Data Science

Marion Favrot, Adélaïde Branger, Léa Cercleron et Hasfa Yassir

November 2025

1 Synthèse de la méthodologie de référence

1.1 Problématique et objectifs de l'article

Pour ce projet, nous étudions l'article "SMOTE: Synthetic Minority Over-sampling Technique. (Chawla, N. V., et al., 2002)". L'article aborde la construction de classificateurs à partir d'ensembles de données déséquilibrés, c'est-à-dire lorsque certaines classes sont largement sous-représentées par rapport aux classes majoritaires. Ce déséquilibre peut atteindre des ratios extrêmes, allant de 100 pour 1 et/ou 100 000 pour 1 selon les applications. Cela est fréquent dans de nombreux domaines tels que la détection de fraude ou encore la détection médicale. Les algorithmes de classifications traditionnels privilégient la précision globale au détriment de la détection des classes minoritaires. Or, ces dernières sont souvent les plus importantes à identifier. Dans ces applications, les erreurs de type « faux négatif » sont souvent bien plus coûteuses que les erreurs de type « faux positif ». Par exemple, pour des données mammographiques, prédire correctement 98% des cas « normaux » peut sembler efficace, mais cela implique de rater la majorité des cas « anormaux » qui sont plus onéreux.

L'objectif de l'article est donc de proposer une méthode de sur-échantillonnage de la classe minoritaire, combinée éventuellement à un sous-échantillonnage de la classe majoritaire, afin d'améliorer la capacité d'un classificateur à détecter la classe rare (minoritaire) sans trop sacrifier la performance globale. Plus précisément, la méthode nommée SMOTE ("Synthetic Minority Over-sampling Technique") est proposée pour créer de nouveaux exemples synthétiques de la classe minoritaire dans l'espace des "features" (des attributs, caractéristiques) plutôt que par simple réplication.

1.2 Description formelle de la méthode proposée

SMOTE crée de nouveaux exemples minoritaires en interpolant linéairement entre un point minoritaire et certains de ses voisins proches. Cette méthode vise à élargir la région de décision associée à la classe minoritaire, réduire le surapprentissage par rapport à la duplication simple, améliorer la détection des cas rares par les classificateurs et accroître la performance globale mesurée par ROC (Receiver Operating Characteristic), l'aire sous la courbe AUC et ROC convex hull (identification des classificateurs optimaux).

Les étapes de la méthode sont les suivantes :

1. Pour chaque observation minoritaire x_i, on identifie ses k plus proches voisins ($k = 5$ par défaut).
2. Selon le taux d'oversampling souhaité, on sélectionne aléatoirement certains de ces voisins.
3. Pour chaque voisin x_{nn} choisi, on génère un exemple synthétique :

$$x_{new} = x_i + \delta \cdot (x_{nn} - x_i),$$

où δ est tiré uniformément dans $[0, 1]$.

Ainsi, chaque nouvel exemple est placé sur le segment reliant x_i et x_{nn}.

De manière algorithmique, pour utiliser la méthode SMOTHE sur des données, nous utilisons la version suivante :

Algorithm SMOTE(T, N, k)*:*
Input: Number of minority class samples T*; Amount of SMOTE* N*; Number of nearest neighbors* k
Output: $(N/100) \times T$ *synthetic minority class samples*

for $i \leftarrow 1$ *to* T
  Compute k *nearest neighbors for* i
  $Populate(N, i, nnarray)$
endfor

Le premier point clé à retenir est que SMOTE utilise la distance euclidienne pour identifier les voisins. Ensuite, les points synthétiques sont créés uniquement à partir des observations minoritaires. Le nombre de points générés dépend du pourcentage d'oversampling choisi. De plus, il est recommandé de combiner SMOTE avec un sous-échantillonnage modéré de la classe majoritaire pour inverser le biais initial. Le sur-échantillonnage par duplication rétrécit la région minoritaire, entraînant un sur-ajustement et donc un sur-apprentissage. SMOTE élargit cette région en créant des points intermédiaires, améliorant ainsi la généralisation.

Par conséquent, les arbres décisionnels induits à partir de données SMOTE adoptent des subdivisions plus globales et moins sensibles aux cas marginaux.

1.3 Résultats et conclusions des auteurs

Les auteurs ont testé SMOTE sur 9 jeux de données variés (Pima, Phoneme, Satimage, Forest Cover, Oil, Mammography, etc.) avec trois classificateurs C4.5 (arbres de décision), Ripper (méthode par règles) et Naïve Bayes. Ils ont comparé différentes stratégies : le sous-échantillonnage seul pour C4.5, le SMOTE + sous-échantillonnage pour C4.5 puis la modification des coûts dans le cas de Ripper et la modification des probabilités a priori dans le cas de Naïve Bayes.

Les résultats obtenus par les auteurs informent principalement sur plusieurs points. Tout d'abord, l'utilisation conjointe de SMOTE et du sous-échantillonnage permet d'améliorer la détection de la classe minoritaire. Puis, SMOTE permet de générer un plus grand nombre de classificateurs considérés comme « potentiellement optimaux » sur le ROC convex hull. De plus, cette méthode surpasse les approches reposant sur les coûts ou les probabilités a priori. Enfin, SMOTE permet d'obtenir des frontières de décision plus larges et plus robustes, contrairement à la simple duplication qui fragmente l'espace des données. Pour finir, sur un total de 48 expériences menées, le classificateur basé sur SMOTE ne s'est révélé non optimal que dans 4 cas.

Malheureusement, cette méthode a ses limites dont certaines sont soulignées par les auteurs. La création d'exemples synthétiques peut engendrer une augmentation des faux positifs, notamment lorsque les variables présentent une grande variance. Le choix du nombre de voisins (k) et du taux d'over-sampling est très important. Un paramétrage trop fort peut engendrer un apprentissage trop général et dégrader la performance du modèle. Ils proposent également une sélection adaptative du nombre de voisins et une automatisation du choix du nombre des plus proches voisins. L'exploration d'éventuelles méthodes alternatives plus sophistiquées pour créer des exemples synthétiques pourrait aider à améliorer la performance. Nous pourrions cibler des voisins, c'est-à-dire prioriser les voisins issus d'exemples mal classés pour renforcer la qualité des zones de décision. Pour réaliser une gestion du voisinage mixte, nous pouvons prendre en compte le cas où un exemple minoritaire est entouré de majoritaires, ce qui peut influencer la redéfinition des frontières de décision.

Plusieurs extensions de cette méthode SMOTE sont présentés par les auteurs : SMOTE-NC (pour données mixtes) et SMOTE-N (pour données nominales). Certaines bases de données testées (ex : Adult) combinaient variables catégorielles et forte variance nécessitant donc SMOTE-NC.

Pour conclure, SMOTE constitue une avancée clé pour les jeux de données déséquilibrés. Le sur-échantillonnage synthétique élargit les régions de décision de la classe minoritaire et améliore sa détection. Combiné à un sous-échantillonnage modéré, il surpasse les méthodes traditionnelles et s'impose comme un outil de référence pour la classification déséquilibrée.

2 Protocole expérimental

2.1 Présentation du cas d'usage actuariel et du jeu de données

Pour utiliser cette méthode dans le cadre de ce projet, nous avons sélectionné deux bases de données issues de Kaggle:

1. *soil_ data.csv* : Ce jeu contient des informations sur le sol et divers indices météorologiques pour différents États américains. La variable cible est le niveau de sécheresse, codé de 0 à 5, correspondant à une intensité croissante de la sécheresse. Ce jeu permet d'identifier l'état du sol et les facteurs environnementaux influençant la sécheresse.

2. *train_ timeseries.csv* : Ce second jeu contient des séries temporelles plus détaillées sur les mêmes régions et périodes, incluant des informations météorologiques supplémentaires (température, précipitations, humidité, etc.). L'objectif est de lier ces informations pour expliquer plus précisément les causes de la sécheresse.

Celles-ci regroupent des informations provenant de plusieurs sources officielles, notamment :

1. Le NASA POWER Project, développé par le NASA Langley Research Center(LaRC) dans le cadre du programme NASA Earth Science/Applied Science.

2. Le U.S. Drought Monitor, un projet collaboratif impliquant le National Drought Mitigation Center (Université du Nebraska-Lincoln), le United States Department of Agriculture (USDA) et la National Oceanic and Atmospheric Administration (NOAA).

Le jeu de données concaténé est structuré sous la forme d'un problème de classification avec six niveaux de sécheresse (0 pour une absence de sécheresse, 1 pour un niveau modéré, et ainsi de suite). Chaque observation correspond au niveau de sécheresse d'un comté américain à une date donnée, accompagné des données de 18 indicateurs météorologiques sur les 90 derniers jours. Parmi ces variables, nous trouvons par exemple la température de surface terrestre en°C (TS) ou encore la vitesse minimale du vent à 10 mètres en m/s (WS10M-MIN). Les États, comtés et autres subdivisions administratives des États-Unis sont désignés par leur code FIPS. Cet identifiant numérique sert principalement à la collecte, au traitement et à l'analyse des données gouvernementales et statistiques. Par exemple, le comté de Travis au Texas a le code FIPS 48453 (48 pour l'État du Texas + 453 pour le comté).

Dans un contexte actuariel, ce cas d'usage vise à modéliser et anticiper les épisodes de sécheresse afin d'évaluer leur impact sur la sinistralité des portefeuilles d'assurance (notamment en assurance agricole), d'améliorer la tarification du risque climatique et de renforcer les stratégies de prévention et de gestion des risques extrêmes.

Dans les sections suivantes, nous détaillerons les résultats des modèles d'apprentissage automatique utilisées pour modéliser et prédire ces événements extrêmes.

2.2 Justification de la pertinence de la méthode pour ce cas d'usage

Le jeu de données utilisé présente un déséquilibre prononcé entre les classes, directement lié à la nature du phénomène étudié. La majorité des observations correspond à une absence de sécheresse (niveau 0), tandis que les niveaux élevés de sécheresse (4 et 5) sont rares. Ce déséquilibre est encore renforcé lorsque la variable cible est transformée en une variable binaire, prenant la valeur 0 en l'absence de sécheresse et 1 lorsqu'un épisode de sécheresse est observé. Dans cette configuration, la classe « sécheresse » devient fortement minoritaire.

Or, les variables explicatives du jeu de données sont principalement continues et météorologiques, agrégées sur une fenêtre temporelle de 90 jours. Cette structure rend la méthode SMOTE particulièrement adaptée, car la génération d'exemples synthétiques par interpolation linéaire entre observations minoritaires produit des profils météorologiques réalistes et cohérents d'un point de vue physique.

L'utilisation de SMOTE permet ainsi de densifier l'espace des observations correspondant aux épisodes de sécheresse, qui est initialement peu représenté. Sans ré-équilibrage, un classificateur entraîné sur ces données tendrait à apprendre principalement les configurations météorologiques associées à l'absence de sécheresse, au détriment des situations menant à des événements de sécheresse. SMOTE contribue à élargir

la région de décision associée à ces épisodes rares, en créant des situations intermédiaires entre différents cas de sécheresse observés dans les données.

Dans le contexte actuariel de l'étude, cette approche est d'autant plus pertinente que l'objectif est d'estimer la fréquence des événements de sécheresse afin d'anticiper leurs impacts économiques et agricoles. Une mauvaise détection des épisodes de sécheresse dans le jeu d'apprentissage conduirait à une sous-estimation du risque. En renforçant la représentation de la classe minoritaire, SMOTE améliore la capacité du modèle à identifier les configurations météorologiques annonciatrices de sécheresse.

2.3 Description succincte de la méthodologie de prétraitement et de modélisation appliquée

La méthodologie mise en œuvre dans ce projet suit une chaîne de traitement classique en apprentissage supervisé, adaptée aux spécificités des données météorologiques et au fort déséquilibre des classes. Elle se décompose en plusieurs étapes successives : préparation des données, ré-équilibrage des classes, apprentissage des modèles et évaluation des performances.

Dans un premier temps, les deux jeux de données (*soil_ data.csv* et *train_ timeseries.csv*) sont fusionnés à l'aide des identifiants communs (code FIPS et date). Cette étape permet d'associer, pour chaque comté et chaque date, les indicateurs de sol avec les informations météorologiques agrégées sur les 90 jours précédents. Les observations incomplètes ou présentant des valeurs manquantes sur les variables explicatives sont supprimées afin de garantir la cohérence du jeu d'apprentissage. Les variables explicatives, essentiellement continues, sont ensuite standardisées (centrage-réduction) afin d'éviter que certaines grandeurs physiques (par exemple la température ou les précipitations) ne dominent artificiellement les calculs de distance utilisés par SMOTE et par certains classificateurs. La variable cible est ensuite transformée en une variable binaire indiquant l'absence de sécheresse (classe 0) ou la présence d'un épisode de sécheresse (classe 1). Cette transformation accentue le déséquilibre initial mais correspond mieux à l'objectif actuariel de détection des événements extrêmes. Le jeu de données est alors séparé en un échantillon d'apprentissage et un échantillon de test, la séparation étant effectuée avant toute opération de sur-échantillonnage afin d'éviter toute fuite d'information.

La méthode SMOTE est appliquée exclusivement sur l'échantillon d'apprentissage. Pour chaque observation appartenant à la classe minoritaire (sécheresse), les k plus proches voisins sont identifiés à l'aide de la distance euclidienne dans l'espace des variables standardisées. Des observations synthétiques sont ensuite générées par interpolation linéaire entre les observations minoritaires existantes et leurs voisins, selon un taux de sur-échantillonnage prédéfini. Cette étape permet de rééquilibrer le jeu d'apprentissage tout en conservant des profils météorologiques réalistes. Dans certains cas, SMOTE est combiné à un sous-échantillonnage modéré de la classe majoritaire afin de limiter la taille du jeu de données et de réduire le biais en faveur de la classe dominante.

Une fois le jeu d'apprentissage rééquilibré, plusieurs modèles de classification sont entraînés afin d'évaluer l'impact de SMOTE sur les performances prédictives. Les algorithmes testés incluent des méthodes classiques d'apprentissage supervisé adaptées aux données tabulaires, telles que les arbres de décision, les forêts aléatoires ou les modèles de régression logistique. Les hyperparamètres des modèles sont sélectionnés par validation croisée sur l'échantillon d'apprentissage rééquilibré.

Enfin, l'évaluation des performances est réalisée sur l'échantillon de test non modifié, afin de mesurer la capacité réelle de la généralisation des modèles. Les métriques retenues ne se limitent pas à la précision globale, mais incluent des indicateurs adaptés aux données déséquilibrées tels que le rappel de la classe minoritaire, la courbe ROC et l'aire sous la courbe (AUC). Cette approche permet d'évaluer précisément l'apport de SMOTE en termes de détection des épisodes de sécheresse, tout en contrôlant l'augmentation potentielle des faux positifs.

3 Analyse des résultats

3.1 Présentation quantitative et qualitative des résultats obtenus

Comparaison des performances des modèles entraînés **avec** et **sans** application de SMOTE :

– Précision globale (accuracy) : généralement peu discriminante dans un contexte fortement déséquilibré

– Rappel de la sensibilité de la classe « sécheresse » : indicateur central pour évaluer la capacité de détection des événements rares

– Précision de la classe minoritaire : mesure de l'augmentation potentielle des faux positifs induite par SMOTE

– Score F1 et aire sous la courbe ROC (AUC) : métriques synthétiques permettant une comparaison globale

Observation attendue :

– Amélioration significative du rappel de la classe minoritaire après application de SMOTE

– Dégradation modérée de la précision globale ou augmentation contrôlée des faux positifs

– Déplacement de la courbe ROC vers le coin supérieur gauche, traduisant une meilleure séparation des classes

Analyse qualitative :

– Les modèles entraînés avec SMOTE semblent mieux capturer la diversité des configurations météorologiques associées aux épisodes de sécheresse

– Les frontières de décision apparaissent plus larges et moins sensibles aux observations extrêmes isolées

À confirmer : Sensibilité des résultats au taux d'over-sampling et au paramètre k de SMOTE

3.2 Interprétation des résultats

Dans le contexte étudié, l'amélioration du rappel de la classe « sécheresse » est cohérente avec l'objectif actuariel de détection du risque, même au prix d'une augmentation limitée des faux positifs. L'effet positif de SMOTE est renforcé par la nature des variables explicatives :

– variables continues,

– relations physiques plausibles entre indicateurs météorologiques,

– continuité temporelle implicite via les agrégations sur 90 jours.

Interprétation métier :

– Une meilleure détection des épisodes de sécheresse permet d'anticiper plus efficacement les risques agricoles et économiques

– Le modèle devient plus conservateur, ce qui est généralement souhaitable dans une logique de gestion des risques

Point de vigilance :

– Une amélioration du rappel ne garantit pas nécessairement une meilleure estimation des probabilités

– Les modèles peuvent devenir trop sensibles à certaines configurations météorologiques intermédiaires générées artificiellement

4 Conclusion et perspectives

4.1 Analyse critique des limites de l'approche dans le contexte appliqué

Limites liées aux données :

– Agrégation temporelle sur 90 jours pouvant lisser certains signaux de court terme

– Possible hétérogénéité spatiale entre comtés non explicitement modélisée

– Transformation binaire de la cible entraînant une perte d'information sur l'intensité de la sécheresse

Limites méthodologiques :

– Absence de prise en compte explicite des dépendances temporelles (approche non séquentielle)

– Validation réalisée sur une seule partition entraînement/test (à confirmer)

4.2 Identification de pistes d'amélioration ou d'extensions futures

Améliorations de la modélisation :

- Passage à une classification multi-classes pour prédire le niveau exact de sécheresse
- Intégration de modèles séquentiels (LSTM, modèles à états cachés) pour exploiter pleinement la dimension temporelle

Perspectives actuarielles :

- Couplage du modèle de classification avec une modélisation de la sévérité des impacts économiques
- Utilisation des probabilités prédites pour la tarification ou la gestion de portefeuille agricole

4.3 Référencement d'autres articles pour étayer l'analyse

- Chawla, N. V., et al. (2002) — *SMOTE: Synthetic Minority Over-sampling Technique*
- He, H., Garcia, E. A. (2009) — *Learning from imbalanced data*
- Han, H., Wang, W.-Y., Mao, B.-H. (2005) — *Borderline-SMOTE*
- Batista, G. E. A. P. A., et al. (2004) — *A study of the behavior of several methods for balancing machine learning training data*
- Fernández, A., et al. (2018) — *Learning from imbalanced data sets*