THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale I2S

UMR TETIS

Mise en relation de données hétérogènes pour le renforcement des systèmes de sécurité alimentaire - Cas de la production agricole en Afrique de l'Ouest

Présentée par Hugo DELÉGLISE

Le 15 décembre 2021

Sous la direction de Agnès Bégué, Roberto Interdonato, Élodie Maître d'Hôtel, Mathieu Roche et Maguelonne Teisseire

Devant le jury composé de

Josiane Mothe, Professeur, Université de Toulouse	Rapporteure
Stan Matwin, Professeur, Dalhousie University, Canada	Rapporteur
Pierre Gançarski, Professeur, Université de Strasbourg	Examinateur
Mathieu Roche, Chercheur HDR, CIRAD - UMR TETIS	Directeur
Maguelonne Teisseire, Directrice de Recherche, INRAE - UMR TETIS	Co-Directrice
Roberto Interdonato, Chercheur, CIRAD - UMR TETIS	Encadrant principal
Elodie Maître d'Hôtel, Chercheuse HDR, CIRAD - UMR MOISA	Encadrante, invitée
Isabelle Mougenot, Maître de conférences HDR, Université de Montpellier	Invitée

UNIVERSITÉ DE MONTPELLIER

Résumé

En français :

Les progrès dans la lutte contre la faim ont été significatifs en Afrique de l'Ouest et au Burkina Faso entre 2000 et 2014, avant que la situation alimentaire ne se détériore à nouveau. Les raisons sont multiples et interdépendantes : des phénomènes météorologiques extrêmes plus fréquents et l'augmentation de la population tendent à réduire la disponibilité alimentaire par habitant ; les déplacements de population dus aux conflits ont pour conséquence la chute de la production agricole et la désorganisation des circuits de distribution ; la pauvreté structurelle des populations est aggravée par un contexte économique mondial difficile. Pour suivre, analyser et prévoir les situations d'insécurité alimentaire aux échelles locales à subnationales, les systèmes de sécurité alimentaire (SSA) intègrent principalement des données agroclimatiques issues d'images satellites et des indicateurs de nutrition, de production et d'économie issus d'enquêtes ménages. Ces enquêtes sont essentielles à la production d'indicateurs clés pour mesurer la sécurité alimentaire (SA), mais sont coûteuses économiquement et en temps.

L'objectif de cette thèse est de fournir des approches innovantes pour l'estimation d'indicateurs de SA et de leurs déterminants, en utilisant des données hétérogènes publiquement accessibles et des approches fondées sur l'intelligence artificielle, dans la perspective d'appuyer les méthodes utilisées par les SSA. Pour cela, plusieurs questions de recherche sont traitées : sur quels indicateurs s'appuyer pour mesurer la SA et quelles en sont les limites ? Comment traiter l'hétérogénéité thématique, temporelle et spatiale des données ? Comment extraire des données leur aspect explicatif, i.e., être capable d'identifier les facteurs d'augmentation ou de diminution de la SA à partir des données ? Pour répondre à ces problématiques, cette thèse propose trois contributions.

Premièrement, nous faisons un état des lieux des nombreux indicateurs utilisés pour quantifier cette notion complexe qu'est la SA. Puis, nous nous concentrons sur des indicateurs de SA issus d'enquêtes ménages (i.e., le score de consommation alimentaire, le score de diversité alimentaire des ménages et l'indice des stratégies de survie réduit) et étudions ce qu'ils nous révèlent sur la SA, leur validité spatiale

et temporelle, ainsi que les biais auxquels ils peuvent être sujets. Nous montrons que malgré leurs biais inhérents, ces indicateurs contiennent des informations spatiales et interannuelles cohérentes qui peuvent être exploitées pour le suivi des crises alimentaires au niveau sub-national.

Deuxièmement, nous proposons des approches originales combinant des méthodes d'apprentissage automatique et profond (i.e., forêts aléatoires, réseaux de neurones convolutifs, réseaux de neurones récurrents à mémoire court-terme et long terme) pour obtenir des approximations d'indicateurs de SA issus d'enquêtes ménages. Ces approches intègrent et combinent des données explicatives hétérogènes. Les données explicatives sont des variables quantitatives (e.g., données météorologiques), des images (e.g., densités de population, occupation des sols), des points GPS (e.g., hôpitaux, écoles, événements violents) et des vecteurs (cours d'eau) avec différentes granularités temporelles et spatiales (e.g., séries temporelles, images à haute résolution spatiale). Nous mettons en évidence la pertinence des approches d'apprentissage automatique selon les données à traiter et constatons l'apport significatif de variables issues de nombreux domaines dans une approche globale.

Troisièmement, nous étudions l'apport des données textuelles, possédant un fort potentiel explicatif, pour effectuer une analyse qualitative de la SA en nous basant sur un corpus de journaux burkinabés. Nous examinons la capacité des méthodes de fouille de textes à extraire automatiquement des informations qualitatives sur la situation alimentaire globale, régionale et annuelle à partir de ce corpus. Ce travail a permis d'obtenir des informations qualitatives spécifiques sur la thématique de la SA et sur ses caractéristiques spatiale et temporelle.

A travers ces trois contributions, cette thèse considère la problématique de l'hétérogénéité des données liées à la SA en mettant l'accent sur les dimensions spatio-temporelles et thématiques qu'elles véhiculent. Les cadres méthodologiques génériques proposés pourront être étendus et adaptés à d'autres domaines.

En anglais :

Progress in the fight against hunger was significant in West Africa and Burkina Faso between 2000 and 2014, before the food situation deteriorated again. The reasons are multiple and interrelated : more frequent extreme weather events and population growth tend to reduce per capita food availability ; population

displacements due to conflicts result in a drop in agricultural production and the disorganization of distribution channels ; structural poverty of populations is aggravated by a difficult global economic context. To monitor, analyze and forecast food insecurity situations at local to sub-national scales, food security systems (FSS) mainly integrate agro-climatic data from satellite images and nutrition, production and economic indicators from household surveys. These surveys are essential for producing key indicators for measuring food security (FS), but are costly in terms of time and money.

The objective of this thesis is to provide innovative approaches for the estimation of FS indicators and their determinants, using publicly available heterogeneous data and artificial intelligence-based approaches, with a view to supporting the methods used by SSA. To this end, several research questions are addressed : which indicators should be used to measure FS and what are their limitations ? How to deal with the thematic, temporal and spatial heterogeneity of the data ? How to extract the explanatory aspect of the data, i.e., to be able to identify the factors of increase or decrease of FS from the data ? To address these issues, this thesis proposes three contributions.

First, we review the numerous indicators used to quantify the complex notion of FS. Then, we focus on FS indicators derived from household surveys (i.e., the food consumption score, the household dietary diversity score, and the reduced coping strategies index) and study what they tell us about FS, their spatial and temporal validity, and the biases to which they may be subject. We show that despite their inherent biases, these indicators contain consistent spatial and inter-annual information that can be exploited for monitoring food crises at the sub-national level.

Second, we propose novel approaches combining machine and deep learning methods (i.e., random forests, convolutional neural networks, recurrent neural networks with short-term and long-term memory) to approximate FS indicators from household surveys. These approaches integrate and combine heterogeneous explanatory data. The explanatory data are quantitative variables (e.g. World Bank economic variables, weather data), images (e.g. population densities, land use), GPS points (hospitals, schools, violent events, weather stations, markets) and vec-

tors (rivers) with different temporal and spatial granularities (e.g. quantitative data, time series, high spatial resolution images). We highlight the relevance of machine and deep learning approaches depending on the data to be processed and note the significant contribution of variables from many domains in a global approach.

Third, we study the contribution of textual data, with a strong explanatory potential, to perform a qualitative analysis of FS based on a corpus of Burkinabe newspapers. We examine the ability of text mining methods to automatically extract qualitative information on the global, regional and annual food situation from this corpus. This work has provided specific qualitative information on the theme of FS and its spatial and temporal characteristics.

Through these three contributions, this thesis considers the problem of the heterogeneity of data related to FS by focusing on the spatio-temporal and thematic dimensions that they convey. The proposed generic methodological frameworks can be extended and adapted to other domains.

Remerciements

Dans l'inconscient collectif, le thésard se retrouve seul face à ses travaux pendant 3 ans. Cette affirmation s'est, en ce qui me concerne, révélée totalement fausse. Ce projet de recherche a germé, puis s'est épanoui dans l'interaction avec des experts de multiples domaines, et mon entourage m'a donné la force de le faire grandir jusqu'à son terme. Pour ces raisons, j'aimerais adresser plusieurs remerciements.

Aux membres du jury qui ont pris le temps d'étudier cette thèse et pour leurs questions et remarques lors de la soutenance qui ont suscité de nouvelles pistes de réflexion.

A mes encadrants : Agnès, Elodie, Maguelonne, Mathieu, Roberto, à qui je dois beaucoup pour leurs conseils et leurs visions inspirées et inspirantes, et qui par leurs qualités humaines ont su trouver la posture adéquate pour que je puisse m'épanouir dans ce projet.

A tous les collègues de la maison de la télédétection, aussi bien aux amis stagiaires et doctorants qu'aux chercheurs et au personnel administratif, qui contribuent à ce climat sain et agréable, au sein duquel je me suis véritablement senti appartenir.

A l'institut #DigitAg qui a apporté un soutien matériel et humain conséquent et qui a contribué à la qualité des résultats obtenus.

A plusieurs rencontres, sur le parcours qui m'a conduit au début de cette thèse et qui m'ont donné l'envie et les capacités nécessaires pour suivre cette voie. Je

pense à Cédric, Sébastien, Michel, Guillemette et à l'ensemble des enseignants de l'excellent parcours MIASHS de l'Université Paul Valéry de Montpellier.

A mes proches, amis comme famille, dont le soutien a été crucial, avec qui j'ai pu continuer à parler, à rire, à sortir et à avoir une vie à peu près normale, même lorsque le projet devenait très éprouvant.

A mes colocs qui m'ont supporté jusqu'au bout !

Table des matières

Abréviations

ACLED	Armed Conflict Location & Event Data Project
ACP	Analyse par composante principale
BERT	Bidirectional Encoder Representations from Transformers
CBOW	Continuous Bag of Words
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
CFSVA	Enquête d'analyse globale de la sécurité alimentaire et de la vulnérabilité
CGIAR	Consultative Group on International Agricultural Research
CNN	Réseau de neurones convolutif
CS	Conjoncturel et Spatial
DCW	Digital Chart of the World
EFM	Echelle de la faim dans les ménages
EPA	Enquête Permanente Agricole
ESA	Explicit Semantic Analysis
FA	Forêt Aléatoire
FAO	Food and Agriculture Organization
FEWS-NET	Famine Early Warning Systems Network
FSPHD	Food Security Prediction based on Heterogeneous Data
GIEWS	Global Information and Early Warning System
GFSI	Indice global de la sécurité alimentaire
GHI	Indice de la faim dans le monde
HRS	Haute Résolution Spatiale
IMC	Indice de Masse Corporelle
ISAr	Indice des Stratégies d'Adaptation réduit
IPC	Cadre intégré de classification de la sécurité alimentaire
LDA	Latent Dirichlet Allocation
LEXA	Lexique sécurité Alimentaire

LEXC	Lexique Crise
LEXG	Lexique Généraliste
LM	Modèle linéaire
LRP	Propagation de pertinence par couche
LSMS	Enquête de mesure des niveaux de vie
LSTM	Réseau de neurones récurrents à mémoire court-terme et long terme
MLP	Perceptron multicouche
MODIS	MODerate resolution Imaging Spectroradiometer
NDVI	Indice de végétation par différence normalisée
NOAA	National Oceanic and Atmospheric Administration
PAM	Programme Alimentaire Mondial
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SSA	Système d'alerte et de Suivi de la sécurité Alimentaire
SCA	Score de Consommation Alimentaire
SDA	Score de Diversité Alimentaire des ménages
SMT	Température de brillance lissée
SONAGESS	Société Nationale de Gestion du Stock de Sécurité Alimentaire
SPI	Indice de précipitation normalisé
SVM	Machine à vecteurs de support
TF-IDF	Term Frequency-Inverse Document Frequency
TIR	Tf-Idf Ratio
TRMM	Tropical Rainfall Measuring Mission
USAID	Agence américaine pour le développement international
VADER	Valence Aware Dictionary and Sentiment Reasoner
VAM	analyse et surveillance des vulnérabilités
W2V	Word2vec

Introduction générale

La faim reste un problème majeur dans de nombreuses régions du monde. Bien qu'une solution généralisée et durable à cette situation soit loin d'être atteinte, des progrès constants ont été réalisés au cours des quinze premières années du siècle actuel. Par exemple, en Afrique de l'Ouest, les années postérieures à 2000 ont vu une diminution de la prévalence de la sous-alimentation de près de 2%, atteignant une valeur relativement basse de 10,4% en 2012, qui est restée presque constante jusqu'en 2014 (10,7%). Néanmoins, une tendance inverse a été observée ces dernières années, avec ce même indicateur atteignant un pic de 15,2% en 2019, et révélant une projection alarmante de 23% pour 2030 (FAO et al., 2020). La même tendance inquiétante peut être observée pour des indicateurs connexes, comme par exemple, la prévalence de l'insécurité alimentaire grave en Afrique de l'Ouest qui a augmenté de près de 10% entre 2014 (20,7%) et 2017 (29,5%) (FAO and ECA, 2018). Parmi les pays d'Afrique de l'Ouest, le Burkina Faso est en proie à l'une des situations les plus critiques, avec une prévalence de la sous-alimentation de 21,3% au cours de la période 2015-2017 (FAO and ECA, 2018). Le Burkina Faso est également l'un des pays les plus touchés par le phénomène communément appelé le "triple fardeau de la malnutrition", caractérisé par la coexistence de la suralimentation, de la sous-nutrition et de carences en micronutriments dans la population (FAO et al., 2018). Les raisons d'une telle situation sont complexes, multifactorielles et interdépendantes. En plus de la pauvreté structurelle, les aléas climatique, exacerbés par le changement climatique, fragilisent la production agricole et la disponibilité des aliments (Tapsoba et al., 2019). Un autre facteur clé est l'augmentation de la population naturelle et migratoire, cette dernière étant causée par le nombre croissant de conflits dans la région du Sahel. Ces mouvements ont un impact majeur sur la chute des circuits de production et de distribution des aliments (Lacher, 2012). Suite à plusieurs crises alimentaires survenues dans les années 70 et 80 dans différentes régions du monde, plusieurs systèmes d'alerte et de surveillance de la sécurité alimentaire (SSA) ont été mis en place par des organisa-

tions gouvernementales et ONG. L'objectif de ces systèmes, qui sont toujours très actifs aujourd'hui, est de prévenir les crises alimentaires et d'aider les pays à planifier des programmes d'aide alimentaire pour optimiser les circuits de production et de distribution de nourriture. Nous pouvons citer le GIEWS (Global Information and Early Warning System) créé par l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO) et FEWS-NET (Famine Early Warning Systems Network) fondé par l'Agence américaine pour le développement international. Mais ces systèmes, bien que largement reconnus et utilisés à travers le monde, comportent des méthodes qui peuvent entraver l'analyse et la prévision rapides et précises des crises alimentaires. D'une part, les SSA s'appuient sur un ensemble important de données, mais ces dernières sont généralement spécifiques à certains domaines. Ces données sont principalement axées sur l'équilibre économique et nutritionnel des ménages dans les zones étudiées, permettant de saisir directement l'état de la sécurité alimentaire locale, et sur les conditions agrométéorologiques de la saison culturale, pour faire des prévisions de récolte. En Afrique de l'Ouest, et tout particulièrement dans la bande sahélienne, la combinaison des récents changements climatiques, démographiques et économiques, couplés aux conflits actuels, font de la compréhension de la sécurité alimentaire une question de plus en plus complexe. De ce fait, nous estimons que les données utilisées en routine par les SSA pourraient être enrichies afin d'obtenir une vision plus complète de cette problématique. D'autre part, ces systèmes sont essentiellement basés sur l'analyse et la synthèse manuelles de chaque source de données intégrée. Les analyses et les prévisions des experts prennent du temps, comportent inévitablement de nombreuses incertitudes et la sophistication des connaissances acquises est tributaire des limites des capacités cognitives humaines.

Compte tenu de la complexité de la question de la sécurité alimentaire et de la nécessité de plus en plus pressante de prévenir et de répondre à l'apparition de famines qui s'intensifient, nous pensons que l'utilisation de méthodes de traitement automatique de données en appui aux SSA peut être appropriée afin d'apporter plus rapidement certaines informations pertinentes. Les méthodes actuelles présentées dans ce mémoire ont très peu été exploitées par les SSA et constituent donc des pistes de recherche particulièrement intéressantes. Mais l'application de ce type d'outil exige une connaissance approfondie de la thématique de la sécurité alimentaire. Tout d'abord, la complexité de ce concept engendre une diversité d'indicateurs de sécurité alimentaire sur différentes composantes (e.g., disponibilité de la nourriture, accès des populations à celle-ci, utilisation), à différentes échelles (e.g., individuelle, des ménages, nationale), auxquels s'ajoutent d'autres

indicateurs sur les conditions climatiques, économiques et politiques. Ces indicateurs possèdent leurs propres variabilités dans l'espace et dans le temps, et certains indicateurs très pertinents dans certaines situations peuvent ne pas permettre d'identifier les crises dans d'autres contextes où d'autres indicateurs peuvent être préférables. Le choix éclairé des indicateurs à intégrer dans un système de traitement des données est donc une condition nécessaire à la production de résultats pertinents. Une autre condition nécessaire concerne les méthodes de traitement à appliquer pour extraire des informations de ces données variées. Plusieurs axes de recherche en mathématiques et en informatique se sont développés depuis le siècle dernier, et ont connu un essor considérable en ce début de siècle pour faire face à la vélocité, au volume et à la variété grandissants des données disponibles (Pandit et al., 2019), dans un large éventail de domaines (e.g., physique, biologie, médecine, télécommunications, sociologie, épidémiologie). Outre les multiples aspects thématiques qui peuvent être pris en compte par ces outils, ceux-ci sont notamment capables de traiter des données qui varient dans l'espace et dans le temps, ce qui est important pour comprendre des crises qui sont ponctuelles par définition. Parmi ces outils nous pouvons mentionner la modélisation mathématique, qui consiste à appliquer des modèles statistiques à des données environnementales afin de simuler leur comportement (Allaire, 2005), et le domaine des réseaux complexes, de la théorie des graphes, qui permet de représenter et d'analyser les relations complexes entre des composantes de la vie réelle par des objets mathématiques interconnectés par des liens (Van Steen, 2010). Basée sur une approche distincte, l'exploration de motifs séquentiels a pour objectif l'identification d'informations récurrentes dans des données complexes, correspondant à des caractéristiques ou des comportements fréquents et/ou inattendus (Masseglia et al., 2004). Enfin, l'apprentissage automatique fait référence à un ensemble de modèles capables d'extraire des règles sous-jacentes de grandes quantités de données, en apprenant à partir de celles-ci, sans être explicitement programmés pour cela (Hadj-Mabrouk, 1992). Au sein de cette discipline, un ensemble plus récent de méthodes offrant des niveaux d'abstraction élevés, qualifiées d'apprentissage profond (Huang et al., 2019), se distinguent par leur efficacité particulière sur des données complexes et hétérogènes dans de nombreux domaines (Valdés, 2018 ; Yuan et al., 2018b ; Wang et al., 2021). Cette famille de méthodes constitue l'un des outils clés de cette thèse. Les modèles résultants possèdent de bonnes capacités prédictives, (i.e., permettent d'approximer efficacement la valeur d'un indicateur d'intérêt), mais sont souvent d'une utilité limitée pour expliquer les facteurs à l'origine des variations de la sécurité alimentaire, car leurs règles de décision abstraites sont généralement difficiles à interpréter. La production d'informations

contextuelles, plus interprétables, nécessite quant à elle l'utilisation de données et de méthodes d'autres types. Pour cela, les données textuelles (e.g., réseaux sociaux, journaux) peuvent offrir une piste intéressante. En effet, celles-ci sont fortement qualitatives grâce à la richesse des informations contenues dans le vocabulaire utilisé, et sont par ailleurs produites par leur auteur dans un contexte particulier, elles sont ainsi également porteuses d'informations spatio-temporelles. La fouille de textes, qui désigne un ensemble de techniques visant à extraire automatiquement des connaissances à partir de données textuelles, peut fournir des informations sur les thèmes, les sentiments ou le vocabulaire spécifique présents dans des textes (Berry and Kogan, 2010). C'est une approche que nous adoptons également dans cette thèse.

L'objectif de cette thèse est la conception de méthodologies permettant la mise en relation de données hétérogènes pour l'analyse de la sécurité alimentaire, et se situe à la croisée de deux domaines de recherche. D'une part, nous portons un regard thématique sur la question de la sécurité alimentaire afin d'identifier l'ensemble pléthorique et hétérogène d'indicateurs qui permettent de la mesurer dans sa complexité. Nous accordons une attention particulière au comportement de ces données dans l'espace et dans le temps ainsi qu'aux biais auxquels elles peuvent être soumises. D'autre part, nous apportons une vision computationnelle en définissant des méthodes originales de science des données spécifiquement conçues pour traiter et combiner des données hétérogènes, afin de fournir des informations de nature prédictive et explicative contextuellement pertinentes.

Cependant, l'aspect "automatique" de ces méthodes prédictives et explicatives est exclusivement lié à leurs traitements. Le choix des données à intégrer, ainsi que des méthodes et de leurs paramétrages constituent des questions majeures. Une réflexion approfondie à plusieurs niveaux doit être entreprise. Tout d'abord sur le choix des données à considérer pour appréhender la sécurité alimentaire sous ses multiples aspects. Ensuite, sur les méthodes les plus appropriées à chaque type de données parmi le grand nombre de méthodes existantes. Enfin, sur les stratégies à adopter pour combiner ces données hétérogènes afin d'obtenir des informations riches de tous ces éléments. A chacune de ces étapes, qui constituent autant d'obstacles méthodologiques, nous devons être guidés par un fil directeur, nous orientant vers des choix préservant et valorisant la triple dimension thématique, spatiale et temporelle propre à notre contexte, de la sélection des données à l'analyse des informations obtenues. C'est sous ce prisme à trois facettes qu'une situation alimentaire peut être pleinement quantifiée et expliquée, mais également située dans l'espace et le temps, ces conditions étant nécessaires pour obte-

nir des informations véritablement pertinentes. Soulignons que le traitement de données hétérogènes reste un problème ouvert et difficile en science des données. Nous tentons de lever ce verrou scientifique à travers deux axes qui s'entremêlent : (1) la mise en avant du triptyque "thématique, spatial, temporel" dans les données ; (2) l'intégration et l'adaptation des méthodes d'apprentissage automatique et de fouille de données.

C'est à la lumière de ce contexte alimentaire complexe et préoccupant, des nouvelles méthodes issues de la science des données qui offrent une opportunité intéressante d'y apporter une contribution, et des différents verrous scientifiques que cela implique que nous avons réalisé le travail présenté ici. Cette thèse est menée autour de trois axes majeurs : mesurer, prédire et expliquer la sécurité alimentaire, qui font l'objet de trois chapitres distincts.

Le chapitre 1 aborde les différentes approches existantes pour mesurer la sécurité alimentaire, ainsi que leurs variations dans l'espace et le temps et leurs limites. Nous détaillons en particulier les indicateurs issus d'enquêtes ménages, qui fournissent des informations essentielles, en examinant et discutant leurs biais inhérents. A partir d'une enquête ménages menée au Burkina Faso au cours de la dernière décennie, et de diverses autres sources de données, nous étudions si les indicateurs de sécurité alimentaire dérivés de ce type d'enquête peuvent, malgré leurs limites, être utilisés pour détecter les crises alimentaires au niveau régional et pour différentes années, et s'ils sont corrélés avec des proxies décrivant le contexte agroclimatique et économique de la région.

Le chapitre 2 se consacre à des méthodes de prédiction numériques d'indicateurs de sécurité alimentaire. Nous proposons un framework basé sur des modèles d'apprentissage automatique et profond, pour l'estimation d'indicateurs de sécurité alimentaire issus d'une enquête ménages burkinabè, dont la validité spatiale et temporelle est démontrée dans le chapitre 1, en utilisant des données explicatives hétérogènes ouvertes. L'ensemble des données intégrées inclut des variables quantitatives (e.g., variables de la Banque mondiale, données météorologiques), des images satellites (e.g., densités de population, occupation du sol), des données géolocalisées (e.g., hôpitaux, écoles, événements violents), des vecteurs lignes (rivières), ainsi que des séries temporelles (e.g., précipitations, prix du maïs). Les résultats expérimentaux montrent des performances très encourageantes dans notre contexte, surclassant les méthodes concurrentes.

Le chapitre 3 examine des méthodes visant à obtenir un contexte explicatif associé aux situations alimentaires à partir de données textuelles. Plus précisément, nous évaluons la capacité des méthodes de fouille de textes à extraire des informations qualitatives, utilisées comme descripteurs de la situation alimentaire régionale et de son évolution au cours des dix dernières années au Burkina Faso, à partir d'un corpus de journaux du pays. L'utilité de cette approche est de proposer un cadre explicatif complémentaire aux sorties des modèles prédictifs appliqués à d'autres types de données (données numériques et images satellites) proposés dans le chapitre 2. Ce travail a permis d'obtenir des informations qualitatives sur le thème de la sécurité alimentaire et sur ses caractéristiques spatiale et temporelle.

Chapitre 1

Mesurer la sécurité alimentaire

1.1 Introduction

La sécurité alimentaire est un concept complexe, résultant de facteurs multiples et interdépendants tels que le climat, l'économie et les guerres. Celle-ci est assurée "lorsque tous les êtres humains ont, à tout moment, un accès physique et économique à une alimentation suffisante, hygiénique et nutritive" (Shaw, 2007).

De cette définition, quatre composantes se dégagent :

1. l'**accès** physique, économique et social de toutes les personnes aux ressources nécessaires pour acquérir les aliments indispensables à une alimentation nutritive ;
2. la **disponibilité** en quantité suffisante d'aliments de nature et de qualité appropriées ;
3. la **qualité** appropriée des aliments aux niveaux sanitaire et nutritionnel ;
4. la **régularité** de l'accès, de la disponibilité et de la qualité de la nourriture dans le temps malgré les chocs naturels ou économiques.

Pour prendre en compte ces multiples facettes, choisir les bons indicateurs n'est pas une tâche aisée. La sécurité alimentaire résulte de l'interaction entre de nombreux facteurs agro-environnementaux, socio-économiques et culturels. Il existe un grand nombre d'indicateurs de sécurité alimentaire dont la catégorisation a fait l'objet de plusieurs études (Pérez-Escamilla and Segall-Corrêta, 2008 ; Jones et al., 2013 ; Carletto et al., 2013) et fait encore l'objet de débats et de controverses. Aucun consensus ne se dégage sur la manière de mesurer la sécurité alimentaire, bien que les indicateurs collectés au

niveau individuel fassent généralement référence (Bobe et al., 2019). En effet, l'alimentation est un acte individuel et l'individu constitue l'élément central de la définition de la sécurité alimentaire, mais celui-ci vit généralement dans un ménage, lui-même inscrit dans un environnement biophysique, économique, social et politique particulier. De ce fait, les données peuvent être collectées : (i) directement auprès des individus, où se fait l'alimentation (e.g., suivi des aliments ingérés ou mesures anthropométriques) ; (ii) auprès des ménages où se raisonne la production, l'achat et la préparation des aliments (e.g., suivi des dépenses engagées et des types d'aliments achetés) ; (iii) au niveau d'espaces géographiques plus larges, comme pour les bassins de production (e.g., état de la végétation agricole, estimation des surfaces cultivées), les marchés agricoles (suivi de prix et de la disponibilité alimentaire) ou encore les politiques publiques mises en place. De fait, les systèmes d'alerte et de suivi de la sécurité alimentaire (SSA) intègrent de multiples sources de données : des données agroclimatiques (e.g., images satellites, données météorologiques), des données socio-économiques recueillies auprès des ménages et des données nutritionnelles individuelles qui sont analysées par des experts en sécurité alimentaire (Fritz et al., 2019). Dans ces systèmes, les données recueillies au travers d'enquêtes ménages représentent une source d'information fondamentale pour la collecte d'indicateurs clés : le Programme Alimentaire Mondial (PAM) déploie ses enquêtes ménages CFSVA (Comprehensive Food Security and Vulnerability Analysis) dans le monde entier (WFP, 2009), la Banque Mondiale collecte des informations sur les niveaux ressentis d'insécurité alimentaire par les ménages dans de multiples régions du globe via le dispositif LSMS (Living Standad Measurement Survey) et la FAO (Food and Agriculture Organization) a développé un outil (ADePT-Food Security Module) pour produire des indicateurs de sécurité alimentaire à partir des données de consommation collectées par des enquêtes ménages (Moltedo et al., 2014). Les indicateurs utilisés par ces systèmes sont généralement agrégés aux niveaux des provinces, régions et pays afin de générer des cartes de sécurité alimentaire qui sont utilisées pour déterminer les cibles des interventions alimentaires [1].

Les enquêtes ménages constituent un outil puissant par leur capacité à couvrir une grande partie de la population et un large éventail de domaines. Celles-ci sont exploitées en routine par les SSA ; dans la plupart des cas la représentativité statistique de ces données est assurée au niveau de divisions administratives sub-nationales et les informations fournies sont considérées comme fiables pour évaluer la situation alimentaire au niveau

1. voir https ://hungermap.wfp.org/

de ces sous-ensembles (Melgar-Quinonez and Hackett, 2008). Cependant, une enquête ménages ne fournit que des estimations de certains indicateurs spécifiques, donnant ainsi une image partielle de la sécurité alimentaire. De plus, comme dans la plupart des enquêtes, la qualité des données obtenues au niveau des ménages peut être affectée par des biais. Plusieurs publications scientifiques ont étudié et classifié les biais qui surviennent généralement dans les enquêtes ménages (Winter, 2004 ; Dussaix, 2009 ; Biemer, 2010), mais très peu d'études examinent en profondeur les biais dans les enquêtes ménages sur la sécurité alimentaire.

Dans ce chapitre, nous analysons des indicateurs de sécurité alimentaire issus d'enquêtes ménages et étudions leur validité pour détecter les crises alimentaires à l'échelle régionale et pour différentes années. Dans un premier temps, en section 1.2 nous passons en revue les principales catégories d'indicateurs de sécurité alimentaire, en mettant l'accent sur les indicateurs issus d'enquêtes ménages et leurs biais inhérents ; nous nous concentrons alors sur trois indicateurs issus d'enquêtes ménages qui sont communément utilisés pour estimer le statut nutritionnel des membres d'un ménage : le SCA (score de consommation alimentaire), le SDA (score de diversité alimentaire des ménages) et l'$ISAr$ (indice des stratégies d'adaptation réduit). Notre étude est basée sur des données provenant d'une enquête ménages officielle (l'enquête permanente agricole) menée au Burkina Faso au cours de la dernière décennie. Dans la section 1.3, nous présentons la méthodologie de cette enquête ainsi que d'autres sources de données liées à la sécurité alimentaire, puis nous en discutons les biais éventuels. En section 1.4, nous explorons les concordances du SCA, du SDA et de l'$ISAr$ aux échelles régionale et annuelle et analysons leur cohérence avec des indicateurs "proxies" liés à la sécurité alimentaire issus d'autres sources de données. La section 1.5 conclut sur les points forts et les limites des enquêtes ménages pour fournir des informations pertinentes sur la sécurité alimentaire.

1.2 Les indicateurs de la sécurité alimentaire

1.2.1 Pluralité d'indicateurs

La sécurité alimentaire peut être évaluée selon ses multiples composantes (accès, disponibilité, qualité, stabilité) et à différents niveaux, au moyen de sources de données aux échelles nationale, régionale, des ménages ou des individus. Il existe un grand nombre d'indicateurs de sécurité alimentaire, et l'utilisation de plusieurs indicateurs est recom-

mandée en raison de la complexité de ce concept (Coates, 2013). Hoddinott (1999) a estimé le nombre d'indicateurs de sécurité alimentaire à environ 450.

Les principaux indicateurs de sécurité alimentaire peuvent être catégorisés comme suit :

Mesures anthropométriques : ces indicateurs sont basés sur le constat que la malnutrition (excessive ou déficiente) a des effets mesurables sur la morphologie d'un individu comme le montrent Leyna et al. (2010). Il s'agit de mesures physiques prises au niveau des individus comme indicateurs de la croissance et de l'état nutritionnel telles que la courbe du poids en fonction de l'âge, le périmètre brachial ou encore l'Indice de Masse Corporelle (IMC) qui correspond au rapport du poids d'un individu par sa taille au carré. Ces indicateurs sont largement utilisés par la FAO (FAO and ECA, 2018) et renseignent sur l'accès des individus aux denrées, ainsi que sur leur qualité nutritionnelle.

Apports caloriques : la quantité de calories ingérées est un indicateur direct de la suffisance alimentaire d'un individu et reflète sa capacité à accéder à l'alimentation. De Araujo et al. (2018) ont démontré que l'insécurité alimentaire réduit la consommation calorique d'aliments sains par les membres des ménages. Les indicateurs privilégiés sont des mesures de l'ingestion d'aliments au niveau individuel (Schiff and Valdes, 1990), mais de plus en plus d'indicateurs au niveau ménage ou même au niveau de la population d'un pays se développent. Au niveau ménage, le suivi sur une période donnée des achats et préparations alimentaires permet d'estimer la quantité de calories ingérées par les individus du ménage (Iram and Butt, 2004). Au niveau de la population d'un pays, la FAO a mis au point une méthode de calcul de la prévalence de la sous-alimentation qui dépend de trois paramètres : la quantité moyenne de calories disponibles dans un pays pour la consommation humaine, l'accès inégal à ces calories dans la population du pays et la quantité moyenne minimale de calories requise par cette population (Wanner et al., 2014).

Fréquence et diversité alimentaires : ce type de mesure permet de quantifier la diversité et la fréquence de consommation de différents groupes d'aliments consommés en les pondérant dans certains cas par la qualité de l'apport nutritionnel du type d'aliment. Ces indicateurs peuvent être collectés au niveau individuel ou au niveau des ménages, la plupart du temps via des questionnaires dans lesquels les répondants doivent retracer sur une durée de 24 h ou 7 jours les aliments qu'ils ont consommés et la fréquence

d'occurrence de ces consommations. Ces indicateurs reflètent la diversité de l'apport et donc la qualité nutritionnelle de l'alimentation, sans nécessairement en refléter la quantité. Cependant, Maxwell et al. (2014) ont observé que ces scores sont significativement corrélés avec les mesures caloriques de la consommation des ménages. Le score de diversité alimentaire des ménages (SDA) est un indice de diversité alimentaire qui a été proposé par l'Agence des États-Unis pour le développement international (Swindale and Bilinsky, 2006) et qui a été largement utilisé depuis. Le score de consommation alimentaire (SCA) est similaire au SDA, mais avec une période de rappel d'une semaine au lieu de 24 heures et incluant une pondération par groupes d'aliments (Wiesmann et al., 2009).

Comportements de consommation : ces mesures saisissent implicitement la sécurité alimentaire en évaluant les comportements de consommation. L'indice des stratégies d'adaptation (ISA) mesure la fréquence et la gravité des comportements que les individus adoptent lorsqu'ils n'ont pas assez de nourriture ou d'argent pour en acquérir, et se base sur les réponses à 12-16 questions posées (Maxwell, 2008). Des travaux plus récents ont permis d'identifier un "indice universel" de 5 comportements d'adaptation considérés comme les plus pertinents, appelé ISA réduit ($ISAr$), largement utilisé par le programme alimentaire mondial (PAM) (Vhurumuku, 2014). Ces comportements concernent la fréquence, la taille et la qualité des repas ainsi que la sollicitation d'une aide pour se procurer de la nourriture. L'existence de ces stratégies est une mesure de l'accès des individus des ménages à l'alimentation, de la régularité de l'accès et de la disponibilité des aliments. L'échelle de la faim dans les ménages (EFM) est un indice de privation alimentaire des ménages, également basé sur l'observation que les privations alimentaires entraînent des comportements prévisibles. (Ballard, 2011).

Dépenses alimentaires : étant donné que les ménages qui s'approchent du seuil de pauvreté ont tendance à consacrer une proportion plus élevée de leurs revenus à la nourriture, l'estimation de la part des dépenses alimentaires est devenue une mesure importante (Smith and Subandoro, 2007). Cet indicateur permet d'estimer l'accès économique des ménages à l'alimentation.

Indices globaux : des indices globaux sont calculés en combinant certains indicateurs présentés ci-dessus et en les synthétisant aux niveaux régional et national. L'indice de la faim dans le monde (GHI), défini par l'Institut international de recherche sur les politiques alimentaires, vise à mesurer la faim à l'aide de trois indicateurs à pondération

égale : la prévalence de la sous-alimentation, la sous-alimentation infantile et la mortalité infantile (International Food Policy Research Institute, 2017). L'indice global de la sécurité alimentaire (GFSI), conçu par l'Economist Intelligence Unit, est un autre outil multidimensionnel permettant d'évaluer les tendances de la sécurité alimentaire au niveau national. Cet indice utilise un total de 30 indicateurs issus de 3 domaines : accessibilité financière des aliments (6), disponibilité (10) et qualité (14) (The Economist Intelligence Unit, 2018).

Il existe également une variété d'indicateurs indirectement liés à la sécurité alimentaire largement utilisés par les SSA (Fritz et al., 2019), parmi lesquels les indices de végétation, les précipitations et les prix des aliments (présentés en section 1.3.3 et analysés en section 1.4.2) sont particulièrement utilisés. Dans cette thèse, nous utilisons le terme "proxies" pour nous référer à ces indicateurs mesurant un aspect indirectement lié à la sécurité alimentaire, par opposition aux indicateurs de la sécurité alimentaire qui mesurent un critère directement lié à ce domaine. Ces proxies saisissent très indirectement la sécurité alimentaire des individus, mais fournissent un contexte climatique, économique et social permettant d'expliquer les situations de famine, et donnent ainsi des pistes pour y répondre. Ces proxies sont donc complémentaires aux indicateurs de sécurité alimentaire catégorisés ci-dessus.

1.2.2 Des indicateurs à différentes échelles

Cette variété dans les types d'indicateurs de la sécurité alimentaire implique une diversité dans les échelles de collecte des données, allant des individus et des ménages dans lesquels ils vivent (e.g., apports caloriques, mesures anthropométriques), aux échelles régionales et nationales (e.g., population et bilans céréaliers du pays) (Tableau 1.1). Les indicateurs dérivés d'observations individuelles sont considérés comme les plus pertinents pour caractériser la sécurité alimentaire, mais ils sont longs et coûteux à obtenir. Les indices globaux disponibles aux niveaux régional et national sont généralement construits à partir d'indicateurs qui peuvent être fournis plus rapidement et à moindre coût que les données individuelles, mais ils ne reflètent que partiellement la consommation réelle. Entre ces deux extrêmes, les indicateurs générés à l'échelle des ménages par le biais d'enquêtes constituent un bon compromis. Ils sont considérés comme un moyen rentable d'évaluer la consommation alimentaire et l'état nutritionnel individuels (comme nous le précisons dans la section suivante) et sont devenus essentiels à l'analyse de la sécurité

alimentaire.

Indicateur	Échelle
Prévalence de la sous-alimentation (Wanner et al., 2014) Indice de la faim dans le monde (GHI) (International Food Policy Research Institute, 2017) Indice global de la sécurité alimentaire (GFSI) (The Economist Intelligence Unit, 2018)	Nationale
Score de consommation alimentaire (SCA) (Wiesmann et al., 2009) Score de diversité alimentaire des ménages (SDA) (Swindale and Bilinsky, 2006) Indice des stratégies d'adaptation (ISA) (Maxwell, 2008) Dépenses alimentaires (Smith and Subandoro, 2007) Échelle de la faim dans les ménages (EFM) (Ballard, 2011)	Ménage
Mesures anthropométriques (Leyna et al., 2010) Apport calorique (De Araujo et al., 2018)	Individuelle

Tableau 1.1 – Vue d'ensemble des indicateurs de sécurité alimentaire à différentes échelles

1.2.3 Les enquêtes ménages pour estimer les indicateurs individuels

Comme nous l'avons évoqué précédemment, les enquêtes individuelles fournissent les informations les plus directes et précises sur la sécurité alimentaire (e.g., mesures anthropométriques, aliments ingérés). Nous nous interrogeons ici sur la pertinence des enquêtes ménages comme moyen efficace d'approcher ces informations individuelles. Pour comparer les avantages de l'utilisation d'enquêtes ménages par rapport aux enquêtes individuelles dans notre contexte, nous nous référons à l'étude de Headey and Ecker (2013), selon laquelle quatre critères sont nécessaires pour évaluer l'intérêt d'un indicateur de sécurité alimentaire : son coût, le délai dans lequel il peut être produit, sa pertinence nutritionnelle et sa validité temporelle et spatiale (i.e., sa capacité à détecter les chocs alimentaires annuels et régionaux). Du point de vue des coûts, les indicateurs obtenus au niveau des ménages sont plus intéressants que ceux obtenus au niveau des individus qui requièrent un suivi très précis et coûteux pour la production de l'information. De même, du point de vue du temps nécessaire à la production des indicateurs, ceux dérivés des enquêtes ménages nécessitent moins de temps pour obtenir les informations nécessaires car celles-ci sont généralement collectées auprès d'une seule personne par ménage (généralement le "chef" du ménage). De plus, les indicateurs dérivés d'enquêtes sur la consommation individuelle impliquent des mesures très fines de la qualité nutrition-

nelle et de paramètres physiques, ce qui nécessite un temps de recueil particulièrement conséquent. En revanche, en termes de pertinence nutritionnelle, les nutritionnistes privilégient les approches où les indicateurs de sécurité alimentaire sont mesurés au niveau individuel (Bobe et al., 2019) ; il est cependant reconnu aujourd'hui que les enquêtes ménages fournissent des informations valides sur le plan nutritionnel (Cafiero et al., 2014). La validité temporelle et spatiale désigne la capacité des indicateurs à détecter les tendances et les chocs alimentaires, d'un espace géographique à un autre et d'une période à une autre. Cette composante a fait l'objet d'études menées à l'échelle du pays (Rautela et al., 2020), de la ville (Tuholske et al., 2020) et du ménage (Wichern et al., 2018). Cependant, celles-ci sont généralement menées sur une année, et les études sur la validité interannuelle des indicateurs de sécurité alimentaire dérivés d'enquêtes ménages font défaut. Dans la suite de ce chapitre, nous nous intéressons particulièrement à ce dernier critère de pertinence des indicateurs issus d'enquêtes ménages comme outil de caractérisation de la sécurité alimentaire, en analysant leur validité spatio-temporelle.

1.2.4 Accent sur trois indicateurs collectés au niveau des ménages

Nous nous concentrons dans la suite de ce chapitre sur trois indicateurs dérivés d'enquêtes ménages : le score de consommation alimentaire (SCA), le score de diversité alimentaire des ménages (SDA) et l'indice des stratégies d'adaptation réduit ($ISAr$). Ces indicateurs fournissent des informations utiles sur la fréquence, la quantité et la qualité de l'alimentation ainsi que sur l'accès économique des ménages à la nourriture et font partie des indicateurs les plus utilisés dans la littérature scientifique et par les organisations internationales et les gouvernements (Jones et al., 2013 ; Maxwell et al., 2014 ; Vhurumuku, 2014).

Score de consommation alimentaire (SCA) : le SCA est une mesure de l'apport en nutriments et en énergie. Il représente une estimation de la fréquence cumulée de consommation de différents groupes d'aliments sur 7 jours pour chaque ménage interrogé. La fréquence de consommation de chaque groupe d'aliments est pondérée par sa valeur nutritive. (Équation 1.1 ; Tableau 1.2). Des seuils sont utilisés pour qualifier la consommation alimentaire des ménages. Nous nous référons aux seuils fixés par le PAM : acceptable (> 42), limite ($28 - 42$) et faible (< 28) ; (Wiesmann et al., 2009).

$$SCA = \sum_{i=1}^{9} x_i.p_i \tag{1.1}$$

$x_i \in$ {Fréquence de consommation pour chaque groupe d'aliments i}, $p_i \in$ {Poids du groupe d'aliments i}

Groupe d'aliments	**Poids (valeur nutritionnelle)**
Céréales et tubercules	2
Légumineuses	3
Légumes et feuilles	1
Fruits	1
Protéines animales	4
Produits laitiers	4
Sucres	0,5
Huiles	0,5
Condiments	0

Tableau 1.2 – Groupes d'aliments et pondérations correspondantes utilisés pour calculer le score de consommation alimentaire (SCA). *Source : Wiesmann et al. (2009)*

Score de diversité alimentaire des ménages (SDA) : le SDA indique le nombre de groupes d'aliments consommés au cours des dernières 24 heures et est considéré comme une estimation acceptable de la consommation alimentaire (Kennedy et al., 2010). Il n'y a pas de consensus sur le nombre de groupes à utiliser et sur les seuils limites à considérer (Kennedy et al., 2013). Dans cette étude, nous nous référons aux recommandations de la FAO et calculons le SDA (Équation 1.2) sur la base de 12 groupes d'aliments (céréales ; racines et tubercules ; légumes ; fruits ; produits carnés ; œufs ; poissons et fruits de mer ; légumineuses, noix et graines ; lait et produits laitiers ; huiles et graisses ; sucreries ; condiments, épices et boissons (Swindale and Bilinsky, 2006)).

$$SDA = \sum_{i=1}^{12} x_i \tag{1.2}$$

$x_i \in$ {0 : l'aliment i n'est pas consommé, 1 : l'aliment i est consommé}

Indice des stratégies d'adaptation réduit ($ISAr$) : l'$ISAr$ prend en compte la sévérité des stratégies que les ménages adoptent pour faire face aux déficits de leur consommation alimentaire. Il s'agit d'une estimation de la fréquence cumulée de cinq stratégies potentielles de réduction de la consommation alimentaire utilisées sur 7 jours au sein de chaque ménage enquêté. La fréquence de chaque comportement est pondérée par sa sévérité (Équation 1.3 ; Tableau 1.3).

$$ISAr = \sum_{i=1}^{5} xi.pi \tag{1.3}$$

$x_i \in$ {Fréquence du comportement i}, $p_i \in$ {Poids du comportement i}

Comportement	**Poids**
Utiliser des aliments moins populaires et moins coûteux	1
Emprunter de la nourriture ou demander l'aide d'un ami ou d'un parent	2
Limiter la taille des portions lors des repas	1
Réduire la consommation des adultes pour nourrir les enfants	3
Réduire le nombre de repas par jour	1

Tableau 1.3 – Groupes d'aliments et poids correspondants utilisés pour calculer l'indice des stratégies d'adaptation réduit ($ISAr$). *Source : Maxwell (2008)*

Nous illustrons dans la suite de ce chapitre comment ces trois indicateurs peuvent être estimés à partir de données issues d'enquêtes ménages, en nous appuyant sur le cas du Burkina Faso. Une enquête (présentée dans la section 1.3.2) sur les ménages ruraux y est menée en routine depuis 1982 (les données sur la consommation alimentaire étant disponibles depuis 2009).

1.3 Matériel et méthodes

1.3.1 La situation alimentaire au Burkina Faso

Le Burkina Faso est un pays en développement situé au cœur de l'Afrique de l'Ouest. Comme dans de nombreux pays à faible et moyen revenus, la sécurité alimentaire y est

étroitement liée à la production agricole. C'est un pays soudano-sahélien au climat semi-aride au nord et subhumide au sud, caractérisé par deux saisons très contrastées : la saison sèche (6 mois au sud à 9 mois au nord du pays) et la saison des pluies qui correspond à la saison agricole. Les précipitations sont généralement faibles, irrégulières et inégalement réparties (Tapsoba et al., 2019). Ainsi, avec 400–600 mm de pluviométrie annuelle dans le nord, ce sont principalement des céréales pluviales (mil, sorgho) et des pâturages qui sont cultivés. Dans le sud, la pluviométrie annuelle (700–900 mm) est suffisante pour assurer une diversification des cultures (sorgho, mil, maïs, riz, coton et arachide) (voir Figure 1.1). La production agricole du pays doit répondre à une demande croissante liée essentiellement à l'augmentation démographique (la population ayant presque doublé en 10 ans, passant de 11,6 millions d'habitants en 2000 à 20 millions en 2019 (World Bank, 2020)), et cela avec des moyens d'intensification agricole limités et le désengagement de nombreux jeunes pour le travail agricole. Ces changements récents et rapides bouleversent l'organisation du pays dans les campagnes, avec 80% des emplois liés à l'agriculture en 2019 (World Bank, 2020). La forte dépendance des populations rurales à l'agriculture pluviale les rend donc particulièrement vulnérables aux variations climatiques. La sécurité alimentaire dépend directement de la production de l'agriculture pluviale : en année "normale", la production nationale couvre la consommation nationale. Mais la multiplication des phénomènes climatiques extrêmes menace la situation de la sécurité alimentaire au Burkina Faso (Tapsoba et al., 2019). Par ailleurs, la menace terroriste se précise et s'aggrave dans les zones frontalières avec le Mali et le Niger qui ont connu des attaques terroristes répétées depuis 2016, déstabilisant le pays sur le plan économique et social (Toros, 2019) et diminuant la production agricole dans les zones de conflits armés. Ces informations contextuelles sont prises en compte dans l'évaluation de la sécurité alimentaire.

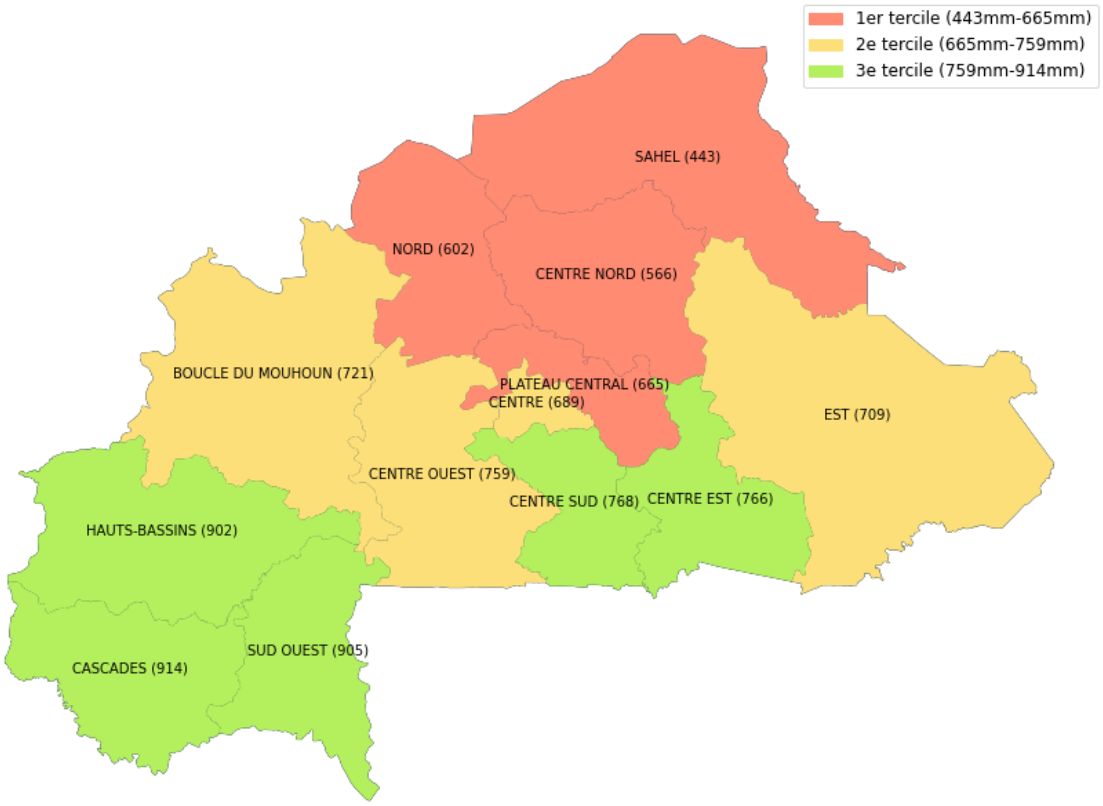


FIGURE 1.1 – Zones climatiques du Burkina Faso, basées sur le cumul pluviométrique de mai à octobre dans les 13 régions du Burkina Faso en moyenne entre 2014 et 2017. *Source : données CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data)*

1.3.2 Les enquêtes ménages

Dans cette section, nous présentons trois enquêtes ménages menées au Burkina Faso. L'enquête permanente agricole (EPA) qui prend une place centrale dans cette thèse, l'enquête d'analyse globale de la sécurité alimentaire et de la vulnérabilité (CFSVA) menée par le PAM et l'enquête de mesure des niveaux de vie (LSMS) qui est l'une des plus importante en matière de sécurité alimentaire en Afrique de l'Ouest.

L'enquête EPA

L'enquête permanente agricole (EPA) est menée chaque année par le ministère de l'Agriculture et des Aménagements Hydrauliques depuis 1982 au Burkina Faso (Permanent Agricultural Survey, 2015). Cette enquête est un dispositif opérationnel utilisé dans le domaine de l'agriculture et de la sécurité alimentaire qui fournit aux décideurs et aux organisations alimentaires des prévisions de récoltes provinciales pendant la saison des cultures (durant le mois d'août) et des estimations de la production agricole en fin de saison (durant le mois d'octobre). A cette fin, l'EPA se focalise sur les ménages agricoles du pays, définis comme pratiquant l'une des activités suivantes : cultures temporaires (cultures pluviales et de contre-saison), fruticulture, élevage d'animaux. L'EPA est réalisée par un échantillonnage stratifié à deux degrés. L'unité primaire est le village administratif, tiré avec une probabilité proportionnelle à sa taille en ménages agricoles. L'unité secondaire est le ménage agricole, les ménages sont regroupés en deux strates homogènes en fonction de leur capacité de production agricole. Le nombre de ménages est choisi pour être représentatif dans chaque province. Depuis 2009, des informations sur la consommation alimentaire des ménages sont également collectées pour calculer les indicateurs de sécurité alimentaire après la récolte (vers le mois de décembre). Les estimations de la production agricole et les indicateurs de sécurité alimentaire sont des éléments clés utilisés par les SSA.

Pour cette étude, nous utilisons un ensemble de données qui couvre la période 2009–2017 et qui contient les informations de 41613 ménages agricoles, soit une moyenne de 4624 ménages agricoles par an répartis dans 342 communes parmi les 351 représentées sur la Figure 1.2.

FIGURE 1.2 – Carte des 351 communes du Burkina Faso (fond de carte : Google Maps).

Nous nous concentrons sur les indicateurs SCA, SDA, $ISAr$ calculés à partir des réponses à l'EPA. Le SCA et le SDA sont moyennés par province et considérés de 2009 à 2017, ce qui représente 405 observations. L'$ISAr$, disponible depuis 2014, est moyenné par province et considéré de 2014 à 2017, soit 180 observations. La distribution des indicateurs moyennés par province est donnée sur la Figure 1.3. Les distributions associées au SCA et au SDA sont proches de lois normales, de moyenne respectivement 52.7 et 5.35. Cela signifie que les distributions des provinces à faible et fort SCA et SDA ont tendance à être équilibrées et symétriques autour des scores moyens. Cependant, pour les deux indicateurs nous notons une légère asymétrie des distributions, avec une queue de valeurs basses plus étalée que celle des valeurs hautes. Cela indique que les provinces sont légèrement plus sujettes à de faibles et très faibles valeurs de SCA et SDA. La distribution associée à l'$ISAr$ s'approche d'une loi exponentielle, avec une majorité de provinces qui possèdent un score très faible et un petit nombre possédant des scores très élevés. Dans la section 1.4.2, le SCA, le SDA, et l'$ISAr$ sont également centrés réduits (i.e., en soustrayant la moyenne et en divisant par l'écart-type) par rapport aux

provinces et aux années pour être comparés avec les proxies de la sécurité alimentaire présentés en section 1.3.3.

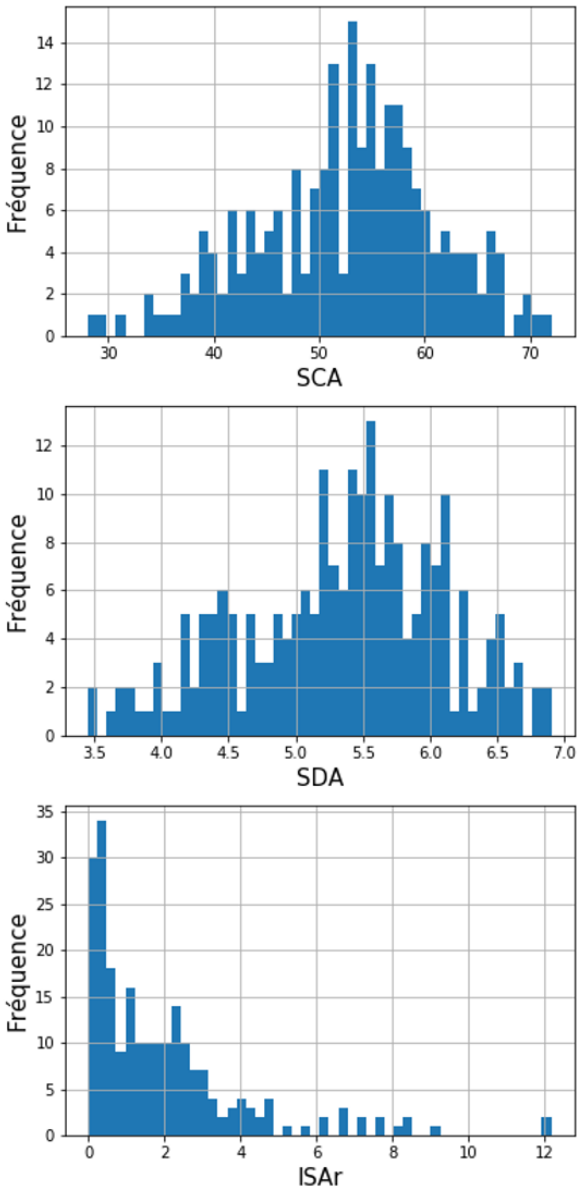


FIGURE 1.3 – Histogrammes des distributions annuelles de SCA, SDA et $ISAr$ de 2014 à 2017, moyennés par province.

L'enquête CFSVA

Le PAM utilise son outil d'analyse globale de la sécurité alimentaire et de la vulnérabilité (CFSVA) pour quantifier la sécurité alimentaire dans certaines régions du monde, mais également pour comprendre les facteurs qui sous-tendent la sécurité alimentaire (WFP, 2009). Les données extraites des enquêtes ménages sont utilisées pour décrire le profil des ménages vulnérables et en situation d'insécurité alimentaire, et élaborer des mesures d'urgence appropriées. Au Burkina Faso, des données ont été collectées auprès de 3670 ménages dans 558 villages en février 2018, à partir desquelles nous avons extrait les indicateurs SCA, SDA et $ISAr$ par province (PAM, communication personnelle, 2019). Les résultats de cette enquête sont comparés à ceux de l'EPA en section 1.4.2.

L'enquête LSMS

Une enquête de mesure du niveau de vie (LSMS) a été menée en 2014 au Burkina Faso par l'Institut National de la Statistique et de la Démographie et financée par la Banque Mondiale. Les enquêtes LSMS appliquent une méthodologie finement élaborée et sont représentatives aux niveaux national et régional (Ambagna, 2018). La méthodologie employée ainsi que les données sont accessibles publiquement [2].

Compte tenu du fait qu'un unique indicateur (le SCA) est comparable entre cette enquête et l'EPA, et que le SCA a été obtenu de manière approchée pour l'enquête LSMS [3], nous n'approfondissons pas l'analyse de la cohérence entre ces deux enquêtes dans la section résultats. Nous avons effectué un test de corrélation de Pearson entre les SCA moyens par province obtenus à partir de l'EPA et de l'enquête LSMS en 2014. Ce test se révèle significativement positif (valeur $p < 0.05$), bien que la corrélation associée ne soit pas élevée (0.3). Cela indique qu'il existe certaines tendances communes dans les résultats relatifs aux SCA issus de ces deux enquêtes.

2. https ://microdata.worldbank.org/index.php/catalog/2538/study-description

3. Le calcul du SCA nécessite de connaître le jour de consommation de chaque type d'aliment. Or, à partir des données LSMS nous n'avons accès qu'à la liste globale des aliments consommés au cours des 7 derniers jours. Ainsi, il est possible que si un type d'aliment est consommé plusieurs fois dans la même journée, il soit comptabilisé plus d'une fois au lieu d'un maximum d'une fois par jour (comme défini dans le calcul du SCA), ce qui peut conduire à une surestimation du SCA

1.3.3 Données de végétation, de précipitations et de prix des denrées

Nous avons identifié des données supplémentaires qui peuvent être interprétées comme des proxies, mesurant des critères indirectement liés à la sécurité alimentaire, à savoir un indice de végétation, des estimations de précipitations et des prix alimentaires issus de marchés. Ces proxies sont utilisés en routine dans la plupart des SSA (Fritz et al., 2019).

1) **Indice de végétation :** l'indice de végétation par différence normalisée (NDVI) est extrait à partir du produit global MOD13Q1. Ce produit, issu du traitement de séries temporelles d'images satellites acquises par le satellite MODIS (MODerate resolution Imaging Spectroradiometer), fournit en tout point du globe des indices de végétation à une fréquence de 16 jours et à 250 mètres de résolution (NASA, 2020). Un masque de culture est utilisé pour ne considérer que les pixels situés dans les zones cultivées : le prototype S2 de la carte d'occupation du sol de l'Afrique en 2016, à 20 m de résolution spatiale (ESA, 2020).

2) **Estimation de précipitations :** les données pluviométriques utilisées sont décadaires et à une résolution de 6 km ; elles sont produites par le Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (CHIRPS, 2020) à partir d'images satellites et de données de terrain.

3) **Prix des aliments :** un historique mensuel des prix du maïs par marché produit par la Société Nationale de Gestion du Stock de Sécurité Alimentaire (SONAGESS) au Burkina Faso (communication personnelle, 2019).

Ces trois proxies sont tout d'abord agrégés mensuellement (en calculant le NDVI maximum et la somme des précipitations) et par province, en considérant les mois de mai à octobre (la saison agricole), puis transformés en anomalies (normalisés) en étant centrés réduits par rapport à l'ensemble des provinces et des mois présentés. Enfin, les anomalies sont agrégées en anomalies moyennes annuelles. Dans la section 1.4.2, nous analysons les corrélations entre les proxies de sécurité alimentaire normalisés (pluviométrie, NDVI et prix du maïs) et les indicateurs de sécurité alimentaire normalisés calculés à partir de l'EPA (SCA, SDA et $ISAr$), au niveau des provinces entre 2014 et 2017 (179 observations par indicateur).

1.3.4 Qualité et biais dans les données

Dans cette section, nous discutons tout d'abord des biais qui peuvent être présents dans les données d'enquêtes ménages, en nous penchant plus particulièrement sur le cas de l'EPA. La qualité des données de cette enquête conditionne directement les résultats de ce chapitre et les méthodes développées dans le chapitre 2. Puis, nous mentionnons les imprécisions qui peuvent être présentes dans les autres types de données utilisées comme proxies de la sécurité alimentaire.

1.3.4.1 Biais dans les enquêtes ménages

Dans toute enquête, différents biais peuvent survenir. Un biais est une erreur systématique, qui peut être méthodologique ou externe et qui affecte la fiabilité des résultats obtenus au cours d'une enquête. Dans le contexte de la sécurité alimentaire, les enquêtes ménages sont une source précieuse d'informations pour quantifier les caractéristiques nutritionnelles et comportementales des ménages en estimant des indicateurs spécifiques (e.g., SCA, SDA, $ISAr$). Pour estimer ces indicateurs avec précision, un certain nombre de biais dans les enquêtes ménages doivent être contrôlés, réduits si possible et, le cas échéant, pris en compte lors de l'analyse des résultats. Dans l'annexe A, nous dressons un panorama des principaux types de biais issus d'enquêtes ménages en nous appuyant sur les classifications proposées par plusieurs études (Winter, 2004 ; Dussaix, 2009 ; Biemer, 2010) : les biais de non-observation liés à la constitution d'un échantillon non représentatif de la population (e.g., erreur de couverture, d'échantillonnage, non réponse de certains participants), les biais d'observation dus à une erreur de mesure durant une enquête, les biais propres aux enquêtes pluriannuelles, ainsi que les biais liés au traitement et à l'analyse des données.

L'enquête EPA n'échappe pas aux biais sus-mentionnés. Les indicateurs de sécurité alimentaire obtenus lors de cette enquête nécessitent par exemple de poser un grand nombre de questions, parfois embarrassantes, aux répondants. Il peut en résulter des réponses non sincères, des non-réponses ou encore des abandons de la part des répondants, ce qui engendre des biais dans l'échantillon et dans la qualité des réponses. Dans l'annexe A, nous illustrons la présentation des biais propres aux enquêtes ménages en faisant une description détaillée de la méthodologie de l'enquête EPA et en donnant un aperçu des biais présents dans cette enquête. Un certain nombre de biais présents sont invérifiables, mais nous devons être conscients de leur existence. Ces biais peuvent se

répercuter sur la qualité des indicateurs de sécurité alimentaire issus de l'EPA, qui sont étudiés dans la suite de ce chapitre et utilisés dans le chapitre 2.

Concernant l'enquête ménages CFSVA, nous n'avons pas accès à la méthodologie employée pour la construction des indicateurs étudiés, mais nous supposons qu'ils sont soumis aux mêmes types de biais.

1.3.4.2 Qualité des autres types de données

La question de la qualité des données se pose également pour les proxies de la sécurité alimentaire utilisés dans cette thèse, dont aucun ne représente parfaitement la réalité à laquelle il se réfère. Par exemple la qualité des images obtenues à partir de capteurs optiques de satellites (e.g., produit NDVI considéré dans ce chapitre) dépendent des conditions atmosphériques au moment de l'acquisition des données, et des prétraitements appliqués aux séries temporelles ; les variables météorologiques mesurées par les instruments de stations météo ou encore la qualité biochimique des sols (que nous utiliserons dans le chapitre 2) évaluée par l'utilisation de microscopes sont dépendants de la fiabilité des instruments de mesure. D'autres données sont issues de modèles mathématiques dont le formalisme et le paramétrage comportent leur part d'erreur (e.g., estimation de précipitations dans ce chapitre, mais également pour le cas de modèles de densités locales de populations appliqués dans le chapitre 2 ou encore d'occupation du sol). Enfin, certains jeux de données ont des valeurs manquantes ou sont disponibles à une granularité spatio-temporelle inadéquate pour l'utilisation visée, ce qui implique le recours à méthodes d'imputation et/ou d'interpolation qui introduisent inévitablement de l'imprécision dans les variables obtenues. C'est le cas, par exemple, des données sur les prix du maïs utilisées dans ce chapitre, qui sont relevées sur les marchés des chefs-lieux provinciaux sur une base hebdomadaire, puis agrégées en moyennes mensuelles sans pondération par le volume des transactions de chaque marché ; ainsi leur variabilité intra-mensuelle (i.e., la variation des prix d'une semaine à l'autre) n'est pas quantifiable, et il en est de même pour leur variabilité intra-provinciale (i.e., les prix d'achat et de vente observés sur les marchés locaux).

Comme pour les indicateurs issus d'enquêtes ménages, les proxies de la sécurité alimentaire provenant de sources de données hétérogènes utilisés dans cette thèse sont donc sujets à des biais et à des imprécisions, dont la plupart sont invérifiables, nous

devons en avoir connaissance lorsque nous les analysons ou quand nous les utiliserons dans le chapitre 2.

1.4 Résultats et discussion

1.4.1 Analyse de la variabilité spatio-temporelle de la sécurité alimentaire

Dans cette section, nous analysons les variations spatio-temporelles des indicateurs SCA, SDA et $ISAr$ issus de l'enquête EPA et discutons de ces variations à la lumière des informations contenues dans des sources de données indépendantes telles que les bulletins et rapports de la sécurité alimentaire.

1.4.1.1 Variation temporelle de la sécurité alimentaire

Pour illustrer la variabilité interannuelle de la sécurité alimentaire au Burkina Faso, nous utilisons les valeurs annuelles du SCA et du SDA calculées à partir des données de l'EPA sur la période 2009-2017 (Figure 1.4). L'$ISAr$ n'est pas représenté ici, car une période de quatre ans (2014-2017) est trop courte pour effectuer une analyse temporelle. À l'échelle nationale, les résultats de l'EPA indiquent une augmentation des SCA et SDA moyens entre 2009 et 2013, suivie d'une forte baisse entre 2013 et 2017 (aboutissant à la valeur la plus basse de la période en 2017).

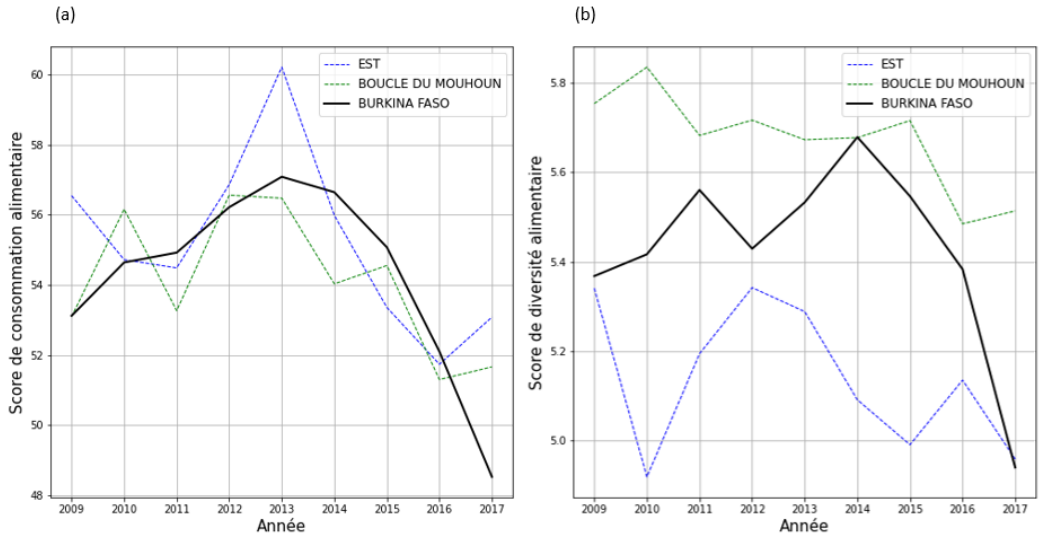


FIGURE 1.4 – Évolution du score de consommation alimentaire (SCA) moyen (a) et du score de diversité alimentaire des ménages (SDA) moyen (b) au Burkina Faso (lignes continues) et dans les régions Est et Boucle du Mouhoun (lignes pointillées) de 2009 à 2017. *Source : données de l'enquête permanente agricole (EPA)*

Depuis 2009, presque chaque année a connu des chocs (e.g., sécheresses, inondations, crises financières) qui ont pu affecter la sécurité alimentaire, mais trois événements se distinguent par leur gravité et ont été plus largement relayés par les journaux et les rapports des ONG : les inondations répétées de 2009 et 2010 (Burkina Faso Government, 2009 ; OCHA, 2010), la grave sécheresse de 2012 qui a conduit à une famine (WFP, 2012 ; World Bank, 2012), et la détérioration générale de la situation alimentaire qui a touché l'ensemble du pays depuis 2014 (FAO, 2017), s'aggravant globalement chaque année pour dégénérer en crise alimentaire et humanitaire (FAO, 2019). Concernant les inondations de 2009 et 2010, deux régions ont été particulièrement touchées, l'Est et Boucle du Mouhoun, qui comptaient parmi le plus grand nombre de victimes en 2009 (Burkina Faso Government, 2009). Cela est confirmé par les données qui révèlent que sur la période 2009–2011, alors que le Burkina Faso a connu une augmentation globale du SCA et du SDA, ces deux régions ont enregistré une diminution de ces indicateurs sur au moins une année entre 2010 et 2011 (Figure 1.4). L'effet de la sécheresse de 2012 est peu visible dans les données de l'EPA. Le SDA a légèrement diminué en 2012, mais le SCA a augmenté sur la même année. Cela peut s'expliquer par le fait que le SCA collecté

lors de cette enquête ne permet pas de capturer la diminution de l'apport calorique, qui est un indicateur clé pour reconnaître une crise alimentaire. L'$ISAr$ aurait pu capter cet effet, mais celui-ci n'est disponible qu'à partir de 2014. L'année 2013 (Burkina Faso Government, 2013) a été favorable sur le plan climatique, ce qui se traduit par une hausse du SCA et du SDA au niveau national. En revanche, les années 2014 (Burkina Faso Government, 2014) et 2015 (Burkina Faso Government, 2015) ont été des années déficitaires en pluviométrie, ce qui se traduit par la diminution du SCA et du SDA sur cette période. Dans l'ensemble, ces deux indicateurs semblent capables de capturer la variabilité interannuelle de la sécurité alimentaire.

1.4.1.2 Variation spatiale de la sécurité alimentaire

Pour analyser la variabilité spatiale de la sécurité alimentaire, nous présentons trois cartes qui représentent la distribution spatiale des indicateurs SCA, SDA et $ISAr$ au Burkina Faso sur la période 2014-2017 (Figure 1.5). La Figure 1.5 (a) illustre la distribution par région de la proportion de ménages ayant un faible SCA (i.e., <28). Les Figures 1.5 (b) et (c) montrent les distributions de l'$ISAr$ et du SDA moyens par région. Il n'existe pas de seuil pour déterminer les valeurs critiques du SDA (INDDEX Project, 2018) et de l'$ISAr$ (Maxwell, 2008), et les seuils de classe dans les Figures 1.5 (b) et (c) ont été fixés de manière à illustrer la variabilité régionale. Les résultats montrent qu'entre 2014 et 2017, les SCA et SDA étaient les plus faibles dans le centre du pays (régions Centre Nord, Plateau Central et Centre) et les plus élevés dans le sud-ouest (régions Cascades et Hauts-Bassins) et les régions sahéliennes, tandis que l'$ISAr$ était le plus critique dans le nord et l'est du pays (régions Centre Nord, Sahel et Est) et meilleur dans le sud-ouest (régions Cascades et Hauts-Bassins).

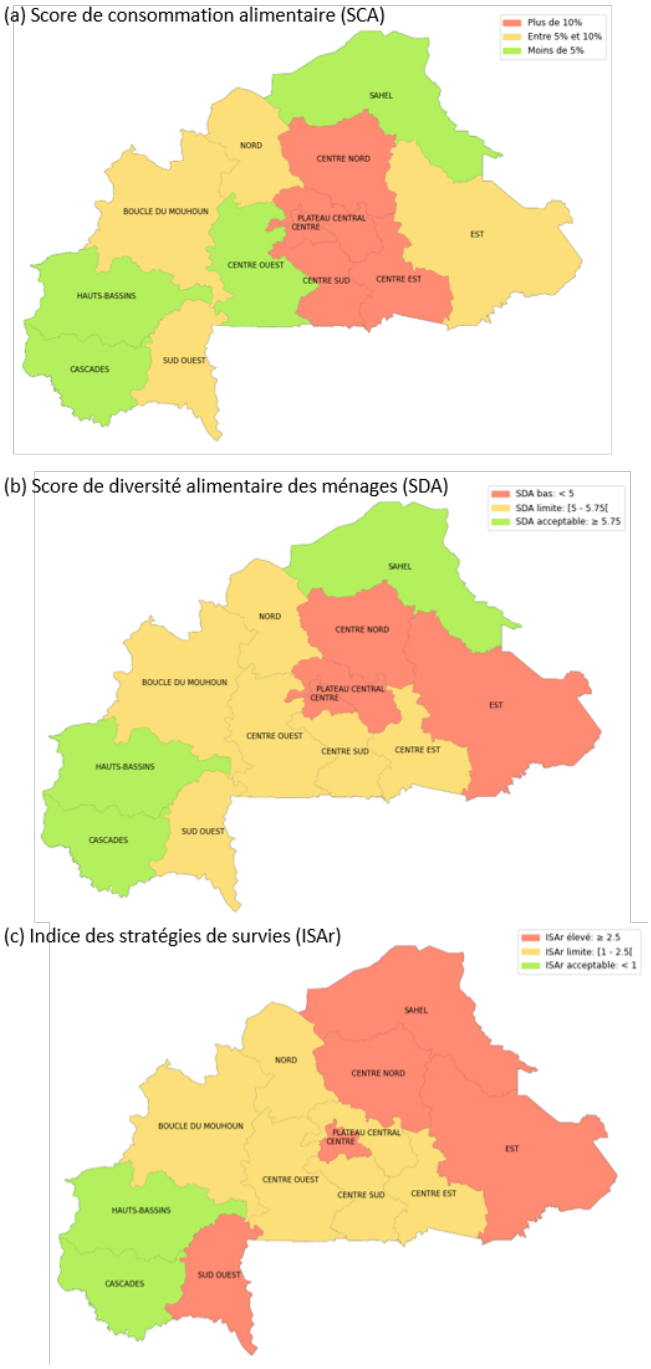


FIGURE 1.5 – Cartes à trois classes (a) de la proportion de ménages ayant un faible score de consommation alimentaire (SCA) (i.e., <28), (b) du score de diversité alimentaire des ménages (SDA) moyen, et (c) de l'indice de stratégies d'adaptation réduit ($ISAr$) moyen, calculés à l'échelle régionale pour la période 2014-2017. *Source : données de l'enquête permanente agricole*

Plusieurs bulletins indiquent que le Centre Nord, l'Est et le Sahel sont les régions où l'insécurité alimentaire est la plus forte. A l'inverse, les régions dans lesquelles la sécurité alimentaire est plus élevée sont les régions Hauts-Bassins, Cascades et Centre (WFP, 2014a ; Zida and Kambou, 2014 ; OCHA, 2015). Ceci reflète directement les conditions agroclimatiques qui sont favorables dans les régions du sud et défavorables dans les régions de l'est et du nord. Pour la plupart des régions, les données obtenues à partir de l'EPA sont cohérentes avec les rapports d'ONG : les régions Cascades et Hauts-Bassins présentent les meilleures valeurs de SDA (6,3 et 5,8, respectivement) et d'$ISAr$ (0,72 et 0,73, respectivement), tandis que les régions Est et Centre Nord ont les valeurs les plus critiques de SDA (4,98 et 4,6, respectivement) et d'$ISAr$ (2,54 et 2,47, respectivement). Dans la région du Sahel, les résultats sont plus contrastés : la région possède l'$ISAr$ le plus haut (2,76), ce qui est cohérent avec les rapports d'ONG et la classification structurelle de cette région comme étant en crise alimentaire, mais possède en même temps un SCA et un SDA plus élevés que la moyenne. Une interprétation de cette divergence peut être que le SCA et le SDA prennent fortement en compte certains groupes d'aliments tels que la viande et les produits laitiers davantage consommés dans cette zone et importants sur le plan nutritionnel, mais qui ne sont pas consistants en termes d'apport calorique. Cet exemple montre combien il est crucial d'utiliser plusieurs indicateurs pour expliquer un phénomène, et il apparaît que le SCA et le SDA sont a priori moins pertinents pour détecter les crises alimentaires au Sahel que l'$ISAr$. Enfin, la région Centre présente des valeurs critiques de SDA et d'$ISAr$ (4,83 et 2,57, respectivement), ce qui contredit ce qui apparaît dans les rapports d'ONG (Zida and Kambou, 2014 ; OCHA, 2015), plaçant cette région comme l'une des moins en proie à l'insécurité alimentaire. Une interprétation de ces divergences peut provenir du fait que dans la région Centre, la population des agriculteurs enquêtés n'est pas représentative de la population totale. En effet, contrairement aux autres régions plus rurales où environ 80% des travailleurs ont des emplois liés à l'agriculture, la région Centre contient la capitale Ouagadougou et son agglomération (Herrera and Ilboudo, 2012). Si les résultats de l'EPA indiquent que les agriculteurs de la région Centre sont en situation d'insécurité alimentaire, ce constat ne peut être généralisé au reste de la population en raison du biais de sous-dénombrement présent dans cette région.

Dans l'ensemble, il semble que le SCA, le SDA et l'$ISAr$ soient des indicateurs valables pour détecter les situations d'insécurité alimentaire qui peuvent survenir dans des régions spécifiques et qu'ils reflètent différentes situations d'insécurité alimentaire.

Alors que le SCA et le SDA tendent à indiquer des situations de régimes à faible densité nutritionnelle, l'$ISAr$ a plutôt pour effet de révéler des situations de faible apport calorique.

1.4.2 Cohérence entre l'EPA et d'autres types de sources de données sur la sécurité alimentaire

Nous examinons dans la suite de cette section si les résultats obtenus à partir de l'EPA sont cohérents avec les résultats obtenus à partir d'autres sources de données : l'enquête ménages CFSVA du PAM réalisée en 2018 d'une part et les proxies obtenus à partir d'images satellites, de données climatiques et de prix du marché alimentaire d'autre part.

1.4.2.1 Comparaison avec les données de l'enquête CFSVA

L'enquête ménages CFSVA est, au même titre que l'EPA, un moyen d'obtenir une image de la sécurité alimentaire à un moment donné sur le territoire burkinabé. Nous supposons que l'existence de concordances entre ces deux enquêtes est une indication supplémentaire de leur fiabilité. Nous comparons les indicateurs calculés à partir de l'enquête EPA de décembre 2017 à ceux calculés à partir de l'enquête CFSVA de février 2018. Les mois de décembre et de février appartiennent tous deux à la période post-récolte où les ménages agricoles ont le meilleur accès à la nourriture : les stocks de récolte sont encore importants et les prix de marché des aliments sont encore bas. Nous comparons les SCA, SDA et $ISAr$ par province (45) calculés séparément à partir de ces deux enquêtes.

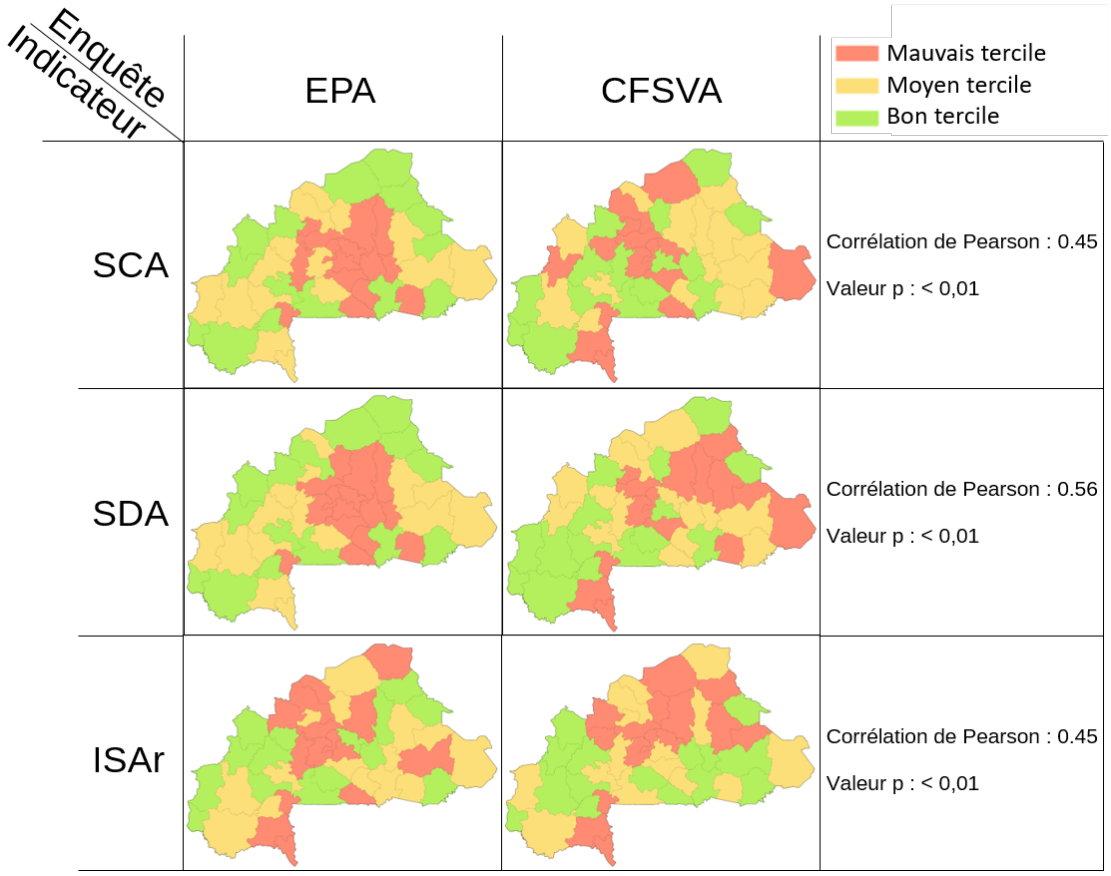


FIGURE 1.6 – Comparaison de la distribution spatiale par province des estimations du score de consommation alimentaire (SCA), du score de diversité alimentaire des ménages (SDA) et de l'indice des stratégies d'adaptation réduit ($ISAr$) de l'enquête agricole permanente (EPA) (décembre 2017) et de l'enquête Comprehensive Food Security and Vulnerability Analysis ($CFSVA$) (février 2018)

.

Pour chacun des trois indicateurs, les estimations obtenues à partir des deux enquêtes ménages sont significativement et positivement corrélées, avec des valeurs p de corrélations de Pearson inférieures à 0,01 (Figure 1.6). Cependant, les corrélations ne dépassent pas 0,56, ce qui signifie que pour un certain nombre de provinces, il existe des différences significatives dans la valeur des indicateurs des deux enquêtes. Nous considérons pour chaque indicateur (SCA, SDA et $ISAr$) et chaque enquête une division en trois classes de fréquence égale comme le montre la Figure 1.6. Pour le SCA et le SDA,

respectivement 18 (40%) et 20 (44%) provinces appartiennent à la même classe, contre respectivement 5 (11%) et 3 (7%) provinces qui ont été classées de manière opposée dans les deux classes extrêmes. Pour ces deux indicateurs, près de la moitié des 45 provinces (22) ont été classées avec une classe de différence (classe 1 et 2, ou classe 2 et 3, ou vice versa). Pour l'$ISAr$, 24 (53%) provinces ont été classées de manière identique, 16 (36%) provinces ont une classe de différence et 5 (11%) provinces sont classées de manière opposée. Nous observons donc une variation au moins modérée des résultats selon l'enquête pour la moitié des provinces. Mais les tendances générales sont les mêmes pour les deux enquêtes : SCA et SDA sont les plus faibles dans les provinces du centre et les plus élevés dans les provinces du sud-ouest et du Sahel, l'$ISAr$ est le plus critique dans les provinces du nord et est le meilleur dans les provinces de l'ouest. Nous concluons que malgré les différences méthodologiques et la présence inévitable de biais qui expliquent une partie des divergences observées, les principaux résultats de ces deux enquêtes sur la sécurité alimentaire des ménages sont globalement cohérents.

1.4.2.2 Comparaison avec trois proxies de la sécurité alimentaire

Dans cette section, nous considérons les corrélations entre les anomalies annuelles de trois indicateurs de sécurité alimentaire (SCA, SDA et $ISAr$) issus de l'EPA et de trois proxies de la sécurité alimentaire que sont les anomalies moyennes annuelles du NDVI, de la pluviométrie et du prix du maïs calculées pour l'année de collecte des indicateurs de sécurité alimentaire et pour l'année précédente (Tableau 1.4). Les indicateurs de sécurité alimentaire obtenus à partir des enquêtes ménages sont globalement peu corrélés avec les proxies étudiés, avec une corrélation de Pearson maximale de 0,3.

	NDVI Yt	**Précipitations Yt**	**Prix du maïs Yt**	**NDVI Yt-1**	**Précipitations Yt-1**	**Prix du maïs Yt-1**
SCA	-0.02	0.13	-0.16*	-0.02	-0.02	-0.13
SDA	0.16*	0.30***	-0.28***	0.17*	0.16*	-0.25***
ISAr	-0.21**	-0.01	0.19*	-0.20**	-0.29***	-0.01

Tableau 1.4 – Corrélations à l'échelle provinciale entre les anomalies annuelles de trois indicateurs de sécurité alimentaire (SCA, SDA et $ISAr$) et les anomalies moyennes annuelles du NDVI, de la pluviométrie et du prix du maïs pour l'année (Yt) au cours de laquelle les indicateurs de sécurité alimentaire ont été collectés et pour l'année précédente (Yt-1) entre 2014 et 2017 (* : valeur p <0,05 ; ** : valeur p <0,01 ; *** : valeur p <0,001)

Le **NDVI** est corrélé positivement avec le SDA et négativement avec l'$ISAr$, ce qui confirme l'effet attendu selon lequel de bonnes conditions culturales conduisent généralement à une production agricole élevée et donc à des revenus agricoles potentiels (grains stockés ou surplus commercialisé), ce qui se traduit par des apports caloriques plus élevés (pas besoin de mettre en place des stratégies de réduction alimentaire, donc l'$ISAr$ diminue) et une consommation alimentaire accrue (possibilité de manger des aliments plus variés, donc le SDA augmente). Cependant, un niveau élevé de consommation alimentaire n'est pas équivalent à un régime alimentaire à haute valeur nutritionnelle car la corrélation entre le NDVI et le SCA n'est pas significative. Cela peut être interprété par le fait que les ménages agricoles ont des régimes alimentaires plus diversifiés, mais que cette diversité peut être liée à la consommation d'aliments qui ne sont pas bénéfiques d'un point de vue nutritionnel, comme les huiles, le sucre et les boissons.

Les **précipitations** sont corrélées positivement avec le SDA et négativement avec l'$ISAr$, et la corrélation avec le SCA n'est pas significative. L'interprétation est la même que celle fournie pour le NDVI (i.e., production agricole plus élevée, moins de stratégies de réduction de la nourriture, consommation alimentaire accrue, mais pas nécessairement plus nutritive).

Les **prix du maïs** sont corrélés négativement avec le SDA et positivement avec l'$ISAr$. Si l'on considère que la plupart des ménages agricoles sont des acheteurs nets (i.e., ceux-ci produisent moins de maïs que ce qu'ils consomment, et doivent donc acheter une partie de leur consommation de maïs sur le marché), cela confirme l'intuition qu'une augmentation des prix du maïs est préjudiciable à la sécurité alimentaire car les ménages ont un accès économique plus faible au maïs et sont obligés de réduire leur consommation de céréales et d'autres produits alimentaires (augmentation de l'$ISAr$ et diminution du SDA).

Dans l'ensemble, il apparaît que le SCA est très peu corrélé avec les proxies de la sécurité alimentaire et que le SDA et l'$ISAr$ sont plus fortement corrélés avec ces proxies. Les corrélations du SDA et de l'$ISAr$ avec les proxies de la même année sont significatives (5 corrélations significatives sur 6) ainsi que les corrélations avec les proxies de l'année précédente (5 corrélations significatives sur 6). Cela indique que la sécurité alimentaire est liée non seulement au contexte climatique et économique de l'année en cours, mais aussi au contexte de l'année précédente.

Nous pouvons conclure de cette analyse qu'il existe des concordances entre les indicateurs et les proxies de la sécurité alimentaire étudiés, mais que ces concordances sont limitées. Tout d'abord, car les trois proxies utilisés dans cette étude sont insuffisants pour capter toute la complexité de la sécurité alimentaire. Mais aussi probablement parce que la sécurité alimentaire est le résultat d'une combinaison complexe de ces proxies qui ne peut pas être entièrement capturée par des calculs de corrélations.

Ces résultats indiquent la nécessité d'intégrer un plus large éventail de données et de les traiter avec des méthodes plus subtiles afin de révéler des concordances plus précises. Dans le chapitre 2, nous mettons en relation les indicateurs dérivés de l'EPA avec un ensemble plus large et hétérogène de proxies de la sécurité alimentaire, en recourant à des méthodes plus sophistiquées de la science des données afin de mettre en lumière des associations plus riches et fines.

1.5 Conclusion

Ce chapitre analyse les contributions des enquêtes ménages pour la compréhension des situations de sécurité alimentaire. Grâce à des études antérieures, nous savions que les indicateurs dérivés d'enquêtes ménages pouvaient être considérés comme fiables pour quantifier la consommation alimentaire, et qu'ils sont moins longs et coûteux à obtenir que les indicateurs de consommation individuelle. Dans ce chapitre, nous avons montré que les indicateurs dérivés d'enquêtes ménages contiennent des informations spatiales et interannuelles cohérentes et qu'ils peuvent être utilisés pour suivre les crises alimentaires à l'échelle sub-nationale malgré leurs biais inhérents. En particulier, nous avons établi que les enquêtes ménages fournissent des informations sur la sécurité alimentaire qui permettent d'identifier des tendances cohérentes avec les SSA, ainsi qu'avec des variables climatiques et économiques liées à la sécurité alimentaire.

Cependant, nous notons par ailleurs deux inconvénients liés aux enquêtes ménages. Le premier inconvénient concerne le temps et le coût nécessaires pour obtenir des indicateurs à partir de ces enquêtes, qui bien que moins élevés que pour les enquêtes individuelles, restent importants. Cela est une conséquence des importants moyens déployés pour effectuer les tâches nécessaires : la définition de la stratégie d'échantillonnage, la conception éventuelle du questionnaire, les enquêtes auprès d'un nombre considérable de ménages, et enfin la saisie, le nettoyage et l'analyse des données par des experts, ce

qui peut également constituer des sources de biais. À l'inverse, les données sur l'état des cultures et les précipitations obtenues à partir d'images de télédétection peuvent être obtenues plus rapidement et souvent à moindre coût, ce qui constitue un avantage majeur à ce niveau. Le second inconvénient provient du fait que les indicateurs issus de ces enquêtes ne fournissent qu'une vision partielle des situations de sécurité alimentaire, qu'il est difficile de relier pleinement à d'autres sources de données par des méthodes classiques. Plus précisément, les variations des indicateurs de sécurité alimentaire mises en évidence dans ce chapitre ne sont que partiellement liées avec les facteurs climatiques et économiques considérés ici, sans que cela ne puisse être entièrement attribué aux biais et imprécisions présents dans toutes ces données. Cela peut plutôt s'expliquer par le fait que, d'une part, les méthodes d'analyse utilisées sont traditionnelles, principalement basées sur des calculs d'anomalies et de corrélations linéaires ne pouvant mettre en évidence des associations complexes, et que d'autre part les trois proxies considérés ici se concentrent sur deux composantes très spécifiques de la sécurité alimentaire (disponibilité et accessibilité économique) et sont insuffisants pour appréhender notre problématique dans toute son étendue.

Pour ces raisons, nous nous concentrons dans le Chapitre 2 sur le développement de méthodes sophistiquées capables d'extraire des règles d'association complexes entre différents types de données : d'une part, les indicateurs de sécurité alimentaire qui contiennent, comme nous l'avons observé, des informations spatiales et temporelles cohérentes sur les tendances et les chocs alimentaires ; d'autre part, un ensemble large et hétérogène de proxies de la sécurité alimentaire (e.g., climatiques, agronomiques, économiques, démographiques, liés à la sécurité) contenant des informations complémentaires. Cela soulève plusieurs questions quant au choix de ces données hétérogènes et des méthodes de traitement et de combinaison pertinentes des données à mettre en œuvre. Pour tenter de répondre à ces questions, nous exploiterons des méthodes d'apprentissage automatique et profond ainsi que le concept de fusion de données, nécessaire pour combiner des données hétérogènes. Ces méthodes d'exploration de données sont de plus en plus utilisées pour extraire des informations pertinentes de données complexes, hétérogènes et spatio-temporelles, mais ont été relativement peu exploitées pour les domaines liés à la sécurité alimentaire. Ce type de recherche peut contribuer aux SSA, afin de permettre une détection plus rapide et précise des famines. Par exemple, via la fourniture, à termes, d'outils capables de prédire les scores d'indicateurs de sécurité alimentaire quelques semaines avant qu'ils ne soient produits, en utilisant des données d'entrée ouvertes et plus

facilement accessibles.

Chapitre 2

Prédire la sécurité alimentaire à partir de données hétérogènes

2.1 Introduction

Pour prédire et analyser les situations d'insécurité alimentaire de la manière la plus complète possible, les systèmes d'alerte et de surveillance de la sécurité alimentaire (SSA) sont principalement basés sur des approches qui requièrent la combinaison et la synthèse manuelles des sources d'information prises en compte (principalement des données agro-climatiques, économiques et nutritionnelles), selon une série de règles prédéfinies (Fritz et al., 2019). Si d'un côté, la nécessité de s'appuyer principalement sur une intervention humaine peut être justifiée par la complexité de la tâche à accomplir, d'un autre côté cela représente également un obstacle à des prédictions des crises alimentaires précises et dans des délais raisonnables. Plus précisément, le processus actuel qui sous-tend ces SSA est très coûteux en temps et ne permet qu'une complexité limitée quant au nombre et à l'hétérogénéité des sources d'informations pouvant être considérées. Par ailleurs, les SSA prennent en compte certains indicateurs clés de la sécurité alimentaire, tels que le score de consommation alimentaire (SCA) ou le score de diversité alimentaire des ménages (SDA) (cf. section 1.2.4 pour une description détaillée de ces mesures), qui comme nous l'avons démontré dans le chapitre précédent, fournissent des informations spatiales et interannuelles cohérentes qui peuvent être utilisées pour suivre les crises alimentaires au niveau sub-national. Mais nous avons également constaté que ces indicateurs nécessitent des enquêtes ménages qui représentent un coût important en termes de temps

et de ressources. Enfin, ces systèmes se basent également sur des indicateurs issus de domaines indirectement liés avec la sécurité alimentaire, que nous appelons "proxies", permettant d'appréhender la sécurité alimentaire et ses causes de la manière la plus complète possible. Les proxies intégrés par ces systèmes sont principalement issus de données agroclimatiques et économiques et pourraient être enrichis par d'autres types de données (e.g., conflits, répartition des hôpitaux et des écoles, dynamique des populations, articles de presse) pour obtenir une vision plus large de ce concept complexe et multifactoriel.

Sachant ce contexte et afin d'apporter une contribution pertinente, nous développons dans ce chapitre plusieurs axes de réflexion. Avec le développement du numérique qui conduit à une ouverture croissante des données sur le web, de plus en plus de données produites par des organismes aussi bien publics que privés sont accessibles à tous et réutilisables. Ces données ont pour avantage de pouvoir être plus rapidement intégrées aux SSA et à moindre coût que les données non ouvertes qui peuvent nécessiter de longs processus d'autorisation avant d'être partagées. Nous pensons que l'intégration d'informations provenant d'autres domaines liés à la sécurité alimentaire et de structures plus complexes (e.g., séries temporelles, images à haute résolution spatiale) issus de ces données ouvertes permettra une description plus précise et plus précoce des aspects structurels et conjoncturels de la sécurité alimentaire. Pour exploiter efficacement ces données hétérogènes en termes de thématique, de structure et de résolution spatio-temporelle (cette notion de niveaux d'hétérogénéité est détaillée dans la section 2.3.2), les méthodes d'apprentissage automatique constituent une approche appropriée, car elles permettent des analyses automatiques guidées par de nombreux types de données sans a priori. Ce domaine contient plusieurs familles de méthodes qui nous seront utiles dans cette thèse. L'apprentissage profond permet de faire des prédictions à partir de données d'entrée traitées avec un haut niveau d'abstraction par des architectures articulées de nombreuses transformations non linéaires (Huang et al., 2019). Ces méthodes permettent de prendre en compte des interactions complexes dans les données explicatives. Le "representation learning", souvent effectué par des méthodes d'apprentissage profond, permet de générer une nouvelle représentation de données étudiées, qui est plus appropriée pour une tâche donnée (Bengio et al., 2013). Cette nouvelle représentation des données peut par exemple être considérée comme entrée de modèles d'apprentissage automatique. D'un point de vue opérationnel, l'aspect automatique de ce type d'approche permet d'augmenter la vitesse de traitement et d'analyse des données sans affecter la qualité des prévisions, ce qui peut constituer un apport pertinent pour les systèmes d'alerte précoce. D'un point

de vue technique, il existe une multitude de modèles d'apprentissage automatique spécifiquement conçus pour traiter chaque type de données de manière efficace, ce qui en fait des outils pertinents pour traiter l'hétérogénéité des données impliquée par l'aspect multifactoriel de la sécurité alimentaire. Par ailleurs, les méthodes d'apprentissage automatique sont guidées par les données et permettent d'extraire des règles de décision qui ne pourraient pas être identifiées par l'œil humain parce que trop complexes ou contre-intuitives. Cependant, ce type d'approche se heurte à plusieurs verrous scientifiques. Premièrement, le choix des méthodes à appliquer à chaque type de données, sachant qu'il existe dans chaque cas une multitude de méthodes présentant des avantages et des limites (e.g., pouvoir prédictif, pouvoir explicatif, sensibilité aux petits échantillons, sensibilité au bruit). Deuxièmement, la calibration des modèles dépend de nombreux facteurs (e.g., taille, type et qualité des données), et il n'existe pas actuellement de règles de calibration spécifiques pour certains types de modèles récents, comme dans le cas de l'apprentissage profond. Troisièmement, aucun modèle d'apprentissage automatique ne peut traiter à lui seul tous les types de données. Plusieurs modèles adaptés à chaque donnée doivent être combinés sans qu'il n'existe non plus de méthode formelle pour y parvenir.

L'objectif principal de ce chapitre est de recourir à des méthodologies originales et efficaces basées sur de l'apprentissage automatique et profond, capables d'estimer le SCA et le SDA à partir de données hétérogènes ouvertes. Plus précisément, nous visons à répondre aux questions de recherche suivantes :

- **Q1** : quels types de données ouvertes doivent être ciblés afin de prédire des scores de sécurité alimentaire.
- **Q2** : comment des données hétérogènes en termes de thématique, de structure et de résolution spatio-temporelle peuvent-elles être considérées afin d'obtenir des prédictions cohérentes sur la sécurité alimentaire pour un site d'étude donné.
- **Q3** : comment les approches d'apprentissage automatique et profond peuvent-elles être exploitées et combinées afin de traiter des données hétérogènes.

Dans le but de répondre à ces questions, nous proposons un framework d'apprentissage automatique, intitulé $FSPHD$ (Food Security Prediction based on Heterogeneous Data), capable d'exploiter des données explicatives hétérogènes utilisées comme proxies afin d'estimer deux indicateurs cibles de la sécurité alimentaire, i.e., le SCA et le SDA. Afin de répondre à **Q1**, nous prenons en compte des données hétérogènes ouvertes multi-

sources telles que des variables quantitatives (e.g., variables économiques de la Banque mondiale, données météorologiques), des rasters (e.g., densités de population, occupation du sol, qualité des sols), des données géolocalisées (e.g., hôpitaux, écoles, événements violents), des vecteurs lignes (cours d'eau), ainsi que des séries temporelles (température de brillance lissée (SMT), estimations des précipitations, prix du maïs). Pour répondre à **Q2**, le framework proposé repose sur un ensemble de techniques sophistiquées de science des données, chacune étant adaptée à des types de données spécifiques : les forêts aléatoires (FA) (Qi, 2012) appropriées aux données quantitatives classiques ; les réseaux de neurones convolutifs (CNN) (Huang et al., 2018) adaptés aux données à haute résolution spatiale ; les réseaux de neurones récurrents à mémoire court-terme et long terme (LSTM) (Song et al., 2020) permettant un traitement efficace de données séquentielles. Dans le but de répondre à **Q3**, c'est-à-dire de trouver une méthode optimale pour combiner les différentes sources de données, nous testons différentes variantes du framework qui diffèrent par le nombre et le type de données d'entrée, et par la manière dont l'ensemble des méthodes appliquées aux données sont traitées et combinées afin d'obtenir un résultat global. Pour cela, nous nous appuyons sur la théorie de la fusion de données, qui catégorise les fusions possibles à trois niveaux : données, caractéristiques et décisions (Brena et al., 2020 ; Hall and Llinas, 2017). Nous effectuons une évaluation expérimentale approfondie centrée sur la zone d'étude du Burkina Faso, en nous appuyant sur les indicateurs SCA et SDA (calculés à partir des données de l'EPA menée par le gouvernement burkinabé) comme vérité terrain.

Le reste du chapitre est structuré comme suit : dans la section 2.2, nous passons en revue les travaux utilisant de l'apprentissage automatique pour des problèmes de sécurité alimentaire, puis pour le traitement de données hétérogènes. Dans la section 2.3, nous décrivons les données utilisées dans l'étude ainsi que le framework proposé, dans la section 2.4, nous présentons, puis discutons des résultats de notre étude expérimentale.

2.2 État de l'art

2.2.1 Apprentissage automatique pour la sécurité alimentaire et les problèmes connexes

Les méthodes d'apprentissage automatique sont de plus en plus utilisées pour extraire des informations pertinentes dans le contexte des problèmes liés à l'insécurité

alimentaire. Plusieurs études exploitent des techniques classiques d'apprentissage automatique (e.g., machines à vecteurs de support, méthodes des k plus proches voisins, arbres de décision et classifieurs Bayésiens naïfs) pour la prédiction d'indicateurs de sécurité alimentaire (e.g., échelle de l'accès déterminant l'insécurité alimentaire des ménages et apport calorifique) (Okori and Obua, 2011 ; Barbosa and Nelson, 2016 ; Lukyamuzi et al., 2018). Mais ces méthodes ne permettent pas de traiter en profondeur des données à la structure complexe propres à cette thématique et de prendre pleinement en compte leur fort potentiel prédictif.

Ces dernières années, les techniques d'apprentissage profond, qui se sont révélées particulièrement efficaces pour analyser des données hétérogènes complexes (Valdés, 2018), ont également été utilisées pour l'analyse de plusieurs sujets liés à la sécurité alimentaire. Yeh et al. (2020) ont abordé la thématique de la pauvreté, ils ont prédit par régression le revenu moyen de ménages africains en appliquant un CNN sur des images satellites multi-spectrales publiquement accessibles. Shailesh et al. (2018) ont également prédit le revenu moyen (de ménages indiens) par l'utilisation de CNN. Dans leur étude, les descripteurs spatiaux utilisés sont des images satellites transformées (type de toit, luminosité de nuit, surfaces en eau). Mumtaz et al. (2018) se sont intéressés au sujet de la sécheresse, ils ont prédit l'indice de précipitation normalisé (SPI) qui est un indicateur de la rareté des précipitations en créant un framework intégrant un modèle linéaire ainsi que des réseaux de neurones classiques nommés "perceptron multicouche" (MLP) (Gardner and Dorling, 1998) et "réseaux de neurones flou" (Buckley and Hayashi, 1994). Les variables explicatives utilisées sont exclusivement des séries temporelles mensuelles de précipitations, de température et d'humidité. Min et al. (2019) ont prédit la valeur de marchés boursiers à partir de séries temporelles journalières des valeurs précédentes. Ils extraient tout d'abord des motifs séquentiels fréquents dans les séries temporelles avec un algorithme nommé "déformation temporelle dynamique" (Berndt and Clifford, 1994). Puis vient une phase d'extraction de caractéristiques à partir des motifs séquentiels avec un CNN. La prédiction est enfin obtenue en appliquant un MLP sur les caractéristiques extraites avec le CNN. Toutes ces études appliquent des méthodes d'apprentissage profond sur des données complexes avec une certaine efficacité, mais celles-ci ne sont pas hétérogènes et les méthodes employées se focalisent exclusivement sur un aspect temporel ou spatial sans combiner ces deux composantes toutes deux essentielles dans notre contexte. Par ailleurs, le réel potentiel de ces méthodes ne s'est pas encore exprimé pour la prédiction d'indicateurs de sécurité alimentaire, et très peu d'études se sont penchées

sur cette question.

Abordons maintenant les études qui se sont penchées sur la prédiction d'indicateurs de sécurité alimentaire. Heisenberg et al. (2020) utilisent un MLP pour prédire un indicateur appelé "cadre intégré de classification de la sécurité alimentaire" (IPC) obtenu à partir d'une enquête ménages menée par l'Agence américaine pour le développement international (USAID) dans la Corne de l'Afrique entre 2009 et 2017. Ils utilisent des données explicatives provenant de divers domaines (e.g., NDVI (indice de végétation par différence normalisée), prix des aliments, conflits et humidité des sols). Un petit nombre d'études se concentrent sur des variables directement liées à la quantité et à la qualité de la consommation alimentaire, telles que le SCA ou le SDA (cf. section 1.2.4), qui sont cruciales pour la compréhension de la sécurité alimentaire. Le principal exemple dans ce contexte est le framework développé par l'équipe VAM (analyse et surveillance des vulnérabilités) du Programme alimentaire mondial (PAM) (WFP-VAM, 2019). Ce framework intègre des techniques d'apprentissage automatique et profond sur des données hétérogènes (i.e., des images satellites à différentes résolutions spatiales et des données géolocalisées) pour prédire le SCA et le SDA dans plusieurs pays. Néanmoins, les résultats varient fortement d'un pays à l'autre, et les prédictions ne sont généralement pas assez précises pour être utilisées dans des cas opérationnels, ce qui semble être une limite commune à la plupart des études comparables. Lentz et al. (2019) prédisent également le SCA et le SDA avec des régressions linéaires utilisant principalement des données ouvertes et gratuites comme prédicteurs. Les variables réponses sont issues des enquêtes sur la mesure des niveaux de vie (LSMS) menées au Malawi en 2010 et 2013. Les données utilisées proviennent de sources diverses : météorologie, précipitations, prix du marché et qualité des sols. Cependant, la qualité des prédictions obtenues est relativement faible, ce qui confirme que la prédiction de ces indicateurs de sécurité alimentaire est une tâche complexe, principalement en raison de leur nature multifactorielle. Dans ce chapitre, nous souhaitons contribuer à la compréhension de ce problème peu étudié.

En outre, l'un des objectifs des deux études mentionnées ci-dessus (i.e., Lentz et al. (2019) et WFP-VAM (2019)) est de pouvoir utiliser des données gratuites et facilement accessibles pour obtenir des estimations d'indicateurs qui ne sont pas communément disponibles car leur calcul nécessite des enquêtes ménages longues et coûteuses. Cette problématique des plus actuelles, qui constitue l'un des objectifs de cette thèse, a été prise en compte par plusieurs autres études qui ont utilisé des données ouvertes comme variables explicatives pour prédire des indicateurs liés à la sécurité alimentaire. Jean

et al. (2016) appliquent un CNN sur des images satellites à haute résolution spatiale publiquement accessibles pour prédire la luminosité de nuit qui est un proxy de l'activité économique régionale. van der Heijden et al. (2018) prédisent l'IPC en utilisant une forêt aléatoire sur des données hétérogènes ouvertes (e.g., altitude, occupation du sol, NDVI, densité de population, prix des denrées). Mais leur approche, comme celles des quelques autres études citées précédemment utilisant des données hétérogènes, sont basées sur le traitement de l'intégralité de leurs données avec une unique méthode d'apprentissage automatique, ce qui ne permet pas d'exploiter le plein potentiel qu'offrent leurs données hétérogènes. Dans la section suivante, nous examinons des études qui ont eu différentes approches possibles pour faire face à cette problématique.

2.2.2 Apprentissage automatique sur des données hétérogènes

En raison de la complexité inhérente au phénomène de l'insécurité alimentaire, la prédiction d'indicateurs nécessite l'utilisation de variables explicatives aux thématiques, structures et échelles spatio-temporelles hétérogènes. C'est pourquoi, pour répondre à cette problématique, il est nécessaire de disposer de méthodologies capables de traiter et de combiner ces variables explicatives de telle sorte que chaque source d'information contribue de manière complémentaire à la prédiction des indicateurs de sécurité alimentaire.

Pour traiter conjointement des données hétérogènes avec des méthodes d'apprentissage automatique, il est nécessaire d'appliquer des techniques de fusion de données car il n'existe pas de méthode d'apprentissage automatique générique adaptée à tous les types de problèmes et de données, mais plutôt des méthodes mieux adaptées aux spécificités de chaque problème (e.g., classification, régression et détection d'anomalies) et de chaque source de données (e.g., variables quantitatives, séries temporelles, images et textes) (Alzubi et al., 2018). La fusion de données est un processus qui consiste à combiner plusieurs sources de données pour produire des informations plus précises, plus robustes et moins redondantes qu'en considérant chaque source de données individuellement (Khaleghi et al., 2013 ; Chandrasekaran et al., 2017). Les données peuvent être fusionnées à trois niveaux (Brena et al., 2020 ; Hall and Llinas, 2017) : 1) au niveau des données (fusion précoce) en concaténant simplement les variables initiales pour obtenir un ensemble de données contenant une plus grande quantité d'informations qui peuvent être mises en

entrée d'un algorithme d'apprentissage automatique, 2) au niveau des caractéristiques (fusion des caractéristiques) en extrayant des caractéristiques plus pertinentes de chaque source de données par des méthodes de "representation learning" et en les concaténant pour les mettre en entrée d'un algorithme d'apprentissage automatique, et 3) au niveau des décisions (fusion tardive) en agrégeant les prédictions des modèles associés à chaque source de données en une prédiction globale. Ce principe de fusion de sources de données multiples a été appliqué à l'apprentissage automatique au niveau des caractéristiques par de nombreuses études (Xue et al., 2017 ; Amin et al., 2018 ; Benedetti et al., 2018). Par exemple, Xue et al. (2017) proposent DeepFusion, un framework d'apprentissage profond multicapteur pour la reconnaissance de l'activité humaine utilisant de l'extraction de caractéristiques au moyen de CNN et de LSTM pour apprendre des représentations porteuses d'informations complexes issues de données sensorielles hétérogènes. La fusion de données a également été appliquée à l'apprentissage automatique au niveau des décisions par un nombre conséquent d'études (Peterson et al., 2018 ; Guo et al., 2019 ; Kumar et al., 2020). Par exemple, Guo et al. (2019) proposent le framework iFusion pour la classification de données médicales, qui utilise des CNN pour traiter et combiner des données hétérogènes en temps réel au niveau des décisions. Ils utilisent séparément chaque type de nouvelles données pour former un nouveau modèle de classification et fusionnent les prédictions des modèles précédemment formés pour obtenir la classification finale.

De telles approches basées sur ces concepts de fusion de données afin d'appliquer des méthodes d'apprentissage automatique qui combinent des sources de données hétérogènes ont été utilisées dans de nombreux autres domaines d'application, et nous en détaillons plusieurs ci-dessous. En médecine, Miotto et al. (2017) passent en revue la littérature récente et déjà abondante sur l'application des technologies d'apprentissage profond (e.g., MLP, LSTM, CNN, Machines de Boltzmann restreintes (Smolensky, 1986)) pour obtenir des connaissances et des informations pratiques à partir de données biomédicales complexes, volumineuses et hétérogènes. Les applications mentionnées dans ce domaine sont diverses : imagerie médicale (Esteva et al., 2017), prédiction des maladies à partir de dossiers cliniques (Cheng et al., 2016) ou encore outils de suivi de la santé de patients en temps réel (Shameer et al., 2017). En biochimie, P. Lewis et al. (2006) font des inférences à partir de grands ensembles hétérogènes de séquences et de structures de protéines en utilisant des techniques de machine à vecteurs de support (SVM). Ils prédisent la fonctionnalité de plusieurs gènes en combinant par somme pon-

dérée deux noyaux de SVM (Mismatch (Leslie et al., 2003) et MAMMOTH (Ortiz et al., 2002)) appliqués conjointement aux séquences et à la structure de protéines. Dans le contexte du traitement de vidéos, Hori et al. (2017) mettent en place un programme de description automatique de vidéos. Les vidéos sont décomposées en séquences d'images et de sons, qui sont mises en entrée de CNN pour en extraire des séquences de caractéristiques. Ces séquences sont appliquées à un mécanisme d'attention (Hernández and Amigó, 2021) dans le but de leur attribuer un poids relatif à leur pertinence, et enfin mises en entrée d'un LSTM qui génère la description textuelle de la vidéo. Dans le secteur routier, Yuan et al. (2018b) réalisent une étude approfondie du problème de la prédiction des accidents de la circulation à l'aide d'un CNN à mémoire court-terme et long terme pour prendre en compte simultanément l'hétérogénéité spatiale et l'autocorrélation temporelle de l'environnement. Dans le domaine économique, Wang et al. (2021) ont pour objectif la prédiction des prix de biens immobilier. Pour prédire le prix de maisons, plusieurs types de données sont exploitées. Des données vectorielles représentant la distribution spatiale de lieux publics (écoles, parcs, stations de transports en commun) sont intégrées, permettant d'obtenir la distance de maisons à ces lieux. Des images satellites multi-spectrales sont traitées par un réseau de transformateurs spatiaux (Jaderberg et al., 2015) suivi d'un CNN qui en extrait des caractéristiques. Enfin, des données quantitatives et qualitatives relatives aux caractéristiques de la maison (e.g., type de maison, ancienneté, superficie) sont prises en compte. Les variables résultantes sont mises en entrée d'un mécanisme d'attention afin d'attribuer automatiquement des poids aux différentes variables selon leur importance, puis intégrées à un MLP pour la prédiction du prix de la maison. Dans le contexte de la télédétection, Benedetti et al. (2018) proposent un framework d'apprentissage profond appelé M3Fusion pour la prédiction de classes d'occupation du sol, capable d'exploiter simultanément les informations conjoncturelles contenues dans des séries temporelles et les informations structurelles présentes dans des données à très haute résolution spatiale à l'aide de LSTM et de CNN. Mentionnons une autre application issue de la télédétection, Han et al. (2021) proposent un outil de classification de types de banquises à partir d'images satellites optiques et radar. Ces deux sources de données sont chacune traitées avec un CNN adapté qui en extraient des caractéristiques : le CNN appliqué aux données radar contient une couche de regroupement des pyramides spatiales (He et al., 2015), et le CNN utilisé pour les données optiques est de type PANet (Liu et al., 2018). Enfin, les caractéristiques obtenues sont regroupées, puis mises en entrée d'un MLP qui effectue la classification du type de banquise.

D'autres études incluent également des données textuelles dans leur framework. Ce type de données peut apporter des informations distinctes et complémentaires de celles obtenues par des données quantitatives ou par des images. Chai et al. (2020) proposent une méthode de prédiction de prix de matières premières qui prend en compte trois types de données : des données d'opérations boursières, des articles de journaux d'actualité ainsi que des commentaires d'investisseurs. Les articles de journaux ainsi que les commentaires d'investisseurs sont associés à des scores de sentiments par une approche basée sur un lexique, puis ces scores sont combinés avec les séries temporelles issues des opérations boursières pour être utilisés comme paramètres dans un modèle de Markov caché (Eddy, 2004) pour la prédiction des prix. Kumar et al. (2020) effectuent dans leur étude de la classification supervisée de sentiments positifs ou négatifs relatifs à des images associées à une description textuelle. Les images sont transformées en caractéristiques avec une méthode de sac de mots visuels (Yang et al., 2007), puis classifiées par un SVM. Les descriptions textuelles sont traitées de deux manières différentes : d'une part, elles sont vectorisées avec l'outil Glove (Pennington et al., 2014), puis un CNN leur attribue un score de sentiment ; et d'autre part l'outil d'analyse de sentiment VADER (Valence Aware Dictionary and Sentiment Reasoner) (Gilbert, 2014) basé sur des règles leur attribue un second score de sentiment. La classification finale est obtenue par fusion sur les décisions en deux temps, tout d'abord en agrégeant les deux scores issus des descriptions textuelles par conversion en valeur angulaire puis sommation, cette valeur obtenue est enfin elle-même agrégée avec la sortie associée à l'image par un système de décision booléen (Kunz et al., 2002). Sheehan (2018) s'intéressent à un sujet connexe à la sécurité alimentaire en utilisant la notion de fusion de données, et prédisent un indice de pauvreté : l'indice de richesse inclusive. Ils utilisent comme descripteurs des images satellites de luminosité de nuit ainsi que le contenu textuel des pages Wikipédia géolocalisés aux points les plus proches des lieux où l'indice de pauvreté est connu. Des caractéristiques sont extraites de ces deux types de données : par application d'un MLP aux images, et par vectorisation des articles avec la méthode Doc2vec (Mikolov et al., 2013). Les caractéristiques obtenues sont enfin mises en entrée d'un MLP pour la prédiction de l'indice de pauvreté. Cette problématique liée à l'intégration de données textuelles semble pertinente dans notre contexte. En effet, ce type de données très différentes de celles communément utilisées par les SSA, peuvent fournir des informations complémentaires, plus qualitatives, et susceptibles de nous éclairer davantage sur les causes et les caractéristiques des famines. Nous aborderons cet aspect textuel plus en détails dans le chapitre 3, qui s'y consacrera.

2.2.3 Apprentissage automatique sur des données hétérogènes pour la sécurité alimentaire

Concentrons-nous enfin sur les études qui appliquent des méthodes d'apprentissage automatique sur des données hétérogènes pour la prédiction d'indicateurs de sécurité alimentaire. La plupart des rares études qui abordent cette problématique effectuent des fusions au niveau des données pour traiter leurs données hétérogènes, i.e., prétraitent leurs données afin de les combiner conjointement avec le même algorithme d'apprentissage automatique ou profond, ce qui est le cas pour les études de van der Heijden et al. (2018), Lentz et al. (2019) et de Heisenberg et al. (2020) décrites ci-dessus. Cependant, la fusion au niveau des données est la plus naïve et la plus simple à mettre en œuvre mais ne permet pas d'extraire les informations de chaque type de données de manière optimisée (Hall and Llinas, 2017). Les fusions aux niveaux des caractéristiques et des décisions ont très peu été exploitées pour les problèmes liés à la sécurité alimentaire, un domaine multifactoriel pour lequel la prédiction d'indicateurs nécessite le traitement approprié de données explicatives de thématiques et d'échelles spatio-temporelles très hétérogènes. L'étude de Sheehan (2018) décrite précédemment effectue de la fusion au niveau des caractéristiques, mais les descripteurs pris en compte (articles Wikipedia et luminosité de nuit) sont insuffisants pour prendre en compte l'ensemble des facteurs liés à cette thématique. L'étude qui, à notre connaissance, approfondit le plus cette problématique est celle menée par le PAM (WFP-VAM, 2019) (cf. section 2.4.1), qui extrait des caractéristiques de sources de données variées à l'aide d'un CNN et les introduit ensuite en entrée d'une régression ridge. Le Tableau 2.1 établit un bilan des niveaux d'hétérogénéité thématique, temporelle et spatiale des données et de la pertinence des méthodes prises en compte par les études présentées dans cette section, qui appliquent de l'apprentissage automatique sur des données hétérogènes dans le contexte de la sécurité alimentaire. L'hétérogénéité structurelle n'est plus évoquée ici, car à ce stade, les variables appliquées aux méthodes d'apprentissage automatique, qui avaient des structures de départ différentes (e.g., données géolocalisées, vecteurs lignes et polygones, pixels), ont été prétraitées et sont maintenant des vecteurs et des matrices dont la structure n'est différenciée que par leurs dimensions spatiales et temporelles. Nous constatons tout d'abord qu'aucune étude citée ne prend adéquatement en compte l'aspect temporel dans leurs données, par exemple par l'intégration et l'analyse de séries temporelles, ce qui est un critère déterminant pour cerner l'aspect conjoncturel propre à notre contexte. De plus, deux études seulement contiennent un large éventail thématique

permettant d'appréhender de manière complète les facettes de la sécurité alimentaire. Enfin, seule l'étude du WFP intègre des données hétérogènes au niveau spatial tout en exploitant des méthodes adaptées à ce type de données, ce qui est essentiel pour saisir l'aspect structurel de la sécurité alimentaire. Dans ce contexte, les contributions de notre étude se situent à la fois dans le choix des données exploitées et dans les méthodes appliquées.

Travaux	**Thématique**	**Temporel**	**Spatial**	**Méthode**
Sheehan (2018)	↓	↓	=	=
van der Heijden et al. (2018)	=	↓	↓	↓
Lentz et al. (2019)	=	↓	↓	↓
Heisenberg et al. (2020)	↑	↓	↑	=
WFP-VAM (2019)	↑	↓	↑	↑
Deléglise et al. (2021)	↑	=	↑	↑

Tableau 2.1 – Comparaison des niveaux d'hétérogénéités thématique, temporelle et spatiale des données et de la pertinence des méthodes prises en compte, parmi les études appliquant de l'apprentissage automatique sur des données hétérogènes dans le contexte de la sécurité alimentaire. Pour les colonnes *Thématique*, *Temporel* et *Spatial*, "↑", "=" et "↓" signifient respectivement une hétérogénéité associée "forte", "modérée" et "faible". Pour la Colonne *Méthode*, "↑", "=" et "↓" signifient respectivement que celle-ci est "très", "moyennement", et "peu" adaptée aux données.

Dans notre étude, nous utilisons des approches d'apprentissage automatique et profond pour faire face à la complexité de nos données, et nous combinons des données explicatives hétérogènes par des fusions aux trois niveaux possibles (données, caractéristiques et décisions) et comparons les performances des différents modèles. À notre connaissance, il s'agit de la première étude à proposer un examen aussi complet de différentes stratégies de fusion de données pour la prédiction des scores de sécurité alimentaire. De plus, c'est l'une des seules études dans le domaine étudié à prendre en compte un aussi grand nombre de sources de données hétérogènes (aux niveaux thématique, temporel et spatial). L'étude existante qui utilise l'ensemble le plus complet de sources d'informations est le framework développé par le PAM, qui n'inclut pas plusieurs sources de données importantes que nous prenons en compte dans ce chapitre, à savoir

les prix du maïs, les densités de population et la qualité des sols. Par ailleurs, leur étude ne prend pas en compte l'aspect séquentiel des séries temporelles, ce qui est le cas pour notre étude.

2.3 Matériel et méthodes

2.3.1 Variables réponses

Dans ce chapitre, nous nous concentrons sur deux indicateurs calculés à partir de réponses issues d'enquêtes ménages : le SCA et le SDA (dont les caractéristiques et la méthode de calcul sont détaillées dans la section 1.2.4). Notons que l'$ISAr$ (indice des stratégies d'adaptation réduit), qui a été décrit et analysé dans le Chapitre 1, n'est disponible que depuis 2014, ce qui produit trop peu d'observations pour appliquer nos méthodes d'apprentissage automatique sur cet indicateur. Les indicateurs SCA et SDA, largement utilisés dans la littérature scientifique et par les organisations gouvernementales et non gouvernementales (Jones et al., 2013 ; Maxwell et al., 2014 ; Vhurumuku, 2014), peuvent être utilisés pour évaluer la fréquence, la quantité et la qualité des aliments dans des zones ciblées et fournissent des informations spatiales et interannuelles cohérentes (comme démontré dans le Chapitre 1). Pour obtenir ces indicateurs, nous prenons en compte les données de l'EPA disponibles de 2009 à 2018 (dont les objectifs et la méthodologie sont détaillés dans la section 1.3.2). Le jeu de données obtenu contient les informations de 46400 ménages agricoles, répartis dans 344 des 351 communes. Notons que les 344 communes considérées ne sont pas incluses dans chacune des années enquêtées par l'EPA. Nous avons utilisé ces données afin de construire notre vérité terrain, en moyennant les indicateurs SCA et SDA par commune, et en considérant une fenêtre temporelle de 2009 à 2018, ce qui génère 3066 observations. Les autres études qui prédisent ces indicateurs (présentées dans la section 2.2) sont basées sur une fenêtre temporelle d'un an. À notre connaissance, notre étude est la seule à ce jour à se baser sur une fenêtre temporelle de 10 ans, ce qui permet d'établir des règles de décision basées sur les variations interannuelles et donc davantage généralisables dans le temps. De plus, ce travail se positionne sur l'analyse approfondie d'un terrain dédié (le Burkina Faso), contrairement aux analyses du PAM, plus étendues géographiquement mais moins exhaustives en termes de données exploitées. Comme dans la plupart des enquêtes, la qualité des données obtenues à partir des enquêtes ménages peut être affectée par des

biais. Ces biais peuvent apporter du bruit aux données et affecter les performances des algorithmes d'apprentissage automatique qui leur sont appliqués. C'est en partie la raison pour laquelle les études (présentées dans la section 2.4) qui ont prédit le SCA et le SDA par apprentissage automatique offrent de faibles performances. L'annexe A propose une classification des différents types de biais ainsi qu'une discussion des biais présents dans l'EPA. Pour la prédiction des variables réponses présentées ici (i.e., le SCA et le SDA), nous exploitons un ensemble de données explicatives ouvertes et hétérogènes dont la méthodologie d'extraction et de constitution est décrite dans la section suivante.

2.3.2 Variables explicatives

Pour répondre à **Q1** ("quels types de données ouvertes doivent être ciblés pour prédire des scores de sécurité alimentaire"), il existe un grand nombre d'indicateurs "proxy", liés à une ou plusieurs composantes de la sécurité alimentaire qui peuvent être pris en compte. Par exemple : les indices de végétation, les précipitations, les prix des aliments, les densités de population locales, la qualité des sols et le nombre d'événements violents, d'écoles et d'hôpitaux (Fritz et al., 2019), que nous prenons en compte lors de la sélection des données explicatives pour notre étude.

L'aspect multifactoriel de la sécurité alimentaire implique l'utilisation de données explicatives hétérogènes pour obtenir une vision aussi complète que possible de la situation alimentaire. Comme cela a été souligné dans la section 2.2.2, les proxies de la sécurité alimentaire utilisés comme variables explicatives peuvent être considérés comme hétérogènes à trois niveaux.

1. Au niveau *thématique*, ces proxies sont liés à des domaines variés tels que la météorologie, l'économie, la démographie, la sécurité ou encore la qualité et l'occupation des sols. Cela implique d'avoir une vision complète des facteurs de famine dans le lieu étudié.
2. Au niveau *structurel*, il existe différents types de données : valeurs quantitatives, données géolocalisées, vecteurs lignes, séries temporelles et rasters. Cela nécessite l'utilisation d'outils et de méthodes adaptés au traitement de chaque type de données.
3. Au niveau de l'*échelle spatio-temporelle*, les données peuvent être disponibles spatialement par région, commune, station ou pixel et temporellement par décennie,

année, mois ou semaine. Ce point nécessite l'utilisation de techniques permettant d'extraire les informations pertinentes à une échelle commune sur laquelle les données peuvent être combinées

Il s'agit donc de choisir l'échelle spatio-temporelle la plus appropriée, ce qui implique par ailleurs de répondre à **Q2** ("comment des données hétérogènes en termes de thématique, de structure et de résolution spatio-temporelle peuvent-elles être considérées afin d'obtenir des prédictions cohérentes sur la sécurité alimentaire pour un site d'étude donné"). Les choix judicieux de l'échelle spatio-temporelle, ainsi que des prétraitements et normalisations des données que cela implique, sont essentiels pour obtenir des variables explicatives pertinentes qui seront appliquées aux modèles d'apprentissage automatique. Cela nécessite une double compétence : thématique (assurant un recul nécessaire sur les données exploitées) et technique (permettant la mise en œuvre des choix effectués) qui a été acquise au cours de cette thèse par des échanges récurrents avec des spécialistes tant de la sécurité alimentaire que de la science des données.

Tout d'abord, les proxies de la sécurité alimentaire sont prétraités pour extraire des variables explicatives pertinentes. Chaque variable explicative doit être agrégée à une granularité spatio-temporelle de référence (parmi 4 granularités proposées et détaillées à la fin de cette section) qui met le mieux en valeur sa structure initiale (e.g., variable quantitative classique, série temporelle, variable à haute résolution spatiale) afin d'être intégrée, avec d'autres variables de même structure, à une méthode d'apprentissage automatique adaptée à ses spécificités. Ces méthodes effectueront leurs prédictions à l'échelle de la commune, qui est la plus petite limite administrative pour laquelle les variables réponses sont spatialisées, permettant ainsi de maximiser le nombre d'observations pour l'apprentissage des modèles. Certains proxies issus de données rasters ou de données géolocalisées ont une granularité plus fine et doivent être agrégés par commune par somme (pluviométrie), moyenne (températures minimales et maximales, qualité des sols), dénombrement (hôpitaux, écoles et événements violents), maximum (température de brillance lissée (SMT), altitude) ou par des agrégations plus complexes (coefficient de Gini, autocorrélation et entropie différentielle des rasters de population). D'autres proxies (données météorologiques accessibles par stations et prix des produits de première consommation accessibles par marchés) sont disponibles à une granularité plus grossière et doivent être interpolés pour chaque commune : dans ces cas, nous avons choisi d'utiliser une interpolation par une approche "K plus proches voisins", qui est une technique à la fois précise et rapide d'exécution, fréquemment appliquée pour effec-

tuer des tâches de ce type dans les domaines liés au climat (Sinta et al., 2014 ; Kiani and Saleem, 2017 ; Yu and Haskins, 2021). Le raster d'occupation du sol utilisé, dont la résolution initiale est de 20 mètres, est rééchantillonné à une résolution de 100 mètres (en calculant pour chaque type de sol considéré la proportion de pixels de 20 mètres contenus dans le pixel agrégé de 100 mètres), afin que sa résolution soit commune avec celle des rasters de population. Des patchs de ces rasters de 10x10 pixels de 100 mètres (soit 1 km carré) seront les entrées d'un CNN pour prédire le SCA et le SDA, restituant des prédictions et des caractéristiques par commune. Les méthodes de mise à l'échelle appliquées sur chacune des variables sont détaillées dans la colonne "mise à l'échelle" du Tableau 2.2. L'indice de végétation par différence normalisée (NDVI), qui est un indice de végétation, est traité par un masque de culture afin de ne considérer que le NDVI dans les zones de culture [1]. Certaines variables sont normalisées par la population (e.g., le nombre d'écoles, d'hôpitaux et d'événements violents) ou par la superficie (nombre et longueur des cours d'eau). Puis, chaque variable explicative est centrée réduite par rapport aux communes et aux années (consiste à soustraire la moyenne et à diviser par l'écart-type). Enfin, les variables explicatives obtenues sont sélectionnées en ne retenant que celles qui sont significativement corrélées avec la variable réponse considérée (valeur p inférieure à 0,05). Les informations relatives à chaque jeu de données sont disponibles dans le Tableau 2.2 ; pour plus de détails sur la variété des données utilisées dans les modèles, voir l'annexe B. Pour tenir compte de l'hétérogénéité spatio-temporelle des données, nous proposons une catégorisation des variables explicatives sélectionnées en 4 groupes ayant une structure similaire afin de traiter chaque groupe avec une méthode d'apprentissage automatique appropriée :

— les **séries temporelles** qui ont plusieurs valeurs par an et une valeur par commune. Celles-ci sont agrégées en séries temporelles mensuelles (mai à novembre de l'année de collecte de l'indicateur FS et de l'année précédente) ;
— les **données conjoncturelles** qui ont une valeur par an et une valeur par commune ;
— les **données spatiales** qui ont une valeur par commune et sont invariantes par année ;
— les **données à haute résolution spatiale (HRS)** qui ont plusieurs valeurs par commune. Les valeurs sont des patchs de 10x10 pixels de 100 m extraits de chaque source de données.

1. Masque de culture : Carte prototype S2 d'occupation du sol à 20 m en Afrique 2016

L'objectif de cette catégorisation en 4 groupes est de rendre les différentes catégories de variables explicatives adaptées à un traitement indépendant par différentes branches du framework, chacune basée sur des techniques d'apprentissage automatique spécifiques (i.e., chaque branche est fondée sur la technique la plus adaptée au type de données spécifique, comme détaillé dans la section suivante).

Variable	Résolution	Fréquence	Source	Mise à l'échelle
Séries temporelles [plusieurs valeurs par an ; une valeur par commune] [70 variables]				
Température de brillance lissée (SMT) [14 variables]	4 km	7 jours	National Oceanic and Atmospheric Administration (NOAA)	Maximum
Précipitations [14 variables]	6 km	10 jours	Tropical Rainfall Measuring Mission (TRMM)	Somme
Températures minimales et maximales moyennes [2x14 vars]	21 km	1 mois	WorldClim	Moyenne
Prix du maïs [14 variables]	64 marchés	1 mois	Société Nationale de Gestion du Stock de Sécurité alimentaire (SONAGESS)	Interpolation par k plus proches voisins
Données conjoncturelles [une valeur par année ; une valeur par commune] [20 variables]				
Données météorologiques [7 variables]	10 stations	1 an	Plateforme Knoema	Interpolation par k plus proches voisins
Densité de population [4 variables]	100 m	1 an	Afripop	Autocorrélation spatiale à 2 km et 5 km, Gini, entropie
Données économiques [7 variables]	Pays	1 an	Banque mondiale	Valeur du pays
Indice de végétation par différence normalisée [2 variables]	250 m	1 an	Modis	Moyenne
Données spatiales [une valeur par commune] [13 variables]				
Hôpitaux, écoles [2 variables]	Vecteurs points	2018	Open Street Map	Comptage
Événements violents [4 vars]	Vecteurs points	2018	Armed Conflict Location & Event Data Project (ACLED)	Comptage
Qualité des sols [3 variables]	1 km	2008	Food and Agriculture Organization (FAO)	Moyenne
Cours d'eau [2 variables]	Vecteurs lignes	2008	Digital Chart of the World (DCW)	Comptage, longueur
Données d'altitude [2 variables]	1 km	2018	Consultative Group on International Agricultural Research (CGIAR)	Maximum, variance
Données à haute résolution spatiale [plusieurs valeurs par commune] [4 variables]				
Densité de population	100 m	1 an	Afripop	-
Occupation du sol (cultures, forêts, zones construites)	20 m	2016	Agence spatiale européenne	rééchantillonné à 100 m

Tableau 2.2 – Résumé des jeux de données

2.3.3 Framework FSPHD

Dans cette section, nous présentons la méthodologie adoptée afin d'exploiter les variables réponses et explicatives définies précédemment. Pour répondre à **Q3** ("comment les approches d'apprentissage automatique et profond peuvent-elles être exploitées et combinées afin de traiter des données hétérogènes ?"), nous définissons le framework d'apprentissage automatique proposé, appelé "Food Security Prediction based on Heterogeneous Data" ($FSPHD$), conçu pour estimer par régression le SCA et le SDA. L'objectif du framework proposé est d'intégrer plusieurs techniques d'apprentissage automatique et profond, capables d'exploiter tout le potentiel du grand nombre de données hétérogènes utilisées en entrée.

Rappelons que les variables explicatives hétérogènes sélectionnées sont réparties en 4 groupes en fonction de leur structure spécifique, afin que des méthodes d'apprentissage automatique adaptées à chaque structure puissent en extraire des informations de manière optimisée. Étant donné que le framework conçu doit permettre d'obtenir une prédiction unique en combinant chaque méthode appliquée, il est nécessaire d'y intégrer le concept de fusion de données discutée dans la section 2.2. À cette fin, nous proposons deux types de modèles de régression (cf. Figure 2.3) pour prédire le SCA et le SDA, qui correspondent à deux variantes différentes du framework proposé, à savoir le modèle *(a)* et le modèle *(b)*. Les modèles *(a)* et *(b)* sont respectivement inspirés d'études d'apprentissage automatique utilisant de la fusion au niveau des décisions (Peterson et al., 2018 ; Guo et al., 2019) et au niveau des caractéristiques (Xue et al., 2017 ; Amin et al., 2018) pour le traitement et la combinaison de variables explicatives hétérogènes. Les approches de fusion au niveau des caractéristiques et des décisions se sont avérées efficaces dans plusieurs domaines d'application (e.g., médecine (Amin et al., 2018 ; Guo et al., 2019), reconnaissance de l'activité humaine (Xue et al., 2017), chimie (Peterson et al., 2018)). Par conséquent, étant donnée la nature exploratoire de ce travail, nous proposons deux variantes du framework intégrant chacune un niveau de fusion distinct.

Un autre point clé de la conception du framework repose sur la construction de chaque branche intégrant une méthode appropriée à chaque type spécifique de données, i.e., en fonction de ce qui a été observé dans la littérature existante. Le défi est d'extraire, avec chaque branche, des informations complémentaires sur la sécurité alimentaire à partir de chaque source de données. Les données conjoncturelles et spatiales (CS) sont des données numériques classiques (i.e., une seule valeur numérique pour chaque observa-

tion de la variable réponse) et sont traitées par une FA (forêt aléatoire), qui est l'une des méthodes d'apprentissage automatique offrant le meilleur compromis entre performances et interprétabilité pour des données de structure non complexes (Qi, 2012 ; Liu et al., 2013 ; Lan et al., 2020). La FA fonctionne en générant une multitude d'arbres de décision (qui est une méthode intuitive d'apprentissage automatique (Wilkinson, 1992)), chacun entraîné sur un sous-jeu de données et de variables explicatives (procédé appelé "bootstrap"). Les prédictions de chaque arbres sont agrégées par moyenne. Les séries temporelles sont traitées à l'aide d'une architecture LSTM (réseau de neurones récurrents à mémoire court-terme et long terme) (illustrée Figure 2.1), qui est une méthode d'apprentissage profond éprouvée pour les problèmes de prédictions à partir de données séquentielles en raison de ses connexions de rétroaction (Song et al., 2020). Le LSTM possède plusieurs mécanismes internes (une "cellule mémoire", une "porte d'entrée", une "porte d'oubli" et une "porte de sortie") pour prendre en compte les dépendances (à court et à long terme) dans des données séquentielles utilisées comme variables explicatives. Plus précisément : chaque neurone est relié à une cellule mémoire qui contient les poids des neurones associés aux états précédents de la séquence ; la porte d'entrée détermine si le poids de son neurone associé est ajouté à la cellule mémoire du neurone de l'état suivant ; la porte d'oubli module l'influence des neurones des états précédents en fonction de leur utilité pour le calcul de l'état actuel d'un neurone ; la porte de sortie module l'influence du contenu de sa cellule mémoire sur la sortie d'un neurone.

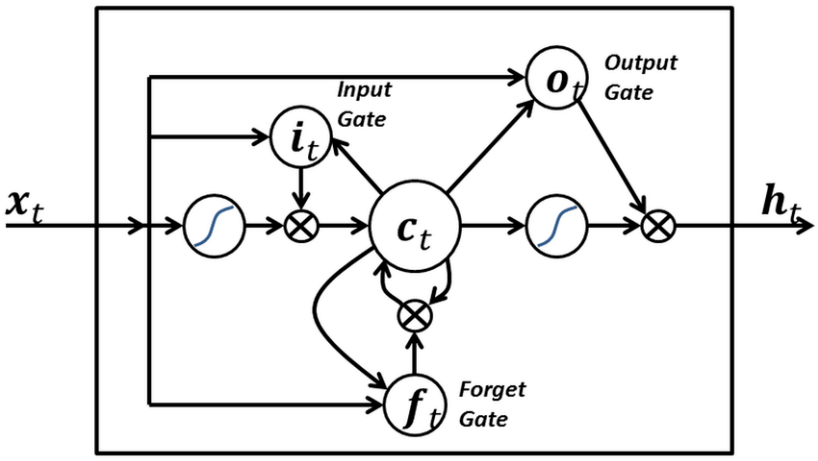


FIGURE 2.1 – Illustration d'une architecture LSTM. Source : Zaytar and El Amrani (2016).

Les données HRS sont introduites dans un CNN (réseau de neurones convolutifs), qui est une méthode d'apprentissage profond adaptée à l'analyse d'images (Huang et al., 2018). L'architecture d'un CNN (illustrée Figure 2.2) est formée par une succession de blocs de traitements appliqués à une image pour extraire des caractéristiques visuelles corrélées à la variable réponse. Le traitement consiste en un empilement de plusieurs couches de neurones (appelées "convolutions"), appliquées successivement à des sous-régions d'une image, qui se chevauchent en la pavant. La première couche de convolution extrait des caractéristiques simples (e.g., lignes, croix, carrés). Puis, chaque nouvelle couche de convolution appliquée à une image intermédiaire issue de la couche précédente permet d'extraire des motifs visuels de plus en plus complexes. D'autres types de couches sont généralement insérées dans un CNN. Par exemple, la couche de correction "ReLU" (Rectified Linear Unit) (Hara et al., 2015) agit comme une fonction d'activation, empêchant une valeur négative à la sortie d'un neurone d'influencer négativement le poids du neurone suivant. Évoquons également la couche de "Pooling", qui compresse l'information en réduisant la dimension d'une image intermédiaire (généralement par sous-échantillonnage, en pavant l'image de groupe de pixels de petite dimension, puis en sélectionnant le maximum de chaque groupe, auquel cas nous parlons de "Max pooling").

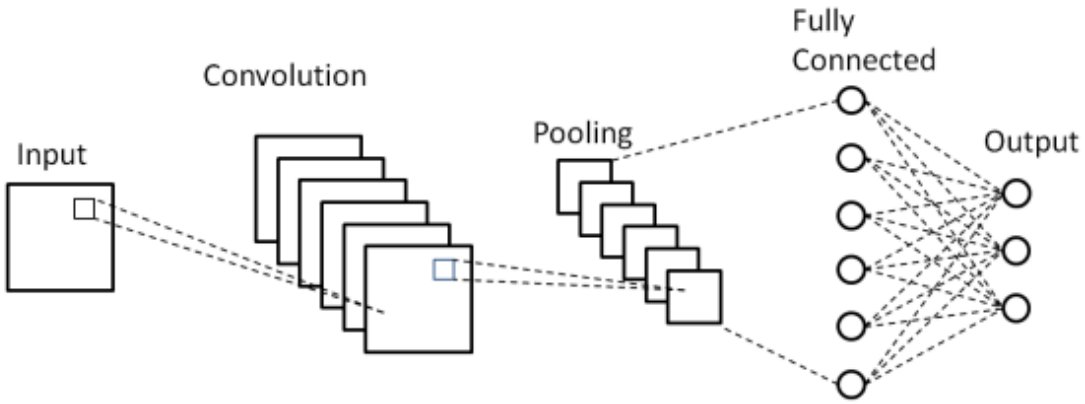


FIGURE 2.2 – Illustration d'une architecture CNN. Source : Phung et al. (2019).

Le framework $FSPHD$ proposé (Figure 2.3) intègre ces méthodes d'apprentissage automatique et profond adaptées à chaque groupe spécifique de variables et fusionne les informations obtenues selon deux stratégies : une stratégie guidée par la fusion au niveau des décisions (modèle *(a)*) et une seconde stratégie guidée par la fusion au niveau des caractéristiques (modèle *(b)*).

Plus précisément, les modèles *(a)* et *(b)* de $FSPHD$ sont structurés comme suit :

— **Modèle (a)** : Nous appliquons un modèle linéaire (LM) avec régularisation ridge sur les réponses des trois modèles d'apprentissage automatique et profond : la réponse du LSTM sur les séries temporelles, la réponse du CNN sur les données HRS et la réponse de la FA sur les données CS. Ce modèle est basé sur de la fusion au niveau des décisions, en agrégeant les prédictions obtenues pour obtenir une prédiction globale plus robuste.

— **Modèle (b)** : Nous appliquons une FA sur les caractéristiques extraites par les modèles d'apprentissage profond. Ce modèle est basé sur de la fusion au niveau des caractéristiques, qui permet aux données complexes (i.e., séries temporelles, données HRS) d'obtenir de nouvelles représentations mieux corrélées à la variable réponse et plus efficacement traitables par une FA.

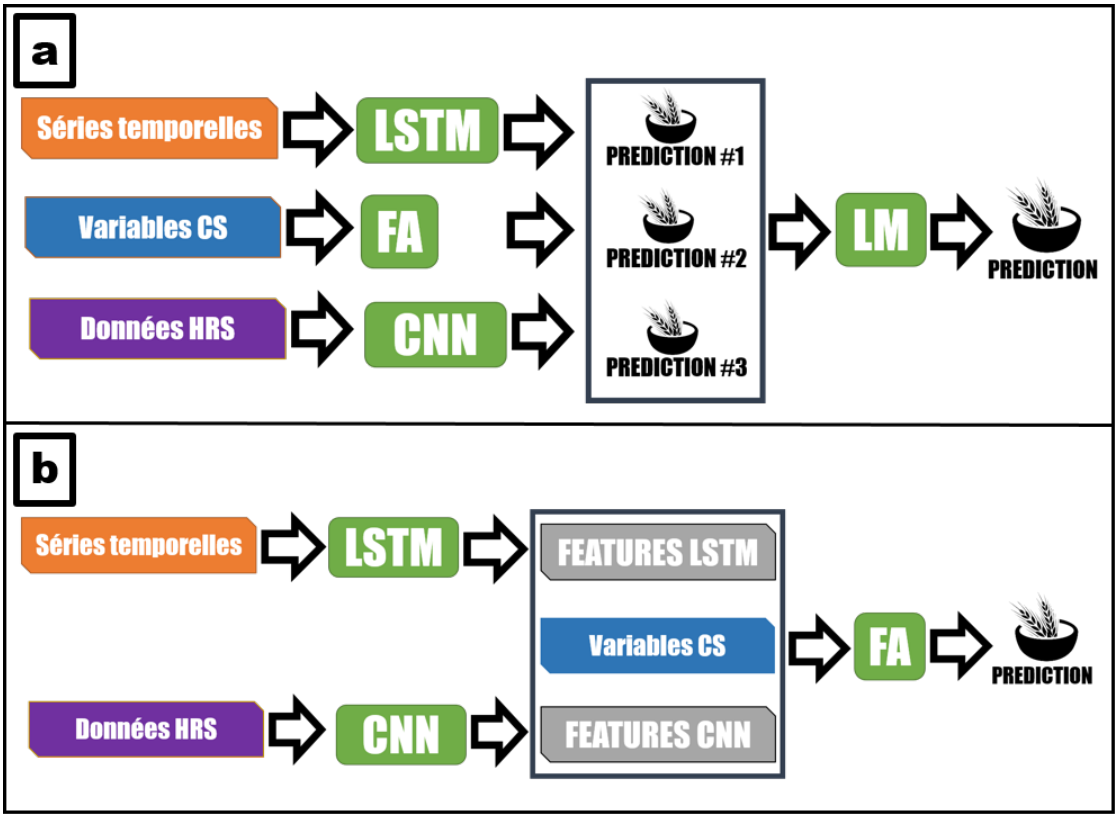


FIGURE 2.3 – Illustration du framework $FSPHD$, qui combine des données explicatives hétérogènes pour prédire le score de consommation alimentaire (SCA) et le score de diversité alimentaire des ménages (SDA). Les modèles (a) et (b) sont respectivement guidés par de la fusion aux niveaux des décisions et des caractéristiques.

2.4 Évaluation expérimentale

2.4.1 Méthodes concurrentes et versions simplifiées

Pour évaluer les performances du framework $FSPHD$, nous le comparons aux performances de plusieurs modèles de référence, versions simplifiées et méthodes concurrentes.

Nous utilisons les deux études indépendantes suivantes comme méthodes concurrentes :

— Comme **première méthode concurrente**, nous choisissons le framework du PAM introduit dans la section 1.3.3 (WFP-VAM, 2019). Ce chapitre porte sur la régression du SCA et du SDA obtenus à partir d'une enquête de février 2018 auprès de 3 650 ménages burkinabés. Les ménages sont regroupés dans 567 villages géolocalisés. L'étude utilise des données provenant de différentes sources (Open Street Map, Google Maps, Sentinel 2, ACLED) comme variables explicatives. Les caractéristiques à l'échelle du village sont extraites de chaque source de données par une méthode appropriée : les images sont traitées par un CNN et les distances les plus courtes d'un village à une école et à un hôpital ainsi que le nombre d'événements violents à 10 km sont extraits de données géolocalisées. Les caractéristiques obtenues sont traitées par une analyse par composante principale (ACP), qui est une méthode permettant de compresser les informations contenues dans un ensemble de variables en produisant un plus petit nombre de variables décorrélées entre elles et synthétisant l'information initiale, appelées "composantes principales". Les 10 premières composantes principales (i.e., les plus importantes car elles expliquent les plus grandes proportions de la variance totale) sont finalement mises en entrées d'une régression ridge pour la prédiction du SCA et du SDA. Nous avons exécuté leur framework (en utilisant leur code public [2]) avec leurs données [3] pour obtenir les résultats.

— La **deuxième méthode concurrente** est issue d'une étude menée par Lentz et al. (2019) (cf. section 2.2.1). Ces travaux prédisent le SCA et le SDA au Malawi en utilisant des données obtenues à partir de l'enquête sur la mesure des niveaux de vie (LSMS) de 2010 pour l'entraînement et en utilisant des données obtenues à partir de l'enquête LSMS de 2013 pour le test. Les données de 2010 (entraînement) et 2013 (test) contiennent respectivement 12 270 et 3 999 observations, qui sont agrégées dans 768 et 204 villages. Ils utilisent des régressions linéaires et comparent les performances de leurs modèles en utilisant uniquement des données ouvertes et en ajoutant les données de l'enquête LSMS précédente. Dans notre étude, nous prenons en compte les performances des modèles intégrant uniquement des données ouvertes conformément à nos objectifs. Les données utilisées proviennent de sources diverses : météorologie, précipitations, prix

2. https ://github.com/WFP-VAM/HRM

3. Ces données ne sont pas publiques, elles nous ont été transmises dans le cadre d'un partenariat

du marché et qualité des sols.

Nous définissons également trois types de modèles de référence (appelés (c), (d1) et (d2)) afin de tester individuellement chacune des techniques d'apprentissage automatique et profond appliquées aux différents sous-ensembles de variables explicatives.

— **Modèle de type (c)** : nous appliquons une FA directement sur les variables initiales : sur les séries temporelles uniquement, sur les variables CS uniquement et sur ces deux types de données conjointement. Ce type de modèle est basé sur de la fusion au niveau des données (Figure 2.4).

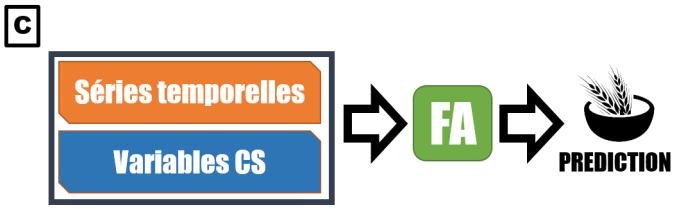


FIGURE 2.4 – Illustration du modèle de type (c)

— **Modèle de type (d1)** : nous appliquons un LSTM adapté au traitement des séries temporelles sur celles-ci (Figure 2.5).

FIGURE 2.5 – Illustration du modèle de type (d1)

— **Modèle de type (d2)** : nous appliquons un CNN adapté aux données HRS sur celles-ci (Figure 2.6).

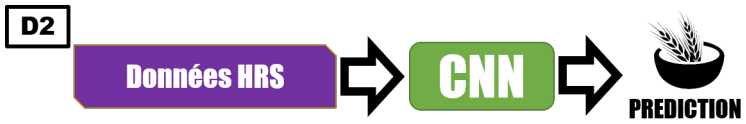


FIGURE 2.6 – Illustration du modèle de type (d2)

Enfin, pour étudier plus en détail les contributions de chaque groupe de caractéristiques sur la réponse finale, nous définissons trois versions simplifiées (ablations) du modèle de type (b) en déclinant comme suit :

— application seulement sur les caractéristiques du LSTM ;
— application seulement sur les caractéristiques du CNN ;
— application seulement sur les caractéristiques du CNN et sur les variables CS.

2.4.2 Cadre expérimental

Le framework $FSPHD$ a été implémenté en utilisant TensorFlow 1.15 sous Python 3.7, le code est en accès libre sur GitHub [4]. Les paramètres choisis et décrits ci-dessous ont été classiquement sélectionnés par 10 validation croisée pour la FA et par 5 validation croisée pour les réseaux de neurones. La FA est configurée avec 900 arbres d'une profondeur maximale de 20. Le LSTM est paramétré avec 2 couches de 64 neurones. La fonction de coût utilisée est l'erreur quadratique moyenne, et l'optimiseur est basé sur l'algorithme FTRL (Hazan and Kale, 2010). Le CNN est configuré avec 3 couches de convolutions possédant 32, 64 et 128 filtres. Un max pooling de dimension 2 est placé après chaque couche de convolution. La fonction de coût utilisée est l'erreur quadratique moyenne, et l'optimiseur est basé sur l'algorithme Adam (Kingma and Ba, 2014). Pour le modèle de type (b) utilisant la fusion de caractéristiques, 64 caractéristiques sont extraites du LSTM, et 128 caractéristiques sont extraites du CNN. LSTM et CNN sont entraînés "from scratch" en utilisant une taille de batch de 250, et en fixant respectivement 1 000 et 100 époques. Cette différence dans le nombre d'époques est due au fait que le CNN commence à sur-apprendre les données plus rapidement que le LSTM. Pour évaluer les performances, nous sélectionnons aléatoirement 85% de l'ensemble de données pour l'apprentissage des modèles et 15% pour le test, en répétant cette procédure 5 fois et en calculant les performances moyennes. Cette méthode de validation, appelée "ré-échantillonnage bootstrap" (Jain et al., 1987), est une alternative à la validation croisée qui est adaptée dans le contexte de jeux de données peu volumineux et de forte variance (Raudys, 1988 ; Beleites et al., 2005), ce qui est le cas dans cette étude. Nous utilisons le R^2 (Nagelkerke et al., 1991) (coefficient de détermination linéaire de Pearson), qui est une mesure classique pour évaluer les performances de régressions. Ce coefficient est compris entre 0 (prédictions hasardeuses) et 1 (prédictions qui correspondent parfaitement

4. https ://github.com/pipapou/FSPHD_Code

aux données prédites).

2.4.3 Résultats

Le Tableau 2.3 présente les performances quantitatives (R^2) des deux variantes (a) et (b) du framework $FSPHD$ et de toutes les méthodes de référence et concurrentes. Nous constatons que $FSPHD$ surclasse toutes les méthodes de référence et concurrentes, avec le modèle *(b)* (i.e., l'approche utilisant la fusion au niveau des caractéristiques) qui surclasse le modèle *(a)* (i.e., l'approche utilisant la fusion au niveau des décisions). Même si les prédictions obtenues avec le framework $FSPHD$ ne sont pas encore assez précises pour être utilisées dans des contextes opérationnels, les valeurs de R^2 obtenues pour le SCA $(0,469)$ et pour le SDA $(0,434)$ sont statistiquement significatives, surclassant toutes les méthodes concurrentes et prouvant ainsi les avantages de l'intégration de différentes techniques de science des données sur un grand nombre de données hétérogènes.

Méthodes concurrentes

Les résultats obtenus par le framework du PAM au Burkina Faso (cf. Tableau 2.3) sont relativement modestes ($0,34$ pour le SCA et $0,30$ pour le SDA). Comme déjà indiqué dans la section 2.2.1, leurs résultats semblent être extrêmement dépendants des données, de sorte que ce même framework obtient des résultats comparables dans d'autres pays (e.g., le Sénégal, la Sierra Leone) (WFP-VAM, 2019). Une explication de ces modestes performances peut être liée au faible nombre d'observations de leurs variables réponses (1 207 pour le Sénégal et 1 292 pour la Sierra Leone, sachant qu'il s'agit des plus larges études qu'ils intègrent), ainsi qu'aux biais inhérents à ce type de données sur les ménages (discutés en détail en annexe A). En ce qui concerne l'étude de Lentz et al, nous notons que le R^2 associé aux SCA et SDA est encore plus faible, ne dépassant pas $0,2$. Ces faibles performances peuvent être fortement attribuées aux méthodes appliquées, qui se limitent à des régressions linéaires.

Modèles de référence

Nous observons que les modèles de type (c) (uniquement basés sur l'utilisation de FA directement sur les variables explicatives) donnent des performances déjà significatives et proches de celles de modèles plus sophistiqués. En intégrant uniquement les variables CS, nous obtenons des R^2 de $0,414$ et $0,401$ respectivement pour le SCA et le SDA.

Cela valide la sélection des sources de données et le prétraitement appliqué aux données intégrées. Les modèles de type (d1) et (d2) visent à traiter des données aux structures complexes (séries temporelles et images HRS, respectivement) avec une méthode d'apprentissage profond adaptée. Le LSTM ne parvient pas à mettre en évidence l'aspect séquentiel des séries temporelles, offrant des performances légèrement inférieures à celles de la FA sur les mêmes données (i.e., modèle (c) sur les séries temporelles seulement) : $0,232$ contre $0,241$ pour le SCA et $0,223$ contre $0,237$ pour le SDA. Notre hypothèse est que dans le cas présent, où nous avons des variables réponses bruitées, le LSTM surinterprète le bruit malgré nos paramétrages et par conséquent se sur-ajuste sur les données. Notons que dans une étude antérieure (Deléglise et al., 2021a), réalisée en effectuant des classifications (en 4 classes) avec les mêmes variables réponses et explicatives, le LSTM a offert des performances légèrement meilleures que la FA (pour le SCA, accuracy de 0.403 contre 0.387 ; pour le SDA, accuracy de 0.402 contre 0.357). Ce résultat, bien que modéré, doit nous encourager à poursuivre dans cette direction. Des travaux futurs devraient donc se concentrer davantage sur les causes des mauvaises performances du LSTM, et se focaliser sur d'autres méthodes pour améliorer la prise en compte des séries temporelles. Le CNN sur les données HRS donne des performances intéressantes (0.34 pour le SCA et 0.392 pour le SDA). Cela confirme que cette capacité connue des CNN à traiter des variables à haute résolution spatiale est applicable dans notre contexte.

Framework $FSPHD$

Les variantes (a) et (b) de $FSPHD$ représentent deux stratégies pour combiner divers types de données en se focalisant sur différentes démarches de fusion. Le modèle de type (a), qui consiste à agréger les réponses des différents modèles avec un modèle linéaire, présente des performances modérées pour le SCA $(0,375)$ et plus élevées pour le SDA $(0,426)$, ce qui témoigne de la contribution du modèle de type (a) utilisant de la fusion au niveau des décisions pour la prédiction du SDA. Le modèle de type (b) consiste à agréger les caractéristiques obtenues par les différents modèles avec une FA. Tout d'abord, les performances relatives au CNN sont significativement améliorées lorsque celui-ci est intégré seul dans le modèle de type (b), i.e., en mettant les caractéristiques extraites du CNN en entrée d'une FA (0.434 pour le SCA et 0.418 pour le SDA). De plus, le modèle de type (b) donne les meilleures performances globales, en combinant les caractéristiques obtenues à partir du CNN avec les variables CS $(0,469$ pour le SCA et $0,434$ pour le SDA). Les performances sont alors significativement meilleures que celles obtenues en traitant séparément les caractéristiques du CNN et les variables CS

avec une FA, ce qui démontre que ces deux types de données fournissent des informations complémentaires sur la sécurité alimentaire, ainsi que la capacité du modèle de type (b) utilisant de la fusion au niveau des caractéristiques à extraire ces informations. Ces bonnes performances du modèle de type (b) sont en phase avec la littérature scientifique présentée dans la section 2.2.2. Dans les sections suivantes de ce chapitre, nous considérons le meilleur modèle pour nos analyses, i.e., le modèle de type (b) combinant les caractéristiques du CNN avec les variables CS.

Modèle	**Type**	**SCA**	**SDA**
Méthodes concurrentes			
Étude du PAM	(b)	0.34	0.30
Étude de Lentz et al.	(c)	0.16	0.18
Fusion sur les données			
FA(séries temporelles)	(c)	0.241	0.237
FA(variables CS)	(c)	0.414	0.401
FA(séries temporelles + variables CS)	(c)	0.339	0.326
Méthodes d'apprentissage profond adaptées			
LSTM(séries temporelles)	(d1)	0.232	0.223
CNN(données HRS)	(d2)	0.34	0.392
Framework $FSPHD$			
Fusion sur les décisions			
LM(réponses des FA, LSTM et CNN)	(a)	0.375	0.426
Fusion sur les caractéristiques			
FA(caractéristiques LSTM)	(b)	0.194	0.181
FA(caractéristiques CNN)	(b)	0.434	0.418
FA(caractéristiques CNN + variables CS)	(b)	**0.469**	**0.434**
FA(caractéristiques CNN,LSTM + variables CS)	(b)	0.455	0.43

Tableau 2.3 – Performances (R^2) du framework $FSPHD$, des méthodes concurrentes et des versions simplifiées pour la prédiction du score de consommation alimentaire (SCA) et du score de diversité alimentaire des ménages (SDA). La colonne "Type" désigne le type de modèle utilisé, selon la catégorisation des modèles proposée dans la section 2.4.1.

2.4.4 Interprétation des modèles

Dans cette section, nous nous concentrons sur l'interprétation des modèles appliqués à chaque groupe de variable explicatives en évaluant, lorsque cela est possible, les variables explicatives qui ont le plus contribué aux prédictions de leur modèle correspondant. En raison des faibles performances associées au LSTM, il n'est pas possible d'en déduire des informations pertinentes et fiables. La description des motifs spatiaux complexes obtenus par CNN est également compliquée en raison de l'effet "boîte noire" inhérent au CNN. Nous pouvons néanmoins constater que les concepts d'occupation du sol et de dynamique des populations pris en compte par le CNN semblent jouer un rôle important pour appréhender la sécurité alimentaire, sachant les bonnes performances du CNN. L'interprétabilité des réseaux de neurones, qui est une problématique d'actualité, devrait faire l'objet de travaux futurs. Pour les variables CS qui sont directement traitées par FA, l'importance des variables peut être évaluée grâce au concept d'importance de permutation (Gregorutti et al., 2017). L'importance de permutation d'une variable est définie comme la diminution du score d'un modèle lorsque les valeurs de la variable sont mélangées de manière aléatoire dans le jeu test. Les tops 10 des variables CS en fonction de leur importance de permutation pour le SCA et le SDA sont présentés dans le Tableau 2.4. Nous remarquons que des variables provenant de multiples domaines sont incluses dans ces deux top 10 : structure du paysage (3 variables), dynamique des populations (2 variables), qualité des sols (2 variables), météorologie (2 variables), végétation (1 variable), insécurité (1 variable), sanitaire (1 variable) ou économie (1 variable), ce qui confirme l'importance de l'utilisation combinée de sources de données provenant de plusieurs domaines. Sept variables sont incluses dans les top 10 du SCA et du SDA conjointement, elles semblent essentielles pour la prédiction de la sécurité alimentaire dans notre cas. Parmi ces variables, nous trouvons :

— le NDVI moyen de l'année précédant l'enquête. Le NDVI, qui est un indicateur de la qualité de la végétation, n'est considéré que pour les zones de culture, celui-ci est donc lié à la qualité des cultures agricoles. Un fait intéressant est que le NDVI moyen de l'année précédente est plus important que celui de la même année, qui se situe aux portes des top 10 ;

— trois variables liées à la structure du paysage : la longueur totale des cours d'eau, qui nous permet d'évaluer la disponibilité en eau, ainsi que le maximum et la variance de l'altitude. La structure du relief a un impact évident sur l'agriculture

(e.g., accessibilité des zones cultivées aux aménagements agricoles et à l'eau, types de plantations spécifiques à certaines altitudes) ;

— deux variables qui expriment la dynamique des populations : les autocorrélations spatiales à 2 km et l'entropie différentielle associée aux populations ; la densité et les mouvements de population pouvant créer des pressions sur la sécurité alimentaire. Ceci confirme, avec les bonnes performances du CNN qui prend en entrée les données d'occupation du sol et de population, l'importance de ces sources de données ;

— une variable de qualité des sols : la capacité de rétention des nutriments, qui est directement liée à la disponibilité des productions agricoles.

Nous constatons également que certaines variables sont spécifiques d'un indicateur de sécurité alimentaire. Le top 5 du SCA contient 2 variables qui sont absentes du top 10 du SDA : les dépenses nationales brutes du pays et le nombre d'événements violents, qui semblent être plus spécifiques de la quantité de nutriments consommés. À l'inverse, la température maximale moyenne par jour est présente dans le top 5 du SDA et absente du top 10 du SCA, cette variable semble donc être spécifiquement liée à la qualité et à la diversité du régime alimentaire. Notre hypothèse est que la diversification des cultures, et par conséquent du régime alimentaire, dépend fortement du climat et des températures.

Rang	SCA	SDA
1	Entropie de la population	Qualité des sols (capacité de rétention des nutriments)
2	Dépenses nationales brutes	NDVI moyen de l'année précédente
3	NDVI moyen de l'année précédente	Altitude maximale
4	Altitude maximale	Température maximale moyenne par jour
5	Nombre total d'événements violents	Entropie de la population
6	Variance de l'altitude	Humidité relative maximale
7	Autocorrélation spatiale de la population à 2 km	Longueur totale des cours d'eau
8	Longueur totale des cours d'eau	Variance de l'altitude
9	Qualité des sols (capacité de rétention des nutriments)	Nombre d'hôpitaux
10	Qualité des sols (conditions d'enracinement)	Autocorrélation spatiale de la population à 2 km

Tableau 2.4 – top 10 des variables CS traitées par FA selon l'importance de permutation pour le SCA et le SDA.

2.4.5 Perspective opérationnelle

Dans cette section, nous évaluons les capacités de notre meilleur modèle (i.e., le modèle de type (b) combinant les caractéristiques obtenues à partir du CNN avec les variables CS) à fonctionner dans un contexte opérationnel, i.e., qu'un utilisateur puisse réaliser des prédictions pour une année à partir des données des années précédentes avec des outils accessibles et exécutables en un temps raisonnable.

2.4.5.1 Définition de la dimension opérationnelle

À cette fin, nous adoptons une approche différente de celle utilisée précédemment pour tester les performances prédictives, où les données de 2009–2018 étaient divisées de manière aléatoire en ensembles d'entraînement et de test. Plus précisément, nous nous plaçons dans une situation où un utilisateur se sert de notre meilleur modèle, qu'il entraîne sur les données de l'EPA (variables réponses) et avec les proxies (variables explicatives) des années précédentes, pour l'appliquer lors de l'année en cours en y mettant en entrée les proxies des mois passés (e.g., images satellites, données économiques) afin de prédire les valeurs de SCA et de SDA par commune quelques semaines avant que les résultats de l'EPA ne soient disponibles. Ceci permettrait la mise en oeuvre plus rapide des plans d'action éventuels.

L'aspect opérationnel de l'outil développé dans ce chapitre doit alors comporter deux composantes : l'une relative à l'applicabilité technique de l'outil, l'autre portant sur la cohérence et l'utilité de ses résultats.

La première composante relative à l'applicabilité technique de l'outil, est elle-même soumise à trois critères : la facilité d'obtention des proxies de l'année en cours, le temps nécessaire à l'exécution de l'outil pour effectuer des prédictions et la simplicité d'utilisation de l'outil. Dans notre contexte, seuls les deux premiers points sont vérifiés. En effet, tous les proxies intégrés par notre outil proviennent de données ouvertes, dont les sources sont accessibles sur internet et actualisées chaque année. De plus, les opérations nécessaires au prétraitement de chaque source de données puis à l'exécution du modèle prédictif sont automatisées par un programme informatique, dont le temps d'exécution ne dépasse pas une heure si le modèle n'est pas entraîné, et quelques minutes s'il l'est déjà. Nous pouvons donc affirmer que notre outil répond aux critères de disponibilité des données d'entrée et de temps d'exécution des algorithmes pour une application dans un cadre opérationnel. Cependant, notre outil est actuellement exécutable dans sa forme brute, c'est-à-dire sous la forme d'un langage de programmation, et ses sorties sont exclusivement des données numériques. Pour être concrètement exploitable par des utilisateurs non-initiés à la programmation, cet outil nécessiterait une interface utilisateur, qui permette d'accéder de manière simple et intuitive aux informations recherchées. L'outil pourrait prendre la forme d'un site web qui, en fonction des requêtes, extrairait automatiquement les données explicatives nécessaires sur les sites web dédiés, puis présenterait les résultats des modèles sous la forme de sorties graphiques pertinentes. Par exemple,

la plateforme FEWS NET développée par l'USAID fournit un accès en temps réel à des informations sur la situation alimentaire de nombreux pays du Sud [5]. La création d'une telle plateforme nécessite des compétences en matière de web design, de programmation web et d'open data ainsi qu'un investissement conséquent. Cet aspect est abordé ici mais sa mise en œuvre exigera un travail spécifique que nous ne traitons pas dans cette thèse.

La deuxième composante concerne la cohérence et l'utilité des résultats de l'outil pour un expert du domaine, qui doivent être significatifs pour justifier l'utilisation d'un tel outil. C'est cette capacité de notre outil à obtenir des résultats qui fassent sens que nous évaluons dans la suite de cette section. Pour cela, nous appliquons ici l'architecture du meilleur modèle pour la prédiction des scores de sécurité alimentaire d'une année donnée à partir des données des années précédentes utilisées comme apprentissage. Dans un premier temps, nous évaluons les performances (R^2) du meilleur modèle pour la prédiction du SCA et du SDA de 2018 moyens par commune à partir des données de 2009-2017, et nous comparons sur des cartes les prédictions obtenues avec les valeurs réelles des indicateurs en 2018. Dans un second temps, nous estimons la dépendance entre le nombre d'années utilisées pour l'entraînement du modèle et ses performances, ainsi que le nombre minimum d'années nécessaires afin d'obtenir des prédictions d'une qualité satisfaisante.

2.4.5.2 Illustration avec l'année 2018

Les performances du meilleur modèle pour la prédiction des données de 2018 à partir des données de 2009-2017 sont significativement plus élevées que les performances obtenues en considérant la moyenne de 5 découpages aléatoires des données, utilisée (dans la section 2.3.3) pour choisir et calibrer le meilleur modèle : 0.57 contre 0.47 pour le SCA et 0.58 contre 0.43 pour le SDA. Ce résultat peut interroger, car nous pouvions nous attendre à des performances moins élevées qu'en faisant des divisions aléatoires de données d'entraînement et de test. En effet, si les données de 2018 ne servent pas pour l'entraînement, alors certaines caractéristiques propres à cette année ne sont pas prises en compte par le modèle lors des prédictions. Ce résultat est de bon augure, car il suggère que l'architecture du modèle utilisé, que nous avons conceptualisé, puis validé sur plusieurs découpages aléatoires des données, est robuste et fonctionne au moins aussi bien avec des données actuelles dans un contexte opérationnel. Toutefois, ce résultat doit

5. https ://fews.net/

être nuancé, aussi bien par la variabilité dans les prédictions qu'implique la taille réduite des données utilisées que par les spécificités contextuelles annuelles du pays. Ce point sera davantage discuté dans la suite de cette section.

La Figure 2.7 est constituée de quatre cartes illustrant pour 2018 la distribution par commune du SCA et du SDA issus de l'EPA, et prédits par le meilleur modèle. Les scores sont représentés en trois classes de fréquence égale. Nous constatons que malgré les performances prédictives moyennes du modèle, l'agrégation en trois classes des prédictions permet de prédire assez finement la distribution des communes à forte/moyenne/faible sécurité alimentaire par rapport aux résultats de l'EPA, et ce pour les deux indicateurs exploités. Parmi les 318 communes considérées en 2018, 211 (66%) communes sont de même classe pour le SDA (resp. 201 (63%) pour le SCA), 100 (32%) communes ont une classe d'écart pour le SDA (resp. 108 (34%) pour le SCA), et 7 (2%) communes sont de classes opposées pour le SDA (resp. 9 (3%) pour le SCA). Pour les deux scores, les prédictions de notre modèle ont donc permis de classer identiquement les communes que pour l'enquête EPA dans une majorité des cas (66% pour le SDA et 63% pour le SCA). À l'inverse, seule une petite minorité des communes (moins de 3% pour les deux scores) ont été classées de manière opposée. Ces résultats ne sont pas encore excellents, mais ils sont très encourageants sachant la complexité de la tâche. De plus, bien que le SCA et le SDA soient très corrélés, leurs distributions spatiales possèdent certaines spécificités visibles sur les cartes calculées à partir de l'EPA, et qui sont également observables sur les cartes obtenues par les modèles. Par exemple : la distribution du SCA est plus déstructurée, moins homogène pour les communes du centre du pays que pour le SDA ; le SCA est moins critique dans certaines communes de la région du Sahel, au nord du pays, que le SDA ; le SDA est moins critique dans la languette qui forme l'extrême sud du pays (région Sud-Ouest) que le SCA. La carte formée par notre modèle permet donc de détecter certaines spécificités spatiales, propres à chaque indicateur, qui sont relativement subtiles. Enfin, soulignons que, bien que nous ne le fassions pas ici, les communes qui n'ont pas été investies par l'EPA en 2018 (en blanc sur les cartes) pourraient être prédites par notre modèle. Cela constitue un intérêt opérationnel évident, puisque les communes qui ne peuvent pas être investies par manque de temps, d'argent, ou par incapacité de s'y rendre (e.g., conflits, climat) peuvent être incluses par notre modèle, et donc l'état de la sécurité alimentaire peut y être estimé.

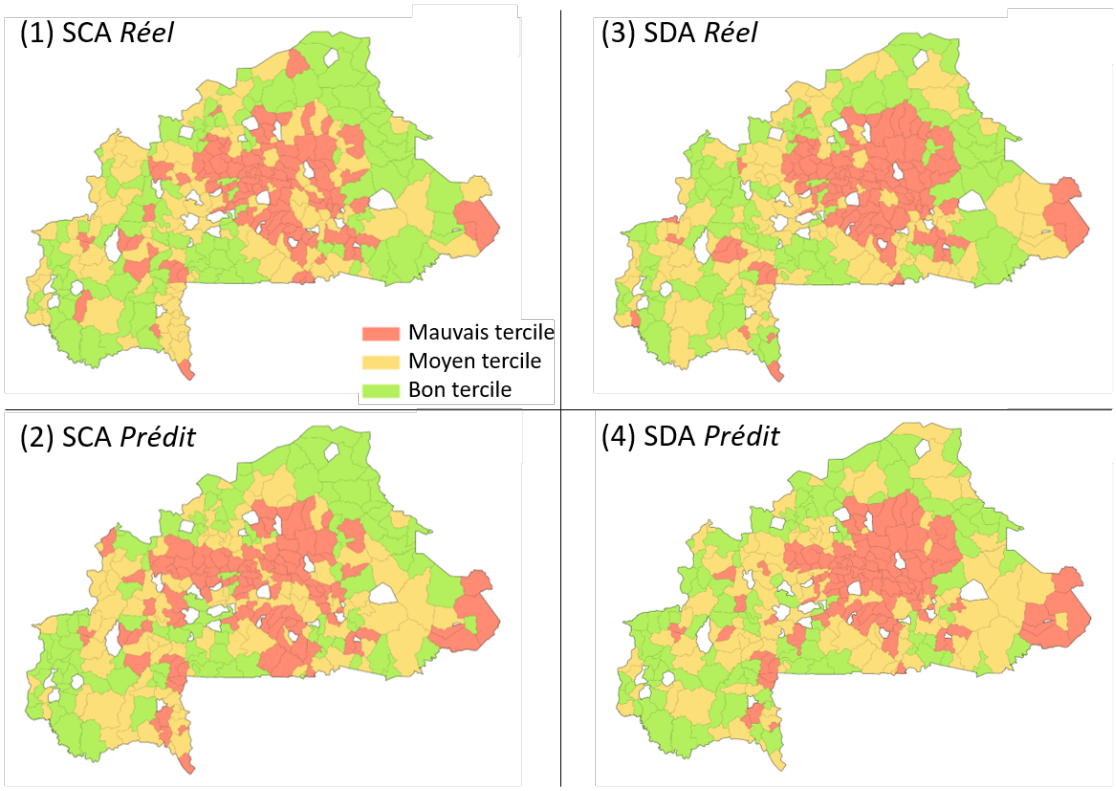


FIGURE 2.7 – Quatre cartes représentant pour 2018 la distribution spatiale du SCA réel (1) et prédit (2) ainsi que du SDA réel (3) et prédit (4), en moyenne par commune. Les prédictions sont réalisées par le modèle b) intégrant les caractéristiques du CNN ainsi que les variables CS. Les distributions sont représentées en 3 classes de fréquence égale.

2.4.5.3 Dépendance de l'approche au nombre de données

Nous nous penchons maintenant sur la dépendance entre la qualité des prédictions des scores de sécurité alimentaire d'une année donnée et le nombre d'années précédentes utilisées pour l'entraînement. L'hypothèse étant que la qualité des prédictions augmente lorsque le nombre d'années utilisées pour l'entraînement, et donc de données, augmente. Nous justifions cette hypothèse en deux points. Tout d'abord, plus le nombre d'années utilisées pour entraîner le modèle est grand, plus celui-ci disposera d'une grande variété d'exemples liés à des phénomènes ayant impacté la sécurité alimentaire sur les années et communes considérées, pour la génération de ses règles de décisions. Le modèle obtenu est donc d'avantage généralisable. De plus, les méthodes d'apprentissage profond

exploitées dans notre modèle ont tendance à sur-ajuster les données quand celles-ci sont peu nombreuses, et fonctionnent de manière optimisée sur des jeux de données de taille conséquente.

Le Tableau 2.5 illustre l'évolution des performances de prédiction du SCA et du SDA dans un contexte opérationnel, en considérant que l'outil (i.e., le modèle (b)) est utilisé chaque année depuis 2014 pour prédire les scores de l'année en cours (i.e., avant que l'EPA ne les fournissent) à partir des données antérieures (i.e., les données de 2009 à $n-1$, où n est l'année prédite). Nous observons que les meilleures performances sont obtenues pour l'année 2018, prédite avec le plus gros jeu d'entraînement (2748 observations). À l'inverse, pour l'année 2014 qui est prédite avec le jeu d'entraînement le plus réduit (seulement 1491 observations), les performances chutent fortement avec des R^2 inférieurs à 0.1, rendant les prédictions inutilisables. Ce seuil d'au moins 6 années représentées dans le jeu d'entraînement semble être dans notre cas une condition minimale pour l'utilisation de notre modèle. Malgré une tendance des performances de prédiction à la hausse avec le nombre d'années utilisées pour l'entraînement, nous constatons que les performances associées à chaque année prédite ne sont pas strictement croissantes. Par ailleurs, la forte augmentation des performances entre 2017 et 2018 doit en particulier nous interroger, car cette augmentation ne peut pas être seulement due au gain en données entre 2017 et 2018. Nous proposons deux justifications à ce phénomène. D'une part, la taille réduite des données utilisées pour l'entraînement implique une certaine variabilité dans les prédictions du modèle utilisé, qui intègre des algorithmes nécessitant une grande quantité de données pour fonctionner de manière optimale. Et d'autre part, les spécificités contextuelles annuelles du pays peuvent rendre certaines années plus facilement prévisibles que d'autres. De ce fait, l'année 2018 peut avoir été une année pour laquelle la prédiction des scores de sécurité alimentaire a été spécifiquement aisée. Malgré cela, l'hypothèse d'un gain de performances du modèle lorsque l'historique des données utilisées pour l'entraînement augmente est fondée, mais cette tendance de notre outil à obtenir des performances croissantes en fonction de la taille des données devra être confirmée lorsque les années passeront et que de nouvelles données de l'EPA viendront enrichir le jeu d'entraînement.

Année prédite	SCA	SDA	Taille d'apprentissage
2018	0.57	0.58	2748
2017	0.29	0.33	2430
2016	0.24	0.38	2112
2015	0.28	0.35	1799
2014	0.09	0.05	1491

Tableau 2.5 – Performances (R^2) de prédiction du SCA et du SDA par commune en utilisant comme jeux de données tests les données de chaque année entre 2014 et 2018, et en utilisant pour l'entraînement les données des années précédentes (i.e., les données de 2009 à $n-1$, où n est l'année utilisée pour le test) et taille des jeux d'apprentissage associés. Nous utilisons le modèle b) intégrant les caractéristiques du CNN ainsi que les variables CS.

2.5 Conclusion

Dans ce chapitre, nous avons proposé le framework $FSPHD$, qui exploite des méthodes d'apprentissage automatique pour prédire par régression deux indicateurs clés de la sécurité alimentaire, normalement obtenus par le biais d'enquêtes ménages longues et coûteuses. Pour valider nos modèles, nous avons utilisé comme vérité terrain une base de données incluant des ménages issus de tout le Burkina Faso de 2009 à 2018, en utilisant principalement des données ouvertes globales comme variables explicatives afin de garantir la reproductibilité et la généralisation de nos méthodes à d'autres pays. Deux verrous scientifiques ont été considérés : 1) l'aspect multifactoriel de la sécurité alimentaire, qui implique un choix de données d'entrée hétérogènes (aux niveaux thématique, structurel et de l'échelle spatio-temporelle) ainsi que des prétraitements appropriés pour maximiser la contribution de chaque type de données. Pour prendre en compte un maximum de composantes de la sécurité alimentaire, nous avons intégré des données issues de différentes thématiques (structure du paysage, dynamique des populations, qualité des sols, météorologie, végétation, insécurité ou économie), encodées selon différents types (valeurs quantitatives, données géolocalisées, vecteurs lignes, séries temporelles et images) et avec différentes granularités spatio-temporelles, et nous avons dû effectuer des traitements adaptés pour extraire des informations pertinentes de ces données (e.g., agrégations, interpolations, normalisations). 2) Il a fallu choisir et combiner des méthodes

d'apprentissage automatique adaptées à chaque type de variable. Nous avons observé que les performances (R^2) obtenues par nos modèles sont statistiquement significatives sans être élevées, ne dépassant pas $0,469$ et $0,434$ de R^2 pour la prédiction du SCA et du SDA, respectivement. Cependant, les résultats de ce chapitre sont supérieurs à la plupart des rares travaux auxquels nous avons pu nous comparer. Ces résultats indiquent que la prédiction de ces indicateurs de sécurité alimentaire est une tâche complexe. Ce travail apporte une contribution supplémentaire à cette question peu étudiée. Nous avons également constaté la contribution significative des modèles d'apprentissage profond (CNN) au traitement des données HRS. L'utilisation de ce type de données dans le contexte de la sécurité alimentaire constitue donc une piste pertinente pour de futurs travaux. En revanche, l'utilisation de modèles d'apprentissage profond (LSTM) pour le traitement des séries temporelles dans le contexte de la sécurité alimentaire n'a pas donné de résultats notables, et des recherches futures devront inclure un meilleur traitement des séries temporelles car l'aspect temporel joue un rôle important dans la sécurité alimentaire. De plus, nous avons démontré l'apport de modèles combinant différents types de méthodes d'apprentissage automatique adaptées à chaque type de données, ce qui confirme la pertinence de ce type d'approche. Les modèles de type (b) basés sur la fusion au niveau des caractéristiques nous ont permis d'obtenir les meilleures performances. Nous avons également observé que les variables indiquées par les modèles comme étant les plus importantes pour la prédiction des indicateurs de sécurité alimentaire proviennent de multiples domaines, ce qui confirme la nécessité de relier la sécurité alimentaire à un large spectre de domaines connexes afin d'obtenir l'image la plus complète possible de ce champ complexe et multifactoriel. Enfin, nous avons discuté de la mise en œuvre de notre outil dans un contexte opérationnel. D'un point de vue technique, le fait que les données nécessaires à l'utilisation des modèles soient ouvertes et que ces derniers puissent être exécutés rapidement et automatiquement au moyen d'un programme informatique permet en théorie de les utiliser en temps réel pour la prédiction d'indicateurs de sécurité alimentaire. Cependant, la qualité des prédictions suscite encore des interrogations, et semble indiquer qu'il est encore actuellement inapproprié d'utiliser un tel outil dans ce contexte. Les résultats obtenus doivent être consolidés, ou nuancés par l'intégration de nouvelles données dans les années à venir. Néanmoins, malgré la complexité de cette tâche, nous avons pu obtenir des résultats significatifs, porteurs de sens et encourageants.

De futurs travaux consisteront à améliorer l'architecture des modèles d'apprentissage profond et à intégrer de nouvelles données afin d'améliorer la qualité de nos modèles et,

à terme, de rendre ce type d'outil exploitable au niveau opérationnel, en appui des SSA. Cependant, si les modèles prédictifs sophistiqués tels que présentés dans ce chapitre ont du sens et une utilité pour cibler les zones à risque et réagir rapidement en cas de famine, ils n'apportent qu'un faible aspect explicatif. Ces modèles prédictifs, en particulier les réseaux de neurones, ont un effet "boîte noire" qui implique souvent un manque de pouvoir explicatif et d'interprétabilité, faisant d'ailleurs actuellement l'objet de tout un axe de recherche (Montavon et al., 2018 ; Khormuji and Rostami, 2021). Pourtant, être capable de comprendre les mécanismes qui sous-tendent les famines est au moins aussi important que leur détection, car c'est en comprenant leurs causes et leurs caractéristiques que l'on peut réellement mettre en place des pistes de réflexion menant à une stabilité alimentaire durable. Cela implique l'utilisation d'approches et de données d'une autre nature, nous permettant d'obtenir des informations qualitatives complémentaires aux prédictions issues de l'apprentissage automatique et des données quantitatives numériques. Dans le prochain chapitre, nous proposons d'aborder cette problématique à travers le prisme des données textuelles et plus précisément des articles de presse, qui comme nous le montrerons possèdent un fort potentiel explicatif, en particulier dans ce contexte de la sécurité alimentaire.

Chapitre 3

Expliquer la sécurité alimentaire à partir de données textuelles

3.1 Introduction

Les systèmes d'alerte et de surveillance de la sécurité alimentaire (SSA) s'appuient principalement sur des données numériques pour effectuer leurs analyses (e.g., données d'enquêtes, prix du marché, données agrométéorologiques, images satellitaires), alors que les données textuelles, plus complexes à traiter, sont peu exploitées (Becker-Reshef et al., 2020). Or, ce type de données de plus en plus abondantes et accessibles sur le web offre une source d'informations pertinentes et complémentaires aux données quantitatives, et pourrait s'avérer d'une grande utilité pour obtenir de l'information sur les signaux de détresse alimentaire de populations qui ne peut être obtenue par le biais de données quantitatives ou d'images satellitaires. L'étude de Li et al. (2017) appuie par exemple le potentiel lié à la complémentarité entre images satellites et données de réseaux sociaux. D'autres travaux abordent les liens entre images satellites et données textuelles pour la détection de changements (Kergosien et al., 2015 ; Neptune and Mothe, 2021).

Les données textuelles issues de sites d'actualités ou de réseaux sociaux permettent d'obtenir des informations qualitatives complémentaires à celles obtenues avec des données quantitatives numériques. Ce fait est particulièrement intéressant dans le contexte de la sécurité alimentaire, dont les causes et les manifestations sont multiples. L'analyse sémantique de ce type de données peut produire des informations précises sur les thèmes abordés, la polarité du discours (i.e., son caractère positif ou négatif), ou encore sur le

vocabulaire spécifique employé. De plus, une donnée textuelle accessible sur le web est généralement localisable dans le temps et fréquemment dans l'espace. Cette association avec un contexte spatial et temporel peut être réalisée de manière basique, si la donnée est directement associée à une date et à une localisation, ou avec des méthodes de fouille de texte si cela est nécessaire (e.g., mise en correspondance avec un lexique géographique (Itoh et al., 2016) et/ou détection d'entités nommées (Xiao et al., 2019)), pour retrouver ces informations dans le texte. Cette nécessité de pouvoir associer des données textuelles à un contexte spatio-temporel est particulièrement pertinente dans le contexte de la sécurité alimentaire, dont les crises sont ciblées dans des zones et à des périodes précises que l'on cherche à expliquer.

Toutes ces informations qualitatives et spatio-temporelles extraites de données textuelles peuvent être utilisées à deux fins :

1. utilisation en entrée de modèles prédictifs (e.g., modèles basés sur des méthodes d'apprentissage automatique et profond) conjointement à d'autres types de données pour augmenter la précision des prédictions d'indicateurs de sécurité alimentaire ;
2. utilisation comme données complémentaires aux modèles prédictifs qui présentent souvent un manque d'explicabilité et d'interprétabilité en apportant un contexte explicatif. C'est sur cet aspect que nous nous focalisons dans ce chapitre.

Concernant l'extraction d'informations à partir de données textuelles, et en particulier dans le cas d'informations liées à la sécurité alimentaire, le nombre de dimensions à prendre en compte lors de l'analyse d'un texte est triple : le message du texte, sa temporalité et sa spatialité. Pour extraire au mieux la sémantique ainsi que le contexte spatio-temporel de données textuelles, la fouille de textes, qui désigne un ensemble de méthodes d'extraction automatique de connaissances à partir de données textuelles, offre une piste intéressante (cf. section 3.2).

Un autre enjeu concerne le choix de la source de données textuelles. Les régions du monde les plus touchées par l'insécurité alimentaire sont également souvent les régions les plus pauvres, pour lesquelles l'usage d'internet est encore très minoritaire, ne permettant pas d'effectuer des analyses pertinentes à partir de messages de réseaux sociaux. Par exemple, le Burkina Faso possède 7,8% de taux de pénétration des médias sociaux contre 16% en moyenne en Afrique et 49% en moyenne dans le monde, ce qui en fait l'un des

pays les moins utilisateurs de réseaux sociaux (Le Kiosque Digital du Burkina, 2020). Les journaux, publiant quotidiennement des articles portant sur des sujets, des lieux et des évènements divers, constituent une riche source d'informations textuelles et offrent une alternative intéressante aux média sociaux. De plus, depuis les années 2000 la plupart des journaux ont créé leur site web pour y publier leurs articles en ligne, cette mise en format numérique permet d'y appliquer des traitements automatiques.

Dans ce chapitre, nous examinons la capacité des méthodes de fouille de textes à extraire et analyser des informations qualitatives utilisées comme proxies de la situation alimentaire globale, régionale et sur son évolution au cours des dix dernières années au Burkina Faso à partir de corpus de journaux du pays. L'utilité de cette approche est de proposer un cadre explicatif complémentaire aux sorties des modèles prédictifs appliqués aux autres types de données (e.g., aux données numériques et images satellitaires) proposés dans le chapitre précédent.

Dans la section 3.2, nous dressons l'état des connaissances concernant les méthodes d'extraction et d'analyse d'informations thématiques et spatio-temporelles à partir de données textuelles dans le contexte de la sécurité alimentaire et des thèmes liés. La section 3.3 présente les données ainsi que la méthodologie d'extraction d'informations déployée. La section 3.4 expose et discute les informations extraites du corpus constitué dans le cadre de ces travaux de thèse.

3.2 État de l'art

Dans cette section, nous passons dans un premier temps en revue des études qui ont utilisé des méthodes d'extraction d'informations thématiques à partir de textes, puis plus précisément d'autres études dont les informations thématiques extraites sont liées à la sécurité alimentaire. Dans un second temps, nous exposons les principales méthodes de traitement de données spatio-temporelles puis explicitons de quelle manière nous prenons en compte l'aspect spatio-temporel des données dans ce chapitre.

3.2.1 Analyse de données textuelles pour la sécurité alimentaire et les crises

Depuis plusieurs années, la multiplication des données textuelles accessibles en ligne, par le biais de réseaux sociaux, de forums ou encore de sites d'actualités couplé aux progrès au niveau des traitements permet l'analyse efficace de ces données complexes. Ce type de données très qualitatives constitue une riche source d'informations thématiques sur la manière de s'exprimer de populations, ainsi que sur leurs sujets d'intérêt ou leurs sentiments. Des outils tirant profit de cette tendance ont été développés dans un large éventail de domaines en mobilisant des techniques de fouille de texte, un domaine qui se concentre sur l'exploration des données textuelles (Berry and Kogan, 2010). Par exemple, en traduction automatique du langage (DeepL GmbH, 2017), dans le domaine de la sécurité (Webb, 2007) ou encore dans le domaine médical (Jenssen et al., 2001).

Nous faisons à présent l'inventaire de l'état de la recherche d'informations thématiques à partir de données textuelles appliquée à la sécurité alimentaire et aux thèmes liés afin d'identifier des proxies "textuels" de la sécurité alimentaire pertinents. Une première source d'information thématique qui peut être intéressante pour évaluer l'état de la sécurité alimentaire est l'étude et le suivi de l'aspect positif ou négatif qui ressort des données textuelles en lien avec la sécurité alimentaire. L'analyse de sentiments est une branche de l'analyse textuelle qui a pour but l'analyse de grandes quantités de données pour en extraire les différents sentiments qui y sont exprimés. Les sentiments extraits peuvent ensuite être exploités pour générer des statistiques sur les sentiments généraux d'une population. Par exemple, Zhang et al. (2011) ont quantifié les sentiments dans des tweets et ont montré que ces sentiments étaient associés à l'évolution d'indices boursiers. Des méthodes d'analyse de sentiments utilisant de l'apprentissage automatique ont également été utilisées dans des domaines plus proches de la sécurité alimentaire. Par exemple, en étant appliquées à l'évolution de prix des denrées, permettant de démontrer que leur augmentation était significativement associée à une polarité davantage négative dans les tweets (Surjandari et al., 2014 ; UN Global Pulse, 2014), la polarité étant une mesure de l'aspect positif ou négatif d'un texte. D'autres études démontrent qu'en cas de stress alimentaire et/ou de crises, les comportements au téléphone (Lukyamuzi et al., 2015 ; Decuyper et al., 2014) et sur les réseaux sociaux (Acar and Muraki, 2011 ; Middleton et al., 2014) sont différents : la population a tendance à émettre davantage d'appels et de messages et ceux-ci portent davantage sur les thèmes de la sécurité alimentaire

et des crises. La proportion de messages liés aux thèmes de la sécurité alimentaire et des crises ainsi que le vocabulaire utilisé dans ces messages peuvent donc être considérés comme des proxies de ces domaines. Ce sont ces différences de comportements perceptibles dans les textes que nous cherchons à isoler et à analyser avec l'outil conçu dans ce chapitre.

Cependant, les difficultés associées à la mise en place de tels outils sont liées à la complexité structurelle des données textuelles et font l'objet d'un grand nombre d'études. Nous détaillons maintenant les approches proposées dans la littérature scientifique pour répondre à ces difficultés. Dans le domaine de l'agriculture qui est étroitement lié à la sécurité alimentaire, l'extraction d'informations à partir de données textuelles est un sujet qui suscite de plus en plus d'intérêt (Drury and Roche, 2019). Dans ce domaine, plusieurs études ont porté sur l'analyse de sentiments (Surjandari et al., 2014 ; Cruz et al., 2015), l'extraction d'entités nommées (i.e., des lieux, des dates ou des individus en lien avec l'agriculture) (Biswas et al., 2015 ; Malarkodi et al., 2016) ainsi que sur l'extraction de vocabulaire spécifique (Martin et al., 2021) (par exemple avec l'utilisation du Thesaurus *Agrovoc*[1] : Roche et al. (2015)), pour évaluer les conditions agricoles. A cette fin, ces études exploitent des méthodes de fouille de texte basées sur de l'apprentissage automatique, du clustering ou encore sur l'utilisation d'ontologies (i.e., des ensembles structurés de termes et de concepts destinés à organiser les connaissances relatives à un domaine). Les applications de cet axe de recherche au contexte agricole sont variées : prédiction des prix des matières premières (Kim et al., 2017), détection de présence d'insectes nuisibles (Bermeo Almeida et al., 2018) ou encore gestion agricole (Liao et al., 2015). Des méthodes d'extraction d'informations à partir de données textuelles dans les domaines de la sécurité alimentaire et des crises font actuellement l'objet de recherches. Concernant les crises, Interdonato et al. (2019) ont eu comme objectif l'extraction non supervisée d'informations pertinentes sur les crises à partir de données de Twitter, ce qui constitue une problématique relativement proche de la nôtre. Afin de produire un classement des tweets les plus informatifs lors d'une situation de crise, l'étude intègre et compare dans un premier temps trois techniques de topic modeling (i.e., en français "modélisation de sujets", ce qui correspond dans ce contexte à du clustering de données textuelles pour identifier des sujets communs à des ensembles de textes) sur le corpus de tweets : 1) Latent Dirichlet Allocation (LDA) (Blei et al., 2003) qui est un modèle probabiliste génératif permettant de mettre en évidence un ensemble de thèmes sous-

1. http ://www.fao.org/agrovoc/fr/access

jacents qui composent un ensemble de textes ; 2) Nonnegative matrix factorization (Wang and Zhang, 2012), basé sur la décomposition d'une matrice textes–termes en deux matrices textes–sujets et termes–sujets ne possédants aucun coefficient négatif ; 3) K-means (Wang et al., 2012), qui propose une division de vecteurs de termes en k groupes de façon à minimiser la distance entre les vecteurs à l'intérieur de chaque groupe, chaque groupe étant considéré comme un sujet. Dans un second temps, les tweets associés aux sujets les plus en lien avec le thème des crises sont sélectionnés puis classés selon leur similarité sémantique avec un lexique de termes du champ lexical des crises, en testant et comparant deux techniques de vectorisation de textes : 1) Word2vec (Mikolov et al., 2013) qui désigne une famille de modèles de traitement automatique du langage utilisant des réseaux de neurones et permettant la transformation de termes et de textes en vecteurs qui contiennent des informations sémantiques et syntaxiques prenant en compte le contexte de chaque terme (i.e., les termes fréquemment rencontrés autour de chaque terme étudié) ; 2) Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) qui permet également d'obtenir une représentation vectorielle de termes et de textes, en utilisant un corpus de documents (en général Wikipédia) comme base de connaissances. L'approche spécifique de cette technique est qu'un terme est représenté comme un vecteur colonne dans la matrice associée au corpus textuel et qu'un document est représenté comme le centroïde des vecteurs représentant ses termes. Les meilleures performances ont été obtenues avec LDA et Word2vec. Notons que les méthodes de vectorisation de textes comme Word2vec peuvent contribuer à identifier des textes traitant d'un thème, en appliquant une mesure de similarité entre un texte et un lexique thématique préalablement vectorisés par ces méthodes, puis en fixant un seuil de similarité à partir duquel le thème du lexique est attribué au texte. C'est d'ailleurs une méthode employée par Basu et al. (2017), qui se concentrent sur le problème de l'identification automatique des tweets informant sur les besoins et les disponibilités en ressources, appelés respectivement "need-tweets" et "availability-tweets" après une catastrophe. Leur méthode est basée sur du plongement lexical : premièrement, les tweets qui contiennent des termes d'intérêt stockés dans un lexique (e.g., "need", "requir", "distribut") sont pré-sélectionnés, puis la similarité sémantique entre chaque tweet et le lexique est calculée avec les méthodes Indri (Strohman et al., 2005) et Word2vec, les tweets sont enfin triés en fonction de leur similarité sémantique avec le lexique. La validation est effectuée avec un ensemble de tweets manuellement labélisés, les meilleures performances sont également obtenues avec Word2vec. Concernant la sécurité alimentaire, Lukyamuzi et al. (2018) proposent une stratégie d'extraction d'informations sur cette thématique à partir de traitement

de tweets par plusieurs modèles d'apprentissage automatique (e.g., support vector machine (Hearst et al., 1998) et multilayer perceptron (Ramchoun et al., 2016)). Leurs performances sont mitigées, possiblement à cause d'un pré-traitement trop pauvre des tweets qui sont seulement transformés en sac de mots par tokenisation, ce qui pointe l'importance de l'étape de prétraitement des données textuelles. Dans notre étude, nous nous inspirons de l'approche utilisée par Basu et al. (2017) et Interdonato et al. (2019), basée sur l'utilisation de Word2vec avec un lexique thématique pour détecter les articles traitant de sécurité alimentaire.

Très peu de recherches portent sur les articles de journaux dans le contexte de la sécurité alimentaire, pourtant ce support textuel peut nous fournir les proxies proposés ci-dessus. Il a été mis en évidence que ce type de support, publié de façon régulière, contient des informations essentielles sur la sécurité alimentaire et les thèmes liés (e.g., changement climatique, récolte de l'eau, production agricole) (Kutyauripo et al., 2021). Un seul article à notre connaissance utilise des méthodes de fouille de textes pour extraire de l'information sur des évènements liés à la sécurité alimentaire à partir de journaux, c'est l'étude de Xiao et al. (2019) qui propose un framework de détection automatique de crises alimentaires. Leur méthode consiste à extraire, pour chaque article, le vocabulaire le plus caractéristique (mots-clés) par tf-idf (term frequency-inverse document frequency) (Salton and Buckley, 1988) qui est une méthode de pondération de termes caractéristiques de textes (détaillée dans la section 3.3.2), puis à extraire les entités nommées avec un framework Bi-LSTM-CNN-CRF (Yu, 2019). Un poids est associé à chaque mot-clé en fonction de sa similarité sémantique (par Word2vec) avec les termes du titre de l'article. Enfin, chaque article, par le biais de l'ensemble des mots-clés pondérés et des entités nommées associées, sont classifiés par single-pass clustering (Papka et al., 1998). La capacité du tf-idf à extraire le vocabulaire pertinent et spécifique des articles de journaux a été mise en évidence dans de nombreuses autres études (Wang et al., 2017 ; Yao et al., 2019 ; Ao et al., 2020).

Dans nos travaux, nous nous concentrons sur des descripteurs textuels de sécurité alimentaire liés à la fréquence d'articles traitant de sécurité alimentaire (obtenue avec l'approche basée sur l'utilisation de Word2vec et d'un lexique thématique), à la polarité de ces articles ainsi qu'au vocabulaire relatif à la sécurité alimentaire et aux crises le plus employé (extraits à partir de lexiques thématiques et pondérés en utilisant la notion de tf-idf).

3.2.2 Analyse spatio-temporelle sur des données textuelles

En plus de l'information sur la thématique des articles et la sémantique des termes qui les composent, qui peut nous permettre de comprendre les caractéristiques des crises alimentaires, la dimension spatio-temporelle est une composante importante de l'extraction d'informations. Comme déjà considéré dans les chapitres précédents, il est crucial pour appréhender la sécurité alimentaire de pouvoir moduler les analyses dans le temps et dans l'espace afin de comprendre l'enchaînement des mécanismes sous-jacents aux crises et aux famines. Or le traitement de données spatio-temporelles est loin d'être trivial, particulièrement pour des données textuelles aux nombreuses subtilités de langages dont le traitement est complexe en soi. Ce type de traitements peut être décomposé en deux phases.

Une première phase se focalise sur l'extraction des informations spatio-temporelles dans les textes, qui peuvent être contenues dans les métadonnées associées (e.g., date et localisation de l'envoi d'un tweet ou de la rédaction d'un article parfois disponibles). Le cas échéant, ces informations peuvent être directement extraites dans le corps des textes en utilisant des méthodes de fouilles de textes adaptées (e.g., mise en correspondance avec un lexique géographique (Itoh et al., 2016), utilisation de gazetiers tels que ceux proposés par OpenStreetMap [2] ou Geonames [3] (Fize et al., 2017), détection d'entités nommées (Xiao et al., 2019)).

Puis vient une seconde phase d'analyse spatio-temporelle de ces textes sur laquelle nous nous concentrons davantage dans ce chapitre. Pour prendre en compte l'aspect séquentiel propre aux données spatio-temporelles, plusieurs méthodes sophistiquées existent et sont issues de différents axes de recherche. Tout d'abord, les réseaux de neurones récurrents, une famille de modèles d'apprentissage profond adaptés aux données séquentielles. Cette méthode permet d'extraire l'information liée à la sémantique des mots, tout en prenant en compte leur ordre dans une phrase en la traitant comme une séquence de termes vectorisés (par plongement lexical), mais également de prendre en compte le contexte spatio-temporel en considérant des suites de textes associés à des lieux comme des séquences temporelles spatialisées (Diaz et al., 2020). Le recours à la théorie des motifs séquentiels est également possible pour le traitement de ce type de données tridimensionnelles (Wong et al., 2000). L'exploration de motifs séquentiels peut

2. https ://www.openstreetmap.fr/
3. http ://www.geonames.org/

être effectuée pour détecter des suites d'évènements qui se produisent fréquemment dans le même ordre au cours du temps et selon la zone géographique (i.e., dans notre contexte des séquences de termes qui reviennent fréquemment dans le temps pour un lieu précis). Enfin, l'approche probabiliste a été explorée dans ce contexte par l'utilisation de modèles probabilistes d'apparition des termes possédant des paramètres spatiaux et temporels (Mei et al., 2006). Ces trois types de méthodes sont intéressantes mais nécessitent un corpus de données dense au niveau spatial (i.e., permettant la mise en relation fréquente des textes avec des localisations), ce qui n'est pas le cas pour des articles de journaux qu'il est souvent difficile d'associer à une localisation. Pour faire face à ce problème de densité spatiale des données, plusieurs études portant sur la détection de crises à partir de données textuelles utilisent des méthodes plus sommaires mais davantage adaptables à des données spatio-temporelles moins régulières. Middleton et al. (2014) proposent des cartographies de catastrophes naturelles, construites en détectant les zones (e.g., rues, places, rivières) citées par des tweets qui contiennent des mots-clés du champ lexical des catastrophes, durant les jours qui suivent une catastrophe. Une zone est considérée comme victime d'une catastrophe si le nombre de tweets associés varie par rapport à une période dite "normale". Itoh et al. (2016) proposent une méthode de visualisation d'événements spatio-temporels à partir de corpus de tweets. Chaque tweet est associé à sa date d'émission disponible dans les métadonnées et spatialisé par détection dans le corps des tweets de localités listées dans un lexique géographique. Pour un intervalle de temps et d'espace donnés, les tf-idf de chaque terme des tweets sont calculés par rapport aux tweets du corpus entier dans le but d'obtenir les termes les plus caractéristiques du contexte spatio-temporel étudié. Enfin, les termes possédant les tf-idf les plus élevés sont visualisés par un nuage de mots.

Dans notre étude, nous cherchons à identifier à partir d'articles de journaux des proxies textuels pertinents pour l'analyse des tendances spatiales et temporelles de la sécurité alimentaire. Compte tenu du peu d'articles localisables dans l'espace, nous nous inspirons de la méthodologie de Itoh et al. (2016) pour traiter l'aspect spatio-temporel des articles de journaux (méthode détaillée dans la section 3.3.2). Contrairement aux travaux cités, le corpus étudié (détaillé dans la section suivante) est défini sur une longue période (10 ans) tout en étant de taille conséquente (plus de 20 000 articles), ce qui permet l'extraction de tendances significatives en matière de sécurité alimentaire et de leur évolution sur le long terme. Enfin, des sorties visuelles adaptées à chaque niveau d'analyse spatiale et temporelle sont proposées : nuages de mots, graphiques de séries

temporelles, graphes de co-occurrences, ainsi que des graphiques radar (i.e., en toile d'araignée) utilisés pour visualiser les variations temporelles du vocabulaire employé dans les articles, se révélant appropriés dans notre contexte. Très peu de travaux ont utilisé cet outil (i.e., le graphique radar) pour la visualisation de données spatio-temporelles (Forlines and Wittenburg, 2010 ; Reski et al., 2020), et parmi ces études, aucune à notre connaissance ne traite de données textuelles. D'autres visualisations qui ne sont pas employées dans ce chapitre pourraient également être exploitées, comme les diagrammes de Venn (Ho et al., 2021).

3.3 Matériel et méthodes

3.3.1 Données

Dans cette section, nous décrivons tout d'abord le corpus de journaux exploité, puis nous exposons de quelle manière ont été conçus les lexiques utilisés pour l'analyse du corpus.

3.3.1.1 Corpus de journaux

Actuellement, les principaux journaux burkinabés possèdent un site web d'actualités sur lequel ils publient leurs articles. C'est le cas pour les journaux les plus importants tels que l'Observateur Paalga [4], Lefaso [5], Sidwaya [6], Le Journal du jeudi [7], Burkina24 [8], ou encore Fasozine [9], bien que la facilité de recueil automatique des articles soit inégale en fonction des journaux. Pour la création du corpus de journaux, nous nous sommes tournés vers deux journaux burkinabés dont les sites web permettent une bonne accessibilité aux données : Burkina24 et LeFaso. Ces journaux, qui sont des quotidiens d'informations générales (e.g., économie, société, politique, culture), comptent parmi les plus lus dans le pays et possèdent en ligne un grand nombre d'articles sur des thématiques variées. Burkina24 et Lefaso sont disponibles sur le web depuis 2011 et 2003 respectivement ; nous avons pu en extraire 22856 articles au total entre 2009 et 2018 (5595 pour Burkina24

4. http ://www.lobservateur.bf/
5. https ://lefaso.net/
6. https ://www.sidwaya.info/
7. https ://www.journaldujeudi.com/
8. https ://www.burkina24.com/
9. http ://www.fasozine.com/

et 17261 pour LeFaso), période au cours de laquelle la sécurité alimentaire a connu des variations significatives et plusieurs crises (voir chapitre 1 et section 3.4). Les prétraitements appliqués aux articles pour la création du corpus sont détaillés dans la section 3.3.2.3. Une synthèse des caractéristiques générales des articles de ces deux journaux est disponible dans le Tableau 3.1. Le corpus de journaux extrait est disponible en accès restreint pour raison juridique sur Dataverse (Deléglise et al., 2021c).

Journal	**Nb articles**	**Nb mots***	**Ecart type Nb mots**	**Nb caractères***
Burkina24	5,595	539	330	2,967
LeFaso	17,261	831	716	4,547
Corpus	22,856	760	655	4,160

Tableau 3.1 – Caractéristiques générales des articles en français issus des deux journaux étudiés entre 2009 et 2018. (* *en moyenne par article*)

La Figure 3.1 représente le nombre d'articles du corpus par année de 2009 à 2018. L'augmentation du nombre d'articles en fonction du temps, qui est une tendance commune avec les deux journaux étudiés, est dû à la tendance croissante des journaux à publier des articles sur leur site web et au fait que la publication des premiers articles du journal Burkina24 sur le web ne date que de 2011. Les articles extraits au format texte contiennent les informations suivantes : titre, date de parution et corps de l'article. La Figure 3.2 donne un exemple d'article extrait en format texte. Dans la suite de ce chapitre, les ensembles d'articles formés par ces deux journaux sont regroupés pour former le corpus qui sera analysé. Nous justifions cela par le fait que résultats présentés dans la section 3.4 sont comparables lorsque l'on considère les articles associés à chaque journal séparément, et que le fait d'effectuer les analyses sur l'ensemble du corpus allège le propos et rend les statistiques et les conclusions plus robustes.

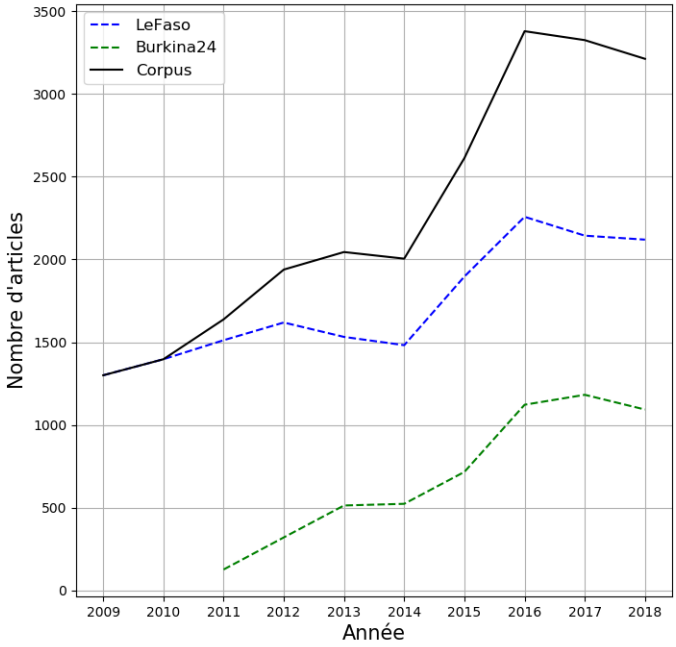


FIGURE 3.1 – Évolution du nombre d'articles par an de 2009 à 2018 dans le corpus étudié (courbe noire) et restreint aux journaux LeFaso (courbe bleue) et Burkina24 (courbe verte).

> Campagne agricole 2014-2015 : Dédougou a produit 932 829 tonnes de céréales.lundi 2 mars 2015 à 00h29min.Campagne agricole 2014-2015 : Dédougou a produit 932 829 tonnes de céréales En attendant les résultats définitifs, la direction régionale de l'Agriculture, des ressources hydrauliques, de l'assainissement et de la sécurité alimentaire estime les résultats prévisionnels de la production céréalière de la Boucle du Mouhoun à 932 829 tonnes. Ce chiffre concerne l'évaluation d'octobre 2014. « 932 829 tonnes ». Pour le moment c'est ainsi que se chiffre la production céréalière de la campagne agricole 2014-2015 de la Boucle du Mouhoun. Ces données sont issues des résultats prévisionnels d'octobre 2014. Les quantités sont reparties comme suite : le sorgho 345 362 tonnes, le mil 273 668 t, le maïs 254 174 t, le riz 49 271 t et le fonio 10.355 tonnes. En attendant les résultats définitifs, il manque 191 873 tonnes pour atteindre les objectifs de 1 million 124 702 tonnes de production céréalière que la direction régionale en charge de l'Agriculture avait fixé. Pour l'ex-directeur régional de l'agriculture de la Boucle du Mouhoun, Jean Marcel Oulé, les difficultés qui expliquent ces résultats sont entre autres : l'installation difficile des pluies, les poches de sécheresse entre juin-juillet occasionnant des ré-semis et entrainant l'abandon de certaines cultures céréalières au profit d'autres spéculations tel le sésame.

FIGURE 3.2 – Illustration d'un article publié le 2 mars 2015 par le journal LeFaso.

3.3.1.2 Lexiques

Comme exposé dans la section 3.2, une méthode efficace pour identifier les articles traitant d'un sujet est l'approche hybride basée sur du plongement lexical et sur l'utilisation d'un lexique thématique, qui s'avère plus efficace que les méthodes strictement basées sur l'utilisation de lexiques ou de Thesaurus, utilisant des ensembles fixes de termes, et ne tenant pas compte de leur aspect sémantique (Dieng et al., 2020). Plus précisément, la méthode choisie consiste à appliquer la technique de vectorisation de textes Word2vec (surclassant les méthodes ESA et Indri) pour calculer la similarité sémantique entre des articles auxquels nous voulons associer ou non un thème, et un lexique généraliste de la thématique examinée afin d'identifier les articles d'intérêt. Dans un deuxième temps, nous examinons le vocabulaire employé dans les articles sélectionnés, en détectant dans ces articles la présence de termes du champ lexical de sous-thèmes d'intérêt par mise en correspondance avec des lexiques thématiques plus détaillés et spécifiques. C'est dans cet objectif que trois lexiques sont exploités dans ce chapitre pour premièrement détecter les articles traitant de "sécurité alimentaire", puis pour examiner le vocabulaire de thèmes "sécurité alimentaire" et "crise" employé dans les articles sélectionnés. La Figure 3.3 résume le rôle des trois lexiques.

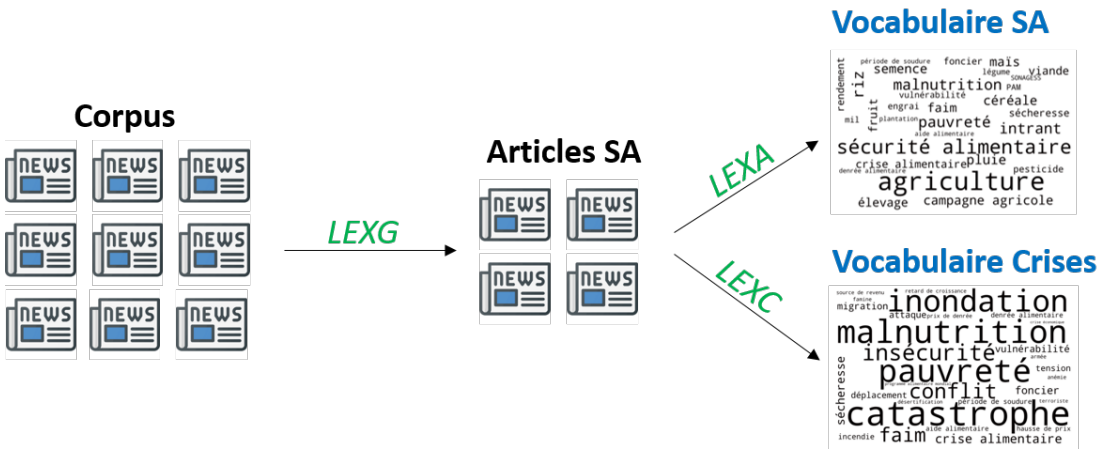


FIGURE 3.3 – Schéma du rôle des trois lexiques $LEXG$, $LEXA$ et $LEXC$.

Un premier lexique généraliste sur la sécurité alimentaire est donc tout d'abord utilisé pour détecter les articles d'intérêt (i.e., des articles mentionnant des événements en lien avec la sécurité alimentaire), nous nommons ce lexique "$LEXG$" (LEXique Généraliste). Puis, pour ces articles d'intérêt, deux autres lexiques plus détaillés sont utili-

sés pour y détecter les expressions de thèmes "sécurité alimentaire" et "crise" utilisées et ainsi obtenir un regard plus qualitatif sur le contenu des articles. Nous nommons ces deux lexiques détaillés sur la sécurité alimentaire et sur les crises respectivement "$LEXA$" (LEXique Alimentaire) et "$LEXC$" (LEXique Crises). Concernant le choix des lexiques thématiques, ce type de ressource étant peu fréquent en langue française, le nombre de lexiques potentiels est relativement faible. De plus, les lexiques et thésaurus les plus connus, comme Agrovoc, ne fournissent pas de vocabulaire suffisamment riche et complet pour le thème de la sécurité alimentaire (e.g., pour Agrovoc, le lexique associé au concept de "sécurité alimentaire" ne contient que 18 expressions). Les expressions du lexique généraliste $LEXG$ (21 expressions) et des deux lexiques détaillés $LEXA$ sur la sécurité alimentaire (87 expressions) et $LEXC$ sur les crises (86 expressions) sont issues du site web *Cultivoo*, davantage étayé dans notre contexte, qui recense du vocabulaire en français de différentes thématiques spécifiques [10]. Les expressions sont des noms communs du champ lexical de la sécurité alimentaire (e.g., "agriculture", "élevage", "céréales", "malnutrition") et des crises (e.g., "malnutrition", "conflit", "catastrophe"). Nous pouvons noter que certaines expressions, comme "malnutrition", appartiennent conjointement aux champs lexicaux de la sécurité alimentaire et des crises. Puis, de manière à inclure des expressions directement liées à ces thématiques dans le contexte socio-économique du Burkina Faso, les lexiques ont été enrichis avec l'aide de deux expertes de la sécurité alimentaire en Afrique de l'Ouest et au Burkina Faso. Les lexiques ont été complétés par démarche itérative, c'est-à-dire par phases successives de propositions d'ajout de nouveaux termes par l'un des experts, suivies de phases de discussions sur la pertinence des nouveaux termes de manière à converger vers des lexiques définitifs. Par exemple, les termes "mil" et "sorgho", qui désignent des céréales couramment consommées dans le pays ont été ajoutés dans le lexique $LEXA$, de même que les termes "criquet" et "pèlerin" qui constituent un véritable fléau pour la sécurité alimentaire en Afrique de l'Ouest. Ces lexiques sont accessibles en accès libre sur Dataverse (Deléglise et al., 2021b).

3.3.2 Méthodes

Dans cette section, nous exposons notre approche visant à identifier des proxies textuels de la sécurité alimentaire capables de fournir des informations à différentes

10. http ://cultivoo.fr/index.php/developpement-durable/agriculture/2590-vocabulaire-sur-la-securite-alimentaire

échelles spatio-temporelles et dimensions thématiques, à partir d'articles de journaux. Nous présentons dans un premier temps les outils issus de la fouille de textes utilisés, puis nous décrivons la méthodologie proposée et mise en place avec l'aide de ces outils pour obtenir et visualiser des proxies de la sécurité alimentaire renseignant sur la situation alimentaire régionale et annuelle du pays avec un angle explicatif. Enfin, nous détaillons les prétraitements effectués sur les articles ainsi que les procédures de validation des seuils associés aux outils de fouille de texte proposés.

3.3.2.1 Outils

Plusieurs outils de fouille de textes sont utilisés dans ce chapitre, ceux-ci ont été choisis, en nous référant aux méthodes présentées dans la section 3.2, car ils permettent de mettre en évidence des informations distinctes et complémentaires sur la sécurité alimentaire.

— **Word2vec**

Word2vec (w2v) (Mikolov et al., 2013) désigne une famille de modèles de traitement automatique du langage permettant le plongement lexical, c'est-à-dire la transformation de termes et de textes en vecteurs. W2v est basé sur des réseaux neuronaux à deux couches et a pour but l'apprentissage de représentations vectorielles de termes dans des textes, de sorte que les termes qui partagent des contextes similaires (i.e., qui sont souvent entourés des mêmes termes) soient représentés par des vecteurs numériques proches. Dans notre étude, une architecture CBOW (continuous bag of words) est utilisée (préférée à l'architecture Skip-gram, qui nécessite davantage de temps d'exécution tout en offrant des performances parfois moins satisfaisantes pour le traitement d'articles de journaux (Jang et al., 2019)). Le CBOW vise à prédire l'apparition d'un terme en utilisant comme proxies les termes qui lui sont proches dans le texte. L'apprentissage du modèle se fait sur un vaste corpus d'entraînement (dans notre étude, un corpus d'articles Wikipedia en français) en parcourant chaque terme ainsi que ses voisins et permettant d'obtenir en sortie un ensemble de vecteurs de caractéristiques qui représentent chaque terme du texte. Ces vecteurs contiennent des informations sur la sémantique des termes, et peuvent être attribués aux termes constituant de nouveaux textes.

— **Polarité d'un terme**

La polarité d'un terme est un critère qui indique son caractère positif, négatif ou neutre (Szabolcsi, 2004). Dans notre contexte, la polarité moyenne de textes traitant de sécurité alimentaire peut nous donner une information pertinente sur leur caractère inquiétant, voire alarmant. Dans ce travail qui intègre des articles de journaux, l'utilisation de la polarité "brute" ne permet pas de différencier les articles selon leur aspect positif ou négatif. Notre hypothèse est que les journaux, se voulant objectifs, ont tendance à avoir un contenu très neutre, ce qui entraîne des polarités souvent proches de zéro, i.e., la majorité des articles sont considérés comme neutres. La variation de l'aspect positif et négatif des articles y est donc difficilement détectable. Cette tendance des articles de presse à une neutralité apparente, qui peut atténuer des informations inquiétantes ou graves, a été soulignée dans la littérature scientifique (Ghazal-Aswad, 2019). C'est pourquoi nous avons choisi de nous concentrer sur l'aspect négatif des articles, que l'on suppose le plus pertinent pour appréhender les crises alimentaires. Cette approche axée sur la négativité a déjà été adoptée dans d'autres études traitant de l'analyse de sentiments dans un contexte de crise (Subramaniyaswamy et al., 2020). Pour évaluer la négativité d'un terme, nous utilisons la version française du modèle d'analyse de sentiments VADER (Valence Aware Dictionary and Sentiment Reasoner) implémenté par le package python vaderSentiment-fr [11]. Ce modèle est basé sur un lexique constitué de 7500 termes classés comme étant positifs ou négatifs ainsi que sur des règles contextuelles pouvant modifier la valence des termes (e.g., l'usage de négation, de ponctuation, de majuscule, d'adverbe). Ce modèle a été choisi car il possède un bon compromis entre sa simplicité d'implémentation et de temps d'exécution et ses performances de classification, faisant mieux que de nombreuses méthodes existantes, dont certaines basées sur l'utilisation d'apprentissage automatique (Gilbert, 2014).

— **tf-idf**

Pour évaluer la discriminance des termes d'un article, nous recourons au concept de tf-idf (term frequency-inverse document frequency) (Salton and Buckley, 1988), qui est une méthode issue du domaine de la recherche d'information permettant de pondérer l'importance d'un terme contenu dans un texte relativement à un

11. https ://pypi.org/project/vaderSentiment-fr/

corpus, i.e., qui mesure à quel point un terme est caractéristique d'un texte en évaluant sa pertinence et sa singularité. Son principe est basé sur une formule dans laquelle deux valeurs : le tf (Term Frequency) et l'idf (Inverse Document Frequency) sont multipliées. Le tf correspond à la fréquence d'un terme dans un texte (i.e., au nombre d'occurrences du terme dans le texte considéré rapporté au nombre total de termes du texte) et augmente donc lorsque le terme est fréquent dans le texte. L'idf mesure l'importance d'un terme non pas en fonction de sa fréquence dans un texte particulier, mais en fonction de sa distribution dans l'ensemble des textes étudiés. Un terme qui apparaît très fréquemment dans quelques textes seulement possède un idf élevé dans ces textes, à l'inverse, un terme qui apparaît dans presque tous les textes d'un corpus possède un idf faible. Cela permet de donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. La formule du tf-idf est donnée dans l'Équation 3.1. Cette méthode trouve sa pertinence dans notre contexte car elle nous permet de mettre en avant des évènements affectant la sécurité alimentaire de manière ponctuelle (e.g., invasion de criquets, incendie) qui auraient été occultés par d'autres termes plus génériques (e.g., "pauvreté", "catastrophe") si la fréquence des termes avait été considérée seule. D'autres méthodes d'extraction de termes spécifiques plus récentes et sophistiquées existent, par exemple les méthodes dérivées du tf-idf comme Okapi BM25 (Whissell and Clarke, 2011) qui apporte une meilleure efficacité sur des corpus comportant des tailles de documents très hétérogènes, ce qui n'est pas le cas du corpus utilisé dans ce chapitre (voir Tableau 3.1). Mentionnons également les approches basées sur l'utilisation de graphes (SingleRank (Wan and Xiao, 2008), TopicRank (Bougouin et al., 2013), Kcore (Rousseau and Vazirgiannis, 2015)). Mais celles-ci n'offrent pas nécessairement de meilleures performances. Ramiandrisoa and Mothe (2016) ont appliqué ces méthodes à base de graphe ainsi que le tf-idf sur un corpus d'articles scientifiques associés à des mots-clés par leurs auteurs, leur résultats ont montré que les termes sélectionnés par tf-idf étaient les plus proches des mots-clés associés.

$$TF - IDF(term, art) = TF(term, art) \times log(\frac{N}{N_{term}}) \tag{3.1}$$

Où *TF-IDF* est le tf-idf de l'expression *term* dans l'article *art*, *TF* est la fréquence de l'expression *term* dans l'article *art*, N est le nombre total d'articles dans le corpus et N_{term} est le nombre d'articles du corpus qui contiennent l'expression *term*.

3.3.2.2 Méthodologie

L'objectif est de réaliser une analyse spatio-temporelle de la sécurité alimentaire basée sur une terminologie de ce domaine, liée aux proxies textuels que nous définissons. Dans ce cadre, nous proposons un processus dédié qui combine différentes approches d'analyse textuelle. C'est à cette fin que nous présentons dans cette section la méthodologie déployée dans le but d'obtenir un contexte explicatif spatial et temporel de la situation alimentaire burkinabè à partir du corpus de journaux étudié. La Figure 3.4 en résume le plan d'analyse.

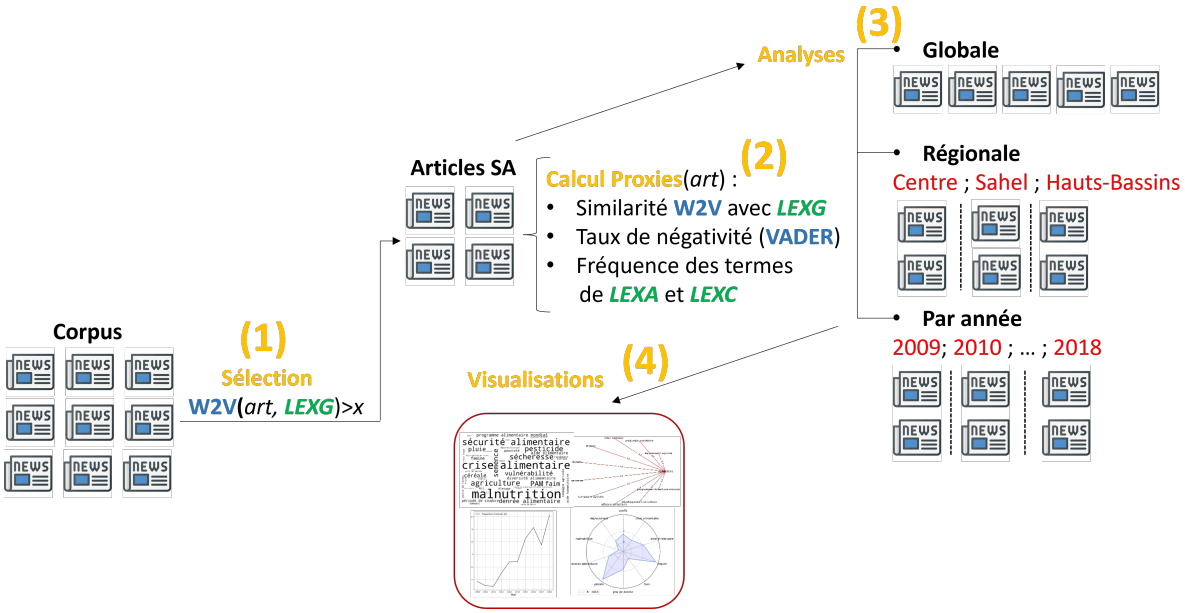


FIGURE 3.4 – Schéma général du plan d'analyse. Les principales étapes sont numérotées : (1) sélection des articles d'intérêt ; (2) calcul des proxies textuels sur ces articles ; (3) analyse globale, spatiale et temporelle ; et (4) Visualisations.

Nous commençons par détailler les méthodes d'extraction d'informations thématiques sur la sécurité alimentaire appliquées au corpus, permettant d'obtenir plusieurs "proxies textuels" de la sécurité alimentaire (i.e., des descripteurs extraits de textes, offrant une information indirecte mais pertinente sur la situation alimentaire), basés sur les outils présentés ci-dessus.

Premièrement, nous présentons l'étape **(1)** de sélection des articles pertinents. Pour cela, nous calculons par w2v la similarité sémantique entre chaque article et le lexique

généraliste $LEXG$, servant de base pour identifier les articles de thème "sécurité alimentaire". Le principe étant de considérer un article comme traitant de sécurité alimentaire si sa similarité sémantique avec $LEXG$ par w2v est supérieure à un seuil x (choisi et validé dans la section 3.3.2.3). Cela a pour but de détecter les articles d'intérêt pour y centrer les analyses.

Deuxièmement, nous établissons dans l'étape **(2)** les proxies textuels de la sécurité alimentaire sur les articles sélectionnés. À cette fin, nous effectuons les opérations suivantes :

— Nous conservons pour les articles sélectionnés leur score w2v calculé lors de l'étape **(1)** qui quantifie leur degré de connexion avec le thème de la sécurité alimentaire, ce qui constitue un proxy de ce domaine (voir section 3.2.1).

— Nous calculons le taux de négativité des articles que nous proposons comme proxy, i.e., la fréquence de termes négatifs dans chaque article (Équation 3.2) afin d'obtenir une information sur le caractère préoccupant du message des articles.

$$Neg(art) = \frac{nb_{terms\_neg}(art)}{nb_{terms}(art)} \tag{3.2}$$

Où *Neg* est le taux de négativité d'un article *art*, nb_{terms_neg} et nb_{terms} représentent respectivement le nombre de termes négatifs et le nombre de termes d'un article *art*, basé sur la version française du modèle VADER (Valence Aware Dictionary and Sentiment Reasoner).

L'hypothèse (appuyée dans la section 3.2.1) suppose que les articles publiés lors de périodes et dans des zones en proie à l'insécurité alimentaire sont associés à des valences davantage négatives que dans un contexte de suffisance alimentaire. Un article est considéré comme négatif si son taux de négativité est supérieur à 0.1 (méthodologie de validation du seuil détaillée dans la section 3.3.2.3).

— Nous étudions le vocabulaire le plus employé dans les articles traitant de sécurité alimentaire dans le but de détecter si le vocabulaire adopté est concordant avec les tendances et crises qui ont affecté la sécurité alimentaire dans le pays et ainsi disposer d'un regard plus explicatif sur les données. Pour cela, nous calculons pour chaque article la fréquence de 119 expressions issues des deux lexiques détaillés $LEXA$ et $LEXC$.

Troisièmement, nous décrivons l'étape **(3)** d'analyse globale, régionale et annuelle des proxies définis en étape **(2)**. Pour prendre en compte l'aspect spatio-temporel de la sécurité alimentaire, les proxies présentés sont par la suite agrégés à différentes granularités pour effectuer des analyses ciblées aux niveaux global, régional et annuel et ainsi pouvoir visualiser les tendances et les crises alimentaires qui ont touché le pays au cours de la dernière décennie. Les proxies sont agrégés à trois niveaux :

— **Niveau global :** ce niveau d'analyse permet d'avoir une vision générale des caractéristiques de la situation alimentaire du pays entre 2009 et 2018, et peut servir de comparaison pour les analyses ciblées (régionale et annuelle). Les proportions d'articles traitant de sécurité alimentaire et d'articles négatifs sont calculées sur tout le corpus. Nous considérons la fréquence moyenne d'apparition de chaque terme des lexiques détaillés $LEXA$ et $LEXC$ dans tous les articles du corpus.

— **Niveau régional :** ce niveau d'analyse vise à donner une représentation de la situation alimentaire et de ses caractéristiques au niveau régional. Nous illustrons nos analyses avec trois régions : les régions Centre, Hauts-Bassins et Sahel. Ces trois régions ont été choisies car elles figurent parmi les plus citées dans les articles du corpus et qu'elles sont associées à des situations sanitaires distinctes (WFP, 2014a ; OCHA, 2015 ; Zida and Kambou, 2014). Notre démarche consiste à considérer un article comme associé à une région si une localité de la région est mentionnée en début d'article (i.e., dans le titre ou dans la première phrase de l'article), en supposant qu'un article qui se consacre à une localité la mentionnera dans cette zone du texte. Un article peut être simultanément associé à 2 ou 3 régions distinctes si des localités de chaque région y sont mentionnées. Nous nous appuyons sur les travaux de Lopez et al. (2010) qui estiment que le titre plus la première phrase d'un texte suffisent pour obtenir les informations essentielles contenues dans le texte. L'association d'un article à une région est réalisée avec la méthode suivante : 1) sélection du titre et de la première phrase de l'article ; 2) détection de la présence de localités appartenant à la région considérée par mise en correspondance du début de l'article avec un lexique géographique constitué des provinces, communes et villages de la région, issue de l'EPA (Permanent Agricultural Survey, 2015) (lexique géographique accessible sur Github [12]). Il n'existe pas au Burkina Faso de source officielle des localités au niveau sub-

12. https ://github.com/pipapou/Burkina_localites

communal, la liste de localités burkinabès issue de l'EPA est la plus complète à notre connaissance (e.g., plus complète que Geonames). Pour chaque début d'article traité, la localité doit être capitalisée (pour éviter un faux positif causé par la présence d'un nom commun de même orthographe) et ne peut être encadrée que de ponctuation ou d'espaces (pour éviter un faux positif causé par l'existence d'un terme qui contient le nom de la localité) ; 3) les articles qui possèdent au moins une localité de la région considérée dans le titre et/ou la première phrase sont considérés comme associés à la région. Une fois les articles associés, quand cela est possible, à une ou plusieurs régions, les proxies de la sécurité alimentaire sont agrégées sur les ensembles d'articles associés à chaque région considérée. Les proportions d'articles traitant de sécurité alimentaire et d'articles négatifs sont calculées pour chacune des 3 régions. Pour extraire le vocabulaire régional caractéristique, nous calculons sur les articles de chaque région considérée le tf-idf de chaque terme des lexiques $LEXA$ et $LEXC$. Le tf-idf permet dans notre contexte de mettre en avant les expressions de sécurité alimentaire et de crises fréquentes dans les articles liés à une certaine région, et qui sont par ailleurs spécifiquement employées dans les articles de la région (i.e., davantage que pour les autres articles).

— **Niveau annuel :** ce niveau d'analyse permet d'obtenir les caractéristiques annuelles de la situation alimentaire burkinabè et de suivre son évolution de 2009 à 2018. Chaque article est associé à son année de publication, dont la valeur est extraite dans les métadonnées liées à l'article. Les proportions d'articles traitant de sécurité alimentaire et d'articles négatifs sont calculées pour chaque année. Pour extraire le vocabulaire annuel caractéristique, nous calculons sur les articles de chaque année considérée le tf-idf de chaque terme des lexiques $LEXA$ et $LEXC$, permettant de mettre en avant les expressions de sécurité alimentaire et de crises spécifiques des articles liés à chaque année. Nous conceptualisons également une nouvelle méthode pour pondérer l'importance d'un terme dans un corpus de textes associés à une temporalité. Cette proposition appelée TIR (Tf-Idf Ratio), est fondée sur le concept de tf-idf et se révèle mieux adaptée dans notre contexte, en permettant de davantage distinguer que le tf-idf les expressions rares et spécifiques d'une année. Plus exactement, nous calculons dans un premier temps pour chaque expression des lexiques $LEXA$ et $LEXC$ le tf-idf de l'expression en moyenne sur les articles de l'année (Équation 3.3), puis dans un

second temps nous calculons le ratio de ce tf-idf par le tf-idf de l'expression en moyenne sur les articles des autres années (ratio TIR ; Équation 3.4).

$$TF-IDF_{moy}(term, A_y) = \frac{\sum_{art \in A_y} TF-IDF(term, art)}{N_y} \tag{3.3}$$

Où $TF-IDF_{moy}$ est le tf-idf moyen du terme "term" sur les articles "art" appartenant à l'ensemble A_y des articles de l'année y, nous notons N_y le cardinal de cet ensemble.

$$TIR(term, A_y) = \frac{TF-IDF_{moy}(term, A_y)}{TF-IDF_{moy}(term, A_z)} \tag{3.4}$$

Où TIR est le ratio du tf-idf du terme "term" en moyenne sur les articles appartenant à l'ensemble A_y des articles de l'année y, par le tf-idf du terme "term" en moyenne sur les articles appartenant à l'ensemble A_z des articles des années différentes de l'année y.

Différentes représentations graphiques sont exploitées dans l'étape **(4)** pour visualiser de manière adaptée les informations obtenues à partir de chaque niveau d'analyse. Des nuages de mots sont utilisés pour faire une synthèse graphique des termes les plus fréquents au niveau global, et régional. Concernant l'analyse annuelle, des graphiques de séries temporelles sont utilisés pour modéliser l'évolution annuelle des proportions d'articles traitant de sécurité alimentaire, d'articles négatifs, et du tf-idf moyen de 5 expressions des lexiques détaillés $LEXA$ et $LEXC$ entre 2009 et 2018. Des graphiques radars, qui ont été exploités dans la littérature scientifique pour la visualisation de données spatio-temporelles (Forlines and Wittenburg, 2010 ; Reski et al., 2020), sont également utilisés pour représenter l'évolution annuelle des expressions les plus caractéristiques (i.e., possédant les plus grands ratios TIR pour chaque année considérée) et permettent de représenter ces informations complexes de manière synthétique et cohérente. Enfin, nous proposons comme perspective d'analyse une visualisation des liens entre les différents éléments de vocabulaire utilisés dans les articles sous forme de graphes de co-occurrences.

Dans la section 3.4, les résultats et visualisations effectuées à chaque niveau d'analyse sont davantage détaillées, puis les différentes sorties y sont comparées, interprétées et discutées.

3.3.2.3 Paramètres

Les traitements statistiques et de fouille de texte ont été effectués avec Python 3.7, en utilisant une machine possédant un processeur Intel i7-8650U CPU @ 2.11 GHz et

32 Go de mémoire vive. Le code permettant d'obtenir toutes les statistiques et sorties graphiques présentées est disponible sur GitHub [13]

Prétraitement des articles

Les articles et informations associées (titre et date) ont été recueillis par crawling (ensemble de techniques qui consistent à extraire de manière structurée des données du code source de pages web) sur les sites web LeFaso.net et Burkina24.com et convertis en format csv. Les articles publiés entre 2009 et 2018 ont été filtrés, puis lemmatisés avec l'outil de lemmatisation du package python *Spacy* basé sur des règles [14].

Validation du seuil x à partir duquel un article est considéré comme traitant de sécurité alimentaire

Nous présentons ici la démarche adoptée pour évaluer le potentiel de w2v à détecter les articles d'intérêt (i.e., de thème "sécurité alimentaire"), puis pour fixer un seuil de séparation optimal entre les articles traitant de la sécurité alimentaire et le reste du corpus.

A) Évaluation de la pertinence de w2v

Premièrement, nous voulons évaluer la capacité de w2v à détecter les articles qui traitent de sécurité alimentaire en testant si les articles possédant les similarités w2v les plus élevées sont les plus liés au thème de la sécurité alimentaire, et si le lien des articles avec la sécurité alimentaire diminue quand leur similarité w2v diminue. Pour cela, les articles sont triés par similarité w2v décroissante, nous sélectionnons un échantillon de ces articles : les articles numéro 1, 2, 3, 4, 5, 100, 200, 300, 400, 500, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000 dans le but d'obtenir un panel de 20 articles constitué aussi bien des articles possédant les plus hauts scores w2v que d'articles avec un faible score w2v, censés ne pas être liés à la sécurité alimentaire. Le contenu de ces 20 articles est disponible sur Github [15].

Chaque article de ce panel est ensuite manuellement annoté comme traitant ou non de la sécurité alimentaire afin de servir de comparaison aux scores attribués par w2v et de

13. https ://github.com/pipapou/analyse_corpus
14. https ://spacy.io/api/lemmatizer
15. https ://github.com/pipapou/20_articles_BF

pouvoir apprécier la qualité de ces scores. La concordance entre les choix des annotateurs est donc cruciale et est également évaluée dans cette partie. Un guide d'annotation (dont la méthodologie est détaillée ci-dessous ; disponible en version pdf sur Github[15]) a été conçu pour permettre aux intervenants d'annoter de manière standardisée, constante et en accord les uns avec les autres.

Une phase préliminaire est tout d'abord consacrée à la création des critères des classes d'annotations et à accorder les annotateurs avec ces critères par l'étude en groupe d'un petit échantillon d'articles.

Puis, les 20 articles du panel sélectionné sont anonymisés (i.e., séparés de leur score w2v associé) et annotés manuellement par trois experts de la sécurité alimentaire. Les classes à annoter sont les suivantes :

— "0" : l'article n'aborde ni la sécurité alimentaire, ni aucun thème lié ;
— "1" : l'article aborde un thème lié à la sécurité alimentaire, i.e., lié à des évènements qui peuvent indirectement améliorer ou empirer la situation alimentaire dans la population (e.g., climat, nuée de criquets, pauvreté, licenciement massif de travailleurs, nouvelle méthode d'agriculture) ;
— "2" : l'article aborde directement l'un des 4 piliers de la sécurité alimentaire (disponibilité de nourriture ; accès des populations aux denrées ; évènement qui dérègle directement la stabilité de l'accès aux denrées (avec mention dans l'article des conséquences alimentaires de l'événement) ; utilisation bonne ou mauvaise des denrées au niveau sanitaire et nutritionnel).

Pour une plus grande robustesse des annotations, la classe majoritaire est choisie. Plus précisément, pour chaque article la classe la plus choisie parmi les trois annotateurs lui est attribuée, dans le cas d'un article pour lequel chaque annotateur choisit une classe différente (i.e., "0", "1", et "2"), c'est la médiane, i.e., la classe "1" qui est attribuée.

Dans le Tableau 3.2, nous constatons comme attendu une tendance nette des articles possédant les plus grandes similarités w2v à être classés "1" ou "2" et des articles aux similarités w2v plus faibles à être classés "0". Parmi les 5 articles possédant les scores w2v les plus élevés du panel, 4 ont une classe majoritaire de "2". Parmi les 10 articles possédant les scores w2v les plus élevés du panel, 8 ont une classe majoritaire d'au moins "1". À l'inverse, 9 des 10 articles possédant les scores w2v les plus bas ont une classe majoritaire de "0". Nous en déduisons que l'utilisation de w2v pour détecter et

classer les articles qui traitent de sécurité alimentaire est pertinente dans notre contexte. De plus, nous pouvons noter un accord significatif entre les choix des annotateurs : 13 articles sur 20 (soit 65%) ont été identiquement labélisés par les 3 annotateurs, et les 20 articles (soit 100%) ont été identiquement labélisés par 2 annotateurs parmi 3. Enfin, le Kappa de Fleiss (Fleiss and Cohen, 1973) est calculé, ce coefficient est une mesure statistique de la concordance de labélisation entre plusieurs annotateurs (" < 0" si la concordance est inexistante, " > 0" si la concordance est au moins faible et " $= 1$" si la concordance est parfaite). Le Kappa de Fleiss est dans notre cas égal à 0.601, ce qui signifie une concordance significative entre les choix des annotateurs. Malgré la faible taille de l'échantillon, la valeur p associée au Kappa est inférieure à 0.01, ce qui nous permet de conclure à sa significativité. Nous en concluons que l'annotation manuelle effectuée par les 3 annotateurs est pertinente pour servir de référence aux scores attribués par w2v et permet d'apprécier la qualité de ces scores.

Rang w2v	Classes Experts 1, 2 et 3	Classe majoritaire
1	2 | 2 | 1	2
2	2 | 2 | 2	2
3	2 | 2 | 2	2
4	0 | 0 | 0	0
5	2 | 2 | 2	2
100	1 | 0 | 1	1
200	0 | 1 | 1	1
300	2 | 1 | 1	1
400	1 | 0 | 1	1
500	0 | 0 | 0	0
1000	0 | 0 | 0	0
1500	0 | 0 | 0	0
2000	0 | 0 | 0	0
2500	0 | 0 | 1	0
3000	0 | 1 | 0	0
4000	0 | 0 | 0	0
5000	0 | 0 | 0	0
6000	0 | 0 | 0	0
7000	0 | 0 | 0	0
8000	1 | 1 | 1	1

Tableau 3.2 – Comparaison des classes assignées par trois experts à 20 articles du corpus et classes majoritaires.

B) Identification des seuils w2v potentiels

Dans un second temps, nous voulons définir un seuil w2v x tel que les articles de similarité w2v supérieure à x (resp. inférieure à x) soient automatiquement classés comme

traitant de sécurité alimentaire (resp. ne traitant pas de sécurité alimentaire). Si ce seuil est trop bas, une proportion importante des articles retenus ne seront pas pertinents. Si ce seuil est à l'inverse trop haut, trop peu d'articles seront retenus pour pouvoir y faire des analyses pertinentes. Il faut trouver une démarche équilibrée sur le niveau de contraintes à fixer pour considérer un article comme traitant de sécurité alimentaire.

Pour prendre en compte cela, nous construisons tout d'abord un intervalle à l'intérieur duquel le seuil w2v final sera choisi. La borne inférieure est fixée de façon empirique, en contrôlant (par lecture) pour plusieurs valeurs "seuils" une trentaine d'articles dont le score w2v est tout juste supérieur à la valeur, la valeur est choisie comme borne inférieure si au moins un quart des articles contrôlés sont en lien avec la sécurité alimentaire. Nous commençons avec la valeur 0.1, puis avançons par pas de 0.05. La première valeur testée pour laquelle nous possédons au moins un quart d'articles liés à la sécurité alimentaire est 0.3. Dans le but d'obtenir un intervalle de seuils potentiels aussi grand que possible nous testons également le seuil 0.28, situé entre 0.3 et 0.25 qui avait été testé comme non suffisant. Pour ce seuil de 0.28, plus d'un quart des articles sont pertinents, nous validons donc cette valeur comme borne inférieure. Pour la sélection de la borne supérieure, dont le seuil associé doit permettre de sélectionner assez d'articles pour les analyses, nous devons examiner attentivement le nombre d'articles sélectionnés par le choix du seuil qui sont associés à chaque région étudiée (i.e., Centre, Hauts-Bassins, Sahel). En effet, contrairement aux années qui peuvent être associées à chaque article (car l'information sur l'année de publication est présente dans les métadonnées), une minorité d'articles du corpus sont associés à au moins l'une des 3 régions étudiées (25% le sont). Nous considérons une borne supérieure comme acceptable tant qu'elle permet d'associer au moins 50 articles à chaque région. Ce nombre de 50 est fixé de manière empirique, il n'existe aucune règle de choix d'échantillon minimal pour les analyses (complexes pour certaines) que nous allons y appliquer (e.g., détection de vocabulaire spécifique dans les articles, calcul de leur tf-idf, ratio TIR). Nous contrôlons pour plusieurs valeurs que cet effectif minimal est vérifié pour chaque région, en commençant par 0.28 (la borne inférieure) et en avançant par pas de 0.02. Le Tableau 3.3 détaille le nombre d'articles sélectionnés comme traitant de sécurité alimentaire sur tout le corpus et pour chacune des 3 régions étudiées en fonction de la valeur choisie comme seuil de détection (0.28, 0.30, 0.32, 0.34, 0.36 et 0.38). La dernière valeur testée pour laquelle nous possédons des effectifs liés à chaque région de taille au moins 50 est 0.36, nous choisissons donc cette valeur comme borne supérieure.

Seuil	Corpus	Centre	Hauts Bassins	Sahel
0.28	6174	1098	320	192
0.30	4694	813	242	153
0.32	3451	574	124	166
0.34	2472	389	114	93
0.36	1675	252	74	66
0.38	1068	152	46	44

Tableau 3.3 – Illustration du nombre d'articles sélectionnés comme traitant de sécurité alimentaire sur tout le corpus et pour chacune des 3 régions étudiées (Centre, Hauts-Bassins, Sahel) en fonction de la valeur choisie comme seuil de détection. Les valeurs en rose correspondent aux effectifs inférieurs à 50.

C) Choix du seuil w2v

Pour valider le seuil x, nous proposons de fixer 3 seuils dans l'intervalle [0.28,0.36] que nous avons construit et de vérifier si les articles correspondant aux zones critiques de ces seuils (articles de similarité w2v tout juste inférieure ou supérieure au seuil) sont pertinents ou non pour la sécurité alimentaire. Pour cela, nous fixons comme seuils x à évaluer : $x = 0.28\,;\, 0.32\,;\, 0.36$, homogènement fixés entre les deux valeurs "limites" de l'intervalle.

Pour chacun des trois seuils x, nous sélectionnons les 12 articles de similarité w2v supérieure à x les plus proches (considérés comme traitant de sécurité alimentaire) et les 12 articles de similarité w2v inférieure à x les plus proches (considérés comme ne traitant pas de sécurité alimentaire). Ce qui donne un total de 24 articles dont 50% sont au-dessus du seuil x et sont donc considérés comme traitant de la sécurité alimentaire. Notons que ce nombre de 24 articles évalués par seuil est insuffisant pour comparer la qualité des seuils de manière statistiquement significative, nous avons effectué un test de χ^2 d'égalité des proportions d'erreurs associées aux 3 seuils qui s'est avéré non significatif (valeur p > 0.05). Ce travail d'annotation doit être étendu (d'au moins une centaine d'articles par seuil), afin d'obtenir des échantillons statistiquement représentatifs et comparables. Cet investissement supplémentaire de la part des experts impliqués sera judicieux pour consolider le choix du seuil détaillé ci-dessous. Les statistiques présentées par la suite

pour valider le seuil x restent donc des tendances et devraient être confirmées par d'autres études plus approfondies.

Les 24 articles sélectionnés (pour chacun des trois seuils x) sont ensuite manuellement annotés avec les mêmes méthodologies, guide d'annotation et classes que présenté ci-dessus.

Nous choisissons le seuil x pour lequel les 24 articles associés ont été le plus souvent classés par seuillage w2v de manière concordante par rapport à l'annotation manuelle. Nous estimons comme classés de manière concordante les articles de similarité w2v inférieure (resp. supérieure) à x et manuellement annotés comme "0" (resp. "1" ou "2") et considérons la F-mesure correspondante comme critère de choix de x.

Le Tableau 3.4 représente pour chaque seuil x la distribution en pourcentage des classes manuellement annotées ("0", "1" et "2") sur les 24 articles correspondants, ainsi que les taux d'erreur, rappel, précision et F-mesure (définis dans l'Équation 3.5) du seuillage w2v par rapport aux classes manuellement annotées. Nous observons que le fait d'augmenter le seuil x ne fait pas diminuer significativement le pourcentage d'articles non pertinents pour la sécurité alimentaire (manuellement annotés comme "0") aux contours du seuil, mais fait en revanche augmenter le pourcentage d'articles très pertinents (classés comme "2"), de 0% (pour x=0.28) à 12.5% (pour x=0.36). Enfin, nous constatons que la F-mesure est maximisée pour x=0.36, nous choisissons donc ce seuil. Ce seuil w2v de 0.36 est, rappelons-le, le seuil maximal que nous nous sommes fixés. En augmentant encore ce seuil, nous pourrions accroître la proportion d'articles pertinents (d'autant plus que pour x=0.36 la précision n'a pas encore amorcé sa diminution), mais le nombre d'articles sélectionnés serait trop faible pour effectuer des analyses robustes (Tableau 3.3).

$$R = \frac{N_{SA,w2v}}{N_{SA}} \quad ; \quad P = \frac{N_{SA,w2v}}{N_{w2v}} \quad ; \quad F = 2 \times \frac{R \times P}{R + P} \tag{3.5}$$

Où R, P et F représentent respectivement le rappel, la précision et la F-mesure. N_{SA} est le nombre d'articles annotés comme de thème "sécurité alimentaire", N_{w2v} est le nombre d'articles classés par w2v comme de thème "sécurité alimentaire" et $N_{SA,w2v}$ est le nombre d'articles annotés et classés par w2v comme de thème "sécurité alimentaire".

	x=0.28	**x=0.32**	**x=0.36**
Proportion de ”0”	62.5	75	66.7
Proportion de ”1”	37.5	20.8	20.8
Proportion de ”2”	0	4.2	12.5
Taux d’erreur	62.5	50	41.7
Rappel	0.33	0.5	0.63
Précision	0.25	0.25	0.42
F-mesure	0.28	0.33	0.5

Tableau 3.4 – Comparaison des distributions (en %) des classes manuellement annotées (”0”, ”1” et ”2”) sur 24 articles, ainsi que les taux d’erreur (en %), rappel, précision et F-mesure associés aux seuils w2v par rapport aux classes manuellement annotées, pour chaque seuil $x = 0.28\,;\, 0.32\,;\, 0.36$.

L’annotation manuelle des textes est un exercice qui exige un temps et une concentration considérables, avec des directives qui doivent être simples et sans ambiguïté mais souvent complexes à appliquer. Compte tenu de la contrainte de temps imposée dans cette thèse, la faible quantité de seuils et de nombres d’articles testés ont ici montré leurs limites. Bien que les résultats obtenus ainsi que le seuil choisi fassent sens, des travaux complémentaires devraient se consacrer aux tests d’un plus grand nombre de seuils avec davantage d’articles associés, permettant par exemple de fixer plus finement le seuil par l’utilisation de critères éprouvés comme les courbes ROC (Receiver Operating Characteristic), qui est un outil reconnu d’aide au choix de seuils (Delacour and Servonnet, 2005).

Nos expérimentations ont permis d’identifier un seuil de 0.36 pour déterminer des textes en lien avec le thème de la sécurité alimentaire véhiculé dans les articles. Ce seuil est appliqué à la phase **(1)** du processus décrit dans la section 3.3.2 et illustré sur la Figure 3.4. Dans un second temps, nous évaluons le seuil de négativité à appliquer pour la phase **(2)**.

Validation du seuil à partir duquel un article est considéré comme négatif

La pertinence du modèle VADER pour la mesure de négativité dans des textes a été mise en évidence dans la littérature scientifique (Gilbert, 2014). Nous souhaitons évaluer

dans quelle mesure des seuils spécifiques associés au modèle VADER permettent de bien prédire l'aspect négatif dans le cas d'articles de journaux traitant de sécurité alimentaire. Ces seuils choisis seront alors appliqués dans le processus proposé dans la section 3.3.2. Nous présentons ici la méthodologie employée pour fixer un seuil de négativité optimal dans notre contexte, à partir duquel un article est considéré comme négatif. Le choix de ce seuil est soumis au même dilemme entre la pertinence des articles sélectionnés d'une part et la quantité d'articles disponibles pour l'analyse d'autre part. Le seuil a été choisi avec une méthodologie analogue à celle utilisée pour le choix du seuil x associé à w2v.

Nous trions dans un premier temps les articles par leur taux de négativité calculé avec le modèle VADER, en nous restreignant aux articles de scores w2v supérieurs à 0.36 (i.e., traitant de sécurité alimentaire). Puis, nous construisons un intervalle à l'intérieur duquel le seuil de négativité final sera choisi. La borne inférieure est fixée de façon empirique, en contrôlant (par lecture) pour plusieurs valeurs "seuils" une trentaine d'articles dont le taux de négativité est tout juste supérieur à la valeur, la valeur est choisie comme borne inférieure si au moins un quart des articles contrôlés sont considérés comme négatifs. Nous commençons avec la valeur 0, puis avançons par pas de 0.025. La première valeur testée pour laquelle nous possédons au moins un quart d'articles négatifs est 0.05, nous validons donc cette valeur comme borne inférieure. Pour la sélection de la borne supérieure, ce sont encore les nombres correspondants d'articles associés à chaque région étudiée qui doivent retenir notre attention, pour les mêmes raisons qu'invoquées précédemment. Nous considérons empiriquement une borne supérieure comme acceptable tant qu'au moins 100 articles négatifs sont associés au corpus, et que chaque région est associée à au moins un article négatif. Pour le choix de ce seuil, la contrainte sur le nombre d'articles retenus est moins forte que pour le seuil x de w2v car l'analyse permise par ce seuil se limite à des calculs de taux de négativité moyens (e.g., par régions ou années). Nous contrôlons pour plusieurs valeurs que ces contraintes minimales sont vérifiées sur le corpus et pour chaque région, en commençant par 0.05 (la borne inférieure) et en avançant par pas de 0.025. Le Tableau 3.5 détaille le nombre d'articles liés à la sécurité alimentaire sélectionnés comme étant négatifs sur tout le corpus et pour chacune des 3 régions étudiées en fonction de la valeur choisie comme seuil de détection (0.05, 0.075, 0.1 et 0.125). La dernière valeur testée pour laquelle nous possédons un effectif sur tout le corpus d'au moins 100 ainsi que des effectifs non nuls liés à chaque région est 0.1, nous choisissons donc cette valeur comme borne supérieure.

Seuil	Corpus	Centre	Hauts Bassins	Sahel
0.05	689	94	18	34
0.75	293	37	6	18
0.1	107	12	1	8
0.125	35	4	0	3

Tableau 3.5 – Illustration du nombre d'articles liés à la sécurité alimentaire sélectionnés comme étant négatifs sur tout le corpus et pour chacune des 3 régions étudiées (Centre, Hauts-Bassins, Sahel) en fonction de la valeur de seuil choisie. Les valeurs en rose correspondent aux effectifs trop réduits pour les analyses futures.

Pour valider le seuil de négativité, nous fixons 3 seuils dans l'intervalle [0.05,0.1] que nous avons construit : 0.05 ; 0.075 et 0.1, homogènement répartis entre les deux valeurs "limites" de l'intervalle. Nous examinons ensuite dans quelle mesure chaque seuil discrimine les articles pertinents en vérifiant si les articles dont le taux de négativité se situe dans la zone critique de chaque seuil (i.e., juste au-dessus) sont négatifs dans une proportion suffisamment importante.

Pour chaque seuil potentiel, les 30 articles possédant les taux de négativité les plus proches et supérieurs à chacun des seuils sont sélectionnés et l'aspect négatif (i.e., la tendance d'un article à parler de sujets graves ou préoccupants en utilisant des termes connotés négativement) de ces articles est vérifié manuellement par un expert et annoté ("0" : article non négatif, "1" : article négatif). Notre critère de choix du seuil consiste d'une part à ce que le seuil ait dans sa zone critique une proportion d'articles pertinents (i.e., annotés comme négatifs) significativement plus élevée que pour tous les autres seuils, et d'autre part à ce que le pourcentage d'articles pertinents dans sa zone critique soit d'au moins 50%. Nous n'effectuons pas ici de calcul de rappel/précision car nous nous concentrons exclusivement sur la capacité du modèle VADER et des seuils potentiels à ne pas classer comme négatifs trop d'articles qui ne le sont pas en réalité (faux positifs), ce qui pourrait biaiser les analyses effectuées sur ce groupe d'articles sélectionnés. Le nombre de 30 articles contrôlés par seuil est dans ce cas suffisant pour obtenir des résultats statistiquement significatifs, nous avons effectué un test de χ^2 d'égalité des proportions d'articles annotés comme négatifs associés aux 3 seuils qui est significatif (valeur $p < 0.05$). Les pourcentages d'articles négatifs associés à chaque seuil illustrés dans le Tableau 3.6 sont donc statistiquement différents et par conséquent comparables,

nous pouvons affirmer que le seuil de négativité de 0.1 permet de maximiser la proportion d'articles négatifs associés (67%), qui est par ailleurs supérieure à la proportion "limite" de 50% fixée au départ. Nous fixons donc le seuil à partir duquel un article est considéré comme négatif à 0.1.

Seuil	**Pourcentage d'articles négatifs**
0.05	30%
0.075	47%
0.1	67%

Tableau 3.6 – Pourcentage d'articles manuellement labélisés par un expert comme négatifs pour 3 groupes constitués des 30 articles possédant les taux de négativité les plus proches et supérieurs aux seuils de négativité 0.05, 0.075 et 0.1.

Comme pour le seuil w2v que nous avons fixé précédemment, des études plus approfondies avec davantage de seuils et d'articles testés seraient souhaitables, de façon à clarifier et renforcer nos choix. Précisons enfin que le seuil de négativité validé ici est propre à notre contexte, lié à des articles de journaux traitant de sécurité alimentaire. Ce seuil est par ailleurs particulièrement faible, nous pensons que cela est dû au type de support textuel, l'article de presse, tenu à un devoir de neutralité dans la manière de présenter les actualités. Pour d'autres supports textuels (e.g., littérature scientifique, bulletins d'ONG, messages de réseaux sociaux) et d'autres thématiques dont les schémas de pensée et les styles d'écriture peuvent être très différents, il n'y a aucune garantie que les seuils de négativité optimaux soient comparables au nôtre. C'est d'ailleurs pour cette raison que nous ne nous sommes pas tournés vers de larges corpus annotés pour régler ce seuil, car en plus d'être en nombre très réduit en langue française, les quelques corpus existants (comme ceux proposés par DEFT [16]) n'étaient pas adaptés à notre contexte.

3.4 Résultats et discussion

Dans cette section, nous présentons les informations que nous avons pu extraire du corpus de journaux afin d'obtenir un cadre explicatif temporel et spatial de la situation alimentaire au Burkina Faso. Des informations distinctes et complémentaires sur

16. https ://deft.limsi.fr/

la thématique de la sécurité alimentaire sont apportées par les proxies proposés dans la section 3.3.2. Afin de mettre en lumière l'aspect spatio-temporel de la sécurité alimentaire, les proxies considérés sont agrégés à différentes échelles spatiales et temporelles pour réaliser des analyses appropriées et ciblées. Les résultats sont décrits, visualisés avec des représentations graphiques adéquates et discutés en trois parties : l'analyse globale qui couvre l'ensemble du pays de 2009 à 2018, l'analyse régionale qui se concentre sur trois régions aux situations agroclimatique et sanitaire contrastées (Centre, Sahel et Hauts-Bassins) et l'analyse annuelle qui distingue les résultats obtenus pour chaque année entre 2009 et 2018. Soulignons que les analyses proposées sont rétrospectives, les résultats obtenus peuvent donc être comparés aux réalités géographiques et historiques à travers des bulletins et des rapports d'ONG afin de valider l'approche d'acquisition automatique d'informations sur la sécurité alimentaire proposée et mise en place dans ce chapitre.

3.4.1 Analyse globale

Dans cette section, nous souhaitons évaluer la capacité des proxies choisis dans ce chapitre à donner une information globale et synthétique sur la situation alimentaire. Pour cela, le cadre d'analyse global proposé permet d'obtenir des proxies de la sécurité alimentaire globaux relatifs aux articles publiés entre 2009 et 2018 dans tout le pays qui donnent une vision générale des caractéristiques de la situation alimentaire du pays. Ces informations peuvent également servir de comparaison aux analyses ciblées (régionale et annuelle).

Les proxies étudiés sont définis en deux étapes : nous commençons par extraire les articles pertinents pour notre analyse (i.e., définis comme thème "sécurité alimentaire" par le seuillage w2v), puis nous calculons les trois proxies sur les articles extraits. Plus précisément, nous calculons tout d'abord la proportion P_{SA} d'articles de thème "sécurité alimentaire" sur tout le corpus en divisant le nombre d'articles de similarité w2v supérieure à 0.36 (i.e., de thème sécurité alimentaire) par le nombre total d'articles (Équation 3.6). Puis, en se restreignant aux articles de thème sécurité alimentaire, nous calculons la proportion $P_{SA,neg}$ d'articles négatifs (Équation 3.7) et deux nuages de mots (pour les expressions du lexique $LEXA$ et pour celles du lexique $LEXC$) dont la taille de police dépend de la fréquence moyenne de chaque expression (Figure 3.5). L'utilisation du tf-idf n'est pas nécessaire pour ce niveau global d'analyse, elle trouve sa pertinence

dans les analyses ciblées (i.e., annuelles et régionales) pour faire ressortir les expressions caractéristiques de certaines années et régions.

$$P_{SA} = \frac{N_{SA}}{N} \tag{3.6}$$

Où P_{SA} est la proportion d'articles de thème "sécurité alimentaire", N_{SA} désigne le nombre d'articles de similarité w2v supérieure à 0.36 avec le lexique généraliste $LEXG$ (i.e., considérés comme de thème "sécurité alimentaire") et N correspond au nombre total d'articles du corpus.

$$P_{SA,neg} = \frac{N_{SA,neg}}{N_{SA}} \tag{3.7}$$

Où $P_{SA,neg}$ est la proportion d'articles traitant de sécurité alimentaire négatifs, $N_{SA,neg}$ désigne le nombre d'articles traitant de sécurité alimentaire négatifs et N_{SA} correspond au nombre total d'articles traitant de sécurité alimentaire.

Sur tout le territoire, entre 2009 et 2018, la proportion d'articles de thème "sécurité alimentaire" (i.e., tels que leur similarité w2v est supérieure à 0.36) est de 7.3%. Parmi ces articles traitant de sécurité alimentaire, la proportion d'articles négatifs est de 6.4%. La Figure 3.5 est constituée de deux nuages de mots représentant les expressions de sécurité alimentaire et des crises les plus fréquentes. Les termes de sécurité alimentaire les plus fréquents sont "sécurité alimentaire" et "agriculture". Concernant les termes de crises, "insécurité", "pauvreté", "inondation", "catastrophe" et "malnutrition" sont les plus employés. Ces pourcentages globaux et nuages de mots, assez génériques, nous disent que la sécurité alimentaire est un thème récurrent dans les articles avec une forte tendance à recourir à un vocabulaire inquiétant, voire alarmant. Ce contexte alimentaire problématique pointé ici est cohérent avec les analyses du chapitre 1 portant sur la situation alimentaire du pays et avec les rapports d'experts (FAO et al., 2020). Les proxies considérés sont dans la suite de ce chapitre analysés par région et par année afin d'obtenir des informations plus fines.

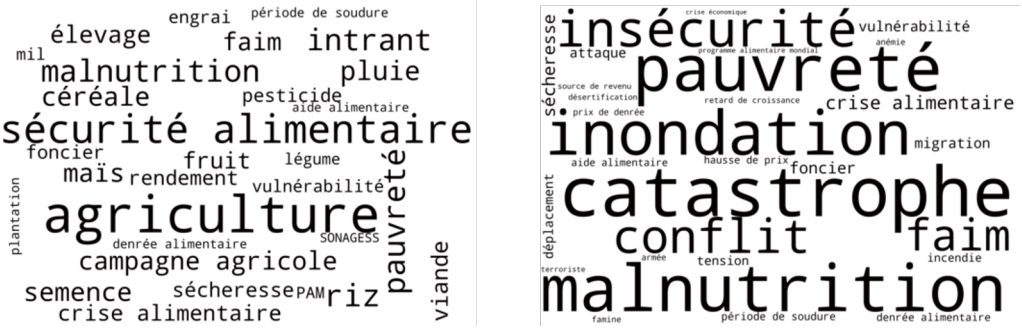


FIGURE 3.5 – Nuages de mots des expressions des lexiques $LEXA$ (gauche) et $LEXC$ (droite), basés sur les articles de thème "sécurité alimentaire" du corpus. La taille des termes est proportionnelle à leur fréquence moyenne.

3.4.2 Analyse régionale

Nous nous concentrons ici sur certaines régions et observons si les proxies de la sécurité alimentaire agrégés sur ces régions sont associés à la situation alimentaire régionale connue. Les trois régions étudiées sont les régions Centre, Hauts-Bassins et Sahel. Comme nous l'avons mentionné précédemment, ces trois régions présentent des conditions sanitaires distinctes que nous précisons ici. La région Hauts-Bassins figure parmi les moins pauvres du pays, son climat subhumide permet une production plus importante et diversifiée de céréales, tandis que la région Sahel située dans la zone nord du pays de climat semi-aride est en proie depuis plus d'une décennie à une situation sanitaire inquiétante. La région Centre, avec la capitale Ouagadougou, est davantage urbaine et connaît une situation contrastée caractéristique des zones à forte densité humaine. Plusieurs bulletins d'ONG décrivent la situation alimentaire de ces régions (WFP, 2014b ; OCHA, 2015 ; Zida and Kambou, 2014).

Pour chacune de ces trois régions, nous sélectionnons ses articles associés pour y centrer nos analyses avec la méthodologie détaillée dans la section 3.3. Sur l'ensemble des articles liés à une région, nous calculons la proportion P_{SA} d'articles de thème "sécurité alimentaire" (Équation 3.6 restreinte aux articles associés à chaque région), puis en se limitant à ces articles nous calculons la proportion $P_{SA,neg}$ d'articles négatifs (Équation 3.7 restreinte aux articles associés à chaque région) et deux nuages de mots (pour les expressions du lexique $LEXA$ et pour celles du lexique $LEXC$) (Figure 3.6). Pour les nuages de mots, la taille de police des expressions doit mettre en valeur les spécificités du

vocabulaire associé à chaque région. Pour cela, nous utilisons comme variable de taille de police le tf-idf moyen des expressions dans les articles liés à chaque région, permettant de mettre en avant les expressions qui sont spécifiquement employés dans les articles liés à une certaine région.

Nous observons maintenant si ces proxies obtenus pour chacune des régions Centre, Sahel et Hauts-Bassins sont associés à la situation alimentaire régionale connue. Dans le Tableau 3.7, nous constatons que la région Hauts-Bassins qui est la moins en proie à l'insécurité alimentaire parmi les régions présentées est associée aux proportions les plus basses d'articles de thème "sécurité alimentaire" et d'articles négatifs. A l'inverse, la région Sahel, qui connaît la situation la plus critique, possède les plus hautes proportions d'articles de thème "sécurité alimentaire" et d'articles négatifs, significativement plus élevés qu'au niveau national. Ces données sont cohérentes avec ce qui est attendu : plus une zone est en proie à l'insécurité alimentaire et/ou aux crises, plus les articles mentionnent ces sujets et sont négatifs.

	Centre	**Hauts Bassins**	**Sahel**	**Burkina Faso**
Pourcentage d'articles de thème SA	6.5	4.9	10.7	7.3
Pourcentage d'articles négatifs	4.8	1.3	12.1	6.4

Tableau 3.7 – Comparaison du pourcentage d'articles de thème "sécurité alimentaire" et du pourcentage d'articles négatifs pour les 3 régions Centre, Hauts-Bassins et Sahel.

Sur la Figure 3.6, nous constatons que pour la région Hauts-Bassins, la moins pauvre, les expressions du champ lexical de la sécurité alimentaire les plus importants sont neutres (e.g., "riz", "agriculture", "campagne agricole"), tandis que dans la région Sahel les expressions de sécurité alimentaire sont davantage négatives (e.g., "malnutrition", "crise alimentaire"). Concernant les expressions du thème des crises, les nuages de mots font ressortir des préoccupations caractéristiques de chaque région. Par exemple, l'expression "inondation" est la plus importante de la région Centre en proie à cette problématique (Lassailly-Jacob, 2015) tandis que l'expression "foncier" est la plus importante de la région Hauts-Bassins pour laquelle la gestion des terres est une problématique majeure (Karambiri, 2018).

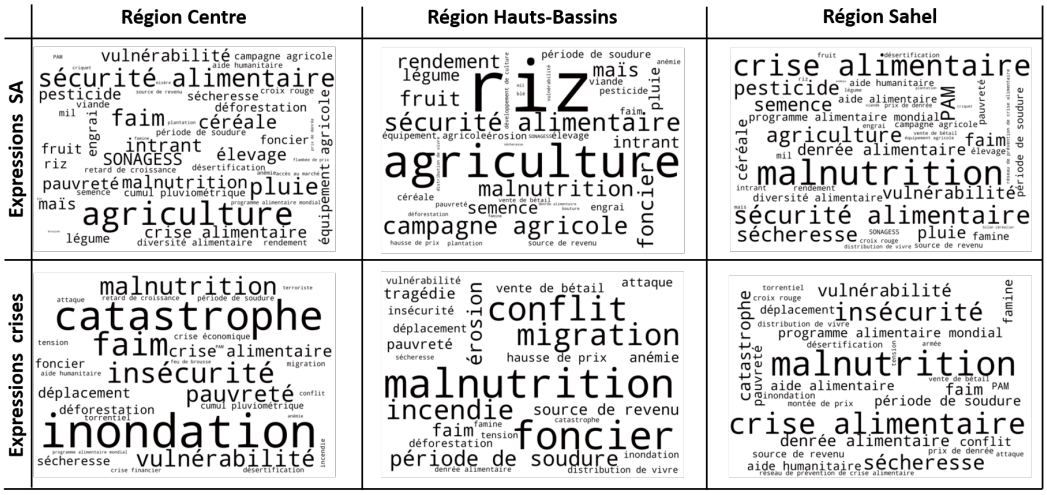


FIGURE 3.6 – Nuages de mots des expressions des lexiques $LEXA$ (Expressions SA) et $LEXC$ (Expressions crises), basés sur les articles de thème "sécurité alimentaire" liées à trois régions (Centre, Hauts-Bassins et Sahel). La taille des termes est proportionnelle à leur tf-idf moyen.

3.4.3 Analyse annuelle

Dans cette section, nous observons si les proxies de la sécurité alimentaire pointent dans le temps des éléments cohérents, nuancés, ou même en contradiction avec des observations et des évènements qui ont eu lieu lors de la dernière décennie et qui ont pu affecter la sécurité alimentaire. À savoir, un recul de la sécurité alimentaire depuis 2013 (Figure 3.7) (FAO et al., 2020) ainsi que des événements impactant négativement la sécurité alimentaire (e.g., inondation, sécheresse, conflit). Nous pouvons noter les inondations de 2009 et 2010 (OCHA, 2010 ; Burkina Faso Government, 2009), la révolte de 2011 (Le Monde, 2011), la sécheresse de 2012 (World Bank, 2012 ; WFP, 2012), de forts déplacements de populations (ONU Info, 2013) et attaques de criquets (Le Hub Rural, 2013) en 2013.

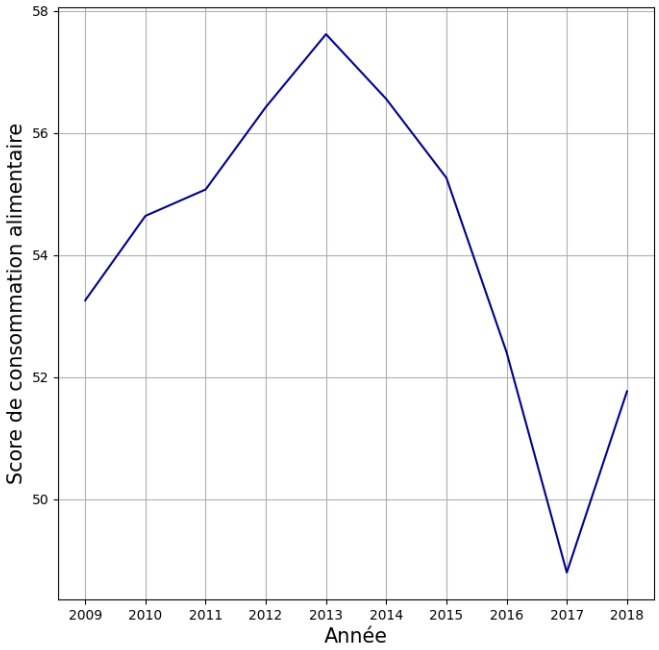


FIGURE 3.7 – Évolution du score de consommation alimentaire moyen annuel de 2009 à 2018 (Deléglise et al., 2020). Ce score a été calculé avec les données de l'enquête permanente agricole burkinabè. Une diminution de ce score est associée à une diminution de la sécurité alimentaire.

L'objectif est ici de connaître les caractéristiques et l'évolution annuelles de la situation alimentaire burkinabè à travers les proxies de sécurité alimentaire considérés. Des analyses temporelles plus fines (i.e., par saison ou par mois) sont possibles car les métadonnées associées à chaque article contiennent la date exacte de publication, celles-ci seraient pertinentes mais demanderaient un travail conséquent qui devrait faire l'objet de travaux futurs. Les articles du corpus sont tout d'abord regroupés par année de 2009 à 2018, en utilisant les dates de publications des articles disponibles dans leurs métadonnées. Pour chaque année, nous calculons la proportion P_{SA} d'articles de thème "sécurité alimentaire" (Équation 3.6 restreinte aux articles associés à chaque année), puis en se limitant à ces articles, nous calculons la proportion $P_{SA,neg}$ d'articles négatifs (Équation 3.7 restreinte aux articles associés à chaque année). Ces proxies de la sécurité alimentaire sont illustrés par les Figures 3.8 et 3.9 (a). Dans le but d'extraire le vocabulaire sur la

sécurité alimentaire et les crises spécifique aux situations des différentes années, nous utilisons de nouveau la notion de tf-idf. Plus précisément, nous calculons pour chaque expression des lexiques $LEXA$ et $LEXC$ le tf-idf de l'expression en moyenne sur les articles de l'année, ainsi que le ratio TIR que nous avons proposé et qui est obtenu par division de ce tf-idf par le tf-idf de l'expression en moyenne sur les articles des autres années. Le rôle de ce ratio TIR est de distinguer davantage les expressions spécifiques d'une année. En effet, les expressions des vocabulaires considérés sont très inégalement employées dans les articles, et certaines expressions très utilisées ont tendance à occulter d'autres expressions globalement plus rares, bien que plus spécifiques de certaines années. Pour ce niveau d'analyse, l'utilisation du tf-idf, qui dépend encore fortement de la fréquence d'occurrence des expressions, n'est pas toujours suffisante pour mettre en évidence certaines expressions très spécifiques mais moins fréquentes.

Le Tableau 3.8 compare les tops 10 des termes du lexique détaillé $LEXA$ les plus spécifiques de 2013 selon 3 mesures d'importance des termes utilisés : la fréquence, le tf-idf et le ratio TIR. Nous savons que cette année 2013 a été marquée par une augmentation significative du nombre de criquets pèlerins dans les pays du Sahel, véritable fléau pour les cultures. Nous constatons dans le Tableau 3.8 qu'en considérant la fréquence des termes, seul le terme "criquet" apparaît en dernière position du top 10. En prenant en compte le tf-idf des termes, le terme "criquet" entre dans le top 5 du classement tandis que "pèlerin" y entre. Enfin, avec le ratio TIR des termes, "criquet" et "pèlerin" occupent les 2 premières places. À l'inverse, les termes "sécurité alimentaire" et "agriculture" du lexique $LEXA$ sont les plus utilisés dans le corpus. En utilisant la fréquence comme mesure d'importance, ces deux termes occupent les premières places pour presque toutes les années, ce qui ne nous renseigne pas sur la survenue d'événements ponctuels. Si l'on considère plutôt le tf-idf et le ratio TIR, ces termes, qui sont toujours présents car ils restent importants, laissent plus de place aux autres.

Fréquence	**tf-idf**	**Ratio TIR**
Sécurité alimentaire	Sécurité alimentaire	Pèlerin
Agriculture	Crise alimentaire	Criquet
Crise alimentaire	Aide alimentaire	Aide alimentaire
Riz	Agriculture	Prix des denrées
Aide alimentaire	Criquet	Crise alimentaire
Céréale	Riz	Céréale
Faim	Céréale	Sécurité alimentaire
Pauvreté	Faim	Faim
Malnutrition	Prix des denrées	Riz
Criquet	Pèlerin	Agriculture

Tableau 3.8 – Comparaison des top 10 des termes du lexique détaillé $LEXA$ les plus spécifiques de 2013 selon 3 mesures utilisées : la fréquence, le tf-idf et le ratio TIR. Les cellules en bleu correspondent aux termes "criquet" et "pèlerin" qui sont très spécifiques aux articles de l'année 2013, les cellules en jaune correspondent aux termes "sécurité alimentaire" et "agriculture" qui sont les termes du lexique $LEXA$ les plus utilisés dans le corpus.

Dans un souci de lisibilité, les logarithmes de ces ratios TIR sont considérés et représentés sur deux graphiques radar (pour les thèmes "sécurité alimentaire" et "crise"), et pour chaque année (soit 20 graphiques radars représentés Figure 3.11). L'idée étant qu'une valeur de TIR supérieure à 0 pour une expression et une année données signifie que l'expression est davantage spécifique des articles de cette année que pour les autres années. Les 10 axes du radar sont occupés par les ratios TIR associés aux 10 expressions possédant le plus grand tf-idf d'une année donnée.

Nous examinons maintenant si les proxies obtenus révèlent des résultats cohérents dans le temps avec les évolutions et les événements qui se sont produits de 2009 à 2018. Sur la Figure 3.8, nous constatons que la proportion d'articles traitant de sécurité alimentaire a significativement augmenté entre 2009 et 2018, en doublant sur la période pour dépasser les 10% en 2018, soit 3% de plus qu'en moyenne sur la décennie. Cette tendance des articles à contenir davantage d'informations sur la sécurité alimentaire en période de crise a été mise en évidence dans la section 3.2 et confirme l'aggravation

connue de la situation alimentaire au cours de la décennie étudiée.

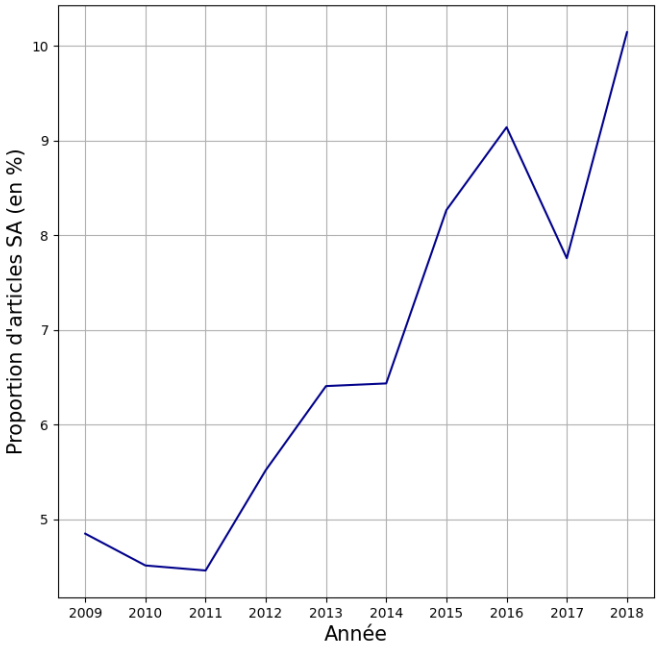


FIGURE 3.8 – Évolution de la proportion (en pourcentage) d'articles de thème "sécurité alimentaire" de 2009 à 2018 sur le corpus étudié.

Nous regardons si l'aggravation connue de la situation sanitaire au Burkina Faso entre 2009 et 2018 est liée à des fluctuations de la proportion d'articles négatifs sur la période. Sur la Figure 3.9 (a) qui représente l'évolution de la proportion d'articles négatifs par année de 2009 à 2018, nous constatons une tendance des articles négatifs à diminuer en proportion. Cela peut sembler contre-intuitif, et peut s'expliquer par une certaine liberté de la presse qui aurait tendance à reculer au cours de la dernière décennie (voir Figure 3.9 (b)).

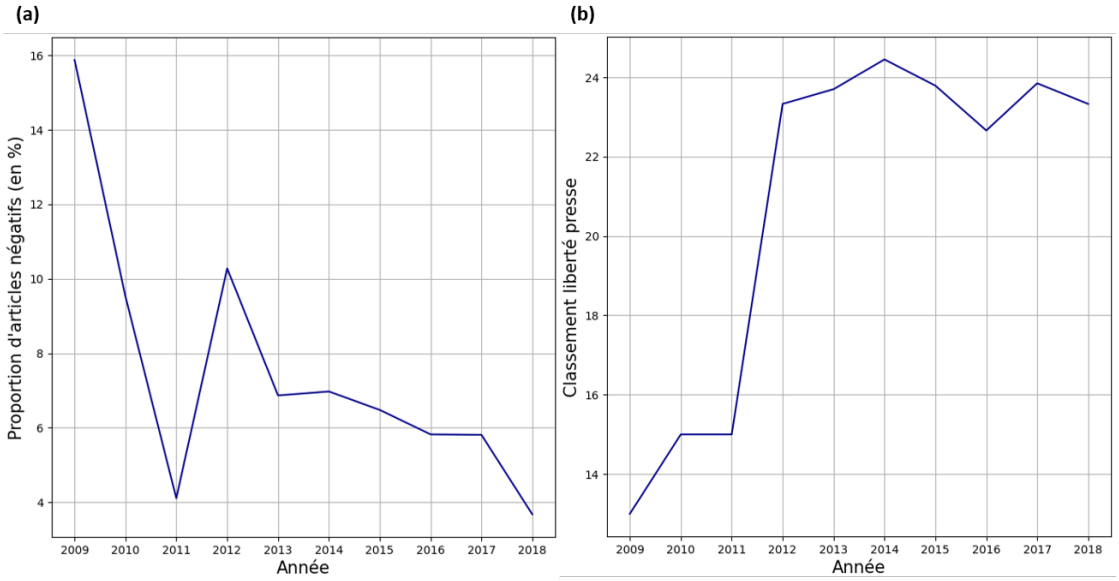


FIGURE 3.9 – Évolution de 2009 à 2018 de la proportion (en pourcentage) d'articles négatifs par année parmi les articles de thème "sécurité alimentaire" (a) et du Burkina Faso dans le classement de la liberté de la presse (Reporters sans frontière [17]) (b).

Nous analysons maintenant l'évolution du vocabulaire de la sécurité alimentaire et des crises employé dans les articles en fonction du temps. La Figure 3.10 montre l'évolution du tf-idf de 5 expressions en moyenne sur les articles de chaque année entre 2009 et 2018. Ces expressions ont été choisies car elles représentent les principales sources de préoccupation alimentaire et sanitaire depuis la dernière décennie (FAO et al., 2020). Le terme "sécurité alimentaire", le plus utilisé dans l'ensemble du corpus, renvoie au sujet de plus en plus crucial et préoccupant de la suffisance et de la qualité de l'alimentation. Une conséquence importante est la "malnutrition", notamment chez les enfants, qui est une préoccupation croissante. Le pays est particulièrement en proie au réchauffement climatique et à ses épisodes de "sécheresse" de plus en plus fréquents. Les "conflits" qui ont éclaté au cours de la dernière décennie ont mis la population sous une tension permanente par peur d'attaques, provoquant également des "déplacements" de populations de certaines régions du pays et des pays voisins, ce qui augmente encore davantage les tensions. L'utilisation du tf-idf se justifie ici par la nécessité de placer les différentes expressions sur la même échelle pour les rendre simultanément visualisables et compa-

17. https ://rsf.org/fr/methodologie-detaillee-du-classement-mondial-de-la-liberte-de-la-presse

rables sur le graphique. En utilisant la fréquence moyenne des termes, certains termes beaucoup plus fréquents que d'autres, comme "sécurité alimentaire", les auraient fait disparaître contre l'axe des abscisses, rendant le graphique illisible. Nous pouvons tout d'abord constater une tendance à la hausse des tf-idf des expressions "sécurité alimentaire" et "malnutrition" qui ont été de plus en plus employés au cours de la dernière décennie. De plus, certains pics correspondent à l'année de survenue d'événements qui ont eu lieu sur la période : le tf-idf de l'expression "sécheresse" est le plus haut en 2012 qui a connu une forte sécheresse. Les tf-idf des expressions "conflit" et "déplacement" atteignent leur maximum en 2013, durant cette année des conflits dans le Sahel ont entraîné des déplacements de populations des pays sahéliens limitrophes au Burkina Faso.

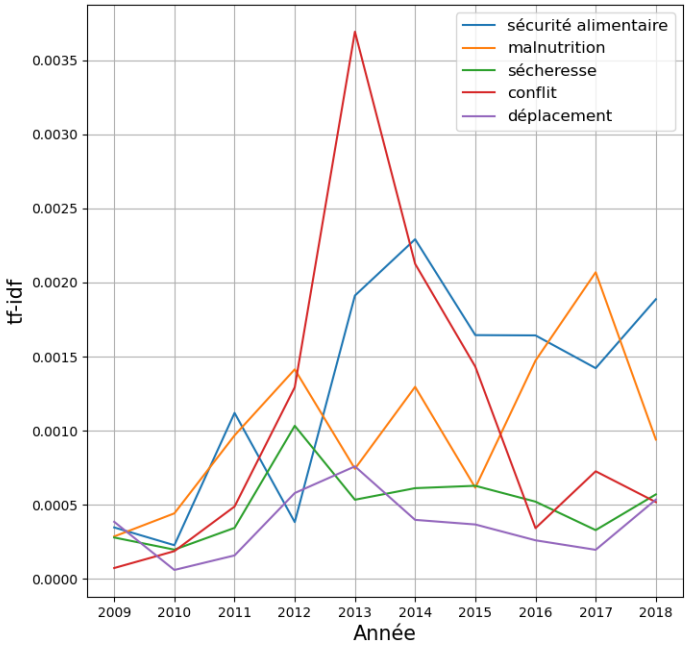
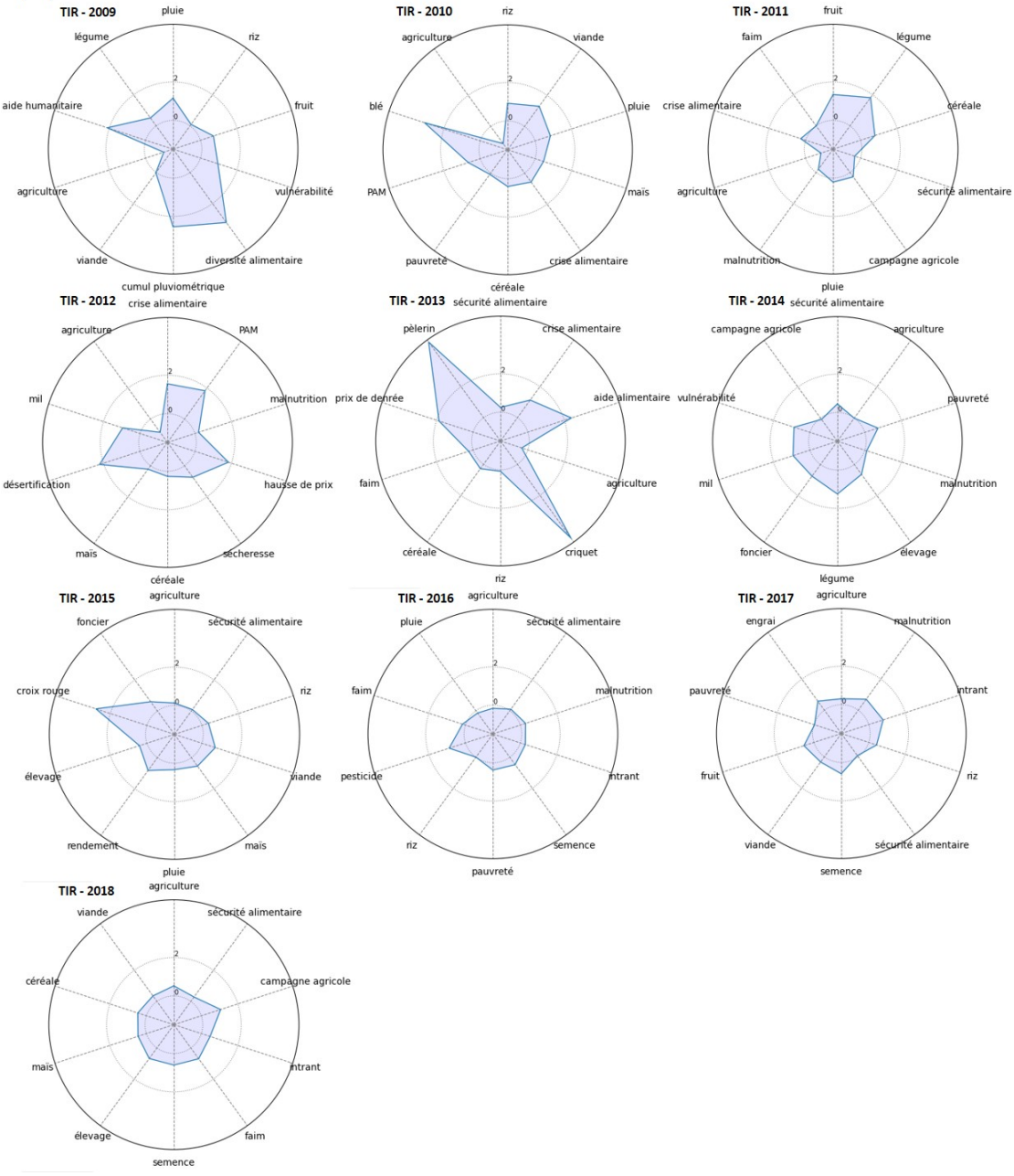


FIGURE 3.10 – Évolution du tf-idf moyen de 5 expressions issues des deux lexiques détaillés $LEXA$ et $LEXC$ entre 2009 et 2018.

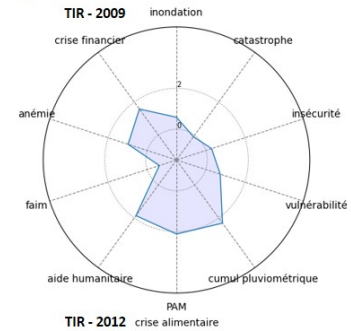
La Figure 3.11 consiste en 20 graphiques radar qui représentent l'évolution des dix expressions les plus caractéristiques de chaque année entre 2009 et 2018 pour les lexiques $LEXA$ et $LEXC$. Pour les expressions de $LEXA$, nous constatons que l'expression "sécurité alimentaire" est de plus en plus utilisée au cours des années, ce qui indique que ce sujet gagne en importance avec les années, nous observons également que les expressions

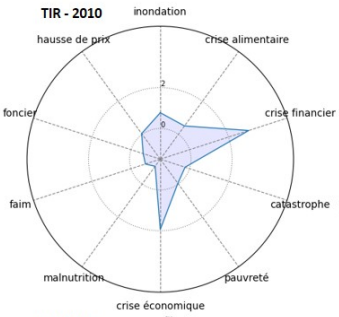
neutres (”riz”, ”fruit”, ”viande”, ”céréale”), les plus caractéristiques des années 2009 et 2010 laissent petit à petit leur place à des expressions davantage négatives (e.g., crise alimentaire, malnutrition, hausse des prix). Concernant les expressions de $LEXC$, les graphiques radar pointent dans le temps un vocabulaire qui correspond avec les crises les plus importantes de la dernière décennie : le terme ”inondation” possède le plus grand tf-idf en 2009 et 2010, années durant lesquelles il y a eu de violentes inondations dans le pays. En 2011, c'est le terme ”incendie” qui possède le tf-idf le plus élevé, cette année-là avait été marquée par une révolte durant laquelle plusieurs commissariats et bâtiments administratifs avaient été incendiés. Nous pouvons enfin noter que le terme ”migration” a le second tf-idf le plus élevé en 2017 et 2018, années à partir desquelles les tensions dans la zone sahélienne et les déplacements de populations qui en résultent se sont intensifiés. D'autres expressions relatives aux crises retiennent notre attention car possèdent un ratio TIR élevé, ce ratio TIR permet de mettre en avant des expressions moins utilisées mais qui sont remarquables pour une année précise. En 2012, les termes ”désertification” et ”sécheresse” ont un ratio TIR élevé (le plus élevé sur l'année pour ”désertification” et significativement supérieur à 0 pour ”sécheresse”). Durant cette année de forts épisodes de sécheresse ont traversé le Burkina Faso. Enfin, en 2013 les termes ”criquet” et ”pèlerin” ont des ratios TIR très élevés (supérieur à 4, ce qui signifie que les tf-idf moyens de ces expressions sur les articles de l'année 2013 sont plus de cinquante-cinq fois plus élevé que leur tf-idf moyens considérés sur les articles des autres années).

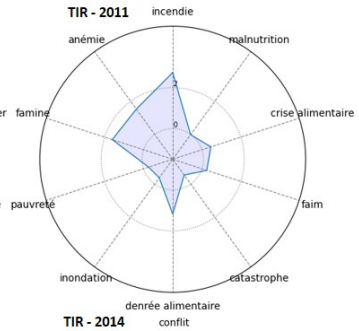
(SA)

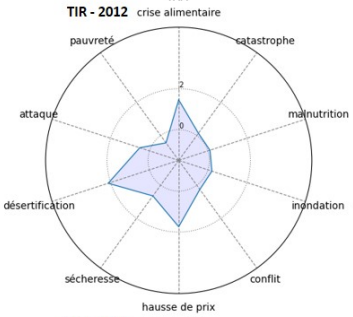

TIR - 2009
pluie
riz
fruit
vulnérabilité
diversité alimentaire
cumul pluviométrique
viande
agriculture
aide humanitaire
légume
TIR - 2010
riz
viande
pluie
maïs
crise alimentaire
céréale
pauvreté
PAM
blé
agriculture
TIR - 2011
fruit
légume
céréale
sécurité alimentaire
campagne agricole
pluie
malnutrition
agriculture
crise alimentaire
faim
TIR - 2012
crise alimentaire
PAM
malnutrition
hausse de prix
sécheresse
céréale
maïs
désertification
mil
agriculture
TIR - 2013
sécurité alimentaire
crise alimentaire
aide alimentaire
agriculture
criquet
riz
céréale
faim
prix de denrée
pèlerin
TIR - 2014
sécurité alimentaire
agriculture
pauvreté
malnutrition
élevage
légume
foncier
mil
vulnérabilité
campagne agricole
TIR - 2015
agriculture
sécurité alimentaire
riz
viande
maïs
pluie
rendement
élevage
croix rouge
foncier
TIR - 2016
agriculture
sécurité alimentaire
malnutrition
intrant
semence
pauvreté
riz
pesticide
faim
pluie
TIR - 2017
agriculture
malnutrition
intrant
riz
sécurité alimentaire
semence
viande
fruit
pauvreté
engrai
TIR - 2018
agriculture
sécurité alimentaire
campagne agricole
intrant
faim
semence
élevage
maïs
céréale
viande

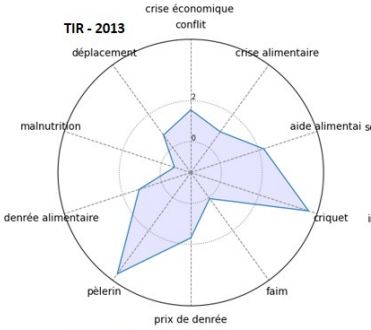
(CR)

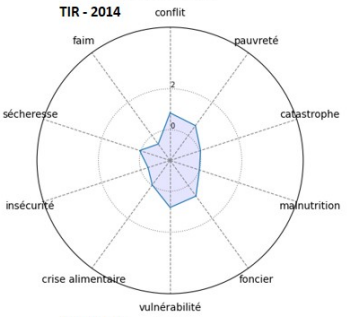

TIR - 2009
inondation
catastrophe
insécurité
vulnérabilité
cumul pluviométrique
PAM
aide humanitaire
faim
anémie
crise financier

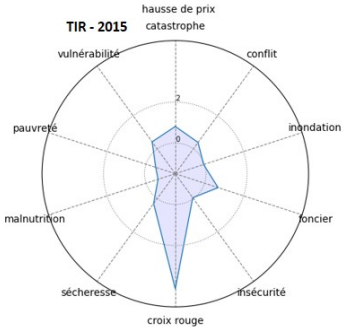

TIR - 2010
inondation
crise alimentaire
crise financier
catastrophe
pauvreté
crise économique
malnutrition
faim
foncier
hausse de prix

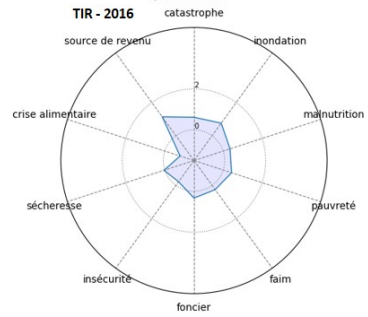

TIR - 2011
incendie
malnutrition
crise alimentaire
faim
catastrophe
denrée alimentaire
inondation
pauvreté
famine
anémie

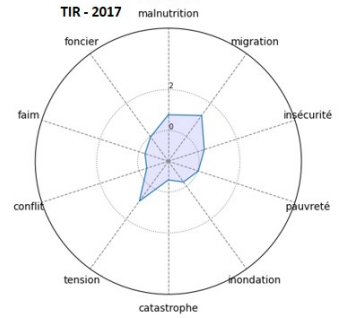

TIR - 2012
crise alimentaire
catastrophe
malnutrition
inondation
conflit
hausse de prix
sécheresse
désertification
attaque
pauvreté


TIR - 2013
conflit
crise alimentaire
aide alimentai
criquet
faim
prix de denrée
pèlerin
denrée alimentaire
malnutrition
déplacement


TIR - 2014
conflit
pauvreté
catastrophe
malnutrition
foncier
vulnérabilité
crise alimentaire
insécurité
sécheresse
faim


TIR - 2015
catastrophe
conflit
inondation
foncier
insécurité
croix rouge
sécheresse
malnutrition
pauvreté
vulnérabilité


TIR - 2016
catastrophe
inondation
malnutrition
pauvreté
faim
foncier
insécurité
sécheresse
crise alimentaire
source de revenu


TIR - 2017
malnutrition
migration
insécurité
pauvreté
inondation
catastrophe
tension
conflit
faim
foncier

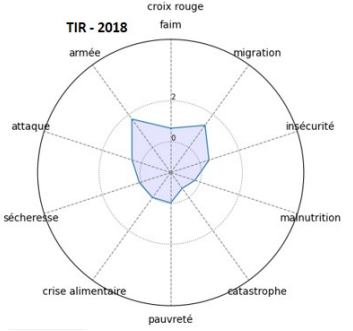

TIR - 2018
faim
migration
insécurité
malnutrition
catastrophe
pauvreté
crise alimentaire
sécheresse
attaque
armée

FIGURE 3.11 – Graphiques radar représentant l'évolution des dix expressions les plus caractéristiques de chaque année entre 2009 et 2018 pour les lexiques $LEXA$ (SA) et $LEXC$ (CR). Les 10 axes sont occupés par les 10 expressions possédant les valeurs de tf-idf les plus élevées sur les articles de l'année, par valeur décroissante dans le sens des aiguilles d'une montre. Les valeurs des axes représentent les ratios des tf-idf des expressions en moyenne sur les articles de l'année par les tf-idf des expressions en moyenne sur les articles des autres années (ratio TIR).

3.4.4 Perspective d'Analyse

Dans cette section, nous proposons une perspective d'analyse intéressante offrant un regard différent sur les données, permettant d'affiner davantage le contexte explicatif obtenu à partir du vocabulaire employé dans les articles. La démarche proposée consiste en l'analyse des co-occurrences dans les articles entre les termes des lexiques $LEXA$ et $LEXC$ afin de mieux comprendre dans quel contexte et pour quels discours certains termes sont utilisés.

De nombreuses études utilisent cette notion de co-occurrence pour extraire des associations sémantiques entre différentes expressions dans des textes (Bordag, 2008 ; Rachakonda et al., 2014 ; Hollis and Westbury, 2016 ; Valentin et al., 2021a). La co-occurrence entre deux termes désigne leur présence simultanée dans une phrase ou dans un texte. L'idée étant que si deux termes apparaissent souvent simultanément, alors ils sont sémantiquement proches. La Figure 3.12 représente sur deux graphes les dix expressions des lexiques $LEXA$ et $LEXC$ les plus co-occurrents, selon la mesure de l'information mutuelle (Feldman et al., 1998), avec les expressions "sécurité alimentaire" et "agriculture" sur tout le corpus. Ces deux expressions sont choisies car elles sont les plus utilisées dans le corpus, par ailleurs leur caractère générique et leur valence neutre peuvent faire ressortir un large éventail de termes aux thèmes et aux valences variés. La mesure de similarité utilisée, permettant de quantifier le degré de co-occurrence entre deux termes, est l'information mutuelle. L'information mutuelle de deux variables aléatoires (dans notre cas ce sont des variables binaires modélisant la "présence / absence" d'un terme dans un article) fait référence à une valeur mesurant la dépendance statistique entre ces variables. L'information mutuelle entre deux expressions vaut 0 si les deux expressions ne se trouvent jamais simultanément dans un même article, et peut croître sans majora-

tion si les deux expressions sont souvent présentes dans les mêmes articles tout en étant absentes dans un grand nombre d'article. Sa formule est donnée dans l'Équation 3.8. Notons que plusieurs autres contextes de co-occurrence peuvent être appliqués, notamment en considérant que deux termes sont co-occurrents s'ils se trouvent dans le même paragraphe, ou dans la même phrase. Plus la contrainte sur la proximité entre les termes est forte, plus il est probable qu'ils soient associés sur des bases pertinentes (i.e., deux termes ont plus de chances d'être des éléments du même propos s'ils sont dans la même phrase que s'ils sont seulement dans le même article), mais moins la formule de similarité utilisée convergera car le nombre d'observations de termes considérés comme apparentés sera plus faible. Le choix du niveau contextuel le plus approprié n'est qu'évoqué ici et devrait faire l'objet de travaux futurs. Ce compromis entre la pertinence et le nombre des données analysées peut être rapproché au compromis que nous avons dû trouver pour le choix des seuils dans la section 3.3.2.3, cette problématique semble être fréquente en fouille de textes.

$$I(Exp1, Exp2) = \frac{P(Exp1, Exp2)}{P(Exp1) \times P(Exp2)} \tag{3.8}$$

Où I représente la mesure de l'information mutuelle entre deux expressions $Exp1$ et $Exp2$, $P(Exp1, Exp2)$ désigne la probabilité pour un article de contenir conjointement $Exp1$ et $Exp2$, $P(Exp1)$ et $P(Exp2)$ sont les probabilités pour un article de contenir $Exp1$ (resp. $Exp2$).

Nous constatons (Figure 3.12) que le terme "sécurité alimentaire", possède comme termes les plus co-occurrents "production céréalière", "SONAGESS" qui est l'organisme burkinabé de gestion des stocks alimentaire et "légumineux". Notons que les termes les plus co-occurrents avec "sécurité alimentaire" sont globalement du champ lexical de l'agriculture, secteur qui est de loin le plus dynamique dans le pays et qui est la principale source de nourriture pour la population. Concernant le terme "agriculture", nous retrouvons dans le top 10 de ses termes les plus co-occurrents plusieurs termes en commun avec "sécurité alimentaire" comme "production céréalière", "développement de cultures" ou "bilan céréalier", mais également d'autres termes propres aux méthodes d'agriculture comme "engrais", "bouture" ou "semence".

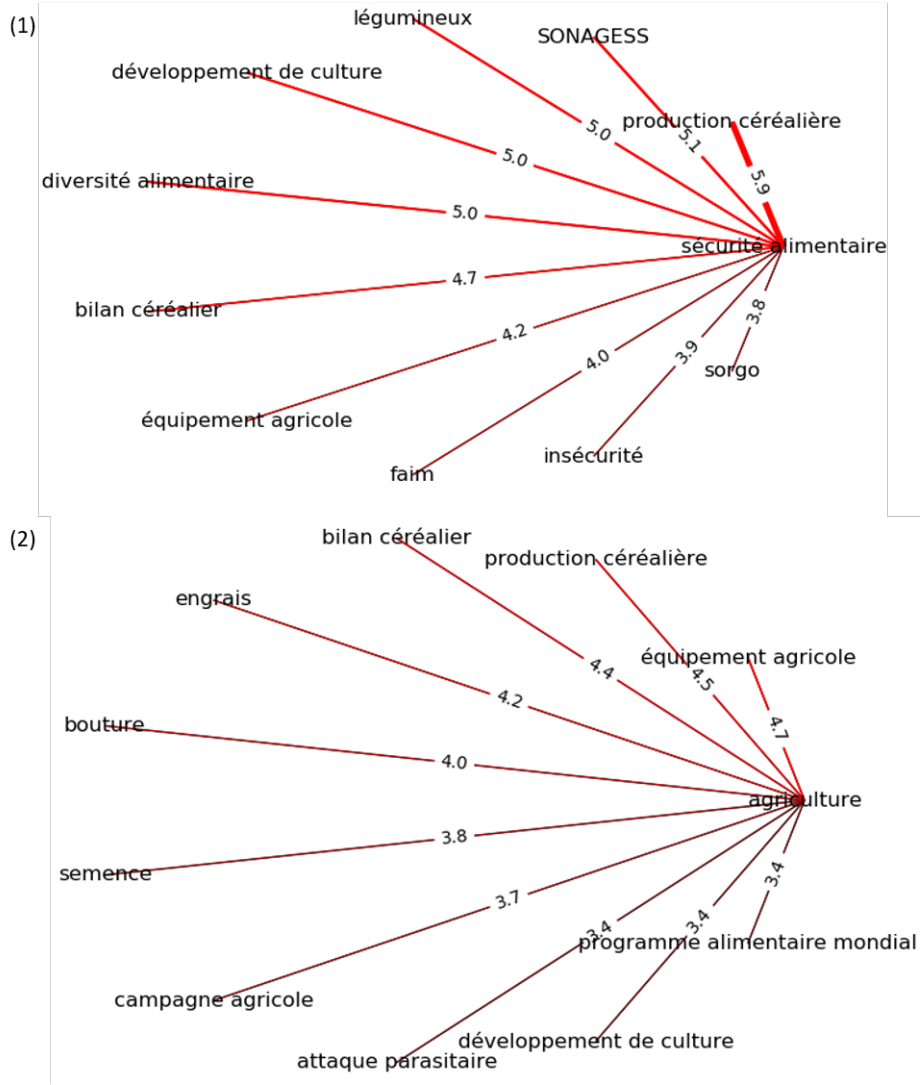


FIGURE 3.12 – Graphes des 10 expressions des lexiques détaillés $LEXA$ et $LEXC$ les plus co-occurrents (selon la mesure de l'information mutuelle) avec les expressions "sécurité alimentaire" (1) et "agriculture" (2) sur tout le corpus.

Une autre mesure de similarité est couramment exploitée dans ce domaine, il s'agit du coefficient DICE (McKeown et al., 1996) (Équation 3.9).

$$DICE(Exp1, Exp2) = \frac{2 \times P(Exp1, Exp2)}{P(Exp1) + P(Exp2)} \tag{3.9}$$

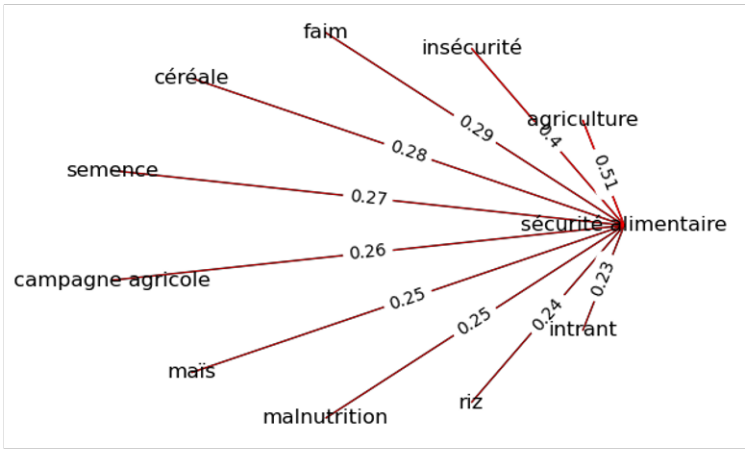
Où $DICE$ représente le coefficient DICE entre deux expressions $Exp1$ et $Exp2$, $P(Exp1, Exp2)$ désigne la probabilité pour un article de contenir conjointement $Exp1$ et $Exp2$, $P(Exp1)$ et

$P(Exp2)$ sont les probabilités pour un article de contenir $Exp1$ (resp. $Exp2$).

Cette mesure de similarité a tendance à privilégier des termes plus fréquents dans les articles, contrairement à l'information mutuelle qui valorise les termes rares (Valentin et al., 2021b). Cela s'explique par les dénominateurs de ces formules (Équations 3.8 et 3.9) : le dénominateur de la formule de l'information mutuelle multiplie les probabilités de présence des deux termes $Exp1$ et $Exp2$, si ces probabilités sont très faibles, le produit le sera d'autant plus, ce qui fera fortement augmenter l'information mutuelle ; inversement, les probabilités de présence des termes $Exp1$ et $Exp2$ sont additionnées dans le dénominateur de DICE, la rareté de ces termes aura donc un effet amplificateur plus faible sur la valeur de DICE.

Nous évaluons maintenant si les deux mesures de similarité DICE et d'information mutuelle fournissent des informations distinctes dans notre contexte. Pour cela, nous examinons les différences de termes mis en avant par l'information mutuelle et DICE en recréant exactement la même sortie que pour la Figure 3.12, mais en remplaçant la mesure d'information mutuelle par le coefficient DICE comme mesure de similarité utilisée. La Figure 3.13 représente sur deux graphes les dix expressions des lexiques $LEXA$ et $LEXC$ les plus co-occurrents, selon la mesure de DICE, avec les expressions "sécurité alimentaire" et "agriculture" sur tout le corpus. Nous constatons tout d'abord que les termes identifiés par ces deux mesures sont bien distincts : pour les termes "sécurité alimentaire" et "agriculture", 8 de leurs termes les plus co-occurrents sur 10 sont différents selon la mesure de similarité utilisée. De plus, comme les propriétés de DICE pouvaient le présager, les termes mis en avant par DICE sont davantage génériques et fréquents dans le corpus (e.g., "sécurité alimentaire", "agriculture", "céréale", "malnutrition", qui n'apparaissent pas en utilisant la formule de l'information mutuelle au profit de termes plus rares : "attaque parasitaire", "développement de cultures", "SONAGESS"). Nous pouvons en conclure que ces deux mesures de similarité permettent de produire des informations distinctes et complémentaires : le coefficient DICE associe un concept étudié avec un vocabulaire co-occurrent fréquent et générique, tandis que l'information mutuelle fait ressortir un vocabulaire plus rare et spécifique. Il peut donc être pertinent d'utiliser ces deux mesures conjointement afin d'obtenir une vision plus complète du vocabulaire lié à un concept.

(1)



(2)

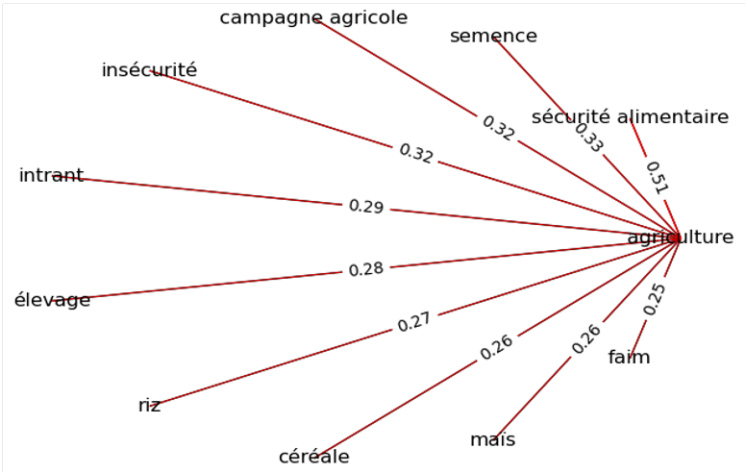


FIGURE 3.13 – Graphes des 10 expressions des lexiques détaillés $LEXA$ et $LEXC$ les plus co-occurrents (selon la mesure de DICE) avec les expressions ”sécurité alimentaire” (1) et ”agriculture” (2) sur tout le corpus.

Pour aller plus loin, les analyses pourraient être diversifiées à plusieurs niveaux :

— En utilisant d'autres mesures de similarité, en complément de l'information mutuelle et du coefficient DICE. Il en existe de nombreuses autres (Lenca et al., 2007), offrant chacune un regard différent sur les termes testés (e.g., coefficient de Jacquard (Hamers et al., 1989), coefficient de recouvrement (Lawlor, 1980)).

— En fixant d'autres fenêtres contextuelles de co-occurrence (e.g., paragraphe, phrase). En effet, comme précisé précédemment cela peut augmenter la pertinence des analyses si la taille de la fenêtre est correctement paramétrée.

— En évaluant d'autres termes dont nous pourrions examiner le vocabulaire le plus co-occurrent. En particulier, il serait intéressant de tester certains termes négatifs, ce qui nous permettrait d'attribuer une polarité négative aux termes les plus co-occurrents du lexique $LEXA$, relatifs à l'agriculture et à la nutrition qui ne sont habituellement pas connotés négativement mais qui pourraient l'être dans certains contextes. Il s'agirait d'une approche complémentaire et plus fine de la négativité que celle basée sur le taux de négativité des articles adoptée dans cette thèse.

— En intégrant des lexiques associés à d'autres sujets liés à la sécurité alimentaire (e.g., l'économie, la santé, la sécurité, la météorologie) pour mettre en évidence des connexions sémantiques plus riches avec les termes d'intérêt.

— En ciblant ce type d'analyse dans le temps et dans l'espace, afin de mettre en évidence des associations sémantiques spécifiques à certaines années et/ou régions.

3.5 Conclusion

Dans ce chapitre, nous avons examiné l'aptitude des méthodes de fouille de texte pour extraire des informations thématiques spatiales et temporelles sur la sécurité alimentaire à partir d'articles de journaux en illustrant avec le contexte du Burkina Faso.

Nous avons proposé, combiné et étendu avec des méthodes de fouille de texte adaptées (le modèle Word2vec de plongement lexical, le modèle VADER d'analyse de sentiments et la méthode tf-idf de pondération de l'importance de termes) trois types de proxies définis sur un ensemble d'articles, permettant d'obtenir des informations distinctes et complémentaires sur la thématique de la sécurité alimentaire : 1) la proportion d'articles qui abordent la sécurité alimentaire, qui donne une indication du niveau d'intérêt et de préoccupation à l'égard de la situation alimentaire ; 2) la proportion d'articles qui abordent la sécurité alimentaire avec un caractère négatif, qui donne une estimation de leur caractère inquiétant, voire alarmant ; 3) le vocabulaire (contenu dans les lexiques

thématiques $LEXA$ et $LEXC$) spécifique aux articles du thème de la sécurité alimentaire, qui donne un contexte explicatif sur les causes et caractéristiques des pénuries et crises alimentaires.

Nous avons pris en compte l'aspect spatio-temporel de la sécurité alimentaire en effectuant des analyses ciblées aux niveaux global, régional et annuel. Les proxies de la sécurité alimentaire ont été agrégés à ces trois niveaux avec des méthodes d'agrégations appropriées (proportions, fréquences, tf-idf, et une nouvelle mesure appelée ratio TIR). Nous avons proposé des représentations graphiques adaptées à chaque proxy et niveau d'analyse permettant de mettre en évidence les informations obtenues (nuages de mots, graphiques de séries temporelles, graphiques radars et graphes). Cela nous a permis de faire ressortir avec succès les tendances alimentaires régionales et annuelles ainsi que les crises qui ont touché le pays au cours de la dernière décennie.

Notons que l'approche proposée dans ce chapitre, basée sur de la reconnaissance de thématique, de l'analyse de sentiments et de l'extraction de vocabulaire spécifique est générique et pourrait être utilisée pour d'autres thématiques en construisant des lexiques portant sur d'autres sujets et pourrait être appliquée à d'autres supports textuels (e.g., réseaux sociaux, publications scientifiques).

Ce type d'approche et les résultats associés peuvent être exploités comme informations complémentaires aux sorties des modèles prédictifs proposés dans le chapitre précédent. En effet, les modèles d'apprentissage automatique et profond appliqués aux autres types de données (e.g., aux données numériques et images satellitaires) possèdent un fort pouvoir prédictif mais présentent souvent un manque d'explicabilité et d'interprétabilité. Ces modèles peuvent alors être validés, nuancés ou encore expliqués par les informations qualitatives issues des données textuelles qui pourraient faire sens auprès des thématiciens et faire avancer leur compréhension des phénomènes complexes liés à la sécurité alimentaire.

Soulignons qu'une analyse rétrospective a été réalisée dans ce chapitre, mettant en évidence plusieurs proxies et seuils associés qui ont été évalués sur un large ensemble de données et qui ont été jugés pertinents dans notre contexte. Ces proxies et seuils pourraient être pris en compte pour la conception d'analyses prospectives, par exemple au moyen de modèles prédictifs. La faible densité spatiale des données issues du corpus de journaux constitué et exploité dans le cadre de ces travaux de thèse ne nous a pas permis

d'y appliquer ces modèles prédictifs. Toutefois, avec la tendance croissante des journaux à publier en ligne via leur site web et la démocratisation de l'utilisation des réseaux sociaux dans les pays du Sud, il sera bientôt possible d'obtenir des corpus de textes exploitables pour cette tâche. Il pourrait alors être pertinent de compléter le framework conçu lors du chapitre 2 avec des modèles prenant en compte ce type de données textuelles, complémentaires aux autres types de données. Plus précisément, nous pourrions y intégrer les proxies textuels proposés dans ce chapitre qui mesurent la tendance des articles à traiter de la sécurité alimentaire, à présenter un aspect négatif et à utiliser un vocabulaire spécifique dans un contexte de crise. Cela soulèverait alors plusieurs questions sur le traitement des données textuelles telles que le modèle à utiliser ou encore le type de fusion à réaliser : faire une fusion dite "sur les données" en y appliquant le moins de prétraitements possible avant de les mettre en entrée d'un modèle prédictif aux côtés des autres types de données ; effectuer une fusion dite "sur les caractéristiques" en y extrayant des descripteurs qui seraient typiquement les proxies présentés dans ce chapitre pour les mettre en entrée d'un modèle prédictif parallèlement aux autres types de données ; réaliser une fusion dite "au niveau des décisions" en appliquant un modèle prédictif adapté à la donnée textuelle, puis en agrégeant la prédiction avec les prédictions des modèles liés aux autres types de données en une prédiction globale.

Pour améliorer la recherche thématique de manière plus fine que les plongements lexicaux appliqués dans ces travaux avec w2v, des technologies fondées sur BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) et les modèles entraînés pour le français comme CamemBERT (Martin et al., 2020) ou FlauBERT (Le et al., 2020) pourront également être intégrés.

Enfin, la fiabilité des informations contenues dans les données textuelles a un impact évident sur la pertinence des analyses effectuées. Il pourrait être judicieux d'intégrer dans les modèles d'analyse et/ou de prédiction appliqués à ce type de données des critères de qualité pondérant les sorties textuelles en fonction de la fiabilité de la source. Par exemple, l'intégration d'articles promotionnels dans les médias est une pratique courante, en particulier au Burkina Faso (Tiao, 2015). Ce type d'informations n'est pas pertinent pour nos analyses et en affectent probablement les résultats. Ces données devraient donc être retirées de nos analyses. Il serait judicieux de tenir compte de cet aspect dans des recherches futures, par exemple en recourant à des méthodes d'apprentissage automatique pour la détection de tels articles.

Conclusion générale

Dans cette thèse, à partir du contexte du Burkina Faso, nous avons d'une part sélectionné, traité et évalué les apports et limites d'un ensemble de données hétérogènes aux niveaux thématique, structurel et spatio-temporel afin d'appréhender la sécurité alimentaire de la manière la plus complète possible. Nous avons par ailleurs défini des méthodes originales de science des données permettant de traiter et de combiner ces données hétérogènes afin de fournir des informations prédictives et explicatives sur la sécurité alimentaire.

Dans le chapitre 1, nous avons passé en revue les indicateurs existants pour mesurer la sécurité alimentaire, puis nous avons analysé les contributions et les limites des indicateurs dérivés d'enquêtes ménages. Nous avons illustré nos propos avec le contexte du Burkina Faso, où une enquête ménages officielle (l'enquête permanente agricole) est conduite depuis une décennie et permet d'obtenir trois indicateurs : le SCA, le SDA et l'$ISAr$, qui fournissent des informations clés sur la sécurité alimentaire. Nous avons montré que ces indicateurs contiennent des informations spatiales et interannuelles cohérentes qui peuvent être exploitées pour le suivi des crises alimentaires au niveau sub-national. Plus précisément, nous avons constaté que ces enquêtes apportent des informations sur la sécurité alimentaire qui permettent d'identifier des tendances cohérentes avec les SSA, ainsi qu'avec des proxies climatiques et économiques liés à la sécurité alimentaire. Ces éléments ont été considérés tout en discutant des biais intrinsèques à ces données, dont une classification a été proposée.

Dans le chapitre 2, nous avons proposé le framework $FSPHD$ (Food Security Prediction based on Heterogeneous Data), qui exploite des méthodes d'apprentissage automatique pour la prédiction d'indicateurs clés de la sécurité alimentaire habituellement obtenus par le biais d'enquêtes ménages longues et coûteuses. Pour considérer un maximum de facteurs de sécurité alimentaire, nous avons intégré des données provenant de différentes

thématiques (structure du paysage, dynamique des populations, qualité des sols, météorologie, végétation, insécurité et économie), encodées selon différents types (valeurs quantitatives, données géolocalisées, vecteurs lignes, séries temporelles et images) et avec différentes granularités spatio-temporelles. Nous avons sélectionné des méthodes d'apprentissage automatique adaptées à chaque type de variable (e.g., FA sur les données CS, LSTM sur les séries temporelles, CNN sur les données HRS) et avons combiné ces méthodes en faisant appel au concept de fusion de données (e.g., aux niveaux des caractéristiques et des décisions). Nous avons montré que dans notre contexte, le CNN appliqué aux données HRS et la fusion des données au niveau des caractéristiques avec une FA sont des approches pertinentes. Nous avons par ailleurs constaté que les variables les plus déterminantes pour la prédiction des indicateurs de sécurité alimentaire proviennent de nombreux champs, ce qui souligne la pertinence de connecter ce domaine à un large éventail de disciplines liées. Bien que les performances atteintes par nos modèles soient perfectibles, les résultats de cette étude sont supérieurs à la plupart des travaux existants.

Dans le chapitre 3 nous avons examiné l'aptitude des méthodes de fouille de texte pour extraire des informations explicatives spatiales et temporelles sur la sécurité alimentaire à partir d'articles de journaux. Nous avons proposé, combiné et étendu de manière générique des méthodes de fouille de texte adaptées (le modèle Word2vec de plongement lexical, le modèle VADER d'analyse de sentiments et la méthode tf-idf de pondération de l'importance de termes). Cette démarche nous a permis de mettre en évidence trois types de proxies permettant d'obtenir des informations distinctes et complémentaires sur la thématique de la sécurité alimentaire : la proportion d'articles qui abordent la sécurité alimentaire, la proportion d'articles de polarité négative, ainsi que le vocabulaire spécifique employé dans les articles. Nous avons pris en compte l'aspect spatio-temporel de la sécurité alimentaire en effectuant des analyses ciblées aux niveaux global, régional et annuel. Les proxies de la sécurité alimentaire ont été agrégés à ces trois niveaux avec des méthodes d'agrégations appropriées (proportions, fréqences, tf-idf et ratio TIR). Nous avons proposé des représentations graphiques adaptées à chaque proxy et niveau d'analyse permettant de mettre en évidence les informations obtenues (nuages de mots, graphiques de séries de temps, graphiques radars et graphes). Cela nous a permis de faire ressortir des tendances alimentaires régionales et annuelles ainsi que les crises qui ont touché le pays au cours de la dernière décennie.

Cependant, nos travaux ont mis en évidence plusieurs limites des approches proposées

pour la compréhension de la sécurité alimentaire, tant au niveau explicatif (i.e., lié à l'extraction d'informations porteuses de sens), que structurel (i.e., lié à la variabilité spatiale) et conjoncturel (i.e., lié à la variabilité temporelle), ouvrant la voie à diverses perspectives de recherche.

Tout d'abord, la composante conjoncturelle de la sécurité alimentaire liée aux séries temporelles n'a pas pu être correctement prise en compte dans les analyses prédictives effectuées dans le chapitre 2. La méthode d'apprentissage profond exploitée au moyen d'un LSTM n'est pas parvenue à mettre en évidence les informations d'ordre séquentiel contenues dans les séries temporelles. Nous pensons que de futurs travaux devraient se diversifier sur d'autres méthodes pour améliorer la prise en compte de ce type de données. Il pourrait être pertinent de se tourner vers la théorie des motifs séquentiels (Masseglia et al., 2004), permettant de rechercher des informations fréquentes dans des données séquentielles, et nous pencher plus précisément sur une sous-branche de cette théorie qui s'intéresse aux motifs spatio-temporels, liés à des séquences temporelles (i.e., possédant un ordre chronologique d'apparition) (Andrienko et al., 2006). Les modèles spatio-temporels ont ainsi été adoptés pour de nombreux types de données associées à une temporalité : dans le cas de variables quantitatives (Kang and Yong, 2010), pour la gestion de relations qualitatives (Fabrègue et al., 2012), pour l'analyse de l'évolution temporelle dans des séries d'images satellites (Julea et al., 2010 ; Wu and Zhang, 2019) ou encore pour la classification de données textuelles (Yuan et al., 2018a). Un autre axe de recherche intéressant peut porter sur l'utilisation des réseaux complexes. Ce type de modèle permet de représenter de grandes quantités d'informations sous forme de graphes composés d'objets interconnectés par des liens et d'accéder à des informations sur les différentes interactions au moyen d'algorithmes dédiés (Van Steen, 2010). Cette approche a été appliquée au traitement de données temporelles dans des domaines liés à la sécurité alimentaire, notamment pour l'étude des migrations humaines (Davis et al., 2013) et des variations temporelles de l'occupation des sols (Zhang et al., 2019). Dans notre contexte, ce type de représentation des données pourrait être appliqué à différents niveaux. Par exemple, pour modéliser l'évolution dans des interactions individuelles (e.g., échanges sur des réseaux sociaux ou réseaux de solidarité dans un contexte de crise alimentaire), ou à des niveaux administratifs, en étudiant la quantité et la nature des échanges entre différentes localités. Notons que ces deux familles de méthodes peuvent également intégrer des informations spatiales, bien que moins fines que celles prises en compte par les CNN.

De plus, la composante structurelle de la sécurité alimentaire liée aux variables HRS a été prise en compte dans les analyses prédictives au moyen de la méthode d'apprentissage profond nommée CNN, et ont offert des performances de prédiction intéressantes. Cependant, l'interprétation des caractéristiques spatiales complexes obtenues par cette approche est difficile en raison de son effet "boîte noire" intrinsèque. Plusieurs méthodes ont récemment été développées pour interpréter les prédictions des modèles de réseaux de neurones, en identifiant les variables en entrée (e.g., pixels) qui contribuent le plus aux décisions de leur modèle (Montavon et al., 2018 ; Khormuji and Rostami, 2021) : l'analyse de sensibilité, basée sur le gradient du modèle évalué localement (i.e., dans le voisinage d'un vecteur de variables explicatives associé à une prédiction) ; l'analyse d'occlusion, qui mesure l'importance des variables explicatives en y introduisant du bruit et en mesurant ensuite l'impact sur la qualité des estimations ; la propagation de pertinence par couche (LRP) fonctionne en appliquant une décomposition de Taylor aux prédictions, permettant de propager l'information liée à la prédiction vers les neurones liés aux variables explicatives en entrée du modèle. L'avantage de ces méthodes est que leurs sorties peuvent être illustrées sur des cartes de chaleur, permettant une analyse assez visuelle et interprétable du rôle de chaque variable. Dans notre contexte, la mise en œuvre de ces méthodes est difficile car celles-ci sont adaptées aux problèmes de classification, et nécessitent donc des recherches intéressantes mais approfondies pour être adaptées à notre étude qui se concentre sur les régressions.

Un autre aspect concerne l'intégration de nouveaux types de données aux modèles prédictifs exposés dans le chapitre 2. Une perspective à moyen terme peut viser à exploiter les proxies textuels proposés dans le chapitre 3 (i.e., proportion d'articles qui abordent la sécurité alimentaire, proportion d'articles qui présentent un caractère négatif, vocabulaire spécifique employé) afin d'alimenter les modèles prédictifs (e.g., modèles basés sur des méthodes d'apprentissage automatique ou profond) présentés dans le chapitre 2. Ces proxies textuels pourraient fournir des informations complémentaires aux autres types de données utilisées conjointement (e.g., données quantitatives, images satellites) et ainsi augmenter la précision des prédictions des indicateurs de sécurité alimentaire. Cette tâche nécessite de trouver un corpus de journaux (ou d'autres médias textuels) garantissant une précision spatiale adéquate pour obtenir des proxies textuels à une échelle sub-nationale, mais nécessite aussi une étude complémentaire de validation des seuils permettant la construction des proxies textuels.

Il convient également d'évoquer la question relative à la qualité des données. Cette pro-

blématique concerne aussi bien les indicateurs de sécurité alimentaire utilisés comme variables réponses que les proxies considérés comme variables explicatives. Nous avons abordé cet aspect lors de la présentation des données, puis à plusieurs reprises afin de nuancer nos résultats. Cependant, nous n'avons pas proposé dans ce travail de méthodes pour prendre cela en compte. La qualité des données a un impact important sur les performances des modèles statistiques qui leur sont appliqués, en particulier pour les approches d'apprentissage automatique. Cette préoccupation était déjà partagée à la fin du siècle dernier (Cortes et al., 1995) et l'est encore aujourd'hui (Gupta et al., 2021). Actuellement, ce problème peut être maîtrisé dans un grand nombre de contextes grâce à la disponibilité de données de plus en plus volumineuses et à la puissance de traitement des machines permettant de les intégrer. En effet, dans le cas de l'apprentissage automatique, un grand volume de données implique davantage de données d'entraînement, ce qui réduit l'impact du bruit des données sur les performances (Brownlee, 2019). C'est sur ce principe que repose par exemple l'approche d'augmentation de données communément utilisée avec les méthodes d'apprentissage profond (Braun and Tashev, 2020). Mais dans notre contexte, où la quantité de données dépend de la couverture des enquêtes ménages réalisées par des moyens humains, la taille des données résultantes reste limitée. Dans ce cadre, les choix sont restreints. Tout d'abord, une approche élémentaire consiste au prétraitement avisé des données en appliquant les transformations adéquates et en supprimant les valeurs aberrantes de l'étude si elles existent (Gupta and Gupta, 2019). Cela confirme une fois encore la nécessité de maîtriser ce sujet en profondeur pour y apporter une vision orientée vers la science des données. Une autre option possible consiste à utiliser des méthodes peu sensibles au bruit dans les données et au surentraînement (Gupta and Gupta, 2019). Mais dans notre cas, l'utilisation de réseaux de neurones et de FA normalement peu sensibles au surentraînement ne suffit pas pour y faire pleinement face. Enfin, une méthode consiste à attribuer un poids à chaque observation en fonction de sa qualité pour moduler sa prise en compte par un modèle d'apprentissage automatique (Byrd and Lipton, 2019). Cependant, l'attribution d'un poids nécessite de disposer d'un critère permettant de détecter et de quantifier quelles observations peuvent être de mauvaise qualité. Cela n'est pas évident lorsqu'une variable présente à première vue des observations non aberrantes, bien que de mauvaise qualité. Pour détecter la qualité d'une observation sans a priori sur la donnée, une approche possible consiste à y appliquer une méthode d'apprentissage automatique ensembliste, afin de détecter les observations dont la valeur prédite varie fortement selon le modèle de la méthode ensembliste (Gupta and Gupta, 2019). Ces valeurs peuvent alors être considérées comme

problématiques. Le recours à cette approche est potentiellement intéressant pour notre contexte.

Une perspective essentielle porte sur la généralisabilité de nos résultats obtenus dans le contexte du Burkina Faso à d'autres terrains d'Afrique de l'Ouest voire à des régions plus lointaines. L'utilisation de données ouvertes et disponibles à l'échelle internationale pour la plupart rend possible cette transposition de nos modèles. Mais chaque région du monde est soumise à sa propre réalité climatique, politique, économique et sociale. Par conséquent, l'ensemble des variables les plus pertinentes pour mesurer la sécurité alimentaire varie selon les régions. Il pourrait être intéressant d'évaluer les variables et modèles identifiés dans cette thèse qui se généralisent le mieux. L'apprentissage par transfert est un domaine de recherche issu de l'apprentissage automatique qui vise à évaluer et à transférer les règles et les connaissances acquises par un modèle dans un contexte précis, afin de les appliquer à de nouvelles tâches ou à des domaines présentant des similitudes (Torrey and Shavlik, 2010). Cet aspect constitue une extension logique du travail réalisé dans cette thèse.

Enfin, n'oublions pas que malgré l'aspect exploratoire de cette thèse, l'un de ses objectifs est que ses résultats puissent conduire au développement d'outils opérationnels utiles. Si l'ouverture des données utilisées, ainsi que les premiers résultats encourageants obtenus et la disponibilité des modèles sous forme de programmes informatiques constituent un premier pas dans cette direction, il reste encore une couche de travail à accomplir, tant au niveau de la qualité des informations exploitables que de la simplicité et de la praticité d'utilisation par des experts non-initiés au code. Quoi qu'il en soit, nous espérons que ce travail sera poursuivi, complété, et débouchera sur des découvertes, qui apporteront directement ou indirectement un soutien substantiel aux acteurs qui réfléchissent et agissent pour réduire la faim.

Références

Adam Acar and Yuya Muraki. Twitter for crisis communication : lessons learned from japan's tsunami disaster. *Int. J. Web Based Communities*, 7(3) :392–402, 2011. doi : 10.1504/IJWBC.2011.041206.

Harold Alderman, Jere R Behrman, Hans-Peter Kohler, John A Maluccio, and Susan Cotts Watkins. Attrition in longitudinal household survey data : some tests for three developing-country samples. *Demographic research*, 5 :79–124, 2001.

Grégoire Allaire. *Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique*. Editions Ecole Polytechnique, 2005.

Jafar Alzubi, Anand Nayyar, and Akshi Kumar. Machine learning from theory to algorithms : an overview. In *Journal of physics : conference series*, volume 1142, page 012012. IOP Publishing, 2018.

Jean Joël Ambagna. *L'utilisation des enquêtes de conditions de vie des ménages pour l'analyse de la consommation alimentaire et de la sous-alimentation : illustrations sur les données camerounaises*. PhD thesis, Montpellier SupAgro, 2018.

Javeria Amin, Muhammad Sharif, Mudassar Raza, and Mussarat Yasmin. Detection of Brain Tumor based on Features Fusion and Machine Learning. *Journal of Ambient Intelligence and Humanized Computing*, 2018.

Gennady Andrienko, Donato Malerba, Michael May, and Maguelonne Teisseire. Mining spatio-temporal data, 2006.

Xiong Ao, Xin Yu, Derong Liu, and Hongkang Tian. News keywords extraction algorithm based on textrank and classified tf-idf. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 1364–1369, 2020. doi : 10.1109/IWCMC48107.2020.9148491.

Terri Ballard. *Household Hunger Scale : Indicator Definition and Measurement Guide*. Food and Nutrition Technical Assistance (FANTA), 2011.

Rommel Melgaço Barbosa and Donald R. Nelson. The Use of Support Vector Machine to Analyze Food Security in a Region of Brazil. *Applied Artificial Intelligence*, 30 : 318–330, 2016. ISSN 10876545.

Moumita Basu, Kripabandhu Ghosh, Somenath Das, Ratnadeep Dey, Somprakash Bandyopadhyay, and Saptarshi Ghosh. Identifying post-disaster resource needs and availabilities from microblogs. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 427–430, 2017.

Inbal Becker-Reshef, Christina Justice, Brian Barker, Michael Humber, Felix Rembold, Rogerio Bonifacio, Mario Zappacosta, Mike Budde, Tamuka Magadzire, Chris Shitote, Jonathan Pound, Alessandro Constantino, Catherine Nakalembe, Kenneth Mwangi, Shinichi Sobue, Terence Newby, Alyssa Whitcraft, Ian Jarvis, and James Verdin. Strengthening agricultural decisions in countries at risk of food insecurity : The geoglam crop monitor for early warning. *Remote Sensing of Environment*, 237 :111553, 2020.

Claudia Beleites, Richard Baumgartner, Christopher Bowman, Ray Somorjai, Gerald Steiner, Reiner Salzer, and Michael G Sowa. Variance reduction in estimating classification error using sparse datasets. *Chemometrics and intelligent laboratory systems*, 79(1-2) :91–100, 2005.

Paola Benedetti, Dino Ienco, Raffaele Gaetano, Kenji Osé, Ruggero Pensa, and Stéphane Dupuy. M3fusion : A deep learning architecture for multi-Scale/Modal/Temporal satellite data fusion, 2018. arXiv.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning : A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8) :1798–1828, 2013.

Oscar Bermeo Almeida, Del Cioppo Morstadt Javier, Mario Cardenas-Rodriguez, Roberto Cabezas-Cabezas, and William Bazán-Vera. Sentiment analysis in social networks for agricultural pests. *Advances in Intelligent Systems and Computing*, 901 : 122–129, 12 2018.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA :, 1994.

Michael W Berry and Jacob Kogan. Text mining. *Applications and Theory. West Sussex, PO19 8SQ, UK : John Wiley & Sons*, 2010.

Paul P. Biemer. Total survey error : Design, implementation, and evaluation. *Public Opinion Quarterly*, 74 :817–848, 2010. ISSN 0033362X.

Payal Biswas, Aditi Sharan, and Ashish Kumar. Agner : Entity tagger in agriculture domain. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1134–1138, 03 2015.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3 :993–1022, 2003.

Magdalena Bobe, Roxana Procopie, Mihaela Bucur, et al. Exploring the role of individual food security in the assessment of population's food safety. *Amfiteatru Econ. J*, 21 : 347–360, 2019.

Stefan Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 52–63, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-78135-6.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. TopicRank : Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

Sebastian Braun and Ivan Tashev. Data augmentation and loss normalization for deep noise suppression. In *International Conference on Speech and Computer*, pages 79–86. Springer, 2020.

Ramon F Brena, Antonio A. Aguileta, Luis A. Trejo, Erik Molino-Minero-Re, and Oscar Mayora. Choosing the Best Sensor Fusion Method : A Machine-Learning Approach. *Sensor*, 20, 2020.

Jason Brownlee. Impact of dataset size on deep learning model skill and performance estimates. *Machine Learning Mastery*, 6, 2019.

James J. Buckley and Yoichi Hayashi. Fuzzy neural networks : A survey. *Fuzzy Sets and Systems*, 66(1) :1–13, 1994. ISSN 0165-0114. doi : https://doi.org/10.1016/0165-0114(94)90297-6.

Bureau central du recensement général de l'agriculture. *Rapport général du module tronc commun ; Phase 2 RGA 2008*. Ministère de l'agriculture et de l'hydraulique du Burkina Faso, 2011.

Burkina Faso Government. *Inondations du 1er Septembre 2009 au Burkina Faso. Evaluation des dommages, pertes et besoins de construction, de reconstruction et de relèvement*. Burkina Faso Government, FAO, 2009.

Burkina Faso Government. *Bulletin agrométéorologique décadaire, période du 11 au 20 août 2013*. Direction Générale de la Météorologie, 2013.

Burkina Faso Government. *Bulletin agrométéorologique décadaire, période du 11 au 20 juillet 2014*. Direction Générale de la Météorologie, 2014.

Burkina Faso Government. *Bulletin agrométéorologique décadaire, période du 1er au 10 juin 2015*. Direction Générale de la Météorologie, 2015.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning ? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

Carlo Cafiero, Hugo R Melgar-Quinonez, Terri J Ballard, and Anne W Kepple. Validity and reliability of food security measures. *Annals of the New York Academy of Sciences*, 1331(1) :230–248, 2014.

Calogero Carletto, Alberto Zezza, and Raka Banerjee. Towards better measurement of household food security : Harmonizing indicators and the role of household surveys. *Global Food Security*, 2 :30–40, 2013. ISSN 22119124.

Lei Chai, Hongfeng Xu, Zhiming Luo, and Shaozi Li. A multi-source heterogeneous data analytic method for future price fluctuation prediction. *Neurocomputing*, 418 :11–20, 2020. ISSN 0925-2312. doi : https://doi.org/10.1016/j.neucom.2020.07.073.

Balasubramaniyan Chandrasekaran, Shruti Gangadhar, and James Conrad. A survey of multisensor fusion techniques, architectures and methodologies. In *SoutheastCon 2017*, pages 1–8. IEEE, 03 2017.

Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk prediction with electronic health records : A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 432–440. SIAM, 2016.

CHIRPS. Description and data of Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) product. `https://www.chc.ucsb.edu/data/chirps`, 2020. Accessed : 2020-04-01.

Jennifer Coates. Build it back better : Deconstructing food security for improved measurement and action. *Global Food Security*, 2 :188 – 194, 2013. ISSN 22119124.

Corinna Cortes, Lawrence D Jackel, Wan-Ping Chiang, et al. Limits on learning machine accuracy imposed by data quality. In *KDD*, volume 95, pages 57–62, 1995.

Laura Cruz, José Ochoa, Mathieu Roche, and Pascal Poncelet. Dictionary-based sentiment analysis applied to a specific domain. In *Information management and big data*, pages 57–68. Springer, 2015.

Kyle F Davis, Paolo D'Odorico, Francesco Laio, and Luca Ridolfi. Global spatio-temporal patterns in human migration : a complex network perspective. *PloS one*, 8(1) :e53723, 2013.

Melissa Luciana De Araujo, Raquel de Deus Mendonça, José Divino Lopes Filho, and Aline Cristine Souza Lopes. Association between food insecurity and food intake. *Nutrition*, 54 :54–59, 2018. ISSN 18731244.

Adeline Decuyper, A. Rutherford, Amit Wadhwa, J. Bauer, G. Krings, T. Gutierrez, V. Blondel, and M. Luengo-Oroz. Estimating Food Consumption and Poverty indices with Mobile Phone Data. Technical paper, Global Pulse, 2014.

DeepL GmbH. Deepl website. `https://www.deepl.com/translator`, 2017. Accessed : 2021-06-07.

Hervé Delacour and Aurélie Servonnet. La courbe roc (receiver operating characteristic) : principes et principales applications en biologie clinique. In *Annales de biologie clinique*, volume 63, pages 145–154, 2005.

Hugo Deléglise, Agnès Bégué, Roberto Interdonato, Elodie Maître d'Hôtel, Mathieu Roche, and Maguelonne Teisseire. Linking heterogeneous data for food security prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 335–344. Springer, 2020.

Hugo Deléglise, Agnès Bégué, Roberto Interdonato, Elodie Maitre D'hotel, and Maguelonne Teisseire. Suivi de la sécurité alimentaire en Afrique de l'Ouest : Quelles méthodes d'analyse de données pour traiter l'interdisciplinarité de la sécurité alimentaire. *Journal of Interdisciplinary Methodologies and Issues in Sciences*, Agriculture Numérique en Afrique, 04 2021a. doi : 10.18713/JIMIS-120221-8-3. URL `https://jimis.episciences.org/7063`.

Hugo Deléglise, Camille Schaeffer, Elodie Maître d'Hôtel, and Agnès Bégué. Lexiques en français sur la sécurité alimentaire et les crises, 2021b. URL `https://doi.org/10.18167/DVN1/C5PU01`. Dataverse CIRAD.

Hugo Deléglise, Camille Schaeffer, Elodie Maître d'Hôtel, Agnès Bégué, Mathieu Roche, Roberto Interdonato, and Maguelonne Teisseire. Corpus de journaux burkinabés en français sur la sécurité alimentaire publiés entre 2009 et 2018, 2021c. URL `https://doi.org/10.18167/DVN1/IVVEQL`. Dataverse CIRAD.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

Juglar Diaz, Barbara Poblete, and Felipe Bravo-Marquez. An integrated model for textual social media data with spatio-temporal dimensions. *Information Processing & Management*, 57(5) :102219, 2020. ISSN 0306-4573. doi : https://doi.org/10.1016/j.ipm.2020.102219.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8 :439–453, 07 2020.

Brett Drury and Mathieu Roche. A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture*, 163 :104864, 2019. ISSN 0168-1699. doi : https://doi.org/10.1016/j.compag.2019.104864.

Anne-Marie Dussaix. La qualite dans les enquêtes. *MODULAD*, 39, 2009.

Sean R Eddy. What is a hidden markov model ? *Nature biotechnology*, 22(10) :1315–1316, 2004.

ESA. Description and data of S2 prototype Land Cover map at 20m of Africa 2016. `http://2016africalandcover20m.esrin.esa.int/`, 2020. Accessed : 2020-04-01.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639) :115–118, 2017.

Mickaël Fabrègue, Agnès Braud, Sandra Bringay, Florence Le Ber, and Maguelonne Teisseire. Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives. In *Inforsid*, pages 123–138, 2012.

FAO. *Vue d'ensemble régionale de la sécurité alimentaire et la nutrition. Le lien entre les conflits et la sécurité alimentaire et la nutrition : renforcer la résilience pour sécurité alimentaire, la nutrition et la paix.* FAO, 2017.

FAO. *Burkina Faso : Aperçu de la réponse - juillet 2019.* FAO, 2019.

FAO and ECA. *Addressing the threat from climate variability and extremes for food security and nutrition.* FAO, 2018. ISBN 9789251311578.

FAO, FIDA, OMS, WFP, and UNICEF. *L'état de la sécurité alimentaire et de la nutrition dans le monde en 2018 : renforcer la Résilience face aux changements climatiques pour La sécurité alimentaire et la nutrition.* FAO, 2018. ISBN 978-92-5-130840-0.

FAO, FIDA, OMS, WFP, and UNICEF. *The State of Food Security and Nutrition in the World - Transforming Food Systems for Affordable Healthy Diets.* FAO, 2020. ISBN 978-92-5-132901-6.

Ronen Feldman, Moshe Fresko, Yakkov Kinar, Yehuda Lindell, Orly Liphstat, Martin Rajman, Yonatan Schler, and Oren Zamir. Text mining at the term level. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 65–73. Springer, 1998.

Jacques Fize, Mathieu Roche, and Maguelonne Teisseire. Spatial textual representation (str) ou comment représenter la spatialité des données textuelles. In *HAL*. Université de Rouen Normandie, 2017.

Joseph L. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3) :613–619, 1973. doi : 10.1177/001316447303300309.

Clifton Forlines and Kent Wittenburg. Wakame : Sense making of multi-dimensional spatial-temporal data. In *Proceedings of the International Conference on Advanced Visual Interfaces, AVI 2010*, pages 33–40, 01 2010. doi : 10.1145/1842993.1843000.

Steffen Fritz, Linda See, and Juan Carlos Laso Bayas. A comparison of global agricultural monitoring systems and current gaps. *Agricultural Systems*, 168 :258–272, 2019. ISSN 0308521X.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, page 1606–1611. Morgan Kaufmann Publishers Inc., 2007.

Matt Gardner and Stephen Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15) :2627–2636, 1998.

Noor Ghazal-Aswad. Biased neutrality : the symbolic construction of the syrian refugee in the new york times. *Critical Studies in Media Communication*, 36(4) :357–375, 2019.

Rodolphe Ghiglione and Benjamin Matalon. Comment interroger. *Les questionnaires (Chapitre 4) In Les enquêtes sociologiques théories et pratiques*, pages 93–138, 1998.

CJ Hutto Eric Gilbert. Vader : A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14).*, 2014.

Peter Glick. How reliable are surveys of client satisfaction with healthcare services ? evidence from matched facility and household data in madagascar. *Social science & medicine*, 68(2) :368–379, 2009.

Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3) :659–678, 2017.

Kehua Guo, Tao Xu, Xiaoyan Kui, Ruifang Zhang, and Tao Chi. iFusion : Towards efficient intelligence fusion for deep learning from real-time and heterogeneous data. *Information fusion*, 51, 2019.

Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4040–4041, 2021.

Shivani Gupta and Atul Gupta. Dealing with noise problem in machine learning datasets : A systematic review. *Procedia Computer Science*, 161 :466–474, 2019.

Habib Hadj-Mabrouk. *Apprentissage automatique et acquisition des connaissances : deux approches complémentaires pour les systèmes à base de connaissances. Application au système" ACASYA" d'aide à la certification des systèmes de transport automatisés.* PhD thesis, Université Polytechnique Hauts-de-France, Université de Valenciennes, 1992.

David Hall and James Llinas. *Handbook of Multisensor Data Fusion : Theory and Practice.* CRC Press, 2017. ISBN 9781315219486.

Lieve Hamers et al. Similarity measures in scientometric research : The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3) :315–18, 1989.

Yanling Han, Yekun Liu, Zhonghua Hong, Yun Zhang, Shuhu Yang, and Jing Wang. Sea ice image classification based on heterogeneous data fusion and deep learning. *Remote Sensing*, 13(4) :592, 2021.

Kazuyuki Hara, Daisuke Saito, and Hayaru Shouno. Analysis of function of rectified linear unit used in deep learning. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2015.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty : Regret bounded by variation in costs. *Machine learning*, 80(2) :165–188, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9) :1904–1916, 2015.

Derek Headey and Olivier Ecker. Rethinking the measurement of food security : from first principles to best practice. *Food security*, 5(3) :327–343, 2013.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4) : 18–28, 1998.

Gernot Heisenberg, Lima Filho da Rocha, L. Caspersen, and S. Wöhrle. Deep learning approach for the prediction of food insecurity. *4th Global Food Security conference*, 2020.

Adrián Hernández and José M Amigó. Attention mechanisms and their applications to complex systems. *Entropy*, 23(3) :283, 2021.

Rémy Herrera and Laurent Ilboudo. *Les défis de l'agriculture paysanne : Le cas du burkina faso*. L'Harmattan, 2012. ISBN 9782336004495.

Sung Yang Ho, Sophia Tan, Chun Chau Sze, Limsoon Wong, and Wilson Wen Bin Goh. What can venn diagrams teach us about doing data science better ? *International Journal of Data Science and Analytics*, 11(1) :1–10, 2021.

John Hoddinott. *Choosing Outcome Indicators Of Household Food Security, Vol. Technical Guide No 7*. International Food Policy Research Institute, 1999.

Geoff Hollis and Chris Westbury. The principals of meaning : Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23, 05 2016. doi : 10.3758/s13423-016-1053-2.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017.

Bo Huang, Bei Zhao, and Yimeng Song. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214 :73 – 86, 2018.

Kaizhu Huang, Amir Hussain, Qiu-Feng Wang, and Rui Zhang. *Deep learning : fundamentals, theory and applications*, volume 2. Springer, 2019.

INDDEX Project. *Data4Diets : Building Blocks for Diet-related Food Security Analysis*. Tufts University, Boston, 2018.

Roberto Interdonato, Jean-Loup Guillaume, and Antoine Doucet. A lightweight and multilingual framework for crisis information extraction from twitter data. *Social Network Analysis and Mining*, 9(1), 2019.

International Food Policy Research Institute. *The concepts of the Global Hunger Index*. International Food Policy Research Institute, 2017.

Uzma Iram and Muhammad S Butt. Determinants of household food security : An empirical analysis for pakistan. *International Journal of Social Economics*, 2004.

Masahiko Itoh, Naoki Yoshinaga, and Masashi Toyoda. Spatio-temporal event visualization from a geo-parsed microblog stream. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, pages 58–61, 2016.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28 :2017–2025, 2015.

Anil Jain, Richard Dubes, and Chaur-Chin Chen. Bootstrap techniques for error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 9 :628 – 633, 10 1987. doi : 10.1109/TPAMI.1987.4767957.

Beakcheol Jang, Inhwan Kim, and Jong Wook Kim. Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8) :e0220976, 2019.

Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353 :790–794, 2016.

Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, and Eivind Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28 :21–28, 2001.

Andrew D. Jones, Francis M. Nguren, Gretel Pelto, and Sera L. Young. What Are We Assessing When We Measure Food Security ? A Compendium and Review of Current Metrics. *Advances in Nutrition*, 4 :481–505, 2013. ISSN 0022-3166.

Andreea Julea, Nicolas Méger, Philippe Bolon, Christophe Rigotti, Marie-Pierre Doin, Cécile Lasserre, Emmanuel Trouvé, and Vasile N Lăzărescu. Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns. *IEEE Transactions on Geoscience and Remote Sensing*, 49(4) :1417–1430, 2010.

Ju-Young Kang and Hwan-Seung Yong. Mining spatio-temporal patterns in trajectory data. *Journal of Information Processing Systems*, 6(4) :521–536, 2010.

Sheila Medina Karambiri. *La gouvernance territoriale par les chartes foncières locales dans la région des hauts bassins/burkina faso.* PhD thesis, Université Paul Valéry Montpellier 3, 2018.

Gina Kennedy, Andrea Berardo, Cinzia Papavero, Peter Horjus, Terri Ballard, Marie-Claude Dop, Jan Delbaere, and Inge D Brouwer. Proxy measures of household food consumption for food security assessment and surveillance : comparison of the household dietary diversity and food consumption scores. *Public health nutrition*, 13(12) : 2010–2018, 2010.

Gina Kennedy, Terri Ballard, and Marie-Claude Dop. *Guide pour mesurer la diversité alimentaire au niveau du ménage et de l'individu.* FAO, 2013. Technical report.

Eric Kergosien, Hugo Alatrista-Salas, Mauro Gaio, Fábio N Güttler, Mathieu Roche, and Maguelonne Teisseire. When textual information becomes spatial information compatible with satellite images. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 301–306. IEEE, 2015.

Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion : A review of the state-of-the-art. *Inf. Fusion*, 14(1) :28–44, January 2013.

Hamed Behzadi Khormuji and H. Rostami. Fast multi-resolution occlusion : a method for explaining and understanding deep neural networks. *Applied Intelligence*, 51 : 2431–2455, 2021.

Kanwal Kiani and Khalid Saleem. K-nearest temperature trends : A method for weather temperature data imputation. In *Proceedings of the 2017 International Conference on Information System and Data Mining*, pages 23–27, 2017.

Jaewoo Kim, Meeyoung Cha, and Jong Lee. Nowcasting commodity prices using social media. *PeerJ Computer Science*, 3 :e126, 07 2017. doi : 10.7717/peerj-cs.126.

Diederik P Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

Akshi Kumar, Kathiravan Srinivasan, Wen-Huang Cheng, and Albert Y. Zomaya. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, 57(1) : 102141, 2020. ISSN 0306-4573. doi : https://doi.org/10.1016/j.ipm.2019.102141.

Wolfgang Kunz, João Marques-Silva, and Sharad Malik. Sat and atpg : Algorithms for boolean decision problems. In *Logic synthesis and Verification*, pages 309–341. Springer, 2002.

Innocent Kutyauripo, Nyaradzo Prisca Mavodza, and Christopher Tafara Gadzirayi. Media coverage on food security and climate-smart agriculture : A case study of newspapers in zimbabwe. *Cogent Food & Agriculture*, 7(1), 2021. doi : 10.1080/23311932.2021.1927561.

Wolfram Lacher. *Organized crime and conflict in the sahel-sahara region.* Carnegie Endowment for International Peace, 2012. Technical report.

Ting Lan, Hui Hu, Chunhua Jiang, Guobin Yang, and Zhengyu Zhao. A comparative study of decision tree, random forest, and convolutional neural network for spread-f identification. *Advances in Space Research*, 65(8) :2052–2061, 2020.

Véronique Lassailly-Jacob. Inondations de 2009 et 2010 au burkina faso. *Mobilité humaine et environnement*, 2015.

Lawrence R Lawlor. Overlap, similarity, and competition coefficients. *Ecology*, 61(2) : 245–251, 1980.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert : Unsupervised language model pre-training for french, 2020.

Le Hub Rural. Afrique de l'ouest : Sahel - agrhymet redoute la reproduction du criquet pèlerin. http://www.hubrural.org/Afrique-de-l-Ouest-Sahel-AGRHYMET.html?lang=fr, 2013. Accessed : 2021-06-07.

Le Kiosque Digital du Burkina. Burkina faso : statistiques sur le digital et les médias sociaux en janvier 2020. https://lekiosquedigitalduburkina.com/2020/02/24/burkina-faso-statistiques-sur-le-digital-et-les-medias-sociaux-en-janvier-2020/, 2020. Accessed : 2021-06-07.

Le Monde. La mutinerie de militaires gagne une quatrième ville du burkina faso, des jeunes manifestent. https://www.lemonde.fr/afrique/article/2011/04/18/la-mutinerie-de-militaires-gagne-une-quatrieme-ville-du-burkina-faso_1509237_3212.html, 2011. Accessed : 2021-06-07.

Philippe Lenca, Benoit Vaillant, Patrick Meyer, and Stéphane Lallich. Association rule interestingness measures : Experimental and theoretical studies. In *Quality Measures in Data Mining*, pages 51–76. Springer, 2007.

Erin Lentz, Hope Michelson, Kathy Baylis, and Yujun Zhou. A data-driven approach improves food insecurity crisis prediction. *World Development*, 122 :399 – 409, 2019.

Christina Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. *Advances in neural information processing systems*, pages 1441–1448, 2003.

Germana H. Leyna, Elia J. Mmbaga, Kagoma S. Mnyika, Akhtar Hussain, and Knut Inge Klepp. Food insecurity is associated with food consumption patterns and anthropometric measures but not serum micronutrient levels in adults in rural Tanzania. *Public Health Nutrition*, 13 :1438–1444, 2010. ISSN 13689800.

Jun Li, Zhi He, Javier Plaza, Shutao Li, Jinfen Chen, Henglin Wu, Yandong Wang, and Yu Liu. Social media : New perspectives to improve remote sensing for emergency response. *Proceedings of the IEEE*, 105(10) :1900–1912, 2017. doi : 10.1109/JPROC.2017.2684460.

Wei-Ting Liao, Luis F Rodríguez, Jana Diesner, and Tao Lin. Improving farm management optimization : Application of text data analysis and semantic networks. In *2015 ASABE Annual International Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2015.

Miao Liu, Mingjun Wang, Jun Wang, and Duo Li. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification : Application to the recognition of orange beverage and chinese vinegar. *Sensors and Actuators B : Chemical*, 177 :970–980, 2013.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

Cédric Lopez, Violaine Prince, and Mathieu Roche. Titrage automatique de documents électroniques par extraction de syntagmes nominaux. In *21èmes Journées Francophones d'Ingénierie des Connaissances*, pages 17–28, France, April 2010.

Andrew Lukyamuzi, John Ngubiri, and Washington Okori. Towards harnessing phone messages and telephone conversations for prediction of food crisis. *International Journal of System Dynamics Applications*, 4(4) :1–16, 2015. doi : 10.4018/IJSDA.2015100101.

Andrew Lukyamuzi, John Ngubiri, and Washington Okori. Tracking food insecurity from tweets using data mining techniques. In *Proceedings of the 2018 International Conference on Software Engineering in Africa - SEiA '18*, pages 27–34, 2018.

CS Malarkodi, Elisabeth Lex, and Lalitha Devi Sobha. Named entity recognition for the agricultural domain. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016) ; Research in Computing Science*, 2016.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi : 10.18653/v1/2020.acl-main.645. URL `http://dx.doi.org/10.18653/v1/2020.acl-main.645`.

Pierre Martin, Thierry Helmer, Julien Rabatel, and Mathieu Roche. Keops : Knowledge extractor pipeline system. In Samira Cherfi, Anna Perini, and Selmin Nurcan, editors, *Research Challenges in Information Science*, pages 561–567, Cham, 2021. Springer International Publishing.

Florent Masseglia, Maguelonne Teisseire, and Pascal Poncelet. Extraction de motifs séquentiels. problemes et méthodes. *Revue des Sciences et Technologies de l'Information-Série ISI : Ingénierie des Systèmes d'Information*, 9(3/4) :183–210, 2004.

Daniel Maxwell. *The Coping Strategies Index ; A tool for measurement of household food security and the impact of aid programs in humanitarian emergency ; Field Method Manual*. Feinstein International Center, 2008.

Daniel Maxwell, Bapu Vaitla, and Jennifer Coates. How do indicators of household food insecurity measure up ? An empirical comparison from Ethiopia. *Food Policy*, 47 : 107–116, 2014. ISSN 03069192.

Kathleen McKeown, Frank Smadja, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics*, 22(1), 1996.

Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06 : Proceedings of the 15th international conference on World Wide Web*, page 533–542. Association for Computing Machinery, 2006. ISBN 1595933239. doi : 10.1145/1135777.1135857.

Hugo Melgar-Quinonez and Michelle Hackett. Measuring household food security : the global experience. *Revista de Nutrição*, 21 :27s–37s, 2008.

Stuart E. Middleton, Lee Middleton, and Stefano Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2) :9–17, 2014. doi : 10.1109/MIS.2013.126.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

Wen Min, Li Ping, Zhang Lingfei, and Chen Yan. Stock market trend prediction using high-order information of time series. *IEEE Access*, 7 :299–308, 2019.

Riccardo Miotto, Fei Wang, Shuang Wang, and Xiaoqian Jiang. Deep learning for healthcare : review, opportunities and challenges. *Briefings in bioinformatics*, 19 :1236—1246, 05 2017.

Ana Moltedo, Nathalie Troubat, Michael Lokshin, and Zurab Sajaia. *Analyzing Food Security Using Household Survey Data : Streamlined Analysis with ADePT Software.* World Bank, 2014. ISBN 978-1-4648-0140-2.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73 :1–15, 2018. ISSN 1051-2004. doi : https://doi.org/10.1016/j.dsp.2017.10.011.

Ali Mumtaz, C. Deo Ravinesh, J. Downs Nathan, and Maraseni Tek. Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting. *Computers and Electronics in Agriculture*, 152 :149–165, 2018.

Nico JD Nagelkerke et al. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3) :691–692, 1991.

NASA. Description and data of MODerate resolution Imaging Spectroradiometer (MODIS) satellite images of composite Normalized Difference Vegetation Index (NDVI) (MOD13Q1) product. `https://lpdaac.usgs.gov/products/mod13q1v006/`, 2020. Accessed : 2020-04-01.

Nathalie Neptune and Josiane Mothe. Automatic annotation of change detection images. *Sensors*, 21(4) :1110, 2021.

OCHA. *Burkina Faso ; Inondations ; période du 1er au 30 septembre 2010.* United Nations Office for the Coordination of Humanitarian Affairs (OCHA), 2010.

OCHA. *Plan de réponse stratégique - Burkina Faso.* United Nations Office for the Coordination of Humanitarian Affairs (OCHA), 2015.

Washington Okori and Joseph Obua. Supervised Learning Algorithms For Famine Prediction. *Applied Artificial Intelligence*, 25 :822–835, 2011. ISSN 2078-0958.

ONU Info. Le hcr annonce un nombre sans précédent de personnes déracinées en 2013. `https://news.un.org/fr/story/2013/12/280332-le-hcr-annonce-un-nombre-sans-precedent-de-personnes-deracinees-en-2013`, 2013. Accessed : 2021-06-07.

Angel R Ortiz, Charlie EM Strauss, and Osvaldo Olmea. Mammoth (matching molecular models obtained from theory) : an automated method for model comparison. *Protein Science*, 11(11) :2606–2621, 2002.

Darrin P. Lewis, Tony Jebara, and William Stafford Noble. Support vector machine learning from heterogeneous data : an empirical analysis using protein sequence and structure. *Bioinformatics*, 22 :2753–2760, 2006.

Vedhas Pandit, Shahin Amiriparian, Maximilian Schmitt, Amr Mousa, and Björn Schuller. Big data multimedia mining : feature extraction facing volume, velocity, and variety. *Big Data Analytics for Large-Scale Multimedia Search*, 61, 2019.

Ron Papka, James Allan, et al. On-line new event detection using single pass clustering. *University of Massachusetts, Amherst*, 10(290941.290954), 1998.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Permanent Agricultural Survey. *Résultats Défnitifs de la Campagne Agricole 2014/2015 et Perspectives de la Situation Alimentaire et Nutritionnelle*. Ministère de l'Agriculture, des Ressources Hydrauliques, de l'Assainissement et la Sécurité Alimentaire, 2015.

Kyle T. Peterson, Vasit Sagan, Paheding Sidike, Elizabeth A. Hasenmueller, John J. Sloan, and Jason H. Knouft. Machine Learning-Based Ensemble Prediction of Water-quality Variables Using Feature-level and Decision-level Fusion with Proximal Remote Sensing. *Photogrammetric Engineering & Remote Sensing*, 85 :269–280, 2018.

Van Hiep Phung, Eun Joo Rhee, et al. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9(21) :4500, 2019.

Rafael Pérez-Escamilla and Ana Maria Segall-Corrêta. Food insecurity measurement and indicators. *Revista de Nutricao*, 21 :15s – 26s, 2008. ISSN 14155273.

Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer, 2012.

Aditya Ramana Rachakonda, Srinath Srinivasa, Sumant Kulkarni, and M.S. Srinivasan. A generic framework and methodology for extracting semantics from co-occurrences. *Data & Knowledge Engineering*, 92 :39–59, 2014. ISSN 0169-023X. doi : https://doi.org/10.1016/j.datak.2014.06.002.

Hassan Ramchoun, Mohammed Amine Janati Idrissi, Youssef Ghanou, and Mohamed Ettaouil. Multilayer perceptron : Architecture optimization and training. *IJIMAI*, 4 (1) :26–30, 2016.

Faneva Ramiandrisoa and Josiane Mothe. Extraction automatique de termes-clés : Comparaison de méthodes non supervisées. In *Conférence en Recherche d'Informations et Applications - Rencontres Jeunes Chercheurs en Recherche d'Information (RJCRI CORIA 2016)*, pages 315–323, Toulouse, FR, 2016. Association Francophone de Recherche d'Information et Applications (ARIA). doi : 10.24348/SDNRI.2016.RJC7.

Sarunas Raudys. On the accuracy of a bootstrap estimate of the classification error. In *9th International Conference on Pattern Recognition*, pages 1230–1231. IEEE Computer Society, 1988.

Garima Rautela, Mohammed K. Ali, Dorairaj Prabhakaran, K. M.Venkat Narayan, Nikhil Tandon, Viswanathan Mohan, and Lindsay M. Jaacks. Prevalence and correlates of household food insecurity in Delhi and Chennai, India. *Food Security*, 12 :1–14, 2020. ISSN 18764525.

Nico Reski, Aris Alissandrakis, and Andreas Kerren. Exploration of time-oriented data in immersive virtual reality using a 3d radar chart approach. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction : Shaping Experiences, Shaping Society*, pages 1–11, 10 2020. doi : 10.1145/3419249.3420171.

Mathieu Roche, Sophie Fortuno, Juan Antonio Lossio-Ventura, Amira Akli, Salim Belkebir, Thinhinan Lounis, and Serigne Toure. Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. *Cahiers Agricultures*, 24(5) :313–320, 2015.

François Rousseau and Michalis Vazirgiannis. Main core retention on graph-of-words for single-document keyword extraction. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, pages 382–393, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16354-3.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5) :513–523, 1988. ISSN 0306-4573. doi : https://doi.org/10.1016/0306-4573(88)90021-0. URL `https://www.sciencedirect.com/science/article/pii/0306457388900210`.

Maurice Schiff and Alberto Valdes. Poverty, food intake, and malnutrition : implications for food security in developing countries. *American Journal of Agricultural Economics*, 72(5) :1318–1322, 1990.

Pandey Shailesh, Agarwal Tushar, and Krishnan Narayanan. Multi-Task Deep Learning for Predicting Poverty from Satellite Images (IAAI18). *The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence*, 2018.

Khader Shameer, Marcus A Badgeley, Riccardo Miotto, Benjamin S Glicksberg, Joseph W Morgan, and Joel T Dudley. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Briefings in bioinformatics*, 18(1) :105–124, 2017.

John Shaw. *World Food Security : A History Since 1945*, volume 1. Palgrave MacMillan, 2007. ISBN 10 : 0230553559.

Evan Sheehan. Utilizing latent embeddings of wikipedia articles to predict poverty. In *Stanford University*, 2018.

Dewi Sinta, Hari Wijayanto, and BJAMS Sartono. Ensemble k-nearest neighbors method to predict rice price in indonesia. *Appl. Math. Sci.*, 8(160) :7993–8005, 2014.

Lisa Smith and Ali Subandoro. *Measuring Food Security Using Household Expenditure Surveys*. International food policy research institute (IFPRI), 2007. ISBN 9780896297678.

Paul Smolensky. Information processing in dynamical systems : Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

Xuanyi Song, Yuetian Liu, Liang Xue, Jun Wang, Jingzhe Zhang, Junqiang Wang, Long Jiang, and Ziyan Cheng. Time-series well performance prediction based on long short-term memory (lstm) neural network model. *Journal of Petroleum Science and Engineering*, 186, 2020.

Trevor Strohman, Donald Metzler, Howard Turtle, and W. Croft. Indri : A language-model based search engine for complex queries. *Information Retrieval - IR*, 01 2005.

Vairavasundaram Subramaniyaswamy, Ravi Logesh, M Abejith, Sunil Umasankar, and A Umamakeswari. Sentiment analysis of tweets for estimating criticality and security of events. In *Improving the Safety and Efficiency of Emergency Services : Emerging Tools and Technologies for First Responders*, pages 293–319. IGI global, 2020.

Isti Surjandari, Muthia Naffisah, and M. Prawiradinata. Text mining of twitter data for public sentiment analysis of staple foods price changes. *Journal of Industrial and Intelligent Information*, 3, 01 2014. doi : 10.12720/jiii.3.3.253-257.

Anne Swindale and Paula Bilinsky. *Household Dietary Diversity Score (HDDS) for measurement of household food access : Indicator guide.* Food and Nutrition Technical Assistance (FANTA), 2006.

Anna Szabolcsi. Positive polarity - negative polarity. *Natural Language and Linguistic Theory*, 22(2) :409–452, May 2004. ISSN 0167-806X. doi : 10.1023/B:NALA.0000015791.00288.43.

Alexandra Tapsoba, Pascale Combes Motel, and Jean-louis Combes. Remittances , food security and climate variability : The case of Burkina Faso. Working papers, HAL, 2019.

The Economist Intelligence Unit. *Global Food Security Index 2018.* The Economist Intelligence Unit, 2018.

Beyon Luc Adolphe Tiao. *Régulation des médias d'Afrique francophone : cas du Burkina Faso.* PhD thesis, Université Michel de Montaigne-Bordeaux III, 2015.

Harmonie Toros. *Informal Governance of Non-State Armed Groups in the Sahel.* NATO Strategic Direction South Hub, 2019.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends : algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

Cascade Tuholske, Kwaw Andam, Jordan Blekking, Tom Evans, and Kelly Caylor. Comparing measures of urban food security in Accra, Ghana. *Food Security*, 12 :1299–1316, 2020. ISSN 18764525.

UN Global Pulse. Mining Indonesian Tweets to Understand Food Price Crises. Methods paper, Global Pulse, 2014.

Julio J. Valdés. Extreme learning machines with heterogeneous data types. *Neurocomputing*, 277 :38–52, 2018. ISSN 18728286.

Sarah Valentin, Renaud Lancelot, and Mathieu Roche. Identifying associations between epidemiological entities in news data for animal disease surveillance. *Artificial Intelligence in Agriculture*, 5 :163–174, 2021a.

Sarah Valentin, Alizé Mercier, Renaud Lancelot, Mathieu Roche, and Elena Arsevska. Monitoring online media reports for early detection of unknown diseases : Insight from a retrospective study of covid-19 emergence. *Transboundary and emerging diseases*, 68(3) :981–986, 2021b.

Wesley van der Heijden, Marc van den Homberg, Martijn Marijnis, Marijke de Graaff, and Hennie Daniels. Combining Open Data and Machine Learning to predict Food Security in Ethiopia. In *UNESCO Chair in Technologies for Development : Voices of the Global South*, 2018.

Maarten Van Steen. Graph theory and complex networks. *An introduction*, 144, 2010.

Elliot Vhurumuku. Food security indicators - WFP. *Integrating Nutrition and Food Security Programming for Emergency response workshop*, 2014.

Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 855–860. AAAI Press, 2008. ISBN 9781577353683.

Pei-Ying Wang, Chiao-Ting Chen, Jain-Wun Su, Ting-Yun Wang, and Szu-Hao Huang. Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. *IEEE Access*, 9 :55244–55259, 2021.

Qian Wang, Cheng Wang, ZY Feng, and Jin-feng Ye. Review of k-means clustering algorithm. *Electronic Design Engineering*, 20(7) :21–24, 2012.

Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization : A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6) :1336–1353, 2012.

Zihuan Wang, Kyu S. Hahn, Youngsam Kim, Sanghyup Song, and Jong-Mo Seo. A news-topic recommender system based on keywords extraction. *Multimedia Tools and Applications*, 77 :4339–4353, 2017.

Nathan Wanner, Carlo Cafiero, Nathalie Troubat, and Piero Conforti. *Refinements To the Fao Methodology for Estimating the Prevalence of undernourishment indicator*. FAO, 2014.

D.C. Webb. *ECHELON and the NSA*, pages 453–468. IGI Global, 01 2007. ISBN 9781591409922. doi : 10.4018/978-1-59140-991-5.ch053.

WFP. *Comprehensive Food Security and Vulnerability Analysis Guidelines*. WFP, 2009. Technical.

WFP. *Burkina Faso ; Avril 2012 : Rapport d'évaluation approfondie sur la sécurité alimentaire des ménages en situation d'urgence (EFSA) dans 170 communes déclarées à risque d'insécurité alimentaire*. WFP, 2012.

WFP. *Burkina Faso : Analyse Globale de la Vulnérabilité, de la Sécurité Alimentaire et de la Nutrition*. WFP, 2014a.

WFP. *Burkina Faso : Analyse Globale de la Vulnérabilité, de la Sécurité Alimentaire et de la Nutrition*. WFP, 2014b.

WFP-VAM. Humanitarian High Resolution Mapping - Complementing Assessments with Remote Sensing Open Data. `https://wfp-vam.github.io/HRM/`, 2019. [Online ; accessed 28-January-2021].

John S Whissell and Charles LA Clarke. Improving document clustering using okapi bm25 feature weighting. *Information retrieval*, 14(5) :466–487, 2011.

Jannike Wichern, Joost van Heerwaarden, Sytze de Bruin, Katrien Descheemaeker, Piet JA van Asten, Ken E Giller, and Mark T van Wijk. Using household survey data to identify large-scale food security patterns across uganda. *PloS one*, 13(12) : e0208714, 2018.

Doris Wiesmann, Lucy Bassett, Todd Benson, and John Hoddinott. *Validation of the world food programme's food consumption score and alternative indicators of household food security*. International Food Policy Research Institute (IFPRI), 2009.

Leland Wilkinson. Tree structured data analysis : Aid, chaid and cart. *Retrieved February*, 1 :2008, 1992.

Joachim Winter. Response bias in survey based measures of household consumption. *Economics Bulletin*, 3 :1–12, 2004.

Pak Chung Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas. Visualizing sequential patterns for text mining. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 105–111, 2000. doi : 10.1109/INFVIS.2000.885097.

World Bank. *Sahel Drought Situation Report No.10*. World Bank, 2012.

World Bank. World bank statistics in burkina faso. `https://www.worldbank.org/en/country/burkinafaso`, 2020. Accessed : 2020-04-01.

Xiaozhu Wu and Ximei Zhang. An efficient pixel clustering-based method for mining spatial sequential patterns from serial remote sensing images. *Computers & Geosciences*, 124 :128–139, 2019.

Kejing Xiao, Chenmeng Wang, Qingchuan Zhang, and Zhaopeng Qian. Food safety event detection based on multi-feature fusion. *Symmetry*, 11(10), 2019. doi : 10.3390/sym11101222.

Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan Yao, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, and Lu Su. DeepFusion : A Deep Learning Framework for the Fusion of Heterogeneous Sensory Data. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206, 2007.

Lu Yao, Zhang Pengzhou, and Zhang Chi. Research on news keyword extraction technology based on tf-idf and textrank. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pages 452–455, 2019. doi : 10.1109/ICIS46139.2019.8940293.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11, 2020.

Haobin Yu. *Named Entity Recognition with Deep Learning.* PhD thesis, Auckland University of Technology, 2019.

Ning Yu and Timothy Haskins. Knn, an underestimated model for regional rainfall forecasting. *arXiv preprint arXiv :2103.15235*, 2021.

Xinglong Yuan, Wenbing Chang, Shenghan Zhou, and Yang Cheng. Sequential pattern mining algorithm based on text data : taking the fault text records as an example. *Sustainability*, 10(11) :4330, 2018a.

Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm : A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 984–992, New York, NY, USA, 2018b. Association for Computing Machinery. ISBN 9781450355520.

Mohamed Akram Zaytar and Chaker El Amrani. Sequence to sequence weather forecasting with long short-term memory recurrent neural networks. *International Journal of Computer Applications*, 143(11) :7–11, 2016.

Min Zhang, Jinman Wang, and Yu Feng. Temporal and spatial change of land use in a large-scale opencast coal mine area : A complex network approach. *Land Use Policy*, 86 :375–386, 2019.

Xue Zhang, Hauke Fuehres, and Peter A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26 :55–62, 2011. doi : https://doi.org/10.1016/j.sbspro.2011.10.562.

Yemdaogo Zida and Sansan H. Kambou. *Cartographie de la pauvreté et des inegalités au burkina faso.* Programme des Nations Unies pour le Développement (PNUD), 2014.

Annexes

A Catégorisation et discussion des biais

Dans cette annexe, nous dressons un panorama des principaux types de biais issus d'enquêtes ménages en nous appuyant sur les classifications proposées par plusieurs articles (Winter, 2004 ; Dussaix, 2009 ; Biemer, 2010) : les biais de non-observation liés à la constitution d'un échantillon non représentatif de la population (e.g., erreur de couverture, d'échantillonnage, non réponse de certains participants), les biais d'observation dus à une erreur de mesure durant une interview, les biais propres aux enquêtes pluriannuelles et les biais liés au traitement et à l'analyse des données. Nous illustrons nos propos en faisant une description détaillée de la méthodologie de l'enquête EPA et en donnant un aperçu des biais présents dans cette enquête. Ces biais peuvent se répercuter sur la qualité des indicateurs de sécurité alimentaire, issus de l'enquête EPA, qui sont étudiés dans la suite de ce chapitre et utilisés dans le chapitre suivant. La catégorisation des biais présentée ci-dessous comporte un niveau élevé de détail, et parmi les types de biais présentés un certain nombre sont invérifiables, que ce soit dans le protocole méthodologique (Permanent Agricultural Survey, 2015) ou dans les données elles-mêmes. Mais il semble important, lorsque l'on travaille sur des modèles sophistiqués appliqués à ce type de données ménages dont la collecte est d'une grande complexité, d'avoir effectué ce travail de réflexion, de lucidité et d'humilité sur tous les facteurs qui peuvent affecter la qualité de ces données, sans lesquels les modèles, aussi complexes soient-ils, ne sont rien.

A.1 Biais de non-observation

Les biais de non-observation, liés à l'utilisation d'un échantillon non représentatif de la population cible, entraînent un écart entre la valeur moyenne "réelle" de l'indicateur sur l'ensemble des réponses obtenues et la valeur moyenne "réelle" de l'indicateur sur la population cible.

A.1.1 Erreur de couverture

Les erreurs de couverture résultent d'une représentation inexacte de la population cible à partir de la base de sondage utilisée pour l'échantillonnage. Certains individus de la population cible peuvent être omis de la base de sondage (sous-dénombrement) ou, au contraire, d'autres individus qui ne font pas partie de la population cible peuvent

être inclus par erreur (sur-dénombrement). Pour éviter cette source de biais, la base de sondage doit être fiable : recensement gouvernemental, enquête menée par une autorité administrative, etc.

L'enquête EPA est réalisée par échantillonnage stratifié à deux degrés (villages et ménages) renouvelé tous les 5 ans. La base de sondage du premier degré est obtenue à partir du module agricole du recensement général de la population de 2006. Cette base a permis de disposer d'une liste de villages (7 871 villages et secteurs) avec 1 219 241 ménages agricoles (en 2008, 1 424 909 ménages au total étaient agricoles, soit 81.5% des ménages (Bureau central du recensement général de l'agriculture, 2011)). La base de sondage des ménages agricoles est créée dans chaque village échantillonné (sélectionnés avec une probabilité proportionnelle à leurs nombres de ménages agricoles) à partir d'une liste de ménages établie chaque année en recensant tous les ménages agricoles du village. Si de nouveaux villages ont été créés depuis 2006, ils ne peuvent pas être sélectionnés car la liste des villages n'est pas mise à jour chaque année. Considérant que nous voulons une estimation de la sécurité alimentaire sur l'ensemble de la population, il existe un léger sous-dénombrement dû au fait qu'un cinquième de la population rurale ne travaille pas dans l'agriculture, cette partie de la population n'est pas représentée par l'enquête EPA. Dans la province urbaine du Kadiogo qui contient Ouagadougou et son agglomération, 85% de la population est urbaine. Le sous-dénombrement y est élevé car la proportion de ménages agricoles est plus faible que dans le reste du pays.

A.1.2 Fluctuations et biais d'échantillonnage

L'échantillonnage désigne les méthodes de sélection de l'échantillon sur lequel la valeur moyenne d'un indicateur est estimée pour l'ensemble d'une population cible. Les fluctuations d'échantillonnage sont des variations aléatoires dans l'estimation de l'indicateur qui se produisent lorsqu'un échantillon est sélectionné. L'erreur aléatoire qui en résulte est donc inévitable lorsqu'on réalise des enquêtes qui ne couvrent pas l'ensemble de la population cible et celle-ci diminue lorsque la taille de l'échantillon augmente. Les erreurs dues à un biais d'échantillonnage sont, au contraire, des erreurs systématiques. Un échantillon est biaisé lorsqu'il présente des caractéristiques différentes de la population qu'il représente et qu'il n'est donc pas représentatif de celle-ci. Cela introduit un biais dans l'estimation de l'indicateur pour l'ensemble de la population cible. Les méthodes d'échantillonnage telles que l'échantillonnage probabiliste ou par quotas permettent de

réduire ce biais.

L'enquête EPA est, pour rappel, réalisée par un échantillonnage stratifié à deux degrés. L'unité primaire est le village administratif, tiré avec une probabilité proportionnelle à sa taille en ménages agricoles. L'unité secondaire est le ménage agricole, les ménages sont regroupés en deux strates homogènes en fonction de leur capacité de production agricole. Le nombre de ménages est choisi pour être représentatif dans chaque province. Six ménages sont sélectionnés par village, par tirage aléatoire simple sans remplacement. Compte tenu du taux de croissance annuel de la population d'environ 3% depuis 2006, les provinces du Nord et du Sahel sont légèrement en sous-effectif en 2012 et 2013. A partir de 2014, tous les échantillons associés à chaque province ont une taille statistiquement significative.

A.1.3 Non-réponse totale et partielle

L'erreur de non-réponse se produit lorsque les informations nécessaires au calcul d'un indicateur ne sont pas recueillies auprès de tous les individus de l'échantillon. La non-réponse est "totale" lorsque l'individu ne répond pas du tout à l'enquête. Les principales causes sont les suivantes : impossibilité de contacter l'individu, impossibilité pour l'individu de répondre ou abandon de l'individu. La non-réponse est partielle si l'individu ne répond pas à certaines questions. Dans ce cas, les causes peuvent être le refus de répondre à certaines questions jugées indiscrètes, l'incompréhension de la réponse de l'enquêté ou l'abandon en cours d'enquête. Tout comme les erreurs d'échantillonnage, ce phénomène peut affecter la précision des estimations en raison de la réduction de la taille de l'échantillon (si les répondants défaillants ne sont pas remplacés) et fausser les résultats de l'enquête si les causes de la non-réponse sont corrélées aux variables de l'enquête (). Il existe plusieurs façons de réduire le biais de non-réponse : informer le répondant de l'enquête à l'avance, établir une relation de confiance avec le répondant ou encore utiliser des incitations comme la rémunération.

Depuis 2009, il y a eu moins de 3% de refus par les ménages à l'invitation à l'enquête EPA. Lors du passage dans les ménages, moins de 1% des ménages ont refusé de répondre à l'ensemble du questionnaire. Concernant les non-réponses partielles, nous ne savons pas quel est le pourcentage de ménages ayant refusé de répondre aux questions qui permettent de calculer le SCA et le SDA car le système de saisie des questions pour

le calcul de ces indicateurs ne permet pas la non-réponse. Pour l'$ISAr$, environ 4% de non-réponses partielles ont été enregistrées, ce qui reste relativement bas.

A.2 Biais d'observation

Les biais d'observation, liés à une erreur de mesure de l'indicateur recherché, provoquent un écart entre la valeur moyenne "estimée" de l'indicateur sur l'ensemble des réponses obtenues et la valeur moyenne "réelle" de l'indicateur sur l'ensemble des réponses obtenues.

A.2.1 Biais liés au questionnaire

Tout instrument de mesure comporte des imprécisions. Dans les enquêtes ménages, l'instrument de mesure est un questionnaire dont la structure et les formulations peuvent entraîner des biais. Sur le fond, il faut s'assurer que le contenu des questions posées permette d'obtenir les informations souhaitées de la manière la plus précise possible. Plusieurs biais cognitifs liés à la forme du questionnaire peuvent se manifester et affecter la qualité des réponses : les effets de primauté et de récence font référence à une préférence du répondant pour le premier et le dernier choix d'une liste ; le biais d'acquiescement reflète une tendance du répondant à répondre plus facilement "oui" que "non" ; l'ordre des questions peut influencer les réponses par un "effet de contamination" ; la longueur du questionnaire peut influencer le nombre de personnes qui accepteront d'y répondre et la précision des réponses ; le choix des mots est essentiel, il existe des mots connotés qui ont une charge émotionnelle suffisamment importante pour influencer les réponses. Par exemple : "Devrions-nous déclarer la guerre" recevrait moins de réponses positives que "Devrions-nous prendre part au conflit ?". Pour réduire cette catégorie de biais, Ghiglione and Matalon (1998) proposent la méthodologie d'élaboration du questionnaire suivante : formulation des objectifs du questionnaire ; définition des informations à collecter et choix des répondants ; formulation des questions ; choix de la structure du questionnaire ; pré-test du questionnaire auprès d'un petit nombre d'individus de la population cible.

Dans le cadre de l'enquête EPA, le choix des données à collecter a été fait avec l'ensemble des acteurs concernés par l'utilisation de ces données. Les questionnaires et les manuels sont contrôlés et validés lors d'ateliers dédiés. Les questionnaires sont révisés chaque

année en fonction des nouveaux besoins d'information et en tenant compte des leçons tirées des années précédentes.

A.2.2 Biais liés à l'enquêteur

Deux types de biais peuvent provenir de l'enquêteur. Premièrement, les biais méthodologiques qui incluent la mauvaise interprétation des questions posées aux enquêtés, l'orientation des réponses ou encore l'attitude verbale et gestuelle qui peut entraîner une posture défensive de l'enquêté. Deuxièmement, les biais techniques liés aux outils utilisés, comme la mauvaise utilisation d'un questionnaire papier ou les erreurs de saisie sur une tablette. L'état physique et mental de l'enquêteur est un facteur de biais à prendre en compte : fatigue, lassitude, dureté et complexité de l'enquête peuvent avoir un impact sur la qualité de la prestation. Pour réduire les biais des enquêteurs, des dispositifs doivent être mis en place avant, pendant et après l'enquête : formation méthodologique et technique des enquêteurs pour les sensibiliser aux sujets et aux objectifs de l'enquête, à l'attitude à adopter pendant l'enquête, à l'utilisation du questionnaire ; accompagnement et assistance des enquêteurs pendant l'enquête ; contrôle a posteriori dans les données et sur le terrain de la qualité du travail des enquêteurs.

Dans l'enquête EPA, les enquêteurs sont recrutés dans chaque village afin d'établir une relation de confiance et d'améliorer la communication avec les ménages locaux. L'enquêteur travaille sous l'assistance directe d'un contrôleur communal qui assure le suivi des opérations de collecte des données. Des superviseurs provinciaux, régionaux et nationaux reproduisent certains entretiens (pour contrôler le travail des enquêteurs) et examinent les questionnaires remplis pour détecter d'éventuelles incohérences. Au début de la campagne agricole, deux sessions de formation sont organisées pour tout le personnel impliqué dans le processus de collecte des données : les formateurs d'enquêteurs sont formés en amont, les instructions des manuels et les différentes variables contenues dans les questionnaires (e.g., leur disposition, leur codage, leur interprétation) sont passées en revue ; les enquêteurs et les contrôleurs sont ensuite formés au niveau régional.

A.2.3 Biais liés au répondant

Le répondant peut donner des réponses inexactes ou erronées pour diverses raisons : mauvaise compréhension de la question, mémoire confuse, fatigue due à la durée de

l'entretien, etc. Les causes peuvent également être liées à la psychologie du répondant : le biais de désirabilité sociale correspond à une tendance du répondant à s'idéaliser et à donner une image valorisante de lui-même ; le biais de conformité sociale induit des réponses que le répondant croit conformes aux normes sociales, à ce qui est attendu ; la contraction défensive à la question personnalisée indique une tendance du répondant à être évasif ou à mentir s'il juge une question trop délicate ou trop personnelle. Certains éléments contextuels comme le lieu ou le moment de l'entretien peuvent affecter les réponses du répondant (Glick, 2009). La conception de questionnaires appropriés et la formation adéquate des enquêteurs permettent de réduire ces biais.

En ce qui concerne l'enquête EPA, les ménages sélectionnés sont informés à l'avance par des enquêteurs qui sont du même village, ce qui facilite l'accès et la confiance des ménages. La sensibilisation des ménages à l'enquête est un processus continu qui commence avant les opérations de collecte de l'EPA et se poursuit jusqu'à la fin afin de maximiser la confiance des répondants et donc le taux de participation et la qualité des réponses. Concernant les informations demandées aux ménages, de nombreuses questions sont nécessaires pour calculer le SCA (119), ce qui représente une charge cognitive importante pour le répondant, cela peut avoir un impact sur la qualité des réponses. Les 5 questions pour le calcul de l'$ISAr$ portent sur des questions de privation alimentaire et peuvent être considérées comme intrusives, embarrassantes (cf. Tableau 4). Par exemple, la majorité des valeurs manquantes dans les questions nécessaires au calcul de l'$ISAr$ sont dues à la question qui concerne les enfants, et c'est pour cette même question que les réponses sont le plus souvent négatives. Est-ce représentatif de la réalité ou les réponses embellissent-elles la situation réelle (biais de désirabilité sociale) ? Nous n'avons aucun moyen de mesurer ce biais mais nous supposons qu'il existe.

A.3 Biais liés aux changements dans le temps

Dans le cas d'enquêtes pluriannuelles, les biais liés à l'échantillon, au questionnaire, à l'enquêteur ou au répondant tendent à augmenter. L'attrition désigne la perte prématurée de répondants au fil du temps, qui peut être due à un déménagement, à la volonté de ne plus participer à l'enquête, etc. En diminuant la taille de l'échantillon, l'attrition peut affecter la significativité des résultats. Si, de plus, l'attrition n'est pas aléatoire et touche des individus aux caractéristiques spécifiques, l'échantillon peut ne plus être représentatif de la population (Alderman et al., 2001). Pour prévenir ce phénomène, une bonne

communication et d'autres moyens tels qu'une compensation financière peuvent être envisagés. Pour remédier à ce phénomène, il est nécessaire de réinjecter dans l'échantillon des individus ayant des caractéristiques similaires à ceux qui ont été perdus de vue. Par ailleurs, le contenu d'un questionnaire peut changer d'une année sur l'autre s'il a été mal défini ou si de nouvelles informations doivent être collectées, ces changements doivent être minimisés car toute modification de la forme du questionnaire peut introduire des biais et rendre difficile la comparaison des indicateurs obtenus entre les différentes années. Enfin, la qualité du travail des enquêteurs et des réponses des répondants peut varier dans le temps pour diverses raisons : fatigue, comportement mécanique des enquêteurs, répondants qui pensent trop bien connaître les caractéristiques de l'enquête et autres changements d'attitudes dus au temps. Il est donc important de s'assurer que les enquêteurs et les répondants participent avec la même application à chaque nouvelle enquête.

Dans l'enquête EPA, l'ensemble des enquêteurs et des ménages enquêtés est renouvelé tous les cinq ans pour tenir compte de la lassitude générée par l'accumulation de tâches répétitives de l'enquête. Des simulations ont été utilisées pour assurer la représentativité et la convergence des nouveaux échantillons en termes de résultats par rapport aux anciens. Les questionnaires sont peu modifiés d'une année sur l'autre ; des modifications significatives des questionnaires ont été apportées en 2014 afin d'obtenir les données nécessaires au calcul de l'$ISAr$.

A.4 Biais de traitement et d'analyse

D'autres sources de biais peuvent être identifiées à chaque étape du traitement informatique des données (Dussaix, 2009). Les biais peuvent provenir d'erreurs de saisie informatique (saisie trop rapide, fiches papier illisibles), d'erreurs de traitement (codage de variables non incluses, traitement des données effectué "à la main", c'est-à-dire sans script automatisé, erreurs de code dans le traitement des données) et d'erreurs d'analyse de données (mauvaise gestion des valeurs aberrantes et manquantes, mise en évidence de résultats non significatifs). Ces biais peuvent être réduits par un travail rigoureux des data managers et des statisticiens en créant des formulaires de saisie avec des contraintes pour minimiser la possibilité d'erreurs de saisie, en mettant en place des doubles saisies, en appliquant des scripts de traitement et de contrôle des données pour vérifier et valider les résultats à chaque étape, en prenant en compte les notions de seuil de confiance et de

marge d'erreur des paramètres estimés et enfin en considérant tous les biais mentionnés ci-dessus dans l'interprétation et la communication des résultats.

La Direction générale des études et des statistiques sectorielles (DGESS) assure le traitement et l'analyse des données de l'EPA. Elle coordonne la conception des programmes de saisie, la formation des agents de saisie et des contrôleurs, l'analyse des données, l'édition et la validation des résultats. Des contrôles d'incohérence sont effectués entre chaque étape du traitement des données : relecture des questionnaires remplis avant la saisie informatique, script de contrôle des incohérences dans la saisie informatique des questionnaires et rapports statistiques des résultats de l'enquête par province et région pour identifier les éventuelles incohérences. Le cas échéant, les procédures de traitement des données incohérentes sont vérifiées et des retours sur le terrain peuvent être envisagés si nécessaire. Concernant le traitement des données réalisé dans le cadre de cette thèse, un prétraitement des données a été effectué avant l'analyse. L'objectif était de détecter d'éventuelles incohérences, puis de les corriger lorsque cela était possible (e.g., un ménage associé à une commune rattachée à la mauvaise région), ou si nécessaire de supprimer les données incohérentes de l'étude. Par exemple, les doublons détectés et certaines valeurs nulles de SCA et SDA (signifiant qu'un ménage n'a consommé aucune nourriture au cours des 7 derniers jours) ont été considérés comme aberrants.

B Jeux de données

Données de séries temporelles
Température de brillance lissée (SMT) mensuelle (mai à novembre)
Précipitations totales mensuelles (de mai à novembre)
Température minimale moyenne mensuelle (°C) (Mai à Novembre)
Température maximale moyenne mensuelle (°C) (Mai à Novembre)
Prix mensuels du maïs (de mai à novembre)
Données météorologiques
Durée moyenne d'ensoleillement par jour
Humidité relative maximale
Humidité relative minimale
Température maximale moyenne par jour
Température minimale moyenne par jour
Évaporation (mm)
Précipitations annuelles (mm)
Données de densité de population
Autocorrélation spatiale de 2 km
Autocorrélation spatiale à 5 km
Indice de Gini
Entropie différentielle
Données économiques de la Banque mondiale
Entrées nettes d'investissements directs étrangers (% du PIB)
Sorties nettes d'investissements directs étrangers (% du PIB)
Dépenses nationales brutes (% du PIB)
Dépenses de consommation finale des ménages (% du PIB)
Dépenses militaires (% du PIB)
Commerce de marchandises (% du PIB)
Croissance du PIB par habitant (% par an)
Indice de végétation par différence normalisée (NDVI)
NDVI moyen de mai à novembre de l'année au cours de laquelle la variable réponse a été collectée
NDVI moyen de mai à novembre de l'année précédant la collecte de la variable réponse
Hôpitaux et écoles
Nombre d'hôpitaux pour 1 000 habitants
Nombre d'écoles pour 1 000 habitants
Événements violents
Nombre total d'événements violents pour 1 000 habitants
Nombre de protestations pour 1 000 habitants
Nombre d'émeutes pour 1 000 habitants
Nombre d'événements violents contre des civils pour 1 000 habitants
Qualité des sols
Capacité de rétention des éléments nutritifs
Conditions d'enracinement
Disponibilité de l'oxygène pour les racines
Cours d'eau
Nombre de cours d'eau
Longueur totale des cours d'eau par km2
Altitude
Altitude maximale
Variance de l'altitude
Données à haute résolution spatiale
patchs de 10x10 pixels de densité de population à 100 mètres de résolution
Occupation du sol - patchs de 10x10 pixels de cultures à 100 mètres de résolution
Occupation du sol - patchs de 10x10 pixels de Forêts à 100 mètres de résolution
Occupation du sol - patchs de 10x10 pixels de zones construites à 100 mètres de résolution

Contributions scientifiques

- **Revues internationales**

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Teisseire, M. (2021) Suivi de la sécurité alimentaire en Afrique de l'Ouest : Quelles méthodes d'analyse de données pour traiter l'interdisciplinarité de la sécurité alimentaire. Journal of Interdisciplinary Methodologies and Issues in Science (JIMIS), Vol. 8.

Deléglise, H. ; Interdonato, R. ; Bégué, A. ; Maître d'Hôtel, E. ; Teisseire, M. ; Roche, M. (2021) Food security prediction from heterogeneous data combining machine and deep learning methods. Expert System With Applications (ESWA), Elsevier, in revision.

- **Actes de conférences internationales**

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2020) Linking Heterogeneous Data for Food Security Prediction. In: Koprinska I. et al. (eds) ECML PKDD 2020 Workshops. ECML PKDD 2020. Communications in Computer and Information Science, Springer, Vol. 1323, pages 335-344.

- **Posters en conférences internationales**

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2020) Linking heterogeneous data for strengthening food security systems. Elsevier, 1 p. 4th International Conference on Global Food Security.

Schaeffer, C. ; Interdonato, R. ; **Deléglise, H.** ; Roche, M. ; Bégué, A. ; Cissé, A. (2020) News mining for food security: the case of Burkina Faso. Elsevier, 1 p. 4th International Conference on Global Food Security.

- **Conférences et Workshops (sans actes)**

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2019) Mise en relation de données hérérogènes pour le renforcement des systèmes de sécurité alimentaire – Cas de la production agricole en Afrique de l'Ouest. CNRIA 2019 : 9e ConféreNce sur la Recherche en Informatique et ses Applications, Saint-Louis, Sénégal.

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2019) Mise en relation de données hérérogènes pour le renforcement des systèmes de sécurité alimentaire – Cas de la production agricole en Afrique de l'Ouest. AgriNumA 2019 : Symposium "Agriculture Numérique en Afrique", Dakar, Sénégal.

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2019) Mise en relation de données hérérogènes pour le renforcement des systèmes de sécurité alimentaire – Cas de la production agricole en Afrique de l'Ouest. Workshop Dispositif de recherche et d'enseignement en partenariat - Information pour la sécurité alimentaire (dP-ISA), Ouagadougou, Burkina Faso.

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2021) Mise en relation de données hérérogènes pour le renforcement des systèmes de sécurité alimentaire – Cas de la production agricole en Afrique de l'Ouest. Workshop Dispositif de recherche et d'enseignement en partenariat - Information pour la sécurité alimentaire (dP-ISA), Ouagadougou, Burkina Faso.

Deléglise, H. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2021) Mise en relation de données hérérogènes pour le renforcement des systèmes de sécurité alimentaire – Cas de la production agricole en Afrique de l'Ouest. Workshop Montpellier Global Days for Science, Education &

Innovation : Africa 2021, Montpellier, France.

- **Données produites**

Deléglise, H. ; Schaeffer, C. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2021) Lexiques en français sur la sécurité alimentaire et les crises. URL:https://doi.org/10.18167/DVN1/C5PU01 ; Dataverse CIRAD.

Deléglise, H. ; Schaeffer, C. ; Bégué, A. ; Interdonato, R. ; Maître d'Hôtel, E. ; Roche, M. ; Teisseire, M. (2021) Corpus de journaux burkinabés en français sur la sécurité alimentaire publiés entre 2009 et 2018. URL:https://doi.org/10.18167/DVN1/IVVEQL ; Dataverse CIRAD.