Étude Scientométrique de la Recherche en Agriculture Numérique

Cas des financements liés à **DigitAg**

Mémoire de Master 2

Modélisation et Analyse Numérique (MANU)

Présenté par : **Rabah DJEBRA**

Stage réalisé à : **INRAE Montpellier, UMR MoISA**

Encadrant : **Jongheon KIM**

Année universitaire : **2024–2025**

Date de soutenance : **09 juillet 2025**

Dédicace

C'est avec une profonde gratitude et un grand respect que je dédie ce travail :

À mon défunt père,
dont l'absence renforce chaque jour mon courage et ma détermination.

À ma chère mère,
pour son amour inconditionnel, ses sacrifices silencieux et son soutien indéfectible tout au long de mon parcours.

À mes amis, sans exception,
pour leur présence, leur bienveillance et leurs encouragements constants.

Rabah

Remerciements

Avant tout, je rends grâce à Dieu pour m'avoir donné la force, la volonté et la persévérance nécessaires à la réalisation de ce travail.

Je remercie sincèrement mes encadrants, Monsieur Jongheom Kim, Madame Karine Gauche, Madame Véronique Bellon-Maurel ainsi que l'équipe DigitAg et mon enseignante référente Madame Hélène Mathis pour leur encadrement de qualité, leur disponibilité et leur confiance, qui ont grandement contribué à la réussite de ce stage.

Je remercie également les membres du jury pour l'honneur qu'ils me font en acceptant d'évaluer ce travail et de contribuer à son enrichissement par leurs remarques et suggestions.

Enfin, je remercie chaleureusement mes amis ainsi que toutes les personnes qui, de près ou de loin, m'ont soutenu, encouragé ou aidé dans l'élaboration de ce travail.

Rabah

Table des matières

Table des figures

Introduction

L'agriculture numérique s'impose aujourd'hui comme un domaine stratégique pour répondre aux défis agricoles du XXI^{e} siècle. Ce champ de recherche interdisciplinaire, à l'intersection de l'agronomie, de l'informatique et des sciences sociales, connaît un développement rapide à l'échelle mondiale. Dans ce contexte, la France a lancé en 2017 l'Institut Convergences DigitAg, un programme ambitieux visant à structurer la recherche en agriculture numérique autour du pôle montpelliérain.

Ce mémoire propose une évaluation scientométrique de l'impact de DigitAg sur la structuration de ce champ scientifique émergent. L'étude repose sur l'analyse systématique des publications scientifiques issues des bases Scopus, Web of Science et HAL, couvrant depuis 1960 jusqu'a 2025. Nous mobilisons des méthodes quantitatives avancées pour répondre à des questions centrales : DigitAg a-t-il favorisé l'émergence d'un réseau de recherche cohérent ? Comment a-t-il influencé les thématiques de recherche ? Quelle est sa place dans le paysage international ?

Notre approche méthodologique combine plusieurs innovations :

— Un pipeline de traitement des données utilisant Python et Power BI pour le nettoyage et la normalisation des métadonnées scientifiques.
— Une modélisation relationnelle des données dans Power BI et MySQL permettant des analyses multidimensionnelles.
— L'application de méthodes scientométriques inspirées des travaux de Callon et al. (1986) et Zupic & Čater (2015), combinant analyse de réseaux, co-mots et performance bibliographique.

Les résultats obtenus éclairent d'un jour nouveau les dynamiques de construction d'un champ scientifique interdisciplinaire. Ils montrent notamment l'émergence d'une signature montpelliéraine distinctive dans l'agriculture numérique, tout en révélant les limites des approches purement quantitatives pour appréhender ce phénomène complexe.

Cette recherche apporte une contribution originale à la compréhension des mécanismes de l'innovation scientifique, avec des implications potentielles pour les politiques de recherche. Elle démontre la pertinence des analyses scientométriques pour évaluer l'impact des investissements ciblés dans l'enseignement supérieur et la recherche.

Chapitre 1

Collecte et traitement des données

1.1 Définitions et objectifs

La scientométrie est une technique d'analyse quantitative de la production scientifique, mobilisée pour mesurer, cartographier et interpréter les dynamiques de recherche. Au-delà de simples indicateurs de performance, elle constitue un outil d'intelligence scientifique permettant de révéler les réseaux de collaboration, les évolutions thématiques ou encore l'impact institutionnel.

Comme le rappellent Callon et al. [2], la scientométrie permet de visualiser les processus de structuration d'un champ scientifique à partir d'analyses de co-mots et de co-auteurs. Elle participe à identifier les controverses, les courants dominants et les reconfigurations épistémiques.

Les principaux objectifs sont :

— **Évaluation de la recherche** : mesurer la productivité et l'impact des chercheurs et institutions via des métriques bibliométriques (publications, citations, h-index, etc.).

— **Appui aux politiques scientifiques** : éclairer les décisions de financement, recrutement ou structuration disciplinaire.

— **Cartographie des connaissances** : visualiser les réseaux thématiques, identifier les domaines émergents et repérer les zones de convergence scientifique.

— **Analyse de tendance** : suivre l'évolution temporelle de la production scienti-

fique et des collaborations.

Toutefois, une utilisation excessive ou mal contextualisée des indicateurs scientométriques peut engendrer des biais, en favorisant la quantité au détriment de la qualité ou en invisibilisant certaines productions non indexées [4].

1.2 Revue des sources de données

Pour cette étude, les données ont été extraites de deux principales bases bibliographiques internationales :

Web of Science (WoS) : plateforme historique de Clarivate, WoS propose des données structurées incluant les facteurs d'impact, les références croisées et les indicateurs issus du Journal Citation Reports.

Scopus : éditée par Elsevier, cette base couvre un large spectre disciplinaire, avec une forte représentation des sciences appliquées, sociales et de l'ingénierie. Elle propose des formats d'export compatibles avec les outils de traitement de données.

Conformément aux recommandations d'Echchakoui [4], nous avons entrepris une procédure de fusion WoS/Scopus en quatre étapes : export des métadonnées, harmonisation des champs, dédoublonnage, puis agrégation dans une base relationnelle. Ce choix permet de bénéficier de la complémentarité des deux bases, et de limiter les biais liés à une couverture inégale des revues.

HAL et publications internes : bien que pertinentes pour l'analyse de la dynamique française et locale, ces sources seront intégrées dans une phase ultérieure, en raison de contraintes de disponibilité et de temps.

1.3 Sources de données et formats

Dans le cadre de cette étude scientométrique, les données ont été collectées en utilisant un ensemble de mots-clés représentatifs des thématiques liées à l'agriculture numérique, tels que : « Digital agriculture », « Smart farming », « Precision agriculture », « E-agriculture », « Agriculture 4.0 », « Agri-tech », « Farming 4.0 », « Agricultural innovation », « Digital farming », « Intelligent agriculture », « Climate-smart agricul-

ture », « Data-driven agriculture », « Sustainable digital farming », « Robotic farming », « Precision breeding » et « Farm automation ». À partir des principales sources :

Web of Science (WoS) : Les données ont été extraites sous forme de fichiers texte. Un script Python a été développé pour parser ces fichiers et les convertir en fichiers CSV, facilitant ainsi leur intégration dans Power BI.

Scopus : En raison du volume important de publications disponibles depuis 1960, plusieurs fichiers CSV ont été téléchargés directement depuis la plateforme.

1.4 Traitement et nettoyage des données

Le nettoyage et la transformation des données ont été réalisés principalement à l'aide de Power Query dans Power BI et Python. Cette étape a permis de :

- **Sélection des champs utiles** : Titre, auteurs, affiliations, mots-clés, année, financement, DOI.
- **Dédoublonnage** : basé sur le titre nettoyé (minuscules, sans ponctuation) et la date. Cela permet d'identifier les doublons inter-bases comme le recommande Echchakoui [4].
- **Standardisation** : les noms d'auteurs, institutions et mots-clés sont normalisés pour éviter les ambiguïtés liées à la casse, aux abréviations ou à la langue.
- **Préparation aux analyses** : les champs de type chaîne (auteurs, affiliations, mots-clés) ont été convertis en formats tabulaires séparés pour alimenter les modèles relationnels.

Ce prétraitement garantit une cohérence des données en amont des analyses bibliométriques (co-citation, co-auteur, clustering thématique), comme recommandé par Zupic & Čater [7].

Chapitre 2

Modélisation des données scientométriques

2.1 Objectifs de la modélisation

L'objectif de cette étape est de structurer les métadonnées bibliographiques de façon formelle afin de permettre une analyse relationnelle avancée. Il s'agit de représenter les entités scientifiques (publications, auteurs, mots-clés, institutions, financements) sous forme de modèles exploitables par des outils de visualisation, de fouille de données et d'analyse de réseaux.

Comme le précisent Zupic et Čater [7], les méthodes bibliométriques peuvent être regroupées en deux grandes familles :

— **Analyse de performance** : basée sur des indicateurs tels que le nombre de publications, les citations, l'indice h, etc.

— **Cartographie scientifique** : qui repose sur des techniques telles que la co-citation, la co-occurrence de mots, le couplage bibliographique et les réseaux de co-auteurs.

Ce chapitre se concentre sur la modélisation relationnelle et la préparation des données en vue d'analyses bibliométriques croisées.

2.2 Modélisation relationnelle initiale

Une modélisation relationnelle a été mise en place pour structurer les données de manière efficace. Pour chaque source, les tables suivantes ont été crées :

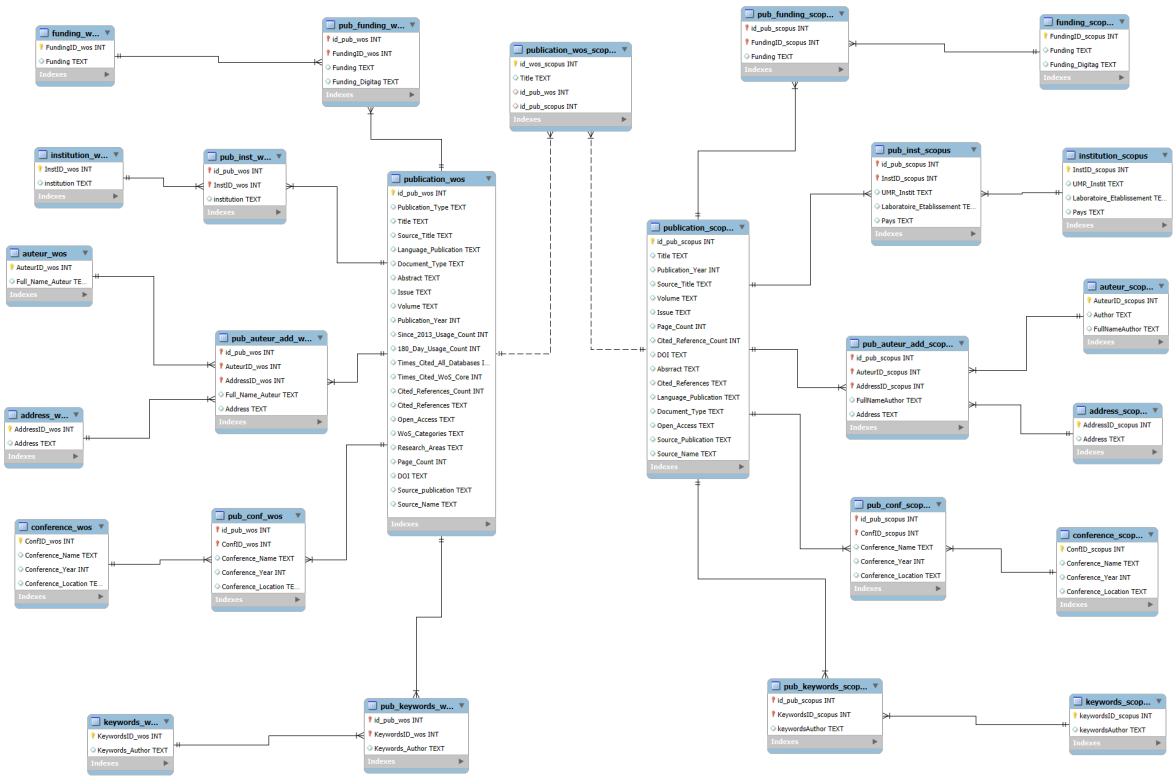


FIGURE 2.1 – Schéma relationnelle wos et scopus de la base de données

Publications : Contient les informations principales de chaque publication.

Auteurs : Liste unique des auteurs.

Address : Liste unique des address associe aux auteurs pour chaque publication.

Institutions : Liste unique des institutions affiliées.

Mots-Clés Liste unique des mots-clés associés aux publications.

Conférences Liste unique des conférences associés aux publications.

Financements : Informations sur les organismes ayant financé les recherches.

Tables de Jointures : Permettent de gérer les relations n à n entre les publications

et les autres entités (auteurs, affiliations, mots-clés, financements).

Tables de correspondances Des tables de correspondance ont été mises en place pour relier les enregistrements dupliqués entre Scopus et WoS, comme préconisé par Echchakoui [4].

2.2.1 Implémentation de la base relationnelle sous MySQL

Dans le cadre de cette étude, une base de données relationnelle complète a été conçue et implémentée dans MySQL Workbench, afin de structurer et interroger efficacement les métadonnées scientifiques issues des bases Web of Science (WoS) et Scopus. La conception a suivi une approche normalisée, en dissociant chaque entité (publications, auteurs, adresses, institutions, mots-clés, conférences, financements) et en modélisant les relations de type n-à-n à l'aide de tables de jointure.

Structuration par source : Deux ensembles de tables ont été créés séparément pour WoS et Scopus. Chaque ensemble comprend une table centrale pour les publications (`Publication_WOS` et `Publication_scopus`) et des tables secondaires pour les auteurs, addresses, mots-clés, institutions, conférences, etc.

Normalisation des entités : Les tables secondaires (`Auteur_WOS`, `Institution_scopus`, etc.) stockent des données uniques. Les relations entre publications et ces entités sont gérées via des tables de jointure comme `Pub_Auteur_Add_wos` ou `Pub_Funding_scopus`, où chaque ligne relie une publication à un ou plusieurs auteurs, affiliations ou financeurs.

Gestion des doublons et fusion : Une table fusionnée `publication_wos_scopus` a été introduite pour relier les publications présentes dans les deux bases, à l'aide de leur `Title` et `DOI` comme clés de correspondance. Les publications y sont enrichies par des champs additionnels, comme les citations, les domaines de recherche ou les identifiants source.

Création de vues consolidées : Plusieurs vues SQL ont été développées pour agréger les informations secondaires :

— `vue_fusion_wos` : regroupe auteurs, mots-clés, affiliations et financements par titre WoS.
— `vue_fusion_scopus` : même agrégation pour Scopus.
— `vue_finale_cortext` : fusion des deux sources, utilisée pour l'export vers Cortext Manager.

Ces vues ont été essentielles pour préparer un fichier d'export au format CSV, structuré selon les attentes de la plateforme Cortext. La commande suivante a permis cet export :

```
SELECT * FROM vue_finale_cortext
INTO OUTFILE '.../publication_cortext.csv'
FIELDS TERMINATED BY ';'
ENCLOSED BY '"'
LINES TERMINATED BY '\n';
```

Résumé technique : L'ensemble du modèle relationnel respecte les bonnes pratiques de modélisation en base de données :

— Utilisation de clés primaires auto-incrémentées ;
— Contraintes de clés étrangères pour assurer l'intégrité référentielle ;
— Groupement avec `GROUP_CONCAT(... SEPARATOR '***')` pour préparer les colonnes multi-valeurs compatibles avec Cortext ;
— Création d'un identifiant de fusion pour chaque publication combinée (`id_fusion`).

Cette modélisation SQL a facilité l'organisation, le filtrage, et l'analyse des publications, en amont des visualisations Power BI et des traitements avancés avec Python ou Cortext.

2.3 Représentation par un graphe orienté pondéré

La base relationnelle est transposée en un graphe orienté pondéré $G = (V, E)$, selon la logique de Callon et al. [2] et des travaux de Loconto et al. [5] en socio-technologie des controverses, où :

— V est l'ensemble des sommets : *Publications*, *Auteurs*, *Institutions*, *Mots-clés*, *Financements*.

— E est l'ensemble des arêtes orientées et pondérées, représentant les liens entre entités (ex. : un auteur est lié à une publication).

2.3.1 Définitions : Graphes orientés et pondérés

Graphe orienté : [1, 3].

— Un graphe est dit *orienté* si ses arêtes, appelées **arcs**, possèdent un **sens de parcours**.

— Un **chemin** est une succession d'arcs mis bout à bout.

— Un **circuit** est un chemin fermé dont les arcs sont tous distincts.

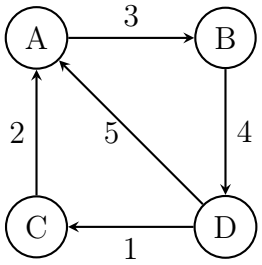
Graphe pondéré : [1, 3].

— Un graphe est dit *étiqueté* si ses arêtes ou arcs portent des **étiquettes** (mots, lettres, symboles, nombres, etc.).

— Lorsque ces étiquettes sont des **nombres**, le graphe est dit **pondéré**, et les étiquettes sont appelées **poids**.

— Le **poids d'une chaîne** (ou d'un chemin) est la **somme des poids** des arcs qui le composent.

FIGURE 2.2 – Exemple de graphe orienté et pondéré



Lecture : Le graphe ci-dessus est orienté (les arcs ont un sens) et pondéré (chaque arc est affecté d'un poids numérique). Par exemple, un chemin de A vers D via B a un poids total de $3 + 4 = 7$.

2.4 Construction des matrices d'adjacence

À partir des relations dans la base, on construit des matrices d'adjacence binaires $A \in \{0, 1\}^{n \times m}$ [6] :

— n : nombre de publications.

— m : nombre d'entités associées (auteurs, mots-clés, institutions, etc.).

$$A_{i,j} = \begin{cases} 1 & si \quad i\mathcal{R}j \\ 0 & sinon \end{cases}$$

Remarque : $i\mathcal{R}j$: Le noeud i en relation avec le noeud j.

2.4.1 Exemples de tables d'adjacence

Matrice Publication–Auteur–Adresse

Publication	Auteur A	Auteur B	Auteur C	Adresse 1	Adresse 2
Pub1	1	1	0	1	0
Pub2	0	1	1	1	1
Pub3	1	1	1	1	1

Lecture : Pub1 est écrite par les auteurs A et B, affiliés à l'adresse 1 ; Pub2 est écrite par les auteurs B et C, affiliés aux adresses 1 et 2 ; Pub3 est écrite par les auteurs A, B et C, affiliés aux adresses 1 et 2.

Matrice Publication–Mot-clé

Publication	Agri-tech	Digital agriculture	Data-driven agriculture
Pub1	1	1	0
Pub2	1	0	1

Lecture : Pub1 utilise les mots-clés Agri-tech et Digital agriculture. Pub2 utilise Agri-tech et Data-driven agriculture.

Matrice Publication–Financement

Publication	#DigitAg	CNRS	UM
Pub1	1	1	0
Pub2	0	1	1

Lecture : Pub1 a été financée par #DigitAg et le CNRS ; Pub2 par le CNRS et l'Université de Montpellier.

Ces matrices sont les points de départ pour l'analyse réseau avec Gephi ou VOSviewer, et pour le calcul de mesures de centralité, de modularité ou de densité Newman [6].

L'approche en graphe permet de croiser les dimensions sociales (collaborations) et cognitives (thématiques), rejoignant l'analyse de controverses développée par Loconto et al. [5] dans leur étude sur les débats agroécologiques.

Chapitre 3

Analyse et interprétation des résultats

3.1 Objectif du chapitre

Ce chapitre présente les principaux résultats issus de l'analyse scientométrique réalisée à l'aide de Power BI, MySQL, des scripts Python et Cortext manager. Il s'appuie sur un corpus de 146 053 publications extraites, nettoyées et fusionnées à partir des bases de données Scopus et Web of Science, selon la méthodologie décrite précédemment. L'analyse s'intéresse à l'évolution temporelle des publications, à la typologie des documents, aux sources de publication et aux principaux financeurs..., en mettant en lumière le rôle du programme DigitAg.

3.2 Évolution temporelle des publications

L'analyse temporelle de la production scientifique constitue une étape clé pour comprendre l'émergence progressive d'un champ de recherche. À l'aide d'un graphique d'aire cumulée, la progression annuelle du volume de publications liées à l'agriculture numérique a été modélisée depuis les années 1960.

Les résultats montrent trois phases distinctes :

— **Phase 1 — Expansion lente (1960–2000) :** La production reste relativement marginale durant cette période, avec moins de 500 publications par an dans le monde. Le champ est alors embryonnaire, centré sur des aspects techniques

spécifiques (capteurs, automatisation de tracteurs).

— **Phase 2 — Croissance progressive (2000–2015) :** La période est marquée par l'introduction des technologies d'information et de communication (TIC) dans l'agriculture, ainsi que par les premiers projets structurants autour de la « precision agriculture ». On observe un doublement quasi tous les 5 ans du volume de publications.

— **Phase 3 — Accélération et explosion (2015–2024) :** Dès 2015, la courbe devient exponentielle. Cette explosion est corrélée à l'émergence des concepts d'« agriculture numérique », à la généralisation des capteurs IoT, à l'intelligence artificielle appliquée à la gestion agricole, et au lancement de programmes tels que DigitAg en France en 2017. Le pic atteint en 2024 (plus de 37 000 publications) reflète la cristallisation du champ au niveau international.

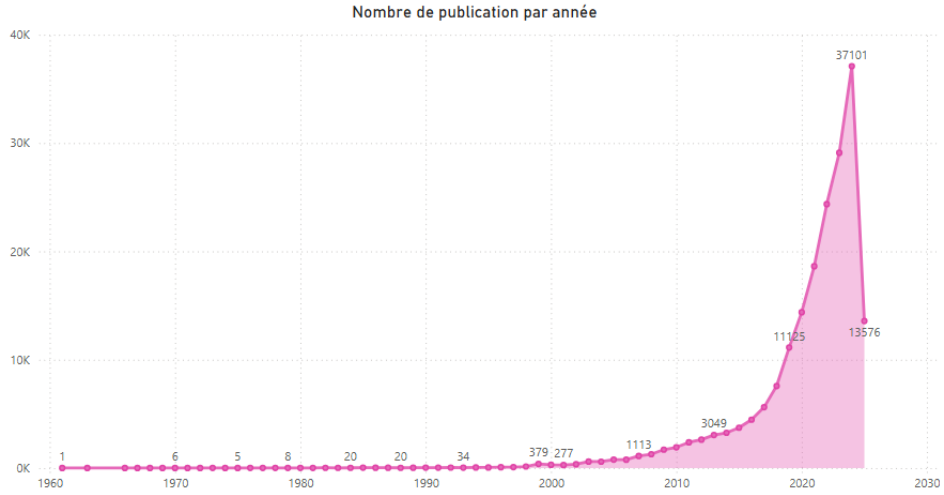


FIGURE 3.1 – Evolution des publications scientifique wos et scopus

Cette dynamique confirme l'hypothèse selon laquelle l'agriculture numérique est devenue un domaine stratégique d'innovation, suscitant l'intérêt croissant des chercheurs, des institutions et des financeurs publics.

3.3 Typologie des documents et des sources

Une typologie documentaire a été réalisée afin de mieux cerner les modalités de diffusion scientifique dans le domaine. À partir d'un graphique combiné (barres empilées par type de document et courbe représentant la source), trois types de publications sont apparus comme dominants :

— **Articles scientifiques (Articles) :** Représentant environ 80 % du corpus, ils constituent le principal canal de diffusion académique. Ce mode de publication est favorisé dans les bases croisées WoS/Scopus et concerne essentiellement des revues à comité de lecture.

— **Communications en conférences (Conference Papers) :** Plus présentes dans Scopus, elles traduisent une orientation technologique, notamment dans les disciplines de l'ingénierie, de la robotique et des technologies de l'information. Elles sont courantes dans les publications issues de projets européens ou asiatiques.

— **Chapitres d'ouvrage (Book Chapters) :** Bien que minoritaires, ils occupent une place significative dans certaines disciplines, notamment les sciences sociales ou les projets collaboratifs transversaux. Ils sont souvent présents dans Scopus, mais rarement indexés dans WoS.

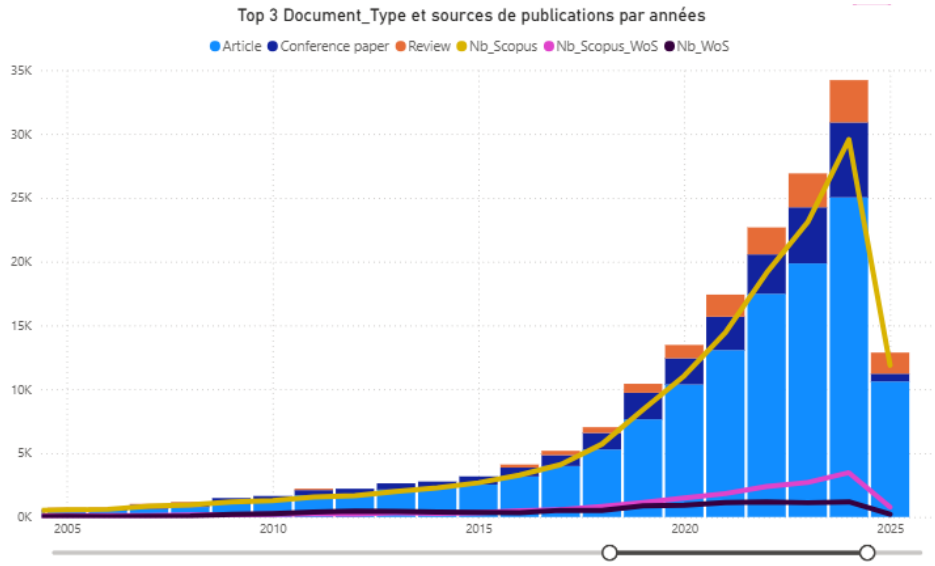


FIGURE 3.2 – Top 3 types de documents et sources de publications par années

Par ailleurs, on note une part croissante de publications figurant simultanément dans Scopus et WoS, signe d'une meilleure visibilité et d'une reconnaissance scientifique accrue. Cette convergence améliore la fiabilité des mesures bibliométriques et traduit une standardisation progressive des pratiques de publication dans le champ.

3.4 Analyse des financements – Focus sur DigitAg

Afin d'analyser les principales sources de financement associées aux publications en agriculture numérique, une visualisation en anneau (donut chart) a été conçue à partir des métadonnées issues des bases bibliographiques. Ce graphique repose sur une mesure DAX personnalisée permettant d'extraire dynamiquement les cinq financeurs les plus fréquemment mentionnés, à laquelle a été ajoutée manuellement la mention explicite du programme `DigitAg [ANR-16-CONV-0004]`, afin de garantir sa visibilité malgré une fréquence d'apparition plus faible.

Cette méthode permet de contourner une limitation observée dans Power BI, où les parts visuelles de chaque catégorie apparaissent équivalentes, indépendamment de leur fréquence réelle. Pour remédier à cela, une mesure intermédiaire a été introduite pour recalculer précisément le nombre de publications par financeur, uniquement pour les entités sélectionnées, assurant ainsi une représentation proportionnelle plus fidèle à la réalité.

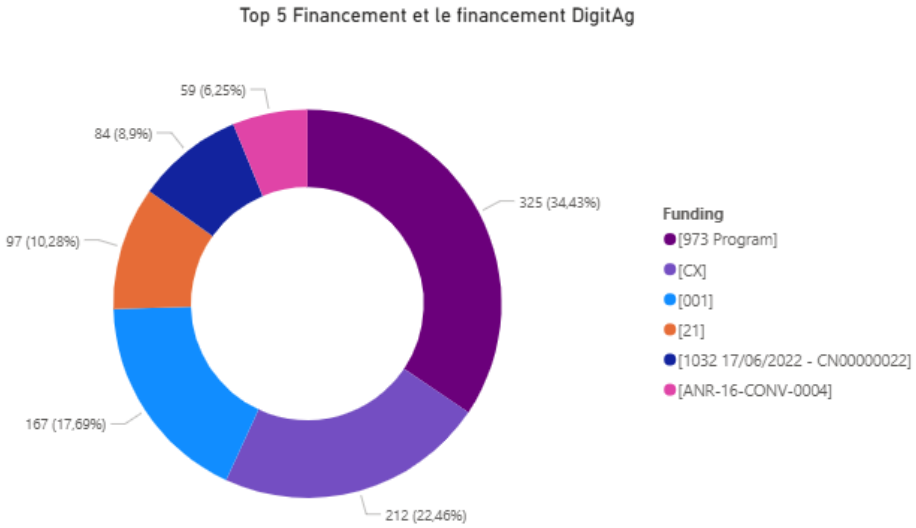


FIGURE 3.3 – Top 5 financements et le financement DigitAg

Il convient toutefois de souligner certaines limites méthodologiques concernant les données analysées. Tout d'abord, cette étude repose exclusivement sur les métadonnées extraites des bases **Scopus** et **Web of Science (WoS)**. Les publications présentes dans la base **HAL** n'ont pas encore été intégrées en raison de contraintes techniques et temporelles. Cette absence peut partiellement expliquer la sous-représentation des organismes nationaux de recherche (ONR) français tels que l'INRAE, le CIRAD ou

l'IRD, ainsi que de l'Agence nationale de la recherche (ANR), principal financeur institutionnel français. Ces acteurs sont souvent mieux référencés dans les archives ouvertes nationales.

Par ailleurs, la qualité de l'extraction automatique des mentions de financement présente des limites. Ces informations sont généralement indiquées entre crochets `[]` ou parenthèses `()` dans les champs textuels, ce qui permet leur détection via des expressions régulières. Cependant, certains cas particuliers, notamment dans des publications chinoises, comportent des parenthèses imbriquées (ex. : `(Supported by the National Program (973) Grant No. ...)`), ce qui complique fortement l'extraction automatique. Malgré l'utilisation de scripts de nettoyage adaptés, ces structures peuvent entraîner des erreurs de segmentation, voire l'omission de certaines mentions. Ces incertitudes doivent être prises en compte dans l'interprétation des résultats présentés.

Les résultats observés à ce stade de l'étude montrent :

— Le programme chinois `[973 Program]` apparaît comme le financeur principal, avec **325 publications**, soit **34,43 %** du total analysé. Ce chiffre reste toutefois susceptible d'être sous-estimé ou surestimé en raison des difficultés d'extraction évoquées.
— Le programme `[CX]` suit avec **212 publications** (**22,46 %**), puis `[001]` avec **167 publications** (**17,69 %**).
— Les financements `[21]` (**97 publications**, soit **10,28 %**) et `[1032 17/06/2022 - CN0000022]` (**84 publications**, **8,9 %**) complètent le Top 5.
— Enfin, le programme `DigitAg` (`[ANR-16-CONV-0004]`) est associé à **59 publications**, soit **6,25 %** du total, ce qui justifie son intégration ciblée dans le visuel.

Bien que numériquement inférieur aux grands financeurs internationaux, DigitAg joue un rôle stratégique dans le paysage français. Son inclusion volontaire dans l'analyse vise à dépasser une lecture strictement quantitative, en soulignant sa fonction structurante dans le développement de l'agriculture numérique.

De plus, des analyses préliminaires montrent que DigitAg est fréquemment associé à d'autres financeurs comme l'INRAE, le CNRS ou des agences nationales. Cette cooccurrence reflète un positionnement transversal du programme, orienté vers le soutien à des projets collaboratifs multi-acteurs. Une exploration croisée avec les affiliations des auteurs permettrait de mieux caractériser ces dynamiques, ce qui fera l'objet d'un approfondissement dans la section suivante.

Enfin, l'intégration ultérieure des publications issues de HAL constitue une perspective essentielle pour améliorer la représentativité de l'échantillon analysé, en particulier pour les productions scientifiques francophones non référencées dans les bases commerciales. Cela permettra également de renforcer l'analyse des financements d'origine nationale et institutionnelle.

3.5 Analyse géographique de la répartition des publications

Pour compléter les analyses thématiques, relationnelles et institutionnelles, une visualisation géographique a été réalisée à l'aide de Power BI. Cette carte permet de représenter les 30 pays ayant le plus contribué à la production scientifique en agriculture numérique, en se basant sur le nombre de publications issues du corpus fusionné WoS/Scopus.

FIGURE 3.4 – Répartition géographique des publications scientifiques (Power BI)

Comme le montre la figure 3.4, la production scientifique se concentre très fortement dans un petit nombre de pays, illustrant les déséquilibres mondiaux en matière de recherche sur l'agriculture numérique.

— **La Chine** domine largement le classement avec plus de 32 000 publications, ce qui témoigne d'un effort massif d'investissement dans la recherche agricole et technologique.

— **L'Inde** suit avec plus de 22 000 publications, traduisant une forte dynamique de recherche sur des technologies adaptées aux systèmes agricoles tropicaux.

— **Les États-Unis** occupent la troisième place avec plus de 18 000 publications, se distinguant par une approche multidisciplinaire (durabilité, IA, politiques agricoles).

— Des pays européens comme **l'Allemagne**, **l'Italie** et **l'Australie** dépassent chacun les 7 000 publications.

— **Le Brésil**, **l'Espagne** et le **Royaume-Uni** affichent chacun un volume supérieur à 5 000 publications.

— **La France**, bien qu'en retrait, contribue avec environ 4 000 publications, notamment à travers des structures comme INRAE, CIRAD ou l'Institut Agro Montpellier (anciennement Montpellier SupAgro).

Cette visualisation renforce les analyses lexicales précédentes en révélant la dimension géographique de la production scientifique. Elle permet de mieux comprendre la distribution des efforts de recherche à l'échelle mondiale et de situer la France dans ce paysage globalisé.

Elle constitue également un outil pertinent pour identifier les pôles régionaux de spécialisation, les zones de collaboration scientifique potentielle, ainsi que les marges d'amélioration en matière de visibilité scientifique pour certains pays.

3.6 Analyse avec Cortext Manager

Dans cette dernière phase, nous avons utilisé la plateforme **Cortext Manager** pour analyser les dynamiques lexicales et thématiques du corpus bibliographique fusionné. Cette analyse s'inscrit dans une démarche de cartographie scientifique, complémentaire aux visualisations temporelles et quantitatives précédemment présentées.

3.6.1 Exportation des données vers Cortext

Une fois les métadonnées nettoyées, harmonisées et fusionnées dans MySQL Workbench (cf. section 2.2), une vue consolidée appelée `vue_finale_cortext` a été générée pour agréger, pour chaque publication, les champs clés suivants :

— `Title`, `Authors`, `Affiliations`, `Keywords`, `Funding`, `Abstract`, `DOI`, `Publication_Year`, `Source`.

Cette vue a été exportée au format `CSV` via la commande SQL (cf. section 2.2)

Le fichier produit respecte les exigences de Cortext Manager, notamment pour l'import du champ multi-valeurs ("Keywords", "Authors", "Affiliations", etc.) séparé par "***".

3.6.2 Analyse lexicale et temporelle avec Epic Epoch

L'analyse dans Cortext a porté sur la dynamique des mots-clés, des sources et des domaines thématiques sur une période longue (1960–2025), en combinant des algorithmes d'extraction lexicale et de mise en réseau. En particulier, le module **Epic Epoch** a été utilisé pour visualiser l'évolution temporelle des concepts dominants.

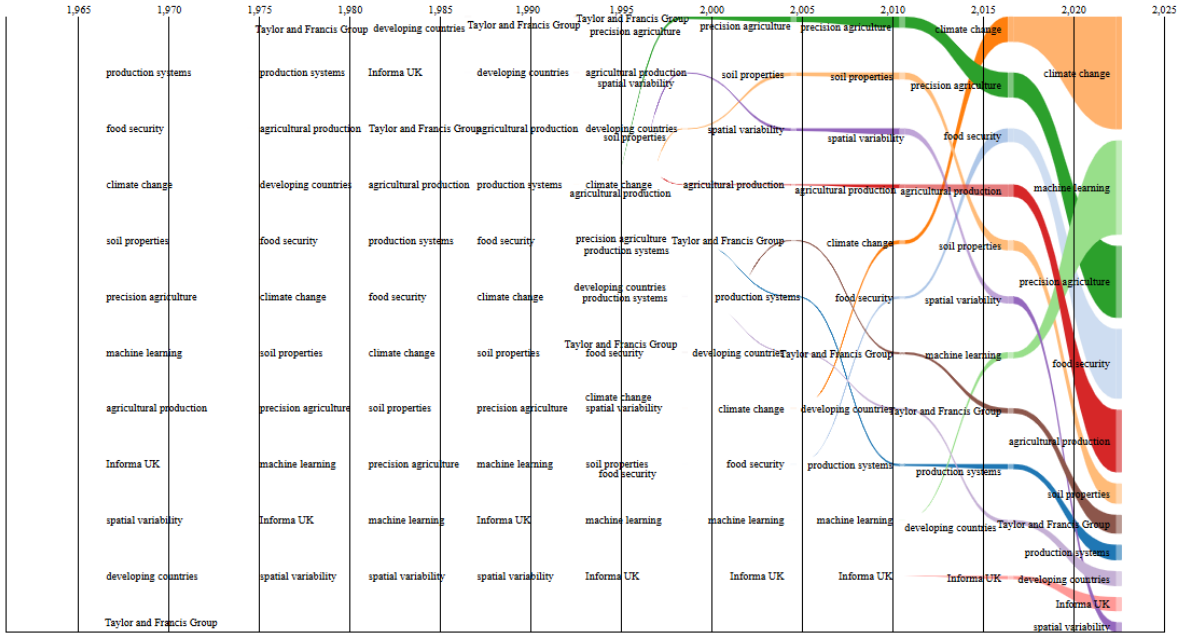


FIGURE 3.5 – Visualisation des thématiques dominantes avec Epic Epoch dans Cortext Manager

La figure 3.5 représente les principales thématiques scientifiques (mots-clés) identifiées dans les publications. Chaque colonne correspond à une période (époque), et chaque nœud à un concept fréquent dans les textes. Les flux entre les colonnes matérialisent la continuité ou l'évolution des thématiques dans le temps.

Les résultats les plus significatifs sont :

— **Precision agriculture**, **Soil properties** et **Spatial variability** sont présents depuis les années 1990 et restent des axes forts.
— **Climate change** et **Food security** s'imposent massivement après 2010, reflétant des préoccupations globales.
— **Machine learning** émerge après 2015, signe d'une intégration croissante de l'intelligence artificielle dans le domaine agricole.

Ces trajectoires thématiques révèlent une hybridation croissante du champ, où se croisent enjeux techniques, environnementaux et sociaux. L'outil Epic Epoch permet ainsi de visualiser non seulement les thématiques dominantes à chaque époque, mais aussi les continuités conceptuelles et les bifurcations stratégiques.

3.6.3 Cartographie des réseaux termes–auteurs

Dans le prolongement des analyses lexicales et temporelles, une cartographie de réseau croisant les termes scientifiques et les auteurs les plus représentés a été réalisée avec **Cortext Manager**. Cette visualisation repose sur un algorithme de co-occurrence basé sur le *Chi2* et une mesure de proximité entre les entités textuelles (auteurs et termes).

La figure 3.6 met en évidence plusieurs structures significatives dans la production scientifique autour de l'agriculture numérique. On y observe :

— Des **clusters thématiques** centrés sur des notions telles que *precision agriculture*, *machine learning*, *soil moisture*, *climate change*, ou encore *IoT*.
— Des **groupes d'auteurs fortement connectés** à certains mots-clés, révélant une spécialisation : par exemple, les auteurs `ZHAO CHUNJIANG`, `ZHANG YU`, `WANG WEI` ou encore `LI DAOLIANG` sont fortement liés aux thématiques liées à l'intelligence artificielle, aux capteurs et aux systèmes d'aide à la décision.
— Une **distribution géographique implicite** : de nombreux auteurs du réseau sont affiliés à des institutions chinoises, ce qui reflète la forte contribution de la Chine à ces thématiques.

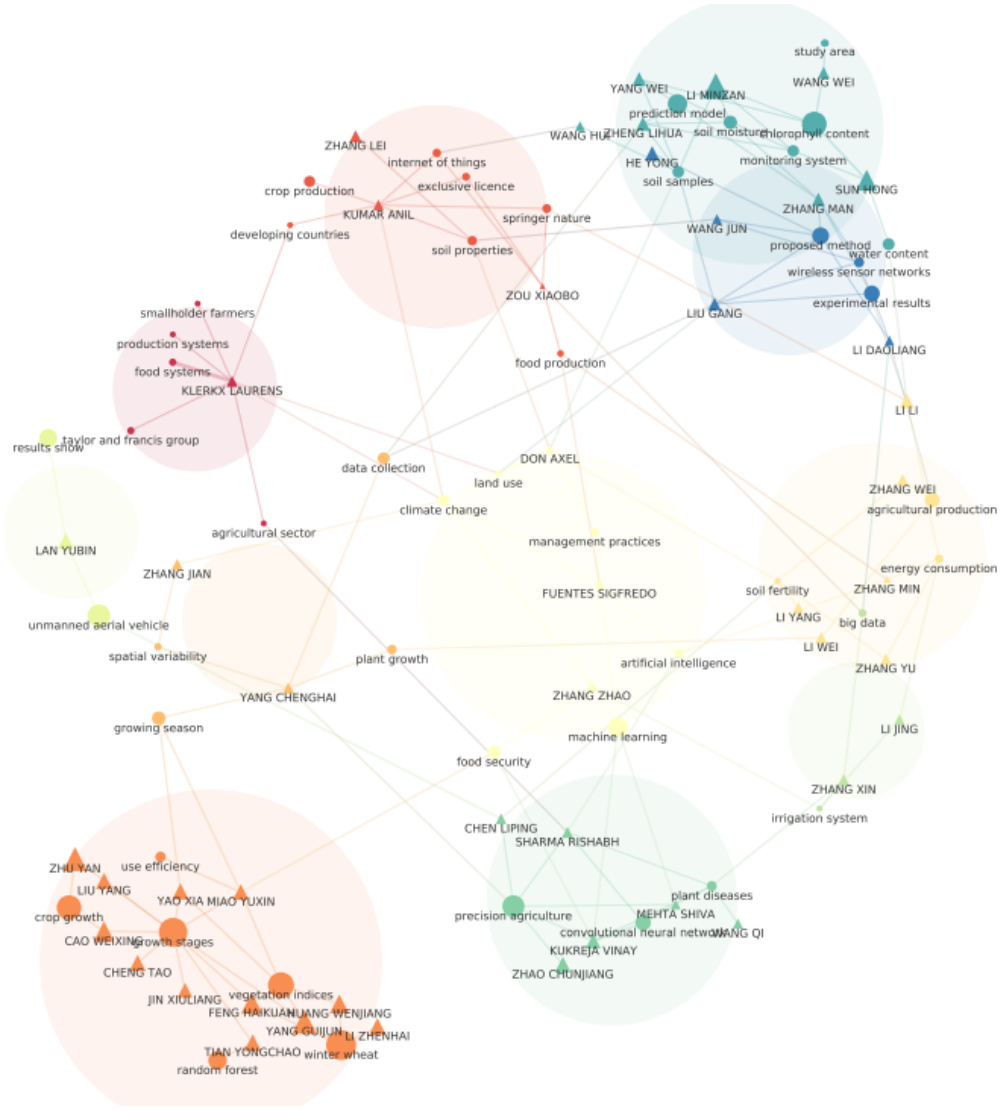


FIGURE 3.6 – Réseau de co-occurrence entre termes scientifiques et auteurs (Cortext Manager)

La construction de ce graphe permet de visualiser les interdépendances entre acteurs et thématiques, et d'identifier les figures structurantes du champ. Les co-occurrences entre auteurs et mots-clés révèlent non seulement des proximités scientifiques, mais aussi des stratégies de positionnement, des écoles thématiques et des partenariats internationaux implicites.

Ce type de visualisation offre un point d'entrée pertinent pour de futures analyses de réseaux de co-auteurs, ou pour suivre l'évolution de groupes de recherche autour de concepts émergents.

3.6.4 Analyse croisée des mots-clés par pays

En complément de l'analyse diachronique menée à l'aide du module Epic Epoch, une analyse croisée entre les pays contributeurs et les mots-clés dominants a été réalisée. Pour cela, une matrice de type *heatmap* a été générée, permettant de visualiser les variations d'occurrence des concepts en fonction des principaux pays producteurs de publications en agriculture numérique. La figure 3.7 présente :

— En ordonnée : une sélection des 15 pays les plus représentés dans le corpus.

— En abscisse : les mots-clés les plus fréquents, standardisés pour éviter les doublons sémantiques.

— En couleur : une échelle de chaleur (`red-blue`) indiquant la fréquence relative par rapport à la moyenne globale. Le rouge indique une surreprésentation, le bleu une sous-représentation.

Cette matrice met en évidence des spécialisations géographiques marquées :

— **La Chine** apparaît fortement liée à des concepts d'intelligence artificielle appliquée à l'agriculture, avec une nette surreprésentation des termes `deep learning` et `smart agriculture`.

— **Les États-Unis** se démarquent par une orientation marquée vers les enjeux de durabilité et d'adaptation climatique, via des mots-clés comme `climate change`, `food security` et `sustainability`.

— **L'Inde** survalorise les concepts d'agriculture connectée, notamment `IoT` et `smart farming`, mais est en retrait sur les thématiques IA ou durabilité.

— **Le Brésil**, **l'Espagne** et **l'Italie** présentent des occurrences spécifiques sur `precision agriculture`, `remote sensing` ou `soil moisture`, illustrant une orientation plus territoriale ou expérimentale.

— **La France** et les pays d'Europe occidentale présentent des profils équilibrés, souvent en lien avec les thématiques transversales (`agriculture`, `remote sensing`, `food security`).

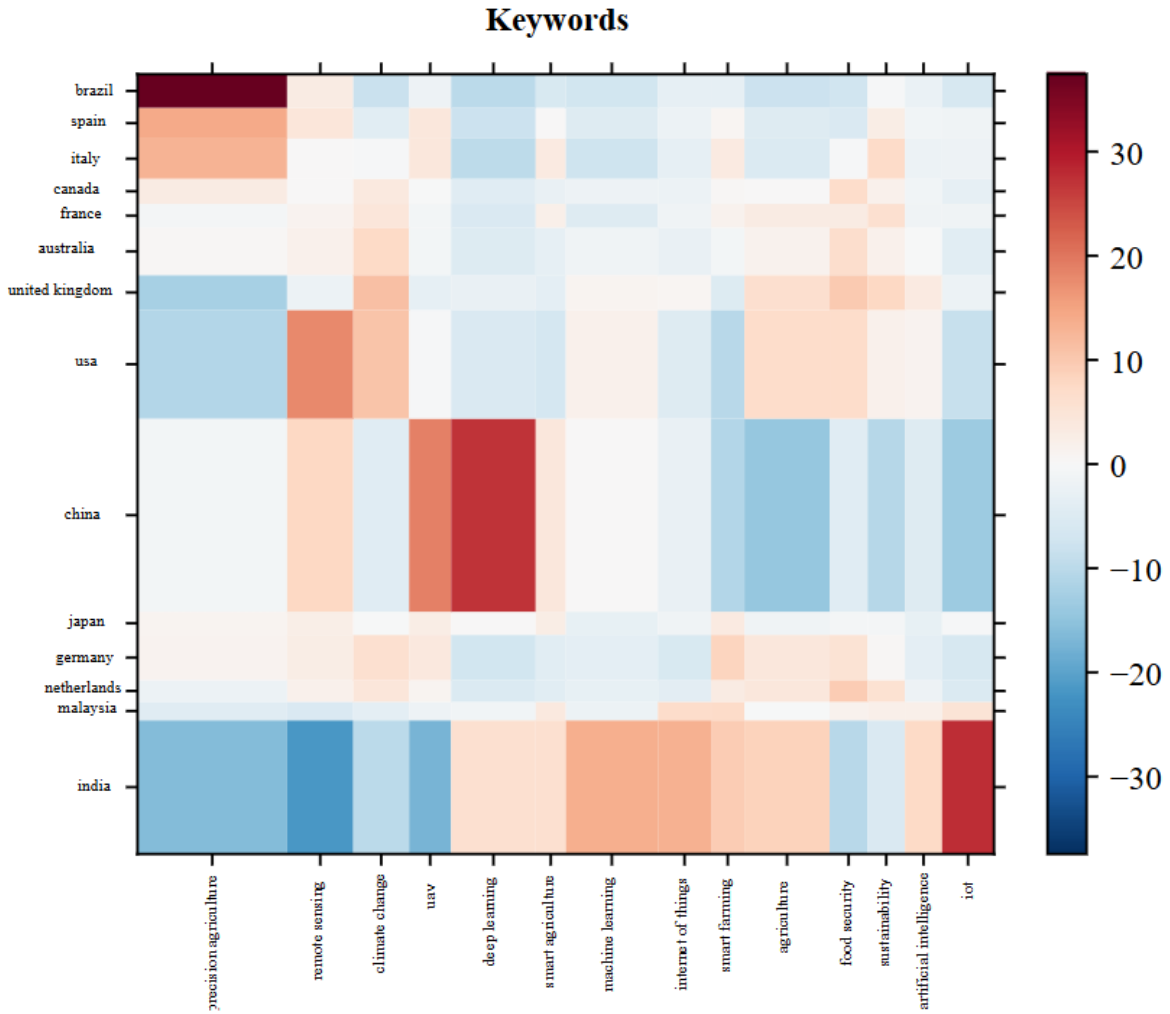


FIGURE 3.7 – Corrélation entre mots-clés dominants et pays contributeurs

Cette visualisation conforte les résultats de l'analyse temporelle et relationnelle, en y ajoutant une dimension géopolitique. Elle permet de mettre en lumière la manière dont chaque pays s'approprie des concepts spécifiques, reflétant à la fois des priorités scientifiques nationales, des stratégies de financement, et des contextes agroécologiques différenciés.

Par exemple, la matrice met en évidence que le changement climatique et la durabilité émergent clairement comme des thématiques majeures dans la recherche française, en cohérence avec les priorités affichées par des établissements tels que l'INRAE et le CIRAD, notamment en matière de biodiversité et de transition agroécologique.

3.7 Synthèse des résultats

L'étude scientométrique présentée dans ce mémoire s'appuie sur l'intégration et l'analyse d'un corpus de plus de 146 000 publications issues des bases Scopus et Web of Science. Grâce à une modélisation relationnelle fine (MySQL), une visualisation interactive (Power BI) et une analyse lexicale et réseau (Cortext Manager), nous avons pu dégager plusieurs résultats majeurs :

- Une **croissance exponentielle** de la production scientifique en agriculture numérique depuis 2015, avec un pic en 2024.
- Une **domination des articles scientifiques**, suivis des conférences, principalement diffusés via Scopus.
- Une **géographie de la recherche** dominée par la Chine, l'Inde et les États-Unis, tandis que la France demeure présente mais sous-représentée dans les bases analysées.
- Des **thématiques structurantes** comme *precision agriculture*, *climate change*, *machine learning*, *IoT*, identifiées grâce aux analyses Epic Epoch et co-termes/auteurs.
- Le programme **DigitAg**, bien que visible dans les données WoS/Scopus, reste quantitativement discret (environ 59 publications), ce qui peut s'expliquer par l'absence des données issues de HAL et des publications internes.

L'ensemble des résultats met en lumière un champ en structuration rapide, à l'interface entre technologies, écologie et politiques agricoles. La méthodologie croisée mobilisée (SQL, DAX, Python, Cortext) a permis d'obtenir une vue multiscalaire temporelle, thématique, institutionnelle et géographique du domaine étudié.

Conclusion générale

Ce mémoire a présenté une démarche complète d'analyse scientométrique visant à explorer la structuration du champ scientifique de l'agriculture numérique, avec un focus particulier sur la contribution du programme DigitAg. En s'appuyant sur une base de données fusionnée provenant de Scopus et Web of Science, enrichie et modélisée dans MySQL Workbench, nous avons pu conduire une série d'analyses à la fois quantitatives (avec Power BI) et lexicales/relationnelles (avec Cortext Manager).

La méthodologie mise en œuvre s'articule autour de trois volets complémentaires :

— **Un traitement rigoureux des données** : extraction, nettoyage, normalisation, fusion, modélisation relationnelle, export structuré pour des analyses ultérieures.

— **Une analyse multidimensionnelle** : dynamique temporelle des publications, typologie des documents, financement, cartographie géographique des affiliations, distribution par pays, spécialisation thématique.

— **Une mise en évidence de logiques scientifiques et institutionnelles** : co-émergence de thématiques (ex. *machine learning*, *remote sensing*), présence dominante de la Chine et des États-Unis, centralité de certains clusters auteurs-termes.

Parmi les constats majeurs :

— Le champ « agriculture numérique » connaît une **croissance rapide**, tirée par l'intégration des technologies numériques dans la recherche agronomique.

— **Les publications francophones**, et notamment celles liées à DigitAg, sont sous-représentées dans WoS et Scopus ce qui limite la portée des conclusions concernant leur visibilité réelle.

— **L'architecture de données conçue** permet cependant de pallier en partie ce biais, en offrant une base exportable et extensible (via l'intégration future des données HAL).

Bien que cette étude ait permis de dégager des tendances fortes à l'échelle mondiale, l'évaluation précise de la contribution de DigitAg appelle plusieurs nuances.

— La place de DigitAg dans le paysage international reste modeste selon les données disponibles, mais elle pourrait être réévaluée à la hausse avec l'intégration des publications nationales non indexées.

— L'impact de DigitAg sur les thématiques reste à préciser, faute de pouvoir isoler ses contributions spécifiques dans le corpus international actuel.

— Les données actuelles ne permettent pas encore d'évaluer pleinement la structuration d'un réseau autour de DigitAg, mais elles en suggèrent les prémices.

Ce travail constitue une base solide pour :

1. développer des tableaux de bord dynamiques de suivi de la production scientifique,
2. orienter les politiques de valorisation et de financement dans les instituts partenaires,
3. conduire des analyses thématiques plus fines, par exemple autour de l'agroécologie, de l'IA ou de l'agriculture de précision.

La prochaine étape logique consistera à intégrer les publications issues de la base HAL, ainsi que les productions internes non indexées, pour compléter l'évaluation de l'impact du programme DigitAg.

Ainsi, cette recherche s'inscrit dans une perspective d'aide à la décision stratégique, offrant des outils et méthodes transférables à d'autres champs scientifiques émergents.

Bibliographie

[1] Claude Berge. *Théorie des graphes et ses applications*. Dunod, 1981.

[2] Michel Callon, Jean-Pierre Courtial, and Hervé Penan. La scientométrie : une introduction à la mesure de la science. *La Découverte*, 1986.

[3] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction à l'algorithmique*. Pearson, 3ème édition edition, 2009.

[4] Sami Echchakoui. Why and how to merge scopus and web of science during bibliometric analysis : The case of salesforce literature. *Journal of Marketing Analytics*, 8 :165–184, 2020.

[5] Allison Loconto, Marion Desquilbet, Théo Moreau, Denis Couvet, and Bruno Dorin. The land sparing–land sharing controversy : Tracing the politics of knowledge. *Land Use Policy*, 96 :103610, 2020.

[6] M. E. J. Newman. *Networks : An Introduction*. Oxford University Press, 2010.

[7] Ivan Zupic and Tomaž Čater. Bibliometric methods in management and organization. *Organizational Research Methods*, 18(3) :429–472, 2015.