







# Modified DBSCAN Algorithm for Microscopic Image Analysis of Wood

Aurora L. R. Martins<sup>1</sup> , André R. S. Marcal<sup>1</sup>  , and José Pissarra<sup>1,2</sup> 

<sup>1</sup> Faculdade de Ciências da Universidade do Porto, Porto, Portugal  
andre.marcal@fc.up.pt

<sup>2</sup> GreenUPorto – Research Centre for Sustainable Agrifood Production, FCUP,  
Porto, Portugal

**Abstract.** The analysis of the intern anatomy of wood samples for species identification is a complex task that only experts can perform accurately. Since there are not many experts in the world and their training can last decades, there is great interest in developing automatic processes to extract high-level information from microscopic wood images. The purpose of this work was to develop algorithms that could provide meaningful information for the classification process. The work focuses on hardwoods, which have a very diverse anatomy including many different features. The ray width is one of such features, with high diagnostic value, which is visible on the tangential section. A modified distance function for the DBSCAN algorithm was developed to identify clusters that represent rays, in order to count the number of cells in width. To test both the segmentation and the modified DBSCAN algorithms, 20 images were manually segmented, obtaining an average Jaccard index of 0.66 for the segmentation and an average index  $M = 0.78$  for the clustering task. The final ray count had an accuracy of 0.91.

**Keywords:** Wood anatomy · Ray width · Image processing · DBSCAN

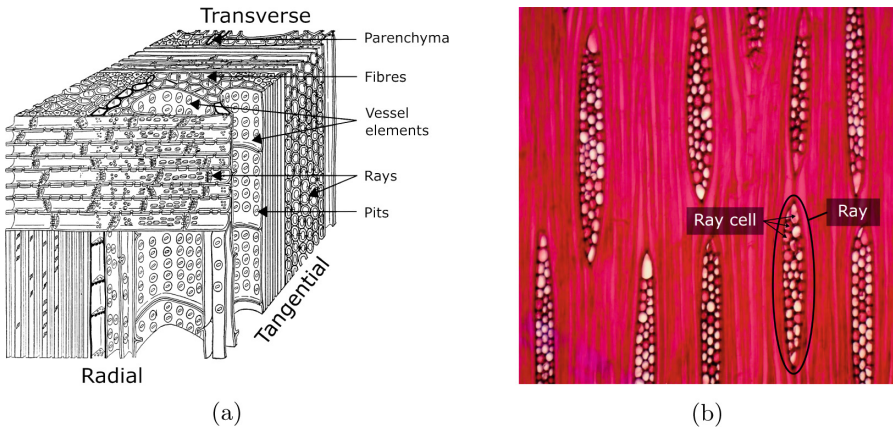
## 1 Introduction

Wood identification is a relevant issue for various applications, such as illegal logging, conservation and restoration, among others [14]. Currently, there are several methods to identify wood species, including chemical, genetic and visual identification [5]. The work presented here focuses on the later, through the analysis of the intern anatomy of wood samples, using microscopic images. The microscopic identification of wood is based in three different cuts in the wood sample that form three different sections: transverse, radial and tangential. A variety of features and pattern recognition techniques can be used to analyse the images of different sections, in order to assist in the identification process.

There are two groups of woods - softwoods and hardwoods. The relevant features for each of these types are significantly different. Softwoods have simpler anatomy than hardwoods, which have many more relevant features and more variety within each feature. In Fig. 1a wood sample from hardwood is presented showing the three sections and some of the main structures found in this type of wood. Each section has important features to support an accurate identification, especially because these features may vary within species. One important feature for the hardwood species is the ray width, which can be seen in the tangential section. This criterion indicates the number of cells that compose a ray in width. These ray cells are usually oval or lens shaped objects. An example of the tangential section of a hardwood sample is presented in Fig. 1b, for a species with bi- to 3-seriate rays.

Ray width can be classified as [15]:

- Uniseriate rays - only one cell wide.
- Bi- to 3-seriate rays - most of the rays are 2-3 cells wide, but rare uni- and 4-seriate rays are present.
- 3- to 5-seriate rays - most of the rays are 3-5 cells wide, but rare uni- and biseriate rays are present.
- Uni- and multiseriate rays - in addition to numerous uniseriate rays, large rays are also present.



**Fig. 1.** (a) Hardwood anatomy, (b) Microscopic image - tangential section

Although the analysis of microscopic wood anatomy is the most reliable method nowadays to provide an initial identification, only experts can perform this task with accuracy. There are only a few wood anatomists in the world and the training needed to educate an experienced one can take decades [5], thus the importance of automation to assist in this process. In recent years there has been an increase in the work developed in this topic, but it is still an area relatively

unexplored. The systems available only analyse the transverse section (ignoring the other two sections) as it is the section that contains most information. Two main tasks are included in existing systems: image processing techniques to segment objects of interest [1, 2, 6, 8, 11] and automatic feature extraction for species classification [3, 9, 10, 13, 16–18]. However, none of the available systems joins the segmentation task and the species classification task with features that are meaningful in a biological sense to perform a species identification.

This work focuses on the use of the tangential section of microscopic images of hardwood, with the aim of providing additional meaningful information for the classification process. The segmentation task is based on an edge detector method, followed by a selection of the objects of interest. In order to separate the clusters that represent the rays, a clustering algorithm is used. Since the number of clusters is unknown, a density based algorithm is used - the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [7]. A modified version of the MATLAB implementation of DBSCAN [4] was developed in order to correctly separate the different clusters by defining an alternative distance function to be used by the algorithm. Finally, the cell count is computed for each cluster.

## 2 Materials and Methods

### 2.1 Wood Anatomy Database

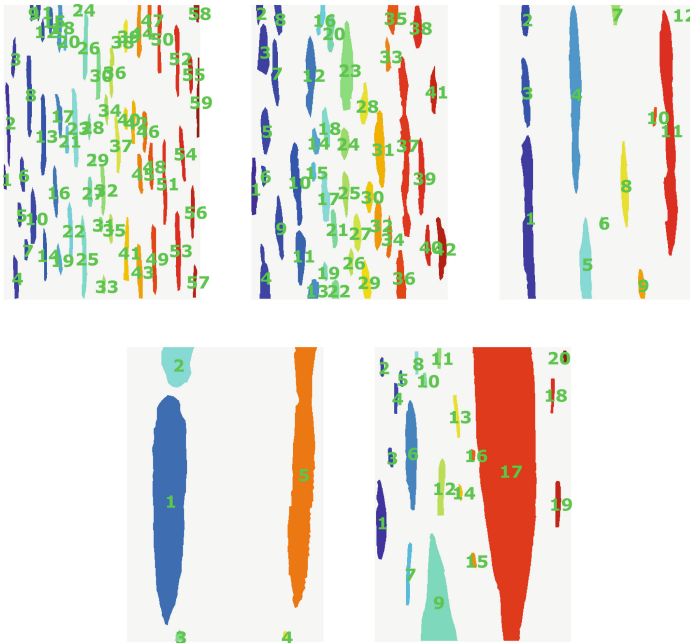
The database used is a portion of a larger database - the Wood anatomy of central European Species [15]. It is composed of 1908 microscopic images from 133 species, with grey scale images of  $600 \times 400$  pixels from the three sections. The database has 383 tangential section images (stemwood samples with a magnification of  $100\times$ ), from 127 species.

To evaluate the performance of the algorithms developed, 20 tangential section images were manually segmented. Four examples of each ray width class (uniseriate, bi- to 3-seriate, 3-seriate to 5-seriate, multiseriate, uni- and multi-seriate) were selected. The manually segmented objects were the rays, instead of the ray cells, meaning that the contours of the ray cells are included in this image. Examples of manually segmented images of each class are present in Fig. 2, where each ray has a different colour.

### 2.2 Segmentation of Ray Cells

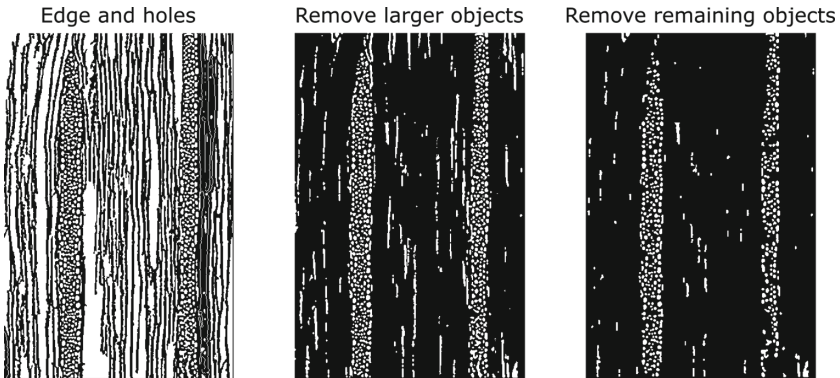
The first step is to segment the ray cells, which can have different levels of intensity. Usually these objects are very bright (white) but quite often they are not (intermediate grey levels), with the same intensity values as other objects that are not ray cells. Furthermore, there are also objects that are not ray cells but that are very bright (nearly white).

An edge detection spacial filter is used - Laplacian of the Gaussian [12] with  $\sigma = 1.5$  and an  $11 \times 11$  kernel, including all zero-crossings. This filter is used



**Fig. 2.** Example of manually segmented images for each ray width class. From left to right - top: uniseriate; bi- to 3-seriate; 3- to 5-seriate; bottom: multi-seriate; uni- and multi-seriate.

because the contours are, in most cases, well defined. The zero-crossing results in closed contours on the output image, which are filled to obtain most of the closed objects of the image. Then, the objects that differ from ray cells in terms of size, shape and intensity values are removed (very large, small, elongated



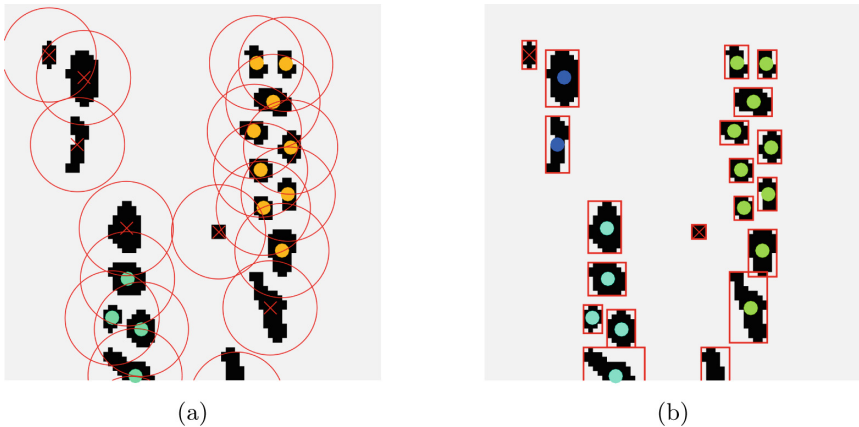
**Fig. 3.** Example of the 3 main segmentation processing stages.

and darker objects). Although not all the unwanted objects are eliminated and some unwanted objects are kept, the final results produced are adequate for the subsequent processing. Figure 3 presents an example of the three main steps of the algorithm.

### 2.3 DBSCAN

The DBSCAN algorithm [7] has two inputs: the minimum number of points a cluster needs to have in order to be considered valid and a fixed radius for the search of neighbours. Basically, the algorithm assigns points to a cluster if they are within the radius of search of at least one of the points already assigned to that cluster. If the cluster does not have the minimum number of points, then those points are considered noise. This algorithm is usually applied to numeric datasets to aggregate sets of points, but here it is applied to the ray cells' centroid coordinates.

Since the objects vary considerably in size and the distance between two objects is dependent on the radii of the two, the choice of the fixed radius needed for the DBSCAN is hard to make. An illustration of the method is presented in Fig. 4a, applied to a small portion of an image of the tangential section. The radius chosen in this case was 10 pixels, with 2 as the minimum number of points. In this example it is noticeable that a radius of 10 pixels might be too small, since many objects were considered noise (represented by a red cross) when they should not have been. However, increasing the radius would introduce real noise points inside of clusters and also join clusters that should be apart.



**Fig. 4.** Illustration of the DBSCAN method: (a) original, (b) modified version. (Color figure online)

The DBSCAN algorithm distance function was thus modified to compute the distance between the bounding boxes of the objects. In this way, instead of

looking for points within a radius, it looks for objects that are closer from the left/right and up/down. A higher distance is assigned to the vertical component and a smaller one to the horizontal component because joining two clusters will only compromise the ray width if they are side by side. The accepted distance is now a two component vector with a fixed horizontal and vertical values of 2 and 10 pixels, respectively, and the minimum number of points is 2. In Fig. 4b presents the result for the same test image, using this modified version, where the bounding box for each object is represented as a red rectangle. The points that do not have neighbours closer than 10 pixel vertically and 2 pixel horizontally are considered noise (marked by a red cross).

The ability of the DBSCAN algorithm in dealing with noise is particularly useful in this context. Objects identified on the segmentation task that are not ray cells and are isolated become noise and are not considered as ray cells.

## 2.4 Ray Width

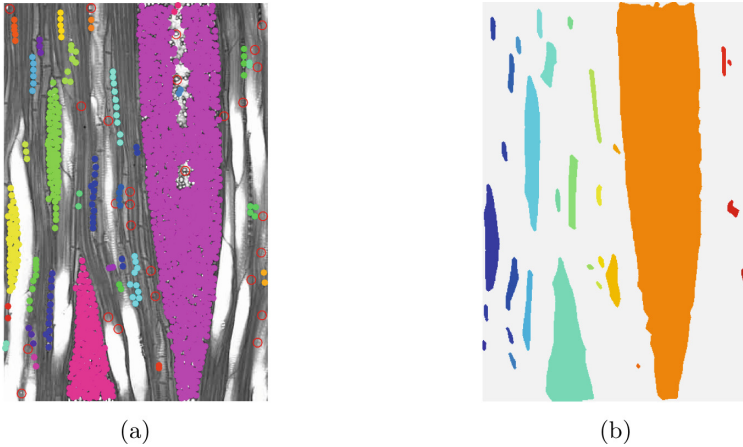
A binary image of each cluster is produced to analyse the number of ray cells. In the example of Fig. 5, two such images are represented with green colour, superimposed on the original grey scale images. A simple but effective approach was implemented: scan horizontal rows and count the number of times a transition from 1 (ray cell - green) to 0 (negative space - not green) occurs. To count the ray cells that are in the margins of these images, zero-padding is previously added. The ray width for each cluster is the highest number of transitions found on each row. In order to reduce the computational effort, only one out of three lines are scanned. Although this sub-sampling can affect the final number of ray cells counted per row, the differences were found to be negligible.



**Fig. 5.** Examples of 2 sub-images of ray clusters (the segmentation is superimposed in green). (Color figure online)

## 2.5 Evaluation

Since the manually segmented images contain closed regions (clusters) representing groups of rays instead of the individual ray cells (the contours of the ray cells are included), the segmented image used to compare the results should also be composed of closed objects, one for each cluster. In order to achieve such image, the contour of each cluster returned by the modified DBSCAN algorithm (excluding the points considered noise) is computed, based on the vertices of the bounding boxes of all the objects that belong to that cluster. The contour is then filled, returning an approximation of the closed object that represents



**Fig. 6.** Example of the computed clusters (a) which are used to create the “cluster image” (b).

the computed cluster. The image that contains all these closed objects is named “cluster image”. Figure 6a shows the clustering results that are then used to create the “cluster image” presented in Fig. 6b. It should be noted that, since the “cluster image” is an approximation, some errors related to the clusters shapes are introduced on the evaluation.

To evaluate the segmentation of the image rays, the Jaccard metric is used. The Jaccard metric, or index, is also known as intersection over union (the intersection of two sets over the union of the two sets). In this case, the sets are the manually segmented image and the “cluster image”.

To evaluate the results of the modified DBSCAN algorithm, the metric should be computed for each cluster, and these values then used to compute a final metric for the whole image. This can be done from the perspective of the clusters on the manually segmented image or from the clusters of the “cluster image”. A one-to-one correspondence is needed between the clusters. Depending on which image we choose as reference, some clusters will be disregarded. For example, if one chooses the manually segmented image as reference, the predicted clusters that are not real ones (false clusters) are never assigned to a cluster. Instead of choosing a reference image, a metric is computed for both images (as reference) and the average result is used.

Let  $I_c$  be the “cluster image” and  $I_m$  the manually segmented image. The metric used for each cluster of the “cluster image” and for each cluster of the manually segmented image are, respectively, given by:

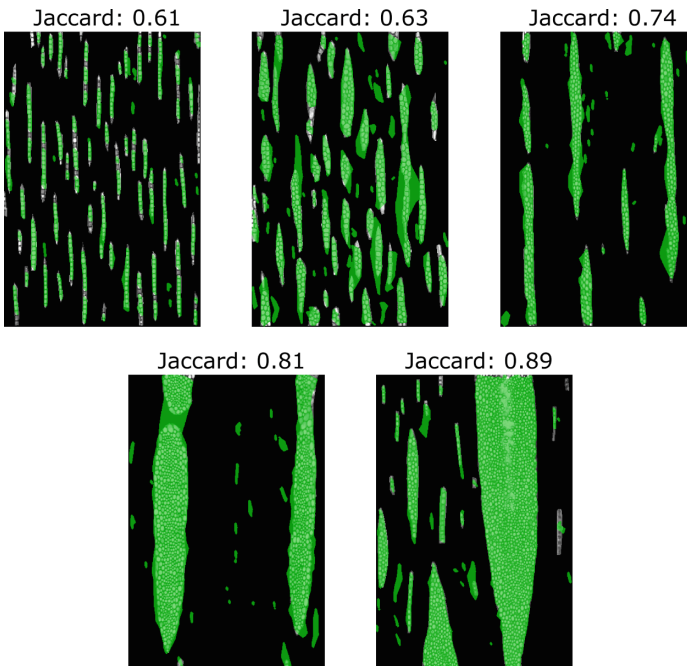
$$M_{c_i} = \frac{I_{c_i} \cap I_m}{I_{c_i}}, \quad M_{m_j} = \frac{I_{m_j} \cap I_c}{I_{m_j}},$$

where  $I_{c_i}$  is the image that contains only the predicted cluster  $i = 1, \dots, N_c$  and  $I_{m_j}$  is the image that contains only the manually segmented cluster

$j = 1, \dots, N_m$ . This metric evaluates how well each cluster is identified, for both the predicted ones ( $M_{c_i}$ ) and for the manually segmented ones ( $M_{m_j}$ ). A weighted mean is applied to the resulting metrics, according to the area of the clusters, since larger clusters are more significant than smaller ones. The final result is a single metric for each image -  $M_c$  and  $M_m$ . A geometric mean is then used to get the final overall metric:  $M = \sqrt{M_c \times M_m}$  (values between 0 to 1, with 1 as the perfect result).

### 3 Results

To illustrate the type of results produced by the segmentation task, Fig. 7 shows one image from each of the classes considered (uniseriate, bi- to 3-seriate, 3- to 5-seriate, multiseriate, uni- and multiseriate). The figure shows the original images (masked by the manual segmentation) with the cluster image superimposed. The values of the Jaccard index are also presented for each image. Although there are some misdetection of ray cells (both false positives and false negatives), the results can be considered adequate for the subsequent processing because they generally capture the width of the rays.

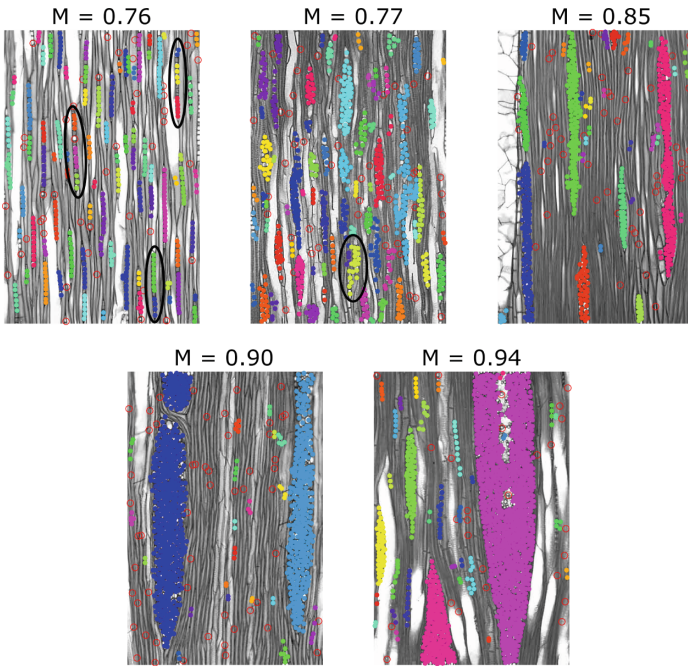


**Fig. 7.** Example of the segmentation results for each ray width class. From left to right - top: uniseriate; bi- to 3-seriate; 3- to 5-seriate; bottom: multiseriate; uni- and multiseriate.



**Table 1.** Segmentation evaluation results - Jaccard index values. O - original DBSCAN; M - modified DBSCAN (in bold when difference is  $\geq 0.05$ ).

Class	Image 1		Image 2		Image 3		Image 4		Mean	
	O	M	O	M	O	M	O	M	O	M
Uniseriate	0.37	<b>0.61</b>	0.12	<b>0.31</b>	0.22	<b>0.70</b>	0.05	<b>0.37</b>	0.19	<b>0.50</b>
Bi- to triseriate	0.49	<b>0.63</b>	0.45	<b>0.52</b>	0.56	<b>0.69</b>	0.22	<b>0.35</b>	0.43	<b>0.55</b>
Tri- to 5-seriate	0.77	0.74	0.50	0.53	0.82	0.80	0.80	0.76	0.72	0.71
Multiseriate	0.84	0.81	0.87	0.88	0.76	0.75	0.17	<b>0.92</b>	0.66	<b>0.84</b>
Uni- and multiseriate	0.88	0.89	0.59	0.59	0.46	<b>0.59</b>	0.39	<b>0.74</b>	0.58	<b>0.70</b>

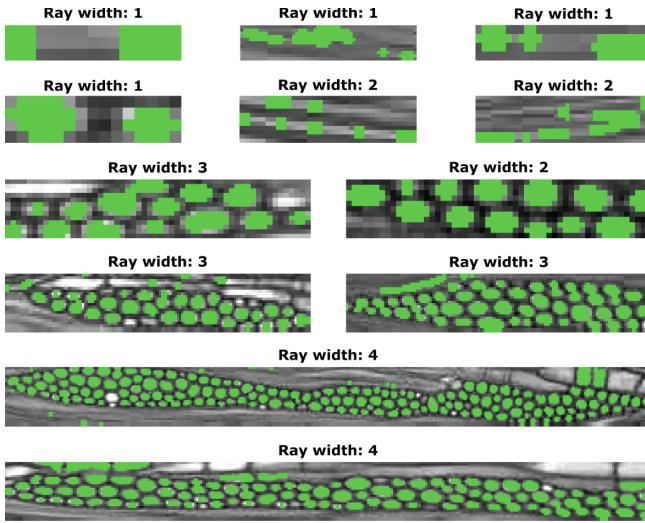


**Fig. 8.** Example of the clustering results for each ray width class. Top: uniseriate, bi- to 3-seriate, 3- to 5-seriate. Bottom: multiseriate, uni- and multiseriate.

The Jaccard index values for the 20 images tested (4 from each class) are present in Table 1, both for the original DBSCAN and for the modified DBSCAN, as well as the mean for each category. The index is higher for images with larger rays, possibly because the shape of smaller rays is more difficult to identify with the method used to produce the “cluster image”. Other factors also have to be considered, such as the fact that darker ray cells tend to be eliminated in the process. The average index for all 20 images is 0.52 for the original DBSCAN and 0.66 for the modified version.

**Table 2.** DBSCAN evaluation results -  $M$  index. O - original DBSCAN; M - modified DBSCAN (in bold when difference is  $\geq 0.05$ ).

Class	Image 1		Image 2		Image 3		Image 4		Mean	
	O	M	O	M	O	M	O	M	O	M
Uniseriate	0.58	<b>0.76</b>	0.27	<b>0.49</b>	0.45	<b>0.82</b>	0.14	<b>0.57</b>	0.36	<b>0.66</b>
Bi- to triseriate	0.67	<b>0.77</b>	0.63	<b>0.69</b>	0.73	<b>0.81</b>	0.37	<b>0.52</b>	0.60	<b>0.70</b>
Tri- to 5-seriate	0.87	0.85	0.68	0.70	0.90	0.89	0.89	0.86	0.84	0.83
Multiseriate	0.91	0.90	0.93	0.93	0.87	0.86	0.41	<b>0.96</b>	0.78	<b>0.91</b>
Uni- and multiseriate	0.94	0.94	0.75	0.75	0.65	<b>0.74</b>	0.62	<b>0.85</b>	0.74	<b>0.82</b>



**Fig. 9.** Ray cells count.

Figure 8 shows the results obtained with the modified version of DBSCAN for one image of each class. The hollow red circles represent points that are considered noise, and the filled coloured circles represent the points that belong to clusters, with different colours representing different clusters. Again, the metric considered ( $M$ ) is displayed for each image. Some rays are split vertically, but the ray width remains the same. The problems arise when the split is horizontal, which introduces two false ray widths. The reverse situation, where clusters are joined, is also detrimental for the vertical case because a larger cluster can hide a smaller one. In the first image we can see that some clusters were split vertically and in the second image we can see an example of two clusters that were joined horizontally, marked with a black ellipse. Fortunately, these situations were not found to be very common.

A visual inspection suggests that the results are adequate, as the clusters are mostly well identified. The quantitative results are presented in Table 2. For classes uniseriate and bi- to 3-seriate (first two rows in Table 2) the values of the M index are lower than the remaining classes. This is consistent with the segmentation results, that have a high influence on the DBSCAN results. The multiseriate images are the ones with the best performance, achieving an average M value of 0.91. In general, the results are fairly good, with an overall average of 0.78 for the modified version of DBSCAN, which is an improvement compared to the original DBSCAN that achieved 0.66.

For the ray width evaluation, clusters were individually analysed to check if the assigned ray width corresponds to the expected values (ground truth). The ray width was evaluated considering only what the algorithm should count and not the actual ray width, meaning that false rays are considered correctly classified in terms of ray width if the count is correct. Figure 9 presents examples of the sub-images obtained for some clusters of a tri- to 5-seriate species, including the predicted ray width. Some of these rays were wrongly identified, such as the first two sub-images, but their count is nevertheless correct.

A confusion matrix (Table 3) was built to compute the accuracy of the ray cells count. It shows that the misclassified cluster have an error of one unit at the most, which is mostly an underestimation. These underestimations happen because only one out of three lines are being analysed, missing the lines that have the maximum width present. The accuracy of the ray width count is 0.91.

**Table 3.** Confusion matrix for the ray width estimation.

Computed	Reference					
	1	2	3	4	5	>5
1	<b>166</b>	6	0	0	0	0
2	1	<b>40</b>	5	0	0	0
3	0	0	<b>31</b>	6	0	0
4	0	0	0	<b>16</b>	6	0
5	0	0	0	0	<b>9</b>	3
>5	0	0	0	0	0	<b>11</b>

## 4 Conclusion

A modified distance function for the DBSCAN method was developed to assist in extracting information from microscopic images of hardwoods (tangential section). The method is used to count the number of ray cells in width. It can be applied to other image processing tasks when there is a need to form an unknown number of groups from the objects present in the image.

The accuracy of the clustering task is obviously dependent on the segmentation task, since missing/additional objects will cause splitting/merging of clusters. In the experimental evaluation carried out, the average segmentation Jaccard index was 0.66 and the average DBSCAN index  $M$  was 0.78. The accuracy of the algorithm developed to identify the ray width was found to be 0.91. As future work the ray cells count may be used to extract features and train a classifier to predict the ray width given an image as input.

## References

1. Arx, G.V., Dietz, H.: Automated image analysis of annual rings in the roots of perennial forbs. *Int. J. Plant Sci.* **166**(5), 723–732 (2005)
2. Brunel, G., Borianne, P., Subsol, G., Jaeger, M., Caraglio, Y.: Automatic identification and characterization of radial files in light microscopy images of wood. *Ann. Bot.* **114**, 829–840 (2014)
3. Cavalin, P.R., Kapp, M.N., Martins, J., Oliveira, L.E.S.: A multiple feature vector framework for forest species recognition. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC 2013, pp. 16–20 (2013)
4. Daszykowski, M., Walczak, B., Massart, D.L.: Looking for natural patterns in data. Part 1: density based approach. *Chemom. Intell. Lab. Syst.* **56**, 83–92 (2001)
5. Dormontt, E.E., et al.: Forensic timber identification: it's time to integrate disciplines to combat illegal logging. *Biol. Conserv.* **191**, 790–798 (2015)
6. Espinosa, L.F., Herrera, R.J., Polanco-Tapia, C.: Segmentation of anatomical elements in wood microscopic images using artificial vision techniques. *Maderas. Ciencia y tecnología* **17**(4), 735–748 (2015)
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Knowledge Discovery and Data Mining, p. 226 (1996)
8. Fabijańska, A., Danek, M., Barniak, J., Piórkowski, A.: Towards automatic tree rings detection in images of scanned wood samples. *Comput. Electron. Agric.* **140**, 279–289 (2017)
9. Martins, J., Oliveira, L.S., Nisgoski, S., Sabourin, R.: A database for automatic classification of forest species. *Mach. Vis. Appl.* **24**, 567–578 (2013)
10. Martins, J.G., Oliveira, L.S., Britto, A.S.B., Sabourin, R.: Forest species recognition based on dynamic classifier selection and dissimilarity feature vector representation. *Mach. Vis. Appl.* **26**, 279–293 (2015)
11. Pan, S., Kudo, M.: Segmentation of pores in wood microscopic images based on mathematical morphology with a variable structuring element. *Comput. Electron. Agric.* **75**, 250–260 (2011)
12. Parker, J.: Algorithms for Image Processing and Computer Vision. Wiley Publishing, Inc., Hoboken (2011)
13. Rosa da Silva, N., et al.: Automated classification of wood transverse cross-section micro-imagery from 77 commercial Central-African timber species. *Ann. Forest Sci.* **74**, 30 (2017)
14. Ross, R.J.: Wood Handbook - Wood as an Engineering Material. Department of Agriculture, Forest Service, Forest Products Laboratory, centennial ed. general technical report fpl; gtr-190. madison, wi: u.s edn. (2010)
15. Schoch, W., Heller, I., Schweingruber, F., Kienast, F.: Wood anatomy of central European Species (2004). [www.woodanatomy.ch](http://www.woodanatomy.ch)

16. Yadav, A.R., Anand, R.S., Dewal, M.L., Gupta, S.: Hardwood species classification with DWT based hybrid texture feature extraction techniques. *Sadhana* **40**(8), 2287–2312 (2015)
17. Yadav, A.R., Anand, R.S., Dewal, M.L., Gupta, S.: Multiresolution local binary pattern variants based texture feature extraction techniques for efficient classification of microscopic images of hardwood species. *Appl. Soft Comput. J.* **32**, 101–112 (2015)
18. Zamri, M.I.P., Khairuddin, A.S.M., Mokhtar, N., Yusof, R.: Wood species recognition system based on improved basic grey level aura matrix as feature extractor. *J. Robot. Netw. Artif. Life* **3**(3), 140–143 (2016)