

Etienne LE NAOUR

ENSAE 3ème année

Stage de fin d'études

Année Scolaire 2019 - 2020

Analyse statistique des transcriptions des réunions du FOMC (Federal Open Market Committee)

**CREST
Palaiseau**

**Maître de stage : Alessandro
RIBONI**

Du 07/07/2020 au 09/10/2020

Table des matières

1	Introduction	2
1.1	Problématique	2
1.1.1	Contexte	2
1.1.2	Notre approche et nos objectifs	4
1.2	Revue de la littérature	4
1.3	Notes de lecture et précisions techniques	6
2	Construction de la base de données et premières statistiques descriptives	7
2.1	Web Scraping	7
2.2	Mise en forme de la base de données	7
2.2.1	Construction des dictionnaires pour chaque réunion	9
2.2.2	Construction de la base de données retraçant la temporalité	9
2.3	Statistique descriptive	10
2.3.1	Statistique sur la taille des réunions et la présence des chairs	11
2.3.2	Statistiques de temporalités	13
3	Analyse de sentiments	16
3.1	Présentation des dictionnaires de sentiments et construction des scores	16
3.1.1	Dictionnaire de Loughran and McDonald	16
3.1.2	Dictionnaire d'Harvard	17
3.2	Analyse de sentiments associés à chaque réunion	18
3.3	Analyse de sentiment relative aux chairs	19
4	Analyse du pouvoir relatif du chair sur les autres membres de la réunion	20
4.1	Régression avec fixed effect	21
4.1.1	Sur le score de positivité	21
4.1.2	Score d'incertitude	22
5	Topics modelling avec l'aide de l'algorithme Latent Dirichlet Analysis	24
5.1	Présentation de LDA	24
5.2	Application de la méthode LDA à notre corpus	26
6	Conclusion	29
6.1	Conclusion du travail de recherche	29
6.2	Conclusion personnelle	30
7	Annexes	31
7.1	Aperçu la temporalité des réunions de Burns, Miller et Volcker	31
7.2	Graphiques score de sentiments	32
7.3	Régression fixed effect avec pour variable de contrôle la valeur du nasdaq au jour de la réunion	33
7.4	Explication de bag of words	33

1 Introduction

Ce stage a été effectué au Centre de Recherche en Économie et Statistiques (CREST) du 07/07/2020 au 09/10/2020 sous la supervision du professeur Alessandro Riboni. Il s'inscrit dans le cadre du stage de fin d'études de troisième année de l'ENSAE et compte également pour le Master data science.

Le but de ce stage est d'appliquer mes connaissances en machine learning (notamment en traitement naturel du langage) à un corpus de texte économique afin d'y extraire des statistiques susceptibles d'alimenter les travaux de recherche de mon maître de stage. Je tiens particulièrement à remercier le Professeur Alessandro Riboni pour son accueil et son accompagnement dans le projet de recherche qu'il m'a proposé. Il a été très disponible et m'a apporté son expertise en économie et notamment en macroéconomie.

1.1 Problématique

1.1.1 Contexte

Le Federal Open Market Committee (FOMC) désigne un organe de la réserve fédérale américaine (FED). Cet organe est en charge des opérations d'achat/vente des bons du trésor américain (opérations d'open market) qui constituent la principale marche de manœuvre de la politique monétaire des États-Unis. De ce fait, le FOMC est responsable de la fixation du taux d'intérêt directeur (taux que les banques commerciales pratiquent entre elles pour les prêts). En effet, après les réunions du FOMC, le directeur du FOMC annonce un taux d'intérêt directeur cible pour lequel la FED va réaliser des opérations d'open market, dans le but de l'atteindre. De plus, le FOMC est aussi responsable des opérations réalisées par la FED sur le marché des changes. L'objectif de la FED est d'assurer la stabilité des prix et d'avoir une croissance économique durable la plus élevée possible.

Il paraît évident que le FOMC possède un impact important sur les marchés financiers et sur l'économie américaine en générale. Il est donc primordiale de comprendre le processus sous-jacent aux décisions prises par les membres du comité des FOMC. Ces derniers se réunissent environ 7 à 8 fois par an afin de prendre des décisions sur la politique monétaire à mener.

Durant ces réunions, l'ensemble des phrases prononcées par les participants sont retranscrites à l'écrit. Ainsi, à l'issue de la réunion, l'intégralité des échanges à été consignée. Cependant, il faut attendre 5 ans avant que le contenu retranscrit soit rendu public. Cette transparence des meetings du FOMC (décidée en 1993) est assez singulière en comparaison aux autres grandes banques centrales. Par exemple, la banque d'Angleterre ou encore la banque centrale européenne ne publient pas le contenu de leurs réunions relatives à leur politique monétaire. Ces transcriptions sont très précieuses pour les chercheurs en macroéconomie car elles permettent de comprendre le processus d'élaboration d'une prise de décision. De plus, il est intéressant d'étudier les relations entre les membres du comité afin d'évaluer le caractère démocratique de la discussion.

Avant de dégager des problématiques qui nous semblent pertinentes, il est nécessaire d'expliquer la structure d'un meeting.

Durant les réunions sont présents :

- Un chair : il est en général le directeur (ou la directrice) de la FED et son rôle est d'animer la réunion (aussi parfois appelé président durant le rapport)
- Un vice chair
- Des gouverneurs
- Des présidents de banques de réserve
- Des membres du personnel de la FED choisis par le conseil
- Des économistes choisis par le conseil

Voici le déroulé chronologique d'un meeting :

1. Un rapport est réalisé par le gestionnaire des opérations d'open market de la Fed de New York chargé de faire converger les taux des fonds fédéraux vers le niveau cible fixé par le FOMC
2. Le chair annonce l'efficacité des opérations d'open market et discute de l'évolution de la situation économique et financière depuis la précédente réunion. Pour ceci, les membres du comité ont à disposition des documents comme le livre vert qui rassemble un ensemble de prédictions pour le futur de l'économie américaine. Ces livres sont distribués une semaine avant le meeting. Il existe d'autres documents similaires mais il me semble peu important de les détailler.
3. La réunion se poursuit ensuite avec le premier des deux "tours de table", qui sont au cœur des réunions du FOMC.
 - Au cours du premier tour, tous les gouverneurs de la FED et les présidents des banques de réserve discutent de leur vision des conditions économiques et financières. Le président du FOMC conclut la discussion et donne son propre point de vue sur l'économie. Ensuite, la discussion politique commence avec le directeur de la Division des affaires monétaires du Conseil de la Réserve fédérale qui présente les différentes options politiques. Puis, le chair résume une proposition basée sur la discussion du comité. Il propose également une déclaration pour expliquer la décision politique. Les différentes possibilités sont une augmentation, une diminution ou une absence de changement du taux cible.
 - Ensuite, il y a un second tour. Les présidents et les gouverneurs des banques de réserve présentent chacun leurs arguments en faveur de l'option qu'ils préfèrent. À la fin de ce tour d'horizon politique, le chair résume à nouveau une proposition basée sur la discussion du comité. Les principaux membres du comité ont ensuite la possibilité de poser des questions ou de faire des commentaires sur l'approche proposée par le chair.
4. À la fin de ces deux tours, les sept gouverneurs des fédérations et les cinq présidents des banques de réserve votent officiellement la décision finale.

1.1.2 Notre approche et nos objectifs

Le fait que les transcriptions des réunions soient disponibles va nous nous permettre de mener des analyses statistiques afin de mettre en lumière des mécanismes sous jacents aux réunions.

Dans un premier temps, nous allons récupérer les données et les mettre sous forme de bases de données exploitables. Puis nous allons mener des statistiques descriptives sur les transcriptions afin de nous familiariser avec les données et d'exhiber des premières informations pertinentes qui permettront de construire un raisonnement.

Par la suite, nous aimerais étudier si le déroulé de la réunion est bien démocratique et si les échanges entre les différents acteurs sont égalitaires. Nous nous demandons notamment si le chair n'a pas trop d'influence sur la réunion. Pour répondre à cette question nous allons essayer d'établir des scores qui vont nous permettre de caractériser la similarité de comportement entre le chair et les autres membres du comité. De plus, nous allons essayer de mesurer directement l'impact du chair sur la réunion en étudiant le vocabulaire utilisé.

Enfin, nous essaierons de faire ressortir les thèmes relatifs à ces réunions et tenterons de construire des mesures de conformité ou de non conformité afin de faire ressortir le caractère démocratique de la réunion.

1.2 Revue de la littérature

Les économistes se sont intéressés à l'analyse statistique des transcriptions de réunion après la parution de l'article S.Hansen [2014] et al. Comme dit précédemment, la décision de rendre le contenu des réunions public à été prise en 1993. Les auteurs de cet article ont utilisé cette expérience naturelle pour comprendre comment la transparence impacte le délibéré en vue de la prise de décision d'une politique monétaire. Les auteurs prévoient un effet négatif sur la conformité du comité (un plus grand nombre de thèmes abordés et un débat plus engagé). Sans rentrer dans les détails, ils trouveront que, effectivement, les comportements des membres des réunions avant 1993 et après 1993 sont bien différents. Ce résultat a été obtenu en construisant des scores de conformité à partir de résultats de l'algorithme Latent Dirichlet Allocation (LDA) que nous présenterons plus tard dans ce rapport.

Récemment, deux économistes de la banque fédérale de réserve de San Francisco ont tenté de prédire les prises de décisions au niveau de la politique monétaire à partir du contenu des réunions. Dans cet article, A.Shapiro and D.Wilson [2019], les auteurs construisent une fonction d'objectif du comité qui est résultante de la façon dont les membres du comité se sont exprimés au cours d'une réunion.

Enfin, nous citerons un dernier article, ou plutôt une analyse statistique menée en ligne, Alexander Ng [2019]. Cette analyse a pour but principal d'utiliser les comptes rendus des réunions (des textes d'environ une page résumant la réunion) afin d'établir des corrélations avec des indices financiers. Leur première approche consiste simplement à utiliser des statistiques descriptives basiques, comme la longueur des comptes rendus afin de comparer les chairs entre eux.

Puis, ils utilisent des dictionnaires de sentiments (que nous utiliserons également dans ce rapport) afin de montrer des corrélations entre l'emploi d'un vocabulaire négatif (ou positif) et le cours de certains indices financiers. Le graphique ci-dessous issu de leurs analyses montre une forte corrélation entre le Russel 1000 (un indice prenant en compte 92% de la capitalisation boursière des actions cotées aux Etats-Unis) et le score de sentiments construit à l'aide des comptes rendus. Ces deux courbes sont comparées sur la période 2007 - 2019.



FIGURE 1 – Scaled Russel 1000 (vert) vs score de sentiment (rouge) sur la période 2007-2019

Cette analyse est particulièrement intéressante et nous nous demandons si nous trouverons des résultats similaires en prenant les transcriptions des réunions et pas uniquement les comptes rendus.

Comme les dates de parution des articles cités précédemment le prouvent, la recherche en statistique et en traitement du langage naturel sur les transcriptions des réunions du FOMC est très récente. Cependant, il nous tenait à cœur de ne pas reproduire des résultats mais d'essayer de prouver de nouvelles choses. Après avoir mené plusieurs analyses que nous montrerons ultérieurement dans ce rapport, mon maître de stage et moi-même avons décidé d'étudier comme axe d'analyse principal de ces réunions : **l'impact du chair sur les autres membres de la réunion**.

Pour cela, nous montrerons de simples corrélations à travers des statistiques descriptives. Ensuite, si possible, nous essaierons de prouver des causalités. La complexité du sujet vient du fait qu'avec la problématique choisie, les données ne sont pas étiquetées. Il faudra donc construire des labels intelligemment pour pouvoir utiliser des méthodes d'apprentissage supervisé.

1.3 Notes de lecture et précisions techniques

Dans ce rapport de stage, les conventions mathématiques adoptées seront les suivantes :

- n désignera le nombre d'observations et i indexera ces observations ($i \in \{1, \dots, n\}$)
- k désignera le nombre de variables et j indexera ces variables ($j \in \{1, \dots, k\}$)
- T désignera le nombre de périodes et t indexera ces variables ($t \in \{1, \dots, T\}$)
- X_i désignera le vecteur des variables predictives pour l'individu i lorsqu'on sera dans le cadre de l'apprentissage supervisé
- Y_i désignera la variable à prédire pour l'individu i lorsqu'on sera dans le cadre de l'apprentissage supervisé

Ce projet a requis l'utilisation de beaucoup de code car il fallait récupérer les données, les mettre en forme et appliquer des algorithmes. J'ai, en grande majorité, utilisé le language python et les packages suivants :

- pandas : pour la gestion des bases de données
- gensim et nltk : pour la manipulation de données textuelles et notamment toutes les fonctions de nettoyages, lemmatizations existantes
- requests : pour pouvoir scraper les données depuis le site web du FOMC
- pdftotext : pour transformer les fichiers pdf en fichier texte
- sklearn : pour notamment appliquer l'algorithme LDA
- datetime : pour la gestion de la date
- defaultdict et json : pour manipuler les dictionnaires en python que j'ai beaucoup utilisé durant ce stage
- matplotlib : pour sa sous classe pyplot avec laquelle j'ai fait la plupart de mes graphiques
- seaborn : pour les graphiques également
- numpy : pour les calculs mathématiques basiques
- PIL : pour la gestions des images en python et créer des animations utiles à la compréhension des données

J'ai aussi utilisé le language R lorsque j'ai dû faire des régressions avec *fixed effect*. Je trouve R plus intuitif lorsqu'il s'agit d'utiliser des outils économétriques.

2 Construction de la base de données et premières statistiques descriptives

Dans cette section, nous allons voir les différentes étapes qui nous ont permis de construire des bases de données exploitables en passant par le scraping, le nettoyage et la mise en forme.

2.1 Web Scraping

Premièrement, l'intégralité des transcriptions des réunions se trouvent sur le site de la FED au lien suivant : https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm. Mon maître de stage m'a demandé de m'intéresser à la période **d'août 1976 à décembre 2014** (dernière réunion disponible).

Nous devons alors pouvoir trouver un moyen de télécharger l'intégralité des 319 transcriptions des réunions de façon rapide. Il faut remarquer simplement que le l'url, pour accéder au pdf, s'écrit : 'https ://www.federalreserve.gov/monetarypolicy/files/FOMC'+str(date)+'meeting.pdf'. Il suffit alors d'écrire une fonction stockant dans une liste l'intégralité des dates que nous désirons puis télécharger l'ensemble des fichiers voulus grâce à la librairie *requests* de python.

Une fois ceci effectué, il faut pouvoir transformer les pdfs en données textuelles afin d'analyser le contenu des réunions. Après quelques recherches, il s'est avéré que le package *pdftotext* était la solution. Après cela, nous obtenons pour chaque réunion un fichier texte avec l'intégralité des mots prononcés durant celles-ci.

Ensuite, un nettoyage des fichiers textes a été nécessaire. En effet, les textes en tant que tels possèdent des paragraphes, des espaces en trop, des annotations de la personne retranscrivant la réunion à l'écrit ainsi que des numéros de page . De plus, chaque fichier comporte une introduction présentant la réunion ainsi que les membres qui la compose (voir la figure 2). Dans l'objectif que nous nous sommes fixés, nous sommes uniquement intéressés par les échanges entre les membres du comité. Il a donc fallu écrire un code python permettant de nettoyer cela.

De plus, nous choisissons dans un premier temps de stocker chaque réunion sous forme de liste où un élément est une combinaison de string et nous passons à l'élément d'après lorsqu'il y a retour à la ligne (voir la figure 3).

2.2 Mise en forme de la base de données

Comme observé sur la figure 3, on constate qu'il y a encore des éléments à nettoyer tels des dates, des espaces en trop, etc. Nous ferons cela par la suite. À présent, il faut trouver une forme convenable pour exploiter les données. De ce fait, nous allons construire deux bases de données qui seront complémentaires dans l'information qu'elles nous apportent.

La première forme de base de données aura pour but de transformer chaque réunion en dictionnaire python où la clé sera l'individu ayant pris la parole et la valeur sera une liste comprenant l'intégralité des mots prononcés durant la réunion.

A meeting of the Federal Open Market Committee was held in the offices of the Board of Governors in Washington, D.C., on Tuesday, January 24, 2012, at 10:00 a.m., and continued on Wednesday, January 25, 2012, at 8:30 a.m. Those present were the following:

Ben Bernanke, Chairman
William C. Dudley, Vice Chairman
Elizabeth Duke
Jeffrey M. Lacker
Dennis P. Lockhart
Sandra Pianalto
Sarah Bloom Raskin
Daniel K. Tarullo
John C. Williams
Janet L. Yellen
James Bullard, Christine Cumming, Charles L. Evans, Esther L. George, and Eric Rosengren, Alternate Members of the Federal Open Market Committee
Richard W. Fisher, Narayana Kocherlakota, and Charles I. Plosser, Presidents of the Federal Reserve Banks of Dallas, Minneapolis, and Philadelphia, respectively
William B. English, Secretary and Economist
Deborah J. Danker, Deputy Secretary
Matthew M. Luecke, Assistant Secretary
David W. Skidmore, Assistant Secretary
Michelle A. Smith, Assistant Secretary
Scott G. Alvarez, General Counsel
Thomas C. Baxter, Deputy General Counsel
Steven B. Kamin, Economist
David W. Wilcox, Economist
David Altig, Thomas A. Connors, Michael P. Leahy, William Nelson, Simon Potter, David Reifschneider, Glenn D. Rudebusch, and William Wascher, Associate Economists Brian Sack, Manager, System Open Market Account
Michael S. Gibson, Director, Division of Banking Supervision and Regulation, Board of Governors
Nellie Liang, Director, Office of Financial Stability Policy and Research, Board of Governors

January 24–25, 2012
Jon W. Faust and Andrew T. Levin, Special Advisors to the Board, Office of Board Members, Board of Governors
James A. Clouse, Deputy Director, Division of Monetary Affairs, Board of Governors
Linda Robertson, Assistant to the Board, Office of Board Members, Board of Governors
Daniel E. Sichel, Senior Associate Director, Division of Research and Statistics, Board of Governors

FIGURE 2 – Aperçu de la transcription de la réunion du 25/01/12 en format texte

[
' CHAIRMAN BERNANKE, Good morning, everybody. I'd like to start by recognizing\n',
'our colleague, Larry Slifman, who is at his last meeting before his planned retirement. Larry
is\n', 'still fairly junior, having been on the Board staff almost 42 years. [Laughter] He has
attended\n', '183 FOMC meetings over 30 years. At one day per meeting, that's almost exactly six
months of\n', 'FOMC meetings. [Laughter] Larry has shown great economic insight but has also
excelled in\n', 'mentoring others in the art of presenting complex material to the Board in the
clearest and most\n', 'logical manner. Larry, those of us around the table and many predecessors
have benefited\n', 'greatly from your dedicated service. Congratulations and best wishes for the
next phase. Thank\n', 'you very much. [Applause]\n', ' CHAIRMAN BERNANKE. I'd like to
welcome, of course, President Lacker, President\n', 'Lockhart, President Pianalto, and President
Williams to the FOMC. We will have a formal\n', 'organizational part of the meeting a little bit
later this morning, but I thought it would be useful\n', 'first to begin with our special topic,
which we're looking forward to. The topic is the role of\n', 'financial conditions in economic
recovery: lending and leverage. This was a highly favored\n', 'pick of FOMC participants when we
polled you last year about what you would like to talk\n', 'about. I particularly want to thank
Glenn Rudebusch in San Francisco for organizing this session\n', 'and acknowledge the
presenters, John Duca from Dallas, Andrew Haughwout from New York,\n', 'and Daniel Cooper from
Boston. Let me call on John.\n', ' MR. DUCA. 1 Thank you, Mr. Chairman. I will be
referring to the handout on\n', ' lending and leverage. This presentation, coauthored with
Anthony Murphy, links the\n', ' sluggish recovery in personal consumption expenditures (PCE)
to financial factors\n', ' and then shows how movements in consumption reflect long- and
short-run shifts in\n', '1\n', ' The materials used by Mr. Duca, Mr. Haughwout, and Mr. Cooper
are appended to this transcript (appendix 1).\n', '\n', 'January 24–25, 2012
5 of 314\n', ' wealth and the availability of consumer and mortgage credit. We end by
discussing\n', ' how recent consumer spending has been bolstered by some stabilization of
household\n', ' balance sheets, coupled with an upturn in the supply of consumer credit.\n', '
To provide a benchmark, exhibit 1 plots real per capita consumption normalized\n', ' around
the prior five major business cycle peaks. Consumer spending barely fell\n', ' during these
recessions—whether in terms of the average of those cycles (the black\n', ' line) or their
range (the shaded gray area). In the current cycle (the red line),\n', ' consumer spending
declined by nearly 5 percentage points before hitting bottom.\n', ' Moreover, in the earlier
episodes, consumption recovered rapidly. By comparison,\n', ' per capita consumer spending has

FIGURE 3 – Transcription après nettoyage et mise sous forme de liste

La deuxième base de données rassemble simplement le numéro de l'intervention prononcée ainsi que le nom de l'interlocuteur. On ajoutera également la taille de l'intervention en nombre de mots.

Lorsqu'on croise ces deux bases de données, nous possérons toutes les informations nécessaires pour exploiter les réunions. Les mettre sous cette forme est bien plus simple pour coder et pour construire des statistiques descriptives.

2.2.1 Construction des dictionnaires pour chaque réunion

Dans un premier temps, nous voulons pouvoir décomposer les transcriptions des meetings en base de données de manière à identifier pour chaque phrase l'interlocuteur qui l'a prononcée. J'ai ainsi choisi de transformer les transcriptions en dictionnaire où la clé sera le nom de l'interlocuteur et la valeur sera une liste contenant l'intégralité de mots (et dans l'ordre) prononcés par l'interlocuteur. Voici un court exemple :

Avant (début du texte lorsque le chair prend la parole) :

CHAIRMAN BERNANKE. Good afternoon. This is Mark Sniderman's last FOMC meeting before he retires from the Federal Reserve Bank of Cleveland in January. Mark started attending FOMC meetings in 1985, and, including today's meeting, he has attended 115 meetings since then, which makes him the current record holder for both staff and policymakers.

Après (début de la liste pour la clé "CHAIRMAN BERNAKE" dans le dictionnaire) :

"CHAIRMAN BERNANKE." : ["CHAIRMAN", "BERNANKE", "Good", "afternoon", "This", "is", "Mark", "Sniderman", "last", "FOMC", "meeting", "before", "he", "retires", "from", "the", "Federal", "Reserve", "Bank", "Cleveland", "in", "January", "Mark", "started", "attending", "FOMC", "meetings", "in", "and", "including", "today", "meeting", "he", "has", "attended", "meetings", "since", "then", "which", "makes", "him", "the", "current", "record", "holder", "for", "both", "staff", "and", "policymakers",

On constate que dans la liste (la valeur correspondante à la clé d'un intervenant), son nom est répété à chaque nouvelle intervention. Dans notre cas, le nom est "CHAIRMAN", "BERNANKE". "CHAIRMAN", "BERNANKE" sera répété à chaque fois que le chairman Bernanke prend à nouveau la parole. On peut les supprimer si l'on s'intéresse au contenu effectivement dit mais il est aussi utile de les garder si on veut compter le nombre d'interventions différentes durant un meeting.

Note : Dans l'exemple ci-dessus cela n'apparaît qu'une seule fois car nous voyons seulement le début de la première intervention du Chairman Bernanke pour la réunion 2013/12/18.

Il est important de noter qu'il y'aura donc un dictionnaire par réunion car les intervenants peuvent changer et donc les clés ne seront plus les mêmes. De plus, le fait que le nombre d'intervenants change rend difficile la construction d'une base de données agrégée pour toutes les réunions. Ainsi, pour faire des statistiques agrégées on itérera sur les dictionnaires construits.

2.2.2 Construction de la base de données retraçant la temporalité

En reprenant l'exemple de la réunion du 2013/12/18 la deuxième base sera celle de la figure 4 ci-dessous.

À présent, nous possédons deux bases de données, riches en informations. La première par son contenu textuel et la seconde par des informations agrégées et sur la temporalité de la réunion. Nous allons essayer d'extraire une première vague d'informations à l'aide de statistiques descriptives.

Date	interlocutor_name	statement_size	statement_number
2013-12-18	CHAIRMAN BERNANKE.	125	statement_0
2013-12-18	MR. SNIDERMAN.	42	statement_1
2013-12-18	CHAIRMAN BERNANKE.	60	statement_2
2013-12-18	MR. PLOSSER.	84	statement_3
2013-12-18	CHAIRMAN BERNANKE.	34	statement_4
2013-12-18	MR. POTTER.	1563	statement_5
2013-12-18	MS. LOGAN.	1074	statement_6
2013-12-18	CHAIRMAN BERNANKE.	14	statement_7
2013-12-18	VICE CHAIRMAN DUDLEY.	93	statement_8
2013-12-18	CHAIRMAN BERNANKE.	4	statement_9
2013-12-18	MR. LACKER.	56	statement_10
2013-12-18	MR. POTTER.	150	statement_11
2013-12-18	MR. LACKER.	18	statement_12
2013-12-18	MR. POTTER.	11	statement_13
2013-12-18	MR. LACKER.	14	statement_14
2013-12-18	MR. POTTER.	134	statement_15
2013-12-18	MR. LACKER.	42	statement_16
2013-12-18	MR. POTTER.	107	statement_17
2013-12-18	MR. LACKER.	37	statement_18
2013-12-18	MR. POTTER.	68	statement_19
2013-12-18	MR. LACKER.	12	statement_20
2013-12-18	VICE CHAIRMAN DUDLEY.	6	statement_21
2013-12-18	MR. POTTER.	12	statement_22
2013-12-18	MR. LACKER.	35	statement_23
2013-12-18	MR. POTTER.	57	statement_24
2013-12-18	MR. LACKER.	22	statement_25
2013-12-18	MR. POTTER.	18	statement_26
2013-12-18	MR. LACKER.	33	statement_27
2013-12-18	MR. POTTER.	53	statement_28
2013-12-18	MR. LACKER.	57	statement_29

FIGURE 4 – Base de données retraçant la temporalité d'une réunion et la taille des interventions

2.3 Statistique descriptive

Comme énoncé en introduction, nous nous concentrerons sur la période d'août 1976 à décembre 2014. Le principal objectif de ce rapport étant de s'intéresser à l'influence du chair sur les autres membres de la réunion, il est essentiel de voir les chairs ayant dirigé les réunions sur la période étudiée. Voici une frise chronologique présentant cela.

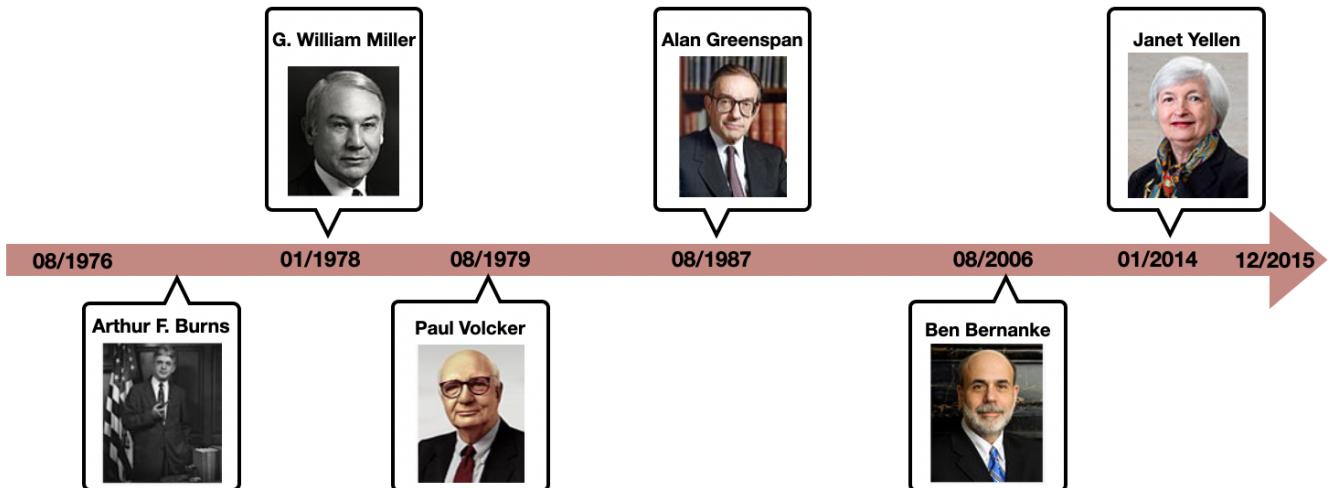


FIGURE 5 – Historique des chairs sur la période étudiée

Sur la période 1976-2015 il y a donc eu **6 chairs différents** pour un ensemble de **319 réunions**. Dans un premier temps, nous allons nous intéresser à la longueur des réunions en terme de nombres de mots et notamment regarder la place que prend chaque chair au sein de ses réunions et voir si on observe des inégalités.

2.3.1 Statistique sur la taille des réunions et la présence des chairs

Ensuite, nous allons nous pencher sur le nombre total de mots prononcés à chaque réunion afin d'avoir un indicateur concernant la longueur des réunions. Les couleurs indiquent le nom du chair qui en charge de la réunion à ce moment là.

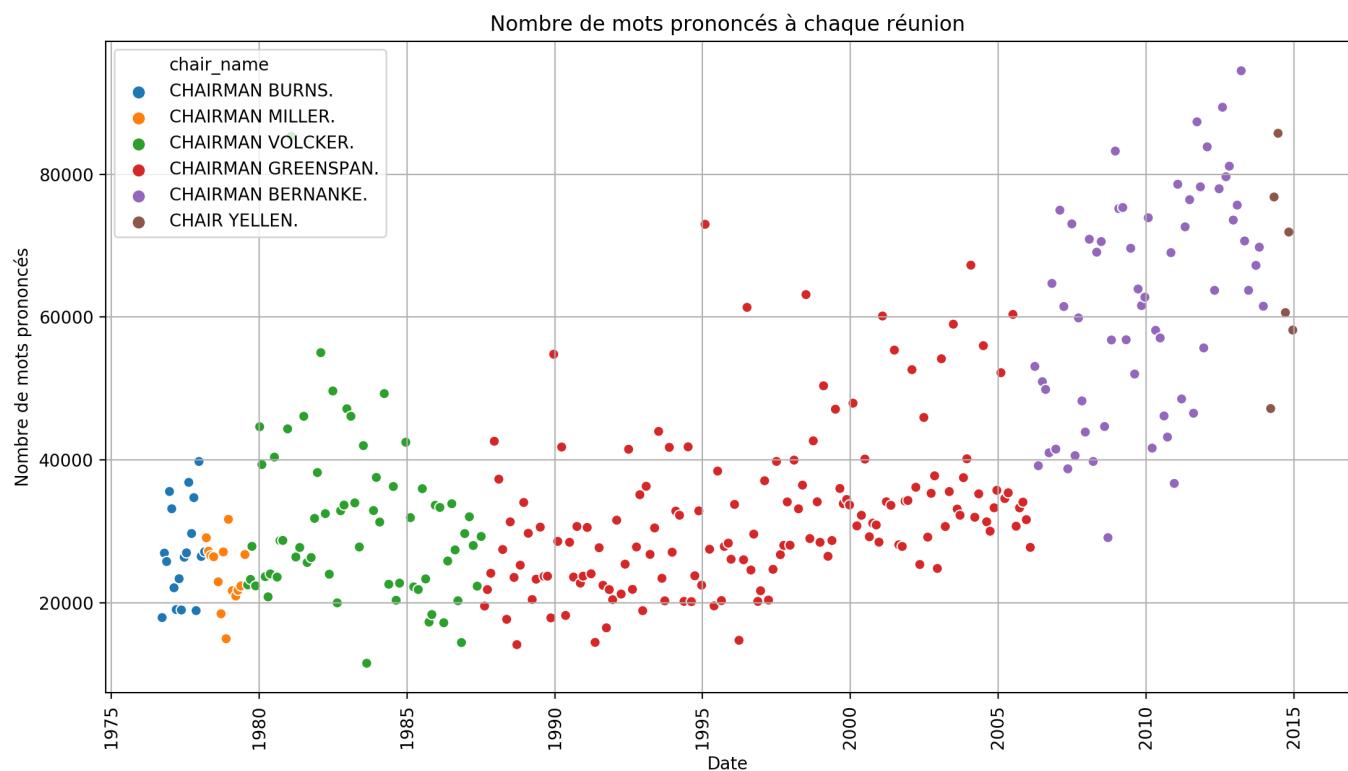


FIGURE 6 – Nombre total de mots prononcés à chaque réunion pour la période 1976-2015

	mean	std
chair_name		
CHAIR YELLEN.	66741.833333	14019.016705
CHAIRMAN BERNANKE.	62265.016129	15362.065525
CHAIRMAN BURNS.	27194.166667	6657.852385
CHAIRMAN GREENSPAN.	32316.570470	11040.312968
CHAIRMAN MILLER.	24121.642857	4483.142658
CHAIRMAN VOLCKER.	31107.437500	11658.593564

FIGURE 7 – Moyenne et écart-type du nombre de mots total prononcés (les statistiques sont regroupées par chair)

En regardant la figure 6 et la figure 7, on constate une forte variance du nombre de mots sur la plage étudiée, quel que soit le chair, même si cela est plus marqué pour le Chairman Bernanke (ceci pouvant s'expliquer peut-être par la crise financière) que pour le Chairman Miller. De plus, on s'aperçoit d'une tendance à la hausse de la durée des réunions notamment à partir du début des années 1990. Ceci peut s'expliquer par le fait qu'à partir de 1993, il a été décidé de rendre public le contenu des réunions (y compris celles avant 1993).

Ainsi, on peut supposer que les réunions se rallongent par soucis de clarté mais aussi pour manifester une rigueur professionnelle. En effet, dans l'article de S.Hansen [2014] et al, il a été prouvé que la transparence des réunions assurait un débat plus riche et engagé.

À présent, nous allons regarder la part du nombre de mots prononcés par le chair par rapport au nombre total de mots prononcés au cours de la réunion. On affichera ainsi un ratio entre 0 et 1 pour chaque réunion en indiquant de quel chair il s'agit.

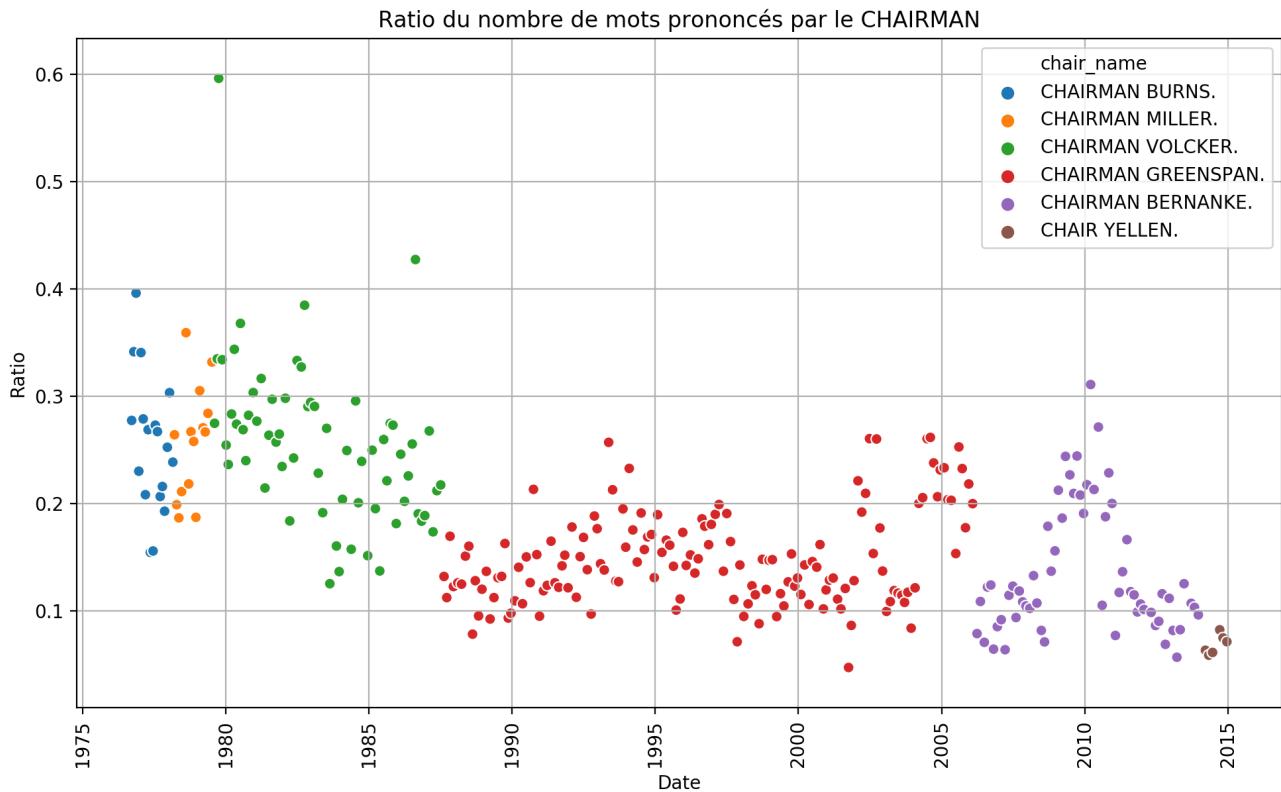


FIGURE 8 – Nombre de mots prononcés par les chairs relativement au nombre total de mots prononcés au cours de la réunion sur la période 1976-2015

chair_name	mean	std
CHAIR YELLEN.	0.068844	0.009105
CHAIRMAN BERNANKE.	0.133066	0.058519
CHAIRMAN BURNS.	0.255772	0.063922
CHAIRMAN GREENSPAN.	0.149052	0.043855
CHAIRMAN MILLER.	0.257889	0.052831
CHAIRMAN VOLCKER.	0.255690	0.075245

FIGURE 9 – Moyenne et écart-type du nombre de mots prononcés le chair relativement au total de mots prononcés

De façon globale, on constate que les chairs ont tendance à prendre de moins en moins la parole au cours des réunions. Cela est notamment marquant avec la Chair Yellen qui prend beaucoup moins la parole que ses prédécesseurs. Néanmoins, cela reste à vérifier car la taille de l'échantillon où Madame Yellen est chair est assez petit.

De plus, on constate que les chairs ont eu plus tendance à prendre la parole au moment des crises. Ceci est marquant après le krash boursier de 2001-2002, période où le Chairman Greenspan a davantage pris la parole.

Ce phénomène est encore plus notable au moment de la crise de 2008 où le Chairman Bernanke, qui intervenait assez peu dans les réunions (avant 2008), va aller jusqu'à occuper 30 % du temps de parole après 2008. Il subsiste malgré tout une volatilité qui est assez difficile à expliquer sans complément d'information contextuelle.

Nous allons désormais nous pencher sur des statistiques descriptives sur la temporalité des réunions ainsi que sur la taille des interventions des participants.

2.3.2 Statistiques de temporalités

Dans cette section, nous allons étudier les interventions d'un chair au cours d'une réunion. Nous allons majoritairement nous intéresser au moment où ces interventions sont prononcées et à la longueur de celles-ci (en nombre de mots). Dans l'objectif de montrer cela à mon maître de stage, j'ai réalisé une animation qui montre réunion après réunion les moments où le chair intervient et en quelle proportion. Ce type de fichier est trop lourd pour un fichier PDF, je vais donc, montrer les cas "types" de réunion pour chaque chair puis essayer de prouver de façon plus quantitative les intuitions montrées par les premiers graphes.

Voici ci-dessous des graphiques représentant le déroulé typique d'une réunion pour la Chair Yellen, le Chairman Bernanke et le Chairman Greenspan. Des graphiques similaires pour les trois plus anciens chairs se trouvent en annexe 7.1.

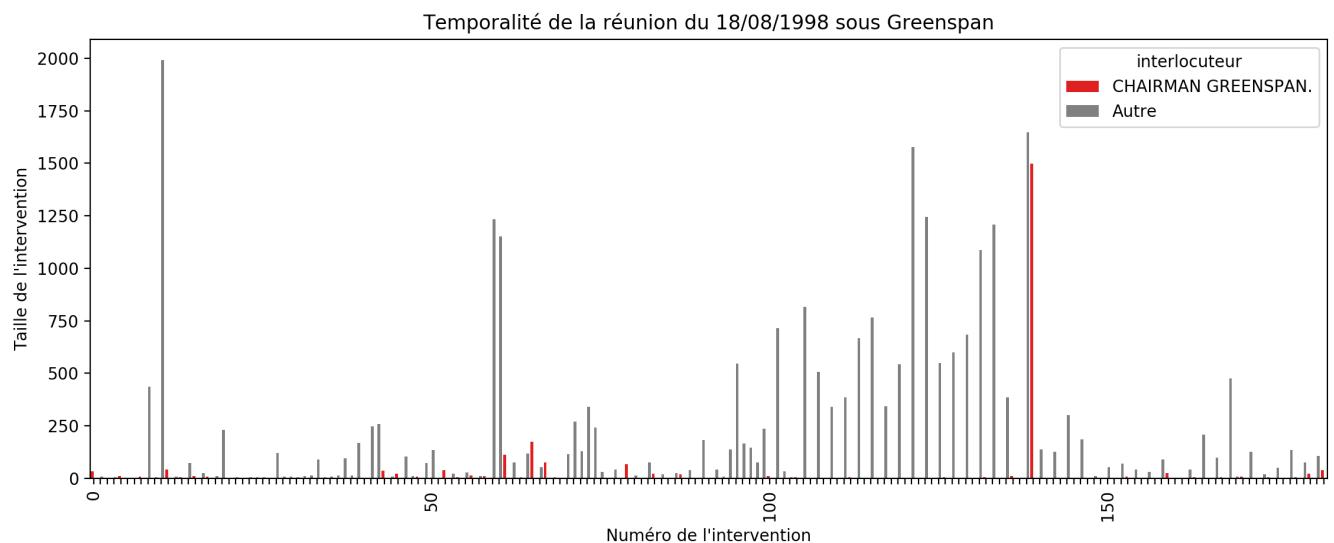


FIGURE 10 – Déroulé temporel d'une réunion sous Greenspan. L'axe des y représente le nombre de mots prononcés pour une intervention.

J'ai choisi pour chaque chair une réunion **représentative de la temporalité des réunions** (les animations sont trop lourdes pour être incluses).

En observant la figure 10, on se rend compte que le Chairman Greenspan intervient assez peu dans la discussion. À l'exception de la grande intervention vers la fin de la réunion. Il apparaît que les interventions de Greenspan sont assez courtes et ont certainement pour vocation de rythmer le débat

et de résumer les points essentiels des deux tours. Néanmoins, on constate un fort pic vers la fin de la réunion. On constate également qu'après ce pic il y a très peu d'interventions excédant 100 mots. En reprenant le déroulé historique des réunions décrit en introduction dans la partie 1.1.1, on peut inférer que ce pic correspond à la clôture de la discussion sur les politiques à mener entre les membres du comité.

Regardons ci-dessous des graphiques similaires pour des réunions que nous pensons représentatives de la tendance pour Chair Yellen et Chairman Bernanke.

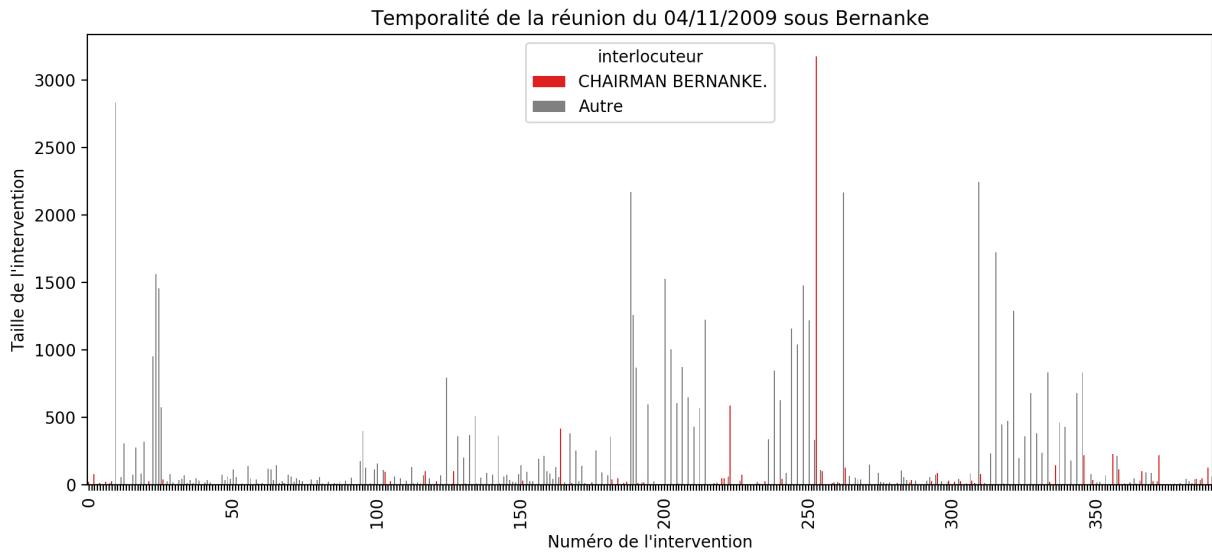


FIGURE 11 – Déroulé temporel d'une réunion sous Bernanke.

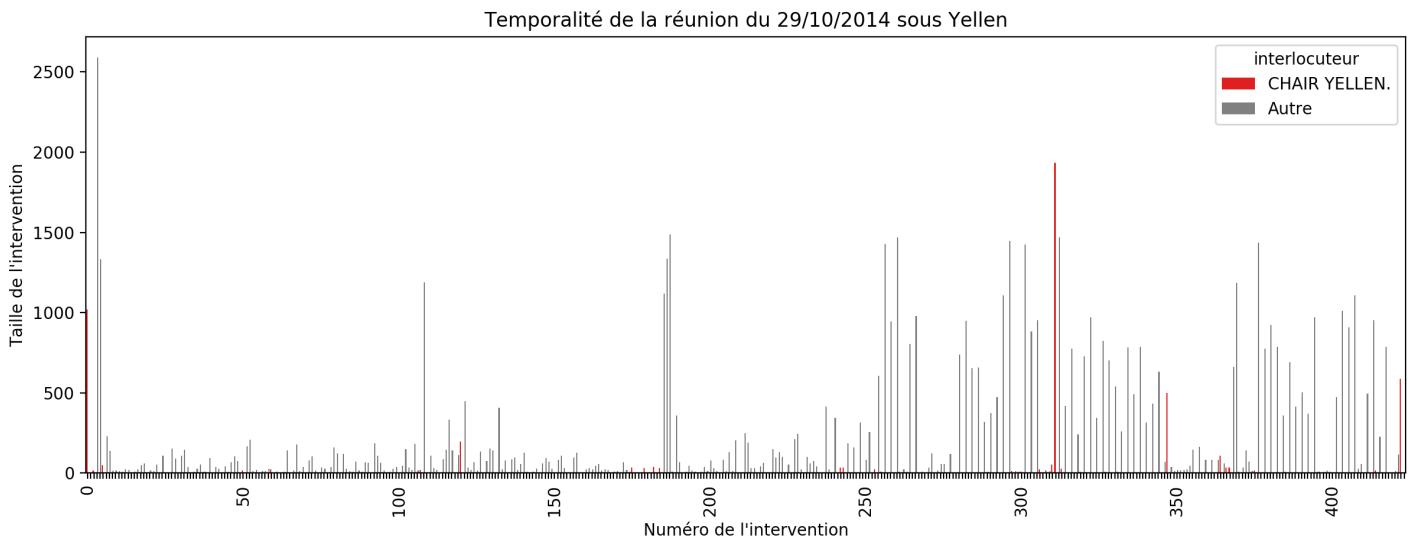


FIGURE 12 – Déroulé temporel d'une réunion sous Yellen

On s'aperçoit qu'un fort pic est également présent pour ces deux chairs aux alentours des 3/4 de la réunion. Cependant, derrière ces pics, il y a encore quelques interventions majeures (voir un assez grand nombre d'interventions pour la Chair Yellen). Ces graphiques représentent l'autorité d'un chair. En effet, si le mot de la fin précédent le vote est toujours prononcé par le Chair Greenspan, qui en profite pour souligner les options qu'il préfère, cela peut avoir un fort impact sur le délibéré final. En comparaison, le comportement de Chair Yellen et Chair Bernanke peut être qualifier de plus démocratique.

Afin d'avoir une démarche scientifique plus juste, et comme les graphiques présentés ci-dessus sont choisis de façon arbitraire, nous allons présenter une statistique pour prouver le phénomène avancé. Nous allons regarder une statistique donnant le pourcentage moyen de mots prononcés (moyenne réalisée sur les réunions) après la grande intervention du Chair identifiable par les grandes pics rouges sur les graphiques ci-dessus.

Nom du chair	Moment moyen du pic	Variance du pic	Ratio moyen de mots prononcés par les autres membres après le pic	Variance de ce ratio
Chairman Burns	0.47	0.29	0.39	0.19
Chairman Miller	0.43	0.19	0.33	0.14
Chairman Volcker	0.47	0.22	0.27	0.19
Chairman Greenspan	0.69	0.15	0.15	0.11
Chairman Bernanke	0.53	0.17	0.34	0.14
Chair Yellen	0.72	0.10	0.26	0.10

TABLE 1 – Ratio du nombre de mots prononcés par les autres membres du comité après le pic de parole du chair (par rapport au nombre total de mots prononcés).

Comme attendu, on constate qu'après le pic de parole du Chairman Greenspan clôturant le discussion, peu de mots sont échangés. On s'en aperçoit notamment en comparant avec la chair Yellen. De plus, ce tableau confirme qu'au fur et à mesure du temps, les chair prennent la parole de plus en plus tard dans la réunion, ils passent d'un rôle d'encadrement à un rôle de synthèse.

Les différentes statistiques proposées dans cette partie confirment qu'il y a une différence de comportement entre les chairs au sein des réunions. Il semblerait que certain sont plus autoritaires, d'autres laissent plus la parole etc. Dans la section suivante, nous allons essayer de mesurer l'impact d'un chair sur les autres membres de la réunion.

3 Analyse de sentiments

Un des problèmes majeurs du corpus de texte est que les données en tant que telles sont non labélisées. On aimerait construire une fonction partant des données et donnant en sortie une mesure de l'impact d'un chair sur les autres membres de la réunion. Cependant, avec les données que nous possédons, il est difficile de construire ce genre de mesure. On peut essayer de construire un score basé sur le nombre de mots ou sur la temporalité des réunions, toutefois la construction de ce type de score ne paraît pas simple.

Pour remédier à ce problème, nous allons puiser dans des dictionnaires de mots construits par des chercheurs. Ces dictionnaires regroupent des catégories de mots consignées à la main.

3.1 Présentation des dictionnaires de sentiments et construction des scores

3.1.1 Dictionnaire de Loughran and McDonald

Pour analyser la polarité des réunions, nous allons nous servir d'un dictionnaire construit par Loughran and McDonald s'appliquant au domaine de la finance. On le retrouve au lien suivant <https://sraf.nd.edu/textual-analysis/resources/>. Ce dictionnaire se présente sous la forme d'une base de données dont les premières lignes sont montrées ci dessous.

A	B	C	D	E	F	G
Negative	Positive	Uncertainly	Litigious	StrongModal	WeakModal	Constraining
ABANDON	ABLE	ABEYANCE	ABOVEMENTION	ALWAYS	ALMOST	ABIDE
ABANDONED	ABUNDANCE	ABEYANCES	ABROGATE	BEST	APPARENTLY	ABIDING
ABANDONING	ABUNDANT	ALMOST	ABROGATED	CLEARLY	APPEARED	BOUND
ABANDONMENT	ACCLAIMED	ALTERATION	ABROGATES	DEFINITELY	APPEARING	BOUNDED
ABANDONMENT	ACCOMPLISH	ALTERATIONS	ABROGATING	DEFINITIVELY	APPEARS	COMMIT
ABANDONS	ACCOMPLISHED	AMBIGUITIES	ABROGATION	HIGHEST	CONCEIVABLE	COMMITMENT
ABDICATED	ACCOMPLISHES	AMBIGUITY	ABROGATIONS	LOWEST	COULD	COMMITMENTS
ABDICTATES	ACCOMPLISHING	AMBIGUOUS	ABSOLVE	MUST	DEPEND	COMMITS
ABDICATING	ACCOMPLISHME	ANOMALIES	ABSOLVED	NEVER	DEPENDED	COMMITTED
ABDICTION	ACCOMPLISHME	ANOMALOUS	ABSOLVES	STRONGLY	DEPENDING	COMMITTING
ABDICTIONS	ACHIEVE	ANOMALOUSLY	ABSOLVING	UNAMBIGUOUSI	DEPENDS	COMPEL
ABERRANT	ACHIEVED	ANOMALY	ACCESSION	UNCOMPROMISI	MAY	COMPELLED
ABERRATION	ACHIEVEMENT	ANTICIPATE	ACCESSIONS	UNDISPUTED	MAYBE	COMPELLING
ABERRATIONAL	ACHIEVEMENTS	ANTICIPATED	ACQUIREES	UNDOUTBEDLY	MIHT	COMPELS
ABERRATIONS	ACHIEVES	ANTICIPATES	ACQUIRORS	UNEQUIVOCAL	NEARLY	COMPLY
ABETTING	ACHIEVING	ANTICIPATING	ACQUIT	UNEQUIVOCALLY	OCCASIONALLY	COMPULSION
ABNORMAL	ADEQUATELY	ANTICIPATION	ACQUITS	UNPARALLELED	PERHAPS	COMPULSORY
ABNORMALITIES	ADVANCEMENT	ANTICIPATIONS	ACQUITTAL	UNSURPASSED	POSSIBLE	CONFINE

FIGURE 13 – Dictionnaire de polarité

On constate ainsi que ce dictionnaire comporte 6 catégories : "positive", "negative", "uncertain", "strong modal", "weak modal", "litigious" et "constraining". Les auteurs ont ainsi listé un grand nombre de mots en anglais qui appartiennent à ces catégories. On note qu'il y a un biais car les catégories ne comptent pas le même nombre de mots.

Negative	Positive	Uncertainly	Litigious	Strong Modal	Weak Modal	Constraining
2356	355	298	905	20	28	185

TABLE 2 – Nombre de mots par catégories

A partir de ces mots, nous allons construire **des scores de positive, d'affirmation et d'incertitude**. Ces scores vont être calculés de la façon suivante (calculs simples qui peuvent être améliorés mais nous privilégiions la comparaison entre les réunions qui nous intéressent).

- $ScorePositivite = \frac{\sum_{word} (\mathbf{1}_{word \in Positive} - \mathbf{1}_{word \in Negative})}{\sum_{word} (\mathbf{1}_{word \in Positive} + \mathbf{1}_{word \in Negative})}$
- $ScoreAffirmation = \frac{\sum_{word} (\mathbf{1}_{word \in StrongModal} - \mathbf{1}_{word \in WeakModal})}{\sum_{word} (\mathbf{1}_{word \in StrongModal} + \mathbf{1}_{word \in WeakModal})}$
- $ScoreIncertitude = \frac{-\sum_{word} \mathbf{1}_{word \in Incertitude}}{NbWord}$

On note que les scores ne seront pas centrés en 0, car comme dit précédemment, il y a des catégories bien plus grandes que d'autres (par exemple les mots négatifs par rapport aux mots positifs). On pourrait modifier cela en rajoutant un coefficient plus faible aux mots négatifs (inversement proportionnel au nombre de mots négatifs par rapport au nombre de mots positifs). Cependant ce n'est pas vraiment cela qui nous intéresse mais plutôt la tendance entre les réunions et l'évolution du score dans le temps.

3.1.2 Dictionnaire d'Harvard

Il existe un autre dictionnaire recensant un grand nombre de catégories de mots réalisé par les chercheurs d'Harvard. Toutes ces catégories sont visibles sur le site suivant : <http://www.wjh.harvard.edu/~inquirer/homecat.htm>. On peut retrouver les catégories suivantes :

- Positive, Negative
- Weak, Strong
- Active, Passive

Ce type de mots se trouve déjà dans le dictionnaire de McDonald. Toutefois, on trouve également des catégories de mots liés à des professions ou à des sujets précis comme :

- Du vocabulaire économique
- Du vocabulaire académique

Cet éventail de vocabulaire nous permettrait de construire des scores variés mais nous allons nous concentrer dans un premier temps sur le dictionnaire de Loughran and McDonald.

3.2 Analyse de sentiments associés à chaque réunion

Dans cette partie, nous allons d'abord regarder l'évolution du score global de positivité et d'incertitude globale des réunions sur la période 1976-2015.

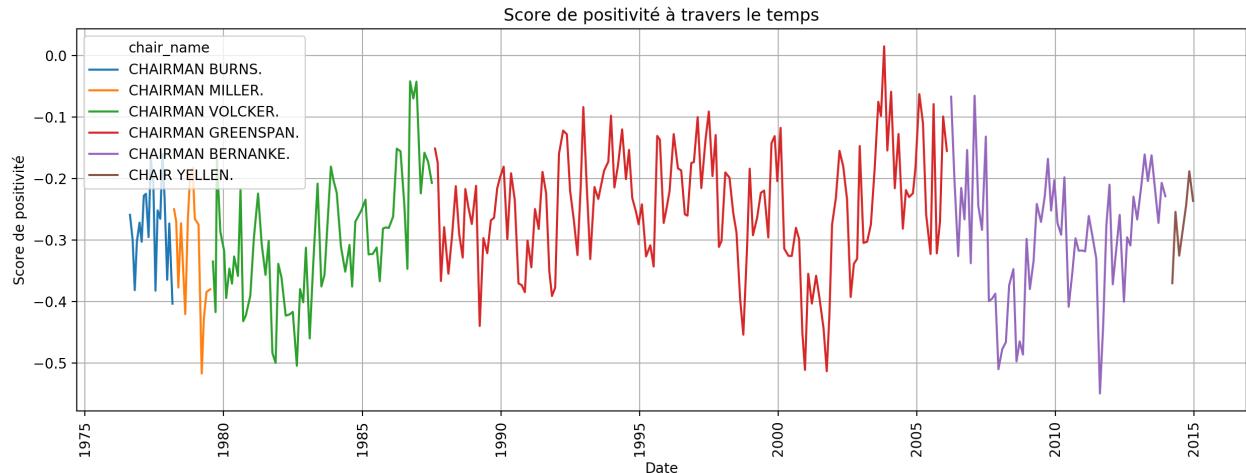


FIGURE 14 – Score global de positivité

On constate que le score de positivité est la très grande majorité du temps négatif. Cela est normal car le vocabulaire des mots négatifs est plus conséquent que celui des mots positifs et que, les membres du FOMC ont peut-être plus tendance à utiliser un vocabulaire négatif. Cependant, il est intéressant d'observer l'évolution de ce score entre les réunions. On constate que le score est au plus bas au moment des crises (par exemple 2002 et 2008) et lorsque la conjoncture économique est mauvaise. Cela est assez remarquable lorsqu'on compare avec l'indice du Dow Jones en 2002 et 2008 (voir annexe section 7.2). On constate malgré tout que le score de positivité est assez volatile.

À présent regardons le score global d'incertitude.

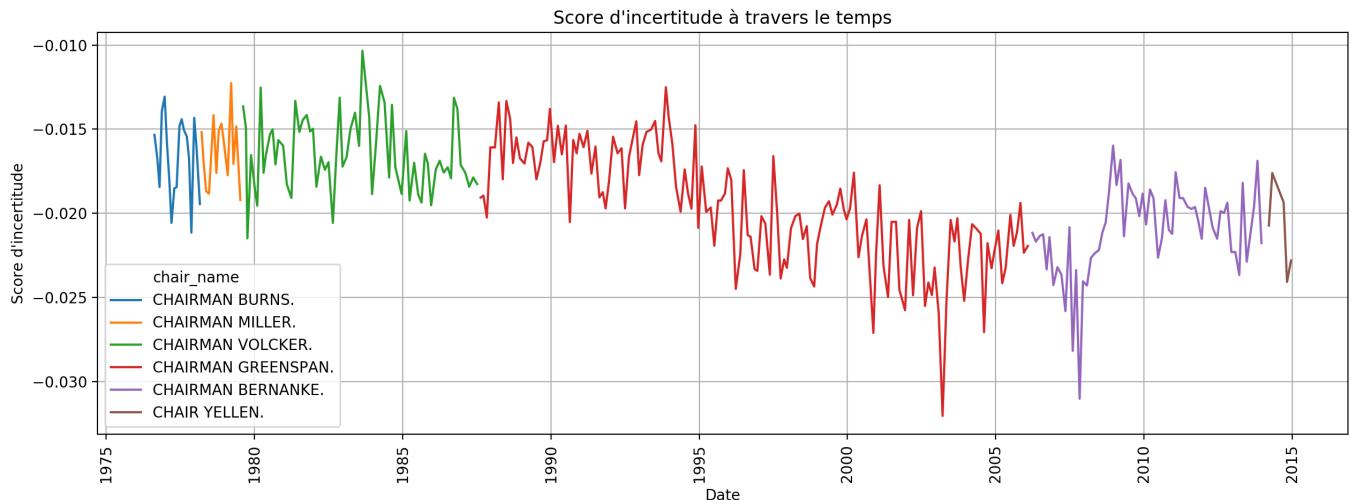


FIGURE 15 – Score global d'incertitude

Le score d'incertitude qui désigne l'emploi d'un vocabulaire incertain tel *approximately*, *confuses*, *could* (mais qui ne sont pas des verbes modaux) marque des tendances assez claires. Là encore, il semble y avoir une corrélation entre cet indice et la conjoncture économique, il suffit de comparer avec l'indice du Dow Jones. De plus ce score semble moins volatile que le score de positivité.

3.3 Analyse de sentiment relative aux chairs

Nous allons maintenant comparer **les scores des chairs relativement aux scores globaux**. Ce ratio va nous permettre de voir si il y a des tendances différentes entre les chairs est les autres membres de la réunion.

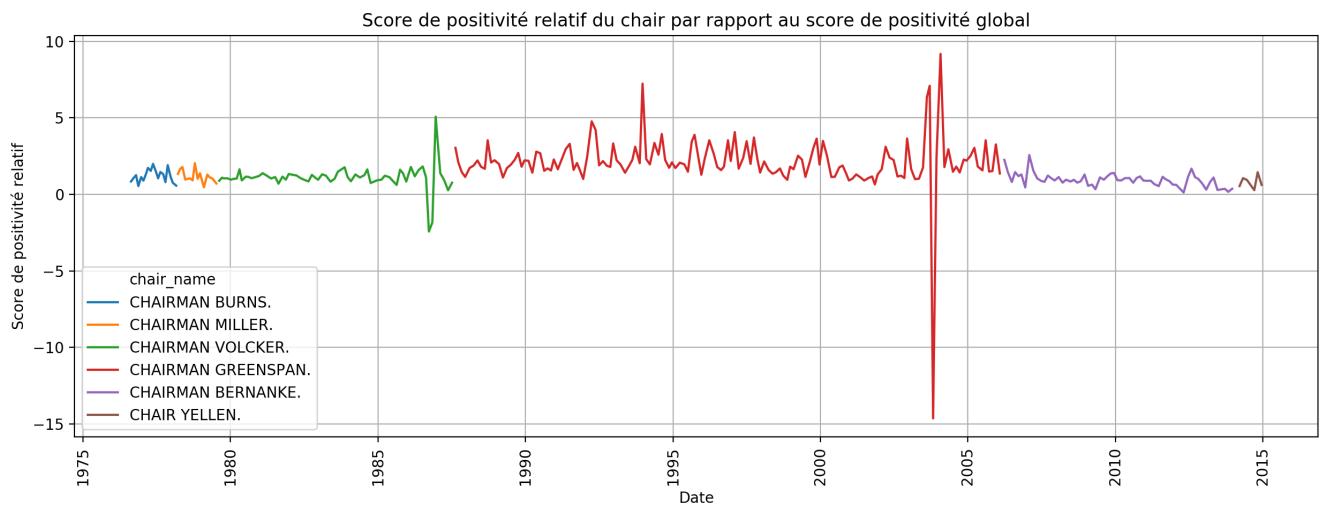


FIGURE 16 – Score de positivité relatif au chair par rapport au score de positivité global affiché en figure 14

Avant d'interpréter ce graphe, il faut se rappeler que ce score est un ratio où le dénominateur est quasiment tout le temps positif. Quand le score est supérieur à 1, cela signifie que le score du chair est plus négatif que le score global de la réunion et inversement quand le score est compris entre 0 et 1. Un score inférieur à 0 signifie que les deux scores du ratio sont contraires. Cependant cela n'arrive que très rarement (deux fois), nous considérerons que ce sont des outliers.

On en déduit donc que Greenspan semble, en moyenne, employer un vocabulaire plus "négatif" que les autres membres de la réunion. En effet, le score correspondant à Greenspan est souvent aux alentours de 2 et parfois plus élevé. À l'inverse, il semblerait que les scores des autres chairs soient plus stables, avec un score aux alentours de 1 avec parfois des variations entre 0 et 2.

Dans la figure 17, on s'aperçoit que le score d'incertitude est relativement stable. On rappelle que les numérateurs et dénominateurs sont tous les deux négatifs. Il semblerait que les chairs ont tendance à employer un vocabulaire moins incertain que les autres membres de la réunion. Cela peut s'expliquer par le fait qu'ils ont un rôle d'organisateur de la réunion et doivent se montrer rassurant. Malgré tout, on aperçoit des disparités entre les chairs. Le Chairman Volcker semble plus enclin à employer un vocabulaire incertain que les autres et le Chairman Greenspan possède un score d'incertitude plus volatile que celui des autres.

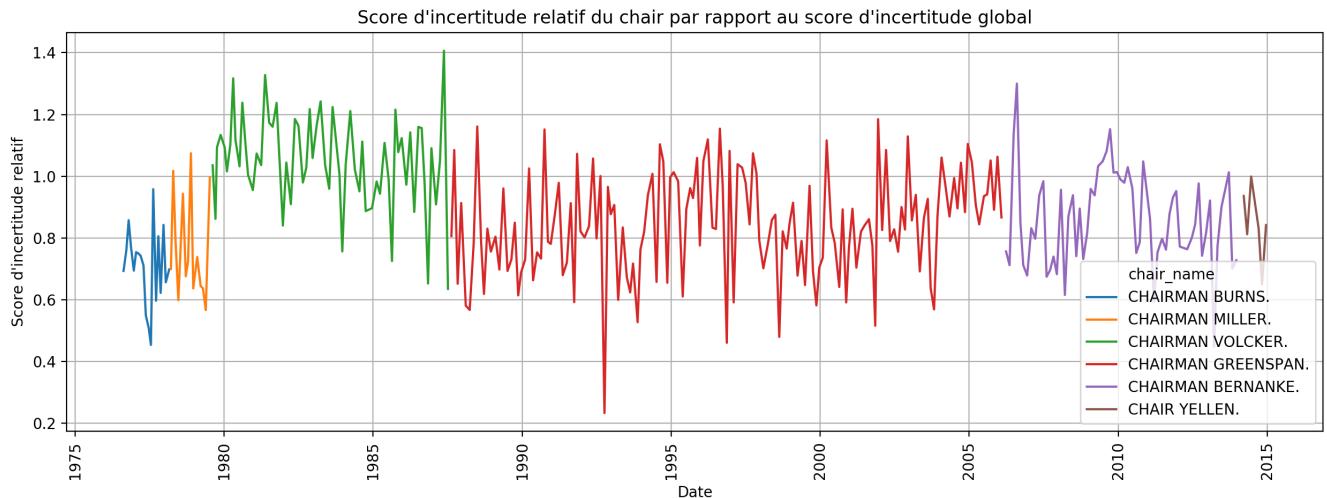


FIGURE 17 – Score d’incertitude relatif au chair par rapport au score d’incertitude global affiché en figure 15

Avec les informations que nous avons obtenues dans cette section, nous allons désormais essayer d’utiliser ces scores pour estimer des effets causaux au travers des régressions avec *fixed effect*. On note qu’on peut s’intéresser également à d’autres scores présents dans le dictionnaire d’Harvard. On peut, par exemple, construire un score du degré académique de la réunion par rapport au vocabulaire utilisé. Des chercheurs ont réussi à montrer avec ce genre de score que les membres des réunions emploient un vocabulaire plus académique depuis 1993, date à laquelle les réunions ont été décidée d’être rendues publiques.

4 Analyse du pouvoir relatif du chair sur les autres membres de la réunion

Dans cette section nous allons essayer de mesurer de façon causal l’impact d’un chair sur les autres membres de la réunion et pour ceci nous allons utiliser des régressions avec fixed effect. Nous allons notamment comparer des scores propres aux membres du comité présent sous Bernanke et Greenspan.

Ainsi, nous allons mesurer l’effet net de l’indicatrice *Greenspan est Chairman* ($0 \Rightarrow$ Bernanke est Chairman) sur le score des individus tout en essayant d’inclure des variables de contrôle comme le taux d’intérêt par exemple ou encore la valeur du nasdaq (indice boursier).

Nous nous intéresserons à la période Greenspan - Bernanke car de nombreux membres du comité ont été présents durant la gouvernance de Greenspan et celle de Bernanke. Il pourra être intéressant de reproduire les résultats sur d’autres périodes.

4.1 Régression avec fixed effect

Pour rappel, une régression avec fixed effect s'écrit de la façon suivante :

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_t + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}$$

où

- X_{it} désigne la variable explicative. Dans notre cas cette variable sera une indicatrice valant 1 quand le chairman est Greenspan et 0 quand le chairman est Bernanke.
- Y_{it} désigne la variable à expliquer. Dans notre cas cette variable désignera un score propre à un membre i du comité à un moment t que nous préciserons par la suite
- Z_t désigne les variables de contrôles. Dans notre cas cela pourra être le taux d'intérêt.
- β_0 désigne la constante du modèle, β_1 désigne le coefficient à la variable explicative d'intérêt, β_2 désigne le coefficient associé à la variable de contrôle.
- Les Dj_i désignent des indicatrices qui captent les effets fixes propres à un individu i .
- Les γ sont les coefficients associés à ces effets fixes.

4.1.1 Sur le score de positivité

La première variable Y_{it} à laquelle nous allons nous intéresser pour la période Greenspan - Bernanke désigne le score de positivité pour les individus présents au comité du FOMC, à la fois pour Greenspan et pour Bernanke. Pour ce faire, nous construisons la base de données ci-dessous :

Score	Name	Date	Chair	TauxInteret
-0.111111	MR. STERN.	1987-08-18	CHAIRMAN GREENSPAN.	6.70
0.333333	MR. KOHN.	1987-08-18	CHAIRMAN GREENSPAN.	6.70
-0.166667	MR. STERN.	1987-09-22	CHAIRMAN GREENSPAN.	7.54
-0.600000	MR. KOHN.	1987-09-22	CHAIRMAN GREENSPAN.	7.54
-0.294118	MR. KOHN.	1987-11-03	CHAIRMAN GREENSPAN.	6.06
-0.500000	MR. STERN.	1987-11-03	CHAIRMAN GREENSPAN.	6.06
-0.441860	MR. KOHN.	1987-12-16	CHAIRMAN GREENSPAN.	5.52
-0.052632	MR. STERN.	1987-12-16	CHAIRMAN GREENSPAN.	5.52
-0.142857	MR. GUYNN.	1987-12-16	CHAIRMAN GREENSPAN.	5.52
-0.739130	MR. KOHN.	1988-02-10	CHAIRMAN GREENSPAN.	5.72
0.230769	MR. STERN.	1988-02-10	CHAIRMAN GREENSPAN.	5.72
-0.500000	MR. KOHN.	1988-03-29	CHAIRMAN GREENSPAN.	6.57
-0.111111	MR. STERN.	1988-03-29	CHAIRMAN GREENSPAN.	6.57
-0.333333	MR. STERN.	1988-05-17	CHAIRMAN GREENSPAN.	6.98
0.500000	MR. KOHN.	1988-05-17	CHAIRMAN GREENSPAN.	6.98
-0.764706	MR. KOHN.	1988-06-30	CHAIRMAN GREENSPAN.	8.27
-0.200000	MR. STERN.	1988-06-30	CHAIRMAN GREENSPAN.	8.27
-1.000000	MR. KOHN.	1988-08-16	CHAIRMAN GREENSPAN.	8.22
-0.250000	MR. STERN.	1988-08-16	CHAIRMAN GREENSPAN.	8.22
0.000000	MR. STOCKTON.	1988-09-20	CHAIRMAN GREENSPAN.	8.06

FIGURE 18 – Aperçu de la base données nécessaire à la régression avec effet fixe

En complément de cette base de données nous pouvons apporter quelques informations. Premièrement, $n = 29$, cela signifie que nous nous intéressons à 29 personnes différentes dans notre panel. De plus, on précise que la personne la moins représentée, Ms. Smith, apparaît 17 fois et que la personne la plus représentée, Mr. Sterne, apparaît 176 fois. Deuxièmement, $T = 210$, cela signifie que la régression va couvrir 210 réunions différentes. En tout la base de données possèdent 2022 lignes.

Nous allons maintenant passer à la régression avec fixed effect (données de panel). Pour ce, la variable Name va être décomposée en indicatrice ainsi que la variable chair. La régression sera effectuée à l'aide du language R qui est plus adapté que python pour les études économétriques. On regardera principalement la p-value pour déterminer si une variable à un effet significativement différent de 0 sur la variable $Y_{i,t}$.

Coefficients:		Estimate	Std. Error	t value	Pr(> t)							
(Intercept)		-0.234570	0.031910	-7.351	2.86e-13 ***							
Chair CHAIRMAN GREENSPAN.		0.104378	0.018646	5.598	2.47e-08 ***							
NameMR. GUYN.		-0.024841	0.045609	-0.545	0.58606							
NameMR. HOENIG.		0.122950	0.038552	3.189	0.00145 **							
NameMR. KAMIN.		-0.223975	0.072975	-3.069	0.00218 **							
NameMR. KOHN.		-0.190977	0.037866	-5.043	4.99e-07 ***							
NameMR. KOS.		-0.086483	0.055260	-1.565	0.11774							
NameMR. LACKER.		-0.041460	0.047496	-0.873	0.38281							
NameMR. MADIGAN.		-0.014193	0.060059	-0.236	0.81321							
NameMR. MEYER.		-0.145339	0.054944	-2.645	0.00823 **							
NameMR. MOSKOW.		0.012768	0.043246	0.295	0.76784							
NameMR. OLSON.		0.119981	0.061859	1.940	0.05257 .							
NameMR. POOLE.		-0.080826	0.046393	-1.742	0.08162 .							
NameMR. REIFSCHEIDER.		-0.203902	0.070503	-2.892	0.00387 **							
NameMR. REINHART.		-0.159556	0.055920	-2.853	0.00437 **							
NameMR. SACK.		-0.069195	0.071988	-0.961	0.33657							
NameMR. SHEETS.		-0.121977	0.067420	-1.809	0.07057 .							
NameMR. SLIFMAN.		-0.120461	0.070861	-1.700	0.08930 .							
NameMR. STERN.		0.078221	0.038150	2.050	0.04046 *							
NameMR. STOCKTON.		-0.132413	0.042306	-3.130	0.00177 **							
NameMR. WILCOX.		-0.110330	0.058462	-1.887	0.05928 .							
NameMR. WILLIAMS.		-0.085006	0.073057	-1.164	0.24475							
NameMS. BIES.		-0.052723	0.058446	-0.902	0.36712							
NameMS. DANKER.		0.627387	0.051711	12.133	< 2e-16 ***							
NameMS. JOHNSON.		-0.038615	0.049067	-0.787	0.43139							
NameMS. MINEHAN.		-0.073401	0.042952	-1.709	0.08762 .							
NameMS. PIANALTO.		0.018989	0.045443	0.418	0.67609							
NameMS. SMITH.		0.388638	0.083708	4.643	3.66e-06 ***							
NameMS. YELLEN.		-0.101072	0.043873	-2.304	0.02134 *							
NameVICE CHAIRMAN GEITHNER.		0.102270	0.059195	1.728	0.08420 .							
TauxInteret		-0.003753	0.003641	-1.031	0.30289							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	.	'.'	0.1	' '	1
Residual standard error:	0.3236	on 1991 degrees of freedom										
Multiple R-squared:	0.189,	Adjusted R-squared:	0.1767									
F-statistic:	15.46	on 30 and 1991 DF,	p-value:	< 2.2e-16								

FIGURE 19 – Résultat sur R de la régression présentée ci-dessus

Afin de conforter notre analyse sur un effet positif du chair Greenspan sur le score de positivité des membres des réunions, on réalise la même régression en remplaçant le taux d'intérêt au jour t par la valeur du nasdaq au jour t. Pour voir les résultats de cette régression, on regardera le premier graphique de l'annexe 7.3. En regardant le résultat de cette régression avec cette autre variable de contrôle, on s'aperçoit que le coefficient associé à l'indicatrice valant 1 lorsque le chairman Greenspan dirige la réunion et 0 lorsque le Chairman Bernanke dirige la réunion vaut environ 0.09 et au regard de la p-value, ce coefficient est statistiquement différent de 0. Ce résultat nous conforte dans notre analyse précédente, néanmoins, on peut douter du fait que le nasdaq capture parfaitement les aléas économiques (en regardant la R carré). Toutefois, en comparant la p-value associée au coefficient du nasdaq et celle associée au coefficient du taux d'intérêt, le nasdaq semble être une meilleure variable de contrôle.

4.1.2 Score d'incertitude

La seconde variable Y_{it} à laquelle nous allons nous intéresser pour la période Greenspan - Bernanke désigne le score d'incertitude pour les individus ayant été présents dans le comité FOMC à la fois pour Greenspan et pour Bernanke.

Sur ce tableau de résultats, on voit que le coefficient associé à l'indicatrice valant 1 lorsque le chairman Greenspan dirige la réunion et 0 lorsque le Chairman Bernanke dirige la réunion vaut environ 0.10 et au regard de la p-value paraît statistiquement différent de 0. Cela signifie que toutes choses égales par ailleurs, et en accord avec le modèle, le fait que le Chairman Greenspan dirige la réunion par rapport au Chairman Bernanke augmente le score de positivité en moyenne de 0.10. On aurait pu s'attendre à d'autres résultats sachant que le chairman Greenspan semblait plus autoritaire. Néanmoins pour mon maître de stage, ce résultat peut être dû au fait que le Chairman Bernanke était en poste au moment de la crise de 2008 et que le taux d'intérêt est une mauvaise variable de contrôle pour purger l'effet de la conjoncture économique sur le score de positivité. Une autre idée pourrait être d'inclure comme variable de contrôle des indicateurs de l'économie comme le PIB ou encore des indices boursiers.

Pour faire cette régression nous utiliserons une base de données ayant également 2022 lignes. Là encore nous nous intéresserons à 210 dates différentes et à 29 personnes différentes où Ms. Smith apparaît peu de fois tandis que Mr. Stern est très présent. Le but de cette régression est de voir si le fait que la réunion soit dirigée par le Chairman Greenspan plutôt que le Chairman Bernanke favorise l'emploi d'un vocabulaire incertain. Les résultats de cette régression se trouvent ci-dessous.

Coefficients:		Estimate	Std. Error	t value	Pr(> t)
(Intercept)		1.194e-02	1.019e-03	11.719	< 2e-16 ***
ChairCHAIRMAN GREENSPAN.		4.696e-04	5.953e-04	0.789	0.430304
NameMR. GUYN.		8.638e-03	1.456e-03	5.933	3.50e-09 ***
NameMR. HOENIG.		5.973e-03	1.231e-03	4.853	1.31e-06 ***
NameMR. KAMIN.		6.520e-03	2.330e-03	2.799	0.005183 **
NameMR. KOHN.		1.015e-02	1.209e-03	8.399	< 2e-16 ***
NameMR. KOS.		6.226e-03	1.764e-03	3.529	0.000427 ***
NameMR. LACKER.		6.511e-03	1.516e-03	4.295	1.83e-05 ***
NameMR. MADIGAN.		1.405e-02	1.917e-03	7.328	3.38e-13 ***
NameMR. MEYER.		1.393e-02	1.754e-03	7.942	3.29e-15 ***
NameMR. MOSKOW.		6.119e-03	1.381e-03	4.432	9.84e-06 ***
NameMR. OLSON.		6.259e-03	1.975e-03	3.169	0.001551 **
NameMR. POOLE.		8.977e-03	1.481e-03	6.061	1.61e-09 ***
NameMR. REIFSCHEIDER.		4.838e-03	2.251e-03	2.149	0.031725 *
NameMR. REINHART.		1.045e-02	1.785e-03	5.854	5.60e-09 ***
NameMR. SACK.		1.129e-02	2.298e-03	4.915	9.61e-07 ***
NameMR. SHEETS.		7.514e-03	2.152e-03	3.491	0.000491 ***
NameMR. SLIFMAN.		7.685e-05	2.262e-03	0.034	0.972902
NameMR. STERN.		9.696e-03	1.218e-03	7.962	2.82e-15 ***
NameMR. STOCKTON.		1.112e-02	1.351e-03	8.233	3.26e-16 ***
NameMR. WILCOX.		1.495e-02	1.866e-03	8.009	1.94e-15 ***
NameMR. WILLIAMS.		8.182e-03	2.332e-03	3.508	0.000461 ***
NameMS. BIES.		4.606e-03	1.866e-03	2.469	0.013643 *
NameMS. DANKER.		3.898e-03	1.651e-03	2.361	0.018321 *
NameMS. JOHNSON.		5.588e-03	1.566e-03	3.567	0.000369 ***
NameMS. MINEHAN.		8.697e-03	1.371e-03	6.343	2.79e-10 ***
NameMS. PIANALTO.		1.033e-02	1.451e-03	7.122	1.48e-12 ***
NameMS. SMITH.		-2.648e-03	2.672e-03	-0.991	0.321879
NameMS. YELLEN.		1.438e-02	1.401e-03	10.270	< 2e-16 ***
NameVICE CHAIRMAN GEITHNER.		2.363e-02	1.890e-03	12.507	< 2e-16 ***
TauxInteret		1.021e-04	1.162e-04	0.878	0.379907

Signif. codes:		0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
1 '.'					
Residual standard error:	0.01033	on 1991 degrees of freedom			
Multiple R-squared:	0.1498,	Adjusted R-squared:	0.137		
F-statistic:	11.69	on 30 and 1991 DF,	p-value:	< 2.2e-16	

FIGURE 20 – Résultat sur R de la régression présentée ci-dessus

Bien que les analyses ci-dessous ne fournissent pas pour l'instant des résultats probants, elles permettent de donner un aperçu des travaux réalisables. En effet, on peut facilement mener des analyses similaires en comparant d'autres chairs entre eux. De plus, afin d'obtenir une meilleure causalité dans les effets mesurés, on peut mettre des indicatrices temporelles. J'ai essayé cette méthode en faisant une indicatrice par réunion mais cela crée bien trop de variables pour avoir un résultat final robuste. Enfin, il serait intéressant de construire d'autres scores afin de mesurer l'impact du chair sur ces scores. On peut notamment penser à la création d'un score de degré académique à l'aide du dictionnaire d'Harvard présenté ci dessus.

À présent nous allons nous intéresser aux sujets sous-jacents aux réunions. Nous aimeraisons observer de la variabilité entre les sujets abordés au cours des différentes réunions, en fonction du chair.

Au regard de la p-value, le coefficient associé au chair n'est pas statistiquement différent de 0. Cela signifie qu'il n'y a pas de différences notables entre le Chairman Bernanke et le Chairman Greenspan du point de vue de l'influence sur le score d'incertitude. Cependant là encore, le taux d'intérêt ne semble pas être une bonne variable de contrôle (coefficient associé non significatif). Comme précédemment, nous allons reproduire cette régression en remplaçant le taux d'intérêt à la date t par le nasdaq à la date t. En regardant la deuxième figure de l'annexe 7.3, on s'aperçoit que le coefficient associé au chairman n'est pas significatif. Néanmoins, cette fois ci, la variable nasdaq ne semble pas être une meilleure variable de contrôle que le taux d'intérêt. Ce résultat est donc à interpréter avec précaution.

5 Topics modelling avec l'aide de l'algorithme Latent Dirichlet Analysis

L'algorithme Latent Dirichlet Analysis (LDA) a beaucoup été utilisé pour étudier les réunions du FOMC. Cependant, il n'a jamais été utilisé dans l'optique de comparer les chairs entre eux. Dans un premier temps nous allons décrire le fonctionnement de cette algorithme, puis, nous allons voir comment il peut faire émerger des topics (distincts si possible) sous-jacents aux réunions du FOMC. La première partie de cette section s'attachera au fonctionnement mathématique de l'algorithme et la deuxième partie sera une application sur nos données.

5.1 Présentation de LDA

En traitement du langage naturel, l'allocation de Dirichlet latente (ADL) permet de trouver des similitudes au sein d'un corpus de documents et par la suite de représenter un document sous forme de mélange de topics et de représenter les topics comme mélange de mots. Nous allons par la suite décrire la modélisation mathématique du modèle LDA. Pour cela nous introduisons les notations (conventionnelles) suivantes :

- V désignera la taille du vocabulaire
- M désignera le nombre de documents
- N_i désignera le nombre de mots dans le document i et ainsi $i \in \{1, \dots, M\}$
- K désignera le nombre de topics différents (hyper paramètre à choisir)
- α désignera le prior d'une loi de Dirichlet caractérisant la distribution des topics par document. On note que α est un vecteur de taille K où chaque composante est positive..
- β désignera le prior d'une loi de Dirichlet caractérisant la distribution des mots par topics. On note que β est un vecteur de taille V où chaque composante du vecteur positive.
- θ_i caractérise la distribution des topics pour le document i et sera de taille K . Là aussi chaque composante du vecteur est positive est la somme des composantes vaut 1.
- ϕ_k caractérise la distribution des mots pour le topics k ($k \in \{1, \dots, K\}$). Là aussi chaque composante du vecteur est positive est la somme des composantes vaut 1.
- $z_{i,j}$ caractérise le topic pour le j -ème mot dans le document i ($j \in \{1, \dots, N_i\}$)
- $w_{i,j}$ caractérise le j -ème mot dans le document i

On parle de modèle latent car la seule variable que nous observons effectivement est W (représentant les mots au sein des documents). Les variables latentes du modèle vont nous servir à construire un modèle d'allocations de mots. Une fois ce modèle construit il faudra remplacer les variables générées $w_{i,j}$ par les "vraies" réalisations dans les documents et maximiser la vraisemblance d'un modèle joint que nous détaillerons par la suite.

Le but final étant d'obtenir les vecteurs $\phi_k^* \forall k \in \{1, \dots, K\}$ et $\theta_i^* \forall i \in \{1, \dots, M\}$ qui désignent respectivement les distributions des topics en termes de vocabulaire et les distributions des documents en termes de topics.

Le modèle génératif est le suivant :

Algorithm 1 Modèle génératif de l'algorithme LDA

Initialiser les vecteurs α et β
 Tirer $\theta_i \sim Dir(\alpha) \forall i \in \{1, \dots, M\}$
 Tirer $\phi_k \sim Dir(\beta) \forall k \in \{1, \dots, K\}$
for Pour chaque documents i et pour chaque position de mots j ($j \in \{1, \dots, N_i\}$) **do**
 on tire un topic $z_{i,j} \sim M(\theta_i)$ (M désigne une multinomiale avec un tirage).
 on tire un mot parmi ce topic $w_{i,j} \sim M(\phi_{z_{i,j}})$
end for

Pour comprendre plus simplement cette étape on peut s'aider du graphique ci-dessus qui est souvent utilisé pour expliquer le modèle LDA.

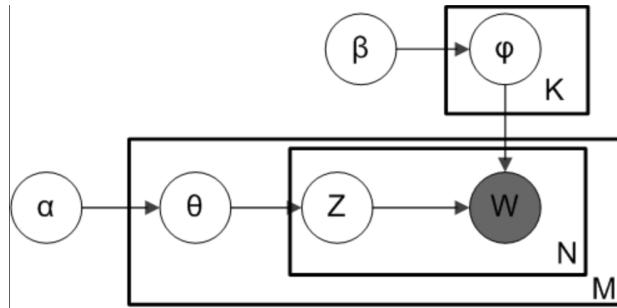


FIGURE 21 – Schéma simplifié des paramètres dans le modèle LDA

Une fois cette étape effectuée, il faut inférer correctement les paramètres du modèle (ϕ et θ). Pour ce faire, il existe plusieurs méthodes. Je vais parler rapidement de la méthode *variational bayes* qui a pour but de maximiser la vraisemblance d'une approximation de la distribution postérieure du modèle. Je parle de cette méthode car c'est cette méthode d'inférence qui est dans le package sklearn que nous allons utiliser pour appliquer l'algorithme LDA à notre corpus.

Pour faire court : dans le modèle LDA nous sommes intéressés par la distribution $P(Z, \Theta, \phi | W, \alpha, \beta)$ des variables latentes pour des paramètres α et β connu ainsi que pour des observations W révélées. Néanmoins, par la formule de Bayes on a :

$$P(Z, \Theta, \phi | W, \alpha, \beta) = \frac{P(W, Z, \Theta, \phi | \alpha, \beta)}{P(W | \alpha, \beta)}$$

La vraisemblance de cette quantité ne peut pas être maximisée à cause du dénominateur. De ce fait, le but de la méthode **variational inference** est d'approximer la distribution postérieure par une famille de distribution (facilement optimisable) dont on peut tuner les paramètres pour approcher la distribution postérieure. Nous nous tiendrons à cette définition superficielle pour rester concis.

5.2 Application de la méthode LDA à notre corpus

Avant d'appliquer la méthode LDA à notre corpus, il faut dans un premier temps le nettoyer dans le sens du traitement naturel du langage. Ainsi, on utilise notamment les packages python *nltk* et *gensim* qui sont spécialisés dans le traitement naturel du langage. Premièrement nous retirons tous les stopwords. Par exemple, les stopwords dans *nltk* sont des petits mots comme ‘*ourselves*’, ‘*hers*’, ‘*between*’, ‘*yourself*’, ‘*but*’, ‘*again*’, ‘*there*’, ‘*about*’, ‘*once*’, ‘*during*’, ‘*out*’, ‘*very*’, ‘*having*’, ‘*with*’, ‘*they*’, ‘*own*’, ‘*an*’, ‘*be*’, ‘*some*’, ‘*for*’, etc. Nous précisons que nous rajoutons des stopwords personnels en plus de ceux des packages *nltk* et *gensim*. En plus de cela nous retirons tous les caractères spéciaux du type "",","*,","?","]", "[",-,!,"?",",,"(",")", "//". On précise que ces caractères avaient déjà été retirés pour les analyses statistiques précédentes.

Après cela nous lemmatisons notre corpus afin de réduire considérablement la taille du vocabulaire, c'est-à-dire que tous les mots issus de la même famille vont être réunis au sein du même mot. Par exemple, les mots "builds", "building" et ""built" vont tous être considérés comme le mot "build". Une fois ces étapes faites nous avons un corpus de réunions nettoyé sur lequel nous pouvons appliquer l'algorithme LDA afin de faire ressortir, les sujets sous jacents aux réunions. Une fois ces opérations effectuées, on se retrouve avec un texte de la forme suivante (exemple avec un **extrait de texte du 2009/01/28**).

```
good afternoon everybody night michelle smith attended swearing tim geithner vice president biden
president obama attendance new treasury secretary dinner tonight meeting honor tim goodbye post
go new year new thing morning federal reserve bank new york announced new president bill dudley
congratulation way table followed welcome look forward working new capacity thank know governor
kroszner resigned anticipation appointment new governor dan tarullo dan cleared senate obviously
won't attending meeting leaf governor embattled think lowest number sitting governor probably long
time quality easy
```

Une fois le texte nettoyé, il est important de choisir la représentation numérique (word embedding) sous laquelle le texte sera donné en entrée à l'algorithme LDA. Dans notre cas nous allons choisir une représentation bag of word car c'est celle qui est la plus souvent utilisée avec l'algorithme LDA classique. Pour la définition du word embedding sous forme bag of word, voir annexe 7.4. En plus de choisir une représentation numérique pour le texte, il faut choisir les différents hyperparamètres de l'algorithme LDA. Voici les hyperparamètres choisis :

- Dans le word embedding (bag of word simple), on retire désormais 10% des mots les plus fréquents et 3% des mots les moins fréquents.
- Le nombre d'itération a été fixé à 100.
- Ici on fixera le nombre de topics à 10 par exemple et le nombre de top words affichés à 5.

Une fois l'algorithme entraîné, il nous est possible d'accéder aux vecteurs ϕ et θ . Dans un premier temps nous allons nous intéresser aux topics en sortie d'algorithme afin de voir si ils sont interprétables et différentiables entre eux. On regardera les cinq mots les plus probables qui les composent.

Topics in LDA model :

Topic 0 : angell jordan asymmetry compensation april
Topic 1 : velocity intervention angell lindsey laware
Topic 2 : transcript intervention angell jordan thats
Topic 3 : asia panel asian tilt equity
Topic 4 : procedure targeting johnson angell heller
Topic 5 : angell asymmetry velocity compensation intervention
Topic 6 : asymmetry swap panel tilt symmetry
Topic 7 : transcript tape edited mexico release
Topic 8 : tunnel cone upper drift virginia
Topic 9 : tilt sentence yen release black

Il est difficile de distinguer tous les topics entre eux. Cela peut notamment s'expliquer par le fait que les documents sont très similaires au niveau du vocabulaire à cause du fait que la temporalité des réunions soit codifié. Pour rappel la temporalité d'une réunion est la suivante : un rapport sur la convergence des taux est réalisé puis le chair prend la parole, puis une discussion économique s'ensuit et enfin par la suite il y a un débat. Il pourrait être intéressant de considérer uniquement la phase de débat lors de l'application de l'algorithme LDA.

De plus, **il semble pertinent de ne plus considérer les réunions comme des documents mais plutôt l'ensemble des phrases prononcées par un membre du comité durant sa carrière au sein des réunions du FOMC.** Ainsi, pour chaque personne correspondra un document désignant l'ensemble prononcés par l'individu durant sa carrière. En faisant ceci pour tous les participants aux réunions, on peut construire un nouveau corpus. En appliquant l'algorithme LDA à ce corpus, on peut faire émerger des pourcentages de topics (vecteur θ) pour les membres du comité.

Les résultats de cette approche sont présents ci-dessous. On précise que l'on s'intéresse aux mélanges de topics relatifs à chaque chair. De plus, on choisit un total de 5 topics qui seront représentés par 5/6 mots. Il est important de préciser que les 20 mots les plus probables pour chaque topics ont été affichés et que cinq mots ont été choisis de façon arbitraire pour que les topics soient interprétables. Les cinq topics sont les suivants :

Topic 0 : president market governor policy committee
Topic 1 : percent rate growth time range
Topic 2 : inflation policy market economy financial
Topic 3 : forecast prices gdp market unemployment data
Topic 4 : economy year time district people

On peut interpréter les topics ainsi :

- Le topic 0 semble relatif à **des personnes, des groupes relatifs au FOMC.**
- Le topic 1 semble relatif à **un vocabulaire quantitatif relatif à des indices financiers.**
- Le topic 2 semble relatif à **des variables financières.**
- Le topic 3 semble relatif à **des variables économiques assez globales.**
- Le topic 4 semble relatif à **des variables temporelles, spatiales**

On rappelle que la méthode LDA dans ce cas présent traite les individus comme des documents. Nous pouvons ainsi voir de quels topics se composent (en pourcentage) Chairman Burns, Chairman Miller, Chairman Volcker, Chairman Greenspan, Chairman Bernanke et Chair Yellen.

Nom du chair	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Chairman Burns	4%	94%	2%	0%	0%
Chairman Miller	5%	95%	0%	0%	0%
Chairman Volcker	0%	100%	0%	0%	0%
Chairman Greenspan	80%	2%	0%	1%	16%
Chairman Bernanke	48%	0%	52%	0%	0%
Chair Yellen	34%	0%	66%	0%	0%

TABLE 3 – Distribution des topics (arrondis à deux chiffres) pour chaque Chair

Au vu de ces résultats, on s'aperçoit que les Chairs se concentrent en très grande majorité sur les trois premiers topics. On s'aperçoit que le Chairman Burns, le Chairman Miller et le Chairman Volcker sont fortement reliés au topic 1 qui désigne un vocabulaire quantitatif relatif à des nombres (*percent, rate, growth, time, range*). Il semble que le Chairman Greenspan soit plus relié au topic 0 qui désigne un vocabulaire désignant les personnes et groupes relatifs au FOMC (*resident, market, governor, policy, committee*). Enfin, le Chairman Bernanke et la Chair Yellen semblent être reliés au topic 0 et (en majorité) au topic 2 qui désigne un vocabulaire des variables financières (*inflation, policy, market, economy, financial*).

Nous nous arrêterons ici pour l'approche LDA. On note que cette analyse peut être améliorée notamment dans la façon dont les textes sont représentés de façon numériques et donnés en entrée à l'algorithme. En effet, l'approche bag of word est limitée car elle ne tient pas compte des relations entre les mots. Une idée intéressante serait de tenir compte de la relation entre les mots dans le modèle LDA avec un word embedding du type word2vec. Pour ce faire, il existe des articles comme l'article écrit par C.Moody [2016], qui explique comment mettre en oeuvre un word2vec avec un modèle LDA.

Un concept bien plus poussé consisterait à utiliser les topics en sortie de l'algorithme LDA pour construire des mesures de conformités entre les membres de la réunion et le chair en charge de la réunion. Ainsi on pourrait étudier si le chair influence fortement ou non le comité en fonction des topics résultant de chaque individu.

6 Conclusion

6.1 Conclusion du travail de recherche

Pour rappel, le but de ce stage était de mener des analyses statistiques sur le corpus des réunions du FOMC pour le compte de Professeur Alessandro Riboni.

Tout d'abord, il a fallu récupérer les données en ligne, puis les nettoyer avant de les stocker sous forme de bases de données exploitables. Une fois ceci fait, nous avons mené des statistiques descriptives afin de mieux comprendre les données et de réussir à dégager une problématique novatrice. Pour ce faire, il a fallu faire un travail d'état de l'art en étudiant plusieurs articles qui traitent de l'analyse statistique et notamment du NLP sur les réunions du FOMC.

Une fois ceci effectué, nous avons décidé de nous intéresser à l'impact du chair sur les autres membres de la réunion et sur la teneur "démocratique" du débat. Nous avons alors mené des statistiques descriptives dans ce sens afin de dégager plusieurs pistes qui permettent de mesurer cet impact du chair sur les autres membres. Nous nous sommes notamment intéressés à la longueur des phrases, au timing, au vocabulaire utilisé et aux topics sous jacents aux réunions.

Après cette phase descriptive, nous avons tenté de mesurer l'impact "causal" du chair via des scores de sentiments construits au préalable. Pour ce, nous avons utilisé un modèle de régression linéaire avec fixed effect. Bien que cette régression apporte des premiers éléments de réponse, l'effet estimé ne semble pas causal. Néanmoins, cette modélisation permet d'ouvrir des perspectives pour de futures analyses allant dans ce sens.

Par la suite, nous avons tenté de déterminer les topics sous-jacents aux réunions. Il s'avère qu'il était plus pertinent de considérer un document (au sens de l'algorithme LDA) comme l'ensemble des phrases prononcées par un membre du comité au cours de sa carrière. Une fois ceci fait, nous avons observé des différences dans les éléments de langage employés par les différents chairs. L'étape d'après serait de réussir à construire des mesures de conformité en fonction de la distribution des topics relative à chaque individu.

À présent, je vais citer différentes pistes de travail qui me semblent intéressantes pour poursuivre cette analyse statistique des réunions du FOMC et qui pourraient être utiles au nouvel assistant chercheur du Professeur Alessandro Riboni. Voici les poursuites envisageables :

- Construire des scores de conformités à l'aide des outputs de l'algorithme LDA.
- Mener des analyses avec la régression fixed effect pour de nouveaux score et d'autres chairs.
- S'intéresser à la notion de timing dans les réunions. En effet les comportements des chairs en terme de timing et de taille d'intervention semblent bien différents. Il serait intéressant de construire un modèle permettant de capturer cela et notamment d'étudier le comportement du Chairman Greenspan.
- Automatiser l'acquisition des données de façon à ce que toute nouvelle transcription d'une réunion disponible sur le site de la FED, soit traitée et ajoutée aux bases de données.

J'aimerais expliquer dans ce paragraphe les différents apports que j'ai fourni à mon maître de stage. Je pense avoir aidé à l'avancement de ses travaux par :

- La lecture et la synthèse de différents articles de recherches relatifs à l'analyse statistique des réunions du FOMC.
- Le web-scrapping des données, la mise en forme et la construction de base de données prêtes à être exploitées
- L'implémentation des analyses statistiques qu'il souhaitait voir.
- L'initiation et la réalisation d'analyses statistiques suivies de la présentation des résultats et du raisonnement utilisé.
- Des questionnements et des propositions allant dans le sens de la problématique que nous nous étions fixée.
- En lui soumettant régulièrement des rapports écrits sur mes avancées.

6.2 Conclusion personnelle

En juillet 2020, j'ai débuté un stage de fin d'études dans le domaine du conseil. Néanmoins, j'ai rapidement vu que cela n'allait pas me convenir et que le contenu scientifique et intellectuel proposé était inadéquat avec les objectifs du stage de fin d'études. J'ai décidé de quitter ce stage et c'est alors que j'ai vu que le professeur Alessandro Riboni cherchait des étudiants en statistique pour analyser les réunions du FOMC.

Je suis reconnaissant envers Alessandro Riboni et le remercie de m'avoir permis de mener un projet de recherche à l'intersection de la statistique et de l'économie. Ce stage était stimulant par son besoin en prise d'initiative et en inventivité. Le fait que le Professeur Riboni soit un professeur d'économie et non de statistique m'a poussé à essayer d'expliquer de façon concise et claire les méthodes que je souhaitais mettre en place. Par ailleurs, ses connaissances en macroéconomie m'ont aidé à comprendre les résultats que j'obtenais et ont facilité les interprétations. J'estime que l'intérêt principal de ce stage vient de sa variété. En effet, j'ai pu programmer, m'intéresser à de nouveaux concepts mathématiques que je ne connaissais pas (notamment pour l'algorithme LDA), lire et étudier des articles récents et novateurs dans leur domaine et j'ai dû me familiariser avec une culture économique et financière propre à la banque centrale américaine.

Enfin, ce stage m'a permis de mûrir ma réflexion quant à ma future vie professionnelle. Pour mon stage d'application de deuxième année, j'avais déjà été assistant chercheur pour le compteur du Docteur Matteo Fasiolo à l'université de Bristol. J'avais pu avoir un bon aperçu des différentes étapes jalonnant un projet de recherche. Après un rapide passage dans le monde du conseil, j'ai une nouvelle fois réalisé un stage en tant qu'assistant chercheur. Je trouve que le monde de la recherche est stimulant pour plusieurs raisons. Premièrement, la réflexion mise en place est réelle et profonde, elle n'est pas conditionnée à une garantie de résultat immédiat. Deuxièmement, l'autonomie laissée au chercheur lui permet de travailler sur un sujet qu'il affectionne et pour lequel il est libre d'exploiter pleinement ses idées. Enfin, l'apprentissage y est constant. Pour ces raisons, je souhaiterais poursuivre mes études par un PhD. Je suis actuellement dans un processus de recherche d'offre de thèse et je suis en contact avec des laboratoires.

7 Annexes

7.1 Aperçu la temporalité des réunions de Burns, Miller et Volcker

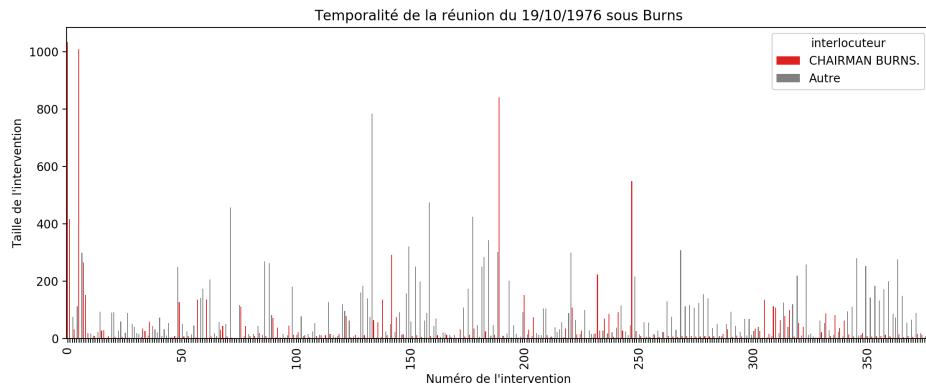


FIGURE 22 – temporalité d'une réunion sous Burns

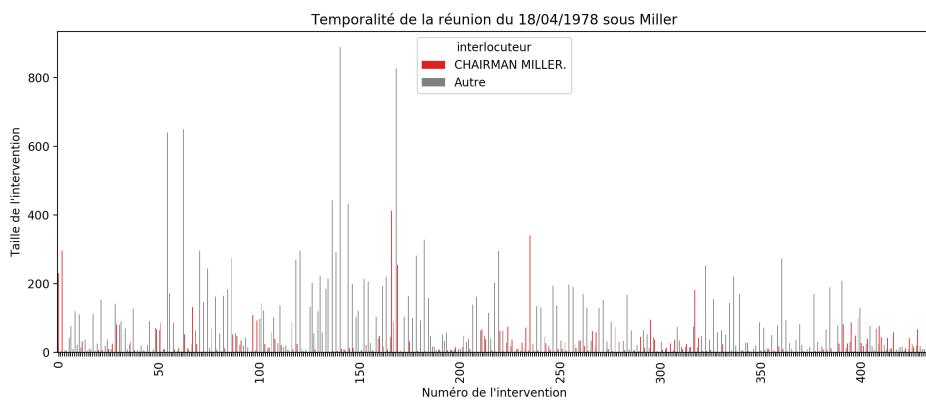


FIGURE 23 – temporalité d'une réunion sous Miller

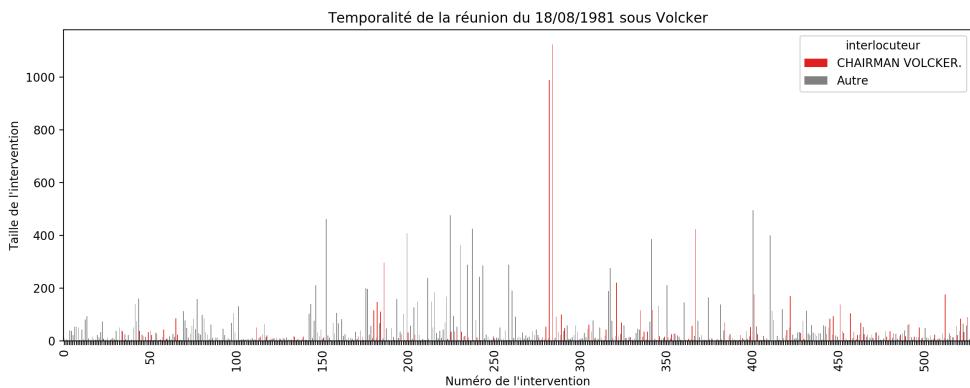


FIGURE 24 – temporalité d'une réunion sous Volcker

7.2 Graphiques score de sentiments

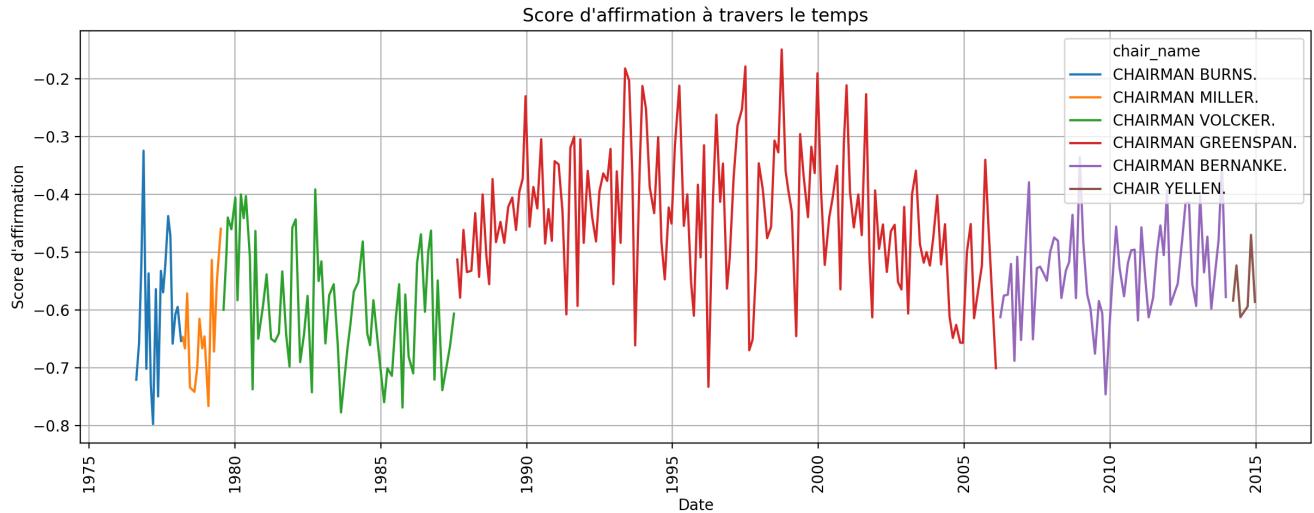


FIGURE 25 – Score global d'incertitude

Il est difficile d'observer une tendance dans le score d'affirmation néanmoins les valeurs qu'il prend nous informe de quelque chose. En effet, les StrongModal (du type Always, Best, Clearly, ...) sont au nombre de 20 et les WeakModal (du type Almost, Apparently, Appeared) sont au nombre de 28). Les classes sont donc plutôt équilibrés néanmoins le score est clairement négatif. Cela nous informe que les membres des meetings ont tendance à employer un vocabulaire "flou" et "sans affirmation" plutôt que des termes "tranchants" et "affirmés".



FIGURE 26 – Indice du Dow Jones sur la période 2000-2019

7.3 Régression fixed effect avec pour variable de contrôle la valeur du nasdaq au jour de la réunion

Coefficients:		Coefficients:	
	Estimate Std. Error t value Pr(> t)		Estimate Std. Error t value Pr(> t)
(Intercept)	-2.382e-01 3.787e-02 -6.291 3.87e-10 ***	(Intercept)	9.281e-03 1.204e-03 7.709 1.99e-14 ***
ChairCHAIRMAN GREENSPAN.	9.966e-02 1.839e-02 5.420 6.69e-08 ***	ChairCHAIRMAN GREENSPAN.	8.762e-04 5.846e-04 1.499 0.134089
nasdaq_value	-2.303e-06 7.426e-06 -0.310 0.75647	nasdaq_value	9.565e-07 2.361e-07 4.052 5.28e-05 ***
NameMR. GUYNN.	-2.486e-02 4.564e-02 -0.545 0.58604	NameMR. GUYNN.	8.460e-03 1.451e-03 5.830 6.44e-09 ***
NameMR. HOENIG.	1.229e-01 3.861e-02 3.184 0.00147 **	NameMR. HOENIG.	6.210e-03 1.227e-03 5.060 4.59e-07 ***
NameMR. KAMIN.	-2.151e-01 7.254e-02 -2.965 0.00307 **	NameMR. KAMIN.	6.102e-03 2.306e-03 2.646 0.008217 **
NameMR. KOHN.	-1.948e-01 3.804e-02 -5.120 3.35e-07 ***	NameMR. KOHN.	1.077e-02 1.209e-03 8.904 < 2e-16 ***
NameMR. KOS.	-8.299e-02 5.516e-02 -1.505 0.13259	NameMR. KOS.	6.230e-03 1.754e-03 3.553 0.000390 ***
NameMR. LACKER.	-3.675e-02 4.728e-02 -0.777 0.43707	NameMR. LACKER.	6.425e-03 1.503e-03 4.275 2.00e-05 ***
NameMR. MADIGAN.	-1.135e-02 6.001e-02 -0.189 0.84996	NameMR. MADIGAN.	1.420e-02 1.998e-03 7.445 1.43e-13 ***
NameMR. MEYER.	-1.474e-01 5.506e-02 -2.678 0.00747 **	NameMR. MEYER.	1.341e-02 1.750e-03 7.661 2.86e-14 ***
NameMR. MOSKOW.	1.144e-02 4.324e-02 0.265 0.79129	NameMR. MOSKOW.	6.097e-03 1.374e-03 4.436 9.66e-06 ***
NameMR. OLSON.	1.261e-01 6.156e-02 2.048 0.04068 *	NameMR. OLSON.	6.287e-03 1.957e-03 3.212 0.001337 **
NameMR. POOLE.	-8.008e-02 4.649e-02 -1.723 0.08512 .	NameMR. POOLE.	8.616e-03 1.478e-03 5.829 6.47e-09 ***
NameMR. REIFSCHEIDER.	-1.956e-01 7.011e-02 -2.790 0.00532 **	NameMR. REIFSCHEIDER.	4.469e-03 2.229e-03 2.005 0.045102 *
NameMR. REINHART.	-1.565e-01 5.585e-02 -2.802 0.00513 **	NameMR. REINHART.	1.047e-02 1.775e-03 5.895 4.40e-09 ***
NameMR. SACK.	-5.950e-02 7.134e-02 -0.834 0.40438	NameMR. SACK.	1.123e-02 2.268e-03 4.953 7.94e-07 ***
NameMR. SHEETS.	-1.157e-01 6.714e-02 -1.724 0.08493 .	NameMR. SHEETS.	7.694e-03 2.134e-03 3.605 0.000320 ***
NameMR. SLIFMAN.	-1.233e-01 7.098e-02 -1.737 0.08250 .	NameMR. SLIFMAN.	6.542e-04 2.257e-03 0.290 0.771923
NameMR. STERN.	7.380e-02 3.828e-02 1.928 0.05405 .	NameMR. STERN.	1.033e-02 1.217e-03 8.491 < 2e-16 ***
NameMR. STOCKTON.	-1.331e-01 4.234e-02 -3.144 0.00169 **	NameMR. STOCKTON.	1.131e-02 1.346e-03 8.405 < 2e-16 ***
NameMR. WILCOX.	-1.044e-01 5.820e-02 -1.794 0.07290 .	NameMR. WILCOX.	1.476e-02 1.850e-03 7.977 2.50e-15 ***
NameMR. WILLIAMS.	-7.427e-02 7.249e-02 -1.025 0.30570	NameMR. WILLIAMS.	7.565e-03 2.305e-03 3.282 0.001047 **
NameMS. BIES.	-4.887e-02 5.833e-02 -0.838 0.40222	NameMS. BIES.	4.693e-03 1.854e-03 2.531 0.011451 *
NameMS. DANKER.	6.311e-01 5.159e-02 12.232 < 2e-16 ***	NameMS. DANKER.	3.842e-03 1.640e-03 2.342 0.019259 *
NameMS. JOHNSON.	-3.723e-02 4.915e-02 -0.757 0.44885	NameMS. JOHNSON.	5.231e-03 1.562e-03 3.348 0.000829 ***
NameMS. MINEHAN.	-7.478e-02 4.294e-02 -1.742 0.08175 .	NameMS. MINEHAN.	8.729e-03 1.365e-03 6.394 2.01e-10 ***
NameMS. PIANALTO.	2.415e-02 4.516e-02 0.535 0.59287	NameMS. PIANALTO.	1.026e-02 1.436e-03 7.144 1.26e-12 ***
NameMS. SMITH.	3.920e-01 8.366e-02 4.686 2.97e-06 ***	NameMS. SMITH.	-2.638e-03 2.660e-03 -0.992 0.321295
NameMS. YELLEN.	-9.903e-02 4.385e-02 -2.258 0.02404 *	NameMS. YELLEN.	1.456e-02 1.394e-03 10.447 < 2e-16 ***
NameVICE CHAIRMAN GEITHNER.	1.021e-01 5.921e-02 1.725 0.08474 .	NameVICE CHAIRMAN GEITHNER.	2.368e-02 1.882e-03 12.578 < 2e-16 ***
---		---	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	
Residual standard error: 0.3237 on 1991 degrees of freedom		Residual standard error: 0.01029 on 1991 degrees of freedom	
Multiple R-squared: 0.1886, Adjusted R-squared: 0.1763		Multiple R-squared: 0.1564, Adjusted R-squared: 0.1437	
F-statistic: 15.42 on 30 and 1991 DF, p-value: < 2.2e-16		F-statistic: 12.31 on 30 and 1991 DF, p-value: < 2.2e-16	

FIGURE 27 – Régression fixed effet sur le score de positivité avec pour variable de contrôle la valeur du nasdaq à la place du taux d'intérêt

FIGURE 28 – Régression fixed effet sur le score d'incertitude avec pour variable de contrôle la valeur du nasdaq à la place du taux d'intérêt

7.4 Explication de bag of words

Cette méthode consiste à représenter un texte par un vecteur donnant les mots qu'ils contient. Pour un ensemble de mots représenté par un vecteur $\Theta = (\text{This}, \text{meeting}, \dots, \text{is})$ de longueur égale au nombre de mots existant dans le corpus considéré, une phrase P est représentée par un vecteur p de même taille que Θ tel que la i -ème composante de p soit égale au nombre de fois où le i -ème mot de Θ apparaît dans la phrase P .

Exemple : $\Theta = (\text{This}, \text{meeting}, \text{last}, \text{answer}, \text{Yellen}, \text{rate}, \text{FOMC}, \text{Sniderman}, \text{economics}, \text{Mark}, \text{is})$ et $P = \text{"This is Mark Sniderman's last FOMC meeting"}$

Alors $p = (1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1)$

Cette méthode a l'avantage de fournir une représentation simple et rapide à implémenter. Elle permet aussi de représenter la redondance de certains mots ou marques de ponctuations dans une phrase (exemple : un seul ! ou 3 !!!).

Cependant, on perd l'information portant sur le sens et la similarité des différents mots ainsi que sur leur ordre d'apparition dans la phrase. Cette information constitue un élément important dans la compréhension d'une phrase. En effet, on voit rapidement que sous cette représentation, étant donnés deux mots x et y tel que $x \neq y$, on a $x \cdot y = 0$. Cela implique notamment que toutes les paires de mots ont une similarité nulle, puisqu'ils sont tous orthogonaux deux à deux. Cela semble être très contraignant et peu réaliste. En effet, il est par exemple souhaitable que les mots *good* et *great* ne soient pas orthogonaux.

Références

- A.Prat S.Hansen, M.Mcmahon. Transparency and deliberation within the fomc : A computational linguistics approach. *The Quarterly Journal of Economics*, 2014.
- A.Shapiro and D.Wilson. Taking the fed at its word :a new approach to estimating central bank objectives using text analysis. *Federal Reserve Bank of San Francisco*, 2019.
- Henry Otuadinma Jagdish Chhabria Alexander Ng, Arun Reddy. *Extracting insights from FOMC statements*, 2019. URL <https://rpubs.com/jackv13/extracting-insights-from-fomc-statements>.
- C.Moody. Mixing dirichlet topic models and word embeddings to make lda2vec. *Online access only*, 2016.

Note de synthèse : stage de fin d'études

**Sujet du stage : Analyse statistique des transcriptions des réunions du FOMC
(Federal Open Market Committee)**

Étudiant : Etienne Le Naour

Organisme d'accueil : CREST

Ce stage a été effectué au Centre de Recherche en Économie et Statistiques (CREST) du 07/07/2020 au 09/10/2020 sous la supervision du professeur Alessandro Riboni. Il s'inscrit dans le cadre du stage de fin d'études de troisième année de l'ENSAE et compte également pour le Master data science.

1. Introduction et problématique

Le Federal Open Market Committee (FOMC) désigne un organe de la réserve fédérale américaine (FED). Cet organe est en charge des opérations d'achat/vente des bons du trésor américain (opérations d'open market) qui constituent la principale marche de manœuvre de la politique monétaire des États-Unis. Les membres du FOMC se réunissent environ 7 à 8 fois par an dans le but de décider de la politique monétaire à mener. Le FOMC possédant un impact important sur l'économie américaine, il est primordiale de comprendre les processus sous-jacent aux décisions prises par les membres du comité des FOMC. Par chance, durant ces réunions, l'ensemble des phrases prononcées par les participants sont retranscrites à l'écrit. Ainsi, à l'issue de la réunion, l'intégralité des échanges à été consignée et est rendue publique par la suite.

L'objectif de ce stage est de mener des analyses statistiques sur les transcriptions des réunions du FOMC. Plus précisément, nous aimerions déterminer si le déroulé de la réunion est bien démocratique et si les échanges entre les différents acteurs sont égalitaires. Nous nous demandons notamment si le directeur du FOMC n'a pas trop d'influence sur la réunion.

2. Déroulement du stage

Le stage a été jalonné par les étapes suivantes :

1. Récupérer les transcriptions sur le site internet du FOMC, nettoyer les données puis les stocker sous forme de base de données exploitables.
2. Réaliser des statistiques descriptives dans l'optique de comprendre des mécanismes sous-jacents aux données et de dégager une problématique novatrice.
3. Faire un état de l'art de l'analyse statistiques des transcriptions des réunions du FOMC. On s'est ainsi aperçu que des chercheurs avaient travaillé sur le mécanisme de prise de décision au cours de la réunions ainsi que sur l'impact de la transparence des réunions sur le comportements des membres de la réunion. C'est à ce moment que nous avons décidé d'étudier l'influence du directeur de la réunion sur les autres membres.

4. Modéliser les données de façon à répondre à la problématique. Nous avons notamment effectué les réalisations suivantes :

- regardé le nombre de mots prononcés par les directeurs relativement aux autres membres de la réunion. Nous avons observé cette statistique dans le temps et nous constatons que les directeurs semblent prendre plus la parole au moment des crises économiques.
 - étudié la temporalité au sein des réunions en regardant les moments où le directeur prenait la parole et en quelle proportion (en terme de nombre de mots) il la prenait. On a ainsi observé que pour les différents chairs les comportements n'étaient pas similaires et que certains semblaient plus organisateurs de la réunion quand d'autres semblaient plus décisionnaires.
 - construit des scores de sentiments (score de positivité, score d'incertitude,...) en rapport avec le vocabulaire employé au cours de la réunion. Nous avons ainsi étudié leur évolution dans le temps et pour les différents directeurs.
 - représenté l'ensemble des réunions comme un ensemble de topics grâce à l'algorithme LDA. Après avoir identifier et expliquer clairement les topics, nous avons pu caractériser chaque directeur du FOMC comme un mélange de topics. Ceci nous permet de caractériser un directeur et d'expliquer son comportement.
5. Estimer l'effet causal de la présence du directeur Greenspan ou du directeur Bernanke (deux directeurs successifs du FOMC) sur les scores de sentiments des autres membres de la réunion. Nous avons prouvé que le fait que ce soit l'un ou l'autre influence l'emploi d'un vocabulaire positif mais n'impacte pas l'emploi d'un vocabulaire relatif à l'incertitude.
6. Donner des pistes de réflexion pour pouvoir poursuivre les analyses menées.
7. Fournir des bases de données (et des codes) propres contenant les différents scores calculés durant ces analyses pour que mon maître de stage puisse poursuivre ses travaux de manière autonome.

3. Un point sur les connaissances, approches (intellectuelles et pratiques) et compétences acquises

Au cours de ce stage j'ai amélioré mes compétences en lecture d'articles de recherche. Je suis désormais capable d'extraire l'information importante plus rapidement. Ensuite, j'ai développé mes compétences en programmation dans plusieurs domaines : récupérer des données en ligne et les mettre en forme de base de données exploitables, implémenter des algorithmes de machine learning. De plus, j'ai essayé de faire preuve d'initiative en apportant des analyses complémentaires à celles qui m'étaient commanditée par mon maître de stage.

4. Conclusion

Je suis reconnaissant envers Monsieur Riboni et le remercie de m'avoir permis de mener un projet de recherche à l'intersection de la statistique et de l'économie. Ce stage était stimulant par son besoin en prise d'initiative et en inventivité. J'estime que l'intérêt principal de ce stage vient de sa variété. En effet, j'ai pu programmer, m'intéresser à de nouveaux concepts mathématiques que je ne connaissais pas (notamment pour l'algorithme LDA), lire et étudier des articles récents et novateurs dans leur domaine. Et enfin, j'ai su me familiariser avec une culture économique et financière propre à la banque centrale américaine.

Internship's summary analysis : end of studies internship

Topic : Statistical analysis of FOMC (Federal Open Market Committee) meeting transcripts.

Student : Etienne Le Naour

Research Center : CREST

This internship was carried out at the Center for Research in Economics and Statistics (CREST) from 07/07/2020 to 09/10/2020 under the supervision of Professor Alessandro Riboni. This project counts for the end-of-studies internship at ENSAE and for the Master Data Science.

1. Introduction and problematic

The Federal Open Market Committee (FOMC) is an organ of the U.S. Federal Reserve (FED). This committee is in charge of buying and selling US treasury bills (open market operations), which are the main instrument of US monetary policy. FOMC members meet about 7-8 times a year to decide on monetary policy. Because the FOMC has a significant impact on the U.S. economy, it is important to understand the processes underlying the decisions made by the FOMC committee members. Luckily, during these meetings, all of the sentences spoken by the participants are transcribed in a document. Thus, at the end of the meeting, the entire discussion is recorded and subsequently made public.

The purpose of this internship is to conduct statistical analysis of FOMC meeting transcripts. More specifically, we would like to determine whether the meeting proceedings are democratic and whether the exchanges between the different actors are egalitarian. In particular, we would like to determine whether the FOMC director has too much influence on the meeting.

2. The Internship process

The internship was marked by the following steps :

1. Retrieve transcripts from the FOMC website, clean up the data and store it as a usable database.
2. Produce descriptive statistics in order to understand the mechanisms underlying the data.
3. Provide a state of the art statistical analysis of FOMC meeting transcripts. It was found that researchers had been working on the decision making mechanism during the meeting as well as the impact of meeting transparency on the behaviour of meeting members. It was at this point that we decided to study the influence of the meeting director on other members.
4. Model the data in order to answer the problem. In particular, we have carried out the following projects :

- looked at the number of words spoken by the chairs in relation to the other members of the meeting. We have observed this statistic over time and we see that chairs seem to speak more at times of economic crises.
 - studied temporality within meetings by looking at when the chair spoke and in what proportion (number of words relative to the total number of words). It was observed that for the different chairs, the behaviors were not similar.
 - constructs feeling scores (positivity score, uncertainty score,...) in relation to the vocabulary used during the meeting. We studied their evolution over time and for the different chairs.
 - represented all meetings as a set of topics thanks to the LDA algorithm. After identifying and clearly explaining the topics, we were able to characterize each FOMC chair as a mixture of topics. This allows us to characterize a chair and explain his behavior.
5. Estimate the causal effect of Chairman Greenspan's or Chairman Bernanke's presence (two successive FOMC chairman) on the feelings scores of other meeting members. We have shown that either influences the use of positive vocabulary but does not impact the use of vocabulary related to uncertainty.
 6. Provide ideas for future analysis to be conducted
 7. Build databases (and codes) containing the different scores calculated during these analyses so that my tutor can continue his work autonomously.

3. Acquired skills

During this internship I improved my skills in reading research articles. I am now able to extract important information more quickly. Then, I developed my programming skills in several areas: retrieving data online and formatting it into a usable database, implementing machine learning algorithms. In addition, I tried to show initiative by bringing complementary analyses for my supervisor.

4. Conclusion

I am grateful to Mr. Riboni and thank him for allowing me to conduct a research project at the intersection of statistics and economics. This internship was stimulating because by its need for initiative and inventiveness. I believe that the main interest of this internship comes from its variety. Indeed, I was able to program, to be interested in new mathematical concepts that I did not know (especially for the LDA algorithm), to read and study recent and innovative articles in their field. And finally, I was able to familiarize myself with an economic and financial culture specific to the American central bank.