Journées de méthodologie statistique de l'Insee 2022

Extraction automatique de données issues d'images scannées : une illustration par les comptes sociaux d'entreprises

Laura GAIMARD (*), Adem KHAMALLAH (**)

(*) Insee, Direction des statistiques d'entreprises
(**) Insee, PSAR Analyse Territoriale

laura.gaimard@insee.fr
adem.khamallah@insee.fr

Mots-clés : Deep learning, Machine learning, Classification, Open Data

Domaines : Analyse des données et data science

Résumé (entre 350 et 900 mots environ)

Afin de produire les statistiques structurelles sur les entreprises, deux principales sources de données sont mobilisées : les sources administratives avec les déclarations fiscales et les sources d'enquêtes avec la réponse aux enquêtes annuelles (Enquête Sectorielle Annuelle et Enquête Annuelle de Production). Ces sources peuvent être incomplètes pour produire certaines informations, notamment sur le profilage des groupes de sociétés.

Une autre source de données, les comptes sociaux des entreprises, est également exploitée manuellement depuis quelques années afin de corriger ou de compléter certaines informations. Or, depuis fin 2019, l'Institut National de la Propriété Industrielle (INPI) met à disposition gratuitement en *Open Data* les informations provenant des greffes des tribunaux de commerce, dont les comptes sociaux déposés. Ces documents peuvent être récupérés massivement sous la forme d'un ensemble d'images scannées.

Cet article traite de l'expérimentation qui a été menée pour automatiser l'extraction des données d'intérêt issues de ces documents, avec l'exemple du tableau des liens capitalistiques (ou tableau des filiales et participations) entre unités légales. Ce tableau comporte des informations importantes pour les travaux de contrôle et d'amélioration de la qualité de certaines données. Cependant, le retrouver manuellement est particulièrement chronophage du fait de la longueur d'un compte social ainsi que du nombre de comptes à examiner, qui se compte en dizaines de milliers de documents.

Ce processus d'extraction est composé de deux étapes : une première de recherche de la page d'intérêt comportant le tableau, puis une seconde d'extraction des données du tableau au sein de cette page. Chacune de ces deux problématiques sera traitée *via* l'apprentissage supervisé. Les échantillons de travail sont construits en vue de l'entraînement de ces modèles.

Ce sujet de récupération automatique d'informations sur les entreprises est essentiel pour plusieurs instituts nationaux de statistiques. Par exemple, les équipes de Statistique Canada ont travaillé sur des problématiques similaires, comme l'extraction du chiffre d'affaires et des effectifs des documents comptables des sociétés.

Au delà de la thématique des comptes sociaux, ces modèles peuvent être répliqués dans d'autres problématiques pour d'autres usages, par exemple pour extraire des données issues de factures scannées ou de comptes publiés par des associations.

Abstract en Anglais (Texte de 5 à 10 lignes)

Elaboration of business structural statistics is based on two main data sources : surveys and administration. Those sources are incomplete : for instance, they lack information about financial links between legal units from a same group of companies.
Complementary data can be found in firms' social accounts, provided in *Open Data*. However, this documents are oftenly scanned. Thus, a way to extract the data of interest has to be used. In this paper, the focus is made on subsidiaries and affiliates table.
The method comes in two stages : the first is a model that finds the page of the table of interest inside the social account ; then a second model extracts and transforms it into structured exploitable data.

Introduction

Dans le cadre de la production des statistiques structurelles des entreprises et depuis une dizaine d'années, l'Insee met en oeuvre le concept d'entreprise au sens économique (ou entreprise profilée - EP). Ce concept est introduit dans la loi de modernisation de l'économie (LME) de 2008 dans les termes suivants : "**la plus petite combinaison d'unités légales qui constitue une unité organisationnelle de production de biens et de services jouissant d'une certaine autonomie de décision, notamment pour l'affectation de ses ressources courantes**".

Cette définition tient compte du fait que certaines unités légales (UL - personne morale ou physique) soient contrôlées par des groupes qui sont dirigés par une société mère.

Considérer l'entreprise au sens économique permet d'évaluer correctement les agrégats sectoriels de valeurs comptables non additives (en UL) comme le chiffre d'affaires ou les achats.

Par conséquent, l'entité *entreprise profilée* (Haag, 2019) est conçue par l'Insee et peut donc être difficilement enquêtée, au contraire de l'UL, brique de base des statistiques d'entreprises ayant une existence juridique et donc principale unité enquêtée.

Les différents agrégats économiques calculés au niveau de l'EP sont élaborés à partir de deux sources de données existantes au niveau des UL : les données d'enquêtes, avec les réponses aux enquêtes annuelles (Enquête Annuelle de Production (industrie) et Enquête Sectorielle Annuelle (autres secteurs)) et les sources administratives, avec les liasses fiscales transmises par la DGFIP [1].

Ces données, collectées au niveau des UL, sont toutefois incomplètes pour la détermination des caractéristiques des EP.

En effet, l'entreprise au sens économique est constituée d'une seule ou d'un groupe d'unité légale. Dans le cas où l'entreprise comporte une seule unité légale, les données de l'UL suffisent à calculer ses caractéristiques. Cependant, dans le second cas, les UL composant l'EP peuvent avoir divers liens commerciaux et financiers entre eux. Ces liens n'ont pas de sens économique. Il faut donc les soustraire des sommes des variables des UL dans le but de construire les EP, c'est le processus de consolidation. Pour ce faire, il faut connaître ces liens commerciaux et financiers, appelés flux, pour chaque EP à construire.

Ces flux ne peuvent pas être complètement calculés à l'aide des données d'UL existantes par manque d'information. C'est pourquoi deux méthodes sont actuellement mises en place pour calculer ces flux :

1. Une enquête spécifique en face-à-face avec 60 des plus grands groupes de sociétés opérant en France
2. Une estimation obtenue à partir de règles métiers à l'Insee

Malgré tout, ces deux méthodes ont des limites, en terme de charge de travail manuelle ou de qualité des données prédites. Elles ne sont pas pleinement satisfaisantes dans l'élaboration des caractéristiques des EP et ce processus est donc perfectible.

Depuis plusieurs années, les gestionnaires exploitent une autre source de données, les comptes sociaux des UL, qui permet de retrouver différentes informations complémentaires afin d'améliorer la qualité des données. Ils comportent toute la documentation comptable et financière de la

1. Direction Générale des Finances Publiques

société qui les publie, sous un format d'images scannées ou de document structuré (format pdf), avec différents tableaux et textes de format différent selon chaque compte social.

Ces informations sont actuellement analysés manuellement par un nombre important de gestionnaires. Ce travail améliore grandement la qualité des données publiées mais est très coûteux en temps et en moyens. Or, depuis fin 2019, les comptes sociaux sont mis à disposition en *Open Data* par l'Institut National de la Propriété Industrielle (INPI), et sont récupérables massivement par le biais d'une Interface de Programmation Applicative (API), ce qui pourrait permettre l'automatisation de la collecte de ces informations.

L'expérimentation se porte sur la recherche d'un tableau précis, celui des liens capitalistiques entre UL (ou tableau des filiales et participations). Ce tableau est assez complexe à rechercher, car il n'a pas une forme pré-définie et il n'est pas présent dans tous les documents analysés.

Après avoir présenté cette nouvelle source *Open Data* de la base du Registre National du Commerce et des Sociétés (RNCS), la deuxième partie de cet article s'articulera autour de la problématique de recherche de la page d'intérêt, avec la construction d'un modèle d'apprentissage supervisé pour prédire la présence ou non du tableau d'intérêt dans le document analysé. Ensuite, la partie sur l'extraction d'information issue d'une page scannée sera adordée, avec la construction de ce second modèle d'apprentissage supervisé, son application sur des images de tableaux. Enfin, ce processus sera illustré par différentes utilisations au sein de la statistique d'entreprise.

1 La base RNCS de l'INPI : une nouvelle source *Open Data* pour l'élaboration des statistiques structurelles

Les comptes sociaux, documents de base pour le processus, sont actuellement disponible dans le Registre National du Commerce et des Sociétés (RNCS) mis à disposition par l'INPI. Dans cette partie, nous présenterons la source Open Data de l'INPI, ainsi que les comptes sociaux et le tableau des liens capitalistiques, tableau d'intérêt dans le cadre de l'expérimentation.

1.1 Le registre national du Commerce et des sociétés

Créé en 1951, l'Institut National de la Propriété Industrielle (INPI) a pour missions de gérer l'ensemble des problématiques liées à la propriété industrielle (délivrer les titres de propriété industrielle, diffuser toutes les informations concernant la propriété industrielle...) ainsi que de centraliser le registre national du commerce et des sociétés. Les informations contenues dans ce registre proviennent de l'ensemble des greffes des tribunaux de commerces, qui gèrent chacun le registre du commerce et des sociétés (RCS).

Le RCS comporte l'ensemble des informations sur les entreprises de chaque greffe du tribunal de commerce. A sa création, l'entreprise a l'obligation de s'immatriculer à ce registre, si elle a une activité commerciale, en vue d'avoir une existence juridique. Il enregistre aussi les différents dépôts de documents réalisé par celles-ci : les actes et statuts lors de la création d'une entreprise qui permettent de définir ses objectif et son fonctionnement, les comptes sociaux, publication annuelle de ses états comptables et financiers, ou la convocation d'une assemblée générale pour les sociétés anonymes (SA).

En 2019, l'INPI a créée le portail data INPI, permettant d'accéder gratuitement à l'ensemble des informations contenues dans le RNCS pour l'ensemble des unités légales françaises ayant une activité commerciale, soit environ 5,9 millions d'unités. En effet, par la loi du 7 octobre 2016 pour une république numérique, l'INPI est dans l'obligation de diffuser l'ensemble des données non confidentielles contenues dans le RNCS. Trois types d'informations sont donc publiés :

1. les caractéristiques de chaque unité légales : l'ensemble de ses établissements, son statut juridique, sa dénomination, toutes les observations juridiques comme le dépôt de ses comptes sociaux ou des informations sur des rachats/fusions avec d'autres unités légales, les bénéficiaires effectifs et les représentants de l'unité
2. les actes et statuts déposés par l'unité
3. les comptes sociaux non confidentiels, présents chaque année depuis 2017

Au delà de cette consultation unitaire des informations sur les UL, l'INPI a aussi mis en place différentes interfaces pour récupérer massivement les informations présentées ci-desssus.

Dans cet article, nous nous intéresserons uniquement aux comptes sociaux, documents qui comportent l'information recherchée.

1.2 Les comptes sociaux : documentation comptable des entreprises

Les comptes sociaux comportent l'ensemble des données comptables et financières en norme internationale de l'unité légale pour l'année de dépôt. Il contient aussi des documents attestant la validité des comptes, ainsi que différents rapports liés au statut juridique de l'entité (notamment pour les Sociétés Anonymes). Cette publication est annuelle, et doit être réalisée dans les six mois après la date de clôture d'exercice de la société.

Les unités légales, selon leur statut juridique, sont soumis à une obligation de dépôt de leur comptes sociaux. Seuls les entrepreneurs individuels et les sociétés ayant un régime fiscal simplifié ne sont pas soumis à cette obligation. Ces comptes déposés peuvent être rendus confidentiels, si la société respecte certains critères de taille en fonction de sa catégorie d'entreprise. Néanmoins, les grandes entreprises sont soumis à l'obligation de publication, ainsi que les petites et moyennes entreprises appartenant à un groupe.

Enfin, le dépôt de ces comptes peut s'effectuer sur place, par courrier ou en ligne. Dans les deux premiers cas, la version imprimée est scannée par le greffe du tribunal de commerce puis envoyés à l'Inpi. Cela concerne la majorité des documents transmis (environ 80% des comptes analysés). Ceux déposés en ligne (20 %) sont des documents structurés.

Dans le cadre du profilage des groupes, seules les données des UL appartenant à un groupe nous intéresse. D'après l'ensemble des obligations de publication énoncés, il est possible de récupérer l'intégralité de leurs comptes sociaux. Ainsi, l'exhaustivité du champ est couvert.

Toutefois, il existe deux limites à l'utilisation de ces comptes sociaux pour l'améliorer la production des statistiques structurelles des entreprises :

— Les données comptables publiées sont celles de la société au niveau monde, alors que les données produites sont au niveau France, ce qui implique des possibles écarts, notamment pour les UL produisant à l'étranger.

— certains groupes ont opté pour la publication de comptes consolidés, en incluant les données comptables de l'ensemble des UL inclus dans son périmètre de consolidation. Ces données peuvent être très utiles si le périmètre de consolidation ne comprend uniquement des UL français ; malheureusement, ce n'est pas toujours le cas.

Un compte social contient une grande quantité d'informations pouvant figurer de manière plus ou moins dense. Cet article abordera en particulier la recherche du tableau des liens capitalistiques entre UL, aussi appelé le **tableau des filiales et participations**.

1.3 Le tableau des filiales et participations

Ce tableau établit la liste des détentions de l'UL *mère*, celle qui publie le compte social, avec l'ensemble des UL *filles*, celles qui sont inscrites dans le tableau. De plus, des informations sur la valeur comptable des titres de participation y figurent, notamment les dividendes perçus. Ces informations sont essentielles dans le calcul de flux financiers et donc dans le calcul d'une entreprise profilée.

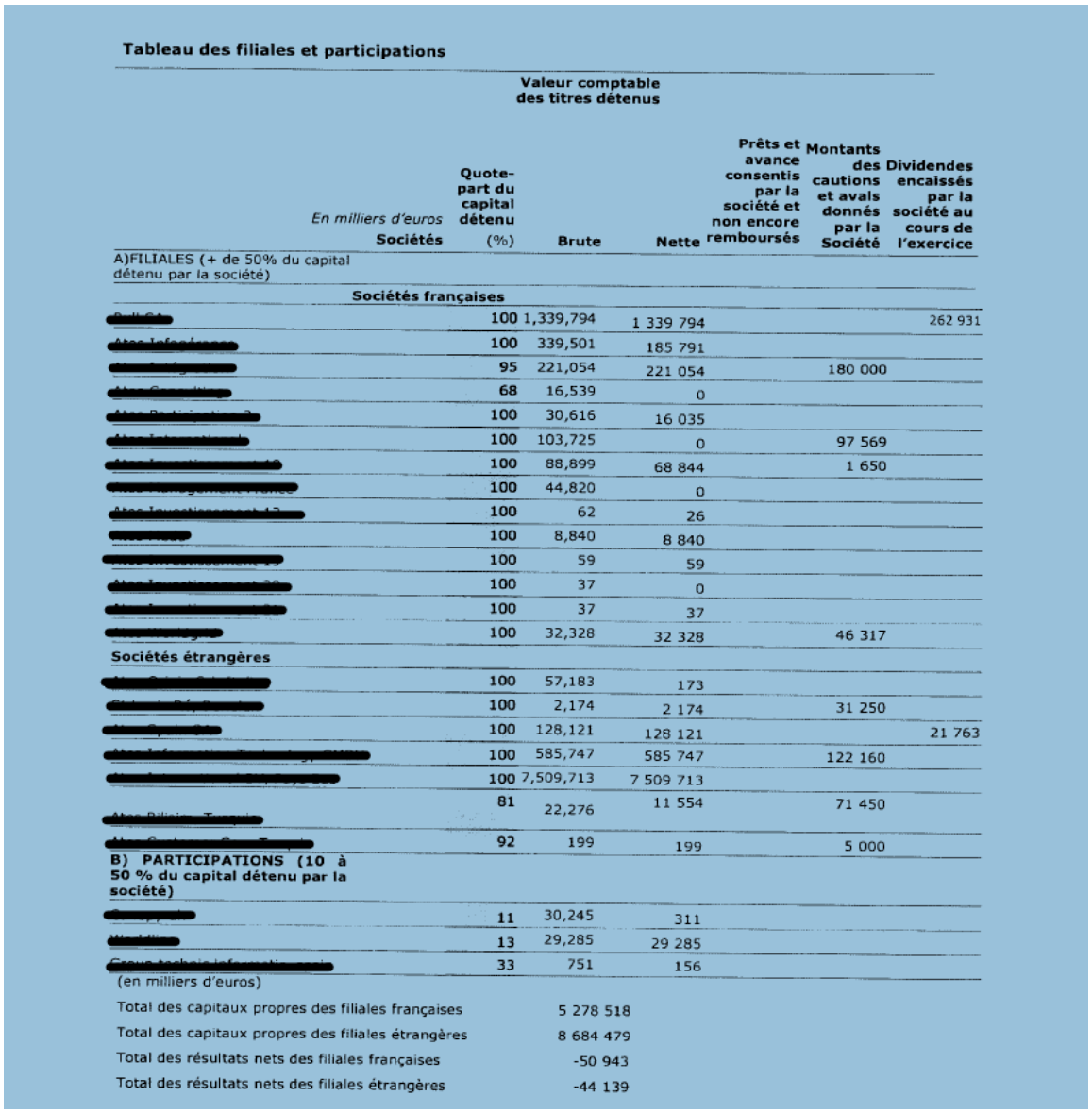

Tableau des filiales et participations

En milliers d'euros Sociétés	Quote-part du capital détenu (%)	Valeur comptable des titres détenus Brute	Nette	Prêts et avance consentis par la société et non encore remboursés	Montants des cautions et avals donnés par la Société	Dividendes encaissés par la société au cours de l'exercice
A)FILIALES (+ de 50% du capital détenu par la société)						
Sociétés françaises						
[redacted]	100	1,339,794	1 339 794			262 931
[redacted]	100	339,501	185 791			
[redacted]	95	221,054	221 054		180 000	
[redacted]	68	16,539	0			
[redacted]	100	30,616	16 035			
[redacted]	100	103,725	0		97 569	
[redacted]	100	88,899	68 844		1 650	
[redacted]	100	44,820	0			
[redacted]	100	62	26			
[redacted]	100	8,840	8 840			
[redacted]	100	59	59			
[redacted]	100	37	0			
[redacted]	100	37	37			
[redacted]	100	32,328	32 328		46 317	
Sociétés étrangères						
[redacted]	100	57,183	173			
[redacted]	100	2,174	2 174		31 250	
[redacted]	100	128,121	128 121			21 763
[redacted]	100	585,747	585 747		122 160	
[redacted]	100	7,509,713	7 509 713			
[redacted]	81	22,276	11 554		71 450	
[redacted]	92	199	199		5 000	
B) PARTICIPATIONS (10 à 50 % du capital détenu par la société)						
[redacted]	11	30,245	311			
[redacted]	13	29,285	29 285			
[redacted]	33	751	156			

(en milliers d'euros)

Total des capitaux propres des filiales françaises	5 278 518
Total des capitaux propres des filiales étrangères	8 684 479
Total des résultats nets des filiales françaises	-50 943
Total des résultats nets des filiales étrangères	-44 139

FIGURE 1 – Exemple de tableau des filiales et participations

Indication de lecture : Chaque ligne correspond à une UL détenue par la société qui publie les comptes sociaux. Pour des raisons de confidentialités, les raisons sociales de ces UL ont été cachées. Chaque colonne correspond à une donnée précise sur les liens entre l'UL mère et les UL filles : taux de détention de l'UL, valeur nette et brute des titres de participation, prêts et dividendes

Cependant, la difficulté réside dans le format, qui n'est pas standard pour tous les tableaux de liens capitalistiques. En effet, ce tableau peut comporter plus ou moins d'informations selon la société qui publie, avec des différences notables sur la structure du tableau (labels des variables, positions, orientation, etc.). Par exemple, les données des UL filles peuvent être au niveau des lignes comme des colonnes, bien que ce dernier cas soit plus rare.

Dans la partie suivante, nous allons construire une méthode qui permet de repérer automatiquement le tableau d'intérêt au sein du compte social, en indiquant la page du document où il se trouve.

2 Comment retrouver la page du tableau d'intérêt au sein d'un document ?

Les comptes sociaux sont des documents comportant entre plusieurs dizaines et plusieurs centaines de pages, sous forme structurées ou d'images (*cf* Les comptes sociaux : documentation comptable des entreprises). Afin de récupérer la page comportant le tableau recherché, un modèle de classification (supervisé) est mis en place à partir de pages labelisées de la présence ou non du tableau d'intérêt. Cette partie développera l'ensemble du processus permettant de retrouver la page comportant le tableau parmi les 490 pages constituant la base de travail :

1. Pré-traitements de récupération des mots de chaque page
2. Constitution de la base d'entraînement des pages de comptes sociaux
3. Entraînement et application du modèle d'apprentissage supervisé pour l'analyse textuelle.

2.1 Pré-traitements des images

2.1.1 Océrisation

Comme expliqué au sein de ce document, la majorité des comptes sociaux sont sous forme d'images scannées. Ce format est inexploitable dans le cadre de travaux d'analyse textuelle. Il convient d'**extraire** les données de l'image pour que les caractères soient reconnaissables par l'ordinateur : cette transformation est appelée une **océrisation** [2].

Ce procédé se heurte à deux limites qui vont nous intéresser particulièrement :

1. La qualité de l'océrisation dépend fortement de la qualité de l'image. Dans notre cas, une image "mal" scannée, *ie.* une image "de travers" ou avec des caractères difficilement reconnaissables, donnera des résultats de mauvaise qualité.
2. C'est un processus coûteux en temps machine selon la taille de l'image.

Ainsi, appliquer cette procédure à chaque page de nos comptes sociaux n'est donc pas optimal, car une partie est scannée, une autre est issue de pages structurées ou mélange les deux types de formats au sein d'une même page. Il s'agit donc d'océriser la page seulement si cela est nécessaire. Ainsi, un critère d'océrisation pour une page donnée est construit à partir de son nombre de mots. En effet, une image aura un nombre de mots très faible par rapport à une page structurée. La page sera donc à océriser si son nombre de mots est inférieur à un certain seuil s.
Ce seuil s est déterminé par un pseudo bootstrap, tel que : [3]

1. L'ensemble des pages composant la base de travail vont être tirées et analysées, afin de récupérer le nombre de mots contenu dans chaque page et son type (scannée ou structurée)
2. On tire B = 50 échantillons de N pages avec remise pour chaque type de page (structurée et scannée). Pour chaque échantillon, N correspond au nombre de pages scannées et de pages structurées de l'échantillon initial.

2. Du sigle anglais OCR, pour Optical Character Recognition.
3. La méthode de détermination reste à perfectionner.

3. Pour chaque échantillon B_k de pages structurées et scannées, on calcule la moyenne du nombre de mots par pages de l'échantillon bootstrap
4. A partir des moyennes du nombre de mots obtenus pour chaque échantillon bootstrap, on calcule un intervalle de confiance à 95% du nombre de mots pour une page scannée et pour une structurée par la méthode du percentile bootstrap (Efron, 1982)

Pour une page scannée, l'intervalle de confiance à 95% du nombre moyen de mots est [1.01 ; 45.51] et pour une page structurée, il est égal à [78.14 ; 358.15]. En moyenne, le nombre de mots d'une page scannée a 95% de chance d'être compris entre 1.01 et 45.51 et celui d'une page structurée a 95% de chance d'être compris entre 78.14 et 358.15.

Nous devons donc choisir un seuil du nombre de mots pour l'océrisation compris entre 45.51 et 78.14, afin d'avoir le maximum de pages scannées et le minimum de pages structurées à océriser. Nous avons choisi ici le seuil de 70 mots par page.

Le moteur d'océrisation utilisé est **tesseract** (Smith, 2005), dont l'utilisation est très répandue dans divers domaines.
Ce moteur comprend différents retraitements pour améliorer la qualité de l'océrisation, ainsi qu'un indicateur de précision de l'extraction du caractère, qui sont deux points essentiels pour l'obtention de la qualité optimale d'extraction.

2.1.2 Approche bag-of-words

La représentation par bag of words (Zhang et al., 2010) (ou sac de mots en français) est une description de document très utilisée en recherche d'information dans un document. Le bag of words est le résumé du document par les fréquences des mots qui le composent. Par exemple, on comptera le nombre d'apparitions du mot "entreprise" dans le compte social.

Ainsi, pour tous les comptes sociaux, on comptera ces fréquences pour chaque mot existant, en ayant au préalable nettoyé le document des stop words (ou mots creux en français). Il s'agit juste, dans notre cas, de retirer tous les mots non signifiants qui polluent l'analyse textuelle plus qu'ils n'apportent de sens comme les articles ("le", "des", ...). De plus, tous les caractères spéciaux, les différences de casse ainsi que les chiffres sont supprimés, car ils n'apportent pas d'information utilisable. Toutefois, le nombre de chiffres dans la page est calculé avant leur suppression, car c'est un déterminant dans la recherche du tableau.

Ainsi, pour chaque page d'un compte social, un vecteur de fréquences sera obtenu. En concaténant tous ces vecteurs, la **matrice documents-termes** est construite. Cette matrice est composée de chaque page de l'échantillon de travail, avec un compteur de l'ensemble des mots restant pour chaque page. Elle va constituer l'ensemble des covariables qui seront associées la variable d'intérêt (présence du tableau des liens capitalistiques dans la page).

Cependant, il apparaît naturellement (par la nature du langage et du champ restreint que sont les comptes sociaux) que certains mots sont très rares et ne sont pas utiles dans la recherche de la page d'intérêt. Par souci d'efficacité au niveau de la vitesse de calcul, de l'espace de stockage et de la performance de l'analyse textuelle, il convient de réduire la dimension de cette matrice, en supprimant les mots avec une fréquence d'apparition inférieure à 2%.

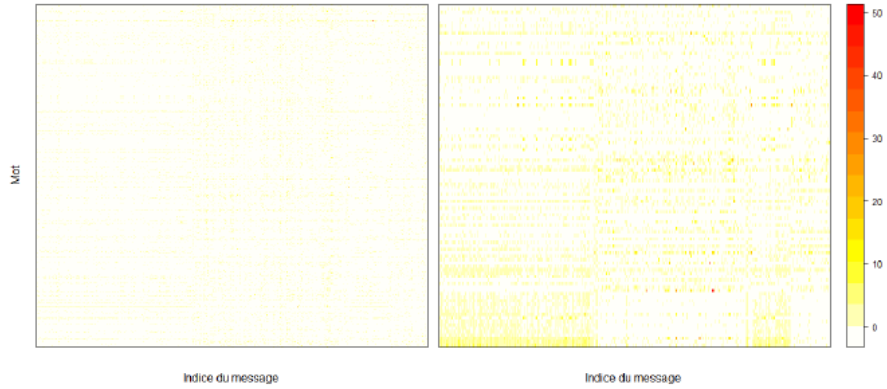

FIGURE 2 – Représentation graphique de la matrice documents-termes après suppression des mots rares

Lecture : La fréquence de l'ensemble des mots est affiché en fonction du document. Un mot rare sera affiché en blanc, car sa fréquence est proche de 0. Après la suppression des mots rares, on peut observer la présence de bien plus de points jaunes, soit des mots présents plus fréquemment.

Cette suppression permet une optimisation du temps de calcul sans perte significative d'informations permettant la reconnaissance des pages comportant le tableau des filiales et participations.

2.2 Analyse de la présence du tableau dans la base de travail

Le paragraphe précédent concernait la construction des covariables. Il reste à construire la variable d'intérêt pour l'ensemble des pages constituant la base de travail.

Les pages constituant cette base de travail n'ont pas été sélectionnées aléatoirement. En effet, le tableau recherché est rare dans les comptes sociaux : il est présent sur une page parmi plusieurs dizaines lorsqu'il existe dans les comptes sociaux). Pour obtenir un modèle avec une bonne prédiction de la présence ou non du tableau sur la page analysée, les pages contenant le tableau sont sur-représentées dans ce jeu de données par rapport à leur présence dans les comptes sociaux. Il faudra donc faire attention au taux de faux positif lors de l'entraînement et de la validation du modèle.

Parmi les pages du jeu de données, 50 pages ont été annotées à la main de la présence ou non du tableau. Ensuite, un premier modèle a été entraîné sur ce jeu de données de 50 pages. Ce modèle est ensuite appliqué sur 420 comptes sociaux afin de récupérer la page ayant la plus grande probabilité de présence du tableau recherché dans chaque document. Pour chaque page obtenue, un contrôle manuel a été réalisé par le biais d'un script R pour annoter la présence effective ou non du tableau.

Une fois ce travail réalisé, la base de données contient 490 pages annotées ou non de la présence du tableau.

2.3 Construction du modèle de détection

Une fois la variable d'intérêt obtenue, ainsi que la matrice document-termes comportant l'ensemble des covariables du modèle, le modèle de détection de la présence du tableau des

filiales et participation peut être construit.

Trois modèles de détection du tableau au sein d'une page ont été confrontés.

1. Support Vector Machine (SVM) (Schölkopf et al., 2002)
2. RandomForest (RF) (Breiman, 2001)
3. Adaboost (Freund et al.,1997)

Les critères de comparaison seront le taux de bien classés et le taux de faux positif parmi les pages. Autrement dit, on considérera que le modèle le plus performant sera celui qui se trompe le moins dans le classement des pages analysées. Pour rappel, un faux positif est la détection à tort du tableau pour une page donnée.

Les trois modèles, une fois entraînés, ont été testés sur un échantillon test de 88 pages, dont 44 comportant le tableau des filiales et participations :

	Random Forest	**SVM**	**Adaboost**
Taux de bien classés	0.9318	0.8977	0.8182
Taux de vrais positifs	0.9545	0.9773	0.9773
Taux de faux positifs	0.0909	0.1818	0.3409
Bien prédites (%)	0.8181	0.7272	0.7954

Le "taux de vrais positif" est la part de pages comportant le tableau d'intêret et bien prédites
"Bien prédites" étant la part de pages bien prédites parmi les comptes sociaux contenant un tableau.

D'après le tableau suivant, le meilleur modèle s'avère être le RF [4]. Le taux de faux positif (présence du tableau à tort dans la page) est aussi intéressant à analyser, notamment à cause de la constitution de la base de travail qui comporte une sur-représentation des pages contenant le tableau des filiales et participations. Sur ce critère, le RF est aussi le meilleur modèle parmi les trois, puisqu'il a le taux de faux positif le plus faible.

Les covariables du modèle étant la fréquence des mots de chaque page, il est intéressant de regarder les mots les plus discriminant pour la présence du tableau des filiales et participation. En regardant l'importance des variables du modèle, *ie.* quels sont les mots qui permettent le plus de réduire les erreurs de classement des pages, on constate que certains mots sont très discriminants, comme les mots *filiales* ou *participations* :

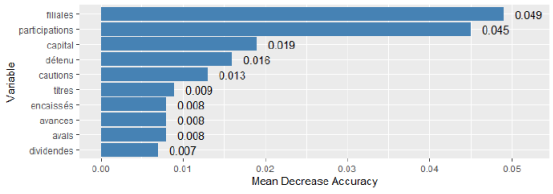

FIGURE 3 – Graphique des 10 variables ayant l'importance la plus élevée dans le modèle RF

Note : La Mean Decrease Accuracy d'une variable mesure quelle est la perte en précision du modèle, si cette variable est exclut des covariables

Lecture : La variable filiales est celle ayant la Mean Decrease Accuracy la plus élévée, sa présence dans une page va fortement orienter le modèle pour indiquer la présence du tableau dans la page

4. Cela peut s'expliquer par la petite taille de l'échantillon par rapport à la dimension de la matrice document-termes

Le taux de pages bien prédites est de 82% pour le modèle RF. En analysant les erreurs de prédictions de ce modèle, deux types d'erreurs sont fréquentes :

— des pages comportant un vocabulaire proche de celui contenu dans le tableau des filiales et participations (comme les mots *filiales*, *participations*, *capital* ou *titres*). Ces pages vont être analysées manuellement pour mieux observer les différences avec les pages comportant le tableau recherché. De plus, elles seront annotées et ajoutées à l'échantillon d'apprentissage

— des pages comportant le tableau des filiales et participations vide. Il faudrait prendre en compte le nombre de chiffre de la page dans le modèle, afin de ne pas renvoyer ces pages.

Un travail d'enrichissement de la base de travail doit donc être réalisé, afin d'avoir de meilleures prédictions du modèle.

Ce modèle a été entraîné et validé sur des extraits de comptes sociaux. Toutefois, le but final de ce processus est son application sur un compte social entier. Ce tableau n'étant pas présent sur tous les comptes sociaux, il faut trouver un critère permettant de récupérer uniquement les pages avec une forte probabilité de présence de ce tableau. Pour cela :

— le modèle entraîné est appliqué sur l'ensemble des pages du compte social, puis la page ayant la plus forte probabilité de présence du tableau est récupérée

— si cette probabilité de présence est supérieure à un seuil s, on considère que le tableau est présent sur cette page

Pour déterminer ce seuil s, on récupère l'ensemble des probabilités de présence du tableau de l'échantillon de travail, puis on analyse la distribution de ces probabilités selon la présence ou non du tableau. Le seuil fixé doit permettre de :

— récupérer le maximum de pages comportant le tableau

— récupérer le minimum de pages ne comportant pas le tableau

A partir de ces distributions, ce seuil est fixé à 0.9. Le processus final renvoie la page du compte social qui possède la plus grande probabilité et qui est supérieure à 0.9.

Cette démarche est appliquée à 100 unités légales prises aléatoirement parmi les unités légales appartenant à un groupe de société. On obtient :

Nombre d'unités légales analysées	100
Nombre de comptes sociaux obtenus	74/100
Part moyenne de pages océrisées dans un compte social	87%
Nombre de tableaux trouvés	46/74
Nombre d'erreur	6/46 (faux positifs uniquement)

Les erreurs de classement sont similaires à celles déjà observées précédemment, ce qui constituent les limites actuelles de ce calcul.

D'autres instituts nationaux de statistiques, comme Statistique Canada, ont travaillé sur ce sujet (Bejju et al., 2019). Leur objectif est différent du notre : leur but est de retrouver le résultat net et le total du bilan sur un document financier d'une société. Ces documents ne sont pas au format image, mais structurés, ce qui implique moins de pré-traitements pour obtenir les différentes données de chaque page. Toutefois, leur modèle retrouve la page comportant l'information recherchée avec une précision de 96% .

3 Extraction des résultats d'un tableau

Après avoir présenté notre premier modèle de détection de la page contenant le tableau filiales et participations, notre second modèle a pour tâche d'extraire les informations contenues dans ce tableau.

Pour établir ce processus d'extraction automatique d'informations de ces tableaux, certaines difficultés doivent être prises en compte :

1. la structure du tableau varie selon les différents comptes sociaux (absence de certaines variables, transposition du tableau)
2. les pages extraites des comptes sociaux sont majoritairement des images (environ 80% des pages comportant des comptes sociaux sont des documents scannées, contre 20% de document structuré)

Après une revue de la littérature sur ce sujet d'extraction d'informations issues d'images, nous avons constaté que le modèle TableNet (Paliwal et al., 2020) répondait aux différents besoins auxquels nous étions confrontés. Il permet de détecter et d'extraire les informations contenues dans des tableaux au format image.

Ce modèle, de type *encoder-decoder*, utilise deux réseaux de neurones convolutif (CNN) qui utilise, comme *encoder*, le modèle VGG-19 (Simonyan et al., 2014) pré-entraîné sur la base *ImageNet* (Deng et al., 2015) et comme *décoder*, deux réseaux convolutifs indépendants permettent d'estimer l'emplacement du tableau et celui de ses colonnes. L'architecture du modèle TableNet est représentée dans le schéma ci-dessous, issus de l'article sur le modèle TableNet :

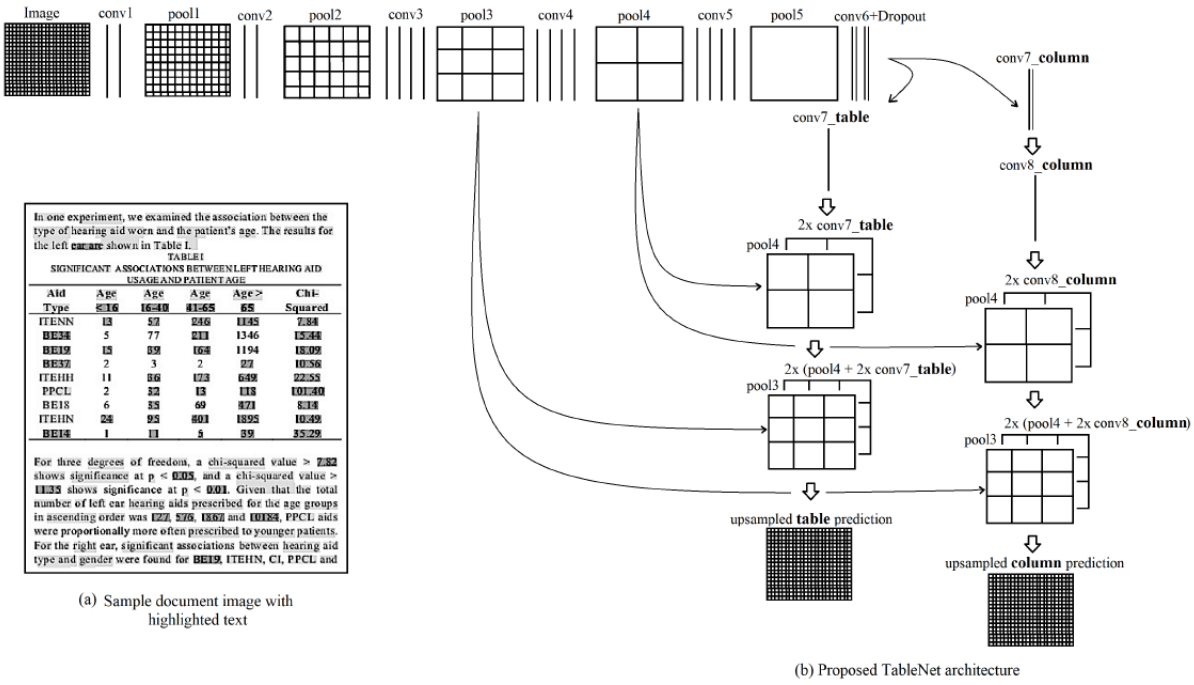

FIGURE 4 – Architecture du modèle TableNet

VGG-19 est un réseau de neurones convolutif, permettant de réaliser de la reconnaissance d'images. Ce modèle a remporté en 2014 la compétition ILSVRC (*ImageNet Large Scale Visual Recognition Challenge*) (Russakovsky et al., 2015) en atteignant une haute précision pour le classement d'images dans différentes catégories. Il a été entraîné durant plusieurs semaines à partir de la base d'images ImageNet avec une importante puissance de calcul. L'utilisation de ce modèle pré-entraîné est donc très intéressante pour un gain de temps d'entraînement du modèle TableNet.

ImageNet est une base de données de plus de 14 millions d'images annotées et réparties en plus de 1000 classes. Cette base est utilisée pour l'entraînement de la plupart des modèles de traitement d'image. Elle fait partie des plus grandes base de données d'images en *Open Data* annotée et de très bonne qualité.

3.1 Collecte des données de travail

Pour la constitution de la base de travail en vue de l'entraînement et de la validation du modèle, nous avons utilisées deux types d'images différentes :

1. le jeu de données Marmot, qui est composé de 2000 pages au format PDF, majoritairement des extraits provenant d'articles de recherches et comportant divers tableaux de données. Dans notre expérimentation, seulement une partie de ce jeu de données a pu être utilisé, soit 500 pages. Ces données proviennent de documents rédigés en anglais.
2. différentes pages extraites à partir du modèle précédent, comportant le tableau des filiales et participations (voir section 2). Une centaine de pages ont été utilisées, parmi l'ensemble des pages prédites dans le modèle précédent.

Les données nécessaires à l'entraînement du modèle Tablenet sont :

— la page à analyser le tableau au format .bmp
— un *masque* du tableau contenu dans la page, afin d'avoir l'emplacement exact du tableau au sein de la page. Ce *masque* est au format .bmp
— un *masque* comportant l'emplacement des colonnes du tableau contenu dans la page, afin d'avoir l'emplacement exact de ces colonnes au sein de la page. Il est également au format .bmp

Le jeu de données Marmot fournit l'ensemble des données utiles pour l'entraînement du modèle. Cependant, pour les pages issues des comptes sociaux, les deux masques du tableau et de ses colonnes ne sont pas disponibles. Nous avons donc dû les déterminer.

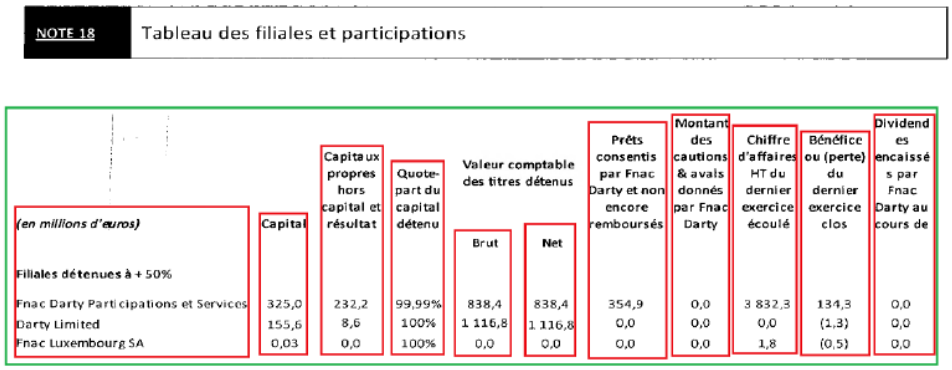

FIGURE 5 – Labellisation manuelle de tableau des filiales et participations

Pour cela, un travail manuel a été réalisé afin de déterminer ces deux masques. Ce travail consiste, à l'aide d'un script *Python*, à sélectionner dans un premier temps l'emplacement du tableau avec un curseur, puis de sélectionner l'ensemble des colonnes. Sur la figure ci-dessus, la tableau sélectionné apparaît en vert, et l'ensemble des colonnes apparaît en rouge. Une fois la sélection effectuée, ce script renvoie les deux masques souhaités. Dans la figure ci-dessous, pour une page comportant un tableau en particulier, les deux masques associés à ce tableau sont affichés. Ces deux masques permettront au modèle TableNet d'identifier les zones des images qui correspondent au tableau et à ses colonnes.

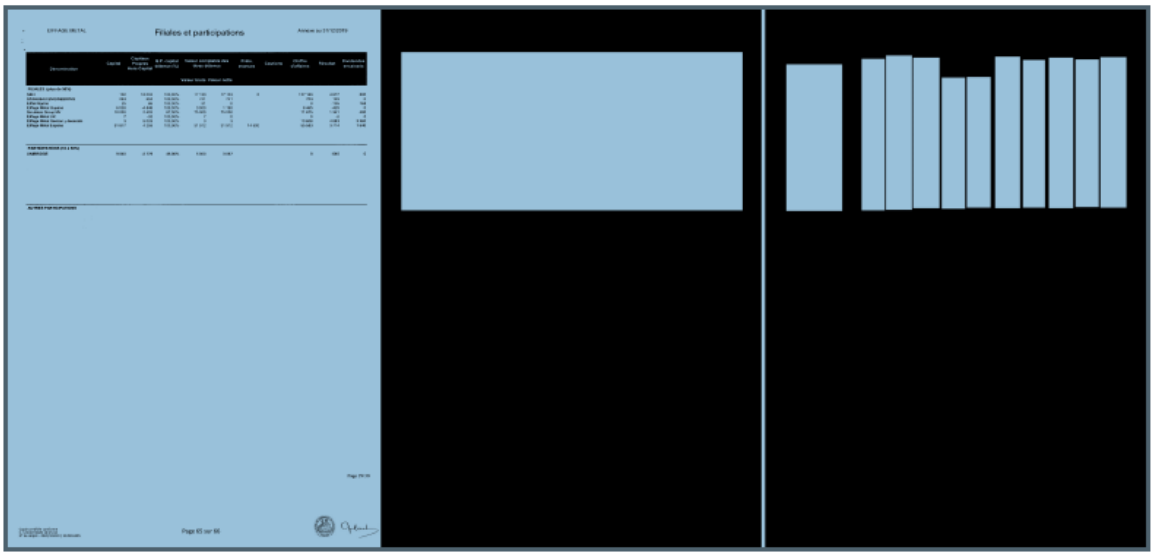

FIGURE 6 – Comparaison masques obtenus avec la page initiale

Nous avons ainsi obtenu pour les deux sources d'images les données nécessaires à l'entraînement du modèle, c'est-à-dire les images avec leurs masques associés.

Pour cette expérimentation, nous avons entraîné le modèle avec deux jeux de données différents :

1. un jeu de données de 500 images comportant uniquement des images du jeu de données marmot
2. un jeu de données *mixte* comportant 400 images de la base Marmot et 100 images de comptes sociaux

3.2 Analyse de la qualité de prédiction du modèle

Le modèle TableNet prédit l'emplacement du tableau et de ses colonnes au sein de l'image analysée en calculant des masques. Les performances de prédiction de ce modèle peuvent se mesurer à l'aide de l'**indice de Jaccard** (Jaccard, 1901) ou *intersection over union*, un coefficient de communauté entre l'aire initiale et l'aire prédite des emplacements du tableau et de ses colonnes.

Soit	S	l'emplacement réel du tableau ou de la colonne
	$\hat{S}$	l'emplacement prédit du tableau ou de la colonne

$$J(S, \hat{S}) = \frac{S \cap \hat{S}}{S \cup \hat{S}}$$

L'objectif est d'avoir la meilleure prédiction possible de l'emplacement du tableau et de ses colonnes. Plus l'indice de Jaccard est proche de 1, et meilleure est la prédiction des emplacements.

Pour chacun des masques (tableau et colonnes), l'indice de Jaccard sera :

Soit	$S_{tableau}$	l'emplacement réel du tableau
	$\hat{S}_{tableau}$	l'emplacement prédit du tableau
	S_i	l'emplacement réel de la colonne i
	$\hat{S}_i$	l'emplacement prédit de la colonne i
	n	le nombre de colonnes du tableau

Pour le tableau	$\mathrm{J}_{tableau}(S_{tableau}, \hat{S}_{tableau}) = \frac{S_{tableau} \cap \hat{S}_{tableau}}{S_{tableau} \cup \hat{S}_{tableau}}$
Pour une colonne i	$\mathrm{J}_i(S_i, \hat{S}_i) = \frac{S_i \cap \hat{S}_i}{S_i \cup \hat{S}_i}$
Pour l'ensemble des colonnes	$\overline{J}_{colonnes} = \frac{1}{n} \sum_{i=1}^{n} J_i(S_i, \hat{S}_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{S_i \cap \hat{S}_i}{S_i \cup \hat{S}_i}$

Un second indicateur de la qualité de prédiction de l'emplacement du tableau et de ses colonnes est le **taux de perte** de la prédiction, *ie* le nombre de pixels mal classés par la prédiction. Les pixels mal classés sont définis comme les erreurs de première et deuxième espèces :

— les pixels classés comme appartenant au tableau ou aux colonnes à tort (donc des pixels en dehors de la zone réelle du tableau ou des colonnes , mais prédit comme appartenant à cette zone)

— les pixels classés comme n'appartenant pas au tableau ou aux colonnes à tort (donc des pixels à l'intérieur de la zone réelle du tableau ou des colonnes, mais prédit comme n'appartenant pas à cette zone)

Soit	$p_{(x,y)}$	indicatrice de présence dans le tableau réel du pixel ayant les coordonnées (x,y) dans l'image
	$\hat{p}_{(x,y)}$	indicatrice de présence dans le tableau prédit du pixel ayant les coordonnées (x,y) dans l'image
	m	le nombre de pixels en largeur de l'image
	n	le nombre de pixels en longueur de l'image

Le nombre de pixel mal classé est défini par :

$$MC = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{p_{(i,j)}=1} * \mathbb{1}_{\hat{p}_{(i,j)}=0} + \mathbb{1}_{p_{(i,j)}=0} * \mathbb{1}_{\hat{p}_{(i,j)}=1}$$

Le taux de perte est défini par :

$$P = \frac{MC}{nm}$$

Ces deux indicateurs nous permettent d'analyser la qualité de la prédiction du modèle pour l'emplacement des tableaux et des colonnes. Ils seront calculés pour chacune des prédictions (pour l'emplacement du tableau et de ses colonnes).

3.3 Résultats de la prédiction du modèle TableNet

Nous avons effectué deux entraînements du modèle TableNet, avec les deux bases de travail déterminées précédemment : une première constituée uniquement du jeu de données Marmot, une seconde *mixte* composée d'images provenant de Marmot et des comptes sociaux.

3.3.1 Résultats avec les données de la base Marmot

Pour rappel, la base de travail est constituée de 500 images provenant du jeu de données Marmot. Parmi ces images, 400 ont servi pour l'entraînement du modèle et 100 pour la validation.

Dans l'exemple ci-dessous, le modèle entraîné est testé sur l'image de gauche qui comporte le tableau des filiales et participations. Les deux images de droite sont les masques du tableau et de ses colonnes prédits par le modèle TableNet. Pour chacun des masques, la zone en jaune correspond à l'emplacement prédit du tableau et de ses colonnes.

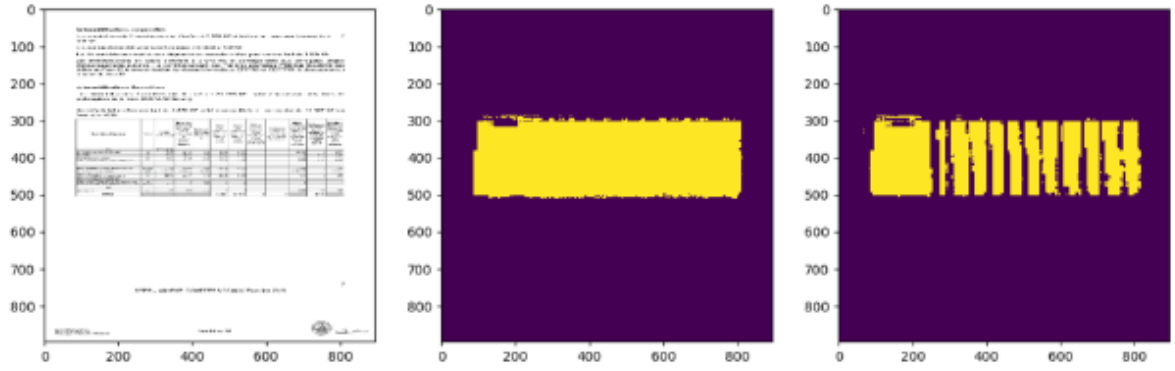

FIGURE 7 – Exemple des masques prédit avec le modèle entraîné avec le jeu de données Marmot

Les zones prédites par le modèle sont similaires à la zone initiale du tableau et des colonnes dans l'image. Cependant, quelques artefacts existent visuellement au niveau de la limites des zones prédites : au niveau du masque du tableau, les bords ne sont pas très nets et ils le sont encore moins pour les bords des colonnes.

Les indicateurs de qualité cités précédemment sont calculés pour l'ensemble des images composant l'échantillon test. Nous obtenons les résultats présentés dans le tableau ci-dessous :

	Table		**Colonnes**	
	Indice de Jaccard	Taux de perte	Indice de Jaccard	Taux de perte
Base de test	0.87	0.06	0.80	0.10

En moyenne sur l'échantillon test, l'indice de Jaccard est assez proche de 1 pour la prédiction du tableau et de ses colonnes. Le taux de perte est lui aussi assez faible : 6% des pixels sont mal classés dans les images de la base de test. Les prédictions réalisées par le modèle sont donc d'une très bonne qualité. Ceci n'est pas surprenant car le modèle utilise les poids pré-entraînés du modèle VGG-19 sur la base d'image *ImageNet* (Simonyan et al., 2014).

Toutefois, l'application de ce modèle sur des images issues de comptes sociaux prédit des masques de moins bonne qualité que sur les données de la base Marmot. L'explication a cet écart proviendrait de la qualité dégradée des images issues des comptes sociaux par rapport aux images de la base Marmot. Une idée d'amélioration serait d'extraire les comptes sociaux pour y entraîner le modèle pour que les champs des données d'entraînement et de test coïncident.

3.3.2 Résultats avec les données mixtes

Une base de travail constituée de 400 images provenant du jeu de données Marmot sont enrichies de 100 images provenant des comptes sociaux. Parmi ces images, 400 ont servi pour l'entraînement du modèle et 100 pour la validation de celui-ci.

Dans l'exemple ci-dessous, le modèle entraîné sur le jeu de données mixte est appliqué à l'image de gauche qui comporte le tableau des filiales et participation. Cette image est identique à celle de la partie précédente, afin de comparer visuellement les masques de prédictions des deux modèles.

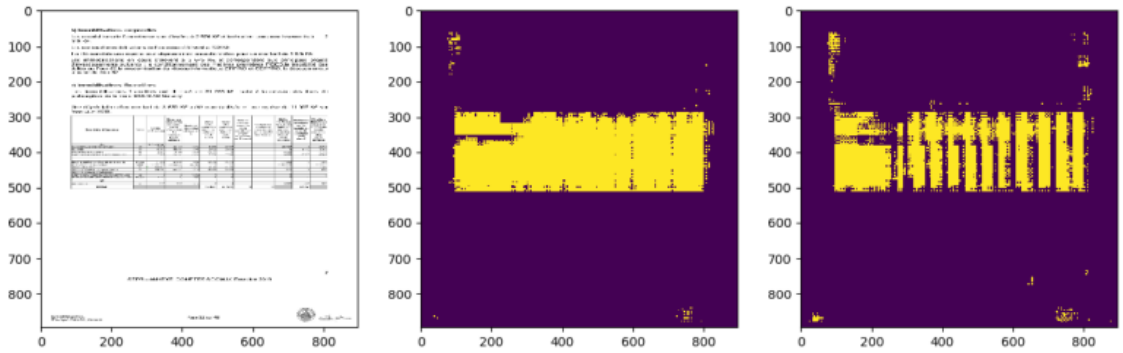

FIGURE 8 – Exemple des masques prédit avec le modèle entraîné avec le jeu de données mixtes

Les zones prédites par le modèle sont similaires à celle initiale du tableau et des colonnes dans l'image. Cependant, les artefacts sont bien plus importants dans ces masques de prédictions par rapport à ceux du modèle précédent : les zones prédites, ainsi que ses bords sont moins nets qu'auparavant, ce qui diminue la qualité des données extraites du tableau.

Les indicateurs de qualité sont calculés pour l'ensemble des images composant l'échantillon test. Nous obtenons les résultats présentés dans le tableau ci-dessous :

	Table		**Colonnes**	
	Indice de Jaccard	Taux de perte	Indice de Jaccard	Taux de perte
Base de test	0.38	0.22	0.30	0.30

En moyenne sur l'échantillon test, l'indice de Jaccard est éloigné de 1 pour la prédiction du tableau et de ses colonnes. Il est bien plus faible que pour le modèle précédent. Ce résultat montre donc que l'ajout d'images des comptes sociaux ne permet pas une meilleure estimation des emplacements du tableau et de ses colonnes.

3.4 Application du modèle aux comptes sociaux

Une fois ces masques obtenus, nous les appliquons à l'image dans le but d'extraire les informations contenues dans le tableau. Les masques provenant du modèle entraîné avec le jeu de données Marmot sont utilisés, car ces prédictions sont meilleures que celles avec le modèle d'images mixtes d'après les différents indicateurs (*cf* 3.4).

A partir de ces masques de prédiction, l'extraction des données du tableau s'effectue en différentes étapes :

1. Les masques de prédictions doivent être retraités, afin d'améliorer la qualité d'extraction des informations du tableau. Pour ce faire, les contours des zones prédites sont lissées, pour supprimer les artefacts liés à la prédiction autour de la zone notamment.
2. Pour chacun des emplacements prédits, leurs coordonnées dans l'image doit être récupérées, afin de lancer les traitements d'extraction dans chaque zone prédite
3. pour chacun des emplacements, un moteur de reconnaissance optique de caractère permet de récupérer l'ensemble des mots contenus de cette zone. Le moteur **tesseract** est utilisé ici.
4. L'ensemble des colonnes récupérées sont ensuite exportées dans un fichier plat. Pour éviter les erreurs d'alignement dans cet exportation, les coordonnées de chaque mot au sein du tableau sont

récupérées grâce à certaines fonctionalités du moteur **tesseract**, puis utilisées lors de l'exportation des colonnes.

Actuellement, cette partie est en cours d'analyse et de développement, ce qui explique l'absence d'indicateurs de qualité de l'extraction.

Toutefois, quelques limites ont été observées avec les premiers résultats de cette application. Certains indicateurs de qualité pourraient être implémentés à partir de ces limites :

- contrôle du nombre de lignes et de colonnes contenues dans le fichier plat d'extraction, en comparaison avec le nombre initial du tableau, pour vérifier le bon alignement des lignes et colonnes du tableau.
- contrôle de la qualité de l'océrisation de l'information, par le moteur **tesseract** : en effet, certains chiffres peuvent être mal retranscrits (0 à la place d'une virgule par exemple). Pour chaque caractère analysé, le moteur d'océrisation renvoie sa précision associée.
- contrôle du type de données pour chaque colonne : est-ce que la donnée est en pourcentage, ou en montant ? est-ce une variable numérique ou caractère ? pour obtenir une donnée cohérente pour une colonne/ligne donnée.

4 Possibilités d'utilisation de cette méthode dans la statistique structurelle d'entreprise

L'objectif premier de cette expérimentation est d'améliorer le profilage des groupes en réduisant le coût de récupération des tableaux des filiales et participations :

- accélération du travail des profileurs de récupération des tableaux d'intérêt au sein des comptes sociaux des UL étudiées pour les très grands groupes (entre 500 et 1000 comptes). Pour une unité légale donnée, si elle a publié ses comptes sociaux, le téléchargement du compte social ainsi que l'extraction de la page comportant le tableau recherché dure en moyenne 30 secondes, d'après les tests réalisés sur la plateforme SSPCloud. Un profileur réalisera le même travail en moyenne en plusieurs minutes.
- Récupération des informations de celles profilées automatiquement pour les petits groupes (plusieurs dizaines de milliers de comptes) et permettant d'améliorer l'estimation des flux intra-groupe.

Le tableau des filiales et participations permet aussi de récupérer les liens capitalistiques entre la société *mère*, celle qui détient des participations dans des sociétés, et celles *filles*, qui sont détenues par elle. Au sein de ce tableau, le lien entre la société mère et ses détentions sont récupérées, avec son taux de détention associé. Cependant, il ne s'agit uniquement d'un lien entre deux raisons sociales : un travail de *sirenisation* devra être effectué afin de retrouver les identifiants (siren) des sociétés.

De plus, d'autres informations contenues dans les comptes sociaux sont utiles pour le contrôle des données structurelles des entreprises :

- ventilation du chiffre d'affaires agrégé de la société
- évènements et activités exceptionnels de la société au cours de l'année
- différents cadres comptables (sous forme de tableaux) afin de comparer avec les données disponibles dans différentes bases de données
- explications de certaines évolutions spécifiques sur l'année comptable de certains postes (chiffre d'affaires, achats, total du bilan, etc.)

Enfin, d'autres applications sont possibles de ces modèles, en dehors des comptes sociaux. En effet,

ce processus pourrait être appliqué afin de pré-remplir certains questionnaires, en parcourant divers documents publiés par les entreprises. Le second modèle peut être aussi utilisé pour le traitements de la facturation des entreprises. Ce processus est aussi utilisable sur d'autres sujets que les statistiques d'entreprises, comme l'archivage de documents par exemple.

Bibliographie

[1] Anurag Bejju, Saeid Malladavoudi, Monica Pickard (2019) Automation of Information Extraction from Financial Statements using Graph-Based Techniques. Travaux pour StatCan.

[2] Karen Simonyan, Andrew Zisserman (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition, University of Oxford.

[3] Leo Breiman (2001) Random Forests. University of California, Berkeley.

[4] Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, Lovekesh Vig (2020) TableNet : Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images. TCS Research.

[5] Jonathan Long, Evan Shelhamer, Trevor Darrell (2015) Fully Convolutional Networks for Semantic Segmentation. Article, Berkeley.

[6] B. Gatos, I. Pratikakis, S.J. Perantonis (2005) Adaptive degraded document image binarization. Article, National Center for Scientific Research "Demokritos", Athens, Greece.

[7] Ray Smith (2007), An Overview of the Tesseract OCR Engine Article, Google Inc.

[8] O. Haag (2019) Le profilage à l'insee - Une identification plus pertinente des acteurs économiques, Courrier des statistiques N2.

[9] Yin Zhang, Rong Jin, Zhi-Hua Zhou (2010) Understanding Bag-of-Words Model : A Statistical Framework, International Journal of Machine Learning and Cybernetics.

[10] Zellig S. Harris (1954) Distributional Structure, WORD, 10 :2-3, 146-162.

[11] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., Fei-Fei, L. (2015) ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3) :211–252.

[12] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. (2009) ImageNet : A large-scale hierarchical image database. In Proc. CVPR

[13] Jaccard, Paul. (1901). Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines.. Bulletin de la Société Vaudoise des Sciences Naturelles. 37. 241-72.

[14] Efront, B. (1984), The Jackknife, the Bootstrap and Other Resampling Plans, 10, 75-91

[15] Yoav Freund, Robert E Schapire (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, Volume 55, Issue 1, 119-139,

[16] Bernhard Schölkopf, Alexander J. Smola (2002) Learning With Kernels : Support Vector Machines, Regularization, Optimization and Beyond, MIT Press, 187-401.