UpDown, une méthodologie d'identification et de caractérisation de perturbations affectant des données longitudinales

Ingrid David, Vincent Le, Tom Rohmer

UPDOWN, UNE MÉTHODOLOGIE D'IDENTIFICATION ET DE CARACTÉRISATION DE PERTURBATIONS AFFECTANT DES DONNÉES LONGITUDINALES

Ingrid David[1], Vincent Le[1,2], Tom Rohmer[1,*]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France*
[2] *Alliance R&D, 35650 Le Rheu, France*
[*] *Corresponding author; tom.rohmer@inrae.fr*

Résumé. Dans de nombreux contextes scientifiques, des données longitudinales sont observées, et organisées en groupe hiérarchique. C'est le cas notamment des données d'élevage issues de capteur haut-débit pour lesquelles les mesures sont réalisées automatiquement sur différents lots d'animaux répartis dans différents enclos. Dans ces systèmes hiérarchiques, des perturbations peuvent affecter un individu ou un groupe d'individus en modifiant la trajectoire attendue des observations. Être capable d'identifier une perturbation affectant un individu ou un groupe d'individus (pouvant avoir des réponses différentes due à leur 'robustesse' propre face à la perturbation) est souvent en enjeu majeur. La méthode UpDown consiste à utiliser l'information à l'échelle des groupes ou des sous-groupes pour faciliter la détection et la caractérisation des perturbations (début, durée, intensité) qui affectent les éléments constituant les groupes. Un package R a été développé permettant de considérer autant de niveaux hiérarchiques que nécessaire pour s'adapter à différentes problématiques scientifiques. Appliqué sur des données simulées mimant des observations issues de système d'élevage porcins observées sur une période de 100 jours, la méthode UpDown a montré une sensibilité pour détecter les perturbations aux différentes échelles allant de 43% à l'échelle individuelle jusqu'à 93% au 3eme niveau hiérarchique, associé à une excellente spécificité à toutes les échelles (>95%). Finalement l'écart médian entre les débuts et fin estimés et théoriques était inférieur à 3 jours. La corrélation entre les intensités théoriques et simulées était supérieure à 0.72 pour les échelles de groupe.

Mots-clés. Détection de perturbation, système hiérarchique, classification non-supervisé

Abstract. In many scientific contexts, longitudinal observations are observed and organized in hierarchical groups. This is the case for example for breeding data from high throughput sensors, where measurements are automatically performed on animals distributed in pens belonging to different batches. In these hierarchical systems, disturbances can affect an individual or a group of individuals by changing their expected trajectories. Being able to identify a disturbance affecting an individual or a whole group of individuals (which may have different responses to the disturbance due to their own 'resilience') is often a major issue. The UpDown method consists in using the information of all the groups or subgroups to facilitate the detection and the characterization of disturbances (beginning, duration, intensity) affecting the elements constituting the groups. An R package has been developed to consider as many hierarchical levels as necessary, thus allowing to adapt to different scientific problems. Applied on simulated data mimicking observed observations over 100 days from pig farming systems, the UpDown method showed a sensitivity to detect disturbances at different scales ranging from 43% at the individual level to 93% at the 3rd hierarchical level,

associated with an excellent specificity at all scales (>95%). Finally, the median difference between the estimated and theoretical beginnings and ends was less than 3 days in these simulations. The correlation between the theoretical and simulated intensities was greater than 0.72 for the group scales.

Keywords. Change-point detection, hierarchical system, unsupervised classification

1 Méthodologie

Les observations longitudinales organisées en groupes hiérarchiques sont étudiées dans de nombreux contextes. Par exemple, dans les systèmes éducatifs, des indicateurs de progrès peuvent être mesurés pour des élèves répartis dans des classes appartenant à des écoles différentes. En analyse de marché, l'évolution du prix d'une propriété peut dépendre de sa localité (quartier intra ville intra pays). Dans les systèmes d'élevage de porcs, les animaux sont répartis dans des enclos répartis eux même en différentes bandes. Dans ces 3 exemples, les observations peuvent alors être considérées à différents niveaux : échelle de l'individu ou du groupe.

Des perturbations qui modifient les trajectoires des observations peuvent également se produire à ces différents niveaux. Elles peuvent modifier une dynamique individuelle ou un groupe d'observations longitudinales. Ainsi, en reprenant l'exemple des élevages, un animal peut être malade à cause d'une maladie qui n'est pas contagieuse. Dans ce cas, la perturbation (la maladie) ne se produit qu'au niveau individuel. Si la maladie est contagieuse, tous les animaux élevés ensemble dans le même enclos (même groupe) seront confrontés à la maladie, la perturbation se produit au niveau de l'enclos, même si tous les animaux ne développeront pas de symptômes en fonction de leur capacité immunitaire (‘robustesse’). La réponse d'un animal à une perturbation dépend ainsi de sa robustesse et des caractéristiques de la perturbation subie (intensité, date de début et durée).

Dans ce contexte, identifier si un animal en particulier a été soumis à une maladie est plus efficace en utilisant les observations de tous les animaux d'un enclos qu’en observant seulement l'animal particulier qui peut ne pas développer de symptômes.

L'algorithme **UpDown** a été développé autour de cette idée. Il consiste à étudier les observations à différentes échelles afin de faciliter la détection et la caractérisation des perturbations. Un **package R** du même nom sera prochainement déposé sur le CRAN.

Lorsque les observations sont recueillies de manière répétée dans le temps (données longitudinales), les perturbations peuvent être identifiées en observant les changements dans la dynamique de l'observation au cours du temps. L'algorithme **UpDown** est capable de détecter et de caractériser les perturbations qui conduisent à une réponse élastique ou plastique (Sauvant and Perez, 2010), c’est -à- dire que la dynamique des observations va changer pendant la perturbation pour revenir à son état d’origine (système élastique) ou à un autre état (système plastique) à l’issu de la perturbation tel qu’illustré dans la Figure 1.

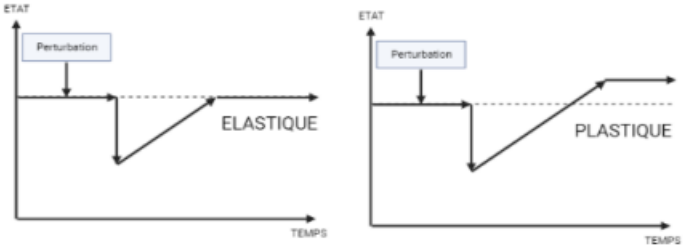

Figure 1 : Illustration d'une réponse élastique et plastique face à une perturbation

La méthode **UpDown** capture les déviations par rapport à la trajectoire théorique (celle qui aurait été observée en l'absence de perturbation), corrigé de leur évolution naturelle, en identifiant les évolutions anormales par rapport à celle attendue. Il se compose de deux parties. L'étape **Up** identifie les éléments confrontés à une perturbation du plus bas niveau hiérarchique (individu) au plus haut. La classification des éléments perturbés est faite en ajustant des modèles de mélange Gaussiens successifs sur les vitesses minimales des dynamiques (via un lissage à noyau Gaussien) individuelles et médianes par groupe, en analysant les probabilités à posteriori d'appartenance à l'ensemble des éléments perturbés et non perturbés. e.g. Figure 2.

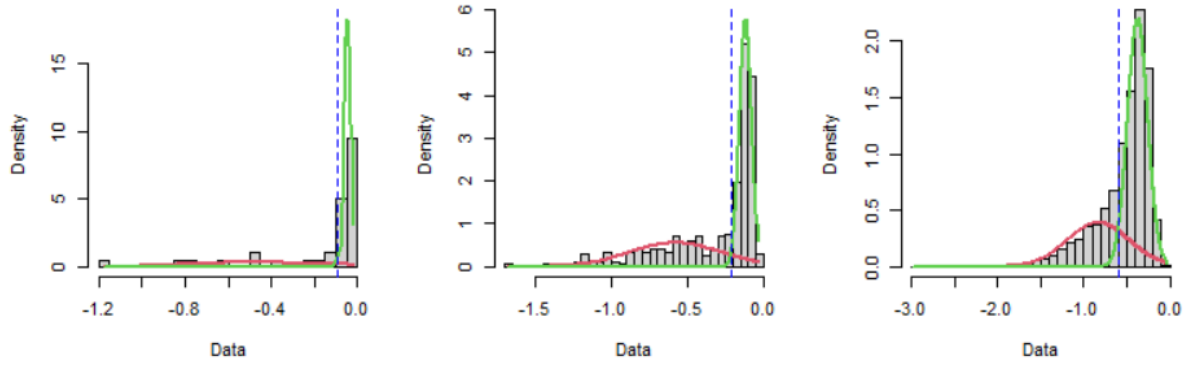

Figure 2 : Modèle de mélange Gaussien pour la classification des éléments perturbés en fonction du niveau hiérarchique, du haut niveau (gauche) au plus bas niveau, i.e. à l'échelle individuelle (droite)

L'étape **Down** valide les groupes hiérarchiques identifiés dans l'étape **Up** comme étant sujets à perturbations, du plus haut niveau au niveau le plus bas (individu) en analysant les débuts de réactions des éléments constituant les groupes. Cette étape permet également d'identifier des groupes ou individus qui subissent plus d'une perturbation (quel que soit l'échelle).

Les débuts de réaction sont estimés par des recherches de minimums locals sur la dérivée de la courbe de lissage. Les fins de réactions sont estimées par des recherche de minimum local sur la trajectoire des observations, survenant après un début de perturbation identifié. Enfin l'intensité de la perturbation sera estimée par la pente entre le début et la fin de la réaction.

2 Résultats

Afin de quantifier la qualité de la méthode pour identifier les perturbations affectant un groupe des observations longitudinales soumises à de telles perturbations ont été simulées (Le et al. 2022) et la sensibilité et spécificité de la méthode ont été évaluées (Le 2022).

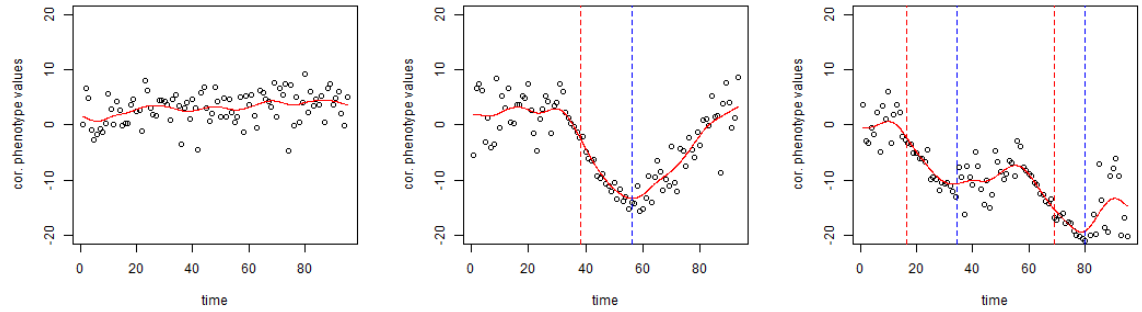

Figure 3 : Exemple d'individus simulés présentant (gauche) aucune perturbation, (milieu) une perturbation individuelle (droite) deux perturbations de groupes. Les courbes de lissages sont représentés en rouge, les débuts et fin de perturbations estimés sont représenté resp. par les lignes pointillées rouges et bleues.

Nous avons généré 1000 réalisations contenant chacune les trajectoires de 6000 individus (93 à 100 observations journalières par individus) réparties en 3 niveaux hiérarchiques. A chacun des niveaux, 20% des éléments ont été soumis à perturbations. Ainsi, chaque individu pouvait subir jusqu'à 3 perturbations simultanément. Les pourcentages de perturbations détectées et non-détectées à raison pour chacun des niveaux hiérarchiques sont résumés dans la Table 1 :

	Niveau 3	Niveau 2	Niveau 1 (individu)
Sensibilité (%)	0.93	0.73	0.43
Spécificité (%)	0.99	0.98	0.95

Table 1 : Qualité de la détection des perturbations en fonction des différents niveaux hiérarchiques

La méthode **UpDown** a ainsi montré une sensibilité de détection des éléments subissant une perturbation qui augmente avec le niveau hiérarchique (de 43% au niveau individuel à 93% au niveau 3) et est associée à une bonne spécificité pour tous les niveaux (>95%). Le plus faible taux de détection au niveau individuelle s'explique du fait des robustesses spécifiques de chaque individu face aux perturbations.

De plus, à chacun des niveaux hiérarchiques, l'écart médian entre la date de début (de fin) estimée et la date réelle était toujours inférieur à 3 jours. Enfin les corrélations entre les valeurs d'intensité estimé et théoriques (qui a du sens uniquement aux échelles du groupe du fait des 'robustesses' individuelles) étaient grandes (>0.72 aux 2 niveaux de groupe

considérés) et plus importantes pour les perturbations de durée longue comme illustré dans la figure 4.

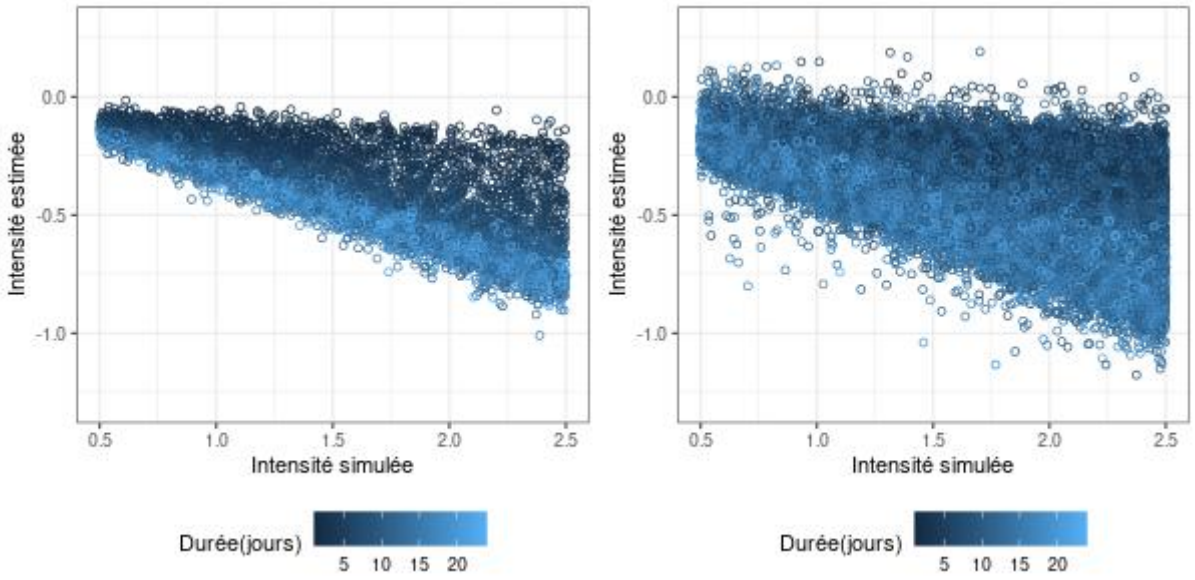

Figure 4 : Intensités simulés vs estimées sur 1000 runs, à l'échelle de la bande (gauche), à l'échelle de la case (droite) pour des durées de perturbations allant de 2 à 25 jours.

Remerciements

Ces travaux ont été financés à 50% par l'unité de Génétique Animal d'INRAE et à 50% par Alliance R&D, et nous tenons à remercier ces derniers par la relecture des manuscrits associés ainsi que par les différents tests effectués sur le package R en soumission prochaine.

Bibliographie

Sauvant D. and Martin O. 2010. Robustesse, rusticité, flexibilité, plasticité... les nouveaux critères de qualité des animaux et des systèmes d'élevage : définitions systémique et biologique des différents concepts. INRAE Productions Animales 23, 5–10

Le V., Rohmer T. and David I., 2022. Impact of environmental disturbances on estimated genetic parameters and breeding values for growth traits in pigs. Animal 16, 100496.

Le V. 2022. Nouvelle mesure de la robustesse des animaux d'élevage par utilisation des données de phénotypage haut-débit. Theses, INPT Toulouse, Nov. 2022.
URL https://hal.inrae.fr/tel-03967884.