Analyse statistique et modélisation prédictive des éboulements côtiers

L'étude vise à comprendre et prédire le nombre d'éboulements sur une zone côtière en fonction d'indicateurs météorologiques, océaniques et atmosphériques. Les données couvrent 5 périodes de durées variables (2001-2008, 2009-2012, 2013-2015, 2016-2019, 2020-2022).

Objectifs :

1. Identifier quels indicateurs expliquent le mieux les éboulements
2. Prédire le taux annuel d'éboulements pour de nouvelles périodes

PARTIE 1 : Sélection des variables explicatives

Normalisation des données

Problème : Les périodes ont des durées différentes (3 à 8 ans). Les variables de comptage (jours de gel, nombre de tempêtes) augmentent avec la durée.

Solution : Normalisation par la durée

- Variables comptages → divisées par la durée (ex: `jours_gel_par_an`)
- Variables moyennes/max → conservées telles quelles

Cela évite des corrélations artificielles dues à la durée.

Métriques calculées pour chaque indicateur

1. Corrélation de Pearson

$$r = \Sigma(xi - \bar{x})(yi - \bar{y}) / \sqrt{[\Sigma(xi - \bar{x})^2 \times \Sigma(yi - \bar{y})^2]}$$

- Mesure la **relation linéaire** entre l'indicateur et le taux d'éboulements
- Valeur entre -1 et +1
- Plus |r| est proche de 1, plus la relation est forte

Attention : Avec n=5 périodes, les corrélations sont très instables. Même r=0.88 n'est pas significatif (seuil critique à 5% = 0.878).

2. Corrélation de Spearman

$\rho = \text{corr}(\text{rang}(X), \text{rang}(Y))$

- Mesure la **relation monotone** (basée sur les rangs)
- Plus robuste aux valeurs extrêmes que Pearson
- Détecte les relations non-linéaires

3. Modèle de Poisson univarié

Pour chaque indicateur, on ajuste :

$\log(E[Y]) = \beta_0 + \beta_1 X$

Coefficient β_1 :

- $\beta_1 > 0$: l'indicateur augmente les éboulements
- $\beta_1 < 0$: l'indicateur diminue les éboulements

IRR (Incidence Rate Ratio) :

$IRR = \exp(\beta_1)$

- IRR = 1.10 → +10% d'éboulements quand l'indicateur augmente de 1 unité
- IRR = 0.90 → -10% d'éboulements

4. Pseudo-R^2 (McFadden)

$R^2 = 1 - (\text{déviance modèle} / \text{déviance modèle nul})$

- Mesure le **pouvoir explicatif** de l'indicateur
- Varie de 0 à 1
- Interprétation (indicative) :
    - 0.20 → bon
    - 0.40 → très bon
    - 0.70+ → excellent (rare avec peu de données)

Résultats : Top 15 des variables

Les 15 meilleurs indicateurs identifiés (par score combiné corrélation + pseudo-R^2) :

1. nb_jours_pmermin
2. rafale_max_kmh
3. nb_seq_depression_3j
4. jours_vent_fort_60
5. plus_longue_serie_tempete_consecutive
6. t02_max (période de houle)
7. marnage_std_m
8. energie_vent_cumulee
9. nb_seq_seche_10j
10. pression_min_hpa
11. nb_jours_vent_dir_ouest
12. t02_moy
13. ifm (indice de fort marnage)
14. jours_pluie
15. nb_combinaisons_critiques

Observation importante : Après normalisation, les corrélations sont raisonnables (0.5-0.7), contrairement aux valeurs suspectes (>0.92) obtenues sans normalisation. Cela confirme l'importance de corriger l'effet durée.

PARTIE 2 : Modélisation prédictive avec LOOCV

Pourquoi éviter l'overfitting ?

Avec seulement 5 périodes et 15 variables potentielles, le risque d'overfitting est majeur. Un modèle trop complexe "apprend par cœur" les données au lieu de capturer les vraies relations, ce qui donne des prédictions catastrophiques sur de nouvelles données.

Trois approches testées

Approche 1 : LASSO (régularisation L1)

- Pénalise les coefficients pour en forcer certains à zéro
- Fait automatiquement de la sélection de variables
- Le paramètre λ (intensité de pénalisation) est optimisé par validation croisée

Approche 2 : Modèle simple (2 variables maximum)

- Sélection de 1-2 variables non corrélées entre elles

- Respecte la règle empirique n/3 (5 périodes → max 1-2 variables)
- Privilégie la parcimonie et l'interprétabilité

Approche 3 : Score composite

- Standardisation et moyenne des meilleures variables
- Réduit à un seul indicateur synthétique
- Combine l'information de plusieurs indicateurs sans surparamétrer

Validation croisée Leave-One-Out (LOOCV)

Protocole pour évaluer la capacité prédictive :

1. Retirer une période (test)
2. Entraîner le modèle sur les 4 autres
3. Prédire le taux d'éboulements de la période test
4. Répéter pour les 5 périodes
5. Comparer prédictions vs observations

Métriques d'évaluation

Pour la prédiction du taux :

- **MAE** (Mean Absolute Error) : erreur moyenne en éboulements/an
- **Erreur relative** : erreur en % du taux observé (< 25% = acceptable avec n=5)
- **R^2** : proportion de variance expliquée (attention : $R^2 > 0.9$ avec n=5 = signe d'overfitting)

Pour la classification (faible/moyen/fort) :

- **Accuracy** : proportion de périodes bien classées
- **Matrice de confusion** : tableau croisant classes observées et prédites

Résultats comparatifs

Modèle	MAE	Erreur relative	R^2	Accuracy
LASSO	6.4	23.5%	0.72	40%
Simple (2 var)	12.0	45.7%	0.90	20%

Composite	6.7	23.8%	0.79	0%

Interprétation :

- LASSO et Composite ont des performances équivalentes (~24% d'erreur)
- Le modèle Simple a un R^2 élevé (0.90) mais une erreur double → c'est de l'overfitting
- L'accuracy de classification est très faible (0-40%) → impossible de classer fiablement en 3 catégories avec 5 périodes

Modèle recommandé : Score Composite

Le score composite est privilégié pour :

- Performance équivalente au LASSO (MAE = 6.7 éboulements/an)
- Interprétabilité supérieure (un seul indicateur synthétique)
- Robustesse conceptuelle (combine plusieurs dimensions météo/marines)
- Risque d'overfitting minimal

Conclusions générales

Les indicateurs météorologiques et marins permettent d'estimer l'ordre de grandeur du nombre d'éboulements avec une erreur relative d'environ 24%. Les périodes sont globalement bien positionnées (prédictions rarement totalement aberrantes), ce qui montre que les processus physiques capturés par les indicateurs ont bien un lien avec l'érosion.

Les variables les plus explicatives sont liées :

- Aux tempêtes et au vent (rafales, séquences de tempêtes, énergie cumulée)
- Aux dépressions atmosphériques (basse pression, séquences de dépression)
- Aux conditions marines (houle, marnage)
- Aux cycles gel-dégel et précipitations

Cela correspond bien aux processus physiques d'érosion côtière connus : l'action des vagues lors des tempêtes, la fragilisation par cycles gel-dégel, et l'influence des marées.

Limites identifiées

Impossibilité de classification fine : Avec seulement 5 périodes, classifier en 3 catégories (faible/moyen/fort) est trop ambitieux. L'accuracy de 0-40% montre que ce n'est pas atteignable.

Corrélations élevées suspectes : Les corrélations initiales >0.92 observées avant normalisation étaient dus à l'effet durée et au faible nombre d'observations. Après correction, les corrélations deviennent raisonnables (0.5-0.7).

Incertitude structurelle avec n=5 : Toute analyse statistique atteint ses limites avec 5 points. Les intervalles de confiance sont très larges, et la généralisation à de nouvelles périodes reste incertaine. Les résultats doivent être interprétés comme des indicateurs qualitatifs de risque plutôt que comme des prédictions quantitatives précises.

Processus non capturés : Les modèles ne prennent pas en compte l'évolution morphologique de la falaise, les propriétés géotechniques locales, ou les interactions non-linéaires entre variables (seuils d'érosion, effets synergiques).

Recommandations méthodologiques

Pour améliorer ces modèles :

- Augmenter le nombre d'observations (subdiviser en années si possible, tout en gérant l'autocorrélation)
- Se limiter à une classification binaire plutôt que ternaire
- Intégrer des variables géomorphologiques et géotechniques
- Explorer les seuils critiques et interactions non-linéaires
- Considérer l'aspect temporel (dépendance entre périodes successives)
-

Synthèse finale

Cette étude démontre qu'il est possible d'expliquer une partie significative de la variabilité des éboulements côtiers à partir d'indicateurs météo-marins, malgré le nombre très limité de périodes observées.

Les modèles développés permettent d'estimer le taux annuel d'éboulements avec une erreur de l'ordre de 24%, ce qui est acceptable dans ce contexte de forte incertitude. Toutefois, la classification fine en catégories de risque reste hors de portée avec si peu de données.

Le principal apport de cette analyse est d'avoir identifié et quantifié les indicateurs les plus pertinents pour expliquer les éboulements, tout en mettant en lumière les limites méthodologiques inhérentes à un petit échantillon et l'importance cruciale de la normalisation des variables pour éviter les biais.