A CRITICAL ANALYSIS OF THE METHODS OF EVALUATING MRI BRAIN SEGMENTATION ALGORITHMS

*Fábio A. M. Cappabianco*** *Paulo A. V. de Miranda*[†] *Jayaram K. Udupa*[‡]

[*] Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, Brazil
[†] Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
[‡] MIPG, Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA

ABSTRACT

Many papers are published every year containing new methodologies for brain tissue segmentation in magnetic resonance images. The evaluation of these methods is fundamental to understand their behavior and to observe their weak and strong points. Even though improvements have been proposed, the analysis of the segmentation results can still lead to incorrect conclusions. This paper contains an investigation of the state-of-the-art in brain tissue segmentation evaluation, which includes tissue classification or segmentation, handling partial volume effect, and evaluation metrics. It uncovers previously unnoticed pitfalls and proposes standard procedures to avoid them. Experiments show that the proposed evaluation strategy gives a better insight about the method's strong and weak points.

***Index Terms*—** Brain Tissue Segmentation, Brain Tissue Classification, Medical Imaging, Partial Volume Effect

1. INTRODUCTION

According to Google Scholar, in searches excluding citations and patents, 174 papers that contain the search keys "brain tissue segmentation" and "brain tissue classification" in their main title have been published in the last 24 years. 48.3% of this amount, i.e. 84 papers, were published in the last four years. And, 2112 papers in the last 24 years contain one of these keywords in their main text. These numbers clearly show that the research topics directly involving tissue segmentation is growing in interest and importance. In fact, a number of applications that include functional magnetic resonance imaging [1], voxel based morphometry [2], and pathology etiology studies [3] strongly depend on the quality of structural image segmentation.

This paper complements previous studies about evaluation of tissue segmentation [4]. Even though the numbers reflect a positive aspect of the research in the area, several works contain a number of uncertainties in their experimental evaluation. We will analyze three of them in Section 2: the imprecisions of tissue segmentation definition, absence of partial volume effect (PVE) in ground-truths, and inappropriate evaluation metrics. We propose more balanced solutions to each of these pitfalls in Section 3. We demonstrate the effectiveness of these solutions with experiments over a popular dataset and state-of-the-art methodologies in Section 4. Finally, we state our conclusions in Section 5.

2. ISSUES IN SEGMENTATION VALIDATION

2.1. Tissue Segmentation Definitions

Before defining brain tissue segmentation we will review brain anatomy. The brain is composed of gray matter (GM) and white matter (WM) tissues, and it is surrounded by cerebrospinal fluid (CSF). The brain may also be divided anatomically into three main regions: the forebrain that contains the telencephalon, i.e. two cerebral hemispheres, and the diencephalon; the midbrain; and the hindbrain that includes the metencephalon and the myelencephalon. The cerebellum is part of the metencephalon.

The importance of CSF delineation depends on posterior analysis purpose. It may be more relevant in the context of brain atrophy estimation [5] than in that of functional analysis of subcortical structures [6]. Because of the distinct expectations and because of the extra effort to segment external CSF the Internet Brain Segmentation Repository (IBSR) [7] [1] discriminate only CSF inside the ventricles (Fig. 1). Valverde et.al[8] propose an interesting study that ignores the sulcal CSF while evaluating the tissue segmentation of a given method. However, this is not the best solution since distinct segmentations would ignore different volumes, depending on the amount of sulcal CSF that is labeled.

The midbrain and hindbrain region delineation is also troublesome. Depending on the aimed application, metencephalon with or without the cerebellum or the myelencephalon may not be important. Also, it is very difficult to set a clear boundary where the myelencephalon ends and the

[*]Thanks to FAPESP(2016/21591-5) and CNPq (486988/2013-9,452985/2017-0) for the financial support.

[1] Images and manual segmentations: Center for Morphometric Analysis at Massachusetts General Hospital (http://www.cma.mgh.harvard.edu/ibsr/)

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

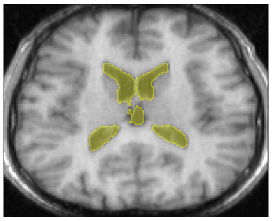
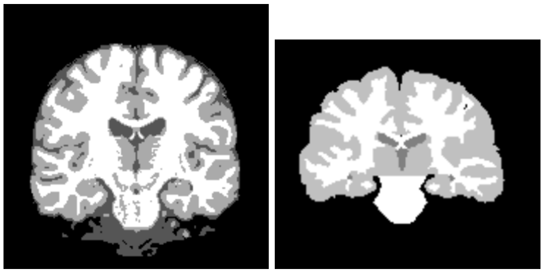


Fig. 1. Sample slice of IBSR dataset images with overlapped CSF ground-truth label.

spinal cord begins. In IBSR dataset, the complete hindbrain and midbrain are labeled as being composed of WM which is an overestimation (Fig. 2). That is not the case in BrainWeb Phantom (BWP) dataset, where a larger portion of the spinal cord appears in the image domain. These concerns are very important since a method that fails to include a relatively large portion of the GM in the cortex region may achieve a better score than a more accurate one just because it segmented a larger or smaller portion of the hindbrain or external CSF.



(a) (b)

Fig. 2. Midbrain and hindbrain labeling in (a) BWP and (b) IBSR dataset images. In (b), the entire region is labeled as WM.

2.2. Strong Influence of Partial Volume Effect

As the voxel is the smallest 3D element of the image, tissue segmentation methods label each voxel with a single class [4]. Tissue classification assigns the fraction of each tissue that the voxels contain. As voxels are not small enough to separate tissues perfectly, this phenomena came to be known as partial volume effect (PVE) [9][10]. The output of some methods includes voxels that are explicitly labeled as voxels with PVE. Still, brain delineation tissue gold standards are extremely burdensome to generate because of the intricate details and complex shape of the sulci and gyri. Including partial volume labels would further exacerbate the task of ground truth generation significantly.

PVE is very relevant when evaluating segmentation/classification methods. IBSR dataset proposed to label the voxels that have at least one adjacent voxel with a distinct label as partial volume voxels. With this strategy, from 26.0% to 38.2% of the voxels contain partial volume of more than one tissue. Therefore, PVE voxels have a considerable influence over the evaluation metric outputs.

2.3. Evaluation metric issues

The last aspect to be analyzed involves evaluation metrics. There are two distinct quantitative evaluation metric types: volume based and border based. They may also be used without a ground-truth as proposed by [4]. The idea is to use any area or border based metric to compare several methods among themselves in order to find the most common among them. However, this strategy results in an output that strongly favors methods that are most similar and may penalize methods that outperform the others with a more accurate answer.

The most commonly used border distance metrics are the Hausdorff distance (HDD) [11, 12] and the mean Euclidean distance error(EDE) [13, 14]. These measurements are not very precise for brain tissue evaluation because of the complex shape of the brain tissue regions.

Volume based metrics make use of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) labeled voxels, for comparing the ground-truth with the segmentation output of a method. In [15, 16] the authors use the similarity metrics Observed Agreement and κ measurement. These metrics consider the TN pixels that may be much larger than the brain itself, resulting in a less sensitive measurement. Even if we consider a perfect skull stripping, the TN of CSF consists of all GM and WM voxels, which outnumbers CSF voxels.

The same can be said about more elaborated metrics such as the receiver operating characteristic (ROC) curve [17] and the delineation operating characteristic (DOC) curve [18, 19]. Even though the influence of $|TN|$ is reduced by a normalization, these curves are more sensitive to under or over segmentation depending on the ratio between the size of the object and non-object labels.

Finally, Dice(DC) [20] and Jaccard(JC) [21] metrics given by Equations 1 and 2, respectively, do not make use of TN. Their disadvantage is to be more sensitive to under segmentation than to over segmentation, especially for more than 10% of misclassified voxels. For instance, in a segmentation of a tissue having 100,000 voxels, a method that outputs a result with 0 FN and 20,000 FP will score 0.83 and 0.91 in JC and DC metrics, respectively. If another method misses the same number of voxel, but with 20,000 FN and 0 FP, it will score 0.80 and 0.89, respectively. These are values commonly found in the literature [22] and they clearly favor FP over FN errors.

$$DC = \frac{2TP}{2TP + FP + FN} \tag{1}$$

$$JC = \frac{TP}{TP + FP + FN} \tag{2}$$

One may argue that higher FN incurs in lower TP, while FP does not. Note that labeling the entire image as object or background is trivial and completely incorrect. Supposing that object and background have the same size, segmenting the entire image as object will produce 0.67 and 0.5 scores in DC and JC metrics, respectively. On the other hand, labeling it completely as background would achieve 0.0 scores by both metrics.

The only difference between these metrics is a constant 2 that multiplies TP in the numerator and in the denominator of DC fraction. This constant factor makes DC more balanced than JC for smaller fractions of FP and FN. The major drawback of DC is that this constant factor also reduces its FP sensibility, that is, the range of the output for larger fractions of FP. For instance, 0.95 and 0.94 DC coefficients may originate from very distinct results. Fig. 3(a) shows the difference between the metric outputs by fixing either FP or FN to 0 and changing the other one by the same amount. Fig. 3(b) presents the sensibility of the metrics to increasing FP fractions, fixing FN to 0. While DC is more balanced for a FP or FN fraction between 0.0 and 0.7 than JC (fig. 3(a)), it is less sensible to variations in FP (Fig. 3(b)).

3. IMPROVED EVALUATION PROCESS

We will discuss here a more balanced evaluation of tissue segmentation methods and give solutions to the issues described in Section 2.

With respect to the midbrain/hindbrain, we propose to present the accuracies with and without the midbrain and hindbrain. This way, one can verify this strategy's influence on the final accuracy of the segmentation. This procedure may be applied to other brain structures that are considered less important.

We propose a novel metric to handle external CSF where this is absent in ground-truth. This metric also improves the understanding of PVE influence. We name it Partial Volume Effect Evaluation (PVEE) and it may be employed with any volume based metric. The idea is to consider the label of a segmented voxel correct if it has the same tissue as its corresponding ground-truth voxel or of one of its 6-adjacent neighbors in the ground-truth. This is reasonable because partial volume effect occurs mostly between adjacent voxels with distinct tissues. Results should still be provided both with and without the PVEE, making it clear if a method outperforms others because of the PVE. Note that while checking the 6-adjacent neighbors, background voxels should be considered as CSF in datasets without external CSF, such as IBSR.

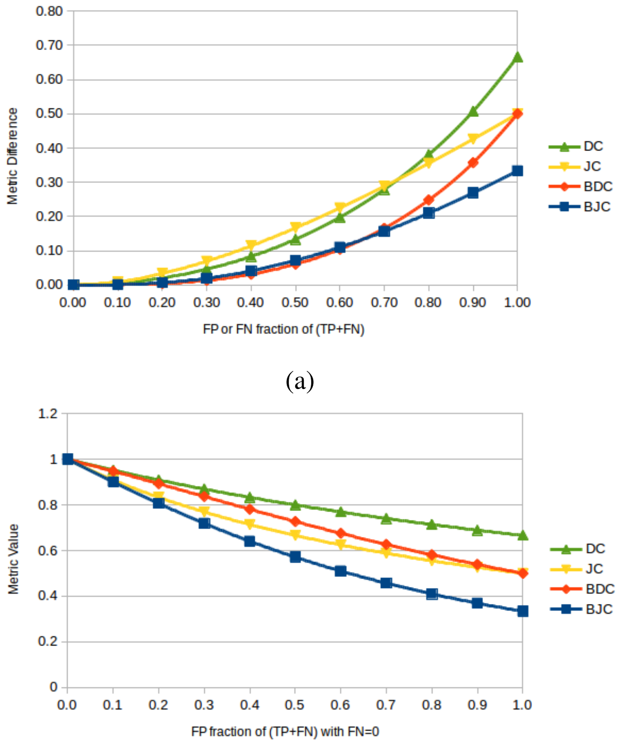
Finally, we propose a variation of DC and JC metrics named Balanced Dice(BDC) and Balanced Jaccard(BJC). They have the property of behaving more similarly to variations in FP and FN fractions. Also, they enhance the sensibility of the original metrics, especially for higher coefficients. The idea is to adapt the original formulas adding a regularizing term in a more distinct position than proposed by DC, as shown in Equations 3 and 4.

$$BDC = \frac{2TP}{2TP + FN + S * FP} \tag{3}$$

$$BJC = \frac{TP}{TP + FN + S * FP} \tag{4}$$

$$S = 1 + \frac{FP}{TP + FN}$$

The term S is automatically adjusted, depending on the FP and FN fraction with respect to TP. Therefore, there is no need of manual tunning of the parameters. The plots in Figure 3 shows that BDC and BJC are fairer than the original metric with respect to equal variations in FP and FN, that is, similar variations in FP and FN produce more similar accuracy values. BDC and BJC also have a wider output range for lower fractions of FN. Variations in FP have the same effect for the original and for the proposed metrics.



(a)

(b)

Fig. 3. (a) The difference between the output value given by DC, JC, BDC, and BJC metrics when methods mislabel the same fraction of voxels in FP or FN. Values close to 0 are better. (b) The output of metrics by setting FP to 0 and increasing fraction of FN. The closer to the anti-diagonal line the more balanced is the method.

Note that increasing the constant factor of DC metric would not provide the same result as the regularizing term

S. Increasing the constant would make the method more balanced with respect to FP and FN errors, but it would make the metric output even less sensitive to high accuracy values.

4. EXPERIMENTS

In this section, we present the results of the proposed evaluation methodology by employing the most popular tissues segmentation methods over the 18-subject IBSR dataset that is publicly available. The tested methods are PVC [10] from BrainSuite[2], FAST [23] from FSL[3], and OPF [19] from the Brain Image Analysis Library (BIAL)[4]. We used the Brain Extractor Tool from FSL for skull stripping and we corrected the inhomogeneity effect applying N4ITK [24]. An exception is FAST because it has its own inhomogeneity correction procedure running alongside the tissue segmentation. All methods were executed with default parameters.

Tables 1, 2, and 3 present the mean and standard deviation output value of DC, JC, BDC, and BJC for CSF, GM, and WM tissues, respectively. The tables contain the evaluation of these metrics over the whole brain (BRAIN) and over the brain without the midbrain and hindbrain, keeping the cerebellum (NO-STEM). We show the traditional (TRAD) and the PVEE analysis.

Notice an average increase of OPF accuracy in Table 3 of 1.49% to 2.35% in JC and BJC metrics when comparing BRAIN and NO-STEM values. This result suggests that OPF was penalized because it presented a higher quantity of FN WM voxels. The proposed methodology gave a better insight about the weakest point of OPF as compared to FAST and PCV.

Also, the supposed superiority of OPF over the others in labeling GM was confirmed by PVEE evaluation. While the accuracy of PVC and FAST increased from 10.04% to 14.16% from TRAD to PVEE, the accuracy of OPF increased from 9.88% to 17.62%. The opposite happened in terms of WM labeling. The maximal superiority of FAST and PVC over OPF decreased from 10.61% to 5.19%. In fact, OPF became more accurate than PVC according to BDC and BJC in PVCC and NO-STEM scenario. Meaningful metric values were generated for CSF segmentation using PVEE approach.

Finally, we can notice the good effect of the proposed Balanced metrics by comparing OPF with PVC and FAST in Tables 1. Using DC and JC, OPF outperformed PVC and FAST. As the balanced metrics are applied, the accuracy difference between OPF and FAST in BRAIN with TRAD fell from 10.85 to 5.30. This makes clear that OPF classified less CSF voxels in the cortical area, decreasing FP voxel fraction. Therefore, the supposed superiority of OPF in CSF classification is due to the absence of external CSF in IBSR dataset.

Table 1. Evaluation of PVC, FAST, and OPF over CSF voxels. (Mean|Standard deviation)

	BRAIN			NO-STEM		
TRAD	PVC	FAST	OPF	PVC	FAST	OPF
DC%	13.53|5.80	11.59|5.27	22.44|11.37	12.28|5.86	10.49|5.36	19.79|9.61
JC%	7.36|3.45	6.23|3.08	13.12|8.04	6.64|3.43	5.62|3.09	11.30|8.51
BDC%	1.54|1.54	1.07|1.21	6.37|8.72	1.28|1.41	0.90|1.11	4.49|5.84
BJC%	0.78|0.79	0.54|0.62	3.51|5.21	0.65|0.72	0.46|0.57	2.38|8.20
PVEE	PVC	FAST	OPF	PVC	FAST	OPF
DC%	60.67|7.17	53.43|7.13	80.22|7.08	59.87|7.31	52.58|7.45	78.87|2.79
JC%	43.91|7.44	36.77|6.91	67.54|10.20	43.10|7.51	36.01|7.14	65.84|4.87
BDC%	40.97|11.21	30.30|10.19	73.32|11.18	39.72|11.32	29.22|10.46	71.04|2.77
BJC%	26.36|9.05	18.29|7.58	59.08|14.54	25.38|9.03	17.55|7.68	56.74|4.83

Table 2. Evaluation of PVC, FAST, and OPF over GM voxels. (Mean|Standard deviation)

	BRAIN			NO-STEM		
TRAD	PVC	FAST	OPF	PVC	FAST	OPF
DC%	74.87|6.20	72.44|3.03	85.12|2.15	75.39|5.94	72.62|2.91	85.42|2.29
JC%	60.18|7.39	56.87|3.73	74.16|3.28	60.82|7.15	57.09|3.59	74.62|3.46
BDC%	74.78|6.16	72.26|3.13	84.18|2.29	75.32|5.91	72.47|3.00	84.67|2.48
BJC%	60.06|7.33	56.66|3.84	72.74|3.46	60.73|7.11	56.91|3.68	73.49|3.70
PVEE	PVC	FAST	OPF	PVC	FAST	OPF
DC%	84.99|5.42	82.76|2.32	95.01|1.51	85.43|5.17	82.87|2.32	95.32|1.56
JC%	74.23|7.61	70.66|3.40	90.52|2.72	74.88|7.32	70.82|3.40	91.10|2.81
BDC%	84.97|5.42	82.75|2.32	94.91|1.54	85.43|5.17	82.87|2.31	95.27|1.58
BJC%	74.21|7.60	70.64|3.40	90.36|2.78	74.87|7.32	70.81|3.39	91.01|2.85

Table 3. Evaluation of PVC, FAST, and OPF over WM voxels. (Mean|Standard deviation)

	BRAIN			NO-STEM		
TRAD	PVC	FAST	OPF	PVC	FAST	OPF
DC%	87.30|6.58	88.61|1.75	81.31|6.49	87.30|6.58	88.61|1.75	82.45|6.03
JC%	77.99|9.45	79.59|2.79	68.98|9.15	77.99|9.45	79.59|2.79	70.55|8.51
BDC%	84.82|10.75	87.69|2.59	81.16|6.33	84.82|10.75	87.69|2.59	82.25|5.84
BJC%	74.88|14.06	78.16|3.99	68.74|8.89	74.88|14.06	78.16|3.99	70.23|8.20
PVEE	PVC	FAST	OPF	PVC	FAST	OPF
DC%	95.26|3.70	96.45|1.09	93.50|3.43	95.26|3.70	96.45|1.09	94.85|2.79
JC%	91.16|6.37	93.16|2.02	87.97|5.94	91.16|6.37	93.16|2.02	90.32|4.87
BDC%	94.66|4.81	96.28|1.25	93.46|3.41	94.66|4.81	96.28|1.25	94.81|2.77
BJC%	90.21|8.05	92.85|2.29	87.91|5.90	90.21|8.05	92.85|2.29	90.24|4.83

5. CONCLUSION

We conducted an investigation on the evaluation of brain tissue segmentation methods. We identified flaws related to the procedure definition, partial volume effect, and evaluation metrics. Solutions to each of these issues were proposed, including a more balanced volume based metric. Experiments, provided a deeper understanding about real and apparent advantages of tested methods. Future works include a deeper analysis of the used metrics and datasets, experiments with other methodologies, and ground-truth generation issues.

[2] http://brainsuite.org/

[3] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

[4] https://github.com/GIBIS-UNIFESP/BIAL

6. REFERENCES

[1] S.M. Smith, M. Jenkinson, M.W. Woolrich, et al., "Advances in functional and structural mr image analysis and implementation as fsl," *Neuroimage*, vol. 23, pp. S208–S219, 2004.

[2] J. Ashburner and K.J. Friston, "Why voxel-based morphometry should be used," *Neuroimage*, vol. 14, no. 6, pp. 1238–1243, 2001.

[3] J. Acosta-Cabronero, G.B. Williams, G. Pengas, and P.J. Nestor, "Absolute diffusivities define the landscape of white matter degeneration in alzheimer's disease," *Brain*, vol. 133, no. 2, pp. 529–539, 2010.

[4] S. Bouix, M. Martin-Fernandez, L. Ungar, et al., "On evaluating brain tissue classifiers without a ground truth," *Neuroimage*, vol. 36, no. 4, pp. 1207–1224, 2007.

[5] N.C. Fox and P.A. Freeborough, "Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease," *Journal of Mag. Res. Imaging*, vol. 7, no. 6, pp. 1069–1075, 1997.

[6] Faria, F. A. and Cappabianco, F. A. M. and Li, C. R. and Ide, J. S., "Information Fusion for Cocaine Dependence Recognition using fMRI," in *23rd Int. Conference on Pattern Recognition*, 2016, (to appear).

[7] J. Acosta-Cabronero, G.B. Williams, J.M.S. Pereira, G. Pengas, and P.J. Nestor, "The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry," *Neuroimage*, vol. 39, no. 4, pp. 1654–1665, 2007.

[8] S. Valverde, A. Oliver, M. Cabezas, et al., "Comparison of 10 brain tissue segmentation methods using revisited ibsr annotations," *Journal of Magnetic Resonance Imaging*, vol. 41, no. 1, pp. 93–101, 2015.

[9] J.P. Chiverton and K. Wells, "Adaptive partial volume classification of MRI data," *Physics in Medicine and Biology*, vol. 53, no. 20, pp. 5577–5594, 2008.

[10] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg, and R.M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, 2001.

[11] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[12] C. Fennema-Notestine, I.B. Ozyurt, C.P. Clark, et al., "Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: Effects of diagnosis, bias correction, and slice location," *Human brain mapping*, vol. 27, no. 2, pp. 99–113, 2006.

[13] F.A.M. Cappabianco, P.A.V. de Miranda, J.S. Ide, et al., "Unraveling the Compromise Between Skull Stripping and Inhomogeneity Correction in 3T MR Images," in *SIBGRAPI*, 2012, pp. 1–8.

[14] P.A.V. Miranda, F.A.M. Cappabianco, and J.S. Ide, "A case analysis of the impact of prior center of gravity estimation over skull-stripping algorithms in MR images," in *IEEE International Conference on Image Processing*. IEEE, 2013, pp. 675–679.

[15] G. Hripcsak and D.F. Heitjan, "Measuring agreement in medical informatics reliability studies," *Journal of Biomedical Inform.*, vol. 35, no. 2, pp. 99–110, 2002.

[16] Trevor F Cox and Michael AA Cox, *Multidimensional scaling*, CRC Press, 2010.

[17] D.J. Goodenough, K. Rossmann, and L.B. Lusted, "Radiographic applications of receiver operating characteristic (ROC) curves," *Radiology*, vol. 110, no. 1, pp. 89–95, 1974.

[18] J.K. Udupa and Y. Zhuge, "Delineation operating characteristic (DOC) curve for assessing the accuracy behavior of image segmentation algorithms," in *Proceedings of SPIE*, 2004, vol. 5370, pp. 640–647.

[19] F.A.M. Cappabianco, A.X. Falcão, C.L. Yasuda, and J.K. Udupa, "Brain tissue MR-image segmentation via optimum-path forest clustering," *Computer Vision and Image Underst.*, vol. 116, no. 10, pp. 1047–1059, 2012.

[20] L.R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[21] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[22] S.P. Awate, T. Tasdizen, N. Foster, and R.T. Whitaker, "Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification," *Med. Image Analysis*, vol. 10, no. 5, pp. 726–39, 2006.

[23] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[24] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, et al., "N4ITK: improved N3 bias correction," *IEEE tran. on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.