Estimation paramétrique de ruptures dans des données censurées à gauche

Clément Laroche, Madalina Olteanu, Fabrice Rossi

▶ **To cite this version:**

Clément Laroche, Madalina Olteanu, Fabrice Rossi. Estimation paramétrique de ruptures dans des données censurées à gauche. JDS 2021 : 52èmes Journées de Statistique de la Société Française de Statistique (SFdS), Jun 2021, Nice, France. hal-03912210

HAL Id: hal-03912210
https://hal.science/hal-03912210v1

Submitted on 23 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation paramétrique de ruptures dans des données censurées à gauche

Clément Laroche [1] & Madalina Olteanu [2] & Fabrice Rossi [3]

[1] *Université Paris I- SAMM, Clement.Laroche@univ-paris1.fr*
[2] *Université Paris-Dauphine - Ceremade, Madalina.Olteanu@dauphine.psl.eu*
[3] *Université Paris-Dauphine - Ceremade, Fabrice.Rossi@dauphine.psl.eu*

Résumé. La phytopharmacovigilance a pour objectif de surveiller les effets indésirables des produits phytopharmaceutiques disponibles sur le marché et couvre notamment la contamination des milieux. Cette surveillance s'exerce notamment en recherchant des anomalies ou des ruptures dans les concentrations de substances actives de ces produits. La mesure d'une concentration chimique dépendant de la précision la machine de mesure, les données observées sont soumises à une censure à gauche. Nous proposons ici une approche paramétrique permettant, en présence de données censurées et à seuil de censure connu, d'estimer le nombre et les positions des changements dans la concentration d'une substance. Les performances de cette approche sont comparées à celles d'une méthode de détection non-paramétrique, notamment sur des données réelles de concentrations du prosulfocarb en région Val de Loire.

Mots-clés. Détection de ruptures, censure à gauche, programmation dynamique, phytopharmacovigilance

Abstract. Pesticide effects are routinely monitored to detect potential health hazards. Pesticide residues in the environment are of particular interest. A natural monitoring strategy consists in looking for anomalies or change points in the concentration of active substances in a given environment. Due to the quantification limit of most chemical analyses, concentration observations are left censored which introduces some difficulties. We propose in this paper a method to detect change points, both in number and positions, in left censored concentration data that follow a parametric distribution (knowing the censoring threshold). The method is compared to a reference non parametric method on simulated and real world data (prosulfocarb concentration in Val de Loire).

Keywords. Changepoint detection, left censorship, PELT, pesticide monitoring.

1 Introduction

Le suivi de concentration de pesticides est un enjeu majeur des agences de sécurité environnementale et de santé publique. En France, il est rendu possible par la collecte de

mesures de concentration pour un grand nombre de substances, dans différents environnements, sur l'ensemble du territoire. Ces données, aujourd'hui publiques, nécessitent des méthodes d'analyse prenant en compte leurs spécificités [6]. Parmi celles-ci, on peut noter une grande hétérogénéité de collecte avec un rythme de relève irrégulier et une densité de collecte très variable en fonction de l'échelle géographique considérée.

De plus, les données de concentration présentent un phénomène de censure à gauche induit par les limites de précision des techniques d'analyse. Lorsque la concentration dans le prélèvement est trop faible, on observe la *limite de quantification* (LQ) de la technique employée et pas la véritable concentration sous-jacente.

Pour suivre les concentrations des substances actives, nous proposons de rechercher des ruptures dans ces valeurs, en adaptant les méthodes classiques [7] au cas de la censure à gauche. Il s'agit de fournir aux analystes des plages de concentration homogène et d'attirer l'attention sur des changements brutaux. Nous nous plaçons en outre dans un cadre paramétrique car les concentrations présentent souvent des distributions de type exponentiel.

La suite de l'article est organisée de la manière suivante. Nous présentons en section 2 le modèle de détection de ruptures ainsi que la procédure d'estimation des paramètres. La section 3 est consacrée à des expériences sur des données simulées. Nous concluons l'article par une section dédiée à une application sur des données réelles.

2 Détection de ruptures dans des données censurées à gauche

On suppose dans la suite que l'on dispose d'une série de réalisations $y_1, ..., y_n$ des variables aléatoires indépendantes $Y_1, ..., Y_n$. On notera par ailleurs $Y_{a:b} = (Y_a, ..., Y_b)$.

2.1 Modélisation

On suppose que les Y_i suivent des lois exponentielles censurées à gauche avec un seuil de censure commun, a. En pratique, ce seuil correspondant à la limite de quantification de l'analyse, est fixé et connu a priori. Les intensités des lois sont supposées constantes par morceaux. Plus précisément, on suppose qu'il existe K^* ruptures associées à $K^\star + 1$ intervalles définis par les instants $\boldsymbol{t}^\star = (0 = t_0^\star < t_1^\star < ... < t_{K^\star-1}^\star < t_{K^\star}^\star < t_{K^\star+1}^\star = n)$. Sur l'intervalle $]t_k^\star, t_{k+1}^\star]$, les Y_i concernées ont une intensité $\lambda_k^\star$.

On cherche à déterminer $K^\star$, $\boldsymbol{t}^\star$ et $\boldsymbol{\lambda}^\star = (\lambda_0^\star, ..., \lambda_{K^\star}^\star)$ à partir des observations. Pour ce faire, on utilise une approche classique de vraisemblance pénalisée, ce qui conduit au critère suivant

$$\tilde{\mathcal{C}}_{Y_{1:n}}(K, \boldsymbol{t}, \boldsymbol{\lambda}) = \sum_{k=0}^{K} W(Y_{(t_k+1):t_{k+1}}, \lambda_k) - \beta_n K \ , \tag{1}$$

où β_n désigne un terme de pénalité associé au rajout d'une nouvelle rupture, et $W(Y_{(t_k+1):t_{k+1}}) = \sum_{i=t_k+1}^{t_{k+1}} \ln f_{\lambda_k}(Y_i)$ la log-vraisemblance du segment $Y_{(t_k+1):t_{k+1}}$.

Le terme β_n peut être choisi de manière à obtenir les pénalités usuelles type AIC ou BIC, ou via des méthodes de calibration non-asymptotiques comme l'heuristique de pente [1].

L'estimateur de maximum de vraisemblance pénalisée est

$$(\hat{K}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{\lambda}}) = \arg \max_{K=1,\dots,K_{\max}, \boldsymbol{t}\in\mathcal{T}_K^{\Delta}, \boldsymbol{\lambda}\in\mathbb{R}_+^K} \tilde{\mathcal{C}}_{Y_{1:n}}(K, \boldsymbol{t}, \boldsymbol{\lambda}) \ , \tag{2}$$

où $\mathcal{T}_K^{\Delta} = \{\boldsymbol{t} = (0 = t_0 < t_1 < \dots < t_{K-1} < t_K = n)\}$. Notons que K est contraint à être inférieur à une valeur $K_{\max}$ fixée par l'analyste sans que cela n'entraîne de perte de généralité. Par ailleurs, selon des arguments similaires à ceux dans [4], l'estimateur $(\hat{K}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{\lambda}})$ est asymptotiquement consistent.

2.2 Procédure d'estimation

Quand K et $\boldsymbol{t}$ sont fixés, l'additivité du critère permet de le maximiser par rapport à $\boldsymbol{\lambda}$ en travaillant segment par segment. On estime donc par maximum de vraisemblance l'intensité d'une loi exponentielle censurée à gauche. En raison de cette censure, il n'existe pas de formule explicite pour l'estimateur $\hat{\boldsymbol{\lambda}}$, on utilise donc une optimisation numérique (méthode de *Newton-Raphson* dans l'implémentation proposée ici). Une fois l'estimateur $\hat{\boldsymbol{\lambda}}$ obtenu, on effectue du *plug-in* pour calculer le score d'un segment.

Pour optimiser le critère par rapport à K et $\boldsymbol{t}$, on utilise l'algorithme *Pruned Exact Linear Time* (PELT [3]). Il permet d'obtenir une segmentation optimale efficacement en combinant programmation dynamique et élagage : le coût de calcul est en $\mathcal{O}(n)$. Notons que le critère retenu doit satisfaire certaines propriétés pour que PELT soit applicable, ce que nous avons vérifié dans le cas présent. L'une d'entre elles consiste en l'augmentation du score d'une séquence de données lors de l'introduction d'une rupture dans celle-ci.

3 Illustration sur données simulées

Dans un premier temps, la méthode paramétrique introduite ci-dessus sera comparée à l'état de l'art, et plus particulièrement à l'approche non-paramétrique *MultRank* décrite dans [5]. Cette dernière est inspirée par l'approche non-paramétrique usuelle basée la recherche de segments homogènes à partir d'un test sur les rangs, et l'adapte au cas où de la censure est présente. On comparera en particulier la capacité des deux méthodes à détecter la présence d'une rupture dans les données. Lorsqu'on se place dans ce cadre, comparer l'approche non-paramétrique à l'approche paramétrique revient à utiliser un rapport de vraisemblance pour la dernière. Remarquons ici que réaliser un test de rapport de vraisemblance ou maximiser la vraisemblance pénalisée introduite dans l'Equation 2 pour $K_{max} = 1$ est équivalent, modulo le choix de la pénalité.

Les données simulées représentent $M = 2000$ échantillons de $n = 200$ réalisations d'une loi exponentielle. La moitié de ces échantillons ne contiendront pas de rupture, alors que l'autre moitié présentera une rupture en position $\frac{n}{2} = 100$. On notera λ_0 le paramètre de l'exponentielle du segment situé à gauche de la rupture et λ_1 celui du segment à droite. Les échantillons ne comportant pas de rupture seront générés selon une exponentielle de paramètre λ_0.

Les statistiques du test non-paramétrique et du test de rapport de vraisemblance seront calculées pour chaque échantillon, et permettront de calculer des courbes ROC et des aires sous la courbe associées, afin de comparer les performances des deux approches. A priori, les performances de l'approche paramétrique devraient être meilleures car le modèle est ici bien spécifié. Néanmoins, l'approche paramétrique devrait être plus sensible dans certains cas, et notamment en raison de la convergence très lente du rapport de vraisemblance et de la faible puissance du test [2] pour la distribution exponentielle.

Les résultats obtenus sur les données simulées sont présentés dans la Figure 1. Pour les deux méthodes, on calcule l'aire sous la courbe ROC, en faisant varier la nombre minimum d'observations avant une rupture. D'après ces résultats, si l'on contraint l'intervalle entre deux ruptures à contenir suffisamment d'observations (un dixième de la taille du signal ici pour $n = 200$), la méthode paramétrique obtient de meilleurs résultats. Par ailleurs, les performances des deux méthodes sont également comparées en fonction du seuil de censure. Dans le cas illustré ici, pour un seuil de censure égal à la médiane de la distribution dans le second segment, les résultats restent comparables, et les performances de la méthode paramétriques sont toujours meilleurs.

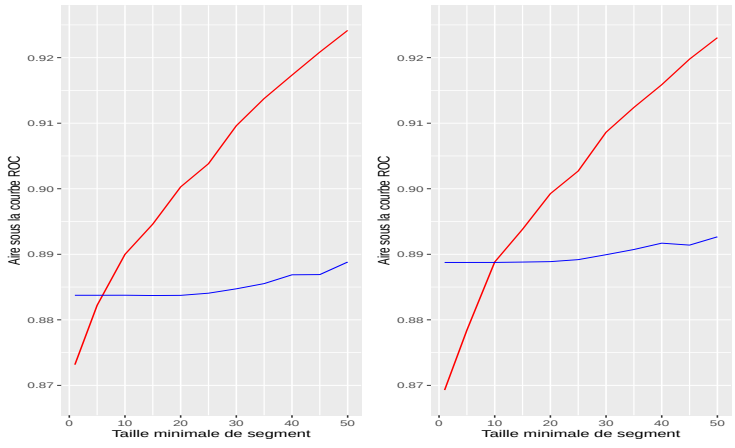

FIGURE 1 – Aire sous la courbe ROC. LR : courbe rouge, MultRank : courbe bleue. $\lambda_0 = 4$, $\lambda_1 = 6$. Gauche : données sans censure. Droite : données avec censure $a = q_{50\%}$ d'une loi exponentielle de paramètre 6

4 Application sur données réelles

On étudie l'évolution de la concentration de prosulfocarbe entre les années 2007 et 2020. Le prosulfocarbe est un herbicide dont l'usage a été ré-autorisé en 2009. Depuis lors, les ventes de prosulfocarbe ont connu une explosion jusqu'à aujourd'hui, en passant de la dix-septième substance la plus vendue à la quatrième en 2017. Seulement les concentrations mesurées en région Centre-Val de Loire seront étudiées.

L'intérêt de ce cas d'étude réside dans le fait que la fenêtre d'observation temporelle dont nous disposons couvre des années où la substance était (normalement) absente des eaux, ainsi qu'une période de réapparition de cette substance active (son usage étant redevenu légal). Toutes les données utilisées dans cette section peuvent être téléchargées depuis l'adresse `http://www.naiades.eaufrance.fr/acces-donnees#/physicochimie`).

Les résultats obtenus via l'optimisation du critère pénalisé introduit dans l'Equation 2, ainsi que les résultats obtenus avec la méthode non-paramétrique *MultRank* sont illustrés dans la Figure 2. Pour la méthode paramétrique, on utilise une pénalité proportionnelle au BIC et une taille de minimale de segment égale à un dixième de la longueur signal total.

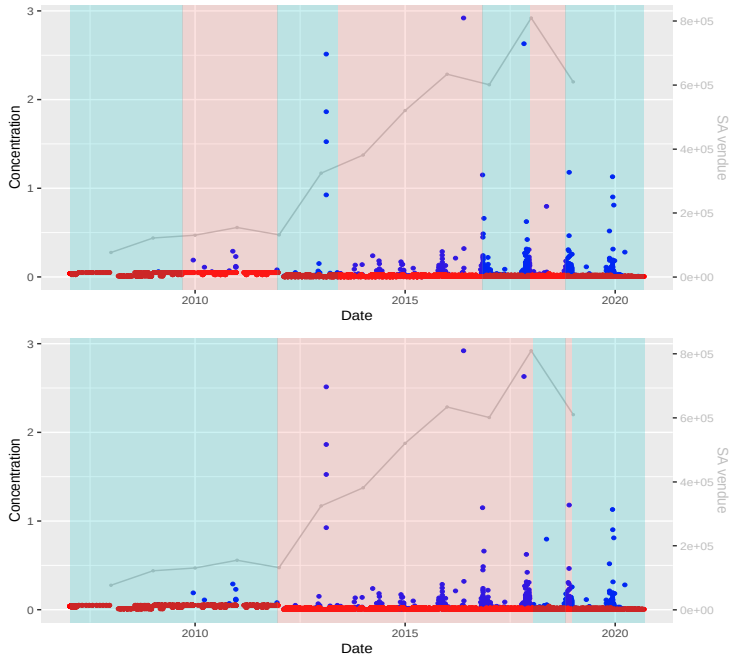

FIGURE 2 – Concentrations (en μg/L) de prosulfocarbe en Centre Val-de-Loire en fonction du temps. Les données censurées sont représentées en rouge. Haut : résultats de la méthode paramétrique. Bas : résultats de la méthode *MultRank*. La courbe grise représente le tonnage des ventes de prosulfocarbe par année en Centre Val de Loire.

Les ruptures détectées par la méthode paramétrique sont disposées de manière plus homogène que celles détectées par *MultRank*. La taille minimale de segment ne permet pas de retrouver le résultat de *MultRank* (voir les deux ruptures formant le segment en 2019 du graphe du bas).

La méthode paramétrique pose des ruptures dès que l'on observe une valeur élevée de concentration. Cela illustre la différence de robustesse entre les deux méthodes.

On peut remarquer que certaines positions de ruptures sont communes aux deux méthodes. Celle ayant eu lieu en 2018 coïncide avec un pic de concentration. Peu de temps précédant cette rupture se trouve la valeur la plus extrême de concentration du signal. Cette détection arrive lors de la plus grosse année de vente également.

Pour finir, la première rupture de *MultRank* en 2012 montre que les deux méthodes sont sensibles aux changements de LQ. Bien qu'elle corresponde au début de la croissance des ventes, on souhaiterait éviter de telles détections car elles ne correspondent pas à une augmentation dans les concentrations de prosulfocarbe (qui n'arrive qu'un an après). Cela s'explique plutôt par un changement de matériel de la part des laboratoires chargés de la mesure.

Références

[1] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics : overview and implementation. *Statistics and Computing*, 22(2) :455–470, apr 2011.

[2] P. Haccou, E. Meelis, and S. van de Geer. The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic Processes and their Applications*, 27 :121–139, 1987.

[3] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500) :1590–1598, oct 2012.

[4] M. Lavielle. Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1) :79–102, sep 1999.

[5] A. Lung-Yut-Fong, C. Levy-Leduc, and O. Cappe. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, 156(4) :133–162, 2015.

[6] J. Réty. Évaluation des risques liés aux résidus de pesticides dans l'eau de distribution : Contribution à l'exposition alimentaire totale. Technical report, Anses, 2013.

[7] Charles Truong. *Multiple change point detection – application to physiological signals.* Theses, Université Paris Saclay, November 2018.