Le package {ProduceR} : faciliter et fiabiliser la production statistique

Vincent Reduron*

Résumé

Le package {ProduceR} propose un ensemble de 5 fonctions conçues pour répondre aux besoins courants pour la production statistique, terme désignant la production de données statistiques fiables et exploitables à partir d'informations brutes. Elles ont été élaborées à partir de l'expérience concrète de travail de son auteur, qui travaille depuis 20 ans dans des services statistiques.

La philosophie sous-jacente est la suivante. La production statistique implique généralement de manipuler des données complexes, notamment quand les informations brutes sont des données administratives (Lefebvre, Soulier et Tortosa, 2024), mais aussi quand ce sont des données d'enquête. Ces informations se trouvent presque systématiquement dans un ensemble de *tables*, car la représentation tabulaire est le modèle canonique pour la statistique (Dondon et Lamarche, 2023). Une production de qualité implique de comprendre la structure de ces tables et leur contenu, mais aussi de maîtriser les relations entre elles. Or, dans un quotidien de travail, on hésite à réaliser des vérifications si elles sont longues à programmer ; une exploration manuelle remplace souvent une vérification exhaustive. Pourtant, il est important de contrôler les données à différentes étapes de la production : par exemple, la jointure entre deux tables doit s'accompagner au préalable de la vérification de leurs clefs uniques, puis du contrôle de l'absence de doublons ou de valeurs manquantes non souhaités.

Ainsi, les cinq fonctions de {ProduceR} ont été développées pour être concises et abordables, pour que le producteur n'hésite pas à beaucoup vérifier. Elles sont volontairement peu nombreuses et ciblées sur les principaux besoins rencontrés en vie réelle (analyse des doublons et des valeurs manquantes, étude synthétique des variables, comparaison de tables).

Le package est disponible sur le CRAN à l'url https://CRAN.R-project.org/package=ProduceR.

Mots-clefs : Statistique – Data – Package

Développement

Le package inclut les fonctions de base suivantes :

- `dup()` : vérifie les *duplicates*.

  Aide à s'assurer qu'une ou plusieurs colonnes constituent une clef unique pour la table. Comprendre la structure des tables revient en général à comprendre *ce qui définit une ligne*. La question de la clef unique des tables est centrale, notamment pour faire des jointures correctes entre différentes tables.

* Ministère de la santé et des solidarités, Direction de la recherche, des études, de l'évaluation et des statistiques (Drees), vincent.reduron@sante.gouv.fr

Exemple : `verif <- dup(base_eu_2025, c("annee", "mois")) # vérifie si la combinaison des colonnes annee et mois est une clef unique pour la table.`

- `miss()` : vérifie les *missing values.*

  Compte les valeurs manquantes pour toutes les colonnes de la table en entrée. Maîtriser l'information contenue dans les tables implique *a minima* de connaître leur taux de bon remplissage. Par ailleurs, les valeurs manquantes sont un sujet particulièrement compliqué à gérer, appelant toujours à des choix méthodologiques.

  Exemple : `verif <- miss(base_eu_2025)`

- `tac()` : tableau de contingence général.

  Compile, dans un tableau unique en sortie, des tableaux de contingence pour l'ensemble des colonnes de la table en entrée. Les valeurs numériques sont synthétisées en 4 modalités (valeur positive, négative, nulle ou manquante).

  Exemple : `verif <- tac(base_eu_2025)`

{ProduceR} inclut aussi les fonctions avancées suivantes :

- `toc()` : compare deux tables supposées proches (par exemple la « même table » sur deux années) et repère les écarts significatifs, qui peuvent être signe d'erreurs.

  La significativité des écarts est jugée selon un critère composite mêlant écart absolu et écart relatif.

  Exemple : `compar_ans <- toc(base_eu_2024, base_eu_2025 %>% filter(Année != 2025)) # écart significatif sur la modalité "Catalogne" (disparition dans la base 2024 par rapport à la base 2025)`

- `chi2_find()` : dans la table en entrée, aide à caractériser des données problématiques (valeur manquantes, aberrantes, …) en suggérant des corrélations avec le reste des données. Utile dans les très grosses tables dont l'exploration manuelle est labyrinthique.

  Le sous-jacent est la réalisation de tests du χ^2 pour juger de la corrélation entre les données problématiques et l'ensemble des modalités de l'ensemble des colonnes de la table. Une très forte corrélation avec une modalité indique une possibilité d'origine de la problématique sur les données.

  Exemple : `detect <- chi2_find(base_eu_2025, criterion = "is.na(PIB)") # Correlation à 100% avec Pays == "Espagne" (les lignes de table avec PIB manquant sont exactement celles avec Pays == "Espagne")`

Références

Lefebvre, O., Soulier, M., Tortosa, T. 2024. « L'accueil des données administratives : un processus structurant ». Courrier des statistiques. Insee. https://www.insee.fr/fr/information/8203046

Dondon, A., Lamarche, P. 2023. « Quels formats pour quelles données ? ». Courrier des statistiques. Insee. https://www.insee.fr/fr/information/7635827