Stochastic Modeling of Multidimensional Particle Properties Using Parametric Copulas

Orkun Furat[1,*], Thomas Leißner[2], Kai Bachmann[3], Jens Gutzmer[3], Urs Peuker[2], Volker Schmidt[1]

[1]*Institute of Stochastics, Ulm University, D-89069 Ulm, Germany*
[2]*Institute of Mechanical Process Engineering and Mineral Processing, Technische Universität Bergakademie Freiberg, D-09599 Freiberg, Germany*
[3]*Helmholtz Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, D-01328 Dresden, Germany*

Abstract

In this paper, prediction models are proposed which allow the mineralogical characterization of particle systems observed by X-ray micro tomography (XMT). The models are calibrated using 2D image data obtained by a combination of scanning electron microscopy (SEM) and energy dispersive X-ray spectroscopy (EDS) in a planar cross-section of the XMT data. To reliably distinguish between different minerals the models are based on multidimensional distributions of certain particle characteristics describing, e.g., their size, shape and texture. These multidimensional distributions are modeled using parametric Archimedean copulas which are able to describe the correlation structure of complex multidimensional distributions with only a few parameters. Furthermore, dimension reduction of the multidimensional vectors of particle characteristics is utilized to make non-parametric approaches like the computation of distributions via kernel density estimation viable. With the help of such distributions the proposed prediction models are able to distinguish between different types of particles among the entire XMT image.
Keywords and Phrases: X-ray micro tomography (XMT), mineral liberation analyzer (MLA), stereology, multidimensional particle characterization, parametric copula

1 Introduction

Processes which separate mixtures of particles based on criteria like size, shape or chemical composition of particles are used in many applications. Often the separation quality can have a critical influence on subsequent processing steps. For example, in the mining

* Corresponding author. E-mail address: `orkun.furat@uni-ulm.de`, Phone: `+49731/50 23555`, *Fax:* `+49731/50 23649`

industry an essential step of ore dressing deals with separation processes which remove unwanted minerals from a system of particles with sizes smaller than 1 mm such that minerals of interest remain. In order to evaluate the separation success a mineralogical characterization of the particles prior and after separation is necessary. Common methods to characterize particles with respect to size, shape and composition rely on two-dimensional (2D) techniques describing a three-dimensional (3D) reality. For example, a combination of scanning electron microscopy (SEM) and energy dispersive X-ray spectroscopy (EDS) can be used (Sunderland and Gottlieb, 1991) to achieve a mineralogical characterization of ore samples. However, SEM-EDS is limited to 2D profiles of samples, which makes the characterization of a whole 3D sample rather expensive or even impossible - at least in a non-destructive manner. Another approach, namely X-ray micro tomography (XMT) provides 3D image data depicting the morphology of particles in a sample. This allows a quantitative analysis of particle systems without the stereological bias of 2D measurement techniques (Wang et al., 2017; Reyes et al., 2018). Recently, Su and Yan (2018) characterized sand particles observed in XMT with various shape descriptors and utilized spherical harmonic functions for both a parametric representation of single particles and a basis for stochastic modeling, see also Feinauer et al. (2015). Besides the morphology of particles, XMT provides information about local material specific constants. More precisely, the grayscale values of XMT images are related to the mass density of the observed material. When the considered minerals have distinct mass densities, the contrast-information from XMT can be utilized to distinguish between them. However, Furat et al. (2018) showed that grayscale values suffice only in a limited way to characterize the observed material in XMT data.

Therefore, for achieving a 3D mineralogical characterization of a sample it is necessary to utilize more information from the XMT data than solely the grayscale values. If we know, for example, that particles composed from a certain type of mineral are more spherical than others, we can utilize the additional information about their shape, which can be obtained from XMT as well, to characterize them.

In the present paper, we consider a sample of milled greisen-type Li-ore which mainly comprises of the minerals zinnwaldite and quartz, though these particles can have imperfections consisting of further components like topaz, muscovite, kaolinite and others, see Leißner et al. (2016). Details on the sample preparation as well as on the analysis of the epoxy block by XMT and SEM-EDS can be found in Furat et al. (2018). We propose prediction models which can characterize 3D particles from XMT based on their size, shape and grayscale values. They can reliably distinguish between consolidated particles which are dominated by either quartz or zinnwaldite grains. The calibration of these prediction models requires SEM-EDS data from only one planar cross-section of the sample. In order to be able to compute size and shape characteristics of XMT image data a particle-wise segmentation of the image data is required, such that it is possible to identify single particles in the XMT image. First, we give a short overview of the image processing steps which were necessary to obtain such a segmentation of our image data. Furthermore, we present several characteristics which can describe the size, shape and grayscale texture of particles. The mineralogical characterization of 3D particles from XMT images will be done based on these characteristics. Additionally to the XMT image, we have SEM-EDS data of the same sample, which lies in a planar cross-section of the XMT image. By utilizing the mineralogical characterization of SEM-EDS we know the mineralogical composition of the 3D particles that hit this cross-section. Therefore, we are able to link the size, shape and grayscale characteristics of these 3D particles to different minerals. The prediction models, proposed in the present paper, require the probability density functions of the considered particle characteristics, for each type of mineral. Since multiple characteristics are necessary to characterize particles, because only grayscale information of the particles does not suffice for characterization, we need multidimensional probability densities of the considered characteristics. An easy way to determine probability densities are so-called kernel density estimators (Scott, 2015), since they do not require the search for a suitable parametric family of distributions (e.g. normal distribution, exponential distribution, etc.). However, due to the "curse of dimensionality", kernel density estimators require huge sample sizes for determining multidimensional densities which can limit their practicality. Therefore, we rigorously show how to determine multidimensional densities by fitting parametric families of distributions to the data. At first we fit one-dimensional parametric densities to single characteristics and, in a second step, we use so-called Archimedean copulas (Nelsen, 2006) to determine joint densities of the considered particle characteristics. As an alternative, we utilize dimension reduction of our multidimensional particle characteristics, in order to make kernel density estimation more viable. With the help of the fitted or estimated multidimensional densities we propose prediction models which misclassify only $3-6\%$ of particles either dominated by zinnwaldite or quartz. In a further step, we validate the prediction models by comparing their predictions with the characterization of SEM-EDS data at a spatially different cross-section which was not used for the calibration of the prediction models.

The rest of this paper is organized as follows. In Section 2.1, we briefly recall some techniques of image processing and segmentation which we recently used in Furat et al. (2018). Then, in Section 2.2, various particle characteristics are explained, which are stochastically modeled later on in Section 2.3. Therefore, in Sections 2.3.2 to 2.3.5, particular emphasis is put on the explanation of the copula approach to parametric modeling of multidimensional probability densities. Section 2.3.6 deals with dimension reduction of data and kernel density estimation, as an alternative to the copula approach. In Section 3 the classification models utilizing either copulas or kernel density estimation are compared and validated. Finally, Section 4 concludes.

2 Materials and Methods

2.1 Image processing

2.1.1 3D XMT data preprocessing and segmentation

In this section, we give a short overview of the image preprocessing steps and the particle-wise segmentation of the 3D XMT data considered in the present paper, for more details see Furat et al. (2018). The first step consists of the reduction of noise in the XMT data with a non-local means denoising algorithm, see Buades et al. (2005). In contrast to the commonly used Gaussian filter, this nonlinear image filter has the advantage of smoothing homogeneous regions while preserving edges. In order to separate the particles from the background we use the local adaptive Sauvola thresholding algorithm, which determines binarization thresholds for each voxel based on its local neighborhood, see Shafait et al. (2008). Such local thresholding techniques can produce better results than global thresholds, due to globally inconsistent grayscale values in XMT images. One crucial and nontrivial task in image processing is the particle-wise segmentation, which allows the extraction of single particles from the image data for quantitative analysis.

In a first step, we use a marker-based watershed algorithm, see Spettl et al. (2015), to get an initial segmentation. This results into some particles being wrongly separated into multiple fragments. To remove such oversegmentations we train a neural network (Hastie et al., 2009) to decide whether adjacent segments should be merged or not. To make this decision the neural network is supplied with information regarding the local morphology and grayscale values around two adjacent fragments. By applying the neural network as a post-processing step on the initial segmentation, we receive our final segmentation, see Figure 1b.

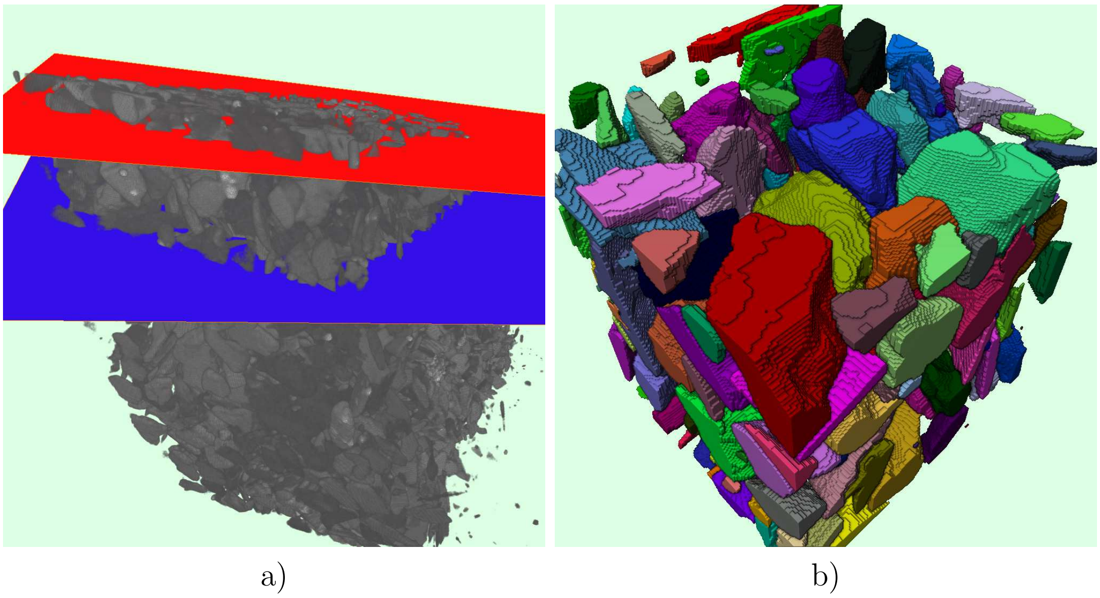


Figure 1: a) Volumetric XMT data (gray) with two registrated 2D SEM-EDS planes (blue and red). b) Cut-out of the particle-wise segmentation of the XMT data.

2.1.2 2D SEM-EDS data and registration

In addition to the 3D XMT image data, SEM-EDS data was considered by Furat et al. (2018). The latter was obtained with a mineral liberation analyzer (MLA), from the same sample at two spatially different planar sections. An illustration of the SEM-EDS data can be seen in Figure 2a. It provides 2D information about the morphology of the particles in planar sections, but in contrast to the XMT data, see Figure 2b, it also provides information about the mineralogical composition of particles. For example, in the false color image of Figure 2a the blue phase indicates zinnwaldite.

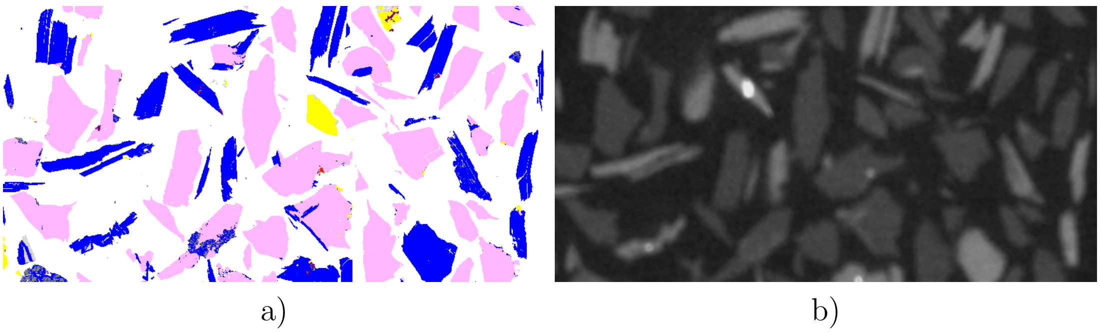


Figure 2: a) 2D SEM-EDS image, the blue phase indicates zinnwaldite and pink indicates quartz. b) The corresponding section registrated in the 3D XMT data.

Thus, by localizing the 2D SEM-EDS data in the segmented 3D image (registration), see Figure 1a, we have knowledge about the mineralogical composition of each 3D particle that intersects with the planar SEM-EDS data. Due to the particle-wise segmentation described in Section 2.1.1 we also know the 3D morphology of each of these particles. Furthermore, we can link the particle-wise segmentation with the grayscale values of the 3D XMT image, since the grayscale values provide valuable information about the composition of particles, as show in Furat et al. (2018). Therefore, both the grayscale values and the morphology of particles in the XMT image that hit the SEM-EDS plane allow us to adjust a classifier that can predict the mineralogical composition of an arbitrary particle solely based on information gained from the XMT image.

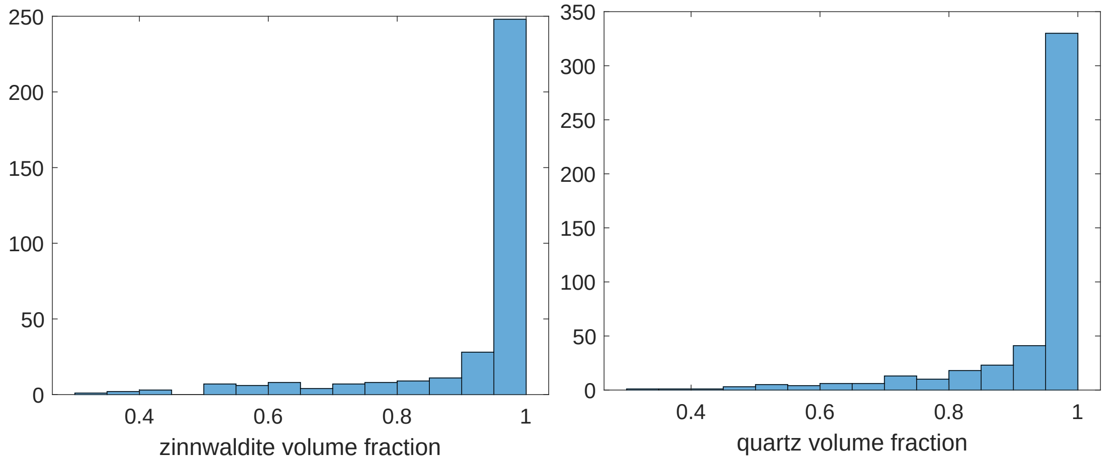
To be precise, let $I\colon W \subset \mathbb{Z}^3 \to [0,1]$ be the grayscale XMT image, where W is a cuboid observation window. The 2D SEM-EDS data can be regarded as a map $L\colon H \cap W \subset \mathbb{Z}^3 \to \{0,1,2,\dots\}$, where H is a set of voxels representing a plane and the values of L indicate different minerals or the background, e.g., we put $L(x) = 0$ if background was observed at $x \in H \cap W$, or $L(x) = 1$ if zinnwaldite was observed at $x \in H \cap W$. Furthermore, let $P_1, \dots, P_n \subset W \subset \mathbb{Z}^3$ be the sets of voxels corresponding to the particles that are obtained by the particle-wise segmentation of the 3D XMT image and that hit the plane H, i.e., $P_k \cap H = \{x_1^{(k)}, \dots, x_{\ell_k}^{(k)}\} \neq \emptyset$ for each $k = 1, \dots, n$. Each particle P_k gets the label

$$L(P_k) = \operatorname{mode}\bigl(L(x_1^{(k)}), \dots, L(x_{\ell_k}^{(k)})\bigr), \tag{1}$$

where the mode of the sample of labels $\bigl(L(x_1^{(k)}), \dots, L(x_{\ell_k}^{(k)})\bigr)$ is the label that appears most often. Thus, we assign to each particle P_k the mineral that is observed most frequently in the intersecting plane H. This means, for example, that a composite particle P_k which consists of zinnwaldite, quartz and other minerals will be referred to as a zinnwaldite composite particle if its main component in the intersection $P_k \cap H$ is zinnwaldite. Since we make this labeling only based on the mineralogical observation in the plane H we assume stationarity of the mineralogical composition of the particles P_k, i.e., that the composition of a particle P_k outside of the plane H is adequately represented by $P_k \cap H$. In our data the set of particles observed in the plane H can be decomposed in three disjoint sets

$$\{P_1, \dots, P_n\} = Z \cup Q \cup O \tag{2}$$

where $Z = \{P_k\colon L(P_k) = 1\}$ are the zinnwaldite composite particles and $Q = \{P_k\colon L(P_k) = 2\}$ are the quartz composite particles. The set $O = \{P_k\colon L(P_k) \geq 3\}$ contains the remaining particles that were observed in H. In the considered plane H contains 861 particles, from which 342 belong to the set of zinnwaldite composite particles Z and 462 to the set of quartz composite particles Q. Since the set O contains only 57 particles, we disregard these observations from now on. Note that due to the labeling done according to (1), a particle $P \in Z$ not necessarily consists solely of zinnwaldite. Figure 3a visualizes the distribution of the zinnwaldite volume fraction for particles labeled as zinnwaldite composite particles in the cross-section $P \cap H$. Analogously, the quartz particles $P \in Q$ are consolidated, see Figure 3b. The goal of the present paper is to develop a method which allows us to predict the main mineralogical component $L(P)$ of particles when the additional planar SEM-EDS data is not available.



a) b)

Figure 3: a) Histogram of the zinnwaldite volume fractions for particles for which the majority mineral phase is zinnwaldite. The mean zinnwaldite volume fraction is 0.925 *with a variance of* 0.018*. b) Histogram of the quartz volume fractions for particles for which the majority mineral phase is quartz. The mean quartz volume fraction is* 0.937 *with a variance of* 0.013*.*

2.2 Particle characteristics

In Section 2.3 below, we will describe a method which allows us to determine the mineralogical characterization of particles solely based on XMT information. This will be done by prediction models, which determine the mineralogical composition of a particle based on grayscale values and shape/size characteristics obtained from XMT data and its particle-wise segmentation. In order to adjust such a prediction model we use the ground truth information from a given planar SEM-EDS section, i.e., the sets of particles Z and Q, for which we know that they consist in majority of zinnwaldite and quartz, respectively. After adjusting the prediction model we still used another spatially separated SEM-EDS section to validate the prediction model. In order to make such predictions, we first introduce some particle descriptors, by which a particle $P \subset \mathbb{Z}^3$ will be characterized.

2.2.1 Size characteristics

Relevant size descriptors of a particle $P' \subset \mathbb{R}^3$, which we only observe on the grid as $P \subset \mathbb{Z}^3$, are the volume $\nu_3(P')$ and the surface area $a(P')$. It is clear that the volume can be estimated by counting voxels belonging to the set P and the surface area can be estimated by considering suitably defined voxel configurations, see Schladitz et al. (2006). We denote the estimated particle volume and surface area by $\nu_3(P)$ and $a(P)$ respectively.

For our prediction models we use the following estimator of the volume equivalent radius

$$r(P) = \sqrt[3]{\frac{3}{4\pi}\nu_3(P)}, \tag{3}$$

which is in a one-to-one relationship with the volume. The surface area $a(P)$ will not be directly regarded as a particle characteristic, but it is required for the sphericity factor which is a shape characteristic considered in the next section.

2.2.2 Shape characteristics

In the previous section we only considered the volume equivalent radius r as size characteristic, but this is not enough to reliably distinguish between zinnwaldite and quartz composite particles, since particles of similar sizes can have very different appearances. Therefore we examine some further characteristics for describing the shape of particles. One of the shape characteristics we consider is the sphericity factor

$$s(P) = \frac{\sqrt[3]{36\pi\nu_3^2(P)}}{a(P)}, \tag{4}$$

which takes values in $[0,1]$, where $s(P') = 1$ holds if $P' \subset \mathbb{R}^3$ is a sphere.

Another shape characteristic of interest is the convexity factor

$$c(P) = \frac{\nu_3(P)}{\nu_3(q(P))}, \tag{5}$$

where $q(P) \subset \mathbb{Z}^3$ is the convex hull of P on the lattice $\mathbb{Z}^3$. Like the sphericity factor, the convexity factor takes values in $[0,1]$, where $c(P) = 1$ holds if and only if P is convex on $\mathbb{Z}^3$. There is a causality between the sphericity factor s and the convexity factor c since more spherically shaped particles also have higher convexity factors. Yet, a convexity factor of 1 does not necessarily imply a large sphericity factor, since, e.g., ellipsoids have always a convexity factor of 1, yet their sphericity factor can be arbitrary close to 0. Since we observe a large number of flat particles in our image data, we also consider characteristics that quantify elongation of particles. For each particle $P \subset \mathbb{Z}^3$ with the barycenter $\overline{x} = (\overline{x}_1, \overline{x}_3, \overline{x}_3) \in \mathbb{R}^3$, whose coordinates are given by

$$\overline{x}_i = \frac{1}{\nu_3(P)} \sum_{x \in P} x_i \qquad \text{for } i = 1, 2, 3, \tag{6}$$

where $x = (x_1, x_2, x_3)$, we consider the (positive-semidefinite) covariance matrix

$$C = \Big(\frac{1}{\nu_3(P)} \sum_{x \in P} \big(x_i - \overline{x}_i\big)\big(x_j - \overline{x}_j\big)\Big)_{i,j=1,2,3}. \tag{7}$$

The eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3$ of C, where we assume that $\lambda_3 > 0$, are closely related to the axis lengths of the best fitting ellipsoid $\varepsilon(P)$ corresponding to the particle P, i.e., the axis lengths a_1, a_2, a_3 of $\varepsilon(P)$ are given by $a_i = \gamma\sqrt{\lambda_i}$ for $i = 1, 2, 3$, where γ is some scaling factor. The elongation factor $e(P)$ of the particle P is then given by

$$e(P) = \frac{a_1}{a_3} = \sqrt{\frac{\lambda_1}{\lambda_3}}. \tag{8}$$

Note that, like the previously considered characteristics, the elongation factor $e(P)$ is normalized and takes values in $[0,1]$, where values close to 0 indicate elongated particles like rod- or plate-shaped particles. By additionally analyzing the relationship between λ_2 and λ_3 it is possible to distinguish between rods and plates, but since we observed no rod-like particles in our data set, this was not necessary in the present paper.

2.2.3 Grayscale characteristics

The XMT data does not only provide information about the 3D morphology of particles. In Furat et al. (2018) we have seen that the grayscale values of XMT images contain substantial information about the mineralogical composition of particles. Recall that the grayscale values of a particle $P = \{x_1, \ldots, x_n\} \subset W$ are given by $y_1 = I(x_1), \ldots, y_n = I(x_n) \in [0, 1]$, where I is the XMT image. Furthermore, we assume without loss of generality that the grayscale values are ordered, i.e., $y_1 \leq y_2 \leq \ldots \leq y_n$. This allows us to define the following grayscale characteristics.

A natural choice could be the mean

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k. \tag{9}$$

But, since particles often have some imperfections, like small regions of high density, see Figure 2, the mean is not well suited for representing the dominant grayscale value. Therefore, we used the more robust median

$$m(P) = y_{0.5} \tag{10}$$

as grayscale characteristic, where the median is defined by the empirical quantiles of the sample

$$y_p = \begin{cases} y_{[np]+1}, & \text{if } np \in \mathbb{N}, \\ (y_{[np]} + y_{[np]+1})/2, & \text{if } np \notin \mathbb{N}, \end{cases} \tag{11}$$

for $p = 0.5$. A natural choice for measuring the variability of grayscale values of a particle could be the sample standard deviation

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}. \tag{12}$$

Similarly to the mean, the standard deviation is not a robust particle descriptor. Therefore we use the interquartile range

$$iqr(P) = y_{0.75} - y_{0.25} \tag{13}$$

to characterize the variability of the grayscale values of the dominant grayscale region of the particle. Further characteristics that describe the texture of particles (Shapiro and Stockman, 2001) will be investigated in a forthcoming paper.

2.3 Stochastic modeling of particle characteristics

In this section we fit parametric distributions to the particle characteristics discussed above. Moreover, we fit parametric distributions separately for zinnwaldite and quartz composite particles, since the mineralogical composition of particles that intersect with the SEM-EDS data is known by the labeling considered in (1). In Figure 4 the fitted one-dimensional distributions of the median grayscale value for both zinnwaldite and quartz are visualized.

Note that the large variance of the median grayscale values of zinnwaldite composite particles, in comparison to the median grayscale values of quartz composite particles, can have multiple reasons. One reason might be the variable chemical composition of zinnwaldite, see Rieder et al. (1970), which can lead to said increase of the variance. Another reason for a broader median distribution are partial volume effects of XMT data (Boas and Fleischmann, 2012). More precisely, Figure 5 indicates that zinnwaldite composite particles have a smaller sphericity than quartz composite particles which in turn implies that the former have a relatively large surface area or more surface voxels in the XMT image. Due to partial volume effects the grayscale values of such surface voxels are influenced by neighboring particles or the background. Nevertheless, we can see in Figure 4 that both minerals have quite distinct distributions. Still, there are overlapping regions which make a classification merely based on these distributions difficult. To be more precise, let $f^z_{\mathrm{m}}, f^q_{\mathrm{m}}$ be the fitted probability density functions of the particle-wise median grayscale values of zinnwaldite and quartz, respectively. Using a likelihood approach for classification, we say that a particle $P \subset \mathbb{Z}^3$ with median grayscale value $x = m(P) \in [0, 1]$ is mainly composed of zinnwaldite if

$$f^z_{\mathrm{m}}(x) > f^q_{\mathrm{m}}(x), \tag{14}$$

otherwise we say that P is a quartz composite particle. The probability of misclassifying zinnwaldite composite particles is then given by

$$\mathbb{P}\big(f^z_{\mathrm{m}}(X) \leq f^q_{\mathrm{m}}(X)\big) = \int_0^1 f^z_{\mathrm{m}}(x)\mathbb{1}_{f^z_{\mathrm{m}}(x)\leq f^q_{\mathrm{m}}(x)}\mathrm{d}x, \tag{15}$$

where X is a random median grayscale value with probability density f^z_m and $\mathbb{1}_{f^z_{\mathrm{m}}(x)\leq f^q_{\mathrm{m}}(x)}$ denotes the indicator function, i.e.,

$$\mathbb{1}_{f^z_{\mathrm{m}}(x)\leq f^q_{\mathrm{m}}(x)} = \begin{cases} 1, & \text{if } f^z_{\mathrm{m}}(x) \leq f^q_{\mathrm{m}}(x), \\ 0, & \text{if } f^z_{\mathrm{m}}(x) > f^q_{\mathrm{m}}(x). \end{cases} \tag{16}$$

This misclassification probability corresponds to the blue area in Figure 4 and has the size of 0.19. Furthermore, Table 1 shows a confusion matrix of the prediction rule given in (14), where we can see that 73 out of 342 zinnwaldite composite particles in the SEM-EDS plane were wrongly classified as quartz composite particles.

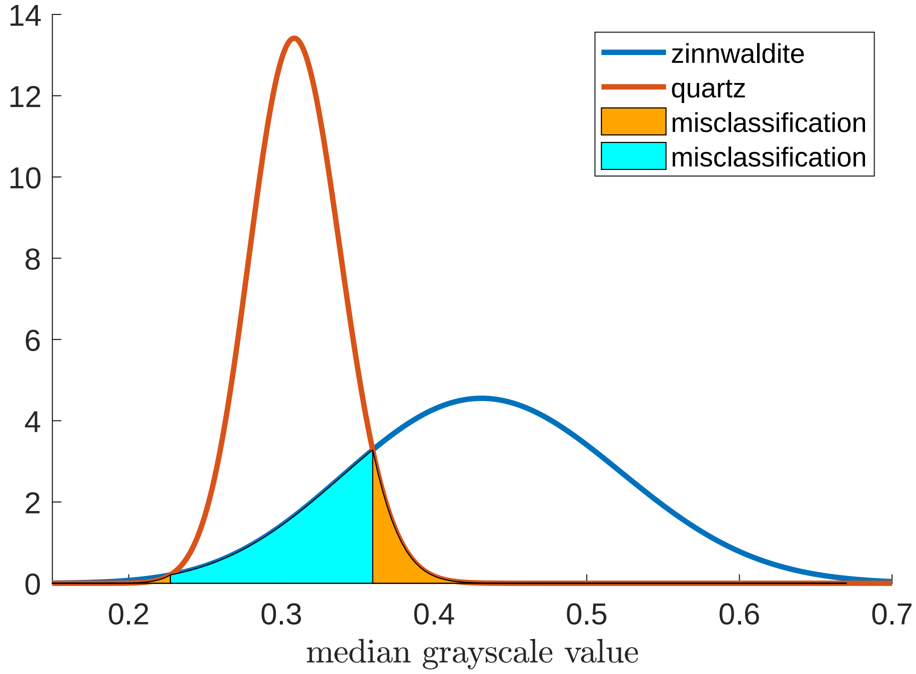


Figure 4: One-dimensional probability densities of the median grayscale value of zinnwaldite (blue curve) and quartz (red curve) particles. The empirical densities were fitted with Beta distributions, see Section 2.3.1. The orange area provides the probability of classifying a quartz composite particle as zinnwaldite, whereas the blue area gives the probability of classifying zinnwaldite as quartz.

Table 1: Confusion matrix for predicting the particle composition based on one-dimensional densities of the median grayscale value, where the particles observed in the SEM-EDS data have been used for adjusting the prediction model.

	zinnwaldite	**quartz**
predicted zinnwaldite	269	26
predicted quartz	73	436

In order to reduce the misclassification probabilities we additionally consider multi-dimensional distributions of vectors of particle characteristics. For example, in Figure 5 the joint probability density of the sphericity factor and the median grayscale value is visualized for zinnwaldite (left) and quartz (right) particles. We can see that these two-dimensional distributions are much more distinct, in comparison to the one-dimensional distributions of Figure 4. In accordance with this, the consideration of bivariate probability densities leads to a remarkable drop of the zinnwaldite misclassification probability from 0.19 to 0.06, where we used a two-dimensional analogue of the decision rule given in (14), see also (41) for the general definition of a multidimensional decision rule. Note that the number of wrongly classified zinnwaldite composite particles dropped from 73 to 27, see Table 2, whereas the number of wrongly classified quartz composite particles only increased slightly. Thus, by considering multidimensional probability distributions of particle characteristics we can achieve better prediction results. In the following we describe in detail how to construct such two- or even higher-dimensional probability distributions and their corresponding decision rules.

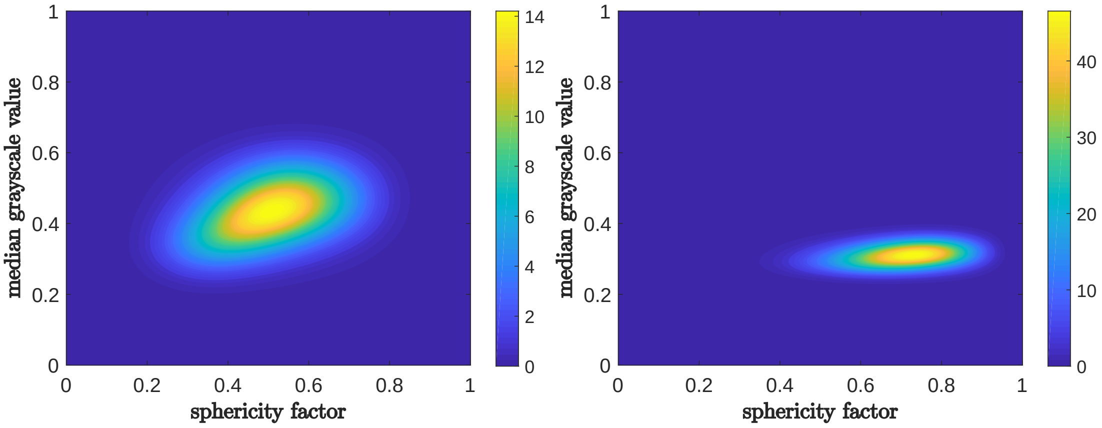


Figure 5: Joint probability density of sphericity factor and median grayscale value for zinnwaldite (left) and quartz (right) particles.

Table 2: Confusion matrix for predicting the particle composition based on joint densities of the sphericity factor and the median grayscale value, where the particles observed in the SEM-EDS data have been used for adjusting the prediction model.

	zinnwaldite	**quartz**
predicted zinnwaldite	315	33
predicted quartz	27	429

2.3.1 Modeling of single particle characteristics

We now fit parametric probability distributions to the one-dimensional particle characteristics described in Section 2.2. To be precise, we first fit parametric distributions to the volume equivalent radius, sphericity factor, convexity factor, elongation factor, median grayscale value and the interquartile range for the 3D zinnwaldite composite particles that intersect with the SEM-EDS plane. The chosen families of parametric distributions and the corresponding parameters that were determined to model the distributions of these characteristics are listed in Table 3. Note that we used gamma- and beta-distributions, whose probability density functions are given by

$$f_{\text{gamma}}(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} \exp\big(-\frac{x}{\theta}\big) \mathbb{1}_{x>0} \tag{17}$$

and

$$f_{\text{beta}}(x) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{1}_{x\in[0,1]}, \tag{18}$$

respectively, where $k, \theta, \alpha, \beta > 0$ are some model parameters and the gamma function $\Gamma\colon [0,\infty) \to [0,\infty)$ is defined by the integral

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \mathrm{d}x \qquad \text{for each } \alpha > 0. \tag{19}$$

For modeling the volume equivalent radius we used the gamma distribution. Note that there are also other families of distributions which result in a good fit, like the log-normal distribution. The remaining characteristics were modeled with beta distributions, since these particle characteristics have only values in the interval $[0, 1]$ which coincides with

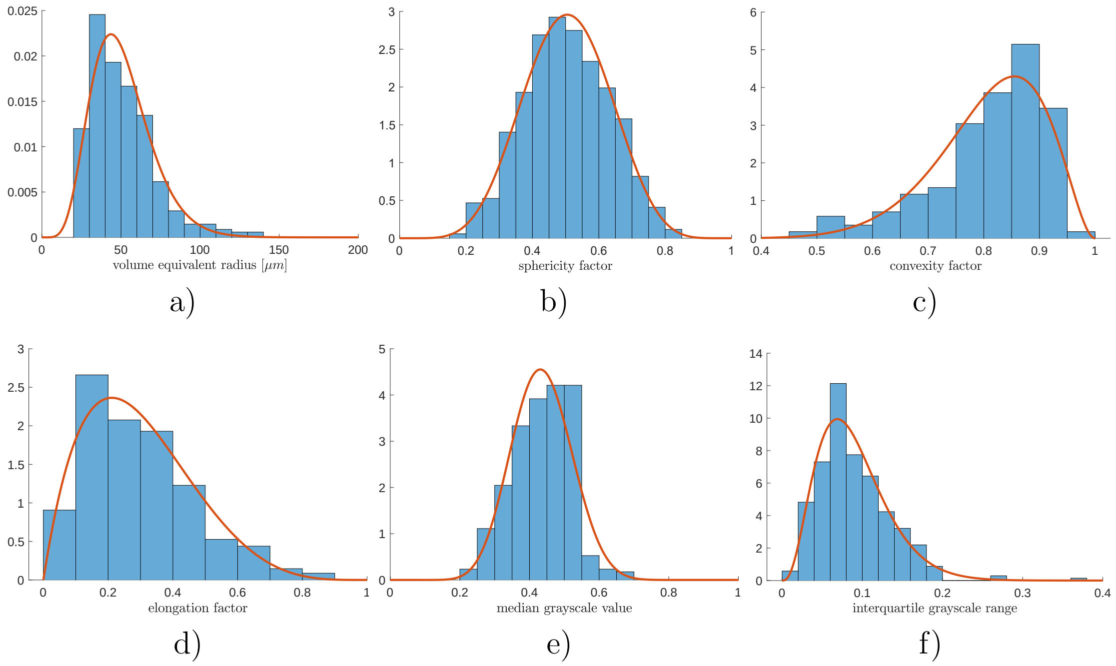


Figure 6: Histograms (blue) and fitted probability densities of zinnwaldite composite particle characteristics for a) volume equivalent radius, b) sphericity factor, c) convexity factor, d) elongation factor, e) median grayscale value, f) interquartile range, see also Table 3.

the support of the beta distribution. A visual comparison with the histograms of particle characteristics is given in Figure 6. In order to determine the model parameters (k, θ for the gamma- and α, β for the beta distribution) for these one-dimensional fits we used maximum likelihood estimators (Held and Bové, 2014). Analogously, the one-dimensional distributions of single particle characteristics were fitted for quartz composite particles, see Figure 7 and Table 3.

Table 3: Parameters of fitted distributions.

characteristic	**distribution**	**zinnwaldite**	**quartz**
volume equivalent radius	gamma	$k = 7.16, \theta = 6.84$	$k = 6.84, \theta = 9.8$
sphericity factor	beta	$\alpha = 7.17, \beta = 7.05$	$\alpha = 10.73, \beta = 4.63$
convexity factor	beta	$\alpha = 12.61, \beta = 2.97$	$\alpha = 17.35, \beta = 2.73$
elongation factor	beta	$\alpha = 2.00, \beta = 4.73$	$\alpha = 2.90, \beta = 2.74$
median grayscale value	beta	$\alpha = 14.12, \beta = 18.32$	$\alpha = 74.95, \beta = 166.85$
interquartile range	beta	$\alpha = 3.78, \beta = 38.54$	$\alpha = 4.52, \beta = 112.11$

2.3.2 Modeling of multidimensional particle characteristics

When modeling the joint distribution of independent random particle characteristics X, Y with probability density functions f_X and f_Y, respectively, the joint distribution is immediately given by $f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$. However, this approach is not available for correlated characteristics, as it is the case for our data, where, for example, we obtained a correlation coefficient of 0.36 for the sphericity factor and median grayscale value of zinnwaldite composite particles.

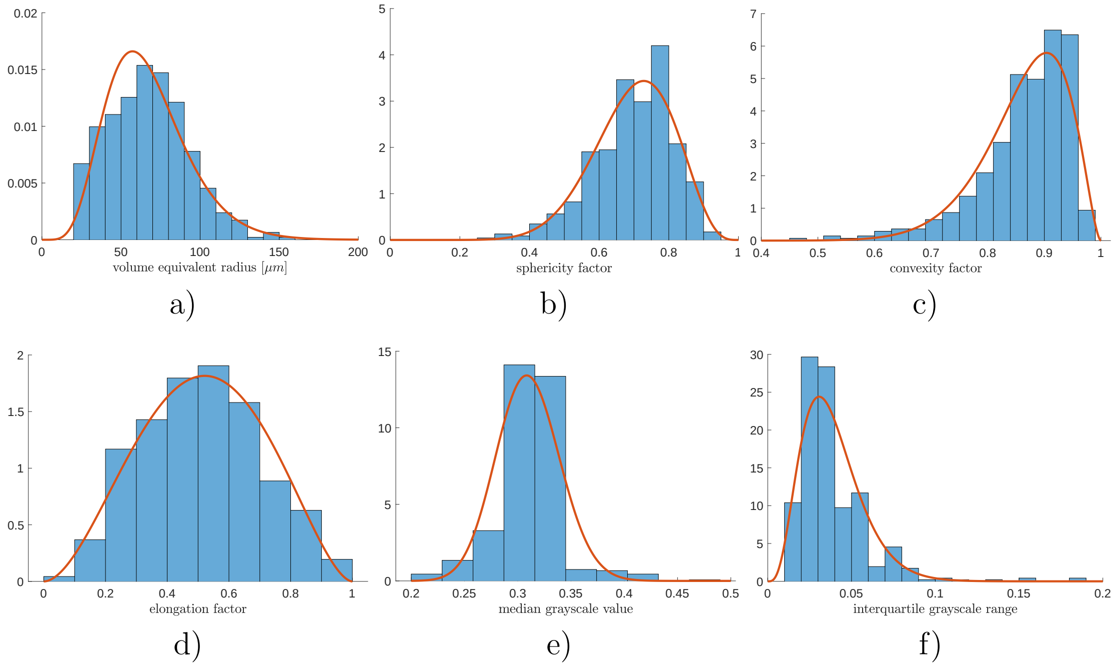


Figure 7: Histograms (blue) and fitted probability densities of quartz composite particle characteristics for a) volume equivalent radius, b) sphericity factor, c) convexity factor, d) elongation factor, e) median grayscale value, f) interquartile range.

Thus, in order to describe the joint distribution of vectors of particle characteristics, whose marginal distributions are given by the one-dimensional distributions we fitted in Section 2.3.1, see Table 3, we consider a more general approach, using so-called Archimedean copulas, see Nelsen (2006).

To make the paper more self-contained, we first explain some basic ideas of the copula approach, which will be used to construct multivariate probability distributions with non-normal marginals. Therefore, for some $d \geq 2$, let $f_1, \ldots, f_d \colon \mathbb{R} \to [0, \infty)$ denote the one-dimensional probability densities of particle characteristics for which we want to determine the d-dimensional joint probability density $f \colon \mathbb{R}^d \to [0, \infty)$ whose marginal densities are given by $f_1, \ldots, f_d$. In order to construct the multivariate density f we need to consider, for technical reasons, the one-dimensional (cumulative) distribution functions $F_1, \ldots, F_d \colon \mathbb{R} \to [0, 1]$ which are given by

$$F_i(x) = \int_{-\infty}^{x} f_i(y)\, \mathrm{d}y \qquad \text{for } i = 1, \ldots, d. \tag{20}$$

Note that $F_1(x)$ denotes the probability that the particle characteristic described by the density f_1 does not exceed the value $x \in \mathbb{R}$ and that the density can be obtained from the distribution function F_i by

$$f_i(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_i(x) \qquad \text{for } i = 1, \ldots, d. \tag{21}$$

Analogously, the d-dimensional density f has a cumulative distribution function $F \colon \mathbb{R}^d \to [0, 1]$ which is given by

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(y_1, \ldots, y_d)\, \mathrm{d}y_d \ldots \mathrm{d}y_1 \qquad \text{for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{22}$$

The joint density f is then obtained by the d-fold partial derivative

$$f(x) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} F(x) \qquad \text{for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{23}$$

In order to show the connection between the multivariate distribution function F and the notion of copulas we begin with the definition of the latter. A d-dimensional copula is a multivariate cumulative distribution function $K\colon \mathbb{R}^d \to [0,1]$, whose one-dimensional marginal distributions are uniform distributions on the interval $[0,1]$. For example the marginal cumulative distribution function $K_1\colon \mathbb{R} \to [0,1]$ of the first component is given by

$$K_1(x_1) = \lim_{x_2,\ldots,x_d \to \infty} K(x_1, x_2, \ldots, x_d) = \begin{cases} 0, & \text{if } x_1 < 0, \\ x_1, & \text{if } x_1 \in [0,1], \\ 1, & \text{if } x_1 > 1. \end{cases} \tag{24}$$

Copulas are of special interest because of Sklar's theorem (Nelsen, 2006). It states that F is a d-dimensional cumulative distribution function with marginal distribution functions $F_1, \ldots, F_d$ if and only if there is a copula function K such that

$$F(x) = K(F_1(x_1), \ldots, F_d(x_d)) \qquad \text{for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{25}$$

This means that every multivariate distribution function F can be represented by its marginals $F_1, \ldots, F_d$ and a copula function K. Thus, when modeling the joint distribution function F of random vectors whose marginal distributions are given by $F_1, \ldots, F_d$ it suffices to model the copula K. Note that there are numerous classes of copula functions, for which many of them do not have an analytical representation. In the present paper we limit the search for an appropriate copula function to so-called Archimedean copulas.

2.3.3 Archimedean copulas and differential variant of Sklar's theorem

For that purpose, let $\varphi\colon [0,1] \to [0,\infty]$ be a continuous, strictly decreasing and convex function with $\varphi(1) = 0$ and $\varphi(0) = \infty$. Note that φ is called an Archimedean generator. Then, it can be shown that the function $K\colon [0,1]^d \to [0,1]$ given by

$$K(u) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_d)) \qquad \text{for } u = (u_1, \ldots, u_d) \in [0,1]^d \tag{26}$$

can be uniquely extended to a copula on $\mathbb{R}^d$, i.e., a function possessing the properties mentioned in Section 2.3.2. It is called an Archimedean copula (Nelsen, 2006). With the copula function K given in (26) which still depends on some abstract function φ we can construct a d-dimensional distribution function F by means of (25) which has the marginal distribution functions $F_1, \ldots, F_d$. Since, finally, we are interested in the joint density f of particle characteristics, applying (23) to (25) we get that

$$f(x) = f_1(x_1) \ldots f_d(x_d) k(F_1(x_1), \ldots, F_d(x_d)) \qquad \text{for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \tag{27}$$

which can be seen as a differential variant of Sklar's theorem given in (25), where the function $k\colon \mathbb{R}^d \to [0,\infty)$ is the d-fold derivative

$$k(u) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} K(u) \qquad \text{for } u = (u_1, \ldots, u_d) \in \mathbb{R}^d. \tag{28}$$

Recall that the d-dimensional probability density f given by (27) has the marginal densities $f_1, \ldots, f_d$ and, besides this, depends on the choice of a suitable copula function K or, equivalently, the choice of its derivative k. Thus, the task in modeling multidimensional probability distributions for given (one-dimensional) marginal densities is the determination of a suitable copula function K which describes the correlation structure of the one-dimensional marginals. Since in the present paper we only consider Archimedean copulas, which are defined by their generator φ, this task is equivalent to finding a suitable Archimedean generator.

2.3.4 Parametric families of Archimedean generators

Instead of analyzing single generator functions φ one by one, we can instead consider parametric families $\{\varphi_\theta : \theta \in \Theta\}$ of generators. For such families the function φ_θ is an Archimedean generator for each parameter $\theta \in \Theta$, where Θ is some set of admissible parameters. This has the advantage that we can consider a whole range of generators for which it is rather easy to determine an optimal choice of a generator among the family $\{\varphi_\theta : \theta \in \Theta\}$ for modeling the joint density f.

An example of a one-parametric family of generators is the Ali-Mikhail-Haq generator, see Nelsen (2006), which is given by

$$\varphi_\theta(u) = \log\Big(\frac{1-\theta(1-u)}{u}\Big), \tag{29}$$

for each $u \in [0,1]$ and some $\theta \in \Theta = (-1,1)$. Note that the parameter θ regulates the correlation between the uniformly distributed marginals of the corresponding copula K_θ. The relationship between the copula parameter θ and the common correlation coefficients, namely, the Pearson, Kendall and Spearman correlation coefficients (Nelsen, 2006) is visualized in Figure 8 for two-dimensional Ali-Mikhail-Haq copulas. Note that in the

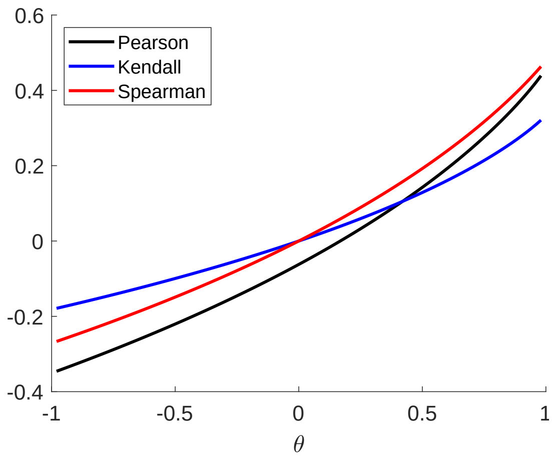


Figure 8: Relationship between the copula parameter θ and the Pearson, Kendall and Spearman correlation coefficients of the uniformly distributed marginals of two-dimensional Ali-Mikhail-Haq copulas.

present paper other one-parametric families of copulas are considered as well, e.g., the Frank copula whose generator is given by

$$\varphi_\theta(u) = -\log\left(\frac{e^{-\theta u}-1}{e^{-\theta}-1}\right), \tag{30}$$

for $\theta \neq 0$. An extensive list on even further copulas, like the Clayton, Gumbel, Joe and Plackett copula, can be found in Joe (1997). Beyond that, it is possible to obtain further copulas by rotating a given two-dimensional copula by a multiple of 90°. Thus, there are four rotated versions of the aforementioned parametric families of copulas, which increases the list of considered parametric copula families even further.

Fitting a model from a parametric family of generators $\{\varphi_\theta : \theta \in \Theta\}$ is equivalent to determining an optimal parameter $\hat{\theta}$. This problem can be described as an optimization problem using maximum-likelihood methods in the following way.

The family of generators $\{\varphi_\theta : \theta \in \Theta\}$ induces a family of copulas $\{K_\theta : \theta \in \Theta\}$ given by

$$K_\theta(u) = \varphi_\theta^{-1}(\varphi_\theta(u_1) + \cdots + \varphi_\theta(u_d)) \qquad \text{for } u = (u_1, \ldots, u_d) \in [0,1]^d, \tag{31}$$

for which the optimal parameter $\hat{\theta}$ has to be determined. Assuming that K_θ is differentiable we can consider the d-fold derivative

$$k_\theta(u) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} K_\theta(u), \tag{32}$$

and the corresponding d-dimensional distribution function F_θ which is given by (25) has the probability density function

$$f_\theta(x) = f_1(x_1) \ldots f_d(x_d) k_\theta(F_1(x_1), \ldots, F_d(x_d)) \qquad \text{for } x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \tag{33}$$

where $f_1, \ldots, f_d$ and $F_1, \ldots, F_d$ are the one-dimensional densities and distribution functions fitted in Section 2.3.1 to the image data considered in the present paper.

Note that, for a sample of $\ell \in \mathbb{N}$ observations $x^{(1)}, \ldots, x^{(\ell)} \in \mathbb{R}^d$ of the considered d-dimensional vector of particle characteristics, the so-called log-likelihood function is given by

$$\log \mathcal{L}(\theta | x^{(1)}, \ldots, x^{(\ell)}) = \sum_{k=1}^{\ell} \log\big(f_\theta(x^{(k)})\big) \qquad \text{for } \theta \in (-1, 1). \tag{34}$$

This leads to the maximum-likelihood estimator

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log \mathcal{L}(\theta | x^{(1)}, \ldots, x^{(\ell)}), \tag{35}$$

which can be considered as the optimal choice for the parameter θ when the sample $x^{(1)}, \ldots, x^{(\ell)}$ was observed. The fitted joint density f is then given by $f = f_{\hat{\theta}}$.

Until now, the procedure discussed above solely explains how to estimate the copula parameter θ for a given parametric family of copulas. In order to select an adequate model type among several parametric copula families, one can use the so-called Akaike information criterion (AIC) which utilizes the log-likelihood function considered in (34) to compare parametric families of copulas and select the best performing model. Note that, besides utilizing the log-likelihood function considered in (34), the AIC additionally takes into consideration the number of model parameters to avoid overfitting models. For more details regarding the AIC and model selection, see e.g. Held and Bové (2014).

2.3.5 Application to particle characteristics

In the context considered in the present paper, the estimation procedure described above can be applied in the following way.

In order to obtain the (two-dimensional) joint density of the sphericity and median grayscale value of zinnwaldite composite particles, see Figure 5, we utilize the 2-dimensional vectors of particle characteristics $x^{(j)} = (s(P_j), m(P_j))$ which have been computed in Section 2.2 for each zinnwaldite composite particle $P_1, \ldots, P_\ell \in Z$ that hits the SEM-EDS plane. For both characteristics the parameters of the corresponding one-dimensional probability densities f_s^z, f_m^z for zinnwaldite composite particles are given in Table 3. Now we consider, for example, the family of Frank copulas, assuming that it is able to describe the two-dimensional density of the sphericity and median grayscale value. Then, we utilize the formula given in (33) and the previously determined one-dimensional probability densities f_s^z, f_m^z to obtain a parametric family of possible candidates for the two-dimensional joint density of the considered particle characteristics. By inserting the computed vectors of particle characteristics $x^{(j)}$ into (34), we obtain the log-likelihood function $\log \mathcal{L}$, see Figure 9, which after maximization provides us the copula parameter $\hat{\theta} = 2.45$. Thus, by means of (33) we obtain the two-dimensional probability density of zinnwaldite composite particle characteristics

$$f_{s,m}^z(x_1, x_2) = f_s^z(x_1) f_m^z(x_2) k_{\hat{\theta}}(F_s^z(x_1), F_m^z(x_2)), \tag{36}$$

where F_s^z and F_m^z are the distribution functions corresponding to the densities f_s^z and f_m^z, respectively. Note that we found out that the family of Frank copulas indeed provides a good fit, in comparison to other parametric families, by comparing the AIC of various families of copulas mentioned in Section 2.3.4. Analogously, we showed that the Frank copula rotated by 90° provides a good fit for the two-dimensional probability density $f_{s,m}^q$ of the sphericity and median grayscale value of quartz composite particles with the copula parameter $\hat{\theta} = -1.14$, see Figure 5.

Recall that the quality of the decision, whether a particle is comprised mainly of zinnwaldite or quartz, improved significantly once we considered the two-dimensional probability densities $f_{s,m}^z, f_{s,m}^q$ over one-dimensional densities, compare Tables 1 and 2. Furthermore, recall that the estimation procedure described in Section 2.3.4 is not limited to the simultaneous consideration of two particle characteristics. But, in the same way, it can be applied for general d-dimensional vectors of particle characteristics. For example, we can investigate the six particle charactistics introduced in Section 2.2, thus considering the case $d = 6$.

For simplicity, we first fit one-parametric copulas to the six-dimensional data, which leads to better prediction results than just considering two-dimensional particle characteristics, see Table 4. However, we are aware of the fact that one-parametric copulas do not entirely capture the correlation structure of the six-dimensional data, because the empirical (pairwise) correlations coefficients can vary distinctly for each individual pair of particle characteristics. It turned out that, among the one-parametric copula families mentioned in Section 2.3.4, the Ali-Mikhail-Haq copula yielded the best fit. In particular, using both the six-dimensional vectors of characteristics $x^{(j)} = (r(P_j), s(P_j), c(P_j), e(P_j), m(P_j), iqr(P_j))$ of zinnwaldite paticles $P_1, \ldots, P_\ell \in Z$ and the corresponding one-dimensional probability densities $f_r^z, f_s^z, f_c^z, f_e^z, f_m^z, f_{iqr}^z$, see Section 2.3.1, by means of (31)-(34) we obtain the copula parameter $\hat{\theta} = 0.69$, and thus a parametric fit for the six-dimensional density f_z of zinnwaldite composite particle characteristics.

Analogously, we obtained the Ali-Mikhail-Haq copula with parameter $\hat{\theta} = 0.47$ and thus the multidimensional density function f_q for the characteristics of quartz particles.

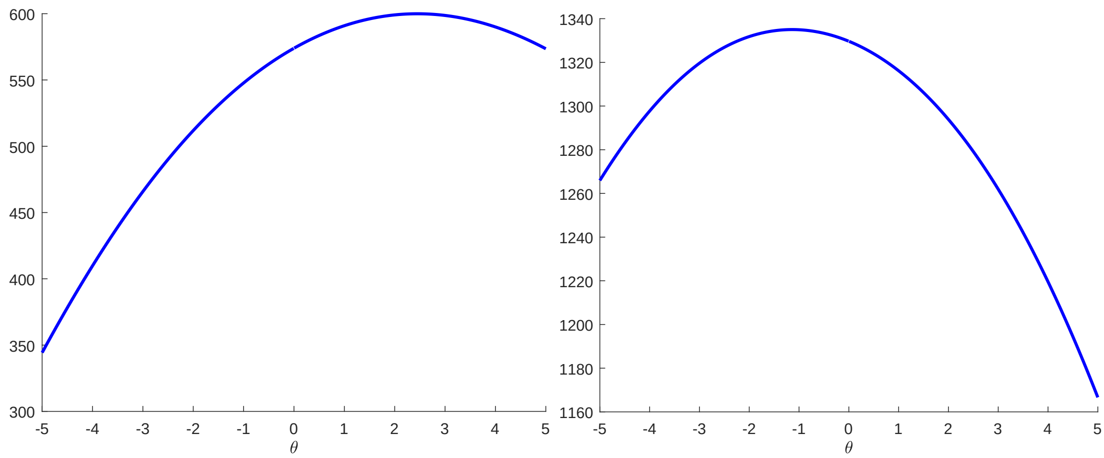


Figure 9: Log-likelihood functions for fitting two-dimensional copulas to the sphericity and median grayscale value of zinnwaldite composite particles using a Frank copula (left) and of quartz composite particles using a Frank copulua rotated by 90° *(right).*

One way to achieve even better prediction results would be to utilize multi-parametric copula families to fully capture the correlation structure of the six-dimensional data, like, e.g., vine copulas, see Bedford and Cooke (2002). However, in the following, we consider yet another approach to achieve this goal, which uses the six-dimensional data, but projects it to a suitably chosen two-dimensional subspace. In this way, we still exploit information of the 6-dimensional vector of particle characteristics while being able to consider two-dimensional copulas, or even two-dimensional kernel density estimators.

2.3.6 Kernel density estimation and dimension reduction

Recall that for describing multivariate probability densities of particle characteristics, which are necessary for improving the decision rule of the one-dimensional case considered in (14), we used a parametric copula approach in Section 2.3.5. Alternatively, such densities could be estimated in a nonparametric way, using so-called kernel density estimators (Scott, 2015). To be precise, from a sample $x^{(1)}, \dots, x^{(\ell)} \in \mathbb{R}$ of a one-dimensional characteristic of zinnwaldite composite particles a non-parametric density $\hat{f}_z$, can be estimated by

$$\hat{f}_z(x) = \frac{1}{\ell} \sum_{j=1}^{\ell} \kappa\Big(\frac{x - x^{(j)}}{h}\Big) \qquad \text{for } x \in \mathbb{R}, \tag{37}$$

where $h > 0$ is called the bandwith and $\kappa \colon \mathbb{R} \to [0, \infty)$ is a kernel function, e.g.,

$$\kappa(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) \qquad \text{for } x \in \mathbb{R}. \tag{38}$$

Analogously, it would be possible to estimate non-parametric multivariate densities for vectors of particle characteristics. However, note that due to the "curse of dimensionality" the required sample size for adequate kernel density estimation increases exponentially with the considered dimension, see Scott (2015). Thus, it is difficult to utilize kernel density estimators for computing accurate six-dimensional densities representing our six-dimensional data.

Nevertheless, the issue concerning the sample size of the six-dimensional vectors of particle characteristics, caused by the "curse of dimensionality", can be remedied by dimension reduction. Common techniques like the principle component analysis (PCA) project the data to a lower dimensional subspace which maximizes the variance of the data (Hastie et al., 2009). However, PCA does not ensure a good separability of the two classes, namely zinnwaldite and quartz composite particles, after projection of the data. Other methods, like the linear discriminant analysis (LDA) can reduce the dimension of the data such that classification can still be performed reliably. Note that LDA always projects data on a $c-1$ dimensional subspace, where c is the number of classes. In the setting of the present paper, LDA would project the six-dimensional vectors of zinnwaldite and quartz composite particle characteristics on a one-dimensional subspace such that the normalized discrepancy measure

$$\Delta_1 = \frac{(\mu_z - \mu_q)^2}{\sigma_z^2 + \sigma_q^2} \tag{39}$$

is maximized, where $\mu_z, \mu_q \in \mathbb{R}$ are the means of the projected characteristics vectors of zinnwaldite and quartz, respectively. Analogously, the values $\sigma_z^2, \sigma_q^2 \in \mathbb{R}$ are their variances. Heuristically speaking, the LDA maximizes the distance between clusters, while minimizing their variances. In order to fit multidimensional probability densities or to perform multidimensional kernel density estimation on the data after dimension reduction we modified the expression considered in (39) such that we reduce the data to two dimensions. More precisely, we project the six-dimensional data on a two-dimensional subspace such that the normalized discrepancy measure

$$\Delta_2 = \frac{\|\tilde{\mu}_z - \tilde{\mu}_q\|^2}{\sigma_{z,1}^2 + \sigma_{z,2}^2 + \sigma_{q,1}^2 + \sigma_{q,2}^2} \tag{40}$$

is maximized, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^2$ and $\tilde{\mu}_z, \tilde{\mu}_q \in \mathbb{R}^2$ are the means of the clusters corresponding to zinnwaldite and quartz composite particles after projection. The values $\sigma_{z,1}^2, \sigma_{z,2}^2$ are the respective variances of the first and second component for vectors of characteristics of zinnwaldite after projection. Analogously, $\sigma_{q,1}^2, \sigma_{q,2}^2$ denote the variances of the components for the quartz case. By maximizing the expression given in (40) we obtain the dimension reduction by projecting for each particle P the corresponding vectors of characteristics $x = (r(P), s(P), c(P), e(P), m(P), iqr(P))$ onto the plane with span vectors $v_1 = (0.0014, 0.3189, 0.0189, 0.0562, -0.9455, -0.0285)$ and $v_2 = (0.0001, -0.0944, 0.1410, -0.0193, -0.0598, 0.9835)$. Figure 10 indicates that the vectors of characteristics for zinnwaldite and quartz composite particles after projection are relatively distinct. Furthermore, this reduction of dimension makes kernel density estimation more viable again. Note that the "ragged" probability densities in Figures 11a and 11b still indicate too small sample sizes for adequate kernel density estimation. However, the methods discussed in Sections 2.3.1 and 2.3.4 can be used to fit parametric copulas to the two-dimensional data after dimension reduction, see Figures 11c and 11d.

3 Results and Discussion

At the beginning of Section 2.3 we showed a way how the mineralogical composition of particles could be predicted on the basis of one-dimensional probability distributions of single particle characteristics introduced in Section 2.2. However, in Figure 4 and Table 1 we have seen that one-dimensional distributions do not suffice for making reliable

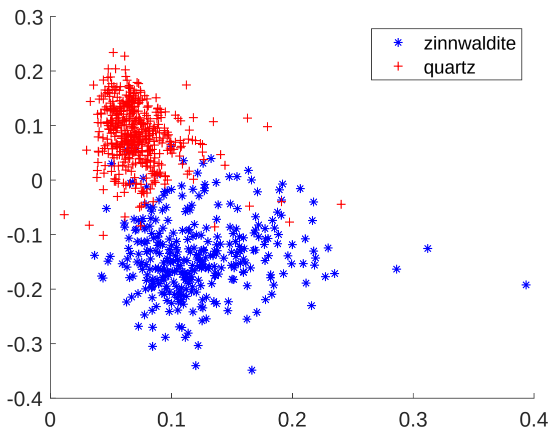


Figure 10: Six-dimensional data after dimension reduction obtained by maximizing cluster distance, while minimizing cluster variances.

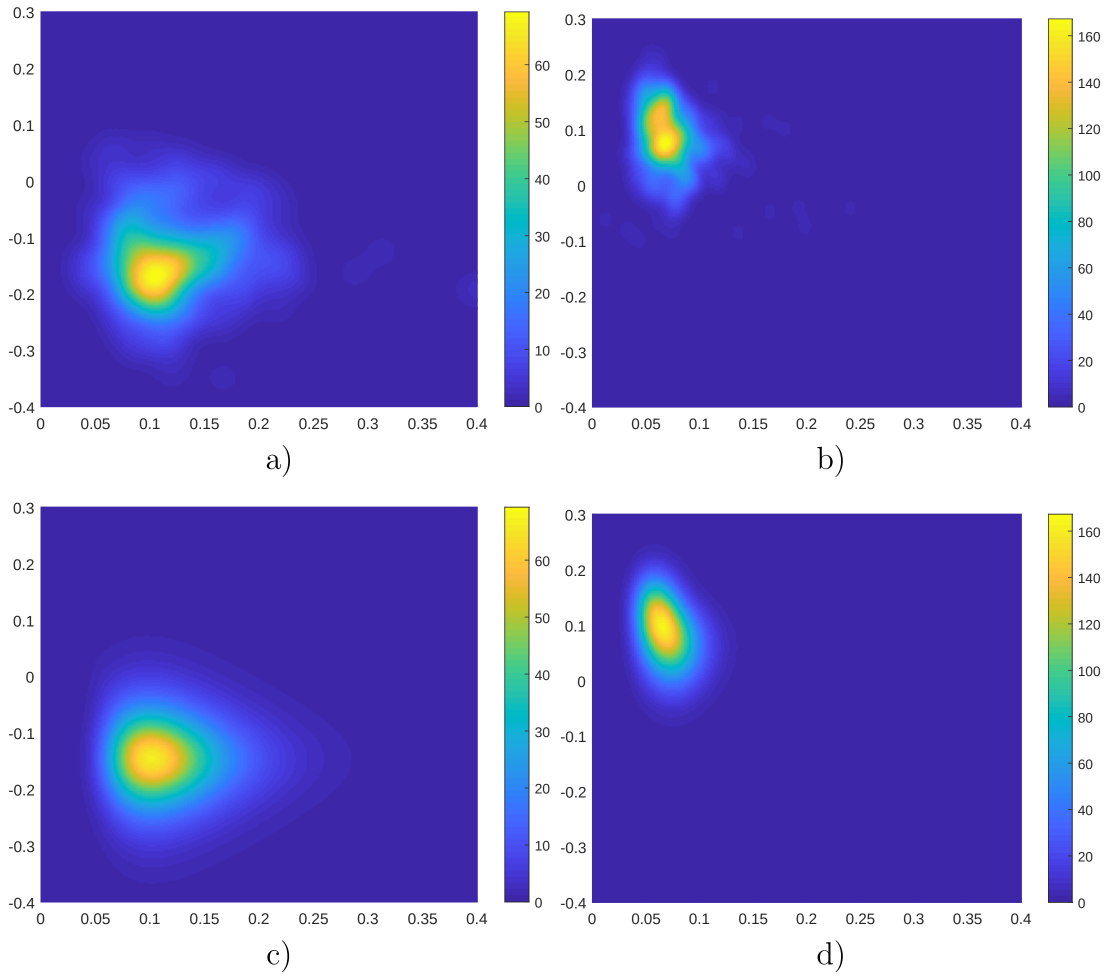


Figure 11: Kernel density estimation of vectors of characteristics after dimension reduction for a) zinnwaldite and b) quartz composite particles. The correpsonding probability densities obtained by fitting parametric copulas for c) zinnwaldite and d) quartz composite particles.

decisions, whereas Figure 5 and Table 2 indicate that the predictive power of our decision rule can increase when we consider bivariate vectors of characteristics. Moreover, in Section 2.3.5 we showed how multidimensional density functions f_z and f_q for 6 characteristics of zinnwaldite and quartz composite particles can be determined. Then, analogously to the one-dimensional case, it is possible to predict the composition of a particle $P \subset \mathbb{Z}^3$ observed in the XMT image data, using the following decision rule: If

$$f_z(r(P), s(P), c(P), e(P), m(P), iqr(P)) > f_q(r(P), s(P), c(P), e(P), m(P), iqr(P)), \tag{41}$$

then we assume that P is mainly composed of zinnwaldite, otherwise of quartz. In other words, the decision rule given in (41) states that a particle P with the vector of characteristics $x = \big(r(P), s(P), c(P), e(P), m(P), iqr(P)\big)$ is classified as a zinnwaldite composite particle if the configuration of characteristics x is more likely to occur according to the distribution of characteristics of zinnwaldite composite particles given by f_z.

Table 4: Confusion matrices of the copula-based decision rule given in (41) for particles observed in the SEM-EDS section which has been used for adjusting the prediction model and for the particles observed in the SEM-EDS section which was withheld for validation.

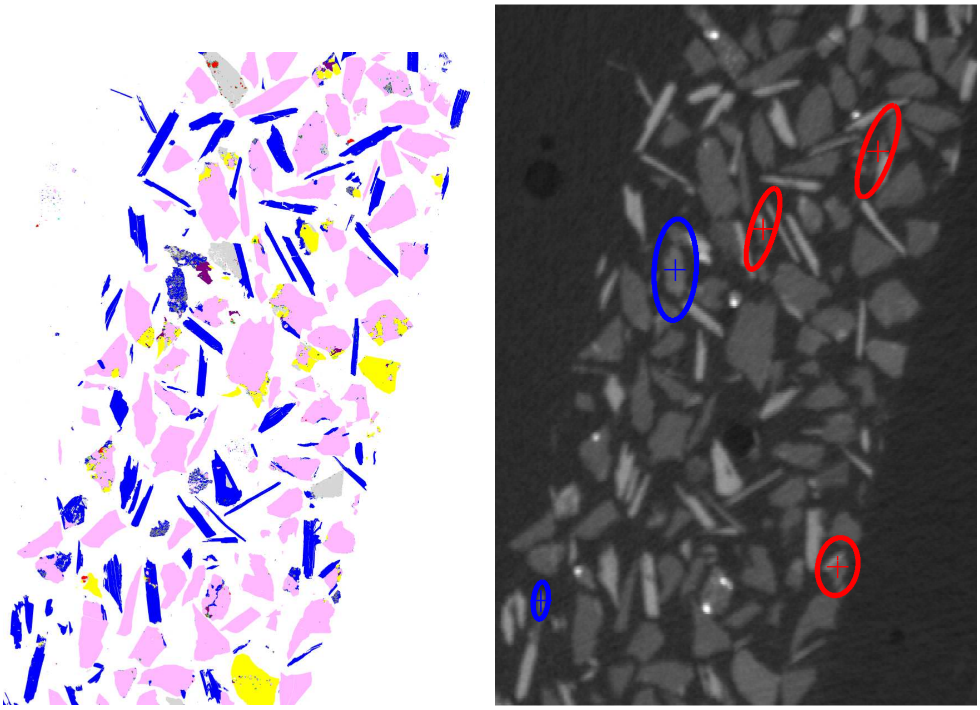
	calibration section		**validation section**	
	zinnwaldite	**quartz**	**zinnwaldite**	**quartz**
predicted zinnwaldite	333	25	204	19
predicted quartz	9	437	12	244

Table 4 shows that the decision rule given in (41) is rather accurate, i.e., only 2.7% of zinnwaldite composite particles are wrongly classified as quartz composite particles and 5.7% of the quartz composite particles are wrongly classified as zinnwaldite composite particles. In comparison to the decision rule which only used the sphericity factor and the median grayscale value, see Table 2, the number of wrongly classified zinnwaldite composite particles dropped from 27 to 9, see also Figure 12. This is a significant improvement compared to the prediction based on two characteristics shown in Table 2. We suppose that the prediction results can become even better when we consider multi-parametric copulas.

Alternatively, we can use the dimension reduction considered in Section 2.3.6, where the six-dimensional vectors of particle characteristics were projected onto a two-dimensional subspace and the corresponding probability densities required for classification were determined using either kernel density estimation (Figures 11a and 11b) or Archimedean copulas (Figures 11c and 11d). The classification results using the latter procedure are listed in Table 5, where the errors for the particles used from the SEM-EDS section used for calibration are similar to the errors occurring by the usage of six-dimensional copulas. The approach using kernel density estimation performs best on the calibration data, see Table 6. This was to be expected, since the entire sample information is encoded in the estimated densities.

Table 5: Confusion matrix of the copula-based decision rule utilizing dimension reduction. The corresponding fitted probability densities are visualized in Figures 11c and 11d.

	calibration section		**validation section**	
	zinnwaldite	**quartz**	**zinnwaldite**	**quartz**
predicted zinnwaldite	323	17	207	8
predicted quartz	19	445	9	255



a) b)

Figure 12: a) Cutout of the SEM-EDS section used for the calibration of the prediction model, where the blue phase indicates zinnwaldite and pink indicates quartz. b) The corresponding XMT section. Blue ellipses mark zinnwaldite composite particles which were wrongly classified as quartz composite particles by the decision rule given in (41). Analogously, red ellipses indicate wrongly classified quartz composite particles.

Table 6: Confusion matrix of the decision rule utilizing dimension reduction followed by kernel density estimation. The corresponding probability densities are visualized in Figures 11a and 11b.

	calibration section		validation section	
	zinnwaldite	**quartz**	**zinnwaldite**	**quartz**
predicted zinnwaldite	331	12	205	8
predicted quartz	11	450	11	255

To validate the predictive capability of the presented models, we used an additional planar SEM-EDS data set measured at a spatially different location of the 3D sample, see red plane in Figure 1a. For particles that hit this second plane we determined the particle characteristics and predicted the composition of the particles using the method based on six-dimensional copulas described in Section 2.3.5 and the classification method based on dimension reduction followed by fitting parametric copulas or kernel density estimation. Due to the SEM-EDS data the true composition of these particles is known. A comparison between the ground truth and the predictions can be seen in Tables 4 to 6, where the predictions were again rather good.

Since in the latter case the mineralogical information of particles that hit the second SEM-EDS plane was not used for calibrating the prediction models, we can assume, based on the results of Tables 4 to 6, that the predictions will remain accurate for each particle of the entire 3D XMT image. Furthermore, the validation results show that the prediction models based on dimension reduction followed by fitting parametric copulas or kernel density estimation perform best on the validation section, which indicates that the approach based on six-dimensional copulas can be improved by using more general, higher-parametric copula models.

Note that, even though the prediction method using multidimensional probability densities was presented for a classification problem with two classes, it is not limited to such binary cases. In general, for a prediction problem with $n \in \mathbb{N}$ classes, the probability density f_k of the considered d-dimensional feature vector is required for each class $k = 1, \cdots, n$. Furthermore, it is necessary to have a priori knowledge on the occurrence probability p_k of each class k. Then a vector of characteristics $x \in \mathbb{R}^d$ is assigned to class i if

$$f_i(x)p_i > f_k(x)p_k \tag{42}$$

for each $k \neq i$. Note that we omitted the a priori occurrence probabilities p_k in the present paper, since we had roughly the same amount of zinnwaldite and quartz composite particles. For more details regarding such classification methods we refer the reader to Duda et al. (2001).

4 Conclusions & Outlook

We presented methods for the mineralogical characterization of particles observed in 3D XMT image data which mainly consists of the minerals zinnwaldite and quartz. The grayscale value $I(x)$ of a voxel $x \in W \subset \mathbb{Z}^3$ in the 3D XMT image already provides some information about the mass density of the observed mineral at the location x. However, this information does not suffice for reliably distinguishing between zinnwaldite and quartz composite particles. Therefore, we additionally considered the morphology of particles to characterize them. The proposed prediction models can then characterize particles based on their 3D morphology and grayscale values extracted solely from XMT data.

For the calibration of the prediction models we had to localize a 2D SEM-EDS data set in the 3D XMT image. The SEM-EDS data provided a mineralogical characterization of particles in a planar cross-section of the sample, such that we know the 3D morphology and the composition of particles that intersect with this cross-section. In order to extract single 3D particles that hit the planar cross-section, we computed a particle-wise segmentation of the 3D image data. This allowed us to determine for each particle its vector of characteristics which is relevant to distinguish between zinnwaldite and quartz composite particles. Among the considered characteristics are the median grayscale value of particles, but also particle size and shape characteristics like the volume equivalent radius and the sphericity factor.

This resulted in vectors of characteristics for each particle. Since the mineralogical composition of particles that hit the SEM-EDS plane is known, we assigned the corresponding vectors of characteristics to the mineral the particle is mostly composed of. Therefore, we had a set of vectors of characteristics which corresponded to particles dominated by zinnwaldite and analogously another set of such vectors corresponding to particles dominated by quartz. The proposed prediction models require joint densities f_z and f_q of these characteristics for particles composed of zinnwaldite and quartz, respectively. Therefore, we fitted parametric copulas to each of the two data sets. This entailed fitting one-dimensional parametric families of distributions to each considered characteristic (separately for zinnwaldite and quartz composite particles). In a second step these one-dimensional densities were combined to obtain multidimensional densities. Since the considered characteristics were correlated, the multidimensional densities had to reflect the correlation structure. We achieved this for two-dimensional densities with the help of parametric Archimedean copulas. Furthermore, we projected our six-dimensional particle characteristics onto a two-dimensional subspace, such that after this dimension reduction the vectors of particle characteristics corresponding to zinnwaldite or quartz composite particles show a good separability. This approach made kernel density estimation more viable again, such that it was possible to utilize the corresponding densities for a further prediction model.

The resulting decision rules characterize a particle with a vector of characteristics $x \in \mathbb{R}^d$ as zinnwaldite composite particle if $f_z(x) > f_q(x)$, whereby the densities f_z, f_q were either fitted using a copula approach or kernel density estimation. We observed that this decision rule became more accurate when higher dimensional vectors of characteristics were considered. However, there was already a significant drop of the misclassification percentage when we considered two-dimensional characteristics instead of just considering single (i.e. one-dimensional) particle characteristics. This error was reduced even further when we considered vectors of six characteristics. However, for simplicity we used only one-parametric copulas when considering six characteristics. We suppose that higher-parametric models for fitting six-dimensional copulas can lead to even better results. As an alternative to higher-parametric copula models, we fitted two-dimensional copulas and also used kernel density estimators to the data after dimension reduction – which led to the best predictive results. To validate the prediction models we used additional SEM-EDS data at a spatially different location than the SEM-EDS data set that was used to adjust the models. This allowed us to compare the predictions of the models with the ground truth obtained by SEM-EDS. Since we observed that the prediction models discussed in Section 3 were quite accurate in this validation step, we can assume that they can reliably distinguish between zinnwaldite and quartz composite particles in the entire XMT image – even for areas where no SEM-EDS data is available.

In a forthcoming study we will use the copula-based modeling considered in the present paper to quantify the success of particle separation processes. To be precise, assume that the multidimensional density f of characteristics of a mixture of zinnwaldite and quartz composite particles is given by the convex-combination

$$f = (1-p)f_z + pf_q, \tag{43}$$

for some $p \in [0,1]$. When the p value drops to 0 after a separation process we can say that the zinnwaldite composite particles are separated from quartz composite particles. However, it is possible that the distribution of zinnwaldite composite particle characteristics itself changes significantly during separation. For example, if the separation process only extracts small zinnwaldite composite particles, the multidimensional density of zinnwaldite characteristics after separation would differ from the original density f_z, prior to separation. In order to track and compare these changes in the densities, it is useful to represent the multidimensional densities f_z prior and after separation with parametric copulas since they are described by only a few parameters. This complexity reduction is an advantage of parametric copulas, since it allows us to compare complicated multidimensional distributions by simply comparing the parameters of the distributions.

5 Acknowledgements

The financial support of the German Research Foundation (DFG) for funding the X-ray microscope (INST267/129-1) as well as the research projects (PE1160/22-1 and SCHM997/27-1) within the priority program SPP 2045 "Highly specific and multidimensional fractionation of fine particle systems with technical relevance" is gratefully acknowledged. The authors would like to thank Sabine Gilbricht for her work at the MLA and Roland Würkert for preparing the epoxy blocks.

References

Bedford, T. and Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068.

Boas, F. E. and Fleischmann, D. (2012). CT artifacts: Causes and reduction techniques. *Imaging in Medicine*, 4(2):229–240.

Buades, A., Coll, B., and Morel, J. M. (2005). A non-local algorithm for image denoising. In *Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, volume 2, pages 60–65, San Diego. IEEE Computer Society.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, Weinheim, 2nd edition.

Feinauer, J., Brereton, T., Spettl, A., Weber, M., Manke, I., and Schmidt, V. (2015). Stochastic 3D modeling of the microstructure of lithium-ion battery anodes via Gaussian random fields on the sphere. *Computational Materials Science*, 109:137 – 146.

Furat, O., Leißner, T., Ditscherlein, R., Sedivy, O., Weber, M., Bachmann, K., Gutzmer, J., Peuker, U., and Schmidt, V. (2018). Description of ore particles from XMT images, supported by SEM-based image analysis. *Microscopy and Microanalysis*, 24:461–470.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2nd edition.

Held, L. and Bové, D. S. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer, Berlin ; Heidelberg.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC, 1st edition.

Leißner, T., Bachmann, K., Gutzmer, J., and Peuker, U. (2016). MLA-based partition curves for magnetic separation. *Minerals Engineering*, 94:94 – 103.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer, New York.

Reyes, F., Lin, Q., Cilliers, J., and Neethling, S. (2018). Quantifying mineral liberation by particle grade and surface exposure using X-ray microCT. *Minerals Engineering*, 125:75 – 82.

Rieder, M., Huka, M., Kučerová, D., Minařík, L., Obermajer, J., and Povondra, P. (1970). Chemical composition and physical properties of lithium-iron micas from the Krušné hory Mts.(erzgebirge). part a: Chemical composition. *Contributions to Mineralogy and Petrology*, 27(2):131–158.

Schladitz, K., Ohser, J., and Nagel, W. (2006). Measuring intrinsic volumes in digital 3D images. In *Discrete Geometry for Computer Imagery: 13th International Conference, DGCI 2006, Szeged, Hungary, October 25-27, 2006. Proceedings*, pages 247–258. Springer.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Shafait, F., Keysers, D., and Breuel, T. M. (2008). Efficient implementation of local adaptive thresholding techniques using integral images. *SPIE Proceedings*, 6815:681510–681510–6.

Shapiro, L. and Stockman, G. (2001). *Computer Vision*. Prentice Hall, Upper Saddle River, NJ, USA.

Spettl, A., Wimmer, R., Werz, T., Heinze, M., Odenbach, S., Krill III, C. E., and Schmidt, V. (2015). Stochastic 3D modeling of Ostwald ripening at ultra-high volume fractions of the coarsening phase. *Modelling and Simulation in Materials Science and Engineering*, 23(6):065001.

Su, D. and Yan, W. (2018). 3D characterization of general-shape sand particles using microfocus X-ray computed tomography and spherical harmonic functions, and particle regeneration using multivariate random vector. *Powder Technology*, 323:8 – 23.

Sunderland, D. and Gottlieb, P. (1991). Application of automated quantitative mineralogy in mineral processing. *Minerals Engineering*, 4(7-11):753–762.

Wang, Y., Lin, C., and Miller, J. (2017). Quantitative analysis of exposed grain surface area for multiphase particles using X-ray microtomography. *Powder Technology*, 308:368 – 377.