of the greatest amount of tomographic but also the greatest efficiency was selected as the final arrangement for the developed method.

The configuration that prevailed was the one that contained six nodes, as it presented greater efficiency than the configuration that contained 10 nodes, despite the speedup value

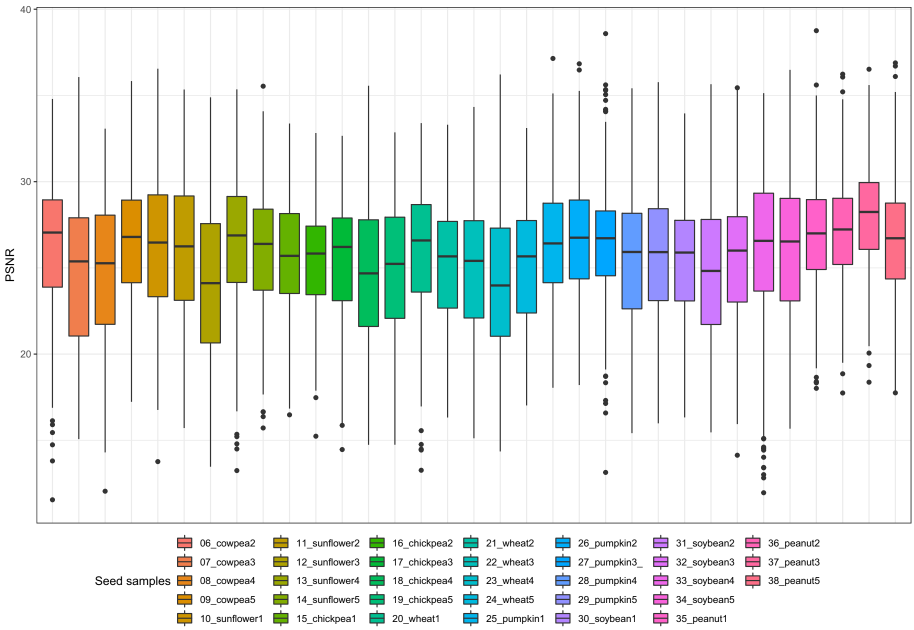


Fig. 25 PSNR analysis of the reconstructed images of the 33 agricultural seed samples

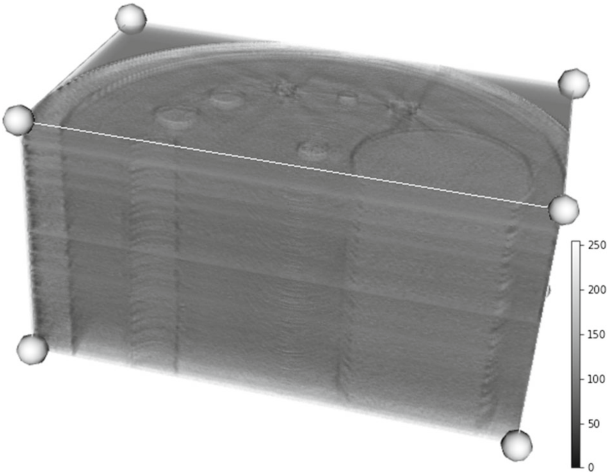


Fig. 26 Cut performed on phantom volumetric visualization

being lower. Therefore, it was possible to verify that a higher speedup value does not necessarily imply greater efficiency. Similarly, it should be noted that efficiency had values greater than 1. Such a result occurred because of better communication between processors and memories, and the use of distributed processing methods. In contrast, to obtain the processing time in parallel, Apache Spark loaded the projection matrices into memory, or a large part of them, having them distributed among the worker nodes. Additionally, the analysis of the cost of sequential/parallel processing corroborated the understanding that the efficiency measure represents more accurate information about the architecture of the environment.

Another relevant aspect presented in this paper is the selection of tomographic projections in each sinogram, the smallest number of projections that best represented the object in the reconstructed image. The energy information of each projection was considered to identify those that were more relevant for obtaining a tomographic reconstruction in two dimensions.

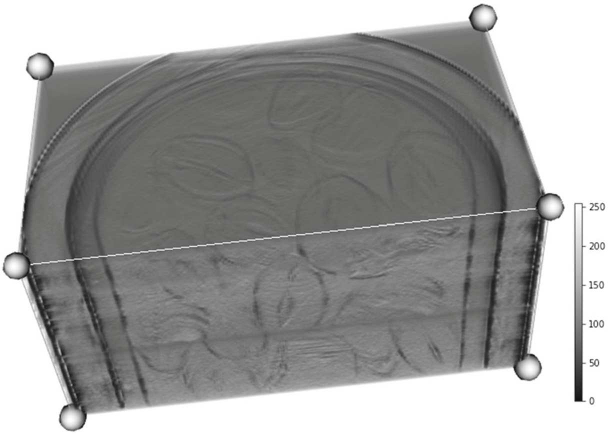


Fig. 27 Volumetric visualization of a sample of cowpeas with cut

Additionally, the organization in energy classes of the tomographic projections proved to be an adequate alternative because it considered the entire spectrum of energies contained in a sinogram. A set of 33 seed samples and a heterogeneous phantom of plexiglass were considered, which totaled $66,640$ projection matrices or 242 GB of tomographic data. A total of $16,932$ projection matrices were considered from this set, or 498 matrices per sample, for the purpose of evaluating the selection of tomographic projections and the quality of the 2D reconstruction. For $16,932$ matrices, the algorithm selected 61.47% to 71.72% of the projections, which implies that there was a reduction of approximately $28\% - 38\%$ by projection matrix analysis. The SSIM metric was calculated for each projection matrix, and the median SSIM value for each sample was observed. Thus, the SSIM analysis showed that tomographic reconstruction of the samples in two dimensions from the selected projections led the SSIM value to be higher than 0.800 for all the samples analyzed. It is also worth noting that the PSNR analysis corroborated the results obtained in the SSIM analysis. Therefore, the results showed that the reduction in the number of projections for the seed samples (i.e., peanut, cowpea, sunflower, chickpea, wheat, pumpkin, and soybean) did not compromise the structure of the information contained in the reconstructed images, as observed by the SSIM and PSNR analyses. The 3D reconstruction (volumetric) established conditions for evaluating seeds of different surfaces and shapes and did not restrict the analysis to flat seeds.

Finally, when considering that the volume of data in the agricultural area has increased considerably, such as in seed analysis, the integration of computer and electronic techniques becomes urgent to seek new solutions that are able to handle this new scenario and meet the needs of agricultural demands. Therefore, the method developed in this study has contributed to the execution of tomographic reconstruction to improve analysis and decision-making in agriculture. In addition, the main contribution of this study was to prepare a framework for the tomographic reconstruction of high-resolution images for the analysis of agricultural samples, which was prepared for execution in a big data environment.

The evaluation of additional strategies for parallelization of the projection matrices and evaluation of other statistical distribution models are proposed for future work, such as Chi-square distribution, which may be useful for the selection of tomographic projections.

Acknowledgements The authors acknowledge the support from Embrapa Instrumentation (CNPDIA), Federal Institute of São Paulo (IFSP), the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Post-Graduation Program in Computer Science of the Federal University of São Carlos (UFSCar).

Author Contributions This work was conducted in collaboration with both authors. Conceptualization, G. M. A., and P.E.C.; methodology, computational model, software and validation; also writing-original draft preparation.

Funding This work was partially supported by Embrapa Instrumentation (CNPDIA) in partnership with the São Paulo Research Foundation (Fapesp), and Federal Institute of São Paulo (IFSP), with grants MPI No. 547 01.14.09.0.01.05.05, Process No. 17/19350-2, and Process No. 23311.000102 /2015-10 respectively.

Data availability statement The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmed MR, Yasmin J, Wakholi C, Mukasa P, Cho B-K (2020) Classification of pepper seed quality based on internal structure using x-ray CT imaging 179:105839. https://doi.org/10.1016/j.compag.2020.105839
2. Ahmed MR, Yasmin J, Collins W, Cho B-K (2018) X-ray ct image analysis for morphology of muskmelon seed in relation to germination. Biosyst Eng 175:183–193. https://doi.org/10.1016/j.biosystemseng.2018.09.015
3. Alves GM, Cruvinel PE (2018) Big data infrastructure for agricultural tomographic images reconstruction. 2018 IEEE 12th International Conference on Semantic Computing (ICSC). https://doi.org/10.1109/icsc.2018.00071
4. Balogun F, Cruvinel P (2003) Compton scattering tomography in soil compaction study. Nucl Inst Methods Phys Res A: Accelerators, Spectrometers, Detectors and Associated Equipment 505(1–2):502–507. https://doi.org/10.1016/s0168-9002(03)01133-1
5. Beutler FJ, Leneman OA (1966) Random sampling of random processes: Stationary point processes. Inf Control 9(4):325–346. https://doi.org/10.1016/s0019-9958(66)80001-3
6. Bronson K, Knezevic I (2016) Big data in food and agriculture. Big Data and Society 3(1). https://doi.org/10.1177/2053951716648174
7. Cruvinel P, Cesareo R, Crestana S, Mascarenhas S (1990) X- and gamma-rays computerized minitomograph scanner for soil science. IEEE Trans Instrum Meas 39(5):745–750. https://doi.org/10.1109/19.58619
8. Cruvinel P, Pereira M, Saito J, Costah LDF (2009) Performance improvement of tomographic image reconstruction based on DSP processors. IEEE Trans. Instrum. Meas. 58(9):3295–3304. https://doi.org/10.1109/tim.2009.2022378
9. Ding C, Wang W, He H, Yang W (2020) Research on tomographic image reconstruction algorithms based on fixed-point rotation x-CT system 79(35–36):25463–25496. https://doi.org/10.1007/s11042-020-08861-2
10. Diniz PSR, Silva EAB, Netto SL (2010) Digital Signal Processing. Cambridge University Press
11. Ditter A, Fey D, Schon T, Oeckl S (2014) On the way to big data applications in industrial computed tomography 792–793. https://doi.org/10.1109/bigdata.congress.2014.125
12. Hajjaji Y, Boulila W, Farah IR, Romdhani I, Hussain A (2021) Big data and IoT-based applications in smart environments: A systematic review 39:100318. https://doi.org/10.1016/j.cosrev.2020.100318
13. Heeraman D, Hopmans J, Clausnitzer V (1997) Three dimensional imaging of plant roots in situ with x-ray computed tomography. Plant Soil 189(2):167–179. https://doi.org/10.1023/b:plso.0000009694.64377.6f
14. Hsieh J (2009) Computed Tomography: Principles, Design, Artifacts, and Recent Advances. John Wiley & Sons Inc
15. Janßen R (1987) A note on superlinear speedup 4(2):211–213. https://doi.org/10.1016/0167-8191(87)90053-6
16. Kak AC, Slaney M (1989) Principles of Computerized Tomographic Imaging. IEEE Press
17. Kamilaris A, Kartakoullis A, Prenafeta-Boldu FX (2017) A review on the practice of big data analysis in agriculture. Comput Electron Agric 143:23–37. https://doi.org/10.1016/j.compag.2017.09.037
18. Kontoghiorghes EJ (2005) Handbook of Parallel Computing and Statistics. Chapman and Hall/CRC
19. Liakos K, Busato P, Moshou D, Pearson S, Bochtis D (2018) Machine learning in agriculture: A review. Sensors 18(8):2674. https://doi.org/10.3390/s18082674
20. Naime JM (1994) Projeto e construção de um Tomógrafo portátil para estudos de ciência do solo e plantas, em campo. Master's thesis, USP
21. Oppenheim AV, Schafer RW (1975) Digital Signal Processing. Prentice Hall
22. Pereira M, Cruvinel P (2015) A model for soil computed tomography based on volumetric reconstruction, wiener filtering and parallel processing. Comput Electron Agric 111:151–163. https://doi.org/10.1016/j.compag.2014.12.006
23. Pires LF, Bacchi OOS (2010) Mudanças na estrutura do solo avaliada com uso de tomografia computadorizada. Pesq Agrop Brasileira 45(4):391–400. https://doi.org/10.1590/s0100-204x2010000400007
24. Pires LF, Borges JA, Bacchi OO, Reichardt K (2010) Twenty-five years of computed tomography in soil physics: A literature review of thebrazilian contribution. Soil Tillage Res 110(2):197–210. https://doi.org/10.1016/j.still.2010.07.013
25. Rangayyani RM (2004) Biomedical Image Analysis (Biomedical Engineering). CRC Press
26. Ribarics P (2016) Big data and its impact on agriculture. Ecocycles 2(1):33–34. https://doi.org/10.19040/ecocycles.v2i1.54
27. Scannavino FA (2013) Tomógrafo de espalhamento Compton para estudos da física de solos agrícolas em ambiente de campo. Ph.D. thesis, USP

28. Serrano E, Garcia-Blas J, Carretero J, Desco M, Abella M (2020) Accelerated iterative image reconstruction for cone-beam computed tomography through big data frameworks 106:534–544. https://doi.org/10.1016/j.future.2019.12.042
29. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
30. Shannon C (1949) Communication in the presence of noise. Proc IRE 37(1):10–21. https://doi.org/10.1109/jrproc.1949.232969
31. Silva AM (1997) Construção e uso de um tomógrafo com resolução micrométrica para aplicações em ciências do solo e ambi. Ph.D. thesis,USP
32. Ullah R, Arslan T (2020) PySpark-based optimization of microwave image reconstruction algorithm for head imaging big data on high-performance computing and google cloud platform 10(10):3382. https://doi.org/10.3390/app10103382
33. Verdu S (1998) Fifty years of shannon theory. IEEE Trans Inf Theory 44(6):2057–2078. https://doi.org/10.1109/18.720531
34. Wang G (2016) A perspective on deep imaging. IEEE Access 4:8914–8924. https://doi.org/10.1109/access.2016.2624938
35. Zhang H et al (2016) Image Prediction for Limited-angle Tomography via Deep Learning with Convolutional Neural Network. ArXiv e-prints. arXiv:1607.08707 [physics.med-ph]
36. Zhao J, Fu Y, Tan Y, Cao F (2013) A reduction algorithm for the big data in 3D surface reconstruction. https://doi.org/10.1109/smc.2013.824

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.