Description des stages longs

1 Stage à la DSDS de l'INSEE, 11/06/18 - 31/08/18

1.1 Contexte

Ce stage a été réalisé à la section Revenu des Ménages de la Direction des Statistiques Démographiques et Sociales de l'INSEE. Ce stage était l'occasion d'acquérir une expérience de recherche en sociologie quantitative. La thématique de recherche a été proposé par Louis-André Vallet, et la supervision du stage a été réalisée conjointement par Céline Goffette (CREST) et Jérôme Accardo (INSEE).

1.2 Problème de recherche

Pour un certain nombre de pays et particulièrement dans le cas de la France, les conclusions des différentes littératures sur la transmission du statut socio-économique ne sont donc pas convergentes : là où les sociologues mettent en évidence une hausse de la fluidité sociale, i.e. un affaiblissement du lien entre origine et position sociales, les économistes mettent au contraire en évidence une transmission substantielle de la capacité à générer des revenus qui semble se perpétuer au fil des générations. L'objectif de cette étude est de proposer une analyse intermédiaire entre les approches sociologique et économique de la mobilité en étudiant dans quelle mesure la capacité à générer des revenus se transmet au fil des générations selon l'origine sociale des individus.

1.3 Méthodologie

Cette étude a pour objet la comparaison des distributions de revenus conditionnelles à l'origine sociale. Nous nous inscrivons dans la lignée du cadre économique et de la procédure statistique proposés par Lefranc, Pistolesi et Trannoy (2004) pour analyser l'évolution de l'inégalité des chances entre 1979 et 2000. Les auteurs proposent de définir l'inégalité des chances en recourant à une expérience de pensée : supposons que les individus ont la possibilité de choisir leur milieu social d'origine. L'inégalité des chances prévaut dès lors qu'un individu rationnel préfère toujours une distribution de revenus à une autre. Et comme le choix hypothétique de l'individu s'apparente à un choix risqué – on compare des distributions de probabilités sur les différents niveaux de revenu – la comparaison des distributions se fait à partir d'un critère de dominance stochastique.

1.4 Données

Nous appliquons ce cadre statistique à la série des enquêtes Revenus Fiscaux (ERF) et Revenus Fiscaux et Sociaux (ERFS) de l'Insee. Ces enquêtes couvrent la période 1996-2015, et permettent en cela de déterminer si la réduction de l'inégalité des chances mise en évidence par les auteurs entre 1979 et 2000 s'observe également au cours des deux dernières décennies.

1.5 Résultats

Nous présentons tout d'abord des statistiques descriptives qui mettent en évidence une inégalité des chances substantielle au niveau statique. Puis nous exposons les résultats issus de l'application de la procédure de Lefranc et al aux ERFS, qui met en lumière une stabilité de l'inégalité des chances entre 1996 et 2015, contrastant avec la forte réduction de l'inégalité à laquelle concluent les auteurs pour la période 1979-2000.

2 Stage à Orange, 11/03/19 - 24/05/19

2.1 Contexte

Ce stage a été ralisé au laboratoire SENSE d'Orange Labs. Il constitue le prolongement de mon stage d'application réalisé conjointement à l'INSEE et à Orange Labs. La raison de ce prolongement a été la possibilité de travailler sur des données mobiles très récentes, permettant d'envisager des estimations de population présente de haute précision spatio-temporelle. Comme le stage d'application, il a été réalisé conjointement sous la supervision de Benjamin Sakarovitch, *data scientist* au SSP-Lab de l'INSEE, et de Zbigniew Smoreda, sociologue au laboratoire SENSE d'Orange.

2.2 Problème de recherche

Le problème de recherche est identique à celui du stage d'application : la localisation géographique des évènements. Cependant, contrairement à l'étude effectuée lors du stage précédent, l'enjeu est ici de s'affranchir complètement de la modélisation par polygônes de Voronoï dans le cadre du *mapping* spatial des évènements. Cette extension est rendue possible par la disponibilité de données précises sur la couverture théorique des antennes.

2.3 Méthodologie

Nous mobilisons à nouveau le modèle de localisation des évènements mobiles proposé par Tennekes (2018) :

$$\mathbb{P}(carreau_i|antenne_j) \propto \mathbb{P}(carreau_i)\mathbb{P}(antenne_j|carreau_i)$$

Nous modifions cependant le terme de vraisemblance $\mathbb{P}(antenne_j|carreau_i)$, qui correspond de fait à la modélisation de la couverture des antennes, afin de tenir compte de l'information technique additionnelle dont nous disposons sur la couverture théorique des antennes. Cette modification témoigne d'une force majeur du modèle bayésien de localisation : il permet de couvrir un large éventail de situations en termes de données disponibles, aussi bien sur la couverture des antennes que pour incorporer de l'information *a priori* sur la grille.

2.4 Données

Nous disposons d'un jeu de données de *signalling* (i.e. non plus seulement des données actives – appels et SMS – mais également passives – connexions fréquentes du mobile aux antennes sans action de l'usager) couvrant l'intégralité des clients d'Orange. Ces données présentent une profondeur temporelle largement plus importante que celles utilisées lors du stage d'application. Les cartes de couverture théorique des antennes sont également fournies par Orange.

2.5 Résultats

La courte durée du stage n'a pas permis de produire de résultats finaux concernant la population présente. De nombreux problèmes computationnels se sont posés : du fait de leur profondeur temporelle considérable, les données de *signalling* utilisées représentent un volume d'environ 1To par jour, ce qui rend très délicat leur exploitation, même en utilisant des outils de calcul distribué adaptés (Spark sur une infrastructure HDFS). Le stage a cependant permis de décrire statistiquement ces données – qui n'ont jamais été exploitées auparavant – ainsi que de coder les scripts informatiques nécessaires à de futures estimations de population présente.

3 Stage à Foundamental (Berlin), 11/06/19 - 06/09/19

3.1 Contexte

Ce stage a été réalisé à Foundamental (Berlin), un fonds de *venture capital* qui investit dans des *startups* innovantes dans le domaine de la construction. Il a été supervisé par Clemens Meyer zu Rheda, qui est *data scientist* à Foundamental. Ce stage a été l'occasion pour moi de réaliser d'une part une expérience dans une équipe internationale, et d'autre part de découvrir un milieu professionnel très différent de ceux de mes précédentes expériences (statistique publique ou laboratoires de recherche).

3.2 Problème de recherche

L'équipe chargée de la *data science* vise à développer des processus permettant de simplifier et de rendre plus efficient le travail des investisseurs, à toutes les étapes du processus de décision de l'entreprise. Dans ce cadre, le principal projet sur lequel j'ai travaillé consistait à détecter la similarité entre *startups* sur la seule base de leur description. L'objectif de ce projet est de pouvoir détecter les potentiels concurrents d'une entreprise dans laquelle Foundamental investit, ce qui est un facteur essentiel dans la décision d'investissement.

3.3 Méthodologie

Comme indiqué précédemment, les seules données disponibles pour procéder à la détection de similarité étaient des descriptions plus ou moins détaillées de *startups*. Le problème se ramenait donc à une application classique de *Natural Language Processing* (NLP) visant à détecter la similarité sémantique entre des textes de longueur inégale par le biais d'outils statistiques. Une des questions majeures a été de trouver le meilleur *embedding*, i.e. la meilleure manière de représenter vectoriellement le texte afin de procéder à leur comparaison. De nombreux *embeddings* ont été implémentés, des plus classiques (comptage binaire, TF-IDF, LDA) aux plus récents (word2vec, doc2vec, réseaux pré-entrainés). Une autre question importante a été de trouver la métrique appropriée pour comparer les textes une fois représentés vectoriellement. Là encore, plusieurs métriques ont été implémentées (similarisé cosine, *clustering* de textes, *word's mover distance*).

3.4 Données

Les descriptions des *startups* provenaient de plusieurs sources : principelement de bases de données commerciales (Crunchbase, Tracxn), mais certaines étaient également issues des

recherches des investisseurs.

3.5 Résultats

Les résultats de la recherche par similarité étaient mitigés. En moyenne, les meilleurs algorithmes permettaient de détecter des entreprises candidates de bonne qualité, mais le nombre de faux-positifs s'est avéré important. La principale raison de ces performances contrastées est l'importante hétérogénéité de la qualité des descriptions, qui ne permet pas de détecter systématiquement des entreprises pertinentes. Face à ce constat, j'ai proposé et implémenté une application dédiée à la détection de similarité, qui permet de visualiser rapidement à la fois la description initiale et les résultats, de modifier à la volée les paramètres importants des modèles, et donc de déterminer rapidement si l'algorithme de détection de la similarité peut s'avérer pertinent pour l'entreprise considérée.