startbox : un package R pour l'analyse et la visualisation des données d'expérimentation en protection de la vigne

Xavier Delpuech 1* Hervé Maire 2† Anne-Sophie Chazalmartin 3‡

Résumé

Le package R `startbox` est dédié à la gestion et à l'analyse des données d'expérimentation en protection phytosanitaire, en particulier pour la vigne. Ce package a été développé par l'Institut Français de la Vigne et du Vin dans le cadre du projet CASDAR STAR (Amardeilh et al. 2025). L'originalité du package est de s'appuyer sur un fichier Excel modèle permettant de standardiser les données produites, en s'appuyant sur un modèle de données et des vocabulaires contrôlés. Il permet de préparer les données (aggrégation, filtrage), de réaliser les statistiques de base et de visualiser les résultats sous forme de graphiques.
Mots-clefs : Statistique - Protection des plantes - Data - ontologie - Package R

Développement

Le développement de `startbox` s'inscrit dans un contexte où l'innovation agricole, notamment pour les biosolutions (biocontrôle, biostimulants), est ralentie par le cloisonnement des informations. Les données expérimentales sont produites de manière massive mais restent souvent stockées en « silos » : chaque projet possède son propre format, sa propre nomenclature et sa propre structure. Cette hétérogénéité génère un déficit d'interopérabilité entre jeux de données, et de fait il est complexe et coûteux de compiler des données provenant de partenaires différents pour réaliser des méta-analyses nécessaires à la génération de connaissances robustes. D'autre part, la qualité même des jeux de données est très variable, en particulier à cause du manque de métadonnées (description du matériel et méthodes, du contexte de production agricole). En conséquence, les données d'expérimentation sont difficilement réutilisables. `startbox` répond à ces enjeux en proposant un cadre structuré qui permet de passer d'une gestion spécifique des données à un processus standardisé et automatisé.

*Institut Français de la Vigne et du Vin, xavier.delpuech@vignevin.com
†Institut Français de la Vigne et du Vin, herve.maire@vignevin.com
‡Institut Français de la Vigne et du Vin, anne-sophie.chazalmartin@vignevin.com

Le principe de `startbox` repose sur la centralisation du flux de travail au sein de l'environnement R, en s'appuyant sur un fichier Excel modèle comme pivot de la structuration. L'idée est de fournir aux expérimentateurs des outils simples et familiers pour structurer leurs données tout en garantissant la rigueur nécessaire à l'exploitation informatique et à la reproductibilité des analyses. Ainsi, le fichier Excel modèle impose une architecture de données pour garantir l'interopérabilité. On y retrouve les informations sur :

- Le contexte général : Informations sur l'expérimentation et ses objectifs.
- Le dispositif expérimental : Description du sites, du plan d'expérience et des traitements expérimentaux.
- Le contexte technique : Détails de l'itinéraire technique (dates d'application, types de pulvérisateurs, doses).
- Les données collectées : Résultats des observations sur les bioagresseurs (intensité d'attaque, fréquence, etc.) données météorologiques.

Pour éviter les ambiguïtés (par exemple, nommer une maladie de trois façons différentes), le modèle Excel s'appuie sur des vocabulaires contrôlés. Dans le cadre viticole, il intègre des référentiels telles que la Vitis Ontology (Duchêne and Pommier 2019). Cette standardisation permet au package `startbox` de reconnaître instantanément les variables et d'automatiser une partie des fonctions. `startbox` propose un ensemble de fonctions couvrant l'intégralité du cycle de vie d'un jeu de données expérimentales :

1. Chargement des données : `startbox` est conçu pour interagir nativement avec le fichier Excel standard, en s'appuyant sur le packahe openxlsx2 (Barbone and Garbuszus 2025)
2. Préparation et nettoyage des données : `startbox` inclut des fonctions pour le filtrage des modalités, l'agrégation des répétitions et la gestion des données manquantes, afin de transformer les données brutes en jeux de données prêts pour l'analyse.
3. Analyses statistiques : `startbox` facilite la réalisation de tests statistiques courants (Anova, test non paramétrique de Kruskal-Wallis) pour comparer les traitements expérimentaux.
4. Visualisation graphique avancée : `startbox` propose des fonctions de « data-visualisation » automatisées pour produire des graphiques, tels que des barplots (Figure 1) ou des heatmaps, essentiels pour l'exploration visuelle des résultats. Ces fonctions s'appuient sur le package ggplot2 (Wickham 2016).
5. Export des données préparées et des résultats statistiques vers Excel.

De futures fonctionnalités de validation de données pourront être implémentées, ainsi que des connecteurs avec d'autres outils et systèmes d'information pour renforcer l'interopérabilité du fichier modèle. Un travail d'alignement avec l'ontologie Experiment on Living Organisms in Agriculture (ELOA) est en cours (Roussey, Tireau, and Neveu 2025).

En conclusion, startbox n'est pas seulement un outil statistique ; c'est un levier de transformation qui permet aux expérimentateurs de s'approprier les principes de la science ouverte tout en simplifiant leur travail quotidien de gestion des données

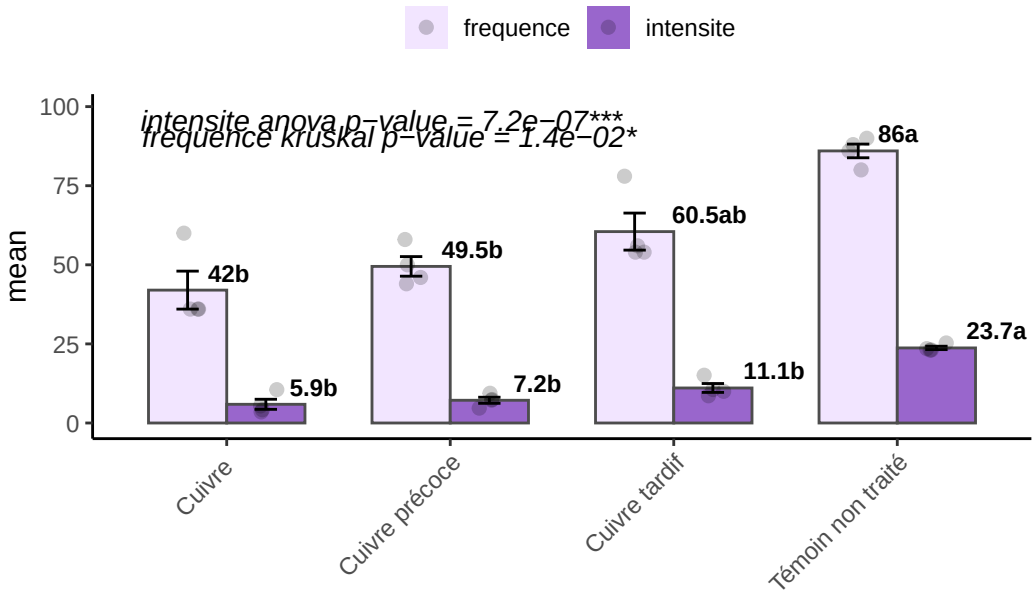

Figure 1: Exemple de barplot généré avec le package startbox

Remerciements

Ce package R a été développé dans le cadre du projet STAR 2024-2027 (France), avec le soutien du ministère de l'Agriculture et de l'Alimentation et la contribution financière du Compte d'affectation spéciale pour le développement agricole et rural (CASDAR).

Références

Amardeilh, Florence, Arnaud Charleroy, Baptiste Darnala, Xavier Delpuech, Catherine Roussey, and Frédéric Salvi. 2025. "Standardiser Les Données Expérimentales Pour Faciliter l'innovation Dans Le Domaine Des Bio Solutions : Le Projet CASDAR STAR." In *8ème Édition de l'Atelier INtégration de Sources/Masses de Données Hétérogènes Et Ontologies, Dans Le Domaine Des Sciences Du VIVant Et de l'Environnement (IN-OVIVE), Adossé à La Conférence Ingénierie Des Connaissances @PFIA*, 5p. Dijon, France: AFIA-Association Française d'Intelligence Artificielle. https://hal.inrae.fr/hal-05074927.

Barbone, Jordan Mark, and Jan Marvin Garbuszus. 2025. *Openxlsx2: Read, Write and Edit 'Xlsx' Files*. https://janmarvin.github.io/openxlsx2/.

Duchêne, Eric, and Cyril Pommier. 2019. "The Vitis Ontology: Sustainable and FAIR (Findable, Accessible, Interoperable, Reusable) for Consistent and Complete Data Description Through Biologist Friendly Ontologies." In. Chania, Greece, 26-28 March 2019, Greece. https://hal.inrae.fr/hal-02947459.

Roussey, Catherine, Anne Tireau, and Pascal Neveu. 2025. "Méthode d'adaptation d'une ontologie d'application : cas des expérimentations agronomiques." In *Actes des conférences hébergées à PFIA 2025, 36es journées francophones d'Ingénierie des Connaissances (IC)*, edited by Fleur Mougin, 73–79. Actes - 36es Journées Francophones d'ingénierie Des Connaissances. Dijon, France: AFIA- Association Française d'Intelligence Artificielle; AFIA-Association Française d'Intelligence Artificielle. https://hal.inrae.fr/hal-05142985.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.