# ELSCP pour la détection d'anomalies dans les flux de données issues de l'agriculture

Juliet Chebet Moso, Stephane Cormier, Cyril de Runz, Hacene Fouchal, John Mwangi Wandeto

# ELSCP pour la détection d'anomalies dans les flux de données issues de l'agriculture

Juliet Chebet Moso*,*** Stéphane Cormier*
Cyril de Runz** Hacène Fouchal* John Mwangi Wandeto***

*CReSTIC EA 3804, Université de Reims Champagne-Ardenne, 51097 Reims, France
juliet-chebet.moso@etudiant.univ-reims.fr, stephane.cormier@univ-reims.fr,
hacene.fouchal@univ-reims.fr
**BDTLN, LIFAT, University of Tours,place jean Jaurès, 41000, Blois, France
cyril.derunz@univ-tours.fr
***Computer Science, Dedan Kimathi University of Technology, Private Bag -10143,
Dedan Kimathi, Nyeri, Kenya
john.wandeto@dkut.ac.ke

## 1 Introduction

Cet article est une version poster de (Moso et al., 2021). L'agriculture bénéficie des développements récents de la technologie des capteurs, de la science des données et des techniques d'apprentissage automatique. Le but de cet article est de trouver des anomalies qui ont une influence sur l'efficacité de la récolte et peuvent être liées à la santé et à l'état des cultures pendant la récolte. En raison de la spécialisation des modèles pour différencier des caractéristiques des données, il est pertinent de combiner les diverses capacités des différentes approches de détection pour générer une décision de consensus (Zimek et al., 2014).

Nous considérons une approche orientée données avec l'objectif de détecter des anomalies à la volée en utilisant des approches de détection non supervisées pour la détection d'anomalies contextuelles locales. Nous proposons un détecteur d'anomalies par ensemble appelé Enhanced Locally Selective Combination in Parallel outlier ensembles (ELSCP). ELSCP est adapté au contexte du streaming en utilisant un système qui convertit les données en un flux et les transmet à ELSCP en utilisant un modèle de fenêtre de référence qui met en œuvre une technique de fenêtre glissante. Cette adaptation permet le traitement des données sous forme de flux, ce qui facilite l'évaluation de notre algorithme dans le contexte du streaming.

La méthode *Locally Selective Combination in Parallel outlier ensembles (LSCP)* (Zhao et al., 2019) construit une petite zone autour d'une instance de test en utilisant le consensus de ses voisins les plus proches. Nous proposons une extension de LSCP nommée ELSCP qui améliore la manière dont la région locale est extraite et la sélection de détecteurs adaptés. Dans ELSCP, pour chaque instance de test, la définition de la région locale est effectuée à l'aide d'une méthode de type ball tree k-plus proche voisin avec une métrique basée sur la distance de Harvesine. Une pseudo vérité terrain locale est construite à l'aide de la région locale, et la corrélation de Kendall est calculée entre les scores de valeurs aberrantes d'apprentissage

de chaque détecteur de base et la pseudo vérité terrain. En utilisant les scores de corrélation, le détecteur ayant les plus grandes corrélations est sélectionné. Le score final est calculé en moyennant les scores du détecteur sélectionné. ELSCP donne la priorité à la sélection des détecteurs en fonction des compétences locales ce qui facilite l'identification des détecteurs de base avec un biais de modèle conditionnellement faible. Il applique trois détecteurs de base : HBOS (Goldstein et Dengel, 2012), MCD (Rousseeuw et Driessen, 1999) et Isolation Forest (Liu et al., 2008).

Nous étudions la détection d'anomalies pour l'agriculture intelligente sur deux axes : la détection des dommages causés aux cultures pendant la récolte, et l'utilisation efficace des moisson-neuses-batteuses. Pour chaque cas, nous utilisons un ensemble de données pertinentes pour identifier les anomalies. Nous proposons des techniques basées sur des tests d'hypothèses sur l'occurrence d'un modèle de mouvement anormal pendant la récolte dans le champ.

Dans l'ensemble de données sur les cultures, notre analyse a montré que 30% des anomalies détectées pouvaient être directement liées aux dommages causés aux cultures. Dans le jeu de données sur les moissonneuses-batteuses, notre approche obtient les meilleurs scores avec une AUC-ROC supérieure de 6,4% à celle de la deuxième approche COPOD (99,8% contre 93,4%) ; une AUCPR de 97,2% alors que les meilleurs scores des autres approches (celui de l'OCSVM) ne sont que de 38,5%. Il convient également de mentionner que notre méthodologie permet de détecter efficacement les comportements déviants des moissonneuses-batteuses. Par conséquent, la détection des anomalies pourrait être intégrée dans le processus de décision des exploitants agricoles afin d'améliorer l'efficacité de la récolte et la santé des cultures.

## Références

Goldstein, M. et A. Dengel (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, 59–63.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation forest. In *Eighth IEEE International Conference on Data Mining ICDM 2008*, pp. 413–422. IEEE.

Moso, J. C., S. Cormier, C. de Runz, H. Fouchal, et J. M. Wandeto (2021). Anomaly detection on data streams for smart agriculture. *Agriculture 11*(11), 1083.

Rousseeuw, P. J. et K. V. Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*(3), 212–223.

Zhao, Y., Z. Nasrullah, M. K. Hryniewicki, et Z. Li (2019). Lscp: Locally selective combination in parallel outlier ensembles. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 585–593. SIAM.

Zimek, A., R. J. Campello, et J. Sander (2014). Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *Acm Sigkdd Explorations Newsletter 15*(1), 11–22.

## Summary

This paper is a poster version of (Moso et al., 2021).