Journal of Statistical Physics manuscript No.
(will be inserted by the editor)

Data-driven selection of tessellation models describing polycrystalline microstructures

Ondřej Šedivý[1] · Daniel Westhoff[1] · Jaromír Kopeček[2] · Carl E. Krill III[3] · Volker Schmidt[1]

Received: date / Accepted: date

Abstract Tessellation models have proven to be useful for the geometric description of grain microstructures in polycrystalline materials. With the use of a suitable tessellation model, the complex morphology of grains can be represented by a small number of parameters assigned to each grain, which not only entails a significant reduction in complexity, but also facilitates the investigation of certain geometric features of the microstructure. However, for a given set of microstructural data, the choice of a particular geometric model is traditionally based on researcher intuition. The model should provide a sufficiently good approximation to the data, while keeping the number of parameters small. In this paper, we discuss general aspects of the process of model selection and suggest several criteria for selecting an appropriate candidate from a certain set of tessellation models. The choice of candidate represents a trade-off between accuracy and complexity of the model. Here, the selected model is used solely to approximate given data samples, but it also provides guidance for developing stochastic tessellation models and generating virtual microstructures. Model fitting is carried out by simulated annealing, applied in a consistent manner to twelve different tessellation models.

Keywords Model selection · Polycrystalline material · Tessellation · Akaike information criterion · Bayesian information criterion · Structural risk minimization

Mathematics Subject Classification (2000) 05B45 · 60D05 · 82D25

E-mail: ondrej.sedivy@uni-ulm.de, daniel.westhoff@uni-ulm.de, kopecek@fzu.cz, carl.krill@uni-ulm.de, volker.schmidt@uni-ulm.de
[1] Institute of Stochastics, Ulm University, Ulm, Germany
[2] Institute of Physics, Czech Academy of Sciences, Prague, Czech Republic
[3] Institute of Micro and Nanomaterials, Ulm University, Ulm, Germany

1 Introduction

In recent literature, a wide range of approaches can be found for representing polycrystalline microstructures using tessellation models. These models include tessellations with convex cells represented by the basic Voronoi model [13] or the Laguerre tessellation [12, 16], models with curved boundaries like the Johnson-Mehl tessellation [7] or spherical grain growth models [20], and, finally, even more complex tessellations based on ellipsoidal grain growth [3, 20] with further extension to generalized balanced power diagrams [2, 18, 19]. However, selection of the model that is most appropriate to a particular data set has often been performed *a priori,* or comparisons have been carried out among a limited number of models, in the absence of general rules for model selection. In [20], the microstructure of an IN100 nickel-based superalloy is approximated by the methods of Voronoi tessellation, spherical grain growth and ellipsoidal grain growth tessellation. In [3], tessellations generated by eleven distinct distance measures are applied to reconstructions of the microstructure of martensitic and bainitic steels. Generally, such investigations have found that models relying on a larger number of parameters provide a better approximation to the data. However, the differences in quality of approximation achieved by the various models are often quite small.

In this paper, the entire process of model selection is discussed in greater detail. In principle, the geometric model that is finally used to approximate the morphology of a given sample's polycrystalline microstructure should meet two requirements. On the one hand, a high accuracy of approximation is desired in both the metrical and topological senses. On the other hand, model complexity should be kept to the lowest possible level. The latter is a reasonable requirement not only from the standpoint of practicality, as the fitting of models with a larger number of parameters is computationally more demanding, but also for subsequent application of the modeled data. For instance, when a tessellation model is intended to serve as the basis for generating virtual microstructures, it is necessary to estimate the distributions of each model parameter, but this becomes increasingly difficult as the dimensionality of the parameter space grows. Finally, the preference for simpler models is consistent with the general scientific goal of expressing complex interrelationships as concisely as possible.

Our approach is based on concepts developed in regression theory, or, more generally, in the framework of statistical learning. The classical techniques use the number of model parameters as a measure for model complexity. This number is an essential component of a penalization term, which is applied in a multiplicative or additive sense to a certain measure for the goodness of fit of the model, usually the residual sum of squares. From these methods, we employ the most commonly used Akaike [1] and Bayesian [15] information criteria. An alternative way of assessing model complexity has been developed by Vapnik and Chervonenkis, whose approach is nowadays called Vapnik-Chervonenkis

theory [22,23]. Here, model complexity is described in a more sophisticated manner by analyzing all possible outcomes of the considered model. With certain necessary simplifying assumptions, this method can be applied to the selection of spatial tessellation models, as shown in the present paper.

For fitting tessellation models to empirical image data, we adopt and modify the approach described in [18]. This methodology, originally applied to generalized balanced power diagrams, is extended here to all of the models considered in the present paper. Furthermore, we introduce additional improvements: notably, we deal only with 'connectified' versions of the tessellations, in which all cells are simply connected sets. Although non-connected cells can easily arise in the aforementioned tessellation models, this fact has often been neglected in previous studies.

2 Tessellations

Tessellations can be viewed as a subdivision of space into non-overlapping sets, which are usually called *cells* or *grains*. Note that some definitions of tessellation appearing in the literature require that the cells be convex. However, we consider a more general class of tessellations that includes those with non-convex cells; for details, see, *e.g.,* [7,11,14].

Definition 1 Consider a countable collection of closed sets, $\mathscr{T} = \{C_i \subset \mathbb{R}^3,\, i = 1,2,\ldots\}$, such that

1. $\mathring{C}_i \cap \mathring{C}_j = \emptyset$ for all $i \neq j$, where $\mathring{C}_i$ is the interior of the set C_i,
2. $\bigcup_i C_i = \mathbb{R}^3$,
3. $\mathscr{T}$ is locally finite, i.e. $\#\{C_i \in \mathscr{T} : C_i \cap B \neq \emptyset\} < \infty$ for all bounded $B \subset \mathbb{R}^3$.

Then $\mathscr{T}$ is called a tessellation of the Euclidean space $\mathbb{R}^3$.

Furthermore, we focus on tessellations generated by a locally finite point pattern, $\mathscr{P} \subset \mathbb{R}^3$. We call the points of $\mathscr{P}$ *seeds* or *generators* of the tessellation and index them by natural numbers. The cell C_i corresponding to a given seed $\mathbf{x}_i \in \mathscr{P}$ is defined to consist of all points in $\mathbb{R}^3$ that are closer to $\mathbf{x}_i$ than to any other seed in $\mathscr{P}$ with respect to an appropriate distance measure d; *i.e.,*

$$C_i = \{\mathbf{x} \in \mathbb{R}^3 : d(\mathbf{x},\mathbf{x}_i) \leq d(\mathbf{x},\mathbf{x}_j) \text{ for all } \mathbf{x}_j \in \mathscr{P}\}. \tag{1}$$

We will further restrict our considerations to finite tessellations of a bounded domain $W \subset \mathbb{R}^3$, consisting of N cells indexed by $1,\ldots,N$.

2.1 General model

A wide class of tessellation models can be obtained if the distance measure is expressed by

$$d(\mathbf{x},\mathbf{x}_i) = [(\mathbf{x}-\mathbf{x}_i)^\top M_i(\mathbf{x}-\mathbf{x}_i)]^{k/2} - w_i, \tag{2}$$

where $M_i \in \mathscr{M}_3^+$, with $\mathscr{M}_3^+$ denoting the set of all positive definite 3×3 matrices, $w_i \in \mathbb{R}$ and $k \in \mathbb{N}$, see, e.g., the discussion of generalized Voronoi diagrams in [5]. Several special cases of distance measure are known from literature, which can be obtained by making particular choices for the parameters. The following list summarizes some typical examples, where I_3 denotes the unit 3×3 matrix:

- $k = 1, M_i = I_3, w_i = 0$: Voronoi tessellation [13] with Euclidean distance

  $$d_V(\mathbf{x},\mathbf{x}_i) = \|\mathbf{x}-\mathbf{x}_i\|;$$

- $k = 2, M_i = I_3$: Laguerre tessellation [4] with power distance

  $$d_P(\mathbf{x},\mathbf{x}_i) = \|\mathbf{x}-\mathbf{x}_i\|^2 - w_i, \quad w_i \in \mathbb{R};$$

- $k = 1, M_i = I_3$: additively weighted Voronoi tessellation [13] with distance

  $$d_A(\mathbf{x},\mathbf{x}_i) = \|\mathbf{x}-\mathbf{x}_i\| - w_i, \quad w_i \in \mathbb{R};$$

- $k = 2, w_i = 0$: ellipsoid-based tessellation [3] with distance

  $$d_E(\mathbf{x},\mathbf{x}_i) = (\mathbf{x}-\mathbf{x}_i)^\top M_i(\mathbf{x}-\mathbf{x}_i), \quad M_i \in \mathscr{M}_3^+;$$

- $k = 2$: generalized balanced power diagram [2] with distance

  $$d_G(\mathbf{x},\mathbf{x}_i) = (\mathbf{x}-\mathbf{x}_i)^\top M_i(\mathbf{x}-\mathbf{x}_i) - w_i, \quad M_i \in \mathscr{M}_3^+,\ w_i \in \mathbb{R}.$$

The positive definite matrix M_i in (2) can be thought of as an ellipsoid centered at the origin. The eigenvectors of M_i correspond to the principal axes of the ellipsoid, and the reciprocals of the eigenvalues denote the squared lengths of the semi-axes. Depending on the number of distinct eigenvalues, we can obtain different models. Let the parameters $k = 2, w_i = 0$ be fixed for now. Then, the case with three distinct eigenvalues leads to the ellipsoidal grain growth model. Besides the three coordinates of each seed, this model has six additional parameters per cell that are necessary to describe each matrix M_i. The model can be reparameterized by three semi-axis lengths and three angles (*e.g.* Euler angles) determining the orientation of the principal axes with respect to a reference coordinate system. If each matrix M_i has only two distinct eigenvalues, we obtain the special case of oblate or prolate ellipsoids. Here, the number of extra parameters per cell

is four, because only two angles and two lengths are required to describe the principal directions, and, therewith, the matrix M_i. If all eigenvalues of each M_i are equal, then the resulting tessellation is the spherical grain growth model with one extra parameter per cell. Finally, if in addition all matrices M_i have the same eigenvalue (*e.g.* equal to 1), we obtain the Voronoi model with no additional parameters. An analogous sequence of models arises for the versions with additive weights $w_i \in \mathbb{R}$.

Let ζ be the maximum number of distinct eigenvalues that are allowed for each matrix M_i. In what follows, we denote each model by $\mathscr{T}_{\alpha\beta}$, where the indices α, β represent the following quantities:

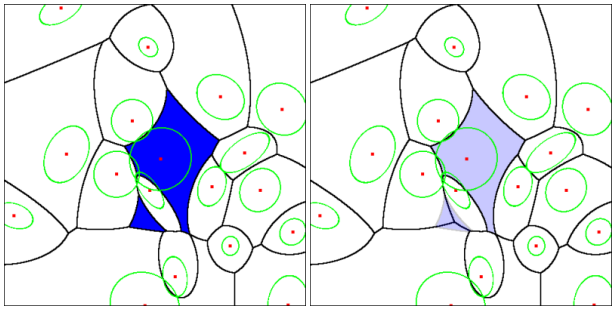
$$\alpha = \begin{cases} 0 & \text{if } w_i = 0 \text{ for all } i, \\ k & \text{otherwise;} \end{cases} \tag{3}$$

$$\beta = \begin{cases} 0 & \text{if all eigenvalues of all matrices } M_i \text{ are equal,} \\ \zeta & \text{otherwise.} \end{cases} \tag{4}$$

For instance, $\mathscr{T}_{00}$ denotes the Voronoi tessellation, $\mathscr{T}_{20}$ is the Laguerre tessellation, and $\mathscr{T}_{23}$ represents the generalized balanced power diagram. Note that in models without additive weights w_i ($\alpha = 0$), the inequality in (1) with distance given by (2) does not depend on the choice of k, thereby allowing for the aforementioned definition of the coefficient α. Furthermore, for the case of $\beta = 0$ we can assume that all matrices M_i are unit matrices, while for $\beta = 1$ they are distinct multiples of unit matrices. The number of additional parameters assigned to each point $\mathbf{x}_i$ in model $\mathscr{T}_{\alpha\beta}$ is $2\beta - \mathbf{1}\{\beta = 1\} + \mathbf{1}\{\alpha > 0\}$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Altogether, we will consider twelve models obtained from the ranges $\alpha \in \{0,1,2\}$ and $\beta \in \{0,1,2,3\}$.

2.2 Connectedness of cells

A negative feature of many non-convex tessellation models is that they can easily include cells that are not simply connected; see Fig. 1. This is an undesirable effect, as it is absent from real microstructures, in which separate regions identified as grains are labeled by different grain indices. There are several possible ways to overcome this problem. One option would be to restrict one's considerations to subclasses of tessellations in which all cells are simply connected. However, this would considerably limit the range of tessellations that could be employed. Another possibility would be to follow the common interpretation of tessellations as growth models, which supposes that each grain grows from a seed with a predefined velocity and direction until it meets another growing grain.



(a) (b)

Fig. 1 (a) Part of a 2D tessellation $\mathscr{T}_{11}$ with a disconnected cell colored in blue. (b) Simply connected modification $\mathscr{T}^*$ of the tessellation obtained after the first iteration of the algorithm described in Section 2.2.

Indeed, in such grain growth tessellation models, all cells are simply connected. However, growth models do not cover the entire class of tessellation models discussed in this paper.

As a slight generalization of grain growth models, we propose the following definition. It relies on the condition that each point $\mathbf{x}$ in cell C_i must be connectable (via a path fully contained within this cell) to some reference point $\mathbf{x}_i^*$, which is assumed to belong to the cell. We will denote this relation by $\mathbf{x} \leftrightarrow \mathbf{x}_i^*$. In principle, an arbitrary point of the cell can be chosen as the reference point. In grain growth models, these reference points can be represented by the seeds from which grains start to grow. However, for the general model (2) the seed does not necessarily lie within the cell it generates. In practice, we want the reference point to be located at a suitably defined central position of the largest connected component of the cell. We choose the reference point of cell C_i to be the center of the largest ball that is fully contained in this cell. Simple connectedness of the cells is then achieved via the following recursive definition. We start from the tessellation $\mathscr{T}^{(0)} = \{C_i^{(0)} = C_i \subset W, i = 1,\dots,N\}$ with cells C_i defined by (1). In each cell we find the region

$$\tilde{C}_i^{(0)} = \{\mathbf{x} \in C_i^{(0)} : \mathbf{x} \leftrightarrow \mathbf{x}_i^*\}$$

consisting of points that are connected to the reference point, and we denote the union of these sets by $\tilde{\mathscr{C}}^{(0)} = \bigcup_{i=1}^N \tilde{C}_i^{(0)}$.

In the second step the points of the complement $W \setminus \tilde{\mathscr{C}}^{(0)}$ are reassigned to the other cells:

$$C_i^{(1)} = \tilde{C}_i^{(0)} \cup \{\mathbf{x} \in W \setminus \tilde{\mathscr{C}}^{(0)} : d(\mathbf{x}, \mathbf{x}_i) \le d(\mathbf{x}, \mathbf{x}_j) \text{ for all } j \in P^{(1)}(\mathbf{x})\},$$

where $P^{(1)}(\mathbf{x})$ is the so-called potential set defined for each point $\mathbf{x} \in W \setminus \tilde{\mathscr{C}}^{(0)}$. The potential set collects indices of all cells of the tessellation $\mathscr{T}^{(0)}$ that are adjacent to the disconnected region to which $\mathbf{x}$ belongs. It means that $\mathbf{x}$ is connectable to each cell C_j, $j \in P^{(1)}(\mathbf{x})$, via a path fully contained within a connected component of the set $W \setminus \tilde{\mathscr{C}}^{(0)}$. Furthermore, the potential set $P^{(1)}(\mathbf{x})$ contains indices of all generators for which the corresponding cell in the tessellation $\mathscr{T}^{(0)}$ is an empty set.

Now, the union of connected regions is $\tilde{\mathscr{C}}^{(1)} = \bigcup_{i=1}^{N} \tilde{C}_i^{(1)}$, where $\tilde{C}_i^{(1)} = \{\mathbf{x} \in C_i^{(1)} : \mathbf{x} \leftrightarrow \mathbf{x}_i^*\}$.

In the m-th step we have

$$C_i^{(m)} = C_i^{(m-1)} \cup \{\mathbf{x} \in W \setminus \tilde{\mathscr{C}}^{(m-1)} : d(\mathbf{x},\mathbf{x}_i) \leq d(\mathbf{x},\mathbf{x}_j) \text{ for all } j \in P^{(m)}(\mathbf{x})\},$$
$$\tilde{C}_i^{(m)} = \{\mathbf{x} \in C_i^{(m)} : \mathbf{x} \leftrightarrow \mathbf{x}_i^*\},$$

where the potential set $P^{(m)}(\mathbf{x})$ is defined for each $\mathbf{x} \in W \setminus \tilde{\mathscr{C}}^{(m-1)}$ on the basis of the tessellation $\mathscr{T}^{(m-1)} = \{C_i^{(m-1)} \subset W, i = 1,\ldots,N\}$ analogously as described above.

Definition 2 Let $\{C_i^{(m)}\}$ be the sequence of sets defined above. The simply connected version of a tessellation $\mathscr{T}$ in a bounded window W is defined as

$$\mathscr{T}^* = \{C_i^{(m^*)} \subset W, i = 1,\ldots,N\}, \quad \text{where } m^* = \operatorname*{argmin}_{m \in \mathbb{N}_0} (\bigcup_{i=1}^{N} C_i^{(m)} = W).$$

Usually, the simply connected version is obtained after the first iteration, *i.e.* with $m^* = 1$. A formal justification that the algorithm terminates would be difficult to carry out; however, in all cases that we examined, the algorithm terminated after a small number of iterations. We do not address the issue of infinite tessellations, in which case it is possible to build tessellations for which the simply connected version remains undefined, owing to infinite recursion. Furthermore, it should be noted that disconnected components of the original tessellation $\mathscr{T}$ can actually be parts of the same cell separated by the face of the bounding box. However, this is impossible to detect without knowledge of the structure outside the bounding box. We neglect such boundary effects and consider each separate region contacting the bounding box to be an individual cell.

2.3 Additional notation

The empirical image data are observed on a cubic voxel grid $W' \subset W$ within a bounded window $W \subset \mathbb{R}^3$ and consist of a finite number of cells indexed by $1,\ldots,N$. Let $I(\mathbf{x})$ denote the grain index at a point $\mathbf{x} \in W'$ in the empirical image data and $I_{\mathscr{T}}(\mathbf{x})$ the index of

the cell that covers the point $\mathbf{x}$ in a tessellation $\mathscr{T}$ fitted to these data. The cells of a tessellation given by empirical data, conventionally called grains in polycrystalline materials, will be denoted by $G_1,\ldots,G_N$, and they can be expressed as $G_i = \{\mathbf{x} \in W' : I(\mathbf{x}) = i\}$. The cells of a fitted tessellation will be denoted by $C_1,\ldots,C_N$, and they can be expressed as $C_i = \{\mathbf{x} \in W : I_{\mathscr{T}}(\mathbf{x}) = i\}$.

Moreover, by B_I we denote the set of boundary voxels in the empirical image data set, *i.e.*,

$$B_I = \left\{\mathbf{x} \in W' : \|\mathbf{x}-\mathbf{y}\| \leq \sqrt{3} \text{ for some } \mathbf{y} \in W' \text{ with } I(\mathbf{x}) \neq I(\mathbf{y})\right\}. \tag{5}$$

The latter condition (with distance measured in units of voxel side length) expresses the requirement that the corresponding cubic voxels make contact at a vertex, edge or face. We denote by $B_{\mathscr{T}}$ an analogous set of boundary voxels in the tessellation model $\mathscr{T}$ projected onto the same voxel grid. Besides the set W', we consider a complementary set $\tilde{W}' \subset W$, the so-called dual voxel grid, consisting of the midpoints of the segments connecting nearest neighbors from W'. In a grid of cubic voxels, the points of W' can be identified with voxel midpoints, while the points of $\tilde{W}'$ can be identified with the midpoints of faces separating neighboring voxels. Grain boundaries are represented by the set

$$\tilde{B}_I = \left\{\mathbf{x} \in \tilde{W}' : \|\mathbf{x}-\mathbf{y}\| = \|\mathbf{x}-\mathbf{z}\| = \frac{1}{2} \text{ for some } \mathbf{y},\mathbf{z} \in W' \text{ with } I(\mathbf{y}) \neq I(\mathbf{z})\right\}. \tag{6}$$

For a tessellation model $\mathscr{T}$, the corresponding set $\tilde{B}_{\mathscr{T}}$ is an off-grid compact subset of W containing all points of the interfaces between the tessellation cells. Finally, we define the subsets $\tilde{B}_I(i,j) \subseteq \tilde{B}_I$ and $\tilde{B}_{\mathscr{T}}(i,j) \subseteq \tilde{B}_{\mathscr{T}}$ of boundary points separating the grains G_i, G_j in the empirical image data I or the cells C_i, C_j in the tessellation model $\mathscr{T}$, respectively.

Assessing the quality of the fitted model with respect to empirical data is possible via the discrepancy measure

$$\mathscr{D}_{I,\mathscr{T}} = \frac{\#\{\mathbf{x} \in W' : I(\mathbf{x}) \neq I_{\mathscr{T}}(\mathbf{x})\}}{\#\{\mathbf{x} \in W'\}}, \tag{7}$$

which provides information regarding the fraction of voxels that are incorrectly assigned. This measure is minimized by the simulated annealing algorithm explained in Section 4. We are also interested in its restriction to the subset of grain boundary voxels:

$$\mathscr{D}^B_{I,\mathscr{T}} = \frac{\#\{\mathbf{x} \in B_I : I(\mathbf{x}) \neq I_{\mathscr{T}}(\mathbf{x})\}}{\#\{\mathbf{x} \in B_I\}}. \tag{8}$$

Consider the relation $G_i \sim G_j$ that grains G_i and G_j are neighbors—*i.e.,* there exists a pair of points $\mathbf{x},\mathbf{y} \in B_I, \|\mathbf{x}-\mathbf{y}\| = 1 : I(\mathbf{x}) = i, I(\mathbf{y}) = j$. The same symbol is used for the

neighborhood of cells in the fitted tessellation projected onto the same voxel grid. Then, for the ith grain the quantity

$$\mu_{I,\mathscr{T}}(i) = \#\{j \in 1,\dots,N : (G_i \sim G_j \text{ and } C_i \not\sim C_j) \text{ or } (G_i \not\sim G_j \text{ and } C_i \sim C_j)\} \quad (9)$$

describes the neighborhood fit by counting the number of disagreements in the list of neighbors between the observed data and the model. By

$$\bar{\mu}_{I,\mathscr{T}} = \frac{1}{N}\sum_{i=1}^{N} \mu_{I,\mathscr{T}}(i) \quad (10)$$

we denote the mean neighborhood fit.

3 Model selection

For the selection of an optimal model, we gain inspiration from the methods used in statistical learning, particularly in regression. Here the response y is assumed to depend on a vector of observed (explanatory) variables $\mathbf{x}$. The dependency is approximated by a class of functions $\{f(\mathbf{x},\theta), \theta \in \Theta\}$, where Θ is an abstract parameter space. The goal is to select a function that best approximates the response.

We will adapt these ideas to our situation, in which a grain structure is approximated by a tessellation model. To this end, consider the response y to be the set of grain boundaries observed in the microstructure. Then the class of approximating functions can be identified with the tessellation models defined in Section 2, where the interfaces between cells are considered to be the model output. Each tessellation is solely determined by its parameters; therefore, in this setting we are missing the usual dependency on explanatory variables. However, this does not limit the range of approximations of the response that are accessible using a suitably defined discrepancy measure. Note, in addition, that we deal only with a single observation. However, statistical inference is still possible, taking information at distinct locations in space into account. This is a common situation in spatial statistics, where the lack of observations is compensated by a sufficient amount of information in the spatial domain. Consistency of statistical estimators is often studied with respect to increasing size of the observation window [9].

3.1 Criteria based on number of parameters

Let us now assume we are given a sample $(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_n,y_n)$ that, in the terminology of statistical learning, plays the role of training data. A common criterion for assessing

model accuracy is the residual sum of squares (RSS) defined by

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \theta))^2. \tag{11}$$

Since our response is given by the voxelized version of grain boundaries, B_I, the residual sum of squares for a tessellation $\mathscr{T}$ can be defined by

$$\mathrm{RSS}(\mathscr{T}) = \sum_{\mathbf{x} \in \tilde{B}_I} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2, \tag{12}$$

where $\tilde{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{y} \in \tilde{B}_{\mathscr{T}}(i,j)} \|\mathbf{x} - \mathbf{y}\|$ is a (not necessarily unique) closest point to $\mathbf{x} \in \tilde{B}_I(i,j)$ belonging to the exact boundaries between cells C_i and C_j in the tessellation model. It remains to define the contribution of points belonging to those boundaries $\tilde{B}_I(i,j)$ for which the corresponding boundary in the tessellation $\mathscr{T}$ does not exist: *i.e.*, $\tilde{B}_{\mathscr{T}}(i,j) = \emptyset$. The penalization of such points could be set, for instance, to the maximum value of $\|\mathbf{x} - \tilde{\mathbf{x}}\|^2$ taken over all points for which it is defined. We employ a more robust version using the 99% quantile instead of the maximum. For ease of notation, we omit the argument θ or $\mathscr{T}$ in the definition of RSS and in related quantities defined in the remaining part of this section.

A natural first choice for the optimal model would be the candidate with minimum RSS. However, we require the model not only to be sufficiently accurate but also simple. It is known from statistical learning theory that an increase in model complexity often leads to the overfitting of training data, which, in turn, can worsen the predictive power of the model. Thus, various ways of penalizing the RSS by a term related to model complexity have been suggested. Many of these approaches are based on the number of parameters being a measure for model complexity. Among these we mention the two most common approaches in the following definition [1,15]:

Definition 3 The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are defined by

$$\mathrm{AIC} = n \ln(\mathrm{RSS}) + 2p, \tag{13}$$

$$\mathrm{BIC} = n \ln(\mathrm{RSS}/n) + p \ln(n), \tag{14}$$

where n is the sample size and p is the number of free parameters of the model.

Good asymptotic consistency of both criteria of Definition 3 is obtained if the number of observations n is much larger than the number of parameters p. In our case, the sample size n is equal to the number of boundary points involved in the computation of the RSS defined in (12). In order to lower the computation time, only a fraction of the points in $\tilde{B}_I$

can be used. Note that smaller sample sizes lead to a stronger effect of penalization. In comparison to that of AIC, the penalization term in BIC increases more rapidly with p, resulting in a stronger preference for simpler models.

The aforementioned criteria have frequently been criticized on the basis that the number of parameters need not properly reflect a given model's complexity. Often, the parameters of a model are found to have differing levels of significance, but these are not considered by the criteria of Definition 3. In the next section, we will introduce a method that aims to assess model complexity in a more objective manner, without requiring explicit determination of the significance of individual model parameters, but taking it into account implicitly.

3.2 Structural risk minimization

An alternative approach toward complexity control—called structural risk minimization (SRM)—was introduced in Vapnik-Chervonenkis (VC) theory [23]. The set $\mathscr{F}$ of approximating functions $f(\mathbf{x},\theta)$ is assumed to consist of nested subsets $\mathscr{F}_k = \{f(\mathbf{x},\theta), \theta \in \Theta_k\}$ with $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \ldots \subset \mathscr{F}$. The complexity of each model $\mathscr{F}_k$ is described by a number $h_k \in \mathbb{R}^+$ called the VC-dimension. In accord with the hierarchy of models $\mathscr{F}_1, \mathscr{F}_2, \ldots$, the VC-dimensions are ordered such that $h_1 \leq h_2 \leq \ldots$. Furthermore, model selection means choosing an optimal element of the structure of nested subsets using so-called VC generalization bounds. These bounds provide an approximate upper limit for the true risk, which is defined as the mean squared error of the approximation. This approximate upper limit for the true risk is called the guaranteed risk and is denoted by R_g. For regression problems, the guaranteed risk is known to be

$$R_g = \frac{\mathrm{RSS}}{n}\left(1 - c\sqrt{\varepsilon_k}\right)_+^{-1}, \tag{15}$$

where n is the number of observations, $(\cdot)_+$ denotes the positive part, *i.e.,* $x_+ = \max(x,0)$, c is a constant, and ε_k is given by

$$\varepsilon_k = \frac{a_1}{n}\left(h_k\left(\ln\frac{a_2 n}{h_k} + 1\right) - \ln\frac{\eta}{4}\right) \tag{16}$$

with two additional constants a_1, a_2, see [23]. The guaranteed risk is an approximate upper bound for the true risk with confidence level $1-\eta$ (see [22]). The constants c, a_1, a_2 depend on the joint distribution of the response and the explanatory variables. However, empirical results suggest that the choice $c = a_1 = a_2 = 1$ provides a good approximation of the guaranteed risk [6].

Furthermore, SRM has often been used in binary classification, where the following expression for the guaranteed risk,

$$R_g = \min\left(1,\ \frac{\mathrm{RSS}}{n} + \frac{\varepsilon_k}{2}\left(1+\sqrt{1+\frac{4\mathrm{RSS}}{n\,\varepsilon_k}}\right)\right), \tag{17}$$

has been obtained [6], once again with ε_k as defined in (16). Here, the constants a_1 and a_2 must be in the range $0 < a_1 \leq 4$, $0 < a_2 \leq 2$, but empirical guidance for selecting more precise values for these constants is lacking. As in the case of regression, the values $a_1 = a_2 = 1$ are frequently used. The RSS is thus a (non-standardized) estimation of the risk based on training data. Note that for binary classification, the RSS equals the number of misclassified cases.

The strategy of the SRM method is as follows. First, the subset $\mathscr{F}_k \subseteq \mathscr{F}$ is identified for which the guaranteed risk is minimal. $\mathscr{F}_k$ represents the model of optimal complexity. Then, the function in $\mathscr{F}_k$ is found that minimizes the empirical risk, evaluated for the training data. This corresponds to the task of parameter estimation.

For the application of VC theory, the models need to be structured as described above. However, for our set of tessellation models, we have only a partial ordering, *i.e.*, $\mathscr{T}_{\alpha\beta_1} \subset \mathscr{T}_{\alpha\beta_2}$ if $\beta_1 < \beta_2$, and $\mathscr{T}_{0\beta} \subset \mathscr{T}_{\alpha\beta}$ for $\alpha > 0$. This partial hierarchy can be schematized as follows:

$$\begin{array}{ccccccc}
\mathscr{T}_{10} & \subset & \mathscr{T}_{11} & \subset & \mathscr{T}_{12} & \subset & \mathscr{T}_{13}, \\
\cup & & \cup & & \cup & & \cup \\
\mathscr{T}_{00} & \subset & \mathscr{T}_{01} & \subset & \mathscr{T}_{02} & \subset & \mathscr{T}_{03}, \\
\cap & & \cap & & \cap & & \cap \\
\mathscr{T}_{20} & \subset & \mathscr{T}_{21} & \subset & \mathscr{T}_{22} & \subset & \mathscr{T}_{23}.
\end{array}$$

Due to incomplete ordering, it is not always possible to pick out an optimal candidate. For instance, the guaranteed risk of models from the first line is not directly comparable with the guaranteed risk of models from the third line, since they do not share a model-submodel relationship. As a complementary rule, a topological fitting characteristic (*e.g.* $\bar{\mu}_{I,\mathscr{T}}$) can be used to decide between models in these cases as well as in situations in which the minimum guaranteed risk is not significantly unique.

The concept of the VC-dimension as an integer-valued measure of model complexity was introduced in [23]. The purpose of this quantity is to quantify in the form of a single number how powerful a set of approximating functions is for approximating the given response. We consider only the class of indicator functions for which the VC-dimension can be defined directly by means of so-called *shattering* [23]. The VC-dimension is the maximum number h of vectors from $\mathbb{R}^d$ that can be separated into two output classes,

zero and one, in all 2^h possible ways using indicator functions. A more precise definition follows.

Definition 4 Consider an abstract parameter space Θ. The VC-dimension of a set of indicator functions $\{q(\cdot,\theta) : \mathbb{R}^d \mapsto \{0,1\}, \theta \in \Theta\}$ is defined as the maximum number h such that there exists a sequence of distinct vectors $\mathbf{x}_1,\ldots,\mathbf{x}_h \in \mathbb{R}^d$ that satisfies

$$\#\{(q(\mathbf{x}_1,\theta),\ldots,q(\mathbf{x}_h,\theta)), \theta \in \Theta\} = 2^h.$$

If such a sequence of vectors exists for any $h \in \mathbb{N}$, then the VC-dimension of this set of functions is equal to infinity.

Alternatively, we say that the vectors can be shattered by the indicator functions. As an example, consider the following set of linear indicator functions,

$$q(\mathbf{x},\mathbf{a}) = \vartheta\left(\sum_{i=1}^{d} a_i x_i + a_0\right),$$

of a d-dimensional vector $\mathbf{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$, where $\vartheta(x)$ is the step-function

$$\vartheta(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0, \end{cases}$$

and $\mathbf{a} = (a_0,\ldots,a_d)$ is a vector of real-valued coefficients. The VC-dimension is then equal to $h = d+1$, since at most $d+1$ vectors can be shattered by functions of this set; see Fig. 2.

We will apply the concept of VC-dimension to tessellations in the following manner. Consider a set of test points $T \subset \tilde{B}_I$ on the empirical grain boundaries. The sampling scheme for this test set will be discussed later in Section 6.2. According to the tessellation setting described in Section 3.2, we will define a binary response for each observation $\mathbf{t} \in T$ that takes the value 0 if $\mathbf{t}$ is a boundary point between the same cells in the fitted tessellation as in the empirical data; otherwise, the response is given the value 1, which occurs if $\mathbf{t}$ is not a boundary point of the fitted tessellation or if $\mathbf{t}$ is a boundary point between different cells in the fitted tessellation than in the empirical data. The VC-dimension h of the model $\mathscr{T}_{\alpha\beta}$ (with fixed α,β) is the maximum number of points of T that can be shattered by a tessellation $\mathscr{T}_{\alpha\beta}$ (with appropriate choice of tessellation parameters) with respect to the aforementioned binary response. Here, the tessellation is assumed to be projected onto the same voxel grid as the observed data.

The idea of estimating the VC-dimension is to generate tessellations repeatedly with varying parameters and to examine points from T. The key finding is whether the tessellation can separate the points of T into two classes with indicator output 0 or 1, respectively,

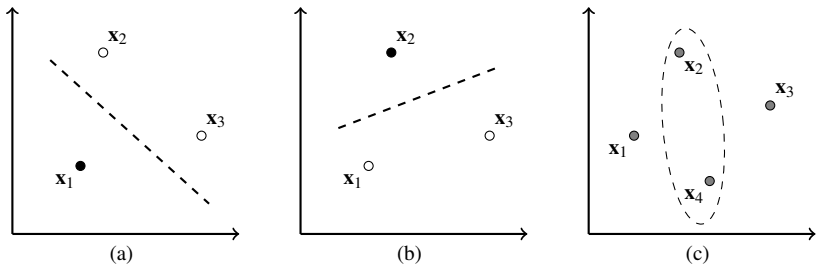


Fig. 2 Linear indicator functions in $\mathbb{R}^2$: (a,b) each observation $\mathbf{x}_i \in \mathbb{R}^2$ is assigned a value 0 (open circle) or 1 (filled circle), depending on the halfspace—generated by the dashed line—to which $\mathbf{x}_i$ belongs. (c) It is not possible to shatter four vectors in $\mathbb{R}^2$. For example, the vectors $\mathbf{x}_1, \mathbf{x}_3$ cannot be separated by a line from the vectors $\mathbf{x}_2, \mathbf{x}_4$. Therefore, the VC-dimension of this example is $h = 3$.

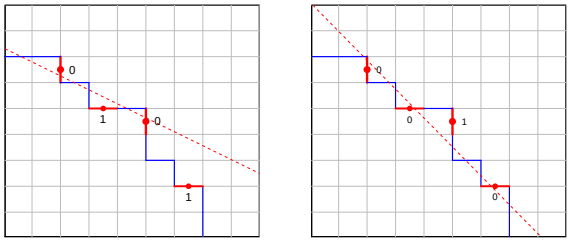
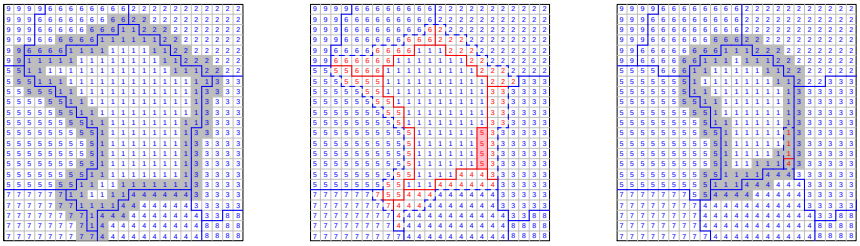


Fig. 3 2D scheme for shattering test points on grain boundaries (blue) by straight lines (dashed), which correspond to interfaces between cells in the Voronoi or Laguerre tessellation. Each test point (filled red circle) is located on a horizontal or vertical line segment (red). The test point is assigned an indicator value of 0 if the segment makes contact with the red dashed line and 1 otherwise. Shattering means obtaining all possible sequences of zeros and ones for the test points.

in all possible ways; see Fig. 3. Thus, the VC-dimension expresses the potential of a given tessellation model to describe the grain boundaries of a given empirical image data set, subject to small errors caused by uncertainty in the recognition of grain boundaries. Evaluation of the VC-dimension is a purely data-driven procedure. Detailed discussion and results of the estimation procedure follow in Section 6.2.

Note that our estimation of the VC-dimension depends strongly on the sampling scheme of the test set T as well as on the cardinality of this set. In Section 6.2, a sampling



(a) (b) (c)

Fig. 4 Schematic illustration of part of the tessellation during a single step of the fitting algorithm. The figure depicts a planar section through the voxelized tessellation, in which each voxel is labeled by the index of the cell to which it belongs. Each cell approximates one grain from the original data. (The original grains are omitted here for ease of graphical representation.) (a) Fitting starts by changing the parameters of the generator of a randomly chosen cell—here, the cell with index 1. Checking for reassignment begins with voxels near the boundaries of cell 1 (shaded grey). (b) The cell indices of reassigned voxels are denoted in red, as are the new cell boundaries that result from voxel reassignment. A small cluster of voxels with cell index 5 (shaded red) is disconnected from the main part of cell 5; the disconnected cluster must be reassigned. (c) Changed configuration following reassignment of the disconnected region, with the new set of boundary voxels colored gray. The disconnected part of the cell with index 5 was reassigned to cells 1 and 4. For more details regarding the simulated annealing algorithm, the reader is referred to [18].

scheme will be suggested that avoids the proximity of test points and the resulting strong correlations, which would negatively influence the estimation. Moreover, the size of T is set to a common value for all models. Under these assumptions, the VC-dimension leads to an objective comparison of different tessellation models.

4 Model fitting

For fitting tessellation models to empirical image data, we use a stochastic optimization algorithm based on simulated annealing. This method was utilized previously with the generalized balanced power diagram model in [18], and its flexibility allows its application to all of the models considered in the present paper. Here, we propose further improvements to the algorithm. The main modification is adaptation of the algorithm to simply connected versions of the tessellations, as described in Section 2.2.

The fitting procedure starts from an initial configuration and proceeds by iterative changes of selected parameters, such that the quality of fit is improved sequentially. Suc-

cessful application of the method relies on several requirements. First, the initial tessellation should provide a sufficiently good fit to the data. We use initial states based on the heuristic approaches described in [2,3,18]. Here, initial parameters are determined from the lengths or directions of principal axes of best-fitting balls, spheroids or ellipsoids that are placed at the centers of mass of the observed grains. These objects have the same volume as the grains, and their orientation is optimized such that they describe the real positions of the grains. Additional details are given in [2,18]. In particular, the initial values of parameters are set as follows. The initial seeds are assigned to the centers of mass of the grains. For the models $\mathscr{T}_{\alpha\beta}$, $\alpha > 0$, the initial weights w_i in (2) are computed as volume-equivalent radii of the grains if $\beta = 0$, and they are set to zero if $\beta > 0$. The initial matrices M_i are computed from best-fitting balls in models with $\beta = 1$ (which are, again, balls with volume-equivalent radii), best-fitting spheroids for $\beta = 2$, and best-fitting ellipsoids for $\beta = 3$. An alternative option for the initial configuration in models $\mathscr{T}_{\alpha\beta}$, $\alpha > 0$, $\beta > 0$, is letting the weights w_i be the volume-equivalent radii and the matrices M_i be unit matrices. The latter option is used if it provides a better initial fit than the previously mentioned configuration.

Our main improvement of the algorithm presented in [18] is adaptation of the latter to simply connected versions of tessellations with non-convex cells. The connectedness of cells is checked, and disconnected regions are removed according to the recursive definition of a simply connected version of a tessellation given in Section 2.2 (Figs. 1 and 4). At first, in each cell we identify a reference point that is assumed to belong to the fitted grain. This point is found as the center of the largest ball that is fully contained within the corresponding cell. Next, we assign a connectivity number 0 to the voxel at each reference point, 1 to all neighbors of these voxels, 2 to neighbors of these neighbors, and so on. This algorithm assigns a connectivity number to all points belonging to $\mathscr{C}^{(0)}$, and any disconnected parts remain unassigned. Furthermore, following Definition 2 we reassign the voxels of $W \setminus \cup_i C_i^{(0)}$ to the other cells for which the voxel-seed distance has the smallest value. This distance is computed using the measure appropriate to each tessellation model. For the reassigned voxels, we apply the same algorithm again and assign connectivity numbers to all components connectable to the reference point. This procedure is repeated until all voxels have been assigned a connectivity number.

Each time a change is applied to the tessellation parameters, the connectedness needs to be checked, and connectivity numbers need to be recomputed. The detection of disconnected regions can be performed quickly using the connectivity number. Each voxel that has been reassigned to another grain in the last step is added to a list of candidates for disconnected voxels. Furthermore, all their neighbors are added to this list, if they do not have a neighbor with the same cell index and lower connectivity number. This procedure is continued iteratively, resulting in a list of candidates for disconnected vox-

els. The connectivity number of all these voxels is deleted. Those voxels in the list that have a neighbor with the same cell index and assigned connectivity number are obviously connectable to the corresponding reference point, so they can be assigned a connectivity number and removed from the list. This procedure is again performed iteratively, until no further candidate can be removed from the list. The remaining voxels are disconnected and have to be reassigned as described in Section 2.2. The procedure described above seems complicated; however, it is computationally much less demanding than recomputing all connectivity numbers in each step.

5 Microstructure data

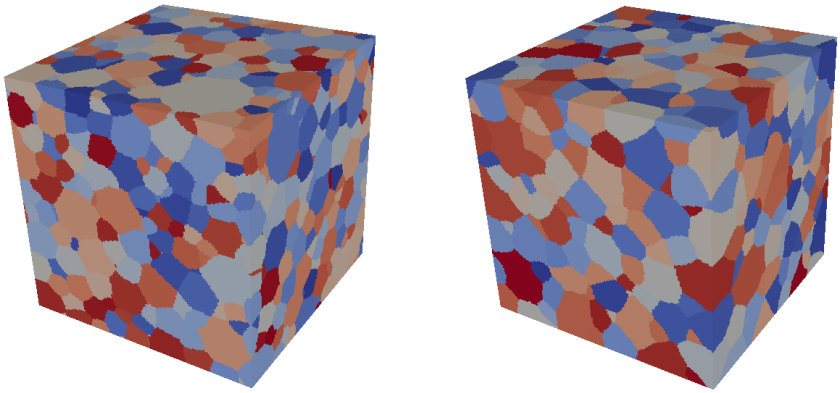
In this section we describe four different microstructure data sets to which our methods have been applied. At first, we compare results obtained from two simulated data sets, which were drawn from different models. For these data we can determine whether our methods are able to identify the correct model. Secondly, we consider two data samples obtained by experimental techniques for characterizing polycrystalline materials in 3D. The length scale of microstructural features differs considerably in the two sample data sets.

5.1 Simulated data

We have simulated two data sets, denoted as S_1 and S_2, in a cube of size $150 \times 150 \times 150$ voxels. These tessellations were generated from two different simply connected models defined in (2): the Laguerre tessellation ($\mathscr{T}^*_{20}$) and the ellipsoidal grain growth model ($\mathscr{T}^*_{03}$). While in the former case, there is only one extra parameter assigned to each seed and all cells are convex sets, in the latter case we have six extra parameters for each seed and interfaces between cells are parts of quadric surfaces, allowing for a wide range of surface curvatures. Both simulated data sets are visualized in Fig. 5.

For the simulation of data set S_1, an algorithm for random sphere packing was used. The packing of spheres is achieved by a collective rearrangement algorithm applied to an initial random system of (overlapping) spheres. During this rearrangement, individual spheres were repeatedly displaced with the aim of reducing and finally eliminating sphere overlap. The algorithm is described in detail in [17]. The centers of spheres together with their radii are considered as marked points, which generate a Laguerre tessellation. In our example, the resulting tessellation (Fig. 5(a)) contains 1128 cells.

The simulated data set S_2 was generated on the basis of a synthetic microstructure builder, which is part of the DREAM.3D software package [10]. This builder enables



(a) (b)

Fig. 5 Simulated data sets (a) S_1 and (b) S_2. Grain colors assigned at random.

microstructures to be simulated with predefined statistical properties, using an algorithm for packing a cubic spatial domain with ellipsoids of given sizes, aspect ratios and orientations. We modified its output such that the resulting model corresponds exactly to the simply connected version of the ellipsoidal grain growth model, as described in previous sections. Details of this modification can be found in [18]. The resulting tessellation (Fig. 5(b)) contains 898 cells.

5.2 Experimental data

The first experimental data sample—denoted E_1—was taken from an aluminum alloy with nominal composition Al-1 wt% Mg. A cylindrical specimen of the material was annealed at 400°C for 1 h, which resulted in a microstructure with mean grain volume of 0.006 mm^3. A 3D image of the microstructure was acquired by 3D X-ray diffraction (3DXRD) microscopy performed at the synchrotron radiation facility SPring-8 in Japan. In this nondestructive characterization technique, a specimen is irradiated with a monochromatic X-ray beam while being rotated up to 360° about an axis perpendicular to the beam, such that each grain in the illuminated volume fulfills the conditions for Bragg reflection numerous times. From the position and shape of the resulting diffraction signals, it is possible to compute the lattice orientation and spatial extent of the diffracting grains. The overall data sample covers a cylinder of approximate radius 0.8 mm and

height 2.7 mm, with a voxel size of $(5\,\mu\text{m})^3$. The number of grains contained in this cylinder is 753.

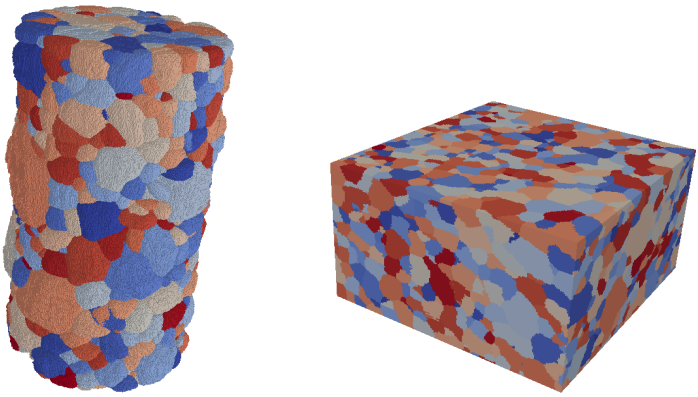
The second experimental data sample—denoted E_2—had nominal composition Al-3 wt% Mg-0.2 wt% Sc. It was processed by 8 passes of equal-channel angular pressing (ECAP) with route B_C at room temperature; see *e.g.* [21] for a full discussion of this procedure. The specimens for ECAP had a cross section of 10 mm × 10 mm and an approximate length of 55 mm. Following ECAP processing the specimen was annealed at 400°C for 1 h, which induced the formation of a fine-grained microstructure with well-defined crystallites. Imaging was carried out using an FEI Quanta 3D FEG field-emission scanning electron microscope (SEM) equipped with a high-speed EDAX/TSL EBSD camera and focused ion beam (FIB). The step sizes of EBSD mapping and FIB slicing were both $0.1\,\mu$m, resulting in a voxel volume of $10^{-3}\,\mu\text{m}^3$. Three-dimensional characterization of the microstructure was accomplished by combining EBSD mapping with sequential micro-milling of the top surface using the FIB. The resulting 3D image consists of a stack of equidistantly spaced planar sections. The number of grains identified in the 3D volume is 3052.

For additional details regarding the preparation of samples E_1 and E_2 and the respective methods of data acquisition, the reader is referred to the following publications. The microstructure of Sample E_1 was analyzed in [8] in order to assess the evolution of crystallographic orientations during particle coarsening. In [18], the microstructure of E_2 was represented by a generalized balanced power diagram, employing the fitting procedure described in Section 4. The morphology of grain boundaries in each sample was described in [19].

6 Results

6.1 Fitting models to data

Simply connected versions of the tessellation models $\mathscr{T}^*_{\alpha\beta}$, where $\alpha \in \{0,1,2\}$, $\beta \in \{0,1,2,3\}$, were fitted to the empirical image data sets described in Section 5 using the simulated annealing methodology of Section 4. We applied 5 million iterations of the algorithm to each model fitted to samples S_1, S_2, E_2 and 2 million iterations to each model fitted to sample E_1, which contains fewer grains. For each fitted model, we first evaluate the summary statistics presented in Table 1. These include the percentage of correctly assigned voxels, $\delta_{I,\mathscr{T}} = (1 - \mathscr{D}_{I,\mathscr{T}})$, the percentage of correctly assigned voxels at grain boundaries, $\delta^B_{I,\mathscr{T}} = (1 - \mathscr{D}^B_{I,\mathscr{T}})$, and characteristics related to the neighborhood of each grain.



(a) (b)

Fig. 6 Tomographic 3D images obtained from polycrystalline specimens of (a) Al-1 wt% Mg (sample E_1) and (b) Al-3 wt% Mg-0.2 wt% Sc (sample E_2). Grain colors assigned at random.

Obviously, the goodness of fit improves with increasing number of parameters involved in the tessellation model. For the generalized balanced power diagram ($\mathcal{T}_{23}^*$), the fraction of correctly assigned voxels achieved for the experimental data via the simulated annealing fitting procedure is about 95% for sample E_1 and about 92% for sample E_2. The discrepancy between these two values can be attributed to two factors. On the one hand, the two data sets come from different materials, and different procedures were followed during preprocessing of the sample data. On the other hand, the average number of voxels per grain differs between the data sets, which affects the level of precision regarding the description of grain boundaries. This also accounts for discrepancies in the goodness of fit with respect to topological characteristics, with a significantly better neighborhood fit being achieved for sample E_1 than for E_2.

In Fig. 7, the distribution of distances between observed and modeled grain boundaries in the experimental data is plotted, on the basis of which the residual sum of squares (RSS) considered in (12) is defined. In general, the empirical risk decreases with increasing number of parameters involved in the model. However, for some pairs of models the discrepancies between grain boundary distance distributions are quite small. In these cases, adding further parameters does not have a significant effect on the overall quality of the fit.

Table 1 Statistics for simply connected tessellations $\mathscr{T}^*_{\alpha\beta}$ fitted to simulated data samples S_1, S_2 and experimental data samples E_1, E_2 by the simulated annealing methodology of Section 4. Percentage of correctly assigned voxels, $\delta_{I,\mathscr{T}} = (1-\mathscr{D}_{I,\mathscr{T}})\cdot 100\%$; percentage of correctly assigned voxels at grain boundaries, $\delta^B_{I,\mathscr{T}} = (1-\mathscr{D}^B_{I,\mathscr{T}})\cdot 100\%$; percentage of grains with all neighbors correct, $\mu^0_{I,\mathscr{T}} = \#\{i : \mu_{I,\mathscr{T}}(i) = 0\}/N \cdot 100\%$; and mean number of incorrect neighbors, $\bar{\mu}_{I,\mathscr{T}}$. The optimal value in each row (*i.e.* the maximum for $\delta_{I,\mathscr{T}}$, $\delta^B_{I,\mathscr{T}}$, $\mu^0_{I,\mathscr{T}}$, and the minimum for $\bar{\mu}_{I,\mathscr{T}}$) is printed in bold type.

		$\mathscr{T}^*_{00}$	$\mathscr{T}^*_{01}$	$\mathscr{T}^*_{02}$	$\mathscr{T}^*_{03}$	$\mathscr{T}^*_{10}$	$\mathscr{T}^*_{11}$	$\mathscr{T}^*_{12}$	$\mathscr{T}^*_{13}$	$\mathscr{T}^*_{20}$	$\mathscr{T}^*_{21}$	$\mathscr{T}^*_{22}$	$\mathscr{T}^*_{23}$
S_1	$\delta_{I,\mathscr{T}}$	77.9	96.6	96.8	97.0	97.8	97.9	97.8	97.0	**98.8**	98.2	98.3	97.2
	$\delta^B_{I,\mathscr{T}}$	56.6	85.7	86.5	87.3	90.9	91.0	90.6	87.2	**95.0**	92.7	92.9	88.0
	$\mu^0_{I,\mathscr{T}}$	21.9	53.3	55.6	57.7	71.6	71.2	76.8	59.2	**88.4**	79.4	81.9	64.6
	$\bar{\mu}_{I,\mathscr{T}}$	1.66	0.70	0.62	0.57	0.36	0.36	0.28	0.55	**0.13**	0.25	0.20	0.48
S_2	$\delta_{I,\mathscr{T}}$	84.2	96.4	97.8	**98.5**	95.7	96.3	97.8	98.4	94.5	96.4	97.7	98.4
	$\delta^B_{I,\mathscr{T}}$	61.5	85.4	91.1	**93.8**	83.0	85.1	90.8	93.4	79.1	85.2	90.6	93.2
	$\mu^0_{I,\mathscr{T}}$	19.1	62.2	74.8	**82.7**	54.1	60.9	75.9	79.8	38.3	59.5	71.9	77.1
	$\bar{\mu}_{I,\mathscr{T}}$	1.74	0.47	0.28	**0.20**	0.64	0.49	0.29	0.24	0.98	0.53	0.34	0.28
E_1	$\delta_{I,\mathscr{T}}$	57.9	90.7	93.9	95.0	89.5	90.5	94.0	95.0	86.2	92.0	94.4	**95.2**
	$\delta^B_{I,\mathscr{T}}$	39.7	60.9	66.5	69.3	60.0	60.6	66.8	69.4	57.5	62.8	67.5	**69.7**
	$\mu^0_{I,\mathscr{T}}$	1.1	9.0	23.4	36.1	22.6	6.8	24.1	38.0	15.0	17.3	29.7	**39.5**
	$\bar{\mu}_{I,\mathscr{T}}$	5.40	2.55	1.66	1.10	1.73	2.74	1.50	1.00	2.14	2.00	1.26	**0.99**
E_2	$\delta_{I,\mathscr{T}}$	57.5	83.8	90.0	91.8	82.7	83.6	90.1	91.8	79.4	84.6	90.2	**92.0**
	$\delta^B_{I,\mathscr{T}}$	42.1	61.4	70.6	74.4	61.3	61.1	70.5	74.3	58.2	62.0	70.7	**74.5**
	$\mu^0_{I,\mathscr{T}}$	0.3	6.9	23.4	**30.8**	9.7	6.8	22.4	29.7	6.9	7.8	23.8	30.7
	$\bar{\mu}_{I,\mathscr{T}}$	6.95	3.28	1.66	1.32	3.13	3.32	1.72	1.34	3.91	3.14	1.66	**1.30**

6.2 Estimation of VC-dimension

The ideas underlying the estimation of VC-dimension have already been sketched out in Section 3.2. Recall that the VC-dimension is intended to characterize the capacity of a tessellation model to describe the grain boundaries appearing in observed microstructures. This capacity is explored through a set of testing points located at grain boundaries, for which shattering—as explained in Section 3.2—is examined. Additional details concerning our estimation procedure, which is based on the simulation of tessellations with varying parameters, now follow.

The first task is to generate a random test set $T \subset \tilde{B}_I$ of points belonging to the grain boundaries of the empirical data sample. In order to avoid an undesirable proximity of test points, we define a hard-core radius $r > 0$ as the lower limit for the pairwise distance between points in T. For the simulation, we use a simple acceptance-rejection algorithm,

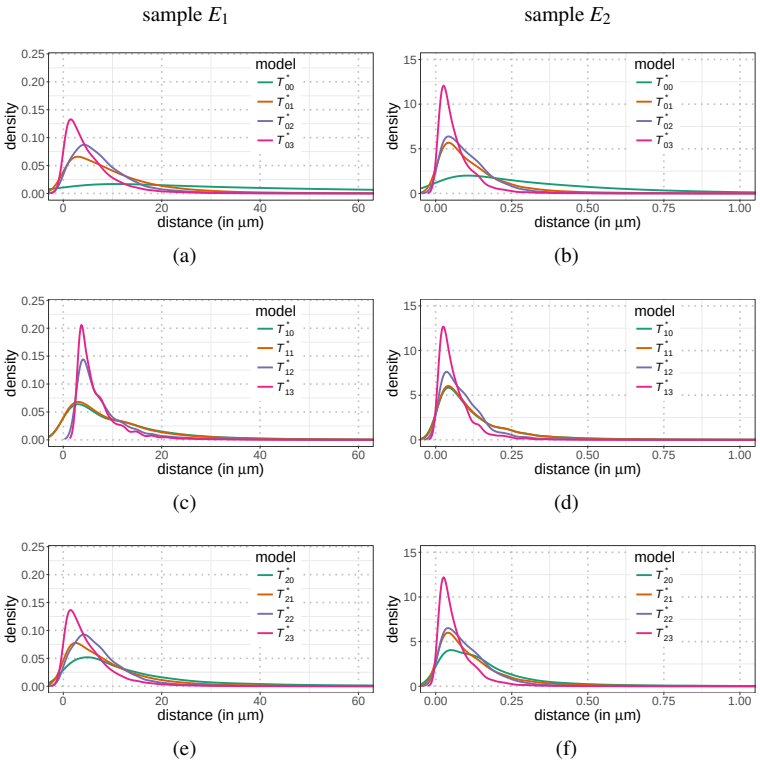


Fig. 7 Distribution of the distance $||\mathbf{x} - \tilde{\mathbf{x}}||$ separating grain boundary points $\mathbf{x} \in B_I$ from the nearest points $\tilde{\mathbf{x}} \in B_{\mathscr{T}}$ of the corresponding grain boundaries in tessellation models: (first row) $\mathscr{T}^*_{00}$ through $\mathscr{T}^*_{03}$; (second row) $\mathscr{T}^*_{10}$ through $\mathscr{T}^*_{13}$; and (third row) $\mathscr{T}^*_{20}$ through $\mathscr{T}^*_{23}$, evaluated for (left column) sample E_1 and (right column) sample E_2.

in which points are generated sequentially and rejected if they do not satisfy the hard-core condition. The generation of points is terminated when a prescribed cardinality of T is achieved.

Next, we need to find a maximum subset of T that can be shattered by tessellations of given type. This necessitates carrying out a deep analysis of the state space of the tessellations. For each element of the state space, the value 0 or 1 is assigned to each point of T, indicating whether the point lies on the corresponding grain boundary of the voxelized version of the tessellation or not. Complete shattering would be achieved if all possible sequences of zeros or ones can be assigned to the points of T by making suitable

choices of tessellation parameters. Obviously, since T contains hundreds or thousands of elements, it would be intractable to investigate all $2^{\#T}$ sequences of zeros and ones. Instead, we search for a maximum subset $T_0 \subseteq T$ having the following two properties:

1. A tessellation exists for which all points of T_0 have the indicator value 0.
2. Each subset of points of T_0 belonging to the same grain boundary can be shattered.

This transforms the problem of shattering test points over the entire network of grain boundaries into an evaluation of the shattering of points on individual grain boundaries.

Instead of varying the tessellation parameters systematically—which would necessitate identifying appropriate parameter ranges and step sizes—we perform a random search through the state space. In particular, we employ the simulated annealing algorithm in modified form to maximize the number of points of T having the indicator value 0. This leads to an approximation of the maximum subset satisfying the first condition. For each proposed change in cell parameters, the indicator values of all points of T are stored and later analyzed with respect to the shattering of points on individual grain boundaries. Thus, *ad hoc* analysis of the simulation allows finding a T_0 that fulfills the aforementioned conditions. It should be noted that the range of (random) parameter values considered by our fitting routine is approximately three times wider than in the simulated annealing algorithm of Section 4, which allows for the deeper search through state space that is required for a correct analysis of shattering.

6.3 Identification of appropriate model

Before proceeding to the presentation of numerical results, we give an overview of parameters used in the computational procedures. In order to reduce computational time, evaluation of the residual sum of squares considered in (12) was based on computing the distance $\|\mathbf{x}-\tilde{\mathbf{x}}\|$ for a randomly chosen subset of points of $\tilde{B}_I$, comprising one tenth of the cardinality of $\tilde{B}_I$. This resulted in the cardinalities n in expressions (13) and (14) ranging from 7.6×10^4 to 2.7×10^5. The number of parameters p therein depends on the number of grains and is much lower, ranging from 2.2×10^3 for the Voronoi tessellation to 3.1×10^4 for the generalized balanced power diagram. Test sets for the estimation of VC-dimension were generated with at most 5 points on each grain boundary, depending on the feasibility of placing 5 points on the particular grain boundary while respecting the hard core condition. This resulted in 1.9×10^3 to 4.5×10^3 test points. VC-dimensions were estimated on the basis of 5 million steps of the simulated annealing methodology mentioned at the end of Section 6.2. Finally, the guaranteed risk considered in (17) was estimated with parameters $a_1 = a_2 = 1$ at confidence level $1-\eta = 0.95$. Here, the RSS differs from the one used in the evaluation of AIC and BIC, because the discrepancy between the response and its approximation is now measured by a binary variable taking

Table 2 Values for the standardized residual sum of squares RSS/n, the Akaike information criterion AIC, the Bayesian information criterion BIC, and the guaranteed risk R_g, evaluated for each simply connected tessellation model fitted to simulated data samples S_1 and S_2. Each quantity is rescaled by the power of 10 indicated in parentheses. The minimum value in each column is printed in bold type.

	sample S_1				sample S_2			
	RSS/n ($\times 10^2$)	AIC ($\times 10^{-4}$)	BIC ($\times 10^{-3}$)	R_g ($\times 10^2$)	RSS/n ($\times 10^2$)	AIC ($\times 10^{-5}$)	BIC ($\times 10^{-4}$)	R_g ($\times 10^2$)
$\mathscr{T}^*_{00}$	664.0	99.0	179.9	83.5	418.2	71.7	74.1	80.7
$\mathscr{T}^*_{01}$	29.7	75.9	-40.3	42.5	21.8	57.2	-70.4	43.7
$\mathscr{T}^*_{02}$	31.1	77.0	1.0	41.0	25.6	58.0	-58.9	30.3
$\mathscr{T}^*_{03}$	20.3	74.2	-5.7	39.5	8.4	52.6	**-111.5**	**23.1**
$\mathscr{T}^*_{10}$	15.8	71.2	-87.5	30.6	29.9	58.7	-54.9	49.2
$\mathscr{T}^*_{11}$	**14.8**	76.8	-21.7	45.1	21.1	57.0	-70.8	44.8
$\mathscr{T}^*_{12}$	20.4	77.8	20.1	41.1	23.7	57.7	-61.6	31.0
$\mathscr{T}^*_{13}$	29.5	77.2	35.2	39.6	**8.2**	**52.5**	-111.3	24.3
$\mathscr{T}^*_{20}$	15.1	**70.8**	**-91.2**	**17.3**	59.2	62.1	-21.2	56.6
$\mathscr{T}^*_{21}$	15.9	74.3	-46.5	38.5	21.7	57.2	-69.4	44.1
$\mathscr{T}^*_{22}$	20.3	77.0	11.6	39.7	27.1	58.3	-54.9	31.5
$\mathscr{T}^*_{23}$	18.8	73.9	1.3	37.9	9.3	53.1	-105.3	24.9

value one for $\mathbf{x} \in \tilde{B}_I$ if $\mathbf{x}$ is not a boundary point between the same cells in the discretized version of the tessellation $\mathscr{T}^*$ and zero otherwise. For this indicator-based RSS, which corresponds to the number of misclassified cases in binary classification, we can use the expression given in (17), employing a VC-dimension based on the same indicator functions. The number of observations n in (17) counts all points of $\tilde{B}_I$ and ranges from 7.6×10^5 to 2.7×10^6.

At first, we focus on the simulated data, for which we know the correct model from which the tessellation was drawn. Results of the fitting were already summarized in Table 1. We note that the correct models—$\mathscr{T}^*_{20}$ for S_1 and $\mathscr{T}^*_{03}$ for S_2—were the most accurate regarding the fraction of correctly assigned voxels as well as with respect to the topological characteristics $\mu^0_{I,\mathscr{T}}$ and $\bar{\mu}_{I,\mathscr{T}}$. Note also that for S_1, the decrease in accuracy from Laguerre tessellation $\mathscr{T}^*_{20}$ to the more general models $\mathscr{T}^*_{21}$, $\mathscr{T}^*_{22}$ and $\mathscr{T}^*_{23}$ can be attributed to the fitting procedure, because there is greater uncertainty when fitting a larger number of parameters. In practice, it is impossible to achieve the same precision for the more general models as for the Laguerre tessellation, as the latter entails optimizing only one parameter per cell in addition to the seed coordinates.

Further insight into the precision of particular models is offered by the standardized RSS given in Table 2 and in Figs. 8(a,b). For the simulated sample S_1, model $\mathscr{T}^*_{11}$ is

even more precise with respect to the standardized RSS than is the Laguerre tessellation. However, both penalization-based criteria, AIC and BIC, prefer the correct model $\mathscr{T}_{20}^*$ (Figs. 8(c,e)). This is also true for the optimal candidate picked out by SRM, for which the preference for the Laguerre tessellation is even more obvious (Fig. 8(g)).

For the simulated data S_2, the lowest values of RSS are attained for the ellipsoidal grain growth model $\mathscr{T}_{03}^*$ and its additively weighted versions $\mathscr{T}_{13}^*$ and $\mathscr{T}_{23}^*$. However, the complexity of these models can be excessive compared to simpler ones. Thus, the model selection criteria described in Section 3 were applied again for identification of an optimal candidate. The results are given in Table 2 and plotted in Fig. 8. Evidently, penalization lowers the differences between models with the same α but different β. However, the increase in accuracy with increasing β outweighs the decrease in simplicity of the models. The ellipsoidal grain growth model is correctly identified by both BIC and SRM, while for AIC the model $\mathscr{T}_{13}^*$ is preferred by a negligible amount. Note that the difference between the non-weighted model $\mathscr{T}_{03}^*$ and its weighted versions $\mathscr{T}_{13}^*$ and $\mathscr{T}_{23}^*$ is small. In this particular case, one extra parameter per cell has only a minor impact on both accuracy and complexity of the tessellation, which implies that the effect of additive weights can be mimicked sufficiently by appropriate modifications of the matrices M_i.

Results for the experimental data sets are provided in Table 3 and Fig. 9. Here, the situation resembles that of the simulated dataset S_2, generated by the ellipsoidal grain growth model. The most complicated models of our collection seem to provide not only the most accurate fitting of polycrystalline data samples, but these models also remain the most appropriate when their increased complexity is taken into account. Again, we observe only negligible differences in the results for different values of α when the parameter β is held at a fixed value greater than zero. This means that the additive weights w_i in these tessellation models have no significant impact on their appropriateness for approximating empirical data.

7 Conclusions

We have applied a simulated annealing methodology to fit a wide class of tessellation models to 3D datasets of polycrystalline materials. Optimal models for each dataset were identified using criteria developed in statistical learning theory. These models meet two requirements: high approximation accuracy and low model complexity.

Validation performed on simulated datasets revealed that several model selection approaches were able to identify the optimal model correctly. For experimental datasets of polycrystalline microstructures, penalization for increasing model complexity according to several criteria does not seem to outweigh the increase in fitting accuracy afforded by tessellation models having a greater number of parameters. We conclude that, in real poly-

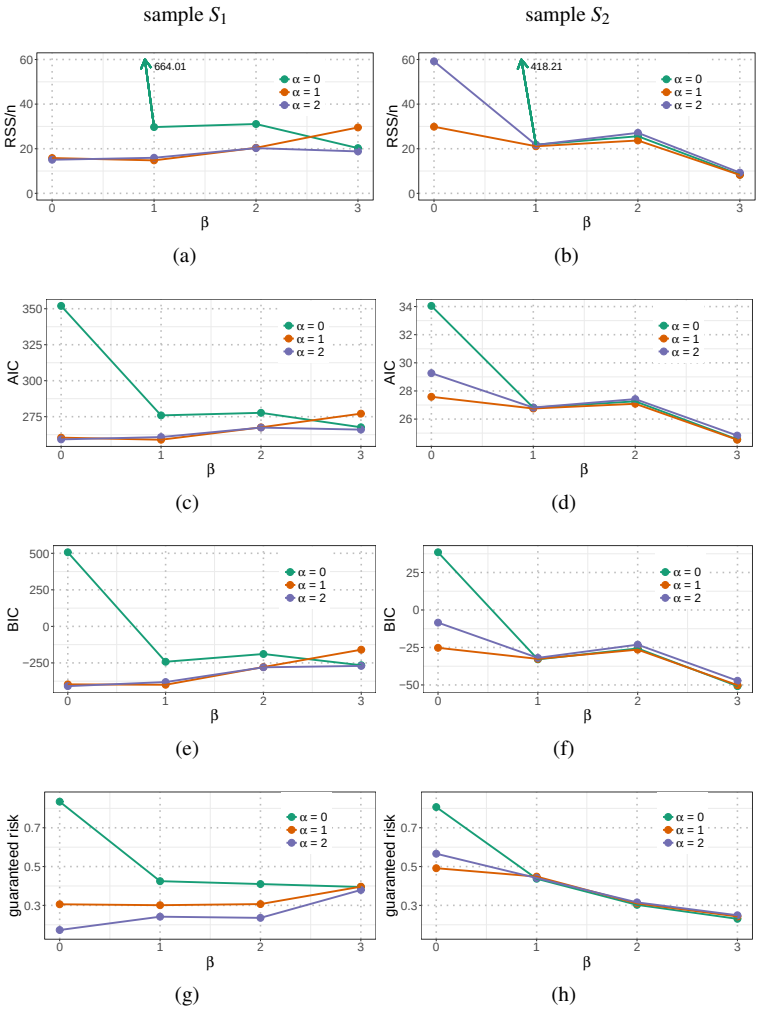


Fig. 8 Evaluation of various criteria used for model selection, applied to simulated data samples S_1 and S_2: (first row) standardized RSS; (second row) Akaike information criterion (AIC); (third row) Bayesian information criterion (BIC); (fourth row) guaranteed risk R_g obtained by SRM. All quantities rescaled as described in the caption of Table 2.

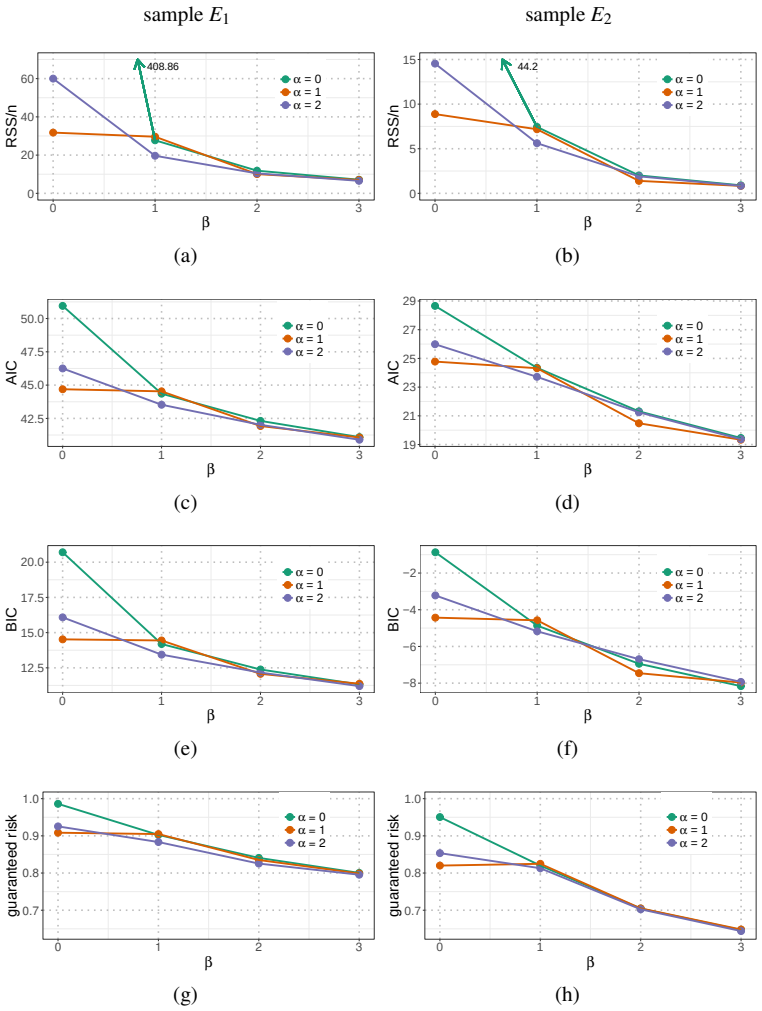


Fig. 9 Evaluation of various criteria used for model selection, applied to experimental data samples E_1 and E_2: (first row) standardized RSS; (second row) Akaike information criterion (AIC); (third row) Bayesian information criterion (BIC); (fourth row) guaranteed risk R_g obtained by SRM. All quantities rescaled as described in the caption of Table 3.

Table 3 Values for the standardized residual sum of squares RSS/n, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the guaranteed risk R_g, evaluated for each simply connected tessellation model fitted to experimental data samples E_1 and E_2. Each quantity is rescaled by the power of 10 indicated in parentheses. The minimum value in each column is printed in bold type.

	sample E_1				sample E_2			
	RSS/n ($\times 10^{-1}$)	AIC ($\times 10^{-5}$)	BIC ($\times 10^{-5}$)	R_g ($\times 10^{-2}$)	RSS/n ($\times 10^{3}$)	AIC ($\times 10^{-5}$)	BIC ($\times 10^{-5}$)	R_g ($\times 10^{2}$)
$\mathscr{T}^*_{00}$	408.9	105.7	41.3	99.4	442.0	60.8	-2.8	95.1
$\mathscr{T}^*_{01}$	27.7	92.4	28.1	91.2	74.5	52.1	-11.2	82.0
$\mathscr{T}^*_{02}$	11.9	88.3	24.2	85.0	20.1	45.8	-16.5	70.4
$\mathscr{T}^*_{03}$	7.1	85.8	21.9	81.1	9.0	41.9	**-19.6**	64.7
$\mathscr{T}^*_{10}$	31.8	93.1	28.8	91.7	88.8	52.9	-10.3	82.0
$\mathscr{T}^*_{11}$	29.6	92.8	28.6	91.5	71.8	51.9	-11.0	82.5
$\mathscr{T}^*_{12}$	10.1	87.5	23.5	84.5	14.0	44.1	-17.9	70.5
$\mathscr{T}^*_{13}$	7.0	85.7	21.9	80.9	**8.3**	**41.6**	-19.6	64.8
$\mathscr{T}^*_{20}$	60.0	96.2	31.9	93.4	145.4	55.4	-7.9	85.3
$\mathscr{T}^*_{21}$	19.7	90.7	26.5	89.2	56.2	50.7	-12.2	81.3
$\mathscr{T}^*_{22}$	10.4	87.7	23.7	83.6	19.1	45.6	-16.3	70.2
$\mathscr{T}^*_{23}$	**6.5**	**85.4**	**21.6**	**80.5**	8.5	41.7	-19.5	**64.4**

crystalline materials, features like the heterogeneity of grain sizes, the local anisotropy of grain shapes and also the curvature of grain boundaries play an important role; therefore, models based on ellipsoids still provide the most reliable approximation of observed microstructures, despite these models' significantly higher complexity.

Acknowledgements This research was funded by the German Science Foundation (DFG) and the Czech Science Foundation (GACR, project number 17-00393J). We are grateful to the Japan Synchrotron Radiation Research Institute for the allotment of beam time on beamline BL20XU of SPring-8 (Proposals 2012A1427 and 2013A1506), and we thank Dmitri Molodov of the Institute of Physical Metallurgy and Metal Physics, RWTH Aachen, for providing the Al-1 wt% Mg specimen.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd International Symposium on Information Theory. Eds: B.N. Petrov and F. Csaki, pp. 267–281. Akademiai Kiado (1973)
2. Alpers, A., Brieden, A., Gritzmann, P., Lyckegaard, A., Poulsen, H.F.: Generalized balanced power diagrams for 3D representations of polycrystals. Philosophical Magazine **95**(9), 1016–1028 (2015)
3. Altendorf, H., Latourte, F., Jeulin, D., Faessel, M., Saintoyant, L.: 3D reconstruction of a multiscale microstructure by anisotropic tessellation models. Image Analysis & Stereology **33**(2), 121–130 (2014)

4. Aurenhammer, F.: Power diagrams: Properties, algorithms and applications. SIAM Journal on Computing **16**(1), 78–96 (1987)
5. Aurenhammer, F., Klein, R., Lee, D.T.: Voronoi Diagrams and Delaunay Triangulations. World Scientific Publishing Co. (2013)
6. Cherkassky, V., Mulier, F.M.: Learning from Data: Concepts, Theory, and Methods, 2nd edn. Wiley-IEEE Press (2007)
7. Chiu, S.N., Stoyan, D., Kendall, W.S., Mecke, J.: Stochastic Geometry and its Applications, 3rd edn. J. Wiley & Sons (2013)
8. Dake, J.M., Oddershede, J., Sørensen, H., Werz, T., Shatto, J.C., Uesegi, K., Schmidt, S., Krill III, C.E.: Direct observation of grain rotations during coarsening of a semisolid Al-Cu alloy. Proceedings of the National Academy of Sciences **113**, E5998–E6006 (2016)
9. Gelfand, A., Diggle, P., Guttorp, P., Fuentes, M.: Handbook of Spatial Statistics. Chapman & Hall/CRC (2010)
10. Groeber, M.A., Jackson, M.A.: DREAM.3D: A digital representation environment for the analysis of microstructure in 3D. Integrating Materials and Manufacturing Innovation **3**(1), 1–17 (2014)
11. Jeulin, D.: Random tessellations and Boolean random functions. In: C.L. Luengo Hendriks, G. Borgefors, R. Strand (eds.) Mathematical Morphology and Its Applications to Signal and Image Processing, pp. 25–36. Springer (2013)
12. Lyckegaard, A., Lauridsen, E.M., Ludwig, W., Fonda, R.W., Poulsen, H.F.: On the use of Laguerre tessellations for representations of 3D grain structures. Advanced Engineering Materials **13**(3), 165–170 (2011)
13. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd edn. J. Wiley & Sons (2000)
14. Scheike, T.H.: Anisotropic growth of Voronoi cells. Advances in Applied Probability **26**(1), 43–53 (1994)
15. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
16. Spettl, A., Brereton, T., Duan, Q., Werz, T., Krill III, C., Kroese, D.P., Schmidt, V.: Fitting Laguerre tessellation approximations to tomographic image data. Philosophical Magazine **96**(2), 166–189 (2016)
17. Spettl, A., Wimmer, R., Werz, T., Heinze, M., Odenbach, S., Krill III, C.E., Schmidt, V.: 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), 065,001 (2015)
18. Šedivý, O., Brereton, T., Westhoff, D., Polívka, L., Beneš V. Schmidt, V., Jäger, A.: 3D reconstruction of grains in polycrystalline materials using a tessellation model with curved grain boundaries. Philosophical Magazine **96**(18), 1926–1949 (2016)
19. Šedivý, O., Dake, J., Krill III, C., Schmidt, V., Jäger, A.: Description of the 3D morphology of grain boundaries in aluminum alloys using tessellation models generated by ellipsoids. Image Analysis & Stereology **36**, 5–13 (2017)
20. Teferra, K., Graham-Brady, L.: Tessellation growth models for polycrystalline microstructures. Computational Materials Science **102**, 57–67 (2015)
21. Valiev, R.Z., Langdon, T.G.: Principles of equal-channel angular pressing as a processing tool for grain refinement. Progress in Materials Science **51**(7), 881–981 (2006)
22. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
23. Vapnik, V.N.: The Nature of Statistical Learning Theory, 2nd edn. Springer (2000)