Data challenge

Drift Cost Detection

Simon Querné, Hugo Maurice, Alix Doineau

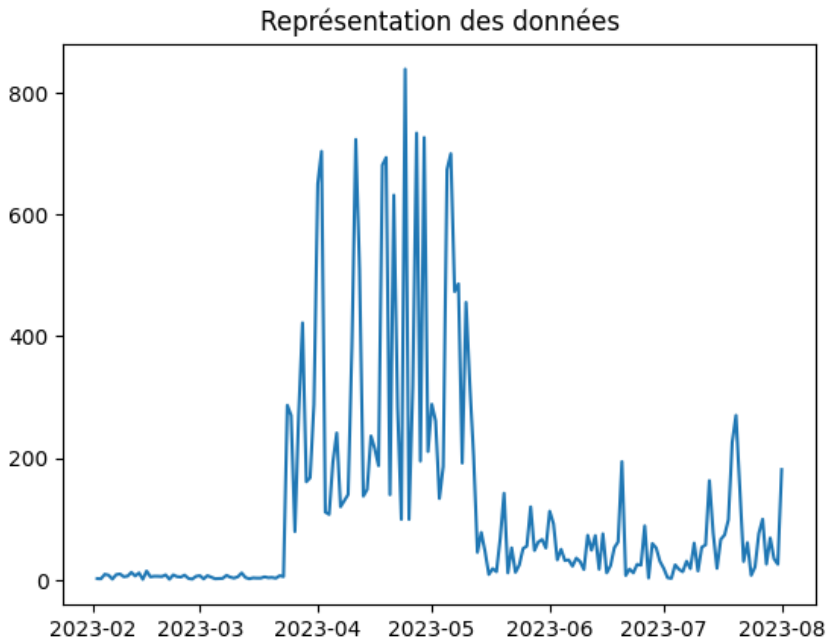
Université Paris-Saclay

11 mars 2024

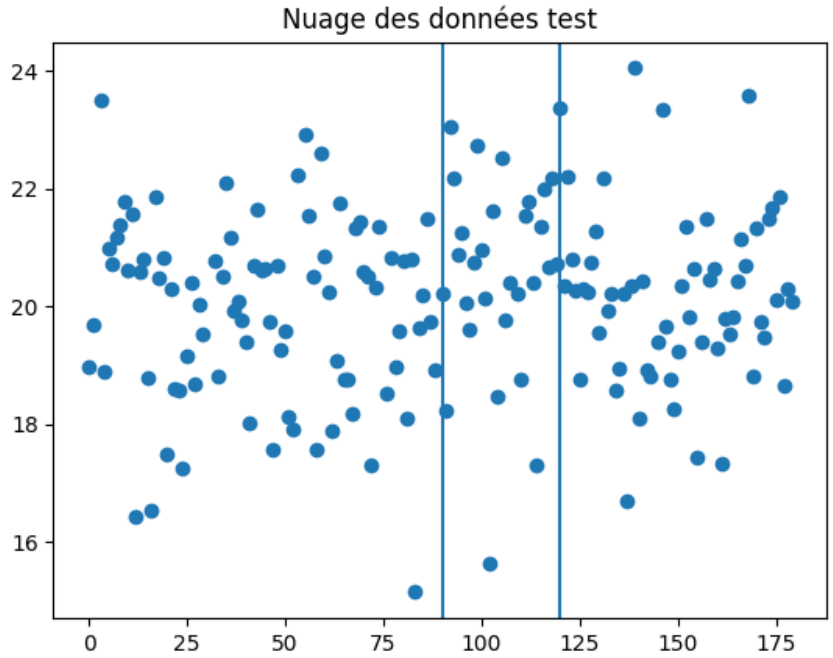
Scan statistics sur des données continues

Si l'échantillon est continu, nous ne pouvons pas construire de test à partir d'une distribution continue. Il est alors possible d'utiliser une approche qui se base sur un test aléatoire de Monte Carlo.
Nous considérons les observations x_i, $i = 1, \ldots, N$, d'une variable aléatoire continue. Chaque observation se situe sur un emplacement s, $s = 1, \ldots, S$. On peut éventuellement avoir plusieurs observations en un même point donc $S \leq N$.

Présentation des données



(a) Coût moyen quotidien associé à un incident technique présentant un cluster anomal



(b) Données simulées normales de moyenne 20 hors cluster et 20,75 dans le cluster (variance = 4)

Démarche et hypothèses

On cherche à détecter un cluster présentant une anomalie, c'est-à-dire un sous-groupe de données distribuées différemment du reste des données. Dans le modèle utilisé, on suppose que les données sont distribuées selon une distribution normale, et que les données du cluster ont une moyenne différente du reste. La recherche du cluster se base donc sur la comparaison de la moyenne du cluster avec les autres données.
Remarque : l'hypothèse normale ne sert qu'à la détection du cluster et n'a pas d'effet sur l'inférence, car le test de Monte Carlo est non-paramétrique.

Détection du cluster

Dans le modèle normal, les EMV de la moyenne et de la variance sont $\mu = N^{-1}\sum_{i=1}^{N} x_i$, $\sigma^2 = N^{-1}\sum_{i=1}^{N}(x_i - \mu)^2$. Alors la formule de la vraisemblance est $\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$. La formule de la log-vraisemblance est alors

$$lnL_0 = -Nln(\sqrt{2\pi}) - Nln(\sigma) - N/2$$

Détection du cluster

Nous introduisons les notations :

- $x_s = \sum_{i \in s} x_i$ la somme au point s.
- n_s le nombre d'observations au point s.
- $X = \sum_{i=1}^{N} x_i$ la somme totale.
- z la fenêtre de centre s et de rayon ω.
- $x_z = \sum_{s \in z} x_s$ la somme sur la fenêtre z.
- $n_z = \sum_{s \in z} n_z$ le nombre d'observations sur la fenêtre z.
- $\mu_z = x_z / n_z$ la moyenne du cluster.
- $\lambda_z = (X - x_z)/(N - n_z)$ la moyenne hors du cluster.

Détection du cluster

Pour détecter le cluster, nous regardons toutes les fenêtres z construites autour de chaque point s et calculons la vraisemblance de chaque fenêtre selon la formule

$$lnL_z = -Nln(\sqrt{2\pi}) - Nln(\sigma_z) - N/2$$

Où σ_z^2 est l'EMV de la variance commune de formule

$$\sigma_z^2 = \frac{1}{N}\left(\sum_{s\in z} x_s^2 - 2x_z\mu_z + n_z\mu_z^2 + \sum_{s\notin z} x_s^2 - 2(X - x_z)\lambda_z + (N - n_z)\lambda_z^2\right)$$

Détection du cluster

On retient le cluster le plus vraisemblable, i.e. qui correspond à la statistique

$$\max_z \{lnL_z / lnL_0\}$$

La rapport est décroissant avec la variance commune σ_z^2, et le cluster correspond donc à la fenêtre qui minimise cette variance.

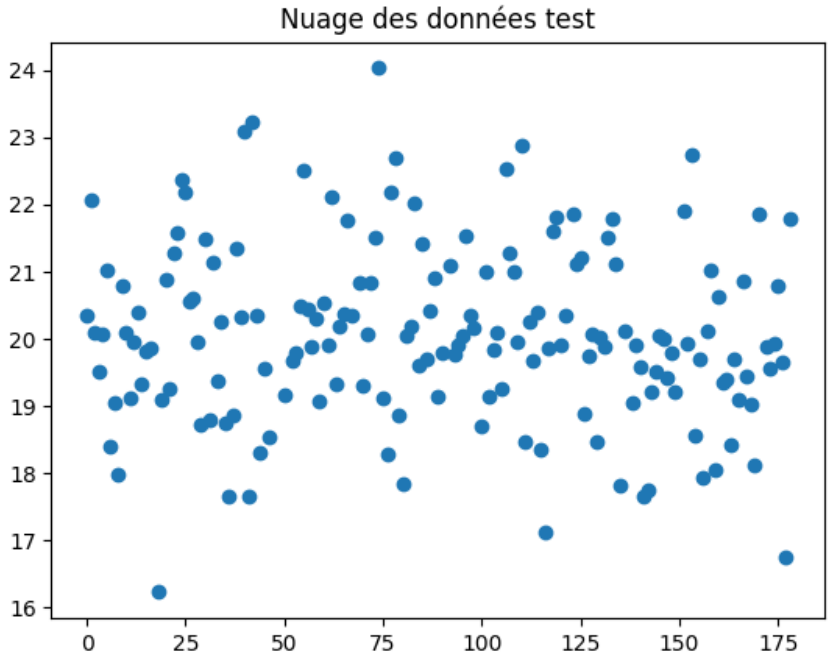
Significativité du cluster

Pour déterminer la significativité du cluster détecté, nous utilisons le test de Monte Carlo : le test consiste à rebattre les positions des N observations un nombre défini M de fois, et retenir la statistique $\max\{z : lnL_z/lnL_0\}$ du cluster le plus vraisemblable à chaque fois.
On note R le rang de la statistique obtenue sur les données d'origines parmi les $M+1$ statistiques calculées au total.
La p-value vaut alors

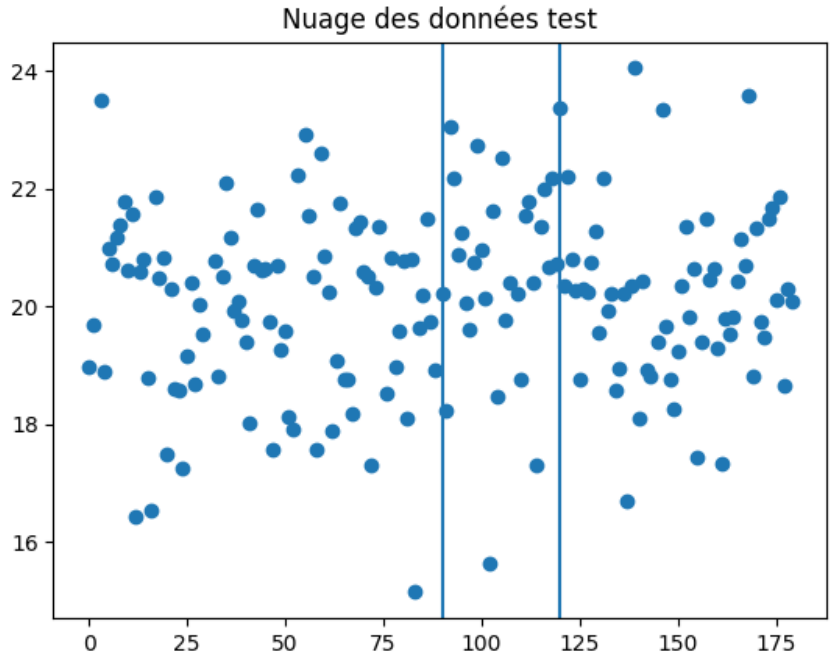
$$P = \frac{R}{M+1}$$

Application à des données test

Soient $N = 500$, $S = 180$. Soient $x_1, \ldots, x_N$ une observation d'un n-échantillon normal de moyenne 20 et de variance 4. Dans une première expérience, on ne crée pas de cluster anomal. Dans une seconde expérience, on crée un cluster où les données sont décalées de 0,75 vers le haut :



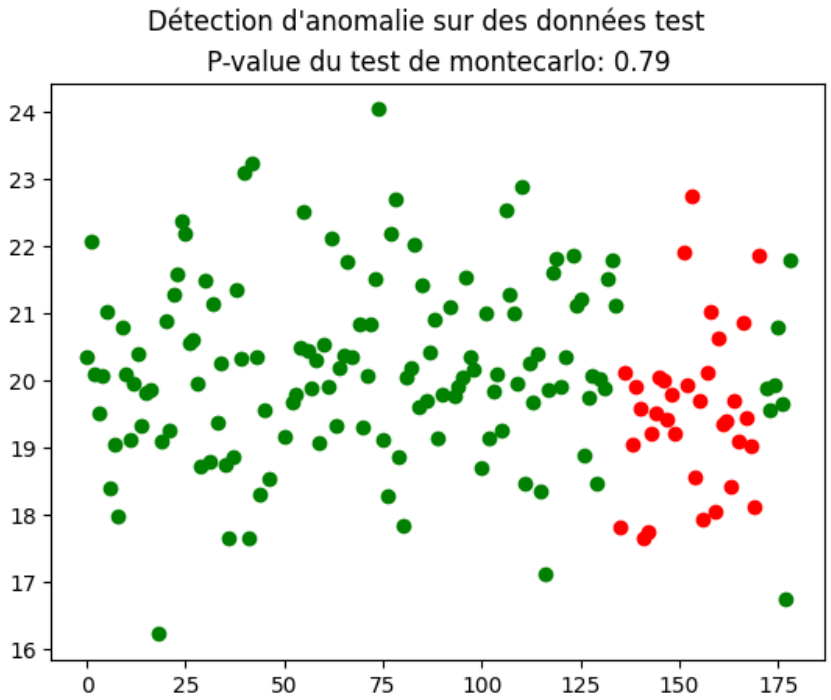
(a) Données sans anomalie



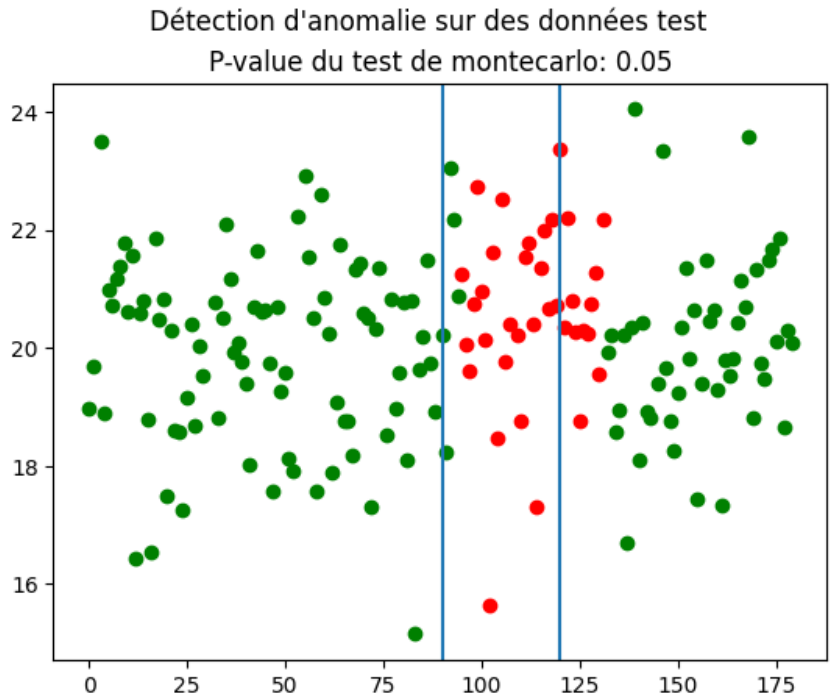
(b) Données avec anomalie

Application à des données tests

Avec une largeur de fenêtre de $180/5$ et $M = 99$, les résultats sont les suivants



(a) Données sans anomalie



(b) Données avec anomalie

Application aux données simulées

Nous considérons un jeu de données de coûts quotidiens associés à un incident technique. Ces données sont simulées avec un cluster anomal.

opening_date_or	daily_cost	daily_counts
2023-02-02	21.583006	9
2023-02-07	44.645135	5
2023-03-18	31.386316	6
2023-04-20	839.625340	6
2023-05-31	52.490612	1

La date correspond aux points s, le coût quotidien aux x_s, et les nombres d'incidents aux n_s.

Application aux données simulées

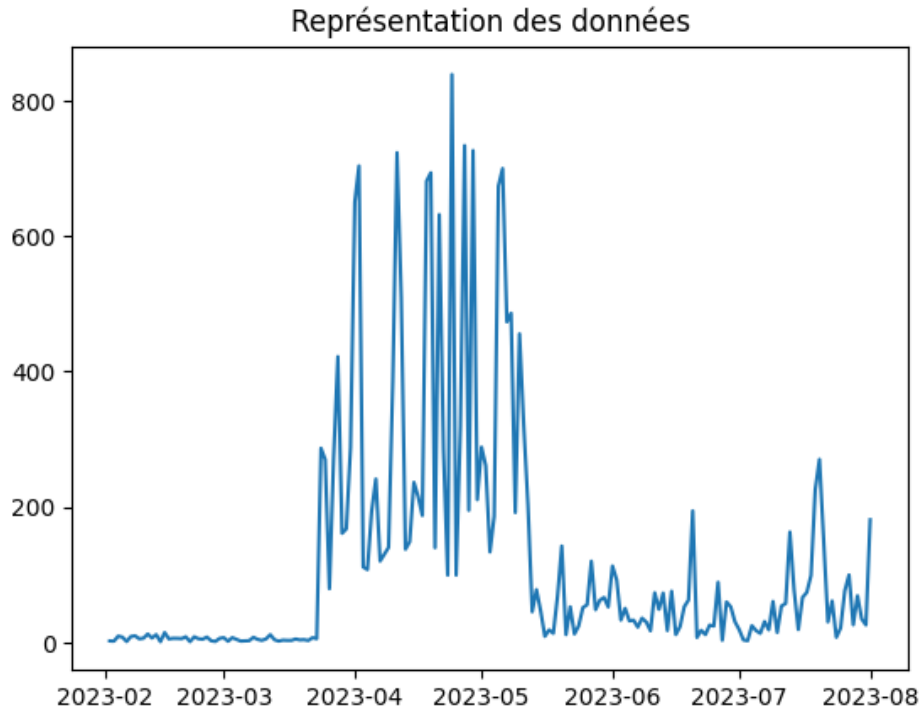


Figure – Coûts moyens quotidiens associées à un certain incident technique

Application aux données simulées

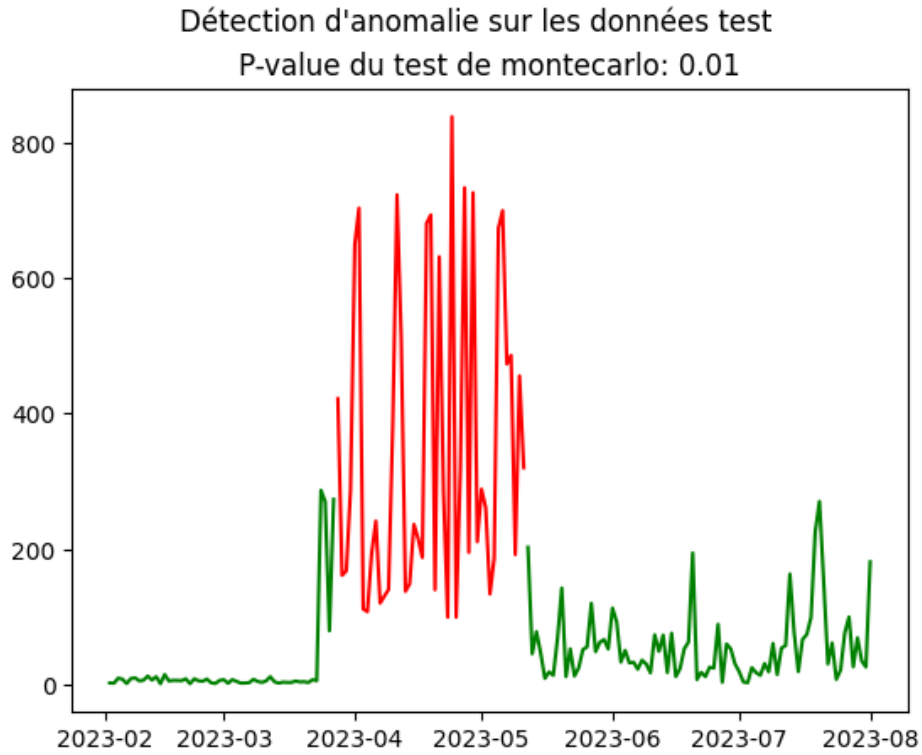


Figure – Cluster et p-value pour une fenêtre de largeur $S/4$ et $M = 99$

Bibliographie

M. Kulldorff, L. Huang and K. Konty, *"A scan statistic for continuous data based on the normal probability model"*, *International Journal of Health Geographics*, number 8 :58, October 2009.