Méthodologie d'étude de la rotation de cultures à partir des données

AMBRE DUPUIS[1,2], CAMELIA DADOUCHI[1,2], BRUNO AGARD[1,2]

[1] LABORATOIRE EN INTELLIGENCE DES DONNEES
DEPARTEMENT DE MATHEMATIQUES ET GENIE INDUSTRIEL,
ÉCOLE POLYTECHNIQUE DE MONTREAL, CP 6079, SUCCURSALE CENTRE-VILLE, MONTREAL, QUEBEC, CANADA
AMBRE.DUPUIS@POLYMTL.CA, CAMELIA.DADOUCHI@POLYMTL.CA, BRUNO.AGARD@POLYMTL.CA

[2] CENTRE INTERUNIVERSITAIRE DE RECHERCHE SUR LES RESEAUX D'ENTREPRISE, LA LOGISTIQUE ET LE TRANSPORT (CIRRELT)

Résumé -
L'agriculture et la science des données fusionnent peu à peu pour promouvoir une agriculture durable : productive et conservatrice de l'environnement. La rotation de cultures est une pratique agricole aux multiplies avantages économiques et environnementaux. Elle est fréquemment étudiée d'un point de vue agronomique cependant, les schémas de rotation de cultures sont encore peu analysés alors qu'ils représentent une composante importante de nombreux modèles agricoles bioéconomiques utiles, entre autres, à l'amélioration de politiques agricoles et au marketing technologique. Le présent article propose une méthodologie d'étude des rotations de cultures à partir des données historiques d'exploitation. Utilisant les principes du Process Mining et les Directly-Follows Graphs (DFG), la méthodologie vise à proposer une représentation et une analyse des successions de cultures, permettant aux décideurs de mieux appréhender les pratiques culturales les plus fréquentes. Les historiques de culture de 394 champs québécois ont été analysés de manière générale puis de façon plus précise en fonction du type d'exploitation et de la zone géographique dont ils sont issus. Notre méthodologie a permis de modéliser les successions directes de cultures les plus fréquentes comme la monoculture de maïs, l'alternance maïs-soja ou encore l'alternance laitue-céleri pour les exploitations maraîchères.

Abstract -
Agriculture and data science are gradually merging to promote sustainable agriculture: productive and environmentally conservative. Crop rotation is an agricultural practice with multiple environmental and economic benefits. It is frequently studied from an agronomic point of view however, crop rotation patterns are still poorly analyzed even though they represent an important component of many bio-economic agricultural models useful, among other things, for improving agricultural policies and technology marketing. This paper proposes a methodology to study crop rotations based on historical farm data. Using the principles of Process Mining and Directly-Follows Graphs (DFG), the methodology aims to propose a representation and an analysis of crops successions, allowing decision-makersto better understand the most frequent cropping practices. The culture history database of 394 Quebec fields was analyzed in a general way and then more precisely according to the type of exploitation and the geographical area from which they originate. Our methodology allowed to model the most frequent direct crop successions such as corn monoculture, corn-soya bean alternation or lettuce-celery alternation for vegetable farms.
Mots clés - Agriculture, process mining, rotation de cultures, DFG
Keywords - Agriculture, process mining, Crop rotation, DFG.

1 INTRODUCTION

Comment nourrir près de 9.8 billions de personnes d'ici 2050 [United Nations, 2017]. C'est l'un des défis les plus importants que rencontre la société du XXIe siècle [Pretty et al., 2010]. Une augmentation de 70% de la production mondiale de nourriture entre 2005 et 2050 semble nécessaire pour assurer une sécurité alimentaire mondiale [Kopittke et al., 2019]. Cependant, les pratiques agricoles actuelles ont un impact néfaste sur l'environnement. L'épandage de fertilisants organiques ainsi que l'apport de fertilisants minéraux et de pesticides dans les champs sont en partie responsables de la diminution de la biodiversité dans les écosystèmes par la pollution des eaux, des sols et de l'air. Le développement de nouvelles stratégies agricoles semble donc nécessaire [Ma, 2019]. Parmi elles, la rotation de cultures est un incontournable de l'agriculture durable [Cui et al., 2019], notamment pour une meilleure utilisation de l'épandage.

Dans le présent article, nous proposons une méthodologie qui vise à identifier les rotations de cultures les plus fréquemment utilisées à partir de l'analyse de données historiques. Cette modélisation pourra, dans un second temps, intégrer des modèles agricoles bioéconomiques plus complexes utiles, entre autres, à l'amélioration des politiques agricoles ou de marketing technologique. Cette méthodologie s'appuie sur les techniques de Process Mining et les historiques de cultures d'exploitations.

L'article se décompose comme suit : la section 2 présente l'état de l'art, dans lequel nous adresserons les fondamentaux agronomiques relatifs à la rotation de cultures (2.1), les recherches actuelles relatives à la modélisation des rotations de cultures (2.2), ainsi que les concepts préliminaires relatifs au Process Mining (2.3). Cette section sera conclue par une synthèse des grands principes proposés dans l'état de l'art (2.4). La section 3 présentera la méthodologie proposée afin de représenter les

rotations de cultures les plus fréquentes. La section 4 exposera un cas d'étude qui a servi à la validation de la méthode. Dans un premier temps, nous présenterons le contexte de l'étude de cas (4.1), dans un second temps l'analyse selon le type d'exploitation sera effectuée (4.2) et, dans un troisième temps, une analyse par zone géographique est réalisée (4.3). Finalement, la section 5 rappelle les résultats et limites de la méthode proposée et présente quelques perspectives.

2 ÉTAT DE L'ART

L'agriculture et la science des données fusionnent pour proposer une agriculture avec le minimum d'impact sur l'environnement tout en maximisant sa productivité [Liakos et al., 2018]. La pratique de la rotation de culture a elle aussi été fréquemment étudiée.

2.1 Rotation de cultures : fondamentaux agronomiques

La rotation de cultures est une pratique agricole consistant à répéter systématiquement, ou du moins fréquemment, une même séquence de culture durant plusieurs saisons de croissances [Reeves, 1994]. Le type de cultures utilisé dépend à la fois des facteurs écologiques (caractéristique physique et chimique des sols, interaction du sol avec l'écosystème, le climat, etc.) et des facteurs économiques [Porter, 2009]. Les avantages de cette pratique sont principalement la lutte contre l'érosion des sols, la conservation de la matière organique, la gestion des nutriments ainsi que le contrôle des maladies [USDA, s.d.]. Ces avantages peuvent être traduits en bénéfices pour les agriculteurs ; la rotation de cultures permet de « diminuer les coûts associés à la fertilisation des champs, réduire les besoins en pesticides et herbicides, améliorer le rendement des champs en valorisant la qualité des sols, prévenir l'érosion et conserver l'humidité des sols, protéger la qualité des eaux et aider à protéger la santé humaine. » [Clark, 2007]. L'agriculture conservatrice favorise ainsi la pratique de la rotation de cultures pour une agriculture durable et productive [Cui et al., 2019].

2.2 Modéliser les rotations de culture

La modélisation des processus permet de les décrire afin de détecter les situations problématiques. La reconception pour l'amélioration des processus modélisés devient alors possible [Biazzo, 2002]. La rotation de culture peut être vue comme un processus à améliorer. Cependant, alors qu'un grand nombre d'études tentent, par différents moyens, de prédire les rotations de cultures [Osman et al., 2015][Klein Haneveld et Stegeman, 2005][Detlefsen et Jensen, 2007][Salmon-Monviolaet al., 2012][Le Ber et al., 2006][Zhang et al., 2019][Yaramasu et al., 2020][Kussul et al., 2017], très peu, à notre connaissance, cherchent à les modéliser à partir de données réelles. Cette observation est d'ailleurs confirmée par [Levavasseur et al., 2015] « peu de modèles proposent de reconstruire des séquences de cultures réelles et de simplifier leur diversité dans un nombre raisonnable de rotations de cultures. ». Pourtant, dès 1994, l'utilité de la modélisation des exploitations agricoles est reconnue par les scientifiques [Edward-Jones & McGregor,1994] à des fins de marketing technologique ou d'évaluation de politique *ex ante*.

La connaissance des rotations de cultures et des interrelations entre les cultures est un des éléments les plus souvent utilisés dans les modèles agricoles bioéconomiques mécanistes. 27 des 48 études revues par [Janssen & van Ittersum, 2007] utilisent des contraintes reliées aux rotations de cultures. Pour répondre à cette problématique de modélisation, [Dogliotti et al., 2003] et [Bachinger et Zander, 2007] proposent des outils basé sur un ensemble de règles agronomiques alors que [Castellazzi et al., 2008] proposent une représentation mathématique des rotations de culture, soit des savoirs théoriques. [Levavasseur et al., 2015] proposent une caractérisation des rotations de cultures à l'aide de données réelles obtenues via des entretiens auprès d'exploitants locaux sur leurs habitudes culturales ainsi qu'une modélisation dérivée des travaux de [Schönhart et al., 2011]. Ces derniers travaux ont permis de développer l'outil CropRota utilisant à la fois les règles agronomiques et des données réelles relatives aux historiques de cultures afin de faire ressortir les séquences de cultures les plus fréquentes [Schönhart et al.,2011].

2.3 Concepts préliminaires relatifs au Process Mining

Le Process Mining est à la croisée des chemins entre la fouille de données et la modélisation des processus d'affaires [van der Aalst, 2011]. L'objectif général est de trouver, à partir des données, des schémas de comportement récurrent [Cook et Wolf, 1995][van der Aalst, 2011]. Le Process Mining permet alors de représenter les principaux flux d'activités au sein d'un processus étudié [Nieves et al., 2020]. Le Process Mining se base sur trois composantes essentielles qui doivent impérativement se retrouver dans les données traitées soit :

- Une activité : événement ou étape bien définis dans le processus.
- Un cas : réalisation unique du processus.
- Un marqueur temporel : indication temporelle de la réalisation de l'activité dans un cas précis.

On appelle *Trace* la succession des activités dans un cas [van der Aalst, 2011].

Ce genre de données peut être utilisé dans 3 types de Process Mining [van der Aalst, 2016] :

- La découverte : permets de générer un modèle à l'aide des données sans informations a priori
- La conformité : permets la comparaison entre un modèle théorique et la réalité contenue dans les données
- L'amélioration : permets de modifier le modèle théorique en fonction de la réalité contenu dans les données

Le résultat obtenu est une représentation graphique (PetriNet, EPC…) de toutes les options possibles permettant de réaliser le processus en question. Il existe, entre autres, 3 grands algorithmes de process mining de type découverte. L'algorithme α, le plus simple, mais limité par ses capacités de détection de boucles courtes (moins de trois activités), d'activités en parallèle et est sensible au bruit. L'*heuristic-miner* prend en compte la fréquence des traces, ce qui permet de limiter l'impact du bruit et la détection des boucles courtes. Finalement, l'algorithme *inductive-minner*, le plus utilisé, permet de détecter les comportements peu fréquents et assure la cohérence du modèle [The Fraunhofer Institute for Applied Information Technology, 2020]. Cela dit, l'ensemble de ces algorithmes est utile lorsque les activités de début et de fin du processus sont identifiables.

Un outil permettant de représenter les activités et leurs relations directes, sans prendre en compte leurs places au sein du processus, est le *Directly Follow Graph* ou *DFG*. Ce graphe a pour nœud les différentes activités présentes dans la base de données et assigne un arc entre deux activités successives. Ainsi le DFG est un graphe orienté dont le poids, par exemple $w_{a,b}$, des arcs représente le nombre de fois dans lequel la succession des activités A puis B a

été enregistrée [Leemans et al., 2018]. Dans le cadre de la modélisation des relations directes entre les cultures, l'utilisation du DFG semble la plus appropriée.

2.4 Synthèse

L'état de l'art montre l'importance de la rotation de cultures pour préserver la qualité des sols et assurer la réduction de l'impact environnemental de l'agriculture sur le sol. De nombreuses méthodes, basées sur le savoir théorique et sur l'acquisition de savoir à partir des données, ont été élaborées afin de prédire les rotations de cultures utilisées par les agriculteurs. Cependant, très peu permettent de modéliser simplement les relations entre cultures, pourtant utiles à l'élaboration d'un modèle agricultural bioéconomique plus complexe permettant une amélioration des politiques agricoles et du marketing technologique. À notre connaissance, aucune méthode n'utilise le Process Mining pour identifier les modèles de rotations de cultures. Pourtant, le Process Mining a pour objectif de trouver des schémas de comportements récurrents à partir de données événementielles. En assimilant la rotation de cultures à un processus, le type de culture planté pourrait être associé à un événement dont la date d'exploitation correspondrait au marqueur temporel nécessaire à l'utilisation des méthodes de Process Mining. Cette technique pourrait permettre de mieux appréhender les habitudes culturales menant à l'exploitation d'un type de culture.

Pour ces raisons, nous proposons une méthodologie d'étude de la rotation de cultures. À partir de l'analyse des données historiques d'exploitation agricole, les successions de cultures sont mises en évidence, puis une analyse des différents graphes DFG permet de mieux comprendre les pratiques de gestion des agriculteurs en matière de rotation de cultures.

3 METHODOLOGIE D'ETUDE DES ROTATIONS DE CULTURES

Cette méthodologie peut être décomposée en 5 étapes détaillées ci-après.

Afin d'illustrer la méthodologie, nous considérons un ensemble de données qui décrit la réalisation successive de différents événements A, B, C, D et E. Par exemple :

Cas 1 $A \rightarrow A \rightarrow A \rightarrow A \rightarrow B \rightarrow B \rightarrow A \rightarrow D \rightarrow B$
Cas 2 $A \rightarrow A \rightarrow A \rightarrow A$
Cas 3 $B \rightarrow B \rightarrow D \rightarrow A \rightarrow B \rightarrow B \rightarrow A \rightarrow A$
…

Le nombre d'événements peut varier d'un cas à l'autre, il peut y avoir des répétitions d'événements, tous les événements ne sont pas obligatoirement présents dans chaque cas.

3.1 Exploration statistique des données

Il est essentiel de comprendre la nature et la distribution des données pour pouvoir, par la suite, les préparer adéquatement. Cette étape permet également « de détecter des sous-ensembles intéressants pour formuler des hypothèses sur les informations cachées » [Pretorius et Matthee , 2006]. Pour connaître la nature des données, une visualisation des attributs disponibles et de leurs types peut être effectuée. Aussi, une visualisation des premiers enregistrements permet d'affiner la compréhension des informations présentes dans la base de données. Ainsi, les attributs relatifs aux activités, aux cas et aux marqueurs temporels (section 2.3) peuvent être identifiés. Par la suite, l'étude de la distribution des différentes activités dans la base de données peut être effectuée à l'aide d'un décompte du nombre d'occurrences.

3.2 Préparation des données

L'exploration statistique des données permet d'orienter la préparation des données. En effet, si le nombre d'activités uniques est trop important, elles peuvent être groupées en catégories plus générales. L'ensemble des données doit être contenu dans un tableau possédant au minimum 3 colonnes contenant les informations présentées dans la section 2.3. Ainsi, une première colonne contient l'identification unique du cas, une seconde colonne doit contenir l'information sur l'activité, une troisième colonne doit contenir la date d'occurrence de la combinaison des deux colonnes précédentes. L'exploration des données (section 3.1) a normalement permis d'identifier les attributs contenant les informations d'intérêt, mais il est possible que certaines transformations soient nécessaires afin d'atteindre le formatage exigé. La création d'un champ permettant l'identification unique des cas par l'association de deux attributs existants dans la base de données peut être nécessaire. Aussi, les données utilisées pour le marqueur temporel peuvent devoir subir une conversion au format « Date ». Finalement, si comme mentionné dans la section 3.1, des sous-ensembles d'intérêts sont identifiés, ils doivent être isolés par une sélection sur l'attribut d'intérêt.

Ainsi le tableau 1 représente le 1er cas de l'exemple, après l'étape de préparation des données.

Tableau 1: Exemple de représentation des données

case:concept:name	concept:name	time:timestamp
1	A	1
1	A	2
1	A	3
1	A	4
1	B	5
1	B	6
1	A	7
1	D	8
1	B	9

L'ensemble des cas sont transformés de la sorte et mis bout à bout.

3.3 Création des graphes (DFG)

Le graphe nommé le « Directly-follows graph » ou DFG (section 2.3) est utilisé pour sa facilité de représentation des relations directes existant entre les activités et les possibilités de transformation postérieure qu'il propose. La réalisation du graphe DFG sur les données brutes est proposée à la figure 1.

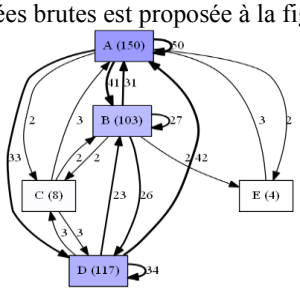

Figure 1: Graphe Direcly-follows (DFG)brute

Ce premier graphe permet de représenter l'ensemble des activités ainsi que leurs interactions directes. On note que les activités A, B et D sont les plus fortement représentées dans la base de données étudiée avec respectivement 150, 103 et 117 alors que les événements C et E sont très marginaux.

Ce graphe permet de visualiser les différentes options possibles à la suite d'une activité. Avec les informations présentes sur la figure, les probabilités conditionnelles de successions de culture peuvent être déduites. Ici, si l'activité A survient, alors l'activité suivante sera l'activité A avec une probabilité de 39%, l'activité D avec une probabilité de 32%, l'activité B avec une probabilité de 26% ou les activités C ou E avec une probabilité de 1.5%.

3.4 *Simplification du modèle*

En général, la création du premier graphe à partir des données traitées est peu lisible, car l'ensemble des transitions sont représentées, et ce sans réelle distinction. Beaucoup d'informations s'enchevêtrent, aboutissant à un graphe surchargé et difficilement compréhensible. Pour pallier ce manque de visibilité, deux traitements sont proposés :

3.4.1 Regroupement des activités

L'illisibilité du graphe issu des données brutes nous pousse à traiter les données afin de limiter la complexité de leurs représentations. Le regroupement de culture consiste à sélectionner l'ensemble des cultures dont le nombre d'occurrences total dans la base de données étudiée est inférieur à un seuil et de les regrouper sous une appellation commune (*e.g.* « Autre »).
Les événements dont le taux d'occurrence est inférieur à un seuil (ici fixé à 10%) sont regroupés. Ici les événements C et E représentent respectivement 2% et 1% des événements enregistrés. Ils sont donc regroupés dans une nouvelle catégorie d'événements appelée « Autre ». Cette nouvelle catégorie représente 3% des événements (12 enregistrements). Un nouveau graphe est réalisé avec ce regroupement (fig 2.).

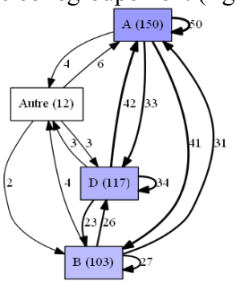

Figure 2 : DFG regroupé (seuil 10%)

Ce regroupement permet d'alléger le modèle en supprimant une case d'activité sans pour autant changer les différentes proportions des transitions entre les événements principaux.
On retrouve les mêmes probabilités de succession à l'activité A que précédemment, à l'exception des probabilités P(A→E) et P(A→C) qui sont transformés en P(A→Autre) =3%.

3.4.2 Filtrage des arcs

Toujours dans l'optique d'améliorer la lisibilité des graphes DFG pour identifier les rotations de cultures les plus fréquentes, nous proposons d'effectuer une opération de filtrage. Le filtrage des arcs consiste à éliminer l'ensemble des arcs dont le taux de transition est inférieur ou égal à un seuil fixé.
Le taux de transition associé à l'arc i (Tt_i) est défini comme suit :

$$Tt_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

Où : *n = nombre d'arcs dans le graphe*
w_i =nombre de transitions associées à l'arc $\forall i \in n$

Puisqu'une transition relie toujours deux activités, on peut noter :

$$\sum_{i=1}^{n} w_i = nb\ activités - nb\ entités$$

D'où

$$Tt_i = \frac{w_i}{nb\ activités - nb\ entités}$$

Le nombre total d'activités est ici de 382 et le nombre total de cas enregistrés est de 53. Ainsi, si l'on fixe le seuil limite à 1%, seuls les arcs comptant plus de 4 transitions sont représentés (fig 3). Cette opération est appelée « filtrage ».

L'opération de filtrage permet de grandement simplifier le modèle puisque seules les activités les plus fréquentes sont représentées. Cependant, cette opération entraîne une légère modification dans les probabilités de successions puisqu'après filtrage :
P(A→A) = 40% ; P(A→B) = 33% et P(A→D) = 27%.
Cela étant dit, l'ordre est conservé.
Les opérations de regroupement et de filtrage peuvent être combinées comme le présente la figure 4. (section 3.4):

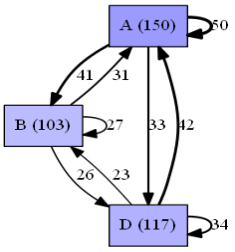

Figure 3 : DFG filtré (seuil 1%)

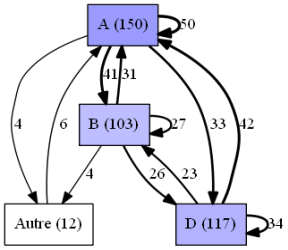

Figure 4 : DFG regroupé et filtré (seuil 10% et 1%)

La combinaison des opérations de regroupement et de filtrage permet de conserver les avantages de l'opération de filtration tout en minimisant son désavantage premier. En effet, ce double traitement permet d'une part de ne conserver que les transitions les plus fréquentes (éliminant de ce fait les activités fortement minoritaires), rendant ainsi le graphe beaucoup plus lisible. Et d'autre part, de limiter l'impact de la filtration sur les probabilités de succession. Dans le cas du regroupement puis du filtrage, les probabilités de succession à l'activité A sont : P(A→A) = 39% ; P(A→B) = 32% ; P(A→D) = 26% et P(A→Autre) = 3%.

3.4.3 Choix des seuils

Les opérations de regroupement et de filtrage permettent de donner plus de lisibilité au graphe DFG en simplifiant sa structure tout en conservant le même ordre dans les probabilités de succession des activités. Ces opérations de traitement sont effectuées en s'appuyant sur des seuils fixés arbitrairement. Une analyse des effets d'une variation sur ces seuils permet de dégager les résultats suivants : un seuil de regroupement tenant compte de la contribution des activités à la base de données permet de produire un graphe identique ou très similaire au graphe brut. *A contrario*, un seuil de regroupement trop élevé engendre une perte d'information trop importante pour une bonne analyse du graphe DFG. Nous proposons de choisir le seuil de regroupement de manière visuelle à l'aide du graphique représentant la distribution des activités dans la base de données. Le seuil devrait être choisi au niveau d'une « marche » afin de séparer les activités principales des activités fortement minoritaires.

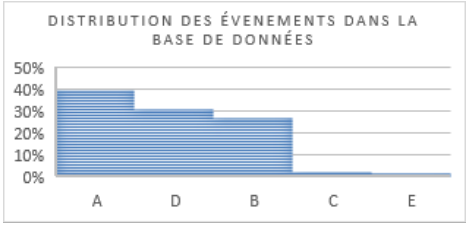

Figure 5 : Distribution des événements

Ainsi, d'après la figure 5, le seuil de regroupement devrait appartenir à l'intervalle [5% ; 25%] pour garder un niveau d'information suffisant tout en simplifiant le modèle. Tout comme

pour le seuil de regroupement, le seuil de filtrage peut être déterminé par l'analyse du graphique de distribution des transitions (fig 6). Le seuil de filtrage devrait appartenir à l'intervalle [1% ; 6%] pour garder un niveau d'information suffisant tout en simplifiant le modèle.

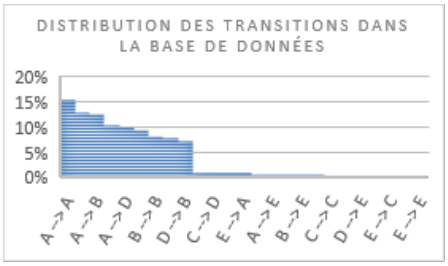

Figure 6 : Distribution des transitions

3.5 Analyse des résultats

L'application de la méthodologie permet de dégager les trois activités les plus fréquentes (A, B et D) ainsi que leurs relations directes. Suite à la réalisation des activités A, B ou D, on retrouve généralement l'activité A. On note la forte tendance à la répétition des événements fréquents. Les évènements C et E sont marginaux et sont regroupés dans la catégorie « Autre ». Ce regroupement montre que, généralement, la succession à une activité « Autre » est l'activité A. Comme précisée dans les sections 3.1 et 3.2, la connaissance du domaine d'application peut être utilisée pour filtrer selon certains critères et obtenir des graphes ciblés. En effet, certains attributs ne faisant pas partie des informations essentielles au Process Mining (section 2.3) peuvent cependant avoir un grand intérêt pour l'analyse puisqu'ils permettent d'ajouter de l'information sur le contexte dans lequel les données évoluent. Ceci peut avoir un impact important sur la succession des activités au sein d'un processus. Le filtrage par attribut dans la phase de préparation des données permet d'ajouter de l'information à notre analyse et de proposer des graphes ciblant des comportements plus spécifiques. L'analyse des graphes permet d'une part de repérer les activités les plus fréquentes au sein d'un processus en limitant la représentation d'activités fortement minoritaires pouvant être perçues comme du bruit et, d'autre part, en analysant des arcs reliant les activités de quantifier les relations directes qu'elles entretiennent entre elles. L'analyse de ces graphes permet d'avoir une représentation lisible des successions directes d'activité au sein d'un processus et de les quantifier.

4 CAS D'ETUDE

En 2018, la province du Québec au Canada comptait 28 919 exploitations agricoles et 1 867 000 hectares cultivés [Institut de la statistique Québec, 2020]. Les zones agricoles sont concentrées sur les berges du Saint-Laurent au sud de la province [Gouvernement Quebec, 2019]. Les données dont nous disposons décrivent les pratiques agricoles de 81 exploitations québécoises réparties sur ces territoires agricoles. Cinq types d'exploitations peuvent être définis soit les élevages de ruminants (bovins, ovins), les élevages de monogastrique (porcs, volailles), les exploitations de grandes cultures, les exploitations maraîchères ainsi que les exploitations de gazon. Ces données retracent les différents types de cultures exploitées dans 394 champs entre 2000 et 2020 au Québec. Les données des 20 dernières années ne sont pas disponibles pour l'ensemble des champs et les données concernant certains champs pour certaines années d'exploitation sont manquantes. 81 types de cultures (exemple : foin de graminées, foin de graminée 100%, foin de graminées 60%, laitue pommée terre noire, laitue frisée transplantée en terre noire …) sont répertoriées.

4.1 Exploration statistique des données

Comme présenté précédemment, 81 types de cultures sont répertoriées cela dit, une grande partie de ces cultures font référence à une catégorie d'espèces particulières. L'exploration des données permet également d'identifier deux attributs pouvant servir de filtre pour définir des sous-ensembles d'analyse. Ces attributs sont les types d'exploitation (voir section 4.3.1) et la situation géographique des champs (voir section 4.3.2). Nous identifions les cultures comme les activités possibles tandis que les champs représentent les différents cas étudiés dans le processus de rotation de cultures. L'année d'exploitation d'une catégorie de culture dans un champ représente le marqueur temporel nécessaire à l'utilisation des techniques de Process Mining (voir section 2.3). L'analyse des données permet d'identifier la première opération de traitement des données à réaliser, soit la catégorisation des cultures selon leurs espèces. En effet, sur 81 types de cultures identifiées, 66 représentent moins de 1% des enregistrements de la base de données. Finalement, certains champs comptent plusieurs enregistrements pour une même année, mais le type de culture exploitée reste le même pour ces enregistrements.

4.2 Préparation des données

Dans un premier temps, l'analyse se porte sur l'ensemble des champs référencés dans la base de données. La première étape de préparation consiste à regrouper les cultures selon leur « espèce ». Ainsi, les cultures de « foin graminées », « foins graminées 100% », « foin graminées 60% » … sont regroupées sous la même appellation « Foin ». Il en est de même pour les autres cultures. Cette opération permet de réduire le nombre d'activités considérées, passant de 81 types de cultures à 36 cultures. Dans un second temps, une clé d'identification unique des champs est créée par une association du numéro de l'exploitation et du numéro du champ. Cette clé d'identification unique à chaque champ est celle utilisée pour l'identification des cas dans l'algorithme de Process Mining. De plus, les enregistrements dupliqués (plusieurs enregistrements d'une même culture dans un même champ pour une même année) sont regroupés sous un même enregistrement soit 6 552 enregistrements. Finalement, seuls les attributs d'intérêt pour l'étude (clé d'identification, type de culture et date d'exploitation) sont sélectionnés et renommés respectivement « case:concept:name », « concept:name », « time:timestamp » pour la réalisation des étapes suivantes.

4.3 Création des graphes

Le premier graphe DFG, crée via la librairie PM4PY disponible sous Python, reprenant l'ensemble des enregistrements, est présenté à la figure 7. Le grand nombre d'activités et d'interrelations le rend relativement peu compréhensible. L'enchevêtrement des arcs demande un certain effort pour pouvoir comprendre les dynamiques entre les différents types de cultures. Cette complexité est appréhendée par les différentes étapes de traitement proposées ci-après.

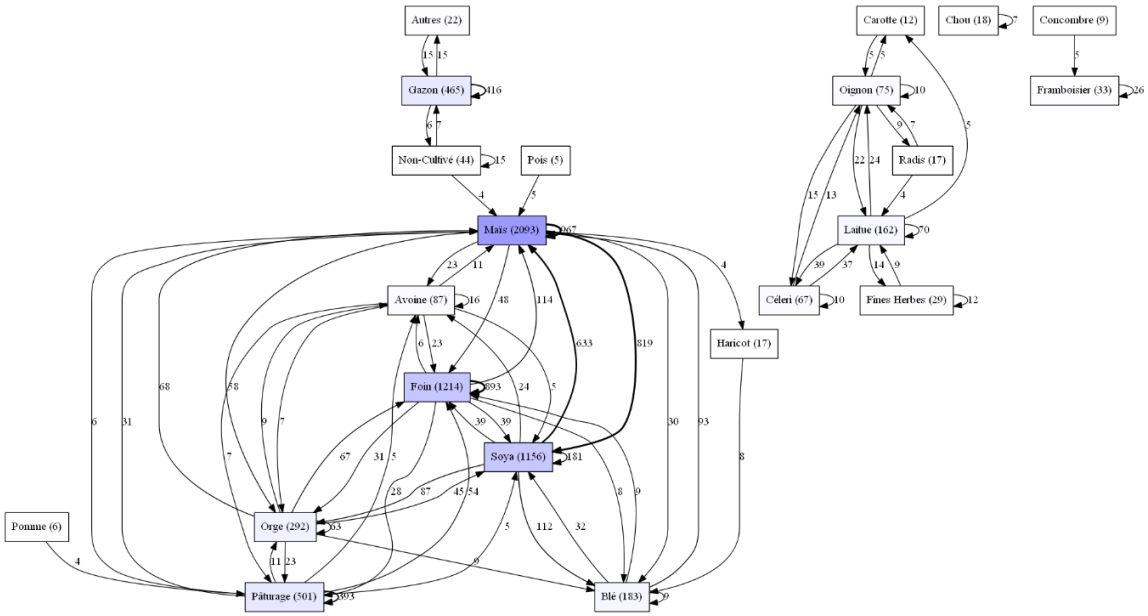

Figure 7 : DFG brut pour les champs étudiés

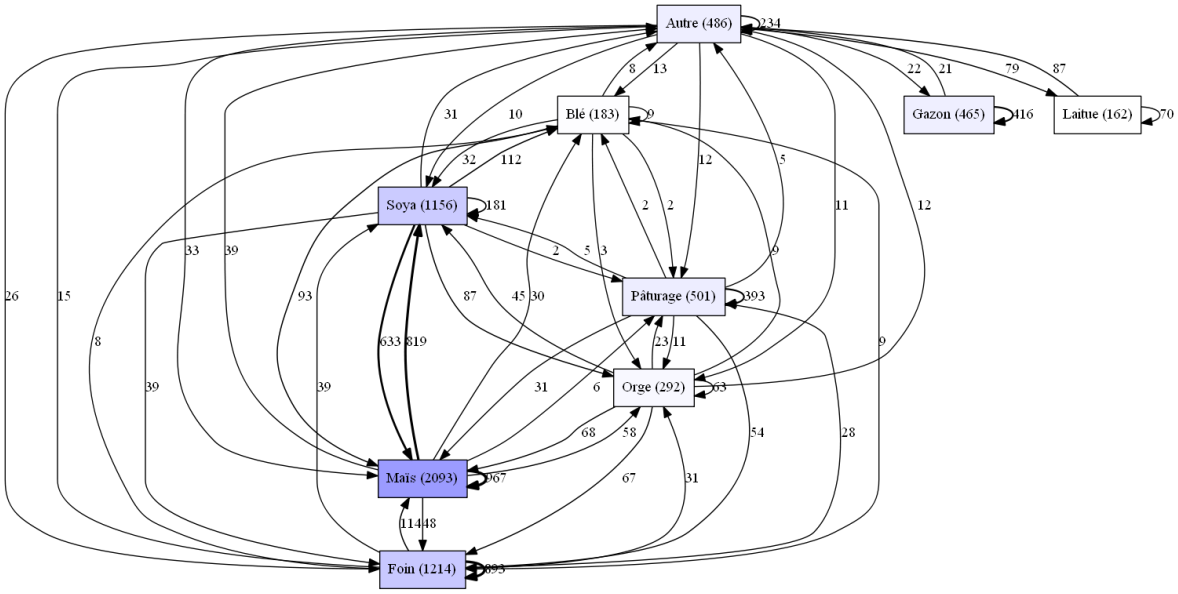

Figure 8 : DFG regroupé (seuil 2%) pour les champs étudiés

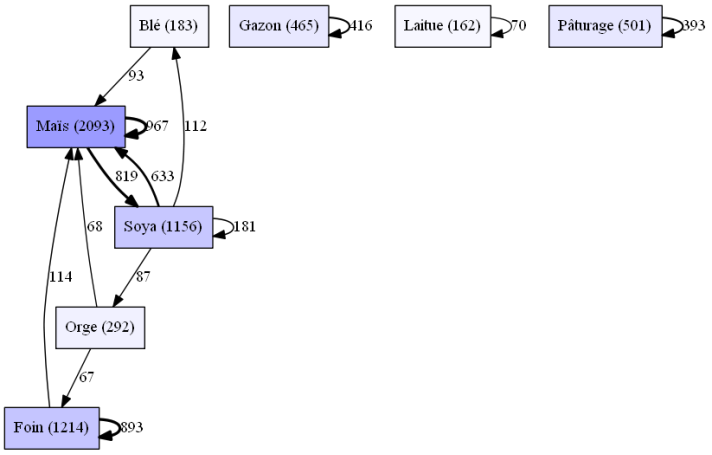

Figure 9 : DFG filtré (seuil 1%) pour les champs étudiés

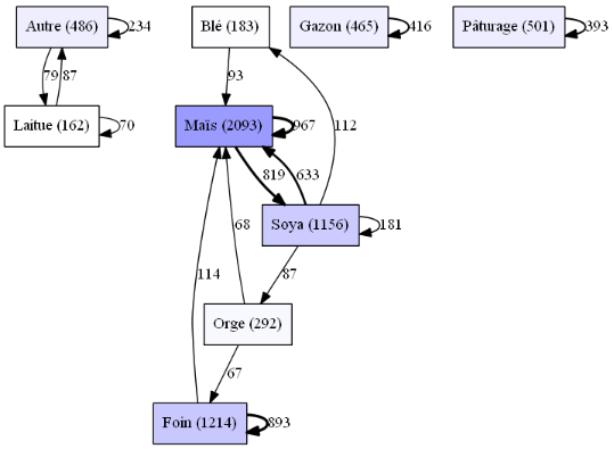

Figure 10 : DFG groupé et filtré (seuils 2% et 1%) pour les champs étudiés

4.3.1 Regroupement des cultures

Afin de regrouper les cultures, le seuil de regroupement a été choisi à la suite de l'analyse du graphique des distributions des activités. Un seuil de 2% a été choisi, soit le regroupement de toutes les cultures comptabilisant moins de 132 occurrences dans la base de données. Le graphe résultant est proposé figure 8.

4.3.2 Filtrage des arcs

Afin de filtrer les arcs, le graphique des distributions des arcs a été réalisé puis analysé. Le choix d'un seuil de 1% a été fait de sorte que seuls les arcs dont le nombre d'occurrences est supérieur à 65 sont visualisés dans la figure 9. Le filtrage permet de simplifier grandement le modèle en ne modélisant que les successions fréquentes. La forte tendance à la monoculture du maïs, du foin ainsi que l'alternance du maïs-soja sont très facilement visibles. On note la diminution du nombre d'activités, particulièrement les cultures de légumes. L'opération de groupement puis de filtrage permet de mettre en valeur un besoin d'investigation des pratiques maraîchères (Fig10).

4.4 Analyse des résultats

On note la forte tendance à l'alternance soja-maïs, ainsi que la forte tendance à la monoculture des cultures de maïs, foin, gazon et pâturage. On peut également noter la faible tendance à la monoculture du soja et du blé. On semble pouvoir observer deux boucles de trois éléments soit maïs-soja-blé et maïs-soja-orge. Cela dit, le graphe DFG ne propose qu'une vision des flux directe donc il est impossible de savoir combien de ces boucles ont réellement été observées. Il semble se dégager un comportement particulier relatif à l'exploitation maraîchère avec une rotation laitue →« Autre » relativement importante. Dans ce cas précis, l'apport des informations relatives au type d'exploitation peut être utile puisque la même méthodologie peut être utilisée sur un ensemble de données ciblé afin de dégager un comportement particulier. C'est pourquoi nous avons réalisé cette analyse par secteur agricole qui est présentée ci-après.

4.4.1 Analyse par secteur agricole

Comme mentionné dans la section 4, il existe 5 types d'exploitations distinctes. L'analyse des rotations de culture générale a dégagé le besoin de cibler les comportements relatifs à l'exploitation de légumes. C'est pourquoi nous utilisons l'information relative au type d'exploitation pour approfondir nos analyses. La base de données précédemment décrite est filtrée pour ne garder que les 30 champs issus d'exploitations de type « Maraîcher ». Seuls les 3 attributs d'intérêts sont conservés. Les colonnes regroupant respectivement les identifiants des champs, l'année d'exploitation et la culture exploitée sont renommées par « case:concept:name », « time:timestamp et « concept:name ».

Le graphe « brute » est réalisé de manière à visualiser l'ensemble des transitions de culture ayant existé entre 2000 et 2020 dans les 30 champs étudiés. Dans ces 30 champs, 23 types de cultures sont représentés dans les 464 occurrences de cultures, d'où un nombre total de 434 transactions.

Afin de faire ressortir les comportements les plus fréquents, un regroupement des cultures représentant moins de 1% des enregistrements est effectué. Ainsi, les cultures présentes moins de 4 fois dans l'ensemble des données sont regroupées dans la catégorie « Autre ». Dans un second temps, les transitions représentant moins de 1% des transitions totales, soit inférieures à 5 occurrences, sont filtrées (Fig 11).

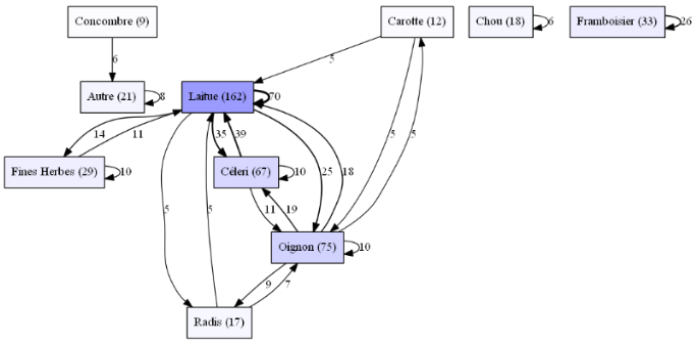

Figure 11 : DFG groupé et filtré (seuils 1%) dans « Exploitations maraîchères »

L'ensemble de ces traitements permet de mettre en évidence, par exemple, la tendance à la monoculture de framboise ce qui parait cohérent à la vue du caractère pluriannuel de cette culture [CRAAQ, 2003]. L'alternance fréquente des cultures de laitue et de céleri peut également être remarquée (23% des transitions issues de la culture de laitue sont à destination de la culture de céleri et 65% des transitions issues de la culture de céleri sont à destination de la culture de laitue). Finalement, les recommandations proposées par [Grignon, 2015] peuvent, en partie, se retrouver dans la figure 11 avec une forte relation de la culture d'oignon à la suite d'une culture de laitue ou de céleri, la relation existante entre la culture de carotte et d'oignon, ou encore l'absence de monoculture de carotte. Cependant on note une divergence entre les pratiques réelles observées à la figure 11 et les recommandations agronomiques puisque la monoculture d'oignon peut être observée dans 16.39% des cas ce qui est déconseillé d'un point de vue agronomique (recommandation d'un minimum de 3ans entre deux cultures d'oignon) [Grignon,2015]. Le même type d'analyse peut être réalisé par rapport à la localisation spatiale des champs.

4.4.2 Analyse par zone géographique

La pédologie des sols, le relief, le climat influencent grandement les pratiques agricoles. Les écodistricts sont des délimitations géographiques du territoire canadien « caractérisées par des assemblages distinctifs de formes du relief, de matériaux de surface, de sols, de plans d'eau, de végétation et d'utilisations du territoire » [AAC, 1995]. C'est le plus faible niveau de généralisation de la classification écologique des terres [Statistique Canada, 2018]. C'est pourquoi il a été choisi comme indicateur spatial associé aux champs.

Les données disponibles sont réparties dans 6 écodistricts distincts dont les limites géolocalisées sont disponibles sur les données ouvertes du Canada [Gouvernement Canada, 2017].

À l'aide du système d'information géospatial QGIS, les différents champs géolocalisés ont été associés à écodistrict auxquels ils appartiennent. L'analyse des rotations de culture en fonction de la zone géographique permet de cibler de manière plus précise les comportements des agriculteurs, mais cette fois sans l'intervention de connaissances extérieures. L'étude des 242 champs présents dans l'écodistrict « Middle St.Lawrence plaine » est entreprise. Comme dans l'analyse par secteur agricole, seuls les 3 attributs d'intérêts sont conservés et renommés. 17 types de cultures sont présents dans ce sous-ensemble de données comptant 4 107enregistrements (occurrences de cultures) et 3 865 transitions. Le graphe « brut » est présenté dans la figure 14.

Le regroupement et le filtrage des données au seuil de 1% est là aussi effectué. Ainsi, les cultures représentant moins de 42

enregistrements sont regroupées est les transitions inférieures à 39 occurrences sont filtrées (Fig 12).

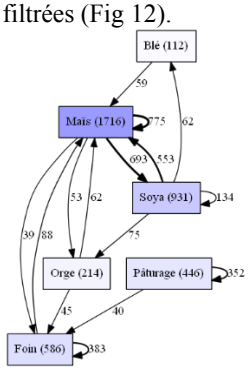

Figure 12 : DFG groupé et filtré (seuils 1%) dans "Middle St.Lawrence Plain"

On note la très grande tendance à la monoculture de maïs (50%) de pâturage (90%) et de foin (81%). La monoculture de foin est en adéquation avec les recommandations agronomiques préconisant une monoculture de foin de 4 à 6 ans [Turcotte, 2018]. La succession soja→maïs est très fortement représentée (67%) tout comme la transition maïs → soja (44%). On note également que le soja a une très faible tendance à la monoculture avec seulement 16% des successions au soja vers le soja. Finalement, deux cycles de trois cultures sont facilement identifiables soit : maïs→soja→blé ainsi que maïs→soja→orge. D'après [Kirkegaard et Ryan, 2014][Kirkegaard et al., 2008] l'insertion d'une culture de légumineuse à grain ou d'oléagineux (ici le soja qui est une oléo-protéagineuse) dans une rotation céréalière (maïs, orge et blé) permet d'augmenter les rendements des cultures céréalières grâce, entre autres, à une amélioration de l'apport en azote et à une diminution des mauvaises herbes et maladies dans les cultures. Les bénéfices de l'apport de la culture de blé dans la rotation maïs→soja ont également été démontré au Canada par [Janovicek et al., 2021]. L'intégration de légumineuse dans une rotation céréalière a également l'avantage de diversifier les revenus des exploitants [Kirkegaard et Ryan, 2014]. Ceci pose la question de l'intérêt de la monoculture de maïs pourtant largement représenté.

5 Conclusion

La méthodologie proposée a permis de représenter les successions de cultures les plus fréquentes, après un découpage par type d'exploitation et par zone géographique. Cette méthodologie s'appuie sur 5 phases : l'exploration statistique des données, la préparation des données, la création de graphes DFG, la simplification du modèle par regroupement et filtrage selon des seuils préalablement définis et, finalement, l'analyse des résultats. Cette méthode permet la création de graphes DFG qui peuvent être utilisés dans un modèle agricole bioéconomique. Les graphes DFG obtenus étant fortement cycliques, une analyse plus approfondie à l'aide d'algorithmes de détection de cycles devra être réalisée. Connaître les habitudes culturales des agriculteurs permet de mieux comprendre et anticiper les schémas de rotation de cultures. Cette connaissance peut aider les agronomes à prévoir l'impact environnemental potentiel des exploitations et orienter leurs services auprès des agriculteurs pour promouvoir une agriculture plus durable et productive. Une limite importante de cette méthodologie est la perte de conformité dans les flux. L'opération de groupement et de filtrage entraîne un éloignement du modèle avec la réalité. Cette question est traitée dans la littérature [Leemans et al., 2019] et devra être étudiée dans de prochains travaux. Cette étude est un premier pas dans la modélisation des rotations de culture à l'aide des techniques de process mining. Ensuite, le but sera de prédire les besoins du sol pour la prochaine culture pour enfin optimiser les opérations d'épandage, en amont. Il existe plusieurs perspectives d'améliorations telle l'utilisation d'ontologie qui pourrait améliorer l'opération de regroupement. Ou encore l'approfondissement de la question du choix des seuils puisqu'ils ont un impact significatif sur les graphes obtenus. L'étape de filtrage devrait faire l'objet d'une réflexion plus poussée puisqu'elle semble nécessaire à la lisibilité du modèle, mais entraîne une perte d'information pouvant être problématique. Finalement des analyses complémentaires telles qu'une comparaison des relations existant entre les cultures pour deux périodes données seraient intéressantes afin de visualiser les modifications des pratiques agricoles en matière de rotation de culture.

6 Remerciements

Nous tenons à remercier notre partenaire industriel pour sa collaboration dans le projet ainsi que IVADO et PROMPT pour leur soutien financier.

7 References

Agriculture and Agroalimentaire Canada. (1995). Cadre écologique national pour le Canada. Ottawa/Hull: Groupe de travail sur la stratification écologique.

Bachinger, J., & Zander, P. (2007). ROTOR, a tool for generating and evaluating crop rotations for organic farming systems. European Journal of Agronomy, 26(2), 130-143.

Biazzo, S. (2002), Process mapping techniques and organisational analysis: Lessons from sociotechnical system theory, Business Process Management Journal, Vol. 8 No. 1, pp. 42-52. https://doi.org/10.1108/14637150210418629

Castellazzi, M., Wood, G., Burgess, P., Morris, J., Conrad, K., & Perry, J. (2008). A systematic representation of crop rotations. Agricultural Systems, 97(1-2), 26-33.

Clark, A. (2007). Benefits of cover crops. Dans Managing cover crops profitably. 3rd ed (p. 9). Beltsville, MD.: Susatinable Agriculture Network.

CRAAQ. (2003). Production de framboises biologiques. pp.4 disponible sur : https://www.craaq.qc.ca/Publications-du-CRAAQ/production-de-framboises-biologiques/p/PABI0004, consulté le 10/02/2021.

Cook, J., & Wolf, A. (1995). Automating Process Discovery through Event-Data Analysis. 17th International Conference on Software Engineering, (p. 73). Seattle, Washington, USA.

Cui, Z., Liu, Y., Huang, Z., He, H., & Wu, G.-L. (2019). Potential of artificial grasslands in crop rotation for improving farmland soil quality. Land Degradation & Development, 30, 2187–2196.

Detlefsen, N., & Jensen, A. (2007). Modelling optimal crop sequences using network flows. Agricultural Systems, 94(2), 566-572.

Dogliotti, S., Rossing, W., & van Ittersum, M. (2003). ROTAT, a tool for systematically generating crop rotations. European Journal of Agronomy, 19(2), 239-250.

Dustdar, S., Hoffmann, T., & van der Aal, W. (2005). Mining of ad-hoc business processes with TeamLog. Data & Knowledge Engineering, 55(2), 129-158.

Grignon, E., (2015). Itinéraires techniques des principales productions maraichères du québec comme outil permettant la gestion et le transfert efficace des connaissances liées à la régie de culture entre les conseillers et les producteurs. MAPAQ, Gouvernement Québécois. Consulté le 10/02/2021. Disponible sur : https://www.mapaq.gouv.qc.ca/sitecollectiondocuments/Agroenvironnement/1589_rapport.pdf

Gouvernement Canada. (2017, 05 03). Cadre Écologique National pour le Canada : Données SIG. (Gouvernement Canada) Consulté le 11 21, 2020, sur http://sis.agr.gc.ca/pages/nsdb/ecostrat/gis_data.html

Gouvernement Québec. (2019, 01 31). Zone agricole du Québec. (Données Québec) Consulté le 11 24, 2020, sur https://www.donneesquebec.ca/recherche/dataset/zone-agricole-du-quebec

Institut de la statistique Québec. (2020). AGRICULTURE. Dans Le Québec chiffres en main (p. 34). Québec: Bibliothèque et Archives nationales du Québec.

Janovicek, K., Hooker, D., Weersink, A., Vyn, R., Deen, B. (2021), Corn and soybean yields and returns are greater in rotations with wheat. Agronomy Journal. DOI:10.1002/agj2.20605

Pretty, J., et al. (2010). The top 100 questions of importance to the future of global agriculture. International Journal of Agricultural Sustainability, 8(4), 219–236.

Kirkegaard, J.A., Ryan ,M.H. (2014). Magnitude and mechanisms of persistent crop sequence effects on wheat. Field Crops Research.Vol.164,pp.154-165

Kirkegaard,J.A., Christen, O., Krupinsky, J., Layzell, D.,(2008). Break crop benefits in temperate wheat production. Field Crops Research.107, pp.185-195

Klein Haneveld, W., & Stegeman, A. (2005). Crop succession requirements in agricultural production planning. European Journal of Operational Research, 166(2), 406-429.

Kopittke, P., Menzies, N., Wang, P., McKennaa, B., & Lombica, E. (2019). Soil and the intensification of agriculture for global food security. Environment International, 132.

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience And Remote Sensing Letters, 14(5).

Le Ber, F., Benoît, M., Schott, C., Mari, J.-F., & Mignolet, C. (2006). Studying crop sequences with CarrotAge, a HMM-based data mining software. Ecological Modelling, 191(1), 170-185.

Leemans, S., Fahland, D., & Van der Aalst, W. (2018). Scalable process discovery and conformance checking. Software & Systems Modeling, 17, 599–631.

Leemans, S., Poppe, E., & Wynn, M. (2019). Directly Follows-Based Process Mining: Exploration & a Case Study. International Conference on Process Mining (ICPM). Aachen, Germany.

Levavasseur, F., Bouty, C., Barbottin, A., Verret, V., & Martin, P. (2015). Characterization of crop rotations variability by combining modelling and local farm interviews. 5th International Symposium for Farming Systems Design. 7-10 September 2015, Montpellier, France

Liakos, K., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. sensors, 18(8).

Ma, Y. (2019). Seed coating with beneficial microorganisms for precision agriculture. Biotechnology Advances, 37(7).

Nieves, D., Ramírez-Quintana, M., Monserrat, C., Ferri, C., & Hernández-Orallo, J. (2020). Learning alternative ways of performing a task. Expert Systems With Applications, 148.

Osman, J., Inglada, J., & Dejoux, J.-F. (2015). Assessment of a Markov logic model of crop rotations for early crop mapping. Computers and Electronics in Agriculture, 113, 234-243.

Porter, P. (2009). Crop Rotations in Organic Production Systems. Dans In Organic Farming: The Ecological System (pp. 49-67). Charles Francis (Ed.).

Pretorius, J., & Matthee, M. (2006). The Impact of Spatial Data on the Knowledge Discovery Process. Proceedings of the Conference on Information Technology in Tertiary Education. Pretoria, South Africa.

Reeves, D. (1994). Principles of Crop Rotation. Dans Cover Crops and Rotations (p. 127). Boca Raton, FL: J.L. Hatfield and B.A. Stewart (Ed.).

Salmon-Monviola, J., Durand, P., Ferchaud, F., Oehler, F., & Sorel, L. (2012). Modelling spatial dynamics of cropping systems to assess agricultural practices at the catchment scale. Computers and Electronics in Agriculture, 81, 1-13.

Statistique Canada. (2018, 01 23). Introduction à la Classification écologique des terres (CET) 2017. (Gouvernement Canada) Consulté le 11 21, 2020, sur https://www.statcan.gc.ca/fra/sujets/norme/ environnement/cet/2017-1

The Fraunhofer Institute for Applied Information Technology. (2020). Process Discovery. (PM4PY) Consulté le 11 25, 2020, sur https://pm4py.fit.fraunhofer.de/documentation

Turcotte,F.(2018). Prairies et pâturages pour chevaux : ce qu'il faut savoir. MAPAQ. Gouvernement Quebec. Consulté le 11/02/2021. Disponible sur : https://www.mapaq.gouv.qc.ca/fr/Regions/monteregie/articles/production/Pages/Prairies_paturages_chevaux_ce_quil_faut_savoir.aspx

United Nations. (2017). World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100. (Department of Economic and Social Affairs) Consulté le 11 24, 2020, sur https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html

USDA. (s.d.). Crop Rotation Practice Standard. (Agricultural Marketing Service) Consulté le 11 22, 2020, sur https://www.ams.usda.gov/grades-standards/crop-rotation-practice-standard

van der Aalst, W. (2011). Process Discovery: An Introduction. Dans Process Mining : Discovery, Conformance and Enhancement of Business Processes (pp. 125-156). Berlin, Heidelberg: Springer.

van der Aalst, W. (2011). Introduction. Dans Process Mining : Discovery, Conformance and Enhancement of Business Processes (pp. 1-25). Berlin, Heidelberg: springer.

van der Aalst, W. (2016). Inductive Miner Based on Event Log Splitting. Dans Process Mining : Data Science in action second edition (p. 223). Eindhoven, The Netherlands: Springer.

Yaramasu, R., Bandaru, V., & Pnvr, K. (2020). pre-season crop type mapping using deep neural networks. Computers and electronics in agriculture, 176.

Zhang, C., Liping, D., Lin, L., & Guo, L. (2019). Machine-learned prediction of annual crop planting in the US Corn Belt based on historical crop planting maps. Computers and Electronics in Agriculture, 166.