Sous la direction scientifique de
Benoit Dostie – Catherine Haeck
Sous la coordination de
Genevieve Dufour

Le Québec économique 10

Compétences et transformation du marché du travail

Chapitre 12

DÉTECTER LES COMPÉTENCES ÉMERGENTES EN UTILISANT L'INFORMATION CONTENUE DANS LES OFFRES D'EMPLOI

Luc Bissonnette
Charles Roy

CIRANO

Comment citer ce chapitre :

Bissonnette, L. et Roy, C. (2022). Détecter les compétences émergentes en utilisant l'information contenue dans les offres d'emploi : une courte introduction à l'utilisation de textes en économétrie. Dans B. Dostie et C. Haeck (dir.), *Le Québec économique 10. Compétences et transformation du marché du travail* (12, p. 259-278). CIRANO. doi.org/10.54932/WJMU9360

Chapitre 12

DÉTECTER LES COMPÉTENCES ÉMERGENTES EN UTILISANT L'INFORMATION CONTENUE DANS LES OFFRES D'EMPLOI

Une courte introduction à l'utilisation de textes en économétrie

Luc Bissonnette
Professeur agrégé au Département d'économique de l'Université Laval

Charles Roy
Professionnel de recherche au Département d'économique de l'Université Laval

Résumé

Avec les développements récents en science des données, les chercheurs de toutes disciplines ont maintenant accès à des outils d'analyse autrefois considérés comme très sophistiqués. Ce chapitre donne un exemple en illustrant comment des techniques d'analyse textuelle peuvent être utilisées afin de brosser le portrait des compétences demandées sur le marché du travail à l'aide du contenu d'offres d'emploi obtenues en ligne. La méthode proposée permet de traiter rapidement l'information contenue dans des centaines d'offres sans qu'un lecteur humain ait à lire chacune d'entre elles, et sans avoir de connaissances préalables sur le marché étudié. L'information obtenue est utilisée dans l'estimation d'un modèle économétrique visant à détecter les compétences émergentes. Un tel exercice permet d'obtenir le portrait d'un marché du travail spécifique, apportant un appui important aux décideurs afin d'assurer l'adéquation entre le marché du travail et les formations offertes dans les établissements d'enseignement. Certaines limites de l'approche proposées, notamment en ce qui a trait à la sélection dans l'échantillon, sont discutées, et des pistes de recherches futures sont proposées.

Introduction

Les nouveaux développements en science des données permettent de traiter une grande variété de types d'information, que ce soit des images, des vidéos ou du texte. Si ces percées ouvrent de nouvelles avenues de recherche aux économistes et analystes de politiques, elles demandent aussi d'adapter nos méthodes de traitement de données afin que ces données soient compatibles avec nos modèles empiriques. Dans l'édition précédente du *Québec économique,* Aubert, de Marcellis-Warin et Warin (2020) ont introduit plusieurs méthodes d'analyse textuelle pouvant enrichir l'analyse économique. Ce chapitre propose un exemple d'application de telles méthodes en illustrant comment les données textuelles peuvent être arrimées à nos modèles économétriques traditionnels. Nous ne présentons pas une revue exhaustive de toutes les possibilités offertes par ces nouveaux développements, mais présentons plutôt une analyse de cas réalisée à la demande du ministère de l'Éducation et de l'Enseignement supérieur et visant à mieux comprendre les besoins de formation pour un domaine d'expertise particulier : l'animation pour le cinéma et la télévision. Nous recommandons au lecteur intéressé à explorer davantage ces nouvelles techniques d'analyse de consulter le survol de l'état des connaissances proposé par Gentzkow et ses collaborateurs (2019), qui expose encore plus de possibilités d'analyse que ce que nous considérons ici[1].

Notre objectif est de comprendre les transformations vécues par un marché spécifique afin de fournir l'information nécessaire pour s'assurer de la pertinence des formations offertes au Québec. La vélocité des transformations du marché du travail fait en sorte que les indicateurs traditionnels comme le plus haut diplôme obtenu, le salaire et le nombre d'heures travaillées ne sont pas des mesures précises pour étudier l'évolution des tendances dans le marché du travail (Frank *et al.,* 2019). Notre profession d'intérêt demande peut-être le même niveau de formation (par exemple, un diplôme d'études collégiales) ou propose peut-être sensiblement les mêmes conditions de travail, mais les tâches et les compétences demandées se sont peut-être transformées au fil du temps. À grande échelle, ce phénomène est l'objet du chapitre 13 de cette édition du *Québec économique.* Les transformations des tâches et des compétences ne sont pas négligeables. Selon Acemoglu et Restrepo (2018 ; 2019), par exemple, 60 % des 50 millions d'emplois créés entre 1980 et 2015 aux États-Unis portaient de nouveaux titres, reflétant l'apparition de nouvelles

tâches. Selon plusieurs experts, les données dites « conventionnelles » ne permettent pas de représenter l'évolution très rapide du marché du travail (voir, par exemple, Frank *et al.,* 2019 ; Atalay *et al.,* 2020). Dans ce chapitre, nous suivons la recommandation de Frank et ses collaborateurs (2019) d'utiliser des données qui proviennent d'offres d'emploi pour analyser le marché du travail, car celles-ci nous informent des changements dans le marché du travail et des compétences recherchées par les employeurs, et ce, en temps réel. À titre d'exemple, Atalay et ses collaborateurs (2020) utilisent les offres d'emploi ainsi que l'analyse de texte pour montrer que le changement se produit au sein des professions plutôt que dans le taux d'emploi parmi les emplois routiniers et non routiniers. L'étude des offres d'emploi est propice à l'utilisation de techniques d'apprentissage automatique, puisque les offres contiennent beaucoup d'informations que l'on peut combiner avec des techniques d'intelligence artificielle afin d'étudier le marché du travail.

Dans l'application proposée ici, nous tentons de déterminer quelles sont les compétences émergentes ou en déclin pour les animateurs de cinéma à l'aide de telles données. L'un des mandats du ministère de l'Éducation et de l'Enseignement supérieur vise à s'assurer qu'il y a adéquation entre les formations offertes dans les différents établissements collégiaux de la province et les besoins en emploi. Pour ce faire, les analystes du Ministère évaluent les besoins pour différentes professions. Avec un marché du travail en évolution rapide, obtenir un portrait de ces différentes professions est de plus en plus difficile. Le nombre de professions à analyser est élevé et il est impossible pour le Ministère d'employer un effectif d'experts pour chacune d'entre elles. L'adéquation entre les besoins des employeurs du milieu de l'animation et les formations offertes par les établissements québécois est un enjeu majeur pour cette industrie en croissance. Il est apparu que ce secteur représentait un choix intéressant pour développer des méthodes visant à quantifier les besoins en ce qui touche aux compétences émergentes qui pourront être appliquées à d'autres secteurs. Cette procédure n'est pas un substitut parfait à une connaissance approfondie d'un secteur d'activité, mais l'approche proposée permet d'obtenir un survol d'un domaine d'expertise pour quelqu'un qui n'a aucune connaissance dans le domaine. Compte tenu des contraintes de temps des analystes, le bénéfice d'une telle approche n'est pas négligeable[2].

La méthode d'estimation proposée peut-être ramenée en termes économétriques standard et le modèle est assez simple. Nous tenterons de prédire l'année de publication d'une offre d'emploi en fonction du contenu de celle-ci. En analysant les mots qui prédisent une publication plus récente, nous obtiendrons une mesure des termes émergents. Notre variable dépendante est l'année de publication. Pour définir notre variable dépendante, il faut introduire la notion de « jeton ». Dans le cadre de ce chapitre, les jetons seront des mots (unigrammes) ou des locutions de deux mots (bigrammes). Nous créons donc des variables dichotomiques indiquant la présence ou non dans une offre donnée de certains jetons choisis. Notre objectif est d'obtenir une matrice de variables indépendantes X qui aura la forme suivante :

	supervision	lead	passionner	ambiance	artiste	équipe	projet	outil
0	1	1	0	0	0	0	0	1
1	0	0	0	1	0	1	0	0
2	0	0	1	0	1	1	1	0
3	0	0	0	0	1	1	0	0
4	1	0	0	0	1	0	1	0

Nous observons si le jeton choisi (colonne) est présent dans une annonce donnée (ligne). Une approche simple alternative aurait été de nous concentrer sur la fréquence de ces mots dans les documents, mais dans le cas qui nous préoccupe, nous souhaitions discuter de la présence ou de l'absence d'une compétence dans une offre, et non du nombre de fois que cette compétence était citée dans l'annonce. Cette hypothèse nous permet notamment de comparer des descriptions de longueurs variées ou qui ont des styles de rédaction très différents.

Il nous faut procéder à quelques manipulations avant de pouvoir estimer un tel modèle. Le reste de ce chapitre présentera ces différentes manipulations et proposera une analyse des résultats obtenus. Dans un premier temps, nous présenterons les données collectées : des offres d'emploi dans le domaine du cinéma d'animation. La section suivante présentera les étapes menant à la création de la matrice X et du vecteur y. L'un des défis que nous rencontrerons est la présence d'un grand nombre de variables explicatives ; nous présenterons donc le modèle de régression LASSO, qui permet de faire à la fois une sélection de variables et l'estimation du modèle, et nous

analyserons les résultats de cette régression. Ce modèle a aussi été utilisé par Nowak et Smith (2017) dans une analyse similaire visant à prédire le prix demandé pour une maison en fonction de la description qui en est faite. En conclusion, nous discuterons de quelques limitations de la méthode proposée et proposerons des avenues de recherche futures.

Collecte des données

Notre analyse repose sur une base de données collectée en ligne et contenant une collection d'offres d'emploi dans le domaine du cinéma d'animation. Ces données proviennent du site de VFX Montréal, spécialisé dans le domaine de l'animation. Les offres se concentrent sur l'industrie des régions de Montréal et de Québec. Nous avons utilisé la technique du moissonnage (*web scraping*), qui consiste à extraire l'information brute d'une page Internet et à la conserver sous forme de base de données. En récoltant l'information sur un domaine précis, nous connaissons la structure exacte du site. Ainsi, nous pouvons indiquer au programme d'aller récolter l'information recherchée à une adresse exacte. Le 2 février 2021, nous avons récupéré 456 offres d'emploi en français pour les périodes de 2014 à 2021[3]. L'échantillon ainsi obtenu n'est pas parfait, puisque les employeurs avaient la possibilité de retirer leur offre lorsque celle-ci était comblée. Une analyse rapide des offres du point de vue de la fréquence et du contenu nous pousse à croire que la plupart des employeurs ont simplement laissé les offres en ligne, puisqu'ils pouvaient obtenir plus de CV pour pourvoir des postes futurs en ne retirant pas leurs offres et qu'il n'y avait pas de coûts à laisser une offre en ligne. Notez qu'entre le moment de la collecte de données et le moment de rédiger ce texte, VFX Montréal a retiré les offres ayant plus de trois mois, puisque les employeurs avaient tendance à les laisser en ligne.

Pour chaque offre d'emploi, nous avons récolté sa description en format brut ainsi que sa date de publication. Notre base de données regroupe des offres d'emploi à plusieurs moments dans le temps et couvrant une liste d'employeurs importants ayant des studios à Montréal et à Québec (voir le tableau **12-1**).

Liste des employeurs contenus dans les données			
A.A. Studios	Alchemy 24	BUF Canada	DNEG
Digital Dimension	Digital Domain	FOLKS	Frima
Felix & Paul Studios	Hybride	L'Atelier Animation	MR. X
Mathematic	Mels	Mikros Image Canada inc.	ON Animation Studios
Oblique FX	Pixomondo	Real by Fake	Raynault VFX
Reel FX Creative Studios	Rodeo FX	SHED	Scanline VFX
Squeeze Studio Animation	Starno	TouTenKartoon Canada inc.	

Tableau t/2022-c**12-1**

À ce stade-ci, quelques remarques s'imposent sur la difficulté d'accès aux données et la représentativité de celles-ci. Dans le cadre des travaux présentés ici, nous désirions obtenir un portrait des compétences demandées dans le domaine de l'animation spécifiquement pour les marchés pertinents pour les établissements d'enseignement québécois. Dans ce cadre particulier, nous n'étions pas préoccupés par le fait que certaines offres pour des postes de niveau intermédiaire ou supérieur ne seraient peut-être pas affichées, une préoccupation qui aurait été centrale si nous avions voulu analyser l'offre de formation continue. Faute de trouver une banque de données parfaite, il nous a semblé que l'information trouvée sur le site de VFX Montréal répondait adéquatement à nos besoins, malgré les risques de sélection décrits plus haut. Il est aussi possible que certains employeurs renommés n'aient pas besoin d'afficher leurs postes sur des sites externes pour attirer de nouveaux candidats. Dans l'exemple que nous donnons ici, il nous a semblé que la sélection d'employeurs présentée au tableau **12-1** était assez diversifiée pour que nous procédions à l'analyse l'esprit tranquille.

Quelle aurait été la solution de rechange pour étudier une discipline pour laquelle nous n'aurions pas pu trouver un agrégateur unique comme le site de VFX Montréal ? Il aurait été possible de faire l'acquisition d'offres d'emploi pertinentes auprès d'un fournisseur de données dont l'information a déjà été extraite et normalisée. Dans le milieu de la recherche

universitaire, les données utilisées proviennent généralement de Burning Glass Technologies (par exemple, Hershbein et Kahn, 2018 ; Deming et Noray, 2020 ; Forsythe *et al.,* 2020) ou de CareerBuilder.com (par exemple, Marinescu et Rathelot, 2018). Ces données sont riches, viennent avec une interface facile à utiliser et les différentes informations ont déjà été extraites, ce qui permet d'obtenir rapidement un portrait d'intérêt du marché du travail. En contrepartie, ces produits ne sont pas toujours adaptés à la réalité du marché du travail québécois ou de domaines très précis, qui représentent des parts négligeables de leurs revenus. Burning Glass ne normalise pas les offres francophones et Career Builder est une compagnie américaine. Cette limitation est particulièrement importante dans le cas des formations collégiales, compte tenu de la particularité québécoise de ces diplômes.

Quelques manipulations

Notre objectif est d'utiliser l'information contenue dans ces textes dans un modèle économétrique traditionnel. Nous voulons en effet prédire l'année de publication à partir de la description d'une offre. Les mots liés à des compétences et prédisant une date de publication plus récente seront interprétés comme étant des compétences émergentes.

Toute l'information contenue dans les offres n'est pas pertinente. De plus, comme il y a beaucoup de mots utilisés dans une collection de documents, nous voudrons réduire le nombre de variables incluses dans l'analyse en ne conservant que les mots ayant un potentiel d'être informatifs. Nous effectuons les manipulations pour utiliser les mots des offres qui conviennent à notre contexte. Un survol des différentes étapes du processus de prédiction, détaillé dans le reste de ce chapitre, est représenté par la figure **12-1**.

Diagramme illustrant la marche à suivre pour réaliser l'analyse

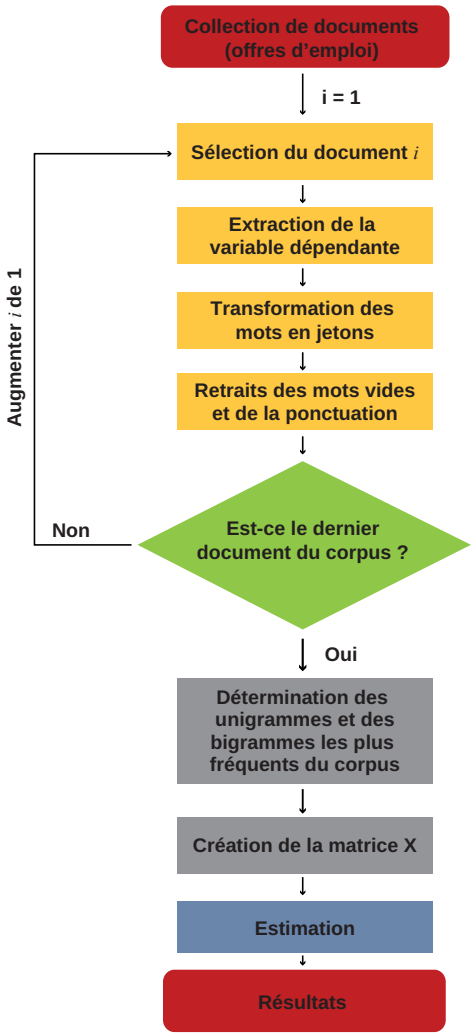

Figure g/2022-c**12-1**

Extraction de la variable dépendante

Commençons par la variable dépendante. Dans l'analyse présentée dans ce chapitre, nous considérons uniquement l'année de publication, extraite à partir des offres. La distribution de publication des offres d'emploi est présentée à la figure **12-2**.

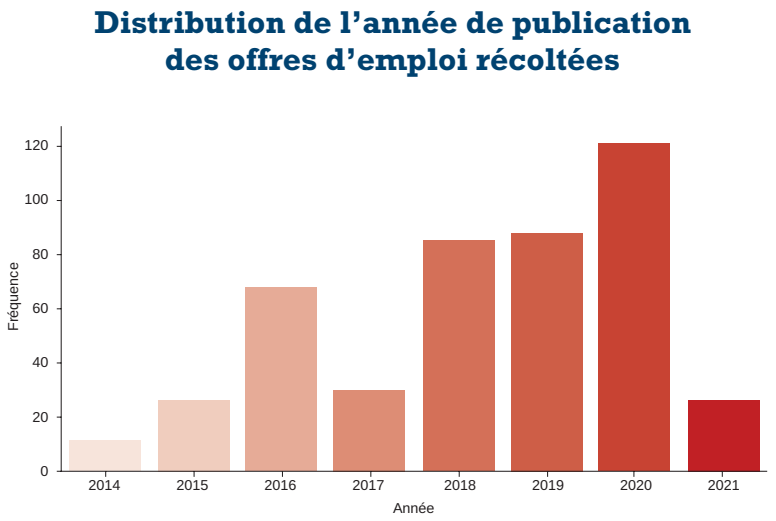

Figure g/2022-c**12-2**

Transformation des mots en jetons

La matrice X est formée avec l'information contenue dans la description des offres d'emploi. On y retrouve la description des tâches à effectuer, les qualifications requises tant professionnelles que relationnelles, la description de l'entreprise et d'autres informations pertinentes. Nous avons décidé d'inclure l'ensemble de ces sections dans notre analyse, ce qui nous permet de contrôler pour différents types de postes ou différents types d'entreprises. Cependant, la section Description des offres d'emploi que

nous avons récoltées est en format brut. Nous devons donc y appliquer des modifications afin de la transformer dans un format matriciel tel que présenté plus haut.

Premièrement, nous devons transformer les mots en jetons. Cette étape est essentielle afin de distinguer les différents mots du texte. La procédure est réalisée à l'aide de Spacy, une bibliothèque en source libre (*open source library*) de traitement automatique des langues (NLP), qui fonctionne avec Python. Nous voulons premièrement regrouper certains mots en un seul jeton. Par exemple, nous voulons repérer toutes les formes d'un même verbe en associant des mots comme « demandé » et « demanderont » sous la seule entrée « demander »; nous voulons également considérer « programmeur » et « programmeuse » comme étant plusieurs formes de la même profession. Nous utilisons pour ce faire des techniques de lemmatisation offertes dans la bibliothèque. Les lettres majuscules sont remplacées par des minuscules, les mots sont modifiés pour être du même nombre de lettres et, lorsque c'est possible, du même genre, et les verbes sont mis à l'infinitif.

Retrait des mots fréquents ou peu informatifs

Ensuite, nous voulons retirer les mots trop fréquents ou peu informatifs. Le modèle en français de Spacy comprend une liste des mots les plus utilisés, appelés mots vides (*stop words*), qui n'ont, *a priori,* aucun pouvoir prédictif dans un contexte d'analyse d'offres d'emploi. On y trouve, entre autres, des prépositions, des pronoms, des adverbes, des verbes fréquemment utilisés comme avoir, être, faire, etc. La deuxième étape du processus est d'uniformiser le texte en retirant les mots vides et la ponctuation. Il reste néanmoins beaucoup de jetons qui pourraient être inclus dans le texte. Nous ajoutons donc quelques critères pour choisir parmi les jetons disponibles qui seront utilisés dans le modèle. Dans un premier temps, nous excluons les mots qui sont encore trop fréquents malgré le retrait des mots vides. Nous nous concentrons sur les jetons qui sont présents au plus dans 80 % de l'échantillon. Les mots qui sont présents à plus haute fréquence peuvent être traités comme des mots vides propres à notre sujet. Parmi ces jetons, nous sélectionnons les 200 jetons les plus fréquents, compte tenu de la taille modeste de notre échantillon.

Analyse des jetons retenus

Le tableau **12-2** illustre quelques jetons représentant des compétences parmi les 200 jetons conservés. On peut y lire les mots ainsi que leur fréquence entre parenthèses. La fréquence est déterminée d'après le nombre d'offres où le mot apparaît au moins une fois. Ce sont les variables qui seront particulièrement intéressantes à surveiller lorsque notre modèle sera estimé.

Quelques jetons retenus			
Compétences relationnelles			
anglais (110)	artiste (313)	artistique (206)	communication (205)
communiquer (155)	créatif (179)	créativité (89)	français (122)
travailler équipe (99)			
Compétences techniques			
cg (127)	compositing (86)	connaissance logiciel (98)	éclairage (116)
informatique (76)	linux (74)	logiciel (250)	logiciel maya (76)
maya (244)	nuke (116)	pipeline (177)	python (131)
résoudre problème (89)	rigging (83)	shotgun (119)	vfx (81)

Tableau t/2022-c**12-2**

Que nous révèle le tableau **12-2**? Nous remarquons deux thèmes principaux dans la partie sur les compétences relationnelles : la communication et la créativité. Ces résultats nous indiquent que, sans tenir compte de l'année de publication, les employeurs cherchent des employés avec ces critères. Sur un total de 456 offres d'emploi récoltées, le mot *artiste* est présent au moins une fois dans 68,64 % des annonces et le mot *communication* est présent au moins une fois dans 44,95 % des annonces. En ce qui concerne les compétences techniques demandées dans les offres d'emploi, on trouve des logiciels tels que Nuke, Shotgun et Maya, ainsi que des compétences techniques générales telles que « Linux », « *Computer Graphics* » (« *CG* »), « *rigging* », « *compositing* », « *pipeline* », « Python », « vfx », « informatique » et « éclairage ». Le signe du coefficient associé à ces compétences pourra nous indiquer si cette compétence cherche à disparaître ou si elle est émergente.

Pour simplifier la présentation, nous n'avons pas rapporté ici tous les mots qui sont inclus dans l'analyse. Parmi ces mots, certains contrôlent pour le type de poste ou le niveau d'expérience demandé. Par exemple, les jetons tels que « *lead* », « gestion », « gestionnaire », « producteur », « directeur », « superviseur », « réalisateur » et « chef » nous indiquent qui est concerné par ces postes. Le travailleur concerné par des postes de gestion n'est pas à sa première année sur le marché du travail et doit posséder un grand nombre de compétences, différentes de celles que possède une personne à sa sortie de l'école. De plus, nous pouvons faire des liens avec les interprétations précédemment énumérées. Par exemple, les postes de gestion requièrent probablement une maîtrise de la plupart des compétences relationnelles discutées.

Création de la matrice X

Finalement, après avoir déterminé la liste des 200 éléments fréquents, nous pouvons former la matrice où les colonnes correspondent aux unigrammes et aux bigrammes et les lignes correspondent aux mots présents dans les offres d'emploi. Il s'agit simplement de déterminer si un jeton est présent ou non dans l'offre en question. Lors de l'estimation, nous ajouterons une colonne de 1 à cette matrice qui permettra d'introduire une constante dans notre estimation. Notez que s'il nous a fallu quelques pages pour vous expliquer notre procédure, celle-ci est facilement applicable à l'aide de quelques lignes de code. Cette relative simplicité implique que la partie la plus difficile d'une telle analyse est d'obtenir les données, non pas de les manipuler.

Économétrie et analyse textuelle

Le modèle économétrique

Lorsque l'information a été ramenée sous forme de matrice, l'analyste peut utiliser celles-ci dans un modèle économétrique standard ou dans un modèle d'analyse de données tel qu'un arbre de classification ou un réseau de neurones. Dans notre illustration, nous tentons de prédire l'année de publication d'une offre d'emploi compte tenu du contenu de celle-ci. Dans le cadre de ce chapitre, nous avons décidé d'illustrer la méthode d'analyse à l'aide d'une régression pénalisée de type LASSO (pour *Least Absolute*

Shrinkage and Selection Operator, voir Tibshirani, 1996), qui semblait appropriée compte tenu du nombre d'observations relativement modeste contenues dans notre base de données. L'estimateur obtenu par la méthode de régression LASSO est donné par l'expression suivante :

$$\arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

où *i* réfère à une offre d'emploi et *j,* à un élément du vecteur de paramètre β. La première partie de l'équation est identique à celle utilisée dans l'estimation de modèle linéaire par moindres carrés ordinaires. La seconde partie est une pénalité additionnelle qui s'appliquera à tout paramètre dont la valeur n'est pas 0. Ainsi, une variable se verra assigner une valeur de paramètre non nulle uniquement si son inclusion dans le modèle de régression mène à une diminution suffisante de la somme du carré des résidus. L'évaluation du modèle mène à un vecteur de paramètre estimé contenant beaucoup de zéros. Le LASSO permet de sélectionner les variables pertinentes à la prédiction et d'effectuer l'estimation du modèle simultanément. Le coût à payer pour une telle sélection est l'ajout d'un biais dans l'estimation en échantillon fini. L'ajout du terme pénalisant, qui est proportionnel à la somme des valeurs absolues des paramètres, permet d'être efficace quant au temps de calcul (Athey et Imbens, 2019) et à la sélection de variables. Le paramètre de pénalité λ doit être calibré. Il existe plusieurs techniques pour déterminer la valeur de ce paramètre. Nous utilisons une procédure par validation croisée pour obtenir cette valeur (pour une explication détaillée, voir par exemple Varian, 2014). Cette procédure est disponible directement dans la fonction de la bibliothèque *SciKit Learn,* utilisée pour l'estimation présentée dans ce chapitre.

L'interprétation du modèle est la même que celle d'une régression linéaire standard. Pour des petites valeurs, le paramètre associé à une variable s'interprétera donc comme la variation attendue de la variable dépendante lorsqu'un mot est présent dans une annonce par rapport à ce qui sera espéré pour une offre ne contenant pas ce mot, toutes choses étant égales par ailleurs.

La question de l'inférence statistique à la suite de l'estimation d'un modèle LASSO est complexe et dépasse le cadre de ce chapitre. Nous n'y parlerons donc pas significativité ou inférence. À notre connaissance, la plupart des méthodes connues supposent que le chercheur s'intéresse

à un sous-ensemble de caractéristiques (par exemple un effet traitement dans le cas de Belloni *et al.,* 2014) ou à l'inférence, et ce, en supposant que le modèle trouvé soit le bon (par exemple Tibshirani *et al.,* 2016). Plusieurs de ces problèmes surviennent puisque le LASSO permet d'inclure un grand nombre de variables explicatives, possiblement plus grand que le nombre d'observations, une propriété utile pour appliquer la méthode proposée ici à des professions plus rares. Notons néanmoins que dans l'application proposée dans ce chapitre, nous avons retenu un nombre de jetons (200) inférieur au nombre d'observations (456), de sorte qu'il aurait été possible d'estimer le modèle par moindres carrés ordinaires. Nous ne présenterons pas ces résultats ici, et nous invitons le lecteur intéressé à consulter le mémoire à l'origine de cet article pour en apprendre plus sur cette comparaison (voir Roy, 2021). Nous soulignons tout de même que les résultats ne sont pas fondamentalement différents lorsque cette méthode alternative est utilisée. L'approche par LASSO simplifie néanmoins l'analyse des données en diminuant le nombre de jetons à analyser. Comme nous l'avons mentionné plus tôt, cette simplification pour l'interprétation des résultats représente un gain de productivité notable pour les analystes du ministère de l'Éducation et de l'Enseignement supérieur, qui doivent étudier des centaines de programmes et de professions. De plus, les méthodes de type LASSO permettent de réduire le surapprentissage, ce phénomène qui se produit lorsqu'un modèle parvient à expliquer presque parfaitement les données étudiées, mais qu'il possède un faible pouvoir prédictif hors de l'échantillon.

Notez finalement que la relative simplicité du modèle a été choisie pour sa facilité d'exposition et la connaissance commune des régressions linéaires. L'analyste voulant appliquer des méthodes économétriques plus poussées ou des méthodes d'apprentissage automatique modernes pourrait facilement le faire à partir des matrices et vecteurs créés à l'aide de la procédure décrite plus haut. Par exemple, plusieurs méthodes présentées dans l'édition précédente du *Québec économique* par Stevanovic (2020) s'appliquent aussi dans le cadre microéconométrique présenté ici.

Résultats

Nous présentons, dans cette section, une synthèse des résultats obtenus dans notre analyse économétrique. Lors de l'estimation, 83 des 200 jetons choisis se sont vu attribuer des paramètres différents de 0. Le tableau **12-3**

présente quelques résultats choisis pertinents pour notre estimation. Nous ne rapportons ici que les résultats relatifs aux compétences choisies dans le tableau **12-1** et certains jetons qui nous ont paru intéressants, omettant ainsi plusieurs coefficients non nuls.

Quelques résultats choisis		
Jeton	**Coefficient**	**Pourcentage contenant le jeton**
Compétences techniques		
Computer Graphic (CG)	-0,263	28 %
Compositing	-0,369	19 %
Éclairage	-0,279	25 %
Logiciel Maya	0,076	17 %
Nuke	-0,112	25 %
Compétences relationnelles		
Anglais	-0,206	24 %
Créatif	0,110	39 %
Travailler équipe	-0,186	21 %
Autres jetons dignes de mention		
Diplôme	0,024	42 %
Masculin	0,232	22 %

Tableau t/2022-c**12-3**

Nos résultats montrent que le logiciel Nuke est présent dans les offres les moins récentes. Si nous regardons deux offres qui contiennent de l'information semblable, dont une contient le mot *Nuke* et l'autre non, alors nous nous attendons à ce que l'offre qui ne contient pas *Nuke* soit de 0,112 an plus récente. Ainsi, l'interprétation est la même pour le jeton « logiciel Maya ». Une offre qui contient « logiciel Maya » comme compétence requise est de 0,076 an plus récente qu'une offre ne contenant pas « logiciel Maya ». Aussi, les résultats nous indiquent que le bigramme *computer graphics* (*CG*) et le mot *compositing* sont moins demandés dans les offres d'emploi récentes, avec des coefficients de -0,263 et -0,369 respectivement. Sur le plan des compétences relationnelles, le mot *anglais* et le bigramme *travailler équipe* ne sont plus présents dans les offres récentes, avec un

coefficient de -0,206 et de -0,186 respectivement. Le mot *créatif* est le seul jeton à coefficient positif qui se retrouve dans la liste des compétences relationnelles précédemment établie, avec un coefficient de 0,110.

Bien que nous ne rapportions pas ici tous les jetons s'étant vu assigner des valeurs non nulles, certains nous ont semblé dignes de mention. Nous remarquons que le modèle estime que le jeton « diplôme » prédit des offres récentes, en faisant un jeton émergent selon notre définition. Les quelque 189 annonces où se trouve le jeton indiquent que les diplômes d'études demandés sont autant collégiaux qu'universitaires. L'analyse révèle que les employeurs ne demandent pas de diplôme en particulier, mais demandent plutôt la possession d'un diplôme dans un domaine artistique ou technique. Il faut noter qu'une variété de programmes de plus en plus importante est offerte dans les établissements d'enseignement collégiaux. De plus, les offres d'emploi qui ne contiennent pas la mention du mot *diplôme* demandent généralement des années d'expérience dans le domaine. L'importance du terme *diplôme* est donc liée à la présence de postes de niveau débutant.

Nous remarquons aussi le jeton « masculin », qui prédit des offres plus récentes. Les conclusions de l'analyse postestimation révèlent que le jeton est toujours employé pour indiquer que les offres sont écrites en utilisant le genre masculin pour alléger le texte. Le coefficient positif relié au jeton dans tous les tableaux indique que les employeurs sont plus inclusifs dans la rédaction des offres. Les offres qui ne contiennent pas le jeton sont majoritairement écrites au masculin, sans aucune mention de l'utilisation du genre. L'émergence de ce mot semble paradoxalement liée à une plus grande sensibilité à la question de la diversité et de l'inclusion. Cet exemple nous rappelle que l'usage de méthodes d'apprentissage statistique ne peut se substituer à la connaissance des données par un humain, puisqu'une interprétation naïve de ces résultats aurait mené à une interprétation erronée de la réalité. Même aujourd'hui, l'économètre ne peut confier tout son travail à la machine.

Quelques limites importantes

Comme dans toute analyse, il est important de garder en tête quelques limites de notre analyse statistique. La première, que nous avons déjà abordée, est que nos résultats sont aussi représentatifs que l'échantillon choisi. Ainsi, si certains types d'emploi ne sont pas affichés sur le site sur

lequel nous avons collecté l'information (par exemple, si le recrutement se fait directement dans les cégeps et les universités ou par contact), notre portrait du marché du travail sera imparfait. De même, il est probable que certains postes destinés à l'interne ne soient pas rendus publics, ce qui pourrait poser problème pour l'analyse d'un marché et masquer l'intérêt pour certaines formations continues destinées à des travailleurs plus expérimentés. Il est aussi possible que la popularité de certaines compétences soit dictée par quelques projets très importants qui ont eu lieu dans les dernières années et ne soit pas un bon prédicteur des années à venir. Finalement, si deux compétences sont pratiquement toujours présentes dans la même offre, le modèle LASSO ne gardera probablement que l'une d'entre elles comme prédicteur. En somme, le modèle LASSO pourra donner des résultats mitigés s'il y a présence de colinéarité, même imparfaite. Il est donc sage de valider d'abord la corrélation entre les variables incluses, quitte à regrouper certains jetons sous une même variable. Il importe ainsi de bien comprendre le domaine d'expertise pour former les étudiants dans ces compétences complémentaires. Comme nous l'avons écrit en introduction, nous proposons ici une méthode permettant au moins à un analyste de se familiariser avec un domaine d'expertise, mais cette analyse sommaire a pour objectif d'amorcer un dialogue avec les acteurs du milieu pour comprendre leurs besoins et non pas de se substituer à cette expertise.

Conclusion

Ce chapitre à vocation méthodologique présente les différentes étapes de la manipulation permettant d'utiliser des données textuelles dans le cadre d'un modèle économétrique traditionnel. La méthode proposée nous a permis d'isoler quelques tendances dans les offres d'emploi concernant un domaine pour lequel nous n'avions aucune connaissance *a priori* : celui des animateurs pour le cinéma et la télévision. Malgré une taille d'échantillon modeste, nous avons été en mesure de repérer quelques compétences techniques et relationnelles qui semblent être plus recherchées. Nous avons été frappés par la relative simplicité d'appliquer des méthodes qui, il n'y a pas si longtemps, auraient largement dépassé nos possibilités techniques.

Le modèle proposé est relativement simple, mais les possibilités sont beaucoup plus riches. Dans le travail présenté ici, nous avons utilisé tous les mots sans égard à leurs catégories. Il est possible et tout aussi simple de créer des modèles prédisant si un mot est une compétence à l'aide

de modèles de reconnaissance d'entités. Nous aurions aussi pu utiliser un modèle qui représente le texte de manière vectorielle afin de préserver le sens des mots dans l'analyse (voir, par exemple, des algorithmes de type Word2Vec). Cela nous aurait notamment permis de réduire le nombre de jetons utilisés dans l'analyse.

Si l'exemple présenté dans ce chapitre est bien aligné avec le thème de cette édition du *Québec économique* sur les compétences et le marché du travail, nous sommes convaincus que les méthodes présentées ici peuvent être utiles au-delà de ce domaine d'expertise. Pour un analyste disposant d'une collection de textes sur son sujet d'intérêt, nous croyons que les méthodes proposées ouvriront la porte à un large éventail de recherche future, pour peu que ce chercheur ait lui-même pris le temps de développer quelques compétences de base en programmation.

Références

Acemoglu, D. et Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares and employment. *American Economic Review, 108*(6), 1488-1542. https://doi:10.1257/AER.20160696

Acemoglu, D. et Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3-30. https://doi:10.1257/jep.33.2.3

Atalay, E., Phongthiengtham, P., Sotelo, S. et Tannenbaum, D. (2020). The evolution of work in the United States. *American Economic Journal: Applied Economics, 12*(2), 1-34. https://doi:10.1257/app.20190070

Athey, S. et Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*(1), 685-725. https://doi:10.1146/annurev-economics-080217-053433

Aubert, B., de Marcellis-Warin, N. et Warin, T. (2020). Science des données, réseaux sociaux et politiques publiques (chap. 11, p. 285-314), dans de Marcellis-Warin, N., Dostie. B. et Dufour, G. (dir.), *Le Québec économique 9,* Montréal, QC : CIRANO.

Belloni, A., Chernozhukov, V. et Hansen, C. (2014), High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives, 28*(2), 29-50. https://doi:10.1257/jep.28.2.29

Deming, D. J. et Noray, K. (2020). Earnings dynamics, changing job skills, and STEM careers. *Quarterly Journal of Economics, 135*(4), 1965-2005. https://doi:10.1093/qje/qjaa021

Forsythe, E., Kahn, L. B., Lange, F. et Wiczer, D. (2020). Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims. *Journal of Public Economics, 189,* 104238. https://doi:10.1016/j.jpubeco.2020.104238

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H. et Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences of the United States of America, 116*(14), 6531-6539. https://doi:10.1073/pnas.1900949116

Gentzkow, M., Kelly, B. T. et Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535-574. https://doi:10.1257/jel.20181020

Hershbein, B. et Kahn, L. B. (2018). Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *American Economic Review, 108*(7), 1737-1772. https://doi:10.1257/aer.20161570

Marinescu, I. et Rathelot, R. (2018). Mismatch unemployment and the geography of job search. *American Economic Journal: Macroeconomics, 10*(3), 42-70. https://doi:10.1257/mac.20160312

Nowak, A. et Smith, P. (2017). Textual analysis in real estate. *Journal of Applied Econometrics, 32*(4), 896-918. https://doi:10.1002/jae.2550

Roy, C. (2021). Prédiction des compétences émergentes par analyse textuelle (mémoire de maîtrise). Université Laval, Québec.

Stevanovic, D. (2020). Prévision macroéconomique dans l'ère des données massives et de l'apprentissage automatique (chap. 12, p. 315-351), dans de Marcellis-Warin, N., Dostie. B. et Dufour, G. (dir.), *Le Québec économique 9,* Montréal, QC : CIRANO.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. https://doi:10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. J., Taylor J., Lockhart R. et Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association, 111*(514), 600-620. https://doi:10.1080/01621459.2015.1108848

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28. https://doi.org:10.1257/jep.28.2.3

Notes

1. Nous sommes reconnaissants au ministère de l'Éducation et de l'Enseignement supérieur pour le financement de ce projet. Ce chapitre présente les résultats d'une analyse réalisée dans le cadre d'un rapport commandé par le Ministère et ayant fait l'objet du mémoire de Charles Roy, l'un des deux auteurs. Le lecteur intéressé à consulter les résultats complets de l'analyse peut consulter le mémoire à l'adresse suivante : corpus.ulaval.ca/jspui/handle/20.500.11794/69195.
2. À la demande du ministère de l'Éducation et de l'Enseignement supérieur, d'autres études similaires ont été réalisées dans le cadre des compétences demandées pour la profession de programmeurs et pour les professions traitant de la cybersécurité. Tous ces domaines sont axés sur l'usage intensif de l'ordinateur et sont en évolution constante et très rapide.

3. Nous nous sommes assurés de ne pas représenter un fardeau pour le serveur sur lequel nous avons fait la collecte de données, notamment en prenant une pause de quelques secondes entre chaque requête. Nous avons suivi les instructions laissées par le gestionnaire du site Web dans le fichier nommé robots.txt, qui prescrit les règles de collectes de données sur le domaine. Il est important de rappeler que les règles d'utilisation de plusieurs sites Internet ne permettent pas la collecte de données par de telles méthodes.