Test de Runs pour données directionnelles : propriétés locales et optimalités asymptotiques.

Maxime Boucher [1] & Christian Franck [2] & Yuichi Gotto [3] & Thomas Verdebout [4]

[1] *Université Libre de Bruxelles (ULB), Belgique, maxime.boucher@ulb.be*
[2] *CREST et Université de Lille, France, christian.francq@ensae.fr*
[3] *Université de Kyushu, Japon, goto.yuichi.436@m.kyushu-u.ac.jp*
[4] *Université Libre de Bruxelles (ULB), Belgique, thomas.verdebout@ulb.be*

Résumé. Dans le travail présenté ici, on s'intéresse au problème de détection des corrélations sérielles dans un contexte de données directionnelles. Motivé par une application sur des données réelles concernant la localisation de tâches solaires au cours des décennies, on définit un concept de runs proprement adapté au contexte directionnel. On montre alors que ce test, basé sur les runs directionnels, possède des propriétés locales et asymptotiques dans le cas d'alternatives locales avec des dépendances sérielles. A l'aide de simulations Monte-Carlo, on expose les propriétés, pour des tailles d'échantillons finies, de notre test et son utilité dans le cadre de l'étude des localisations des tâches solaires pendant les dix derniers cycles solaires.

Mots-clés. runs, données directionnelles, dépendance sérielle, optimalité locale et asymptotique, randomness.

Abstract. In the present work, we tackle the problem of detecting serial correlation in the context of directional data. Motivated by a real data example involving sunspots locations, we define a concept of runs properly adapted to the directional context. We then show that tests based on the latter runs enjoy some local and asymptotic property against local alternatives with serial dependence. We compute the finite sample performances of our tests using Monte Carlo simulations and show their usefulness on a real data illustration that involves the analysis of sunspots locations for various solar cycles.

Keywords. runs, directional data, randomness test, serial dependence, local and asymptotic optimality, randomness.

1 Test de runs directionnels

1.1 Contexte Directionnel

Dans le cas univarié, on définit un run comme une suite consécutive d'observations du même signe. Si l'on dispose d'un échantillon univarié $X_1, \ldots, X_n$ avec pour paramètre de localisation θ, on définit le signe de X_t par $U_t = sign(X_t - \theta)$. Le nombre de runs dans un échantillon peut-être alors calculé de la manière suivante :

$$\sum_{t=2}^{n} U_t(\theta)U_{t-1}(\theta) = N_n(\theta) - \mathbb{E}[N_n(\theta)].$$

où $N_n(\theta) := 1 + \sum_{t=2}^{n} \mathbb{I}[U_t(\theta) \neq U_{t-1}(\theta)]$ est le nombre de runs. On peut donc alors construire un test de runs pour tester la randomness de l'échantillon à l'aide de l'asymptotique gaussien classique.

Dans le cas multivarié, la notion de signe est adaptée par le signe multivarié comme étant, pour un échantillon de p-vecteurs $X_1, \ldots, X_n$, les quantités :

$$U_t = \frac{X_t - \theta}{\|X_t - \theta\|}$$

On définit alors la notion de run dans le cas multivarié par :

$$R_1^{(n)} := \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} U_t'(\theta)U_{t-1}(\theta)$$

où θ est à nouveau le vecteur de localisation. La quantité $R_1^{(n)}$ permet alors de construire un test afin de détecter une possible dépendance dans l'échantillon (une dépendance avec un lag de taille 1 puisque l'on regarde l'angle entre deux vecteurs consécutifs).

Dans ce travail, on se focalise plus particulièrement sur les runs pour un échantillon directionnel. C'est-à-dire que l'on dispose d'un échantillon de p-vecteurs sur la sphère $\mathcal{S}^{p-1}$ de $\mathbb{R}^p$. Dans le cadre directionnel, les données peuvent se décomposer de la manière suivante:

$$X_t = v_t(\theta)\theta + \sqrt{1 - v_t(\theta)^2}\Gamma_\theta S_t(\theta)$$

où $v_t(\theta)$ est la partie projetée de X_t sur la direction θ, Γ_θ est une matrice semi-orthogonale qui vérifie $\Gamma_\theta\Gamma_\theta' = I_p - \theta\theta'$ et $\Gamma_\theta'\Gamma_\theta = I_{p-1}$ et le vecteur $S_t(\theta) = \Gamma_\theta' X_t / \|\Gamma_\theta' X_t\|$ représente la composante tangentielle de X_t. Dans ce cas là, $S_t(\theta)$ est le **signe multivarié**. On peut donc adapter la définition des runs au cadre directionnel avec :

$$R_{1,d}^{(n)} := \frac{1}{\sqrt{n-1}} \sum_{t=2}^{n} S_t(\theta)' S_{t-1}(\theta)$$

Dans cet exposé, on s'intéresse au cas où les données sont à symétrie rotationnelle autour de la direction θ. Cela implique que la densité de l'échantillon est de la forme :

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} c_{p,g} g(x'\theta)$$

où g est appelée la fonction angulaire, $c_{p,g}$ est la constante de normalisation. Dans ce contexte, on connait la densité des composantes projetées $v_t(\theta)$, on sait que les composantes tangentielles sont uniformes sur la sphère $\mathcal{S}^{p-2}$ et enfin que les deux composantes sont indépendantes.

1.2 Test d'hypothèse : randomness contre dépendance sérielle tangentielle

Dans notre travail, on définit la distribution Markov-tangentielle commme étant, pour un paramètre $\lambda > 0$:

$$S_1(\theta), \dots, S_n(\theta) \text{ ont pour densité } (s_1, \dots, s_n) \mapsto c_\lambda^n \exp\left(\lambda \sum_{t=2}^{n} s_t' s_{t-1}\right) \text{ sur } (\mathcal{S}^{p-2})^n.$$

Dans ce cas :

- $S_t(\theta) \sim \mathcal{U}_{\mathcal{S}^{p-2}}$,
- $S_t(\theta)|S_{t-1}(\theta) = s_{t-1} \sim VMF(s_{t-1}, \lambda)$ où $\lambda > 0$ est le paramètre de concentration, s_{t-1} est le paramètre de localisation de la loi de Von-Mises Fisher.

Théorème 1 *Supposons que les composantes $(v_t(\theta))_{t=1,\dots,n}$ sont i.i.d avec pour densité $\tilde{g}$ sur $[-1, 1]$ et sont indépendantes des signes $(S_t(\theta))_{t=1,\dots,n}$ distribués suivant la distribution de Markov-Tangentielle. Alors, $vec(X_1, \dots, X_n)$ a pour densité :*

$$vec(x_1, ..., x_n) \mapsto c_{p,g}^n c_\lambda^n \exp\left(\lambda \sum_{t=2}^{n} s_t' s_{t-1}\right) \prod_{t=1}^{n} g(v_t(\theta))$$

selon la mesure de surface sur $\mathcal{S}^{p-1}$.On note dans ce cas : $vec(X_1, \dots, X_n) \sim P_{\theta,g,\lambda}$

Notre travail consiste alors à étudier le test :

"iidness" contre "dépendance sérielle"

Ce qui revient à tester, grâce au Théorème 1 :

$$H_0 : "\lambda = 0" \text{ contre } H_1 : "\lambda > 0"$$

1.3 Résultats Principaux

On a montré les résultats suivant :

Théorème 2 *Sous l'hypothèse nulle, dans le cas où les $S_1(\theta), ..., S_n(\theta)$ sont bien iidness, et considérant la quantité*

$$s_n(\theta) = tr\left[\left(\frac{1}{n}\sum_{t=1}^{n} S_t(\theta)S_t(\theta)'\right)^2\right],$$

Nous avons :

$$s_n(\theta)^{-\frac{1}{2}} R_{1,n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Nous étudions également ce qu'il se passe sous l'alternative : on considère une perturbation de l'hypothèse nulle $\left(0+\frac{1}{\sqrt{n}}\ell_n, \theta+\frac{1}{\sqrt{n}}\tau_n\right)$ avec deux suites bornées (ℓ_n) et (τ_n) où (τ_n) converge vers $0 \neq \tau \in \mathbb{R}^p$ et (l_n) converge vers $0 \neq l \in \mathbb{R}$. On considère le ration log-vraisemblance :

$$\Lambda_n := \log \frac{dP^{(n)}_{\theta+\frac{1}{\sqrt{n}}\tau_n, g, 0+\frac{1}{\sqrt{n}}\ell_n}}{dP^{(n)}_{\theta,g,0}}$$

Théorème 3 (Résultat LAN) *Soit $u_n := (\ell_n, \tau_n')'$, on a :*

$$\Lambda_n = u_n'\Delta_n - \frac{1}{2}u_n'\Gamma u_n + o_{\mathrm{P}}(1)$$

pour $n \to +\infty$ sous $P^{(n)}_{\theta,0,g}$, où $\Delta_{\theta,n} := n^{-1/2}\sum_{t=1}^{n}\varphi_g(v_t(\theta)(1-v_t(\theta)^2)^{1/2}S_t(\theta)$ et $\Delta_{\lambda,n} := n^{-1/2}\sum_{t=2}^{n} S_t(\theta)'S_{t-1}(\theta)$, la suite centrale $\Delta_n := (\Delta_{\lambda,n}, (\Delta_{\theta,n})')'$ est asymptotiquement gausienne (toujours sous $\mathrm{P}^{(n)}_{\theta,0,g}$) de moyenne nulle et pour matrice de covariance

$$\Gamma := \mathrm{diag}((p-1)^{-1}, \tilde{\Gamma}).$$

De ce résultat LAN, on en déduit le théorème suivant :

Théorème 4 *D'après le résultat LAN précédent, pour le test :*

$$H_0 : "\lambda = 0" \text{ contre } H_1 : "\lambda > 0"$$

Le test asymptotiquement et localement plus puissant est le test $\phi^{(n)}_{opt}$ qui rejette l'hypothèse nulle au niveau $\alpha \in]0,1[$ quand :

$$\Delta_{\lambda,n}\sqrt{p-1} > z_{1-\alpha}$$

où $z_{1-\alpha}$ est le quantile de la loi normale.

Ainsi, nous pouvons construire le test le plus puissant pour l'étude de ce problème avec comme alternative une distribution de Markov-Tangentielle. On en déduit aussi de ce résultat qu'il n'y a pas de coût asymptotique en remplaçant θ par une estimation tant qu'on utilise un estimateur $\hat{\theta}$ qui est $\sqrt{n}$-consistant. On définit également une statistique de test afin de pouvoir détecter des dépendances au-delà d'un lag de 1.

Théorème 5 *On définit le run de lag $h \in \mathbb{N}^*$ par :*

$$R_{h,n}(\theta) := \frac{1}{\sqrt{n-h}} \sum_{t=h+1}^{n} S_t(\theta)' S_{t-h}(\theta).$$

Un test pouvant détecter une dépendance à l'ordre $H \in \mathbb{N}^*$ peut être construit à l'aidre de la statistique

$$s_n^{-1}(\theta) \sum_{h=1}^{H} (R_{h,n}(\theta))^2.$$

Nous avons que $s_n^{-1}(\theta) \sum_{h=1}^{H} (R_{h,n}(\theta))^2$ converge en loi vers

(i) un chi-deux avec H degrés de liberté sous $\mathrm{P}_{\theta,g}^{(n)}$ et

(ii) un chi-deux avec H degrés de liberté et un paramètre de décentralité $(p-1)^{-1}\ell^2$ sous $\mathrm{P}_{\theta,n^{-1/2}\ell_n,g}$, où $\ell := \lim_{n\to+\infty} \ell_n$.

1.4 Simulations Monte-Carlo

Nous appuyons notre travail avec des simulations Monte-Carlo où nous calculons les courbes estimées des puissances de notre test en comparaison avec les tests traditionnels.

Ces courbes (Figure 1) illustrent très bien la propriété de non-coût asymptotique dans le remplacement de θ par une estimation. De plus, on illustre la puissance qui augmente au fur-et-à-mesure que la valeur de λ grandit (et donc que l'on s'éloigne de l'hypothèse nulle). On remarque également que l'on obtient bien la valeur nominale α pour $\lambda = 0$ donc sous H_0.

2 Application aux données solaires

Pour terminer, dans ce travail, nous proposons d'appliquer nos runs à l'étude des positions des tâches solaires pendant les derniers cycles enregistrés (cycle 11 à 24). En Figure 2, on trouve la représentation des positions de ces tâches solaires au cours du temps et sur plusieurs cycles. Le dégradé illustre l'évolution au cours du cycle (rouge en début de cycle et jaune en fin de cycle). Au vu de ces illustrations, on peut supposer que l'hypothèse de symétrie rotationnelle est vérifiée.

Dans la littérature, une dépendance sérielle à déjà était mise en évidence : la loi de Spörer. Elle permet de modéliser la dépendance des observations par leurs latitudes (en

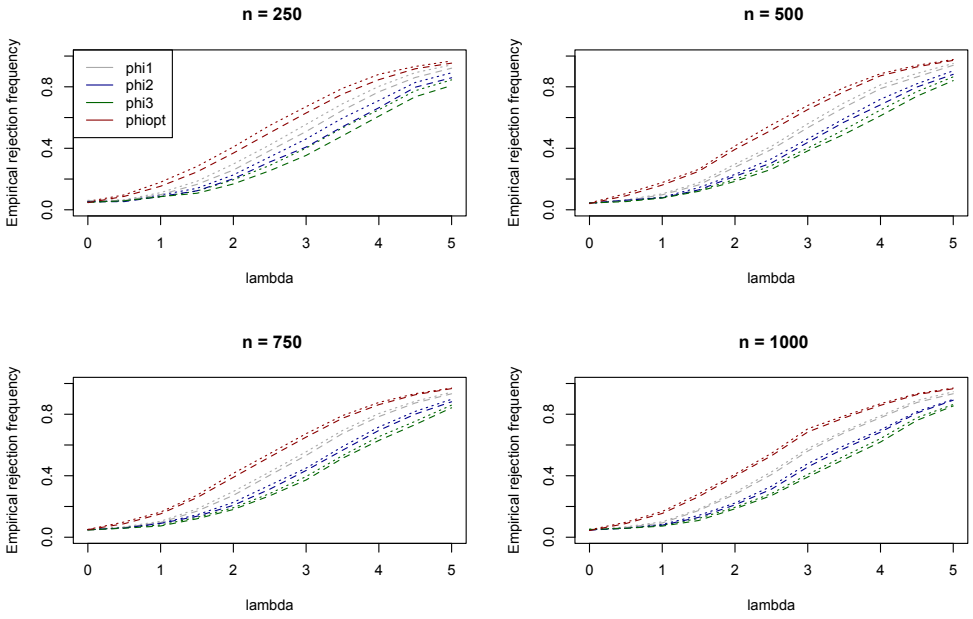

Figure 1: Fréquences de rejets empiriques pour plusieurs tests de runs directionnels : $\phi_1^{(n)}$, $\phi_2^{(n)}$, $\phi_3^{(n)}$ et $\phi_{\text{opt}}^{(n)}$. Avec des pointillés : fréquence de rejet pour un $\theta = (1, 0, 0)'$ alors que la courbe avec des tirets correspond aux calculs de puissances avec un estimateur de θ et avec $\alpha = 5\%$ pour tous.

valeurs absolues). Avec nos travaux, on souhaite étudier si l'on peut détecter une dépendance également via leurs longitudes. Dans les Figure 3 à 5, on présente les boxplot de p-valeurs pour la partie projetée (latitude) selon l'axe du pôle nord de nos observations (pour chaque cycles), les latitudes en valeurs absolues et les longitudes. Chaque boxplot, pour des lags de 1 à 4, ont été construit sur 200 répétitions de nos tests en conservant à chaque fois un sous échantillon aléatoire représentant 75% de l'échantillon de départ. On en conclut qu'en plus de la dépendance sur les latitudes absolues, une dépendance sérielle le long des longitudes est présente. Le cas du cycle 11 reste encore à expliquer bien que le nombre d'observations est très faible en comparaison des autres cycles.

Bibliographie

Hentati-Kaffel, R. and De Peretti, P. (2015). Generalized runs tests to detect randomness in hedge funds returns. *Journal of Banking & Finance*, 50:608–615.

Henze, N. and Penrose, M. D. (1999). On the multivariate runs test. *Annals of statistics*, pages 290–298.

Ley, C., Swan, Y., Thiam, B., and Verdebout, T. (2013). Optimal R-estimation of a spherical location. *Statist. Sinica*, 23(1):305–332

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Wiley Series in Probability and

Figure 2: Positions des tâches solaires pour les cycles 16 à 24.

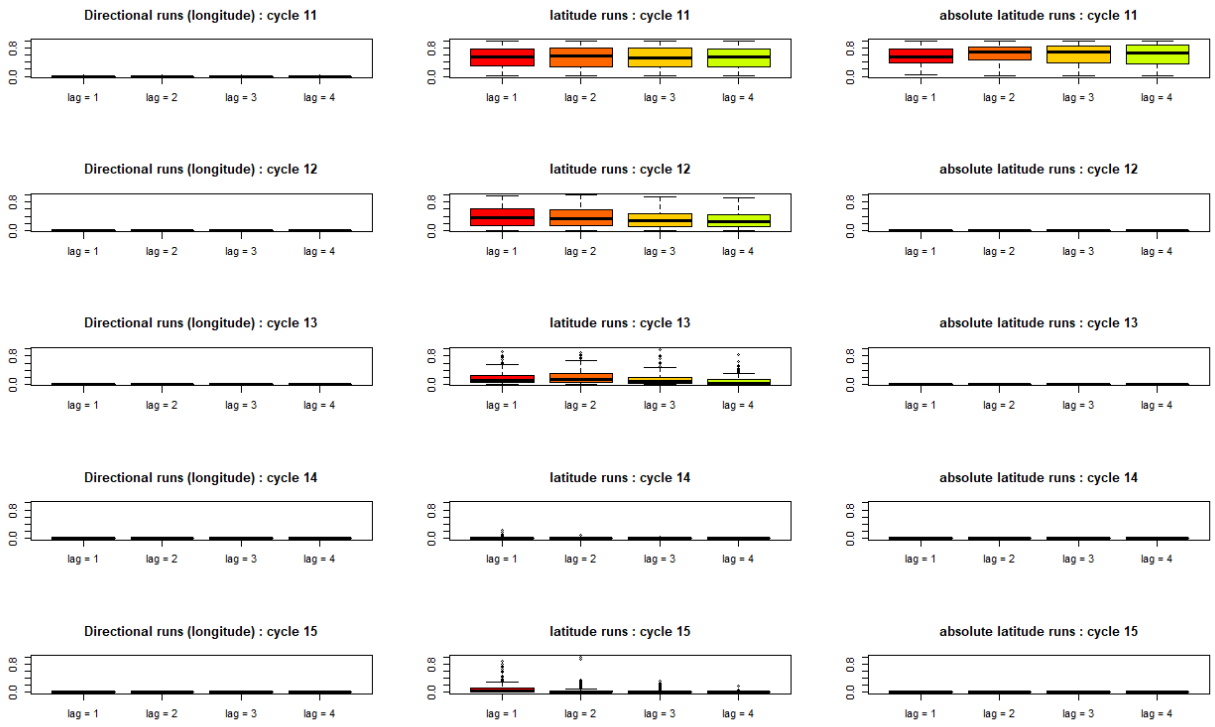

Figure 3: Boxplot des p-valeurs pour les tests de runs pour les latitudes, latitudes absolues et longitudes pour quatre valeurs de lag différentes pour les cycles 11 à 15.

Statistics. Wiley, Chichester.

Paindaveine, D. (2009). On multivariate runs tests for randomness. *Journal of the American Statistical Association*, 104(488):1525–1538.

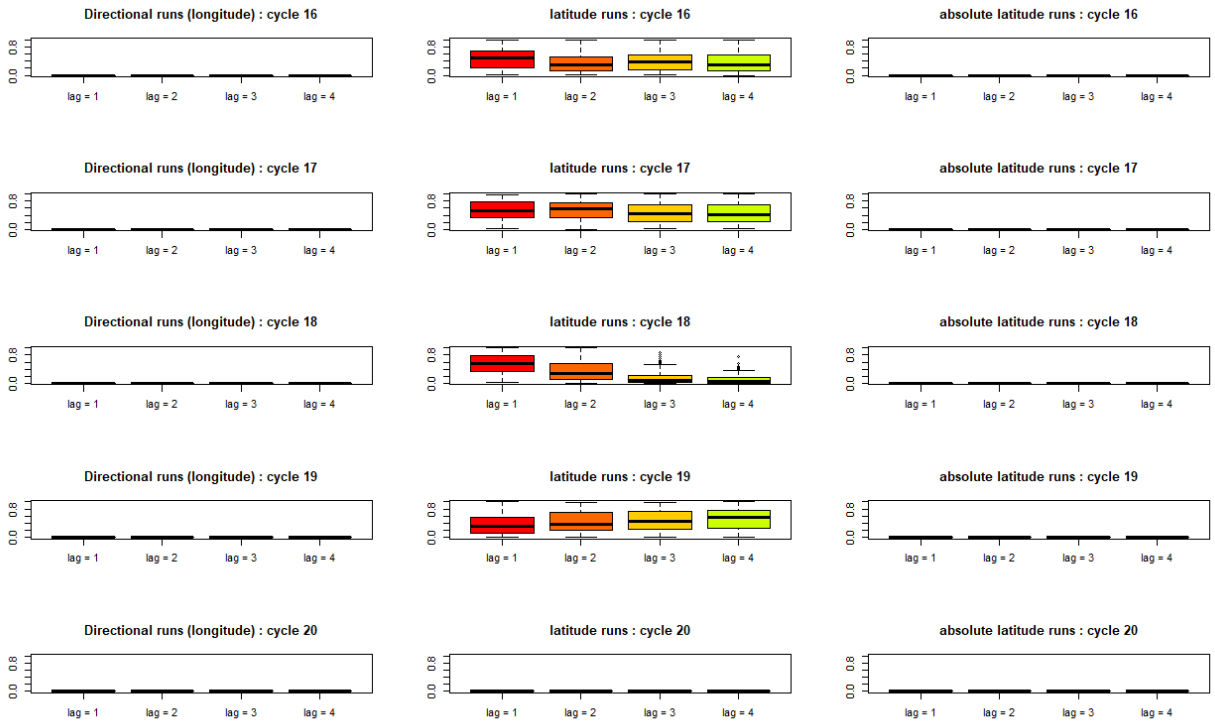

Figure 4: Boxplot des p-valeurs pour les tests de runs pour les latitudes, latitudes absolues et longitudes pour quatre valeurs de lag différentes pour les cycles 16 à 20.

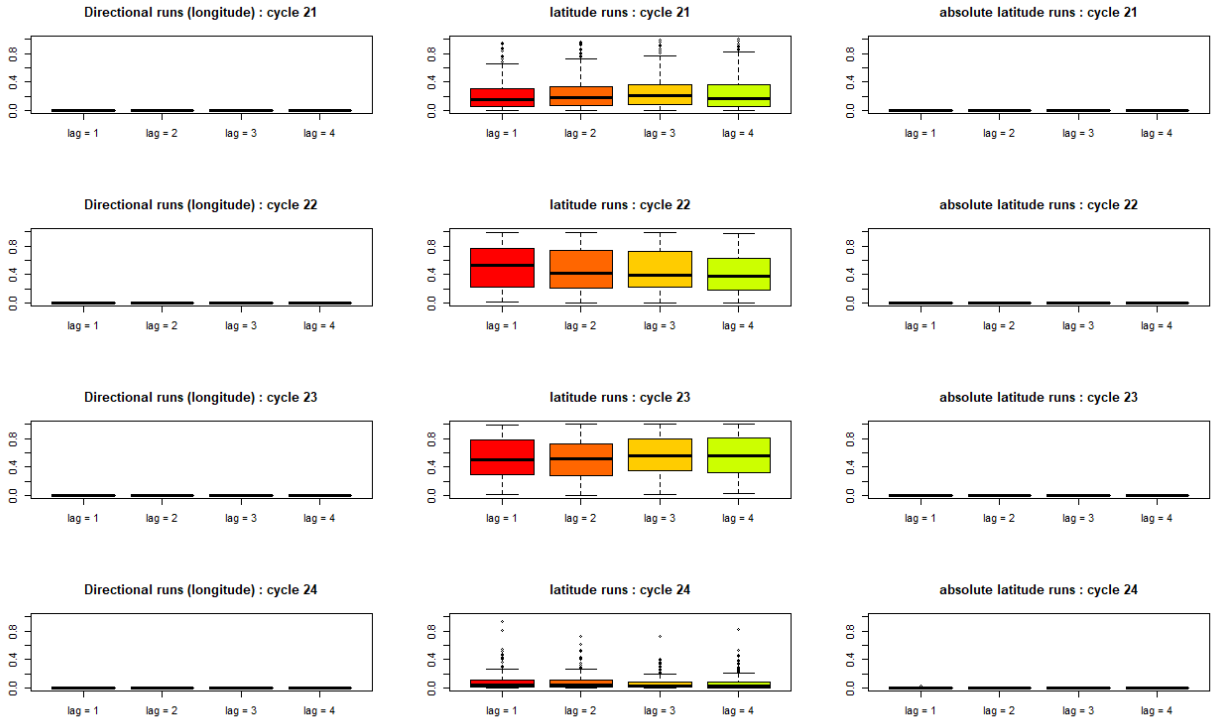

Figure 5: Boxplot des p-valeurs pour les tests de runs pour les latitudes, latitudes absolues et longitudes pour quatre valeurs de lag différentes pour les cycles 21 à 24.