Université Grenoble-Alpes

M2 MIASHS Statistiques et Sciences de Données

Année universitaire 2021 - 2022

Rapport de stage de fin d'études

Enseignants référents : *Jean-François Coeurjolly et Adeline Leclercq Samson*

Couplage d'un modèle de culture avec une plate-forme de capitalisation de données issues d'agroécosystèmes à La Réunion

---

De mars à août 2022

Encadré par

- Sandrine Auzoux
- Mathieu Roche
- Mathias Christina
- Benjamin Heuclin

Réalisé par

- Billy Ngaba

Remerciements

J'aimerais remercier Sandrine Auzoux qui m'a donné la possibilité de faire ce stage à la Réunion et qui a fait tout le nécessaire sur le plan administratif et social pour que mon camarade Elyzé Barbaza et moi soyons bien intégrés. Elle m'a suivi pendant toute la durée du projet semaine après semaine. Elle a aussi diffusé beaucoup de bonne humeur et une bonne ambiance tout au long du stage. Je remercie également Mathieu Roche qui a bien voulu m'encadrer sur le plan technique, depuis le site du CIRAD de Montpellier. Je remercie également Mathias Christina, qui s'est toujours porté disponible pour mettre à disposition les données dont j'avais besoin et Benjamin Heuclin, qui était très intéressé par mon travail. Finalement, je remercie Feder, Feader et la région La Réunion pour le financement de mon stage.

Préambule

Ce rapport présente le travail que j'ai réalisé durant mon stage de seconde année de Master MIASHS, parcours Statistiques et Sciences des Données, au sein du CIRAD (Centre de Coopération International en Recherche Agronomique pour le Développement), de mars à septembre 2022. La mission centrale du CIRAD est de contribuer au développement rural des pays tropicaux et subtropicaux par des actions de recherche. Il travaille en coopération avec plus de 100 pays d'Afrique, d'Asie, du Pacifique, d'Amérique latine et d'Europe. Le CIRAD comprend plusieurs sites en France métropolitaine et aussi en outre-mer français.

Table des matières

1 Introduction **1**

2 Exploration des données et préparation d'un jeu de données normalisé **3**

3 Méthodes de fouille de textes d'aide à la normalisation **6**

3.1 Contexte 6

3.2 Processus global 6

3.3 Description des étapes du processus 8

3.3.1 Prétraitement 9

3.3.2 Mesure lexicale 9

3.3.3 Mesure contextuelle 10

3.3.4 Pistes d'amélioration des performances de la mesure contextuelle 11

3.3.5 Combinaison 12

3.3.6 Règles de normalisation 13

3.4 Méthode d'évaluation 14

4 Enrichissement du contexte pour améliorer la mise en lien **15**

4.1 Extension de la méthode contextuelle par enrichissement du contexte par la fouille dans la littérature scientifique 15

4.1.1 Documents choisis 15

4.1.2 Principe de la méthode de fouille texte 15

4.2 Résultats 16

4.2.1 Sans l'enrichissement contextuel 16

4.2.2 Après l'enrichissement contextuel 17

5 Simulation du modèle STICS **20**

6 Conclusion **23**

7 Accessibilité des données et des codes sources **25**

Références **26**

Annexes **28**

A Exemple de représentation vectorielle de textes avec td-idf : **28**

B Autres annexes **29**

B.1 Plan du dispositif expérimental au niveau de l'essai tout entien 29

B.2 Variables sources contextualisés pour enrichir le vocabulaire global. 30

B.3 Listes des termes clés ciblés dans les descriptions des variables sources contextualisés pour enrichir le vocabulaire global 30

1 Introduction

Contexte du projet

J'ai réalisé mon stage dans l'unité de recherche AIDA (Agroécologie et Intensification Durable des Cultures Annuelles), à La station de la Bretagne, à La Réunion. Durant ces six mois, j'ai été encadré par Sandrine Auzoux , cadre en informatique scientifique spécialisée en base de données, Mathias Christina, chercheur en agronomie, Mathieu Roche, chercheur en informatique spécialisé dans la fouille de texte et Benjamin Heuclin, chercheur en biostatistiques, basés à Montpellier. J'ai utilisé le système d'information AEGIS (Agroecological Global Information System) (Auzoux et al. 2019) développé par le CIRAD et le modèle STICS (Simulateur mulTIdisciplinaire pour les Cultures Standard) (Brisson et al. 1998) développé par INRAE. AEGIS permet de stocker et exploiter en ligne des données provenant d'expérimentations en agroécologie menées dans les pays du Sud . Le modèle STICS est un modèle dynamique, générique et robuste qui permet de simuler la croissance d'une culture dans un système sol-atmosphère-culture .

Objectif du Stage

Le sujet du stage est le couplage d'un modèle de culture avec une plate-forme de capitalisation de données issues d'agro-écosystèmes à La Réunion. L'objectif est d'interfacer la base de données du système d'information AEGIS avec le modèle de culture STICS en associant dynamiquement les variables expérimentales décrites par les chercheurs en agronomie avec les paramètres d'entrée du modèle STICS et de paramétrer dynamiquement les simulations depuis R. Ce processus est illustré par le schéma suivant (Figure 1)

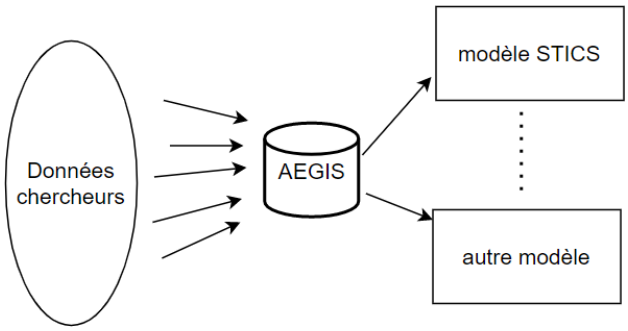


Figure 1 – Cheminement de la donnée du chercheur jusqu'au modèle.

La problématique est la suivante : Chaque chercheur a son propre système de d'annotation des données. AEGIS a basé son système sur du vocabulaire contrôlé provenant d'ontologies en agronomie ou à dire d'expert. Chaque modèle de culture a aussi une terminologie propre pour décrire les paramètres, les variables internes et de sortie. L'objectif est de mettre en relation

des données qui ont la même signification mais qui sont nommées (et/ou décrites) différemment.

Ainsi, mon stage s'est déroulé en quatre phases :

1. Exploration des données et préparation d'un jeu de données normalisé
2. Proposition de méthodes de fouille de textes pour aider à la normalisation
3. Extension de ces méthodes par enrichissement du contexte
4. Automatiser la chaîne de traitement permettant de réaliser une simulation du modèle STICS

2 Exploration des données et préparation d'un jeu de données normalisé

Au cours de ces dix dernières années, les projets CanecoH [1], Ecocanne, Agriecoh menés par eRcane [2] en collaboration avec le CIRAD à La Réunion, ont généré une grande quantité de données expérimentales sur les systèmes de culture à base de canne à sucre en association avec des plantes de services. Les plantes de services remplissent de nombreuses fonctions parmi lesquelles la lutte contre l'enherbement, la lutte contre l'érosion, l'amélioration de la fertilité des sols, la régulation des bio-agresseurs ou l'augmentation de la biodiversité.

Le premier objectif du stage est de capitaliser tous ces jeux de données collectés et d'inciter les chercheurs en agronomie à décrire, structurer, stocker et partager leurs jeux de données dans l'optique d'une réutilisation potentielle. Ce projet s'intègre dans une démarche de partage et d'ouverture des données promue par la Science Ouverte, basée sur les principes FAIR (Wilkinson et al. 2016) [3]Ainsi, les jeux de données doivent avoir une nomenclature et une structure harmonisée pour pouvoir être réutilisés en totalité.

Le jeu de données de référence (voir table 1) sur lequel j'ai travaillé provient de 10 essais [4] en canne à sucre associées à des plantes de service [5] pour limiter la croissance des adventices [6], menées de 2012 à 2020 sur l'île de La Réunion.

1. Canne à sucre économe en Herbicides
2. Acteur de la filière canne à sucre de La Réunion, le centre de recherche, eRcane participe à l'accroissement de la production, à sa rentabilité et à sa pérennité, par une présence à l'avant-garde du progrès génétique et technologique.
3. Findable Accessible Interoperable Reusable ; la démarche FAIR c'est construire, stocker, présenter ou publier des données de manière à permettre que la donnée soit « trouvable, accessible, interopérable et réutilisable ».
4. En recherche agronomique, un essai est une culture expérimentale réalisée en plein champ, c'est-à-dire en milieu non confiné physiquement, destinée à vérifier les informations acquises précédemment par des raisonnements théoriques ou grâce à des expériences en milieu confiné (serre ou laboratoire).
5. En agriculture, une plante de service est une plante installée dans un agroécosystème pour profiter des services écosystémiques qu'elle fournit, mais qui n'est pas récoltée.
6. Adventice : plante qui pousse dans un milieu aménagé sans y avoir été intentionnellement introduite

Essai	Années d'expérimentation
A_Crescence	2015
Baleya	2014
Canne_Bio	2021
O.Fontaine	De 2016 à 2020
P12	De 2018 à 2020
P25_ITK	De 2013 à 2015
P25_LEG	De 2012 à 2016
P25_montagne	De 2018 à 2020
P41	De 2015 à 2021
P42	De 2016 à 2020

Table 1 – Récapitulatif des essais

Ces données ont été collectées par différents chercheurs sur plusieurs sites. Elles incluent des données qui décrivent le dispositif expérimental mis en place, des mesures de croissance (biomasse, taux de recouvrement) de la canne à sucre et des plantes de service ainsi que des observations (notation, relevé floristique) à l'échelle de chaque espèce d'adventices dans les parcelles. En plus de ces données mesurées, le jeu de donnée inclue tout l'itinéraire technique des parcelles expérimentales.

Le Dataverse (King 2007) est une application web open source qui permet de préserver, citer, partager et explorer des données de recherche. Le CIRAD a choisi le Dataverse comme entrepôt de données institutionnel, afin que les scientifiques et les partenaires du Sud puissent déposés les données produites dans le cadre de leurs activités de recherche, d'expertise et de formation en partenariat. La première partie du stage a donc été un travail de structuration, de standardisation et d'homogénéisation des jeux de données destinés à être stockés dans AEGIS et déposés sur le dataverse du CIRAD pour être diffusés et réutilisés.

Dans un premier temps, les fichiers ont été structurés avec un format spécifique. A partir de fichiers récapitulatifs qui rassemblent toutes les informations sur les essais, j'ai construit 33 dossiers. Chacun des dossiers est représenté par un essai sur une année. Chaque dossier a une nomenclature bien précise, à savoir, le nom de l'essai et l'année à laquelle les données ont été recueillies pour cet essai.

Par exemple, 'O.Fontaine_2016' et 'O.Fontaine_2017' sont deux dossiers différents qui regroupent des données particulières de l'essai O.Fontaine sur les années 2016 et 2017. Le nom de ces différents dossiers et des fichiers qu'ils contiennent est aussi très important pour avoir une

information de premier plan. On peut noter par exemple dans le dossier 'A.Crescence_2015' que le fichier "A.Crescence_2015_notes" représente le fichier de notation de l'essai A.Crescence sur l'année 2015. En une lecture du nom, toute l'information est présente.

Pour chaque dossier (chaque essai) :

- Un fichier de plusieurs onglets, parmi lesquels :
  - Un onglet 'design' qui décrit tout le dispositif expérimental de l'essai, à savoir les blocs expérimentaux, les parcelles, tout ce qui se trouve à l'échelle en dessous de la parcelle (rang, inter-rang, plantes de services, cane, adventices et pour chaque parcelle la liste des 125 espèces d'adventices). Sachant qu'un essai contient plusieurs blocs et qu'un bloc contient plusieurs parcelles (Annexe B.1), chaque parcelle respecte le schéma d'arborescence suivant 2

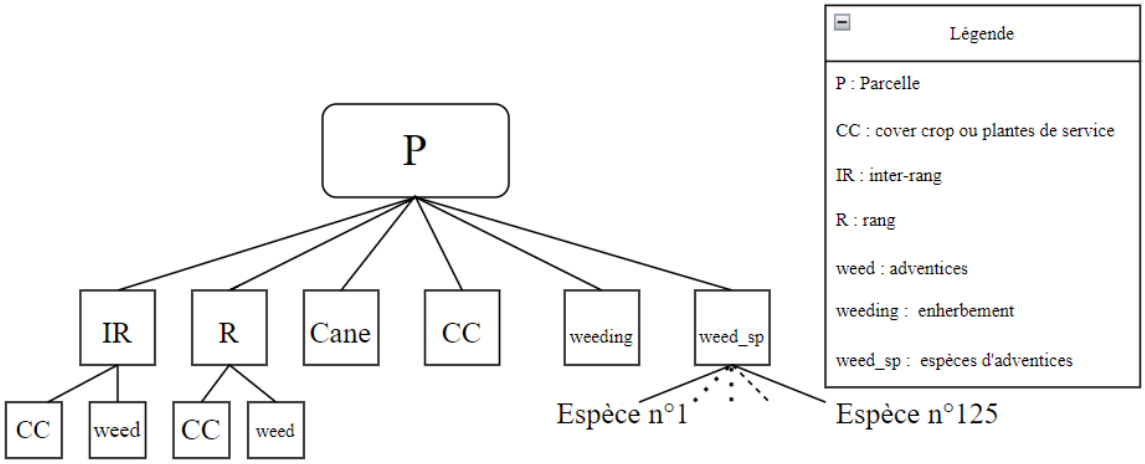


FIGURE 2 – Arborescence du dispositif expérimental.

    La page 'design' suit alors le code de notation de l'arborescence ci-dessus en fonction des parcelles de l'essai (par exemple, P11_R indique rang de la parcelle P11). On se retrouve avec des feuilles de 'design' de plusieurs milliers de lignes
  - Deux pages 'Treatment_Level' et 'Treatment_Unit' qui regroupent respectivement les traitements utilisés (pour lutter contre les adventices) sur l'essai entier et sur chacune des parcelles que contient cet essai.
- Un fichier de notation à l'échelle des unités expérimentales qui indique l'état d'enherbement général, l'état d'enherbement des adventices et l'état d'enherbement des plantes de service à l'échelle de la parcelle, du rang et de l'inter-rang.
- Un fichier d'observation à l'échelle des parcelles qui, pour chaque parcelle, indique le Yield_CAS (rendemment en canne, en tige fraîche et utilisable), le Sugar_CAS (teneur en sucre de la masse de tige fraîche) et l'IFTH (indice de fréquence d'application d'her-

bicides).

J'ai codé des scripts R pour automatiser la restructuration et l'harmonisation des données. Cela m'a permis de gagner du temps et d'être plus performant et efficace. Ces données m'ont servi de base de travail pour la deuxième partie de mon stage sur la fouille de texte.

3 Méthodes de fouille de textes d'aide à la normalisation

3.1 Contexte

Dans le cadre de ses recherches pour favoriser le développement des cultures tropicales, en l'occurrence la canne à sucre, le CIRAD investit énormément pour faire des plans d'expérimentation. Ces plans d'expérimentation sont retranscrits par les chercheurs en fichiers de données et ont pour vocation à être capitalisés dans une base de données générique AEGIS (Agro – Ecological Global Information System). Cependant, il y a autant de fichiers de données qu'il y'a de chercheurs, et chacun d'entre eux a sa propre façon de nommer les variables et de les décrire. Le problème qui se pose est alors de pouvoir faire correspondre les variables de toutes ces données de chercheurs avec les variables communes fixés pour AEGIS pour pouvoir les y intégrer.

N.B : Pour éviter tout souci d'ambiguïté, il est important de noter que par la suite, le terme "variable source" fait référence à "variable chercheur" et le terme "variable candidate" fait référence à "variable commune" (celle qui vient d'AEGIS).

3.2 Processus global

Pour optimiser le processus de lien entre variables chercheurs et variables communes dans AEGIS, l'objectif est de proposer automatiquement pour une variable donnée par un chercheur, les variables les plus proches dans des bases de données existantes.

Le schéma du processus globale est le suivant (Figure 3)

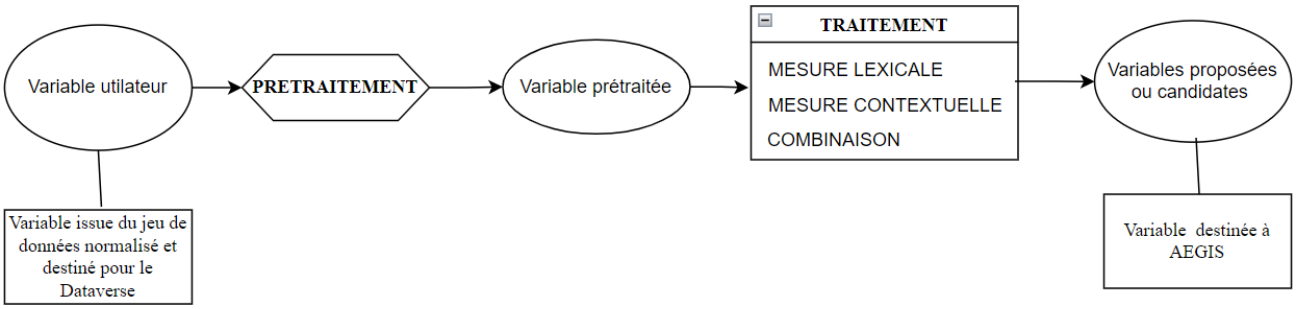


FIGURE 3 – Processus global du traitement des variables.

A ce stade du projet, nous disposons de deux sources de données pour mettre en place nos processus de mise en correspondance, des données textuelles des chercheurs (ou variables utilisateurs) et des données textuelles venant du dictionnaire des variables d'AEGIS (ou variables à proposer).

Ces deux sources de données ont les configurations suivantes :

1. les variables sources (celles qui viennent des chercheurs) ont comme configuration :
   — leur nom de variable
   — leur description
   — leur unité

   Par exemple, ci-dessous trois variables sources (voir Table 2) :

Variable source	Description variable source	Unité
Yield_CAS	Cane yield (in fresh machinable stem)	t.ha-1
Sugar_CAS	Sugar content of fresh stem mass	%
IFTH	Herbicide Application Frequency Index	[0,1]

Table 2 – Trois exemples de variables sources.

2. les variables candidates (celles qui viennent d'AEGIS) ont comme configuration :
   — leur nom de variable
   — leur description
   — leur classe
   — leur sous-classe
   — leur domaine

   Par exemple, ci-dessous trois variables candidates (voir Table 3) :

Variable commune(cand)	Description candidate	Class	Subclass	Domain
root_crop_yield_dm_t.ha1	Measurement of root dry biomass at plot level	experimental variable	Plant	agronomical traits
ferti_K_app_rate_kg.ha1	potassium application rate for soil fertilization	itk	organic fertilization	soil traits
abv_sugar_fm_content_%	percentage of sugar of the fresh matter overground biomass	experimental variable	plant	biomass quality traits

TABLE 3 – Trois exemples de variables candidates.

3.3 Description des étapes du processus

Pour automatiser le lien entre les variables sources (ou variables des chercheurs) et les variables communes ou candidates, nous proposons d'utiliser et adapter des méthodes de fouille de texte (*text mining*).

Nous avons choisi d'utiliser deux principales approches qui peuvent être combinées :

- **Mesure lexicale** : le but de cette approche est de comparer les noms de variables sources et candidates sur la base de leur chaîne de caractères. Pour cette approche, nous avons appliqué la distance de Levenshtein (NOVIKOV 1966) (voir section 3.3.2) qui calcule le nombre de changements entre deux chaînes de caractères, à savoir les noms de variables sources et candidates.
- **Mesure contextuelle** : l'objectif de cette approche est de comparer les variables à travers leur description. Cette description sous forme de "sac de mots" représentera le contexte textuel de chaque variable. Ces contextes pourront être comparés avec des mesures de similarité, en particulier la mesure cosinus (HAN 2012) (voir section 3.3.3).

Enfin, ce contexte sera enrichi en intégrant des corpus textuels (articles scientifiques). Cette extension est décrite dans la section suivante.

3.3.1 Prétraitement

Avant d'appliquer chacune de ces deux méthodes, nous sommes passés par une phase de prétraitement et de normalisation du texte. Cela commence par des opérations élémentaires de normalisation de texte (corpus ou vocabulaire).

1. Mettre le texte en minuscule, mais au risque de perdre les capitales initiales qui indicent les noms communs.
2. Supprimer la ponctuation, au risque de déstructurer les phrases.
3. Rajouter les unités aux variables sources pour améliorer les résultats de Levenshtein.
   En effet, quasiment tous les noms de variables candidates ont leurs unités imbriquées dans leurs noms. Ce qui n'est pas le cas par défaut pour les variables sources.
   Par exemple, la variable candidate « root crop yield dm t.ha-1 » a pour unité 't.ha-1' et on s'aperçoit que cette dernière est clairement imbriquée dans le nom. Par contre, ce n'est pas le cas par défaut pour les variables sources : on voit que la variable source « Yield CAS » est nommée sans son unité qui est 't.ha-1'. Vu que l'approche lexicale (la méthode de Levenshtein) s'appuie principalement sur la comparaison des chaînes de caractères, en rajoutant aussi les unités correspondantes dans les noms des variables sources, les résultats de Levenshtein devraient s'améliorer. Par exemple, (calcul appliqué avec l'équation (2) plus bas) Lev(« Yield CAS », « root crop yield dm t.ha-1 ») = 0,20, alors que Lev(« Yield CAS t.ha-1», « root crop yield dm t.ha-1 ») = 0,44

3.3.2 Mesure lexicale

La distance de Levenshtein (Novikov 1966) est une distance, au sens numérique (ou au sens mathématique), donnant un résultat qui reflète la différence (ou la similitude) entre deux chaînes de caractères. Elle calcule la ressemblance entre les deux chaînes de caractères que nous notons C1 et C2. Il s'agit du coût minimal, i.e. du nombre minimal d'opérations à réaliser sur C1, pour transformer C1 en C2. Les opérations sont les suivantes (Dumas 2016) :

— substitution d'un caractère de C1 en un caractère de C2,

— ajout dans C1 d'un caractère de C2,

— suppression d'un caractère de C1.

Pour obtenir la distance de Levenshtein entre les chaînes de caractères C1 et C2, il faut associer à chacune de ces opérations un coût. Le coût de chaque opération est toujours égal à 1, mais dans le cas où il faut substituer deux caractères identiques (ce qui revient à ne rien changer) ceci vaut 0. Notons que $Levenshtein(C1, C2) \in [0, \infty[$.

Par exemple, si la source est C1 = "écart" et la cible est C2 = "égare", le "c" et le "t" doivent être respectivement changés en "g" et "e" sans suppressions ou insertions supplémentaires, ainsi la distance de Levenshtein sera de 2.

3.3.3 Mesure contextuelle

Le but de l'approche contextuelle est de comparer les variables à travers leurs descriptions représentées sous forme de "sacs de mots". Dans cette section, le terme "contexte" ou "document" fera référence à la description d'une variable chercheur ou d'une variable candidate. Dans cette représentation textuelle, nous utiliserons les notations suivantes :

— d_j : description de la variable j.

— v_j : vecteur associé à la variable j. Ce vecteur est constitué des termes t_i où t_i fera référence au $i$ème terme composant la description d_j.

— D : ensemble des descriptions de variables.

La similarité cosinus est la mesure que nous avons considérée comme mesure contextuelle. Elle permet d'évaluer la ressemblance (ou la similitude) entre deux contextes (dans notre cas entre la description d'une variable chercheur (ou source) (d_j^s) et d'une variable candidate ($d_{j'}^c$)). Elle consiste à calculer le cosinus de l'angle entre les documents vectorisés :

$$cos(v_j^s, v_{j'}^c) = \frac{v_j^s \; v_{j'}^c}{\|v_j^s\| \|v_{j'}^c\|}, \tag{1}$$

où v_j^s est la vectorisation du document d_j^s et $v_{j'}^c$ est la vectorisation du document $d_{j'}^c$. Cette mesure est comprise entre 0 et 1.

La valeur 0 signifie que les deux contextes ne sont pas similaires tandis que la valeur 1 signifie qu'ils sont très proches (voire identiques).

Ainsi, il nous faut convertir chaque description de variable chercheur ou de variable candidate en un vecteur de termes. Plusieurs approches peuvent été proposées afin de pondérer les termes constituant les vecteurs. Dans ce contexte, nous avons appliqué une méthode largement utilisée dans la littérature, à savoir la méthode dite TF-IDF (**term frequency − inverse document frequency** (Salton et Buckley 1988)). L'objectif de cette approche est de donner un poids plus important aux termes discriminants composant les descriptions de variables.

Ainsi, un terme sera favorisé s'il est à la fois fréquent (TF - *Term Frequency* ou fréquence du terme élevé) et spécifique à certaines descriptions (cela correspond à une valeur d'IDF élevée).

La fréquence du $i$ème terme (TF_i) du document d_j est, par abus de langage, le nombre de fois où le terme apparaît dans le document ($TF_i = |t_{i'} : t_{i'} = t_i, \forall t_{i'} \in d_j|$). La fréquence inverse de document du terme t_i (IDF_i, pour *inverse document frequency*) du document d_j est

le logarithme de l'inverse de la proportion de documents du corpus D qui contiennent le terme t_i :

$$IDF_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}.$$

Le poids du $i$ème terme ($TF\text{-}IDF_i$) est alors donné par le produit des deux fréquences du $i$ème terme :

$$TF\text{-}IDF_i = TF_i \ . \ IDF_i.$$

L'ensemble des poids des termes d'un document d_j forme alors le vecteur v_j.

N.B : Un exemple de vectorisation utilisant td-idf est illustré en annexe A.

Récapitulatif des étapes de la méthode contextuelle

Les étapes de notre algorithme pour appliquer la mesure cosinus de similarité (pour les variables sources et candidates) sont les suivantes :

1. Je créé un corpus (ou vocabulaire) de mots constitué des variables candidates et de leurs descriptions
2. Je fais de ce vocabulaire un espace vectoriel appliquant la méthode tf-idf (décrite ci-dessus) pour affecter un poids à chacun des termes de ce vocabulaire
3. Sur cet espace vectoriel, je vectorise l'ensemble de toutes les descriptions des variables avant de pouvoir appliquer le calcul de cosinus qui nous intéresse

N.B : Pour affiner cette méthode, nous avons testé l'hypothèse d'enrichir le vocabulaire avec des informations supplémentaires sur les variables candidates, en l'occurrence leurs 'class', 'subclass' et 'domain'. Ainsi, nous avons testé pour toutes les descriptions des variables sources, simcosinus('description de variable source' , 'class'+ 'subclass' + 'domain'). Cependant, cette hypothèse n'a pas fonctionné comme nos l'espérions, alors nous ne l'avons pas pris en considération et nous avons mis de côté ces informations supplémentaires.

3.3.4 Pistes d'amélioration des performances de la mesure contextuelle

Pour explorer un peu plus l'approche contextuelle, nous avons fixé deux façons de faire :

— En tokenisant

— En tokenisant et avec lemmatisation.

- **Tokenisation**

  Une fois que les données textuelles sont nettoyées, la prochaine phase de traitement de ces données textuelles est la tokenisation. La tokenisation des documents consiste à identifier les unités de textes élémentaires, les termes, qui peuvent être soit des lettres, des syllabes, mots, phrase ou encore des séquences des ces éléments. Dans notre cas, on considère un mot comme l'unité élémentaire. Le vocabulaire (ou corpus) devient alors une liste ordonnée (ou non) de termes élémentaires : les tokens (Balech et Benavent 2019).

- **Tokenisation + lemmatisation**

  La lemmatisation est un traitement lexical qui a pour but de retirer uniquement les fins flexionnels [7] d'un mot et qui permet de renvoyer la forme de base (ou de dictionnaire du mot). Cette forme de base est connue sous le nom de lemme. En d'autres termes, la lemmatisation, consiste à ramener un terme, quels que soient ses accords, déclinaisons, etc. à sa forme la plus simple.

  Pour appliquer la lemmatisation, nous utilisons des balises POS (*Part of Speech*) (Charniak et al. 1993,Brants 2000). Nous ajoutons à chaque mot sa balise définissant son type (verbe, nom, adjectif, etc.). Notons que le balisage POS est un processus qui, pour un texte donné, attribut à chaque mot (en fonction de sa définition et de son contexte) de ce texte une balise ; cette balise est un marqueur qui permet d'indiquer si le mot est un nom, un adjectif, un adverbe ou un verbe, etc.

  Par exemple,

  - Mot + Type (balise POS) → Mot lemmatisé
  - conduite + verbe 'v' → conduire
  - chiens + nom 'n' → chien

3.3.5 Combinaison

Nous disposons maintenant de deux mesures pour comparer le groupe de variables chercheurs avec celui des variables candidates. La mesure de Levenshtein pour les noms de variables et la mesure de similarité cosinus pour les descriptions de variables.

Nous pouvons maintenant les combiner pour obtenir une mesure global de similarité. Toutefois, nous devons d'abord standardiser la mesure de Levenshtein qui est comprise entre 0 et l'infini.

7. Qui sert à marquer la flexion. Élément flexionnel. La désinence, c'est-à-dire la caractéristique flexionnelle ou élément variable de fin de mot qui distingue les formes d'un paradigme nominal ou verbal

Pour ce faire nous appliquons la formule :

$$Lev(C1, C2) = 1 - \frac{Levenshtein(C1, C2)}{\max(longueur(C1), longueur(C2))} \tag{2}$$

Nous avons décidé de fixer une mesure globale que nous avons appelé combi qui serait une combinaison linéaire entre les mesures de Levenshtein et de cosinus.

$$combi \ = \ \alpha \ cos \ + \ (1 - \alpha) \ Lev \ \text{où} \ \alpha \in]0, 1[. \tag{3}$$

Cette combinaison de mesure est comprise entre 0 et 1.

3.3.6 Règles de normalisation

Nous pouvons noter que toutes les variables (sources ou candidates) ont des '_'. Par exemple, Yield_CAS (nom de variable source) et root_crop_yield_dm_t.ha-1 (nom de variable candidate).

Pour regarder s'il y'a des changements au niveau des résultats, il a été proposé de remplacer les '_' par des espaces (' '). Nous avons observé que cela n'avait aucune incidence sur le résultat de la distance de Levenshtein. Par exemple : Lev (« Yield_CAS », « root_crop_yield_dm_t.ha-1 ») = Lev (« Yield CAS », « root crop yield dm t.ha-1 ») = 0,20 . Par contre cela améliore les résultats de similarité cosinus sur les descriptions de variables. En effet, en remplaçant les '_' par des espaces (' '), le vocabulaire s'enrichit ; le mot « root_crop_yield_dm_t.ha-1 » devient les 4 mots « root », « crop », « yield », « dm », « t.ha-1 ».

Par exemple, prenons le cas d'une variable source « WCE_dm » et une de ses variables candidates « stem_nonstalk_crop_yield_fm_t.ha-1 ».

- « WCE_dm » a pour description : « Weed control efficiency in terms of weed dry mass » et « stem_nonstalk_crop_yield_fm_t.ha-1 » a pour description : « stem crop yield fresh mass » Posons V1 et V2 les vecteurs représentants respectivement les descriptions de « WCE_dm » et « stem_nonstalk_crop_yield_fm_t.ha-1 »
- Dans le cas où on garde '_' :
  Lev (« WCE_dm », « stem_nonstalk_crop_yield_fm_t.ha-1 ») = 0,088235 Simcosinus(V1, V2) = 0,381940
- Dans le cas où on remplace '_' par ' ' :
  Lev (« WCE dm », « stem nonstalk crop yield fm t.ha-1 ») = Lev (« WCE_dm », « stem_nonstalk_crop_yield_fm_t.ha-1 ») = 0,088235 Simcosinus(V1, V2) = 0,440242

3.4 Méthode d'évaluation

Pour évaluer chacune des méthodes (Levenshtein, cosinus et la combinaison des deux), nous avons regardé pour chacune des variables sources, s'il y avait parmi les n premiers résultats (variables candidates), la solution pertinente (la vraie correspondance validée par l'expert terrain). Cela revient à calculer la Précision au rang n (P@n).

Notons qu'à ce stade du projet, nous disposons de 84 variables sources et 170 variables candidates. Pour chacune des 84 variables sources, nous appliquons les différentes méthodes de calculs (les mesures Lev et Cos) avec toutes les 169 variables candidates. Notons aussi que des variables sources peuvent avoir une même solution pertinente dans AEGIS .

Pour regarder l'impact ou non de la lemmatisation, nous avons essentiellement regardé les résultats de la méthode cosinus (avec et sans lemmatisation).

Exemple : Résultats obtenus avec la variable source Yield CAS (Voir Figure 4)

variable source	variables candidates avec les meilleurs scores sur cosinus	scores
Yield CAS t.ha-1	stem nonstalk crop yield fm t.ha-1	0.545656
Yield CAS t.ha-1	stem nonstalk crop yield fm t.ha-1	0.545656
Yield CAS t.ha-1	stem plant fm kg	0.517871
Yield CAS t.ha-1	plant stalk height sum m.m-2	0.469830
Yield CAS t.ha-1	stem juice crop yield l.ha-1	0.421439

(a) Sans lemmatisation

variable source	variables candidates avec les meilleurs scores sur cosinus	scores
Yield CAS t.ha-1	stem nonstalk crop yield fm t.ha-1	0.557441
Yield CAS t.ha-1	stem nonstalk crop yield fm t.ha-1	0.557441
Yield CAS t.ha-1	stem plant fm kg	0.533317
Yield CAS t.ha-1	stem juice crop yield l.ha-1	0.439870
Yield CAS t.ha-1	stem crop yield fm t.ha-1	0.358853

(b) Avec lemmatisation

FIGURE 4 – Illustration de l'influence de la lemmatisation sur sur une seule variable : la variable source Yield CAS

Sur cet exemple, nous avons considéré n=5. Nous observons que dans les 2 cas, les variables candidates ayant les meilleurs scores sont les quasiment mêmes (4 sur 5), aussi les scores sont quasiment identiques. La lemmatisation semble donc avoir un grand effet sur le calcul du cosinus. Pour confirmer ou infirmer cette hypothèse, nous avons étudié les résultats de manière globale.

Regardons l'extrait du tableau de toutes les variables sources sur les calculs de Levenshtein et du cosinus (voir Figure 5), pour expliquer notre processus d'évaluation.

	Variable source	Levenshtein				cosinus			
		précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10
0	Yield CAS t.ha-1	0	0	0	1	0	0	1	1
1	Sugar CAS %	0	1	1	1	0	0	1	1
2	IFTH	0	0	0	0	1	1	1	1
3	Rec globale plein %	0	0	0	0	0	0	0	1
4	Rec globale B %	0	0	0	0	0	0	0	1

FIGURE 5 – Extrait du tableau d'évaluation de Lev et Cos.

Toutes les cellules du tableau sont initialisées à 0. Pour une variable source X donnée, on affecte 1 à la cellule correspondante à la méthode M et au rang N si et seulement si la solution pertinente (vraie variable liée à X) fait partie des N meilleur(s) score(s) pour la méthode M. Par exemple, dans le tableau, si nous regardons la ligne n°1, on voit que la solution pertinente pour la variable source « Sugar CAS % » est dans les 3 meilleurs scores selon la méthode de calcul de Levenshtein mais ce n'est pas le cas pour la méthode de calcul de cosinus. Notons que si une variable source à la précision au rang 1, cela signifie que c'est la variable candidate associée ayant obtenu le meilleur score de calcul qui est la solution pertinente.

4 Enrichissement du contexte pour améliorer la mise en lien

4.1 Extension de la méthode contextuelle par enrichissement du contexte par la fouille dans la littérature scientifique

4.1.1 Documents choisis

Pour travailler sur cette approche, notre corpus de documents est constitué d'articles scientifiques, de chapitres d'ouvrage, de rapports et de thèses. Tous ces documents principalement autour des thèmes de la canne à sucre, de la fertilisation des sols, des plantes de service et des adventices. Il est à noter que tous ces documents ont été triés et sélectionnés par ma tutrice Sandrine Auzoux et sont tous rédigés en anglais. J'ai disposé au total de 122 documents. Ces documents contiennent tous, entre 5 et 160 pages.

4.1.2 Principe de la méthode de fouille texte

Pour utiliser cette méthode, nous avons sélectionné quelques variables sources qui, jusqu'à ce niveau du projet, n'avaient pas au moins une précision de rang 10 selon notre critère d'évaluation. Avec ces variables, nous avons appliqué la méthode qui suit :

1. Considérer les descriptions de chacune d'entre elles

2. Dans ces descriptions, je considère les termes clés (ou les mots essentiels de leurs descriptions) (voir annexe B.3)
3. Pour chacun de ces termes, je fouille dans les 122 documents supplémentaires les n mots avant et après qui l'entourent. Par exemple, pour le terme 'cane yield', à chaque fois qu'il apparaît dans un document, je récupère les n mots à gauche et les n mots à droite.
4. Ensuite, je considère les m premiers mots qui apparaissent les plus fréquents pour constituer le corpus associé aux variables sources n'ayant pas une précision de rang 10.

4.2 Résultats

4.2.1 Sans l'enrichissement contextuel

Les figures 6 et 7 donnent les récapitulatifs des performances des méthodes de Levshtein et de similarité cosinus sans et avec la lemmanisation respectivement.

1. Sans lemmatisation :

Nombre de bons résultats (sur les 84 variables sources)							
Levenshtein				**cosinus**			
précision au rang 1	**Rang 3**	**Rang 5**	**Rang 10**	**précision au rang 1**	**Rang 3**	**Rang 5**	**Rang 10**
15.48%	19.05%	23.81%	42.86%	29.76%	44.05%	51.19%	58.33%

FIGURE 6 – Récapitulatif de l'évaluation de Lev et Cos **sans lemmatisation**

2. Avec lemmatisation :

Nombre de bons résultats après lemmatisation (sur les 84 variables sources)							
Levenshtein				**cosinus**			
précision au rang 1	**Rang 3**	**Rang 5**	**Rang 10**	**précision au rang 1**	**Rang 3**	**Rang 5**	**Rang 10**
15.48%	19.05%	23.81%	38.10%	33.33%	44.05%	51.19%	60.71%

FIGURE 7 – Récapitulatif de l'évaluation de Lev et Cos **avec lemmatisation**

Le premier constat que nous pouvons faire est que la méthode cosinus est plus performante que la méthode de Levenshtein.
Notons aussi que pour la méthode de Levenshtein, les résultats sont exactement les mêmes, ce qui est tout à fait normal car cette méthode de calcul ne prend pas du tout en compte la lemmatisation. Par contre, il est clair que la lemmatisation améliore les résultats de calcul de la méthode cosinus.

Nous allons maintenant regarder l'efficacité de la combinaison de mesures (voir équation (3)). Il est question de trouver le α qui donne le meilleur compromis entre les mesures de Levenshtein et de cosinus.
Suivant le même raisonnement (pour l'évaluation des méthodes Levenshtein et cosinus) que précédemment, nous avons fait des récapitulatifs sur la mesure combi suivant différentes valeurs de α entre 0,1 et 0,9 par pas de 0,1. On obtient les résultats suivants :

1. Sans lemmatisation,

Nombre de bons résultats (sur les 84 variables sources)		
Précision	Le α donnant le meilleur résultat	Meilleur résultat
Rang 1	0,3	40,48%
Rang 3	0,5	53,57%
Rang 5	0,5	61,90%
Rang 10	0,5	71,43%

2. Avec lemmatisation,

Nombre de bons résultats (sur les 84 variables sources)		
Précision	Le α donnant le meilleur résultat	Meilleur résultat
Rang 1	0,3	41,67%
Rang 3	0,6	54,76%
Rang 5	0,5	64,29%
Rang 10	0,5	73,81%

Les résultats ci-dessus confirment la tendance selon laquelle la lemmatisation améliore la mesure de cosinus.

4.2.2 Après l'enrichissement contextuel

Il faut noter qu'après une trentaine de configurations du couple (n,m) (voir paragraphe 4.1.2), les valeurs avec lesquelles j'ai obtenu de meilleurs résultats sont pour n = 6 et m = 20.
Après avoir enrichi le contexte par la fouille dans la littérature scientifique tel que décrit dans le paragraphe précédent, on obtient une légère amélioration des résultats par rapport aux résultats

précédents.

1. Sans lemmatisation :

Nombre de bons résultats (sur les 84 variables sources)							
Levenshtein				**cosinus**			
précision au rang 1	**Rang 3**	**Rang 5**	**Rang 10**	**précision au rang 1**	**Rang 3**	**Rang 5**	**Rang 10**
15.48%	19.05%	23.81%	42.86%	34.52%	44.05%	51.19%	58.33%

FIGURE 8 – Récapitulatif de l'évaluation de Lev et Cos **sans lemmatisation**

2. Avec lemmatisation :

Nombre de bons résultats après lemmatisation (sur les 84 variables sources)							
Levenshtein				**cosinus**			
précision au rang 1	**Rang 3**	**Rang 5**	**Rang 10**	**précision au rang 1**	**Rang 3**	**Rang 5**	**Rang 10**
15.48%	19.05%	23.81%	42.86%	34.52%	45.24%	52.38%	60.71%

FIGURE 9 – Récapitulatif de l'évaluation de Lev et Cos **avec lemmatisation**

Nous allons maintenant étudier l'efficacité de la mesure combi (voir équation 3). Nous allons simplement suivre le même principe utilisé ci-dessus où il est question de trouver le α qui donne le meilleur compromis entre les mesures de Levenshtein et de cosinus.

1. Sans lemmatisation,

Nombre de bons résultats (sur les 84 variables sources)		
Précision	Le α donnant le meilleur résultat	Meilleur résultat
Rang 1	0,3	42,86%
Rang 3	0,2	54,76%
Rang 5	0,3	63,10%
Rang 10	0,3	71,43%

2. Avec lemmatisation,

Nombre de bons résultats (sur les 84 variables sources)		
Précision	Le α donnant le meilleur résultat	Meilleur résultat
Rang 1	0,3	44,05%
Rang 3	0,6	55,95%
Rang 5	0,3	64,29%
Rang 10	0,3	73,81%

Avec ces résultats, nous pouvons nous apercevoir qu'avec ces méthodes de fouille de textes d'aide à la normalisation et le traitement que nous proposons, le meilleur résultat que nous pouvons espérer est de 73,81% pour une précision au rang 10. Cela signifie concrètement que pour toute variable chercheur, ce modèle peut proposer 10 variables candidates, et qu'il y aura 73,81% de chance que la solution pertinente (c'est-à-dire la variable candidate qui correspond vraiment à la variable chercheur) se trouve parmi les 10.

A ce stade du projet, le travail qui a été réalisé (à savoir les sections 2, 3 et 4) contribut à faciliter l'intégration des données chercheurs dans la base de données AEGIS. Ceci correspond à la phase 1 de la figure 10. Cependant, une fois les données intégrer dans AEGIS, l'autre challenge du projet est d'automatiser la simulation de STICS à partir de données provenant de AEGIS. Nous essayons de répondre à ce challenge dans la section suivante (à savoir 5). Cela correspond à la phase 2 de la figure 10.

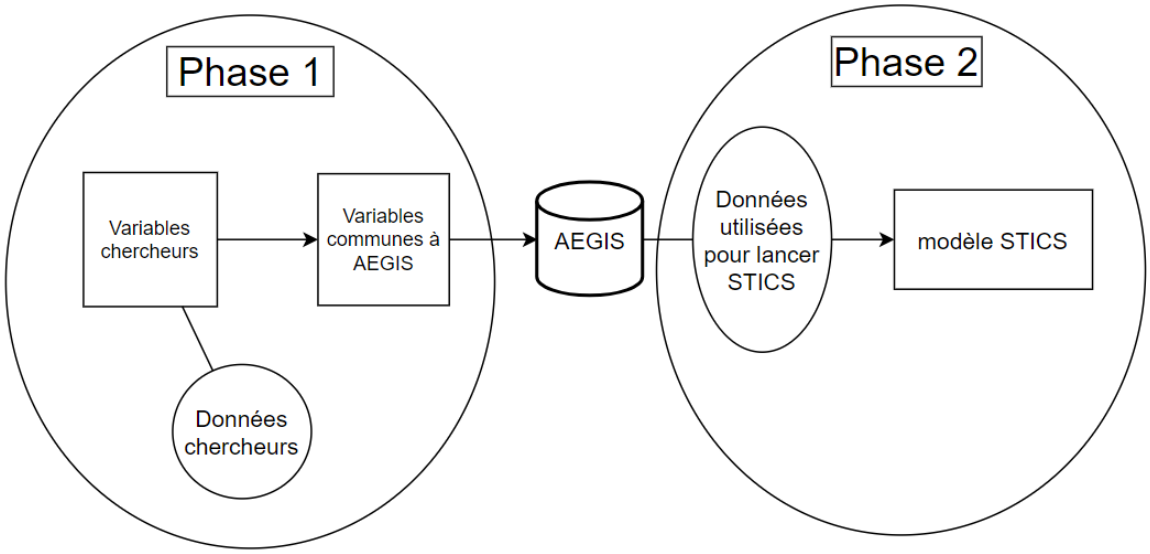


FIGURE 10 – Les deux principales phases du projet

5 Simulation du modèle STICS

Le modèle STICS tel que défini à la section 1 est une plateforme du CIRAD qui permet de simuler en particulier plusieurs facteurs [8] de plusieurs essais donnés. Il est à noter que STICS a été initialement créé sous Fortran, et le modèle a été adapté sous javascript pour faciliter l'utilisation de l'interface graphique.

Pour exécuter une simulation, STICS prend en entrée des fichiers sous forme html. Des packages R ont été développés par l'équipe du projet STICS pour créer ces fichiers html automatiquement à partir de fichiers csv ('input') regroupant les informations suivantes :

— Un fichier 'itk' regroupant les itinéraires techniques [9] du facteur choisi dans les différents essais à simuler,

— Un fichier 'ini' regroupant les conditions initiales du sol (au début des itinéraires techniques) du facteur dans ces essais,

— Un fichier 'obs' d'observation qui donne les mesures expérimentales observées du facteur dans ces essais,

— Un fichier 'weather' regroupant les informations climatiques relatives à ces essais,

— Un fichier 'stat' regroupant les informations des stations météo relatives à ces essais,

— Un fichier 'sol' regroupant les informations sur les états du sol (au cours des itinéraires techniques) de ces essais,

— Un fichier 'USM' qui définit l'unité (ou les unités) de simulation. Pour un essai donné, il renseigne les noms des fichiers qui contiennent les informations ci-dessus ('itk', 'ini', 'obs', 'climat', 'station' et 'sol') liées à cet essai.

Mon travail a été d'automatiser sous R la création de chacun des fichiers décrits ci-dessus en fonction de la simulation qu'un utilisateur aimerait réaliser sur un ou plusieurs essais. Le processus d'automatisation suit les étapes suivantes :

1. Un fichier R contenant les scripts permettant la connexion à la base de donnée AEGIS sous PostGreSQL.

2. En fonction du (ou des) facteur(s) dans l'essai (ou des essais) à simuler, je fais une sélection grâce à des requêtes SQL dans AEGIS des tables qui regroupent les données relatives à STICS ; puis j'enregistre ces tables dans des fichiers CSV.

3. Un fichier R contenant les fonctions permettant de créer les fichiers input décrits ci-

8. toute série d'éléments de même nature conditionnant le phénomène étudié, en agriculture le comportement agronomique de la culture (par exemple la variété, la dose d'engrais, etc.).

9. L'itinéraire technique est un enchaînement logique et ordonné d'interventions techniques culturales permettant de tirer le maximum des potentialités du milieu vis-à-vis d'une production donnée, du semis la récolte.

dessus ; ces fonctions exploitent les tables de l'étape n°2 où se trouvent les informations nécessaires pour la simulation. Chaque fonction correspondant à un fichier en particulier. Par exemple, la fonction INPUTTEC permet de créer un fichier 'Tec'.

4. Une fois ces fichiers créés, ils sont exploités par des fonctions (définies dans un fichier R) qui utilisent les packages R relatifs à STICS et qui permettent de créer les fichiers html lisibles par la STICS.

5. La simulation peut alors être effectuée.

Le processus d'automatisation est alors illustré par la figure 11.

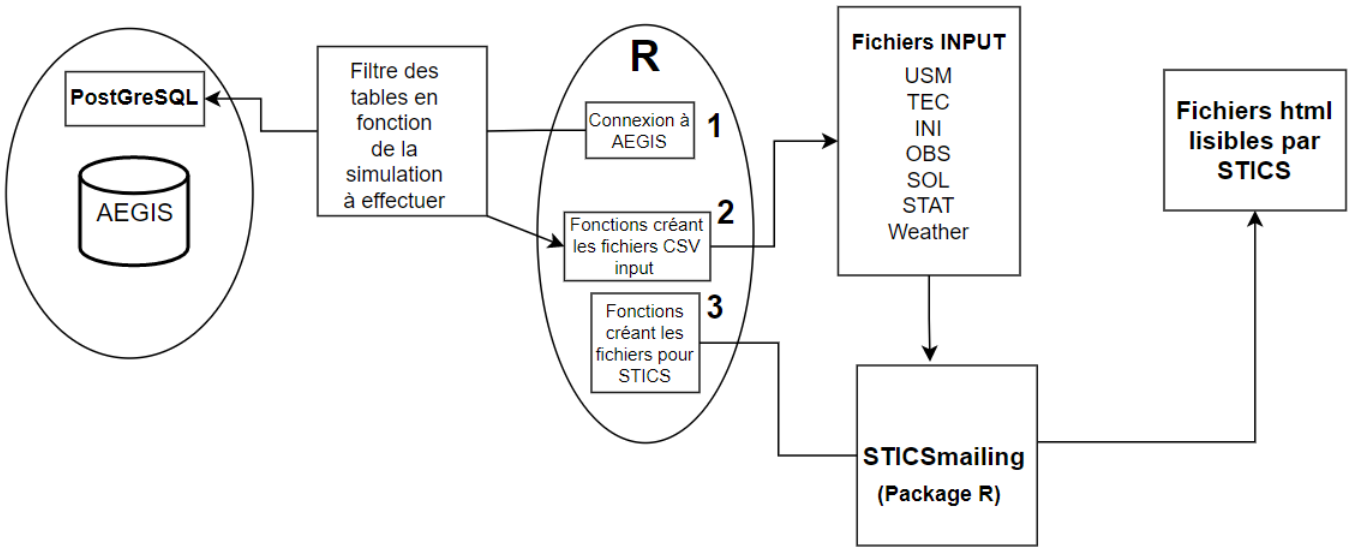


FIGURE 11 – Automatisation de la chaîne de traitement permettant de réaliser une simulation du modèle STICS

A titre d'exemple, nous avons effectué la simulation sous STICS (en utilisant le processus d'automatisation décrit ci-dessus) du facteur variété de canne à sucre sur l'essai nommé Delicas (voir figure 12).

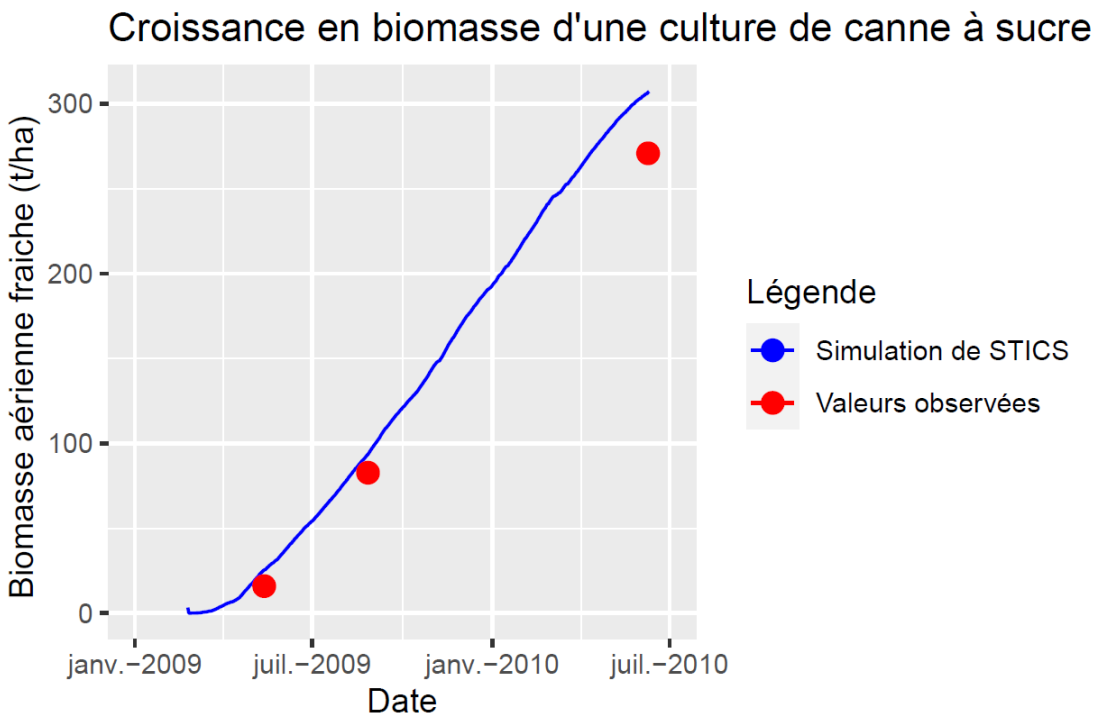


FIGURE 12 – Simulation d'une variété de canne sur Delicas

Cette figure (Figure 12) illustre l'évolution d'une culture de canne à sucre sur l'essai Delicas entre deux coupes [10] (de janvier 2009 à juillet 2010). Elle montre en particulier (en bleu) la simulation de STICS sur la croissance en biomasse aérienne fraîche d'une culture de canne à sucre entre ces 2 coupes et elle met en parallèle 3 valeurs réelles (en rouge) observées respectivement le 13 mai 2009, le 27 août 2009 et le 9 juin 2010.

10. une coupe est une récolte

6 Conclusion

Aspect pédagogique

Au fil du stage, nous avons effectué un travail qui couvrait plusieurs thèmes en lien avec la science des données :

- la structuration, l'harmonisation et le traitement des données, qui a été réalisé sous R,
- la fouille de texte, qui a été réalisé sous python
- la simulation d'un modèle STICS avec R combiné à la gestion de bases de données qui demandait de maîtriser SQL.

Ce stage m'a permis de développer une grande palette de compétences en R, Python et SQL.

Néanmoins, j'ai rencontré plusieurs difficultés au cours de ce stage.

- La première a été de devoir s'adapter à l'évolution de la structure du jeu de données normalisé à construire. En effet, pour pouvoir préparer le jeu de données, il a fallu se référer à une structure particulière à respecter. Cependant, cette structure a été plusieurs fois modifiée et il a fallu à chaque fois tout remettre à plat et s'y adapter.
- La disparité des données textuelles a rendu difficile la mise en place des méthodes de fouille de textes pour aider à la normalisation. En effet, les noms des variables des chercheurs sont généralement très différents de ceux des variables d'AEGIS. Ce qui explique aussi que les résultats de la mesure lexicale (voir les résultats de la distance de Levenshtein dans 4.2) ne sont pas élevés, vu que la mesure lexicale s'appuie sur les noms de variables.

Ce stage a été très enrichissant sur le plan pédagogique. Un Data paper [11] dont je suis co-auteur, est en cours de rédaction afin de valoriser le travail effectué tout au long du stage sur les données.

Résolution de la problématique

Dans le cadre de mon stage, j'ai contribué à :

- Enrichir le dataverse du CIRAD avec les nouveaux jeux de données harmonisés et standardisés.

  Les scripts R réalisés dans cette partie pour créer les fichiers nécessaires à l'importation dans AEGIS pourront être facilement adaptés à d'autres jeux de données expérimentaux.
- Automatiser la correspondance entre les variables des chercheurs agronomes et les variables d'AEGIS.

  Les méthodes de fouille de textes pour aider à la normalisation peuvent être améliorées

11. article scientifique évalué par les pairs et citable.

mais les résultats qui ont déjà été obtenus montrent que ce travail constituera déjà une bonne base pour des futurs travaux qui seront appliqués à ce sujet.

— Etablir la chaîne de traitement permettant de récupérer des données spécifiques dans AEGIS et de réaliser une simulation du modèle STICS.

  Dans cette partie, nous avons réalisé qu'un seul exemple de simulation d'une variété de canne à sucre avec un modèle spécifique. Cependant les travaux qui ont permis cette simulation pourront être généralisés sur plusieurs autres modèles de cultures (comme le modèle MOSICAS [12] par exemple) et sur d'autres types de cultures (comme le riz ou le sorgho).

Perspectives

Afin d'améliorer les méthodes de fouille de texte abordées dans ce projet et de continuer d'aider à l'automatisation de la mise en correspondance entre les variables expérimentales et de simulation de STICS, nous pouvons :

— Construire un package pour faciliter la mise en œuvre du nettoyage des variables

— Enrichir ce contexte de manière encore plus large en faisant du web crawling (recherche sur internet) en plus des articles utilisés pour la fouille dans la littérature scientifique.

— Utiliser des approches de modèles de langues en Deep learning (MacCartney et Manning 2007,Devlin et al. 2018,Lewis et al. 2019)

12. Modèle de fonctionnement de d'élaboration du rendement canne

7 Accessibilité des données et des codes sources

Les jeux de données provenant des 10 essais menées de 2012 à 2020 sur l'île de La Réunion sont mis à disposition dans un sous-dataverse dédié spécifiquement aux études sur les plantes de services, nommé "APEEDAIS", à l'adresse suivante : `https://dataverse.cirad.fr/dataverse/APEEDAIS`.

Le jeu de données expérimental (Ngaba et al. 2022) utilisé pour la fouille de texte a été déposé sur le dataverse Aïda, à l'adresse suivante : `https://doi.org/10.18167/DVN1/XDHKR8`.

Les codes sources pour structurer les jeux de données expérimentaux, mettre en liaison les variables candidates et communes, et lancer une simulation de croissance de la canne à sucre sous STICS à partir d'AEGIS ont été déposés sur github à l'adresse suivante : `https://github.com/bilson98/STAGE_Cirad`.

Références

Auzoux, Sandrine et al. (2019). "AEGIS, an extended information system to support agroecological transition for sugarcane industries". In : ISSCT. doi : https://agritrop.cirad.fr/593440/.

Balech, Sophie et C Benavent (2019). "Les techniques du NLP pour la recherche en sciences de gestion". In : p. 6. doi : https://hal.archives-ouvertes.fr/hal-02400308/document.

Brants, Thorsten (2000). "TnT-a statistical part-of-speech tagger". In : *arXiv preprint cs/0003055*.

Brisson, Nadine et al. (1998). "STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn". In : *Agronomie* 18.5-6, p. 311-346. doi : https://hal.archives-ouvertes.fr/hal-00885888.

Charniak, Eugene et al. (1993). "Equations for part-of-speech tagging". In : *AAAI*. T. 11, p. 784-789.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In : *arXiv preprint arXiv:1810.04805*.

Dumas, Léonard (2016). "Conception de formes de relecture dans les chaînes éditoriales numériques". In : doi : https://ics.utc.fr/~led/co/eart_1_4.html.

Han, Jiawei (2012). "Cosine Similarity". In : *Data Mining (Third Edition)*. doi : https://www.sciencedirect.com/topics/computer-science/cosine-similarity.

King, Gary (2007). *An introduction to the dataverse network as an infrastructure for data sharing*.

Lewis, Mike et al. (2019). "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In : *arXiv preprint arXiv:1910.13461*.

MacCartney, Bill et Christopher D Manning (2007). "Natural logic for textual inference". In : *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p. 193-200.

Ngaba, Billy et al. (2022). *Experimental dataset for mapping researcher variables from service plant trials to AEGIS dictionary variables*. Version V1. doi : 10.18167/DVN1/XDHKR8. url : https://doi.org/10.18167/DVN1/XDHKR8.

Novikov, P.S. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In : *Soviet physics doklady* 10.8, p. 707-710. doi : https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Salton, Gerard et Christopher Buckley (1988). "Term-weighting approaches in automatic text retrieval". In : *Information processing & management* 24.5, p. 513-523.

WILKINSON, Mark D et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In : *Scientific data* 3.1, p. 1-9.

Annexes

A Exemple de représentation vectorielle de textes avec td-idf :

Tiré de https ://dotnettutorials.net/lesson/tf-idf-and-cosinesimilarity-in-machine-learning/

Soit le corpus constitué des 3 documents suivants :

— Document 1 : She loves food with cheese.

— Document 2 : Her favorite food is Italian.

— Document 3 : She lives in the Italian state.

Regardons à quoi ressemblera la "term-frequency" (TF) normalisée :

Document	She	Loves	Food	With	Cheese	Her	Favourite	is	Italian	Lives	in	State
Doc 1	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0	0	0	0	0
Doc 2	0	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0
Doc 3	$\frac{1}{5}$	0	0	0	0	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

FIGURE 13 – Calculs de TF.

Maintenant, regardons les valeurs IDF de ces termes :

Term	IDF
She	0.58
Loves	1.58
Food	0.58
With	1.58
Cheese	1.58
Her	1.58
Favourite	1.58
Is	1.58
Italian	0.58
Lives	1.58
In	1.58
state	1.58

FIGURE 14 – Calculs des IDF.

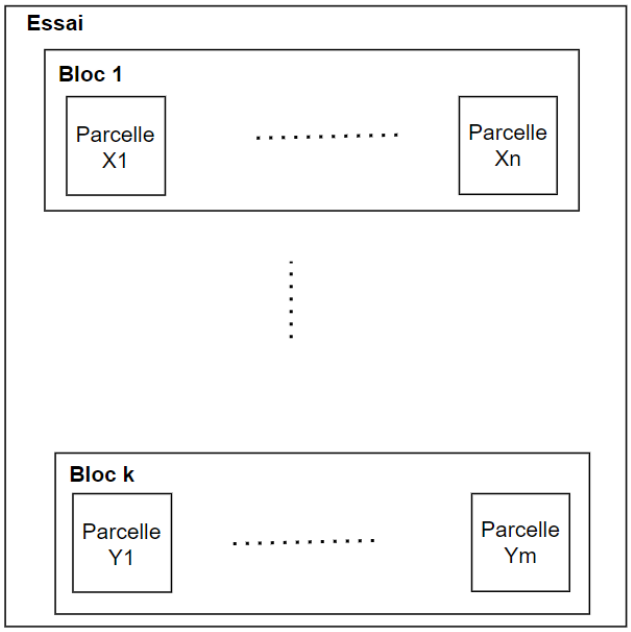
Les valeurs TF-IDF des trois documents sont donc :

Document	She	Loves	Food	With	Cheese	Her	Favourite	is	Italian	Lives	in	State
Doc 1	0.12	0.316	0.12	0.316	0.316	0	0	0	0	0	0	0
Doc 2	0	0	0.12	0	0	0.316	0.316	0.316	0.12	0	0	0
Doc 3	0.12	0	0	0	0	0	0	0	0.12	0.316	0.316	0.316

FIGURE 15 – Représentation vectorielle des 3 documents dans l'espace vectoriel de ce corpus pris en exemple.

B Autres annexes

B.1 Plan du dispositif expérimental au niveau de l'essai tout entien



B.2 Variables sources contextualisés pour enrichir le vocabulaire global.

Variable avec une précision inférieure à 10	Description
Rec adv plein %	full weed coverage
ABV FM t.ha-1	Aboveground fresh mass
HUM perc %	Aboveground humidity
tsha %	Sucrose content in harvested cane
weight tot kg	Total weight of cane harvested on the elementary plot
htvd cm	Height of the last visible ligule of a stem
Variable avec une précision d'au moins 10	Description
Yield CAS	Cane yield (in fresh machinable stem)
IFTH	Herbicide Application Frequency Index

Table 4 – Variables sources contextualisés pour enrichir le vocabulaire global.

B.3 Listes des termes clés ciblés dans les descriptions des variables sources contextualisés pour enrichir le vocabulaire global

— 'weed','coverage',

— 'aboveground','fresh','mass','crop','percentage',

— 'weight','cane','harvested','plot', 'yield',

— 'machinable','stem','herbicide','application','frequency','index',

— 'cane yield','cover crop','cover crops',

— 'elementary plot','plant','Height'.

Liste des tableaux

1 Récapitulatif des essais 4

2 Trois exemples de variables sources. 7

3 Trois exemples de variables candidates. 8

4 Variables sources contextualisés pour enrichir le vocabulaire global. 30