Science Arts & Métiers (SAM)
is an open access repository that collects the work of Arts et Métiers Institute of Technology researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: https://sam.ensam.eu
Handle ID: .http://hdl.handle.net/10985/17597

To cite this version :

Dorian DEPRIESTER, Regis KUBLER - Resolution of the Wicksell's equation by Minimum Distance Estimation - Image Analysis and Stereology - Vol. 38, p.213-226 - 2019

Any correspondence concerning this service should be sent to the repository Administrator : scienceouverte@ensam.eu

Image Anal Stereol 2019;38:213-226 doi:10.5566/ias.2133
Original Research Paper

RESOLUTION OF THE WICKSELL'S EQUATION BY MINIMUM DISTANCE ESTIMATION

DORIAN DEPRIESTER✉ AND RÉGIS KUBLER

MSMP laboratory (EA 7350), École Nationale Supérieure d'Arts et Métiers, 2 cours des Arts et Métiers - 13617 Aix-en-Provence, France
e-mail: dorian.depriester@ensam.eu, regis.kubler@ensam.eu
(Received February 27, 2019; revised June 8, 2019; accepted June 20, 2019)

ABSTRACT

The estimation of the grain size in granular materials is usually performed by 2D observations. Unfolding the grain size distribution from apparent 2D sizes is commonly referred as the corpuscle problem. For spherical particles, the distribution of the apparent size can be related to that of the actual size thanks to the Wicksell's equation. The Saltikov method, which is based on Wicksell's equation, is the most widely used method for resolving corpuscle problems. This method is recursive and works on the finite histogram of the grain size. In this paper, we propose an algorithm based on a minimizing procedure to numerically solve the Wicksell's equation, assuming a parametric model for the distribution (e.g. lognormal distribution). This algorithm is applied on real material and the results are compared to those found using Saltikov or Saltikov-based stereology techniques. A criterion is proposed for choosing the number of bins in the Saltikov method. The accuracy of the proposed algorithm, depending on the sample size, is studied.

Keywords: microstructure, minimization, probability density function, Saltikov, stereology.

INTRODUCTION

THE CORPUSCLE PROBLEM

When one attempts to characterize the grain size distribution of a given granular material, such as metals or ceramics, it is common to perform observation at microscopic scale thanks to optical or electron microscopy. Still, this observation gives informations in 2D sections, and grains appear as surfaces instead of polyhedra. In this case, the 2D apparent equivalent radius of a given grain (denoted r bellow) is usually computed depending on its apparent area (S):

$$r = \sqrt{\frac{S}{\pi}} \,. \tag{1}$$

If the microstructure is equiaxed, each grain can be considered as almost spherical. Under this approximation, it is clear that its apparent radius is always smaller than its real radius. Let R be the radius of a spherical grain cut at random latitude; then, the probability of finding an apparent radius comprised in between r_1 and r_2 is:

$$P(r_1 < r < r_2) = \frac{1}{R}\left(\sqrt{R^2 - r_1^2} - \sqrt{R^2 - r_2^2}\right) \,. \tag{2}$$

If the grains size distribution is not monodisperse (different possible values for R), evaluating the distribution of R from the distribution of r is not straightforward. This introduces the so-called corpuscle problem. Saltikov[1] (1967) has proposed an algorithm to evaluate the distribution of R, knowing that of r. This algorithm uses the finite histogram (finite number of classes for r) to recursively evaluate the classes of R thanks to Eq. 2, starting from the upper values of R. It is worth mentioning that some authors use R as the upper limits of the bins (*e.g.*, Saltikov, 1967; Sahagian and Proussevitch, 1998), whereas others use R as the centers of the bins (*e.g.*, Higgins, 2000; Lopez-Sanchez and Llana-Fúnez, 2016).

Under certain circumstances, it is more convenient to characterize the grain size population using a parametric Probability Density Function (PDF) than using finite histograms. Indeed, parametric descriptions help to compare two distinct populations; in addition, they can be useful for random generation (see for instance DREAM.3D (Groeber and Jackson, 2014) or NEPER (Quey *et al.*, 2011) softwares). Rhines and Patterson (1982) have reported that in many recrystallized polycrystalline materials, the grain volume follows a lognormal distribution; thus, the equivalent grain radii follow a lognormal distribution too. Lopez-Sanchez and Llana-Fúnez (2016) have proposed the so-called two-step method, which consists in fitting a parametric distribution on the histograms given by the Saltikov method; the previous authors have used the lognormal distribution as the

[1] Sometimes misspelled "Saltykov" in the literature.

underlying distribution. However, the Saltikov method depends on the number of classes (or bins) of the histogram; thus, the results may be user-dependent. In addition, using a finite number of classes necessarily lessens the actual distribution of r since it reduces a possibly large amount of radius values into a restricted series of class–frequency pair values. For instance, Lopez-Sanchez and Llana-Fúnez (2016) have reported that the two-step method was efficient for comparative purposes with only 10 to 20 classes.

In a previous work (Depriester and Kubler, 2019), the present authors have numerically generated 3D radical-Voronoï tessellations (Okabe *et al.*, 2009) based on random packs of spheres whose radii follow lognormal distributions. Then, they have characterized the 2D sections in order to evaluate the ability of the two-step method to unfold the 3D distribution. A discrepancy between the results from the latter method and the theoretical distributions has been shown. Thus, they have proposed a set of correction coefficients to find the real parameters related to the lognormal distribution, namely the three-step method. For the sake of understanding, this set is detailed in Appendix. Depriester and Kubler have used experimental data (2D orientation map of uranium dioxide) in order to assess the three-step method. They have shown that assuming a lognormal distribution for the 3D grain size of the uranium dioxide was a reasonable assumption.

THE WICKSELL'S EQUATION

Let f be the PDF associated to the 3D radius of spherical particles. Let $\tilde{f}$ be the PDF associated to the radii of apparent 2D circles when the particles are cut at random latitudes. Wicksell (1925) gives the following relationship:

$$\tilde{f}(r) = \frac{r}{E}\int_r^\infty \frac{f(R)}{\sqrt{R^2-r^2}}\mathrm{d}R\,, \tag{3}$$

where E denotes the expectation of f, that is:

$$E = \int_0^\infty R f(R)\,\mathrm{d}R\,. \tag{4}$$

Bach (1958) has generalized the Wicksell's equation for thin sections of thickness t:

$$\tilde{f}(r) = \frac{2r}{2E+t}\left(\int_r^\infty \frac{f(R)}{\sqrt{R^2-r^2}}\mathrm{d}R + t f(r)\right)\,. \tag{5}$$

The Wicksell's equation can be considered as an integral transform from 3D to 2D domains. The latter author provides a general solution for Eq. 3, that is:

$$f(R) = \frac{-2ER}{\pi}\int_R^\infty \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\tilde{f}(r)}{r}\right)\frac{\mathrm{d}r}{\sqrt{r^2-R^2}}\,. \tag{6}$$

Wiencek *et al.* (2005) have used the Wicksell's equation to unfold the graphite particle size distribution in nodular cast iron from 2D sections. Assuming a Weibull distribution for f, they have used an inverse method to find the parameters for that distribution leading to the best fit between the empirical PDF and $\tilde{f}$, with respect to the least squares criterion. Keiding and Jensen (1972) have used the Maximum Likelihood (ML) method (Krishnamoorthy, 2016, Chap. 1) to evaluate the size distribution of liver cell nuclei from thin sections, thanks to Eq. 5.

AIMS OF THIS WORK

Let Ω_n be a finite sample consisting of n radius values, measured experimentally:

$$\Omega_n = (r_1, r_2, ..., r_n)\,.$$

Then, the empirical Cumulative Density Function (CDF) is:

$$F_n(r) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{r_k \le r}\,, \tag{7}$$

where $\mathbf{1}_X$ is the indicator function of event X (equal to 1 if X is true; 0 otherwise). Let $r_{\min}$ and $r_{\max}$ be the lowest and the largest elements in Ω_n, respectively. According to the definition given in Eq. 7, it is clear that:

$$F_n(r) = 0 \quad \text{if } r < r_{\min}\,; \tag{8a}$$
$$F_n(r) = 1 \quad \text{if } r \ge r_{\max}\,. \tag{8b}$$

Because of the way F_n is defined, it is not differentiable. As a result, the PDF cannot be computed without a regularization procedure, such as linear interpolation or smoothing (Anderssen and Jakeman, 1975). For instance, Jeppsson *et al.* (2011) have used the so-called kernel density estimation to get a continuous description of the empirical PDF. In addition, the estimation of the expectation (E) is not straightforward. As a consequence, Eq. 6 cannot be easily evaluated. For their work, Keiding and Jensen (1972) have assumed that the size distribution of liver cell follows a mixture of three χ-distribution. Indeed, the Bach's transform Eq. 5 of such a mixture can be computed analytically, allowing for ML.

The aim of this work is to provide a new method for the resolution of the corpuscle problem, that is solving Eq. 3. This method should not be user-dependent and it should use the empirical CDF as is, instead of binning it into classes. First, an algorithm is proposed to find the optimal set of parameters of a given parametric distribution by means of Minimum Distance Estimation (MDE). This MDE is computed on the transformed distribution, with

respect to the Wicksell's equation (Eq. 3). Then, the proposed algorithm is applied on an experimental set, provided by a previous work (Depriester and Kubler, 2019), investigating some well-known distributions (monodisperse, uniform, normal, lognormal etc.). Finally, the results are discussed. They are compared with those given by other stereology techniques, such as the Saltikov method, and the influence of the initial guesses for the MDE is discussed. When applicable, the benefit of using analytical equations is shown. Finally, the influence of the sample size on the accuracy of the results is studied.

MATERIAL AND METHODS

NUMERICAL INTEGRATION OF THE WICKSELL'S EQUATION

Working on a finite number of values for the radius, Eq. 3 becomes:

$$\forall i = 1,2,...,m \quad \tilde{f}(r_i) = \frac{r_i}{E}\int_{r_i}^{\infty} \frac{f(R)\,\mathrm{d}R}{\sqrt{R^2 - r_i^2}}\,, \tag{9}$$

where $(r_1, r_2, ..., r_m)$ correspond to the integration points (m increasing values).

Let $\tilde{F}$ be the CDF of $\tilde{f}$:

$$\tilde{F}(r) = \int_0^r \tilde{f}(x)\,\mathrm{d}x\,. \tag{10}$$

Based on Eq. 10, one can evaluate the CDF by cumulative trapezoidal integration, that is:

$$\tilde{F}(r_i) \approx \begin{cases} \frac{\tilde{f}(r_1)r_1}{2} & \text{if } i = 1\,, \\ \tilde{F}(r_{i-1}) + \frac{\tilde{f}(r_i)+\tilde{f}(r_{i-1})}{2}\delta r_i & \text{otherwise,} \end{cases} \tag{11}$$

with:

$$\delta r_i = r_i - r_{i-1}\,,$$

and assuming $\tilde{F}(0) = 0$.

Because of the spatial resolution, the minimum possible value for the measured radius is necessarily greater than 0. In addition, the largest experimental radius is necessarily finite. Let $\tilde{F}^*$ be the truncated CDF of $\tilde{F}$, taking into account the aforementioned bounds. It comes:

$$\tilde{F}^*(r_i) = \frac{\tilde{F}(r_i) - \tilde{F}(r_{\min})}{\tilde{F}(r_{\max}) - \tilde{F}(r_{\min})}\,. \tag{12}$$

[2]Usually denoted $n\omega^2$ in the literature.

Eq. 12 ensures that:

$$\tilde{F}^*(r_i) = 0 \quad \text{if } r_i \le r_{\min}\,; \tag{13a}$$

$$\tilde{F}^*(r_i) = 1 \quad \text{if } r_i \ge r_{\max}\,. \tag{13b}$$

MINIMUM DISTANCE ESTIMATION

Let F be a theoretical CDF and F_n an empirical CDF computed from a sample Ω_n of size n. The Cramér–von Mises (CvM) criterion, denoted $\mathscr{O}$ below[2], can be used as a goodness of fit test. It is defined as follows (Anderson and Darling, 1952):

$$\mathscr{O} = n\int_0^{\infty} [F_n(r) - F(r)]^2\,\mathrm{d}F(r)\,. \tag{14}$$

Let f be a parametric PDF and $(\theta_1, \theta_2, ..., \theta_z) = \theta$ a vector containing its z parameters. Let $\tilde{f}$ be the Wicksell's transform of f, as defined in Eq. 3. One can define the CvM criterion depending on f and θ:

$$\mathscr{O}^f(\theta) = n\int_0^{\infty} \left[F_n(r) - \tilde{F}^*(r \mid \theta)\right]^2 \mathrm{d}\tilde{F}^*(r \mid \theta) \tag{15}$$

where $\tilde{F}^*$ is the truncated CDF of $\tilde{f}$, computed from Eqs. 9, 11 and 12. Eqs. 8 and 13 lead to:

$$F_n(r) = \tilde{F}^*(r) \quad \text{if: } r < r_{\min} \text{ or } r \ge r_{\max}\,.$$

Hence, Eq. 15 becomes:

$$\mathscr{O}^f(\theta) = n\int_{r_{\min}}^{r_{\max}} \left[F_n(r) - \tilde{F}^*(r \mid \theta)\right]^2 \mathrm{d}\tilde{F}^*(r \mid \theta)\,. \tag{16}$$

Eq. 16 can be evaluated using the trapezoidal integration. Thus, for a given PDF f, one can find the optimal set of parameters θ_{opt} leading to the least value of $\mathscr{O}^f(\theta)$. That is:

$$\theta_{\text{opt}} = \underset{\theta}{\operatorname{Arg\,min}}\left(\mathscr{O}^f(\theta)\right)\,.$$

This method is usually known as the Minimum Distance Estimation (MDE). The corresponding algorithm is illustrated in Fig. 1.

For a given PDF f, the present algorithm requires to initialize the set of parameters (θ_0 in Fig. 1). This value is usually referred as the "initial guess" in the literature. Its influence on the accuracy of the method is discussed further throughout this paper.

One may note that ML could also be used here. Nevertheless, MDE is known to be more robust against outliers (Woodward *et al.*, 1984).

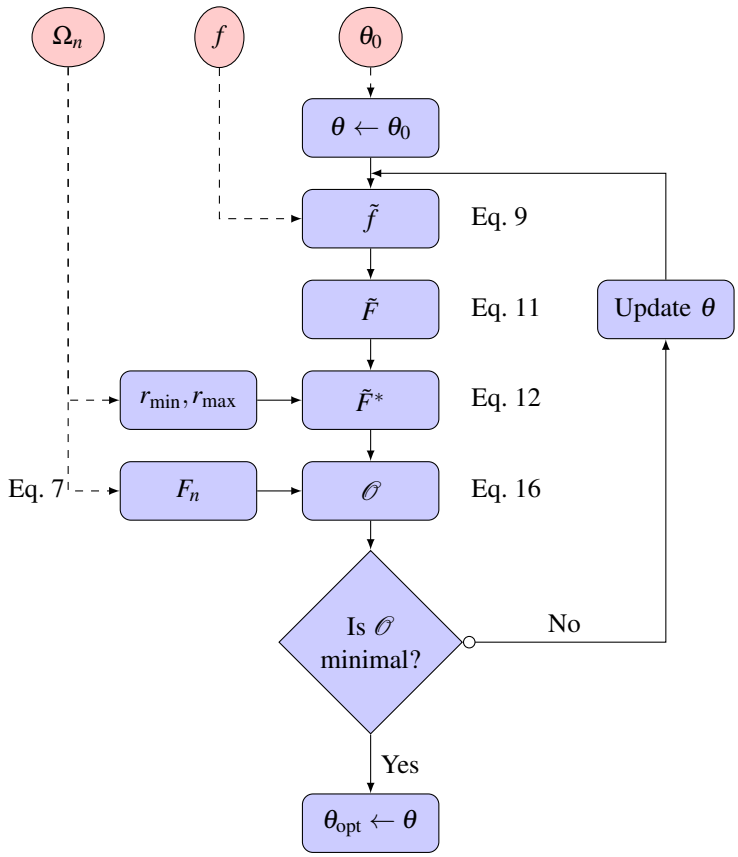

Fig. 1. *Schematic representation of the MDE: assuming a parametric PDF f, the algorithm aims to find the best set of related parameters (θ_{opt}), compared to experimental data (Ω_n), with respect to the CvM criterion ($\mathscr{O}$). θ_0 denotes the initial set of parameters to be used. The way θ is updated depends on the minimizing procedure.*

CRITICAL VALUES FOR THE CVM TEST

At a given significance level α, one can evaluate whether the null hypothesis should be rejected or not with respect to the CvM criterion. Let κ be the critical value of α at which the null hypothesis should be rejected; Csorgo and Faraway (1996) have built a correspondence between the CvM criterion and κ, depending on the sample size (n). Those tabular data are partially summed up in the "supplementary material" section of the online version of the journal. Thus, based on the CvM criterion, κ can be estimated thanks to the aforementioned data, assuming that:

- $\mathscr{O}$ follows a linear interpolation as function of $\log(n)$, where log denotes the natural logarithm;
- κ follows a linear interpolation as function of $\mathscr{O}$.

IMPLEMENTATION

All the methods detailed in the previous sections have been implemented in MATLAB®. Improper integrals (see Eqs. 4 and 9) were computed using the built-in `integral` command whereas the trapezoidal integrations in Eqs. 11 and 16 were computed using the `cumtrapz` and `trapz` commands, respectively. The latter were performed using $m = 1000$ equally spaced integration points between $r_{\min}$ and $r_{\max}$.

The MDE was performed using the `fminsearch` command. This command uses the simplex method (Lagarias *et al.*, 1998) to perform minimization. It is efficient when the investigated space is small (low value of z) and if the cost function has no local minima (Nelder and Mead, 1965).

MATERIAL

Uranium dioxide (UO_2), introduced in a previous paper (Depriester and Kubler, 2019), has been used as a test material. Its microstructure was imaged by Electron Backscattered Diffraction (EBSD) mapping and the EBSD data were processed to reconstruct the grains using of misorientation threshold of 5°. Grains cropped by the region of interest were removed from the data, resulting into $n = 4264$ individual grains. Fig. 2 illustrates the final dataset.

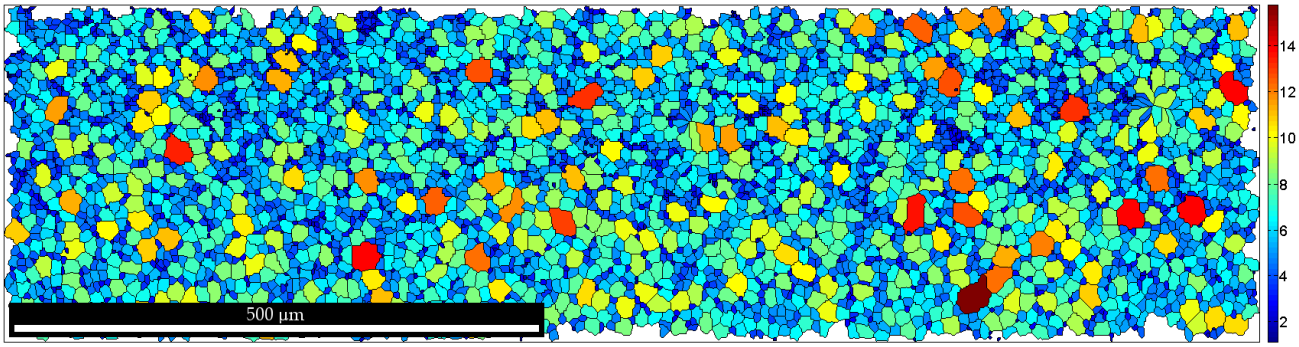

Fig. 2. *Reconstructed grains in* UO_2*: the colour indicates the corresponding equivalent radius (*µm*) (Depriester and Kubler, 2019).*

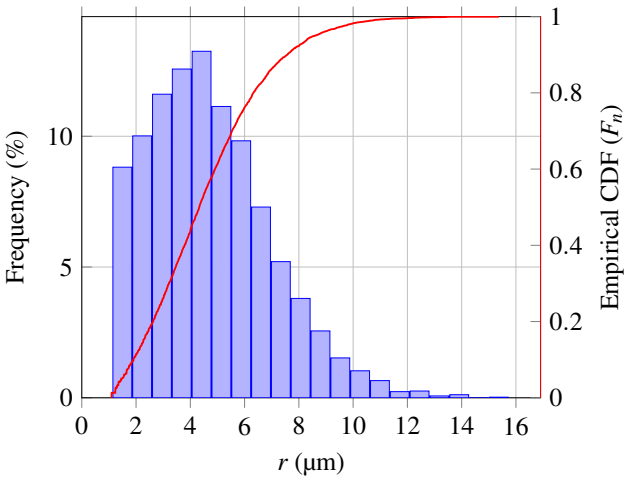

Fig. 3. *Histogram of equivalent radii, computed from the area of 2D apparent grains of* UO_2*. The red solid curve illustrates the empirical CDF.*

Based on their area, their apparent equivalent radii were computed with respect to Eq. 1. Fig. 3 gives the 2D histogram of the apparent radii.

In addition, the empirical CDF, computed from Eq. 7, is plotted in this figure. One may notice the threshold effect due to the resolution of the EBSD data. Indeed, the step size was 2 µm; thus, the minimum area was 4 μm^2. As a result, the equivalent radius, as defined in Eq. 1, was always larger than 1.128 µm. It is worth mentioning that the quality of the original EBSD data allows to consider all grains, even if some of them contain only one pixel. Indeed, the frequency at lower values in Fig. 3 appear reasonably correct ; in other cases, a minimal size may be introduced.

The following parametric distributions are investigated: monodisperse, uniform, positive normal, lognormal, Gamma, Weibull and Rayleigh. All those distributions and their related parameters are detailed in Table 1.

The Rayleigh distribution is a special case of the Weibull distribution (with $k = 2$). Wicksell (1925) has shown that its transform (Eq. 3) keeps the Rayleigh distribution unchanged (*i.e.*, its PDF is an eigenfunction with eigenvalue 1). The Wicksell's transforms of monodisperse and uniform distributions can be computed analytically (see Eqs. 22 and 24 in Appendix, respectively). The benefit of using the analytical equations is detailed bellow.

RESULTS

Results from the MDEs performed on the aforementioned distributions and based on the experimental data are summed up in Table 2. This table gives the optimal parameters (denoted θ_{opt} above) with each investigated type of distributions.

Fig. 4 illustrates the corresponding truncated CDFs ($\tilde{F}^*$).

For the sake of comparison, the empirical CDF is also given. Insets in Fig. 4 help to visualize the differences between the different CDFs. It appears that the monodisperse and uniform distributions lead to large discrepancies. Conversely, the consistency appears to be good with the other investigated distributions since they are hard to distinguish. This result is consistent with the corresponding values of the CvM criterion given in Table 2. Indeed, it appears that the normal distribution results in good correlation ($\mathscr{O} = 0.0935$) compared to the monodisperse and uniform distributions ($\mathscr{O} = 73.5$ and 2.51, respectively).

Since MDE using the Rayleigh distribution can be considered as a constrained MDE using the Weibull distribution (with $k = 2$), the corresponding CvM criterion ($\mathscr{O} = 0.202$) is larger than that of the (unconstrained) Weibull distribution ($\mathscr{O} = 0.123$). Nevertheless, the Weibull distribution found here is

Table 1. *List of parametric Probability functions investigated in this work to unfold the distribution given in Fig. 3.*

Name	Parameter(s)	PDF
Monodisperse	Unique radius: E	$f_{\text{mono}}(R\,\vert\,E) = \delta(R-E)$
Uniform	Lower bound: R_{min} Upper bound: R_{max}	$f_{\text{uni}}(R\,\vert\,R_{\text{min}},R_{\text{max}}) = \begin{cases} \frac{1}{R_{\text{max}}-R_{\text{min}}} & \text{if } R_{\text{min}} \leq R \leq R_{\text{max}} \\ 0 & \text{otherwise} \end{cases}$
Positive normal	Mode: R_{m} Shape parameter: σ	$f_{\mathcal{N}^+}(R\,\vert\,R_{\text{m}},\sigma) = \frac{1}{\sigma(1-\Phi_0)}\phi\left(\frac{R-R_{\text{m}}}{\sigma}\right)$ with: $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)$ and $\Phi_0 = \frac{1}{2}\left[1+\text{erf}\left(\frac{-R_{\text{m}}}{\sigma\sqrt{2}}\right)\right]$
Lognormal	Location: μ Shape: σ	$f_{\log\mathcal{N}}(R\,\vert\,\mu,\sigma) = \frac{1}{R\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln R-\mu)^2}{2\sigma^2}\right)$
Gamma	Scale: θ Shape: k	$f_{\gamma}(R\,\vert\,k,\theta) = \frac{1}{\Gamma(k)\theta^k}R^{k-1}\exp\left(-\frac{R}{\theta}\right)$ with: $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)\,\text{d}x$
Weibull	Scale: λ Shape: k	$f_{\text{W}}(R\,\vert\,\lambda,k) = \frac{k}{\lambda}\left(\frac{R}{\lambda}\right)^{k-1}\exp\left(-\left(\frac{R}{\lambda}\right)^k\right)$
Rayleigh	Mode: R_{m}	$f_{\text{R}}(R\,\vert\,R_{\text{m}}) = \frac{R}{\sigma^2}\exp\left(\frac{-R^2}{2\sigma^2}\right)$

Table 2. *Optimal parameters resulting from MDEs and corresponding CvM criterion values, depending on the parametric distribution used for MDE (see Table 1).*

Distribution	Parameters	CvM test	Crit. significance	Mode
	θ_{opt}	$\mathcal{O}$	κ	R_{m}
Monodisperse	$E = 5.695$	73.5	> 0.999	5.695
Uniform	$R_{\text{min}} = 1.073$ $R_{\text{max}} = 8.250$	2.51	> 0.999	N/A
Normal	$R_{\text{m}} = 3.876$ $\sigma = 2.816$	0.094	0.370	3.876
Lognormal	$\mu = 1.519$ $\sigma = 0.428$	0.853	0.992	3.803
Gamma	$k = 4.724$ $\theta = 1.026$	0.350	0.901	3.822
Weibull	$\lambda = 5.203$ $k = 2.106$	0.123	0.511	3.833
Rayleigh	$R_{\text{m}} = 3.612$	0.202	0.730	3.612

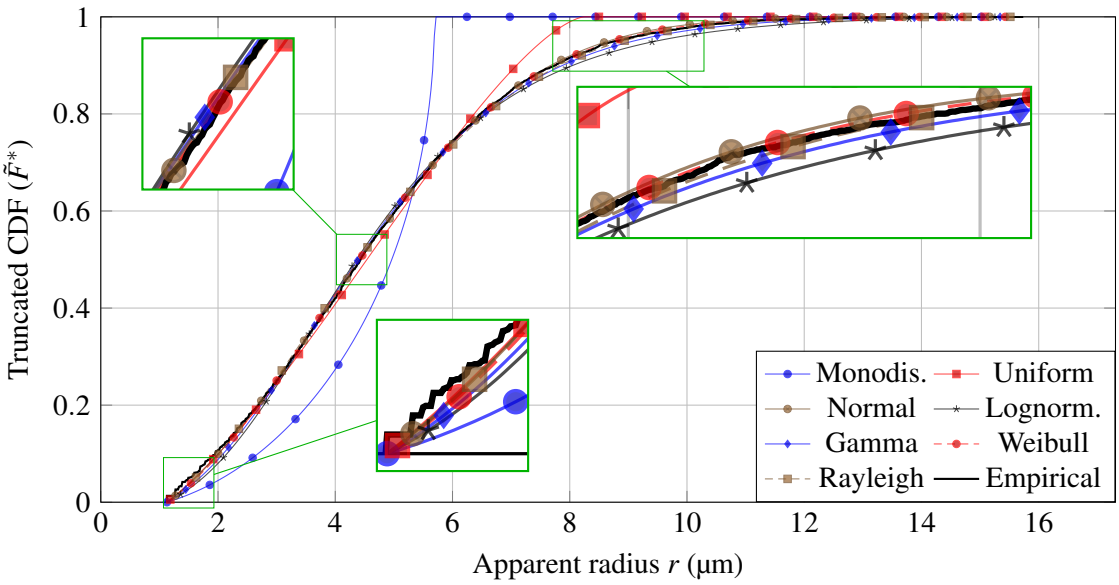

Fig. 4. *Results from fitting by MDE using the standard distributions: transformed CDFs, as compared to the empirical CDF (thick black curve). Insets give details of the gaps near the tails and the median values of the distributions.*

close to a Rayleigh distribution because k is close to 2 ($k = 2.106$).

Table 2 also gives the results from MDEs in terms of critical values for the significance level (κ). Thus, using the normal distribution results in κ around 0.370, whereas that of the lognormal distribution is close to 1. This result is inconsistent with the usual hypothesis that R follows a lognormal distribution in recrystallized material (Rhines and Patterson, 1982). Indeed, sintering usually results in sufficient grain growth for considering that the manufactured material is fully recrystallized and that its grain size distribution follows a lognormal rule (Readey and Readey, 1986). The values for κ found here are quite high, even for the normal distribution. This result may be surprising considering the apparent good fits in Fig. 4. As a reminder, we have here $n = 4264$. For such large sample, the CvM criterion is somehow severe. Other parametric distributions could eventually lead to lower significance levels.

For the sake of comparison, the PDFs found for each investigated parametric distribution (see Table 1) are shown in Fig. 5. Since the monodisperse and the uniform distributions can be considered as irrelevant, they are not plotted in Fig. 5.

In this figure, it appears that all the plotted PDFs reach their maximum value around the same location (mode R_{m}). Indeed, as summed up in Table 2, the

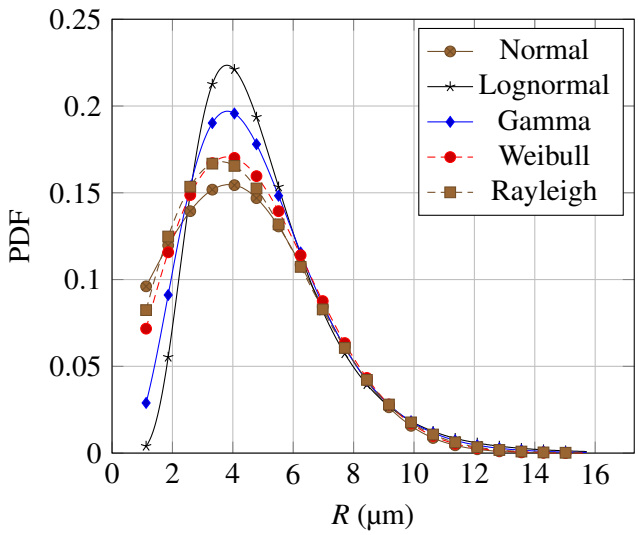

Fig. 5. *Representation of the PDFs using the parameters given in Table 2 (the monodisperse and the uniform distributions are not represented here).*

modes of the normal, lognormal, Gamma and Weibull distributions are equal to around 3.80 µm, Still, the corresponding PDFs are quite different from each other, specially in terms of dispersion (distribution spreads). In Fig. 5, it is also clear that the Weibull and the Rayleigh distributions are almost superimposed; this result is related to the fact that the k parameter of the former distribution is close to 2.

It is worth mentioning that, at first, attempts to perform MDEs using the Kolmogorov–Smirnov

(KS) criterion (Wolstenholme, 2017, Ch. 4) as a distance estimator (denoted $\mathscr{O}$ here) have been made. However, the MDE fails to converge in this case because of the small gap between the empirical and the transformed CDFs at lower radii, as evidenced by the left-hand inset in Fig. 4. Indeed, this gap appears to be almost irreducible. In essence, the KS test only uses a particular value of the gap (maximum absolute difference between the empirical and the theoretical CDFs) whereas the CvM criterion, as defined in Eq. 14, is integrated over the whole domain, resulting in a much more stable criterion.

DISCUSSION

In this section, the following investigations are made:

–comparison with other stereology techniques;

–effect of the initial guess on the results from MDE;

–influence of using analytical Wicksell's transforms instead of using numerical integration;

–influence of the sample size on the accuracy of the results from MDE.

COMPARISON WITH OTHER STEREOLOGY METHODS

The results from MDE can be compared with those from other stereology techniques, namely the Saltikov method (Saltikov, 1967), the two-step method (Lopez-Sanchez and Llana-Fúnez, 2016) and the three-step method (Depriester and Kubler, 2019). Fig. 6 schematically illustrates the workflow to be used in order to utilize those techniques.

In order to compare the results from MDE to those from the Saltikov method, the latter has been applied on the example dataset Ω_n. Since there is no general rule about the number of bins to be used in the Saltikov method (denoted N hereafter), the latter has been applied with 10 to 30 bins. In each case, the CvM test ($\mathscr{O}$), as defined in Eq. 16, has been computed. It is worth reminding that the Saltikov method gives the unfolded radius distribution; thus, it must be *refolded* before computing the CvM test. Eq. 27 allows for refolding a finite histogram thanks to the Wicksell's equation (see Appendix). Fig. 7 illustrates the evolution of $\mathscr{O}$ as a function of N.

Thus, it is clear that the best correlation is reached when using 14 bins (with $\mathscr{O} = 7.80$). This value lies

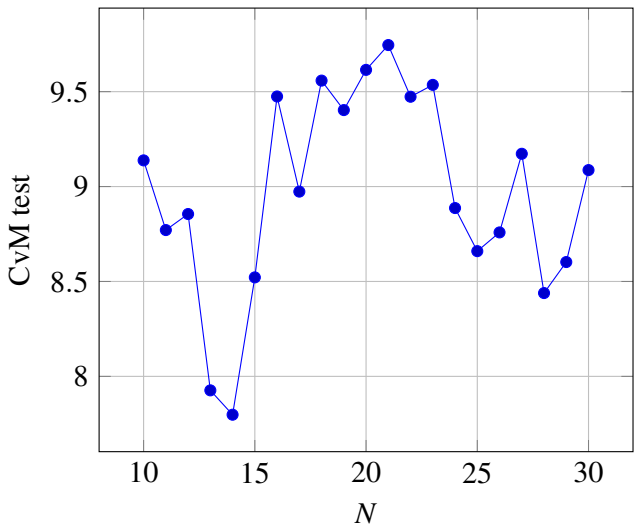

Fig. 7. *CvM test Eq. 16 when using the CDFs given by the Saltikov method, as a function of the number of bins used for this method.*

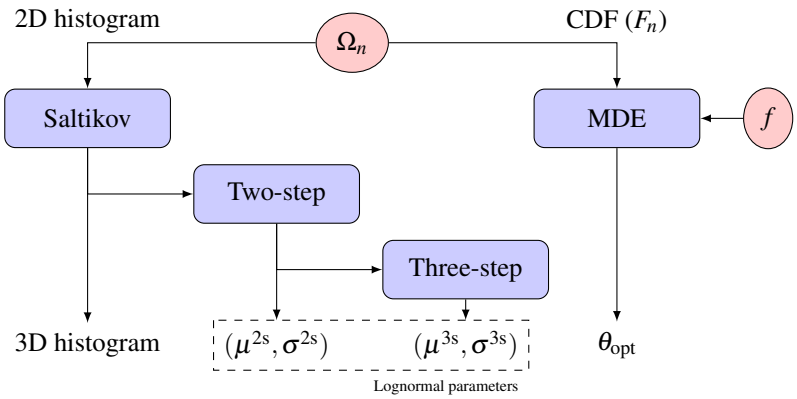

Fig. 6. *Schematic representation of the investigated stereology techniques: the Saltikov method gives the unfolded 3D histogram, whereas the two-step method uses this latter to evaluate the lognormal parameters; the three-step method aims to increase the accuracy of the two-step method.*

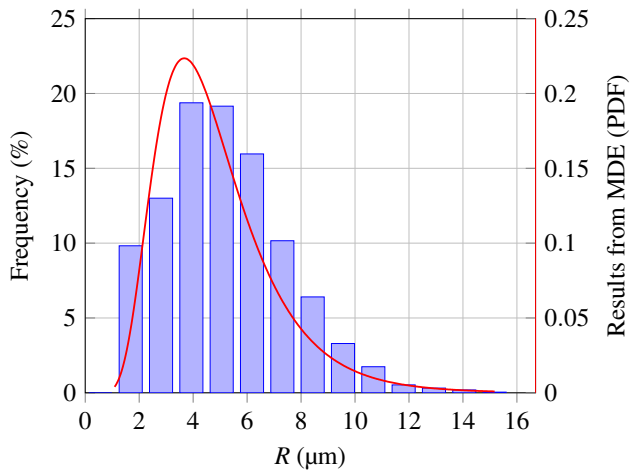

Fig. 8. *Histogram of the unfolded 3D radius distribution given by the Saltikov method using 14 bins (the first one is not visible) and optimal PDF given by MDE (red solid curve)*

within the conventional range, between 10 to 20, for N (Lopez-Sanchez and Llana-Fúnez, 2016). Still, this result may depend on the sample and no general rule may be raised. Fig. 8 shows the resulting histogram from the Saltikov method.

As a comparison, the results from MDE (using normal distribution) has been plotted as well. It appears that the histogram generated by the Saltikov method is slightly shifted toward larger radii, compared to that of the present algorithm. For instance, the Saltikov method results in a null frequency for $R < 1.68\,\mu$m. This lower cut-off has no physical relevance, considering the manufacturing means of the investigated material (sintering).

Based on the previous results, the two-step method (Lopez-Sanchez and Llana-Fúnez, 2016) has been applied using 14 bins on the studied sample. The resulting parameters for the lognormal distribution are given in Table 3.

Table 3. *CvM test given by the Saltikov method (Saltikov, 1967), the two-step method (Lopez-Sanchez and Llana-Fúnez, 2016) and three-step method (Depriester and Kubler, 2019) and corresponding lognormal parameters (when applicable).*

	CvM test	μ	σ
MDE (w/ lognormal dist.)	0.853	1.519	0.428
Saltikov method	7.80	N/A	N/A
Two-step method	39.3	1.630	0.507
Three-step method	3.98	1.612	0.356

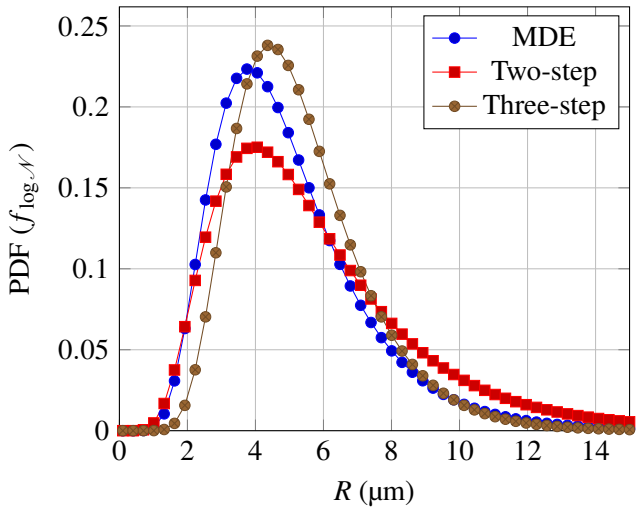

Fig. 9. *Result of MDE performed using the lognormal distribution, compared to that obtained thanks to the two-step method (Lopez-Sanchez and Llana-Fúnez, 2016) and the three-step method (Depriester and Kubler, 2019).*

As a comparison, Fig. 9 shows the resulting PDF, to be compared with that given by MDE.

It appears that the MDE results in a sharper distribution, as evidenced by lower value of σ in Table 3. On the opposite, the modes are almost equal in each case. Indeed, the present work leads to the modal value $R_{\mathrm{m}} = 3.80\,\mu$m whereas the two-step method leads to $R_{\mathrm{m}} = 3.95\,\mu$m. This result is in accordance with the hypothesis that the mode is the key value in the corpuscle problem, as stated before. According to Table 3, the two-step method results in a larger CvM criterion than the Saltikov method. This may be because the fit is not good between the lognormal distribution and the histogram during the two-step process.

Once the two-step method has been applied, its results can be adjusted according to the three-step method (Depriester and Kubler, 2019), as illustrated in Fig. 6. Further details of this method are given in Appendix. The previous values for μ and σ, once adjusted with respect to Eqs. 19 and 21, are given in Table 3. Fig. 9 illustrates the corresponding CDF, to be compared with that given by MDE. It appears that the three-step method slightly overestimates both the modal value and sharpness of the lognormal distribution. According to Fig. 9, the PDF found using MDE seems to be comprised in between those found using the two-step and three-step methods. As evidenced by the low value of the CvM test, as given in Table 3, it is clear that the three-step method gives accurate results, close to the optimal ones given by MDE.

INITIAL GUESS FOR THE MDE

The aforementioned MDEs were performed with different initial guesses (denoted θ_0 in Fig. 1). In each case, they lead to almost the same results. As an example, Table 4 gives the results of MDEs performed with the normal distribution, depending on the initial guess.

Table 4. *Results from MDE using the normal distribution and corresponding computation times (CPU), depending on the initial guess (θ_0).*

θ_0		θ_{opt}		CPU (s)
R_m	σ	R_m	σ	
1.0	0.1	3.876018	2.815846	121.1
10.	0.1	3.876088	2.815800	116.6
10.	10.	3.876057	2.815831	134.2
1.0	10.	3.876082	2.815809	120.5
3.876	2.816	3.876077	2.815807	48.7

Thus, the algorithm proposed in this paper appears to be stable and the simplex method is very efficient here. Table 4 also gives the computation time in each case. It appears that the initial guess only influences the number of minimization iterations before converging. As a conclusion, the initial guess may be arbitrary chosen if the computational time is not a matter of choice.

ON THE USE OF ANALYTICAL EQUATIONS

It is worth mentioning that the computational times required by the MDE (*e.g.*, those given in Table 4) are due to a large extent to the improper integrals Eqs. 4 and 9. Indeed, when applicable, the analytical resolution of the Wicksell's equation (Eqs. 22 and 24 for monodisperse and uniform distributions, respectively) leads to very fast MDEs. For instance, Table 5 gives the results from MDEs using the uniform distribution with numerical resolution of the Wicksell's transform, as detailed above, to be compared with that using the analytical equation Eq. 24.

Table 5. *Results from MDE using the uniform distribution and corresponding computation times (CPU), depending on the resolution method.*

Resolution method	R_{min}	R_{max}	CvM test	CPU (s)
Numerical	1.121	8.253	2.5067	102.93
Analytical	1.089	8.250	2.5071	0.0497

It appears that both the methods are consistent with each other because they lead to almost the same values for the optimal parameters and the CvM criterion. Nevertheless, using the analytical equations is more than 2000 times faster than using the numerical integration.

INFLUENCE OF THE SAMPLE SIZE

It has been shown before that the normal distribution leads to the best fit with respect to the CvM criterion; thus, only this distribution will be used hereafter. The sample used here is composed of 4264 grains. Hence, this dataset has been randomly down-sampled in order to investigate the influence of the sample size (denoted n above) on the accuracy of the results. For each investigated value of n, 100 different random samplings have been done, from which the MDEs have been performed. As a result, 100 values of R_m, 100 values of σ and 100 values of κ have been found for each value of n. In each case, the mean values and the standard deviations of those sets have been computed.

Fig. 10 shows the results from MDEs (mean values of R_m, σ and κ, and their related standard deviations as well) as functions of n.

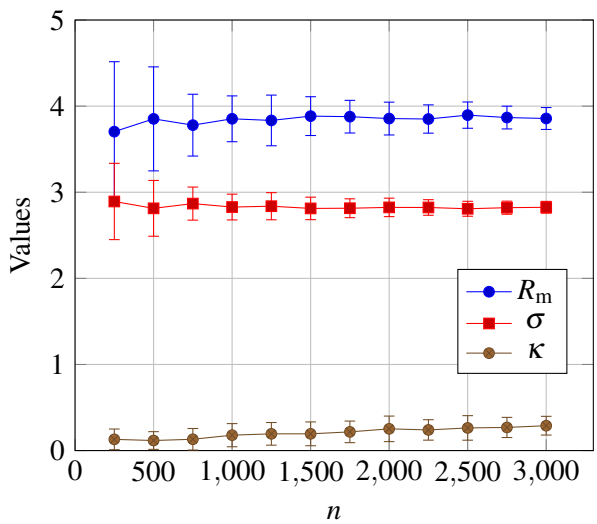

Fig. 10. *Optimal parameters for normal distribution and resulting values for κ, depending on the sample size (n): evolutions of the mean values as functions of n; error bars give the standard deviations (± 1 std).*

It appears that the mean values of R_m and σ are almost constant when $n > 1000$. In addition, when $n > 1000$, their standard deviations are small, compared to the mean values. Indeed, the corresponding coefficient of variations (ratio of the standard deviations to the means) are lower than 7 % in this case. Hence, it can be

concluded that at least 1000 radius values are required to get sufficient accuracy. This number is the same as that required for the Saltikov method (Lopez-Sanchez and Llana-Fúnez, 2016).

Surprisingly, κ appears to be slightly increasing with n; nevertheless, the corresponding standard deviations are almost constant (around 0.126). As a conclusion, sample size larger than 1000 does not result in a significant increase in accuracy for a given PDF. However, it is clear that the larger sample is, the better one can estimate whether the investigated PDF is a good candidate or not.

CONCLUSION

In this work, the Wicksell's equation has been used to evaluate the grain size distribution in a polycrystal from 2D sections. The proposed algorithm is made on the assumption that the grain radii R follow one of the standard distribution functions (e.g normal, lognormal etc.). Then, MDEs are performed on the transformed distributions in order to evaluate the corresponding distribution's parameters. The CvM criterion is used as a goodness of fit test.

The proposed algorithm has been applied on a given material (uranium dioxide, imaged by EBSD). It has been shown that the normal distribution has led to the best results, among all investigated distribution functions. The use of the CvM criterion together with the simplex optimization technique results in a very stable MDE. The results from MDE have been compared with those from other stereology techniques (Saltikov, two-step and three-step methods). A criterion, based on the CvM test, is proposed to choose the best number of bins for the Saltikov method. The following statements have been made:

- the proposed algorithm results in a distribution with larger frequencies at lower radii than the Saltikov method;
- the best accuracy (lowest value of the CvM criterion) was reached when using 14 bins for the Saltikov method on the test material (although this value may be sample-dependent);
- since the two-step method is based on the Saltikov method, it results in a decrease of accuracy (greater value of the CvM criterion), compared to the Saltikov method alone;
- the three-step method, as proposed by Depriester and Kubler (2019), helps to improve the accuracy of the two-step method;
- in all cases (including MDE using different distributions), the modal values of the unfolded distributions were almost the same.

The last statement implies that the proposed algorithm gives very accurate estimation of the mode of the unfolded distribution, compared to its other parameters (*e.g.*, dispersion).

Finally, the influence of the sample size (number of radius values from 2D section) has been investigated. It has been stated that at least 1000 values are required to obtain the PDF parameters with good accuracy. Larger samples may help to determine whether the investigated PDF is a good candidate or not. The values of κ, estimated from the CvM criterion, may be used with caution because of possible large values, specially for large samples.

Since this method uses the empirical CDF as is (no binning), it may be more accurate than any Saltikov-based technique. In addition, the proposed method is purely deterministic (not user-dependent) and may work on various standard distributions. Still, it requires to make some assumptions about the potential parametric distributions to be used.

As a further work, the proposed algorithm could be applied on input data generated from a known 3D distribution (*e.g.*, using random generation of 3D aggregates, then slicing); thus, its results could be compared to the actual 3D distribution. Comparison could also be made between MDE and ML in terms of accuracy and robustness.

DATA AVAILABILITY

Datasets and MATLAB® routines related to this article can be found at https://doi.org/10.24433/CO.7108475.v2, hosted at Code Ocean (Depriester, 2019).

APPENDIX: THE THREE STEP METHOD

This section briefly describes the method proposed in (Depriester and Kubler, 2019) to evaluate the distribution of 3D equivalent radius R, based on 2D sections, namely the three-step method.

When performing radical Voronoï tessellation (Okabe *et al.*, 2009) from a random pack of spheres, one can define the equivalent radius of each Voronoï cell depending on its volume V:

$$R^{\text{cell}} = \sqrt[3]{\frac{3V}{4\pi}} \, .$$

If the sphere radii follow a lognormal distribution (see Table 1) with shape parameter σ^{sph} and expectation E^{sph}, Depriester and Kubler (2019) have shown that the resulting distribution of R^{cell} can be approximated with a lognormal distribution too. Let σ and E be the corresponding shape parameter and expectation, respectively. For $0 \leq \sigma^{\text{sph}} \leq 0.9$, the aforementioned authors have shown that the following relationships apply:

$$\frac{E}{E^{\text{sph}}} = 0.1123\left(\sigma^{\text{sph}}\right)^2 - 0.013\sigma^{\text{sph}} + 1.1587\ ; \tag{17a}$$

$$\sigma = 0.7166\sigma^{\text{sph}} + 0.0228\ . \tag{17b}$$

Depriester and Kubler have made 2D sections from the aforementioned tessellation, then they have applied the two-step method (Lopez-Sanchez and Llana-Fúnez, 2016) on the resulting size distribution in order to evaluate its ability to unfold the distributions. Let $\sigma^{2\text{s}}$ and $E^{2\text{s}}$ be respectively the resulting shape parameter and expectation given by the two-step method. Depriester and Kubler have reported the following approximations:

$$\frac{E^{2\text{s}}}{E^{\text{sph}}} = 0.2225\left(\sigma^{\text{sph}}\right)^2 + 0.1749\sigma^{\text{sph}} + 1.1505\ ; \tag{18a}$$

$$\sigma^{2\text{s}} = 1.0184\sigma^{\text{sph}} + 0.0341\ . \tag{18b}$$

Thus, Eqs. 17b and 18b lead to:

$$\sigma = 0.7037\sigma^{2\text{s}} - 0.0012 \tag{19}$$

whereas Eqs. 17a, 18a and 18b lead to:

$$\frac{E}{E^{2\text{s}}} = \frac{0.5047\left(\sigma^{2\text{s}}\right)^2 - 0.0939\sigma^{2\text{s}} + 5.404}{\left(\sigma^{2\text{s}}\right)^2 + 0.7323\sigma^{2\text{s}} + 5.337}\ . \tag{20}$$

This equation applies for $\sigma^{2\text{s}} \in [0.034, 0.951]$, according to Eqs. 17b and 18b. In this range, Eq. 20 can be approximated as follows:

$$\frac{E}{E^{2\text{s}}} \approx 0.0363\left(\sigma^{2\text{s}}\right)^3 - 0.0680\left(\sigma^{2\text{s}}\right)^2 - 0.1582\sigma^{2\text{s}} + 1.0127\ , \tag{21}$$

with relative error below 1.1×10^{-4}.

For lognormal distribution, it is worth reminding that the expectation can be computed from the location and scale parameters with (Krishnamoorthy, 2016, Ch. 22):

$$E = \exp\left(\mu + \frac{\sigma^2}{2}\right)\ .$$

APPENDIX: ANALYTICAL WICKSELL'S TRANSFORMS

MONODISPERSE DISTRIBUTION

In the case of monodisperse distribution (see Table 1), the Wicksell's equation (Eq. 3) becomes:

$$\tilde{f}_{\text{mono}}\left(r\,|\,E\right) = \begin{cases} \frac{r}{E}\int_r^{\infty} \frac{\delta(R-E)}{\sqrt{R^2-r^2}}\,\mathrm{d}R & \text{if } r < E \\ 0 & \text{if } r \geq E \end{cases}\ .$$

Thanks to the properties of the Dirac function, it comes:

$$\tilde{f}_{\text{mono}}\left(r\,|\,E\right) = \frac{r}{E}\frac{1}{\sqrt{E^2-r^2}} \quad \text{if } r < E\ .$$

Thus, the corresponding CDF is:

$$\begin{aligned}\tilde{F}_{\text{mono}}\left(r\,|\,E\right) &= \begin{cases} \int_0^r \tilde{f}_{\text{mono}}(x)\,\mathrm{d}x & \text{if } r < E \\ \int_0^E \tilde{f}_{\text{mono}}(x)\,\mathrm{d}x & \\ + \int_E^r \tilde{f}_{\text{mono}}(x)\,\mathrm{d}x & \text{if } r \geq E \end{cases} \\ &= \begin{cases} 1 - \frac{\sqrt{E^2-r^2}}{E} & \text{if } r < E \\ 1 & \text{if } r \geq E \end{cases}\ . \end{aligned} \tag{22}$$

This equation is consistent with Eq. 2 since:

$$P\left(r_1 < r < r_2\right) = \tilde{F}_{\text{mono}}\left(r_2\right) - \tilde{F}_{\text{mono}}\left(r_1\right)\ .$$

UNIFORM DISTRIBUTION

In the case of uniform distribution (see Table 1), the Wicksell's equation (Eq. 3) becomes:

$$\tilde{f}_{\text{uni}}\left(r\,|\,R_{\min}, R_{\max}\right) = \begin{cases} \frac{r}{E(R_{\max}-R_{\min})} \cdot \int_{R_{\min}}^{R_{\max}} \frac{\mathrm{d}R}{\sqrt{R^2-r^2}} & \text{if } r \leq R_{\min} \\ \frac{r}{E(R_{\max}-R_{\min})} \cdot \int_r^{R_{\max}} \frac{\mathrm{d}R}{\sqrt{R^2-r^2}} & \text{if } R_{\min} \leq r,\ r \leq R_{\max} \\ 0 & \text{if } r \geq R_{\max} \end{cases}\ . \tag{23}$$

By substitution, it can be demonstrated that:

$$\int \frac{\mathrm{d}R}{\sqrt{R^2-r^2}} = \log\left(R + \sqrt{R^2-r^2}\right) + C\ .$$

In addition, it is clear that:

$$E = \frac{R_{\max} + R_{\min}}{2}\ .$$

Thus, Eq. 23 can be rewritten:

$$\tilde{f}_{\text{uni}}\left(r\,|\,R_{\min},R_{\max}\right)=\begin{cases}\frac{2r}{R_{\max}^2-R_{\min}^2}\log\left(\frac{R_{\max}+\sqrt{R_{\max}^2-r^2}}{R_{\min}+\sqrt{R_{\min}^2-r^2}}\right) & \text{if } r\leq R_{\min}\\ \frac{2r}{R_{\max}^2-R_{\min}^2}\log\left(\frac{R_{\max}+\sqrt{R_{\max}^2-r^2}}{r}\right) & \text{if } R_{\min}\leq r\leq R_{\max}\\ 0 & \text{if } r\geq R_{\max}\end{cases}.$$

The corresponding CDF Eq. 10 can be evaluated using a symbolic computation software, such as Wolfram *Mathematica*® (Wolfram Research, Inc., 2015). Indeed, the latter gives:

$$\tilde{F}_{\text{uni}}\left(r\,|\,R_{\min},R_{\max}\right)=\begin{cases}1-\frac{\gamma(r)+r^2\log\left(R_{\min}+\sqrt{R_{\min}^2-r^2}\right)-R_{\min}\sqrt{R_{\min}^2-r^2}}{R_{\max}^2-R_{\min}^2} & \text{if } r\leq R_{\min}\\ 1-\frac{\gamma(r)+r^2\log(r)}{R_{\max}^2-R_{\min}^2} & \text{if } R_{\min}\leq r\leq R_{\max}\\ 1 & \text{if } r\geq R_{\max}\end{cases}\tag{24}$$

with:

$$\gamma(r)=R_{\max}\sqrt{R_{\max}^2-r^2}-r^2\log\left(R_{\max}+\sqrt{R_{\max}^2-r^2}\right).$$

FINITE HISTOGRAM

This section proposes a procedure to compute the Wicksell's transform of a finite histogram. This procedure is used to refold the distribution given by the Saltikov method before computing the CvM test.

Given an histogram consisting of N classes, it is assumed that within each class k, the distribution is homogeneous between the corresponding lower bound ($R_{\min}^k$) and upper bound ($R_{\max}^k$). In other words, the corresponding PDF is:

$$f_{\text{hist}}\left(R\,|\,\text{bins}\right)=\sum_{k=1}^{N}\text{Freq}^k.f_{\text{uni}}\left(R\,\middle|\,R_{\min}^k,R_{\max}^k\right)$$

where Freq^k is the relative frequency of the k-th class. Thus, the Wicksell's transform Eq. 3 of f_{hist} is:

$$\tilde{f}_{\text{hist}}\left(r\,|\,\text{bins}\right)=\frac{r}{E}\sum_{k=1}^{N}\text{Freq}^k.\tilde{f}^k(r)\tag{25}$$

with:

$$\begin{aligned}\tilde{f}^k(r)&=\int_r^{\infty}\frac{f_{\text{uni}}\left(R\,\middle|\,R_{\min}^k,R_{\max}^k\right)}{\sqrt{R^2-r^2}}\mathrm{d}R\\&=\frac{E^k}{r}\tilde{f}_{\text{uni}}\left(r\,\middle|\,R_{\min}^k,R_{\max}^k\right)\end{aligned}\tag{26}$$

where $\tilde{f}_{\text{uni}}$ is the Wicksell's transform of the uniform distribution, as defined in Eq. 23, and E^k the mid-point of the k-th class:

$$E^k=\frac{R_{\min}^k+R_{\max}^k}{2}.$$

Eqs. 25 and 26 give:

$$\tilde{f}_{\text{hist}}\left(r\,|\,\text{bins}\right)=\frac{1}{E}\sum_{k=1}^{N}\text{Freq}^k.E^k.\tilde{f}_{\text{uni}}\left(r\,\middle|\,R_{\min}^k,R_{\max}^k\right).$$

Here, the expectation is:

$$E=\sum_{k=1}^{N}\text{Freq}^k.E^k.$$

Finally, the corresponding CDF is:

$$\tilde{F}_{\text{hist}}\left(r\,|\,\text{bins}\right)=\frac{1}{E}\sum_{k=1}^{N}\text{Freq}^k.E^k.\tilde{F}_{\text{uni}}\left(r\,\middle|\,R_{\min}^k,R_{\max}^k\right)\tag{27}$$

where $\tilde{F}_{\text{uni}}$ is given in Eq. 24.

REFERENCES

Anderson T, Darling D (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann Math Stat 23:193–212.

Anderssen R, Jakeman A (1975). Abel type integral equations in stereology. II. Computational methods of solution and the random spheres approximation. J Microsc Oxford 105:135–53.

Bach G (1958). Über der Größenverteilung von Kugelschnitten in durchsichtigen schnitten endlicher Dicke. ZAMM Z Angew Math Me 38:256–8. In german.

Csorgo S, Faraway J (1996). The exact and asymptotic distributions of Cramer–von Mises statistics. J R Stat Soc B 58:221–34.

Depriester D (2019). Stereology problems: Solving the Wicksell's equation by minimum distance estimation. Published on Code Ocean. https://doi.org/10.24433/CO.7108475.v2.

Depriester D, Kubler R (2019). Radical Voronoï tessellation from random pack of polydisperse spheres: Prediction of the cells' size distribution. Comput Aided Des 107:37–49.

Groeber MA, Jackson MA (2014). DREAM.3D: A digital representation environment for the analysis of microstructure in 3D. Integr Mater Manuf Innov 3:5.

Higgins MD (2000). Measurement of crystal size distributions. Am Mineral 85:1105–16.

Jeppsson J, Mannesson K, Borgenstam A, Agren J (2011). Inverse Saltykov analysis for particle-size distributions and their time evolution. Acta Mater 59:874–82.

Keiding N, Jensen ST (1972). Maximum likelihood estimation of the size distribution of liver cell nuclei from the observed distribution in a plane section. Biometrics 28:813–29.

Krishnamoorthy K (2016). Handbook of statistical distributions with applications. New York: Chapman and Hall/CRC.

Lagarias J, Reeds J, Wright M, Wright P (1998). Convergence properties of the Nelder-Mead Simplex method in low dimensions. SIAM J Optimiz 9:112–47.

Lopez-Sanchez M, Llana-Fúnez S (2016). An extension of the Saltykov method to quantify 3D grain size distributions in mylonites. J Struct Geol 93:149–61.

Nelder J, Mead R (1965). A simplex method for function minimization. Comput J 7:308–13.

Okabe A, Boots B, Sugihara K, Chiu SN (2009). Spatial tessellations: concepts and applications of Voronoi diagrams, 2nd Ed. Chichester: John Wiley & Sons.

Quey R, Dawson P, Barbe F (2011). Large-scale 3D random polycrystals for the finite element method: Generation, meshing and remeshing. Comput Method Appl M 200:1729–45.

Readey MJ, Readey DW (1986). Sintering of ZrO_2 in HCl atmospheres. J Am Ceram Soc 69:580–2.

Rhines F, Patterson B (1982). Effect of the degree of prior cold work on the grain volume distribution and the rate of grain growth of recrystallized aluminum. Metall Trans A 13:985–93.

Sahagian DL, Proussevitch AA (1998). 3D particle size distributions from 2D observations: stereology for natural applications. J Volcanol Geoth Res 84:173–96.

Saltikov S (1967). The determination of the size distribution of particles in an opaque material from a measurement of the size distribution of their sections. In: Elias H, ed. Stereology. Berlin, Heidelberg: Springer.

Wicksell S (1925). The corpuscle problem: A mathematical study of a biometric problem. Biometrika 17:84–99.

Wiencek K, Skowronek T, Khatemi B (2005). Graphite particle size distribution in nodular cast iron. Metall Foundry Eng 31:167–73.

Wolfram Research, Inc. (2015). Mathematica, Version 10.3. Champaign, IL.

Wolstenholme LC (2017). Reliability modelling: a statistical approach. New York: Routledge.

Woodward WA, Parr WC, Schucany WR, Lindsey H (1984). A comparison of Minimum Distance and Maximum Likelihood Estimation of a mixture proportion. J Am Stat Assoc 79:590–8.