Analyse exploratoire de métriques en segmentation d'images : application en proxidétection

Paul MELKI[1, 2], Lionel BOMBRUN[1], Estelle MILLET[2], Boubacar DIALLO[2],
Hakim ELCHAOUI ELGHOR[2], Jean-Pierre DA COSTA[1]

[1]Université de Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image, F-33405 Talence, France

[2]EXXACT Robotics, 1 rue Vincent Ballu, F-51200 Epernay, France

prenom.nom@ims-bordeaux.fr, prenom.nom@exxact-robotics.com

Résumé – Dans le contexte de la détection automatique de plantes sur des images de proxidétection agricole, nous étudions 12 métriques d'évaluation que nous utilisons pour évaluer 3 modèles de segmentation d'images récemment présentés dans la littérature. Après une présentation de ces métriques, nous explorons leurs relations à l'aide d'analyses multivariées. Nous distinguons 3 groupes de métriques fortement corrélées et, par une inspection visuelle d'images représentatives, nous identifions les aspects de la segmentation que chaque groupe évalue. L'objectif de cette analyse est de fournir quelques indices aux praticiens pour comprendre et choisir les métriques les plus pertinentes pour leur tâche agricole.

Abstract – Working in the context of automatic detection of plants in agricultural proximal sensing images, we study 12 evaluation metrics which we use to evaluate 3 image segmentation models recently presented in literature. After a unified presentation of these metrics, we carry out an exploratory analysis of their relationships using multivariate analyses. We distinguish 3 groups of highly correlated metrics and, through visual inspection of representative images, identify the aspects of segmentation that each group is evaluating. The aim of this exploratory analysis is to provide some clues to practitioners for understanding and choosing the metrics that are most relevant to their agricultural task.

1 Introduction

L'évaluation de la performance des modèles de *machine learning* (ML) est, d'un point de vue applicatif, l'étape la plus importante du pipeline prédictif car elle est souvent présentée comme une étape décisionnelle [1]. Se basant sur les valeurs des métriques d'évaluation, le praticien décide si un modèle est performant ou non. Une telle décision est souvent basée sur la "confiance" attribuée aux métriques et est fondée sur deux hypothèses implicites : les métriques reflètent la véritable performance du modèle et elles témoignent d'aspects de la performance pertinents pour l'application en question. Ces hypothèses reposent sur une compréhension des aspects théoriques et comportementaux des métriques utilisées, une condition qui n'est pas toujours satisfaite, en particulier pour les tâches de ML les plus complexes. Cela conduit les praticiens à adopter des métriques moins élaborées, dont les valeurs numériques sont plus faciles à interpréter et à expliquer [2]. Cela peut être problématique dans les applications industrielles, où les métriques offrant une évaluation bonne mais biaisée d'un modèle peuvent conduire les praticiens à l'appliquer dans des conditions réelles, pour découvrir qu'il est peu performant [2, 3]. Alors que d'autres sujets font l'objet de recherches actives dans la littérature du ML, comme la qualité des données [4] ou l'amélioration de modèles [5], l'étude des métriques utilisées pour évaluer ces pipelines prédictifs y a pris une place relativement minime. Certains chercheurs ont étudié ce sujet sous différents angles mais peu de travaux ont concerné le contexte particulier de la segmentation d'images [6], qui peut être considérée comme un cas particulier de classification de pixels, où chaque pixel est affecté à une classe représentant une entité sémantique. Ce travail propose une méthodologie statistique pour analyser les relations entre les métriques de classification au niveau des pixels dans le contexte de la segmentation d'images. Il fournit des interprétations concrètes de ces relations à la lumière des masques de segmentation produits par les modèles. La méthodologie est illustrée dans un cas d'étude relatif à la segmentation des plantes et du sol dans des images obtenues par proxidétection dans le contexte du désherbage automatique. Les principaux objectifs de ce travail peuvent être résumés comme suit :

— Proposer une approche statistique exploratoire facilement interprétable pour l'étude des relations entre métriques ;
— Réaliser une typologie de 12 métriques issues de la littérature afin d'aider les praticiens à identifier les plus utiles pour leur application.

2 Matériel et Méthodes

2.1 Jeu de données expérimental

Les 153 imagettes de taille 500×500 utilisées pour réaliser les expériences ont été découpées à partir d'images de taille de $2\ 940 \times 1\ 960$ acquises en conditions réelles sur de mul-

tiples parcelles agricoles à travers la France. Toutes les images sont annotées manuellement par des agronomes qualifiés, produisant des masques d'annotation binaires, que l'on appelle masques de vérité terrain (VT). Seules des imagettes contenant des plantes ont été préservées pour notre étude. La distribution spatiale des régions vertes et leurs tailles réelles sont en revanche très variables entre les images. L'ensemble de données est ensuite divisé aléatoirement en deux ensembles d'entraînement et de test, de 16 et 137 imagettes. La figure 2.1 montre un échantillon de 3 images et leurs masques d'annotation, représentant la diversité des conditions dans notre base de données.

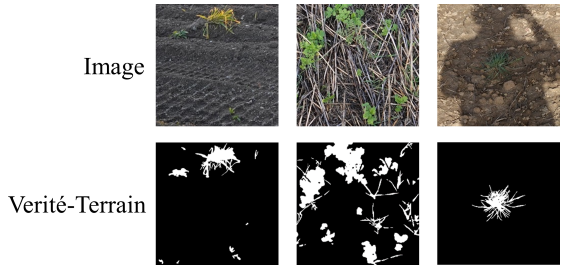

FIGURE 1 – Exemples d'images et de masques de vérité terrain.

2.2 Modèles de segmentation

L'étude et le développement de modèles n'étant pas l'objet du présent travail, nous avons testé trois modèles qui ont montré de bons résultats dans des contextes agricoles. Le premier, *Decision Tree Segmentation Model* (DTSM) [7] entraîne un arbre de décision sur des descripteurs obtenus en transformant les valeurs RVB des pixels dans 6 autres espaces de couleurs (*RVB*, *YCbCr*, *HSL*, *HSV*, *CIELab* et *CIELuv*). La deuxième approche propose de transformer les valeurs RVB des pixels en *CIELuv* puis d'entraîner un classifieur *Support Vector Machine* (SVM) sur les descripteurs obtenus [8]. Le troisième modèle, *Colour Index of Vegetation Extraction* (CIVE), n'est pas un modèle d'apprentissage mais plutôt propose de classifier les pixels entre *plante* et *sol* en appliquant un seuil sur une combinaison linéaire des valeurs RVB de chaque pixel [9]. Pour une comparaison équitable des modèles, ils ont tous été entraînés et testés sur les mêmes ensembles d'apprentissage et de test.

2.3 Métriques d'évaluation

Nous étudions 12 métriques d'évaluation qui peuvent toutes être formulées en fonction des composantes de base de la matrice de confusion : vrais positifs (TP), vrais négatifs (TN), faux positifs (FP), faux négatifs (FN). Par souci de concision, nous ne présentons pas la définition mathématique de chacune de ces métriques, mais renvoyons le lecteur aux références appropriées. Certaines sont des métriques de classification largement connues : (1) la précision (PRC), (2) le rappel (*recall*, RCL), (3) l'*accuracy* (ACC), et (4) la précision sur la classe négative (*negative predictive value*, NPV) [10] ; ainsi que des métriques de classification plus élaborées : (5) la variante équilibrée de l'*accuracy* (*balanced accuracy*, BAC),(6) le *F1-score* (F1S) [10] , et (7) l'indice *Kappa* de Cohen (KAP). Une autre métrique qui tient compte de l'aspect aléatoire est (8) l'*adjusted mutual information* (AMI), basée sur la théorie de l'information [11]. Une métrique largement utilisée pour la segmentation et la détection d'objets est (9) l'*intersection-over-union* (IOU) qui mesure le taux de superposition entre masques de vérité-terrain et de détection [12]. Spécialement conçue pour la segmentation binaire d'images, (10) la *global consistency error* (GCE) [13] est une mesure globale de l'erreur de segmentation entre deux masques de segmentation basée sur l'agrégation des erreurs de cohérence locales mesurées à chaque pixel, et (11) le *relative vegetation area* (RVA) est une métrique simple de recouvrement spatial qui compare les surfaces de végétation dans les masque de détection et de vérité-terrain [14]. Enfin (12) la distance de Hausdorff (HDD) mesure la similarité de forme entre les masques de segmentation [15].

2.4 Méthodologie d'exploration

Après avoir calculé les 12 métriques sur les imagettes de test pour les trois modèles étudiés, nous effectuons une analyse exploratoire afin d'étudier les relations entre ces métriques.

La première approche consiste à étudier la variabilité des métriques sur les images pour chaque modèle séparément. Pour cela, nous effectuons un clustering hiérarchique de variables [16] et une analyse en composantes principales (ACP) qui nous permettent d'identifier les relations entre les métriques. Afin de satisfaire aux conditions de validité de ces méthodes, une transformation en rangs est appliquée aux valeurs des métriques, ce qui permet de s'affranchir des particularités des distributions observées, très asymétriques.

Afin de vérifier que les résultats obtenus ne dépendent pas du modèle, nous avons appliqué une analyse factorielle multiple (AFM) [17] sur les classements produits par les trois modèles.

3 Résultats et discussion

3.1 Analyse mono-modèle

Dans un premier temps, une analyse mono-modèle est réalisée. Nous considérons ici les résultats de segmentation obtenus à l'aide de l'algorithme DTSM et évaluons le comportement des métriques pour ce modèle particulier.

Comme l'illustre le dendrogramme obtenu par le clustering de variables en Figure 3, les métriques étudiées peuvent être séparées en 3 groupes distincts. Le groupe *orange* est composé des métriques NPV, GCE et ACC, le groupe *vert* des métriques BAC, AMI, F1S, KAP, RVA, RCL et IOU et le groupe *bleu* de HDD et PRC. Afin de comprendre pourquoi ces métriques ont été regroupées en trois classes, une analyse en composantes principales est mise en oeuvre.

La figure 2 montre le biplot sur les 3 premières composantes principales de l'ACP qui expliquent 88% de l'inertie du jeu de données. Les groupes de métriques *vert*, *orange* et *bleu* sont

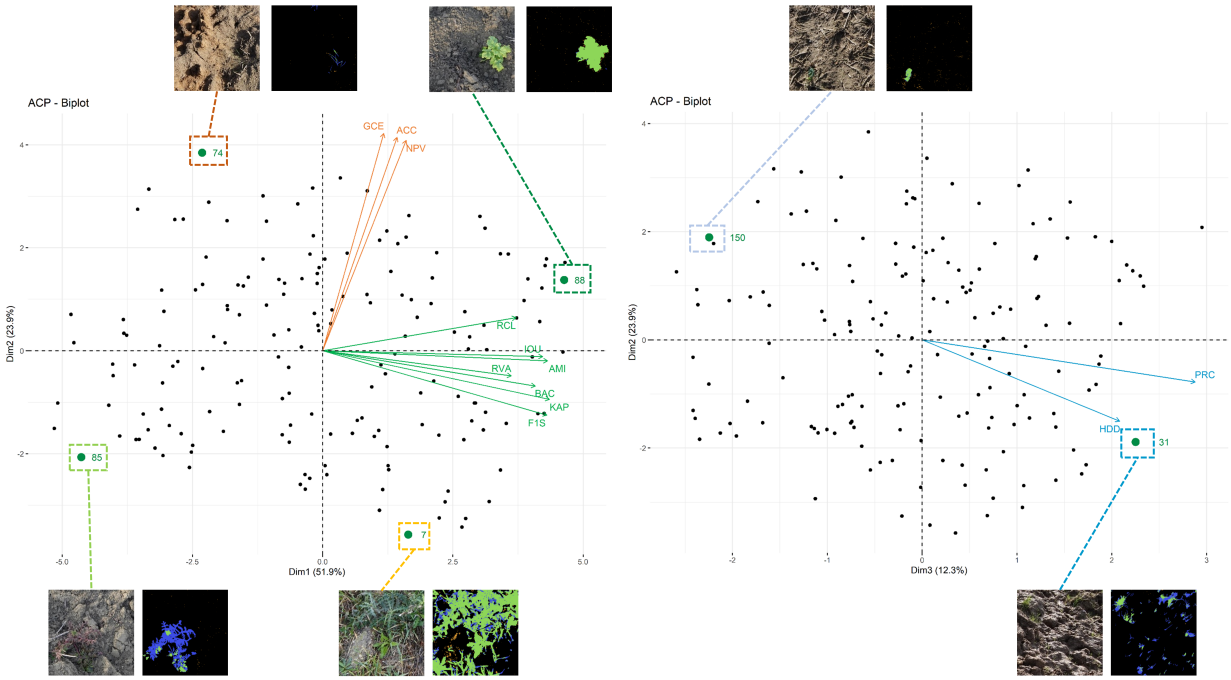

FIGURE 2 – Représentation d'images caractéristiques de chaque groupe de métriques sur les plans factoriels obtenus par ACP. A gauche : CP1 et CP2 ; à droite : CP2 et CP3. Le modèle de segmentation utilisé ici est DTSM.

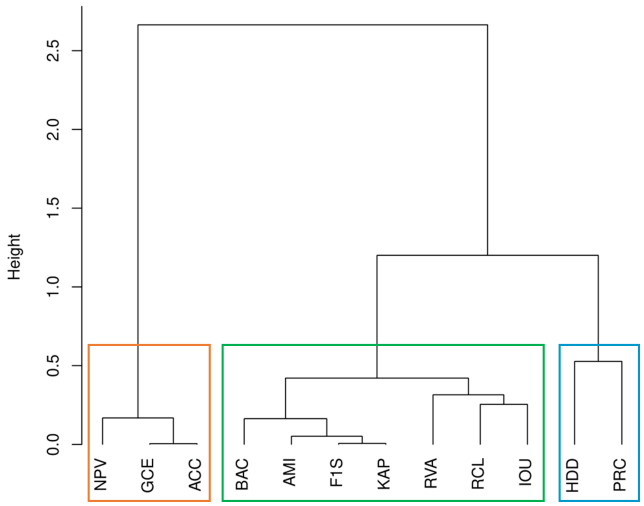

FIGURE 3 – Dendrogramme obtenu par la classification hiérarchique des 12 métriques sélectionnées, calculées après application de l'algorithme DTSM.

respectivement corrélés aux composantes principales 1, 2 et 3. Cette figure montre également quelques images représentatives de chaque groupe de métriques, avec la comparaison des vérités terrains et des segmentations produites (TP, FN et FP). Les images ayant une coordonnée positive sur la première composante principale sont celles qui sont bien évaluées par les métriques du groupe *vert* comme par exemple l'image n°88. Pour ces images, il n'y a quasiment aucune différence entre la vérité terrain et l'image segmentée. A contrario, les images très mal segmentées ont une coordonnée négative sur la première composante principale (image n°85). La deuxième composante principale caractérise les métriques du groupe *orange*. Ces métriques sont sensibles au déséquilibre entre les classes, c'est à dire dans notre contexte applicatif à une grande différence entre la proportion de pixels de plantes et de sol. Les images bien évaluées par ces métriques sont celles qui contiennent plutôt des objets de petites tailles comme l'image n°74 alors que celles qui sont moins bien évaluées ont un meilleur équilibre entre les 2 classes (image n°7). Les métriques du groupe *orange* se montrent donc plus optimistes lorsqu'on est en présence d'un déséquilibre entre les classes. Enfin les métriques du groupe *bleu* contribuent à la troisième composante principale. Les images bien évaluées par ces métriques ont un taux de FP faible (image n°31) tandis que celles qui sont mal évaluées ont un taux de fausses détections relativement élevé (image n°150), distribuées de manière aléatoire dans l'image.

3.2 Analyse multi-modèles

Afin de vérifier si les relations décrites en partie 3.1 sont spécifiques à l'algorithme DTSM ou si elles peuvent être étendues à d'autres approches, une analyse factorielle multiple (AFM) est mise en oeuvre en considérant les trois modèles de segmentation décrits en 2.2. Le tracé des axes partiels de l'AFM (figure 4) montre les projections des composantes 1 à 3 de l'ACP de chaque modèle sur les dimensions 1 à 3 de l'AFM. Les composantes des trois ACP apparaissent très étroitement corrélées, ce qui suggère que les ACP appliquées à chaque modèle séparément ont des structures très similaires. Cela confirme l'hypothèse de départ selon laquelle les observations réalisées sur les métriques et sur leurs relations ne semblent pas dépendre du modèle de classification et peuvent être généralisées.

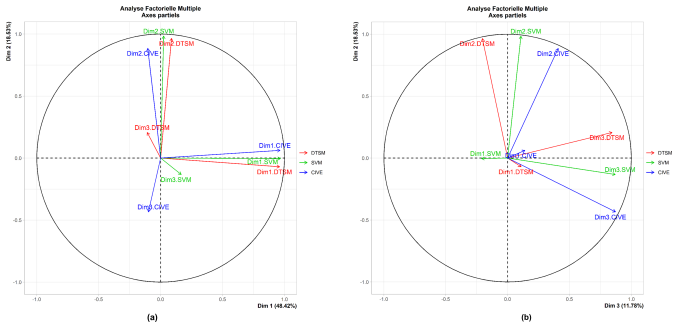

FIGURE 4 – Axes partiels de l'AFM : **(a)** dimensions 1 et 2, **(b)** dimensions 3 et 2. En rouge, vert et bleu : les 3 premières composantes principales des ACPs réalisées après application des algorithmes DTSM, SVM et CIVE.

4 Conclusions

Dans cet article, nous avons présenté et analysé 12 métriques dans le contexte de la segmentation d'images de plantes en proxidétection agricole. L'analyse exploratoire nous a permis d'identifier trois groupes distincts de métriques. Le premier et plus grand groupe (IOU, BAC, F1S, RCL, AMI et KAP) est performant quelle que soit la structure intrinsèque de l'image et est indépendant du déséquilibre entre les classes. Ces métriques peuvent être utilisées pour des tâches de segmentation à usage général, où la segmentation exacte avec des contours bien délimités n'est pas de la plus haute importance. Les praticiens peuvent adopter certaines métriques de ce groupe, qui mesurent différemment la qualité : par exemple l'IOU pour le recouvrement, l'indice Kappa de Cohen pour tenir compte du caractère aléatoire et l'*adjusted mutual information* pour une mesure basée sur la théorie de l'information. Le deuxième groupe (ACC, NPV et GCE) s'est révélé être très sensible au déséquilibre des classes. En effet, l'*accuracy*, qui est une métrique largement utilisée en raison de sa facilité de calcul et d'interprétation, ne devrait être utilisée que dans les cas où les classes sont à peu près équilibrées. Sinon elle aura tendance à être biaisée en faveur de la classe majoritaire. Il en va de même pour NPV et GCE. Le troisième groupe de métriques identifiée est constitué de la précision et de la distance de Hausdorff. Ces métriques se sont avérées très sensibles à la sur-segmentation des images, en particulier lorsque les fausse détections sont uniformément réparties dans l'image. Enfin, les résultats de l'analyse factorielle multiple pour trois algorithmes de segmentation de l'état de l'art (DTSM, SVM, CIVE) ont montré que le comportement et les relations entre les métriques semblent indépendants de l'algorithme de segmentation considéré.

Références

[1] S. L. Salzberg, "On comparing classifiers : A critique of current research and methods," *Data Min Knowl Discov*, 1999.

[2] A. Zheng, *Evaluating Machine Learning Models*. O'Reilly Media, Inc., 2015.

[3] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC : A family of discriminant measures for performance evaluation," in *Advances in Artificial Intelligence*, 2006, vol. 4304, pp. 1015–1021.

[4] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning : Going beyond data cleaning and transformations," *Int. j. adv. software*, vol. 10, pp. 1–20, 07 2017.

[5] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision : A brief review," *Comput. Intell. Neurosci.*, 2018.

[6] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation : analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, Dec. 2015.

[7] W. Guo, U. K. Rage, and S. Ninomiya, "Illumination invariant segmentation of vegetation for time series wheat images based on decision tree model," *Comput Electron Agric*, vol. 96, pp. 58–66, Aug. 2013.

[8] M. Rico-Fernández, R. Rios-Cabrera, M. Castelán, H.-I. Guerrero-Reyes, and A. Juarez-Maldonado, "A contextualized approach for segmentation of foliage in different crop species," *Comput Electron Agric*, vol. 156, pp. 378–386, Jan. 2019.

[9] T. Kataoka, T. Kaneko, H. Okamoto, and S. Hata, "Crop growth estimation system using machine vision," in *IEEE AIM*, vol. 2, 2003, pp. b1079–b1083 vol.2.

[10] J. Lever, M. Krzywinski, and N. Altman, "Classification evaluation," *Nature Methods*, vol. 13, no. 8, pp. 603–604, 2016.

[11] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, p. 2837–2854, Dec. 2010.

[12] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[13] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE ICCV*, vol. 2. Vancouver, BC, Canada : IEEE Comput. Soc, 2001, pp. 416–423.

[14] H. K. Suh, J. W. Hofstee, and E. J. van Henten, "Improved vegetation segmentation with ground shadow removal using an HDR camera," vol. 19, no. 2, pp. 218–237, 2018.

[15] A. A. Taha and A. Hanbury, "An efficient algorithm for calculating the exact hausdorff distance," *IEEE TPAMI*, vol. 37, pp. 2153–2163, 2015.

[16] M. Chavent, V. Kuentz-Simonet, B. Liquet, and J. Saracco, "ClustOfVar : An R Package for the Clustering of Variables," *Journal of Statistical Software*, vol. 50, no. 13, 2012.

[17] B. Escofier and J. Pagès, *Analyses factorielles simples et multiples : objectifs, méthodes et interpétation*. Paris : Dunod, 2008, oCLC : 519685951.