ANALYSE GÉOMÉTRIQUE TEXTUELLE D'UNE QUESTION OUVERTE ISSUE DU BAROMÈTRE DE LA CONFIANCE EN 2015 ET 2016

Mireille Gettler Summa[1] & Frédérik Cassor[2] & Brigitte le Roux[3]
[1] *CEREMADE CNRS, Université Paris Dauphine, summa@ceremade.dauphine.fr*
[2] *CEVIPOF CNRS, Sciences-Po Paris, frederik.cassor@sciencespo.fr*
[3] *CEVIPOF CNRS, Sciences-Po Paris, MAP5, Université Paris Descartes, brigitte.leroux@mi.parisdescartes.fr*

Résumé. L'analyse systématique des questions ouvertes pour décrire un comportement électoral est devenue opérationnelle grâce à des mathématiques, des algorithmes et des logiciels dédiés aux données textuelles

Nous décrivons dans ce travail l'intérêt de l'Analyse Géométrique Textuelle, et en particulier de la visualisation des résultats, ainsi que l'intérêt de l'Analyse Spécifique de Classe pour l'étude de groupes d'individus.
Nous présentons également une méthodologie de traitements informatiques conçue en 2017 pour ces approches statistiques

Un exemple de cette méthodologie est mise en œuvre sur des verbatim provenant d'enquêtes initiées par le CEVIPOF (centre de recherches politiques de Sciences Po Paris), nous en donnerons les principaux résultats.

Mots-clés. Analyse Géométrique Textuelle, Analyse Spécifique de Classe, Question Ouverte, Visualisation, Verbatim Politique

Abstract. Open ended questions were introduced to study elector opinion and political vote intentions. The present presentation focuses on Textual Geometric Analysis and its related visualisation potentiality. Ad hoc formulas for Class Specific Analysis are demonstrated in the context of Correspondence Analysis on a contingency table, thus extending its application from Multiple Correspondence Analysis. We process data coming up from the Confidence Barometer of the CEVIPOF Sciences-Po Paris laboratory, an open ended question carried on for two years, in 2015 and 2016. All results are obtained through an adequate software that integrates the new results of Textual Geometric Analysis since the beginning of 2017.

Keywords. Textual Geometric Analysis, Class Specific Analysis, Open Ended Question, Visualisation, Political Verbatim

1. Introduction

L'analyse systématique des questions ouvertes pour décrire un comportement électoral est devenue opérationnelle grâce à des mathématiques, des algorithmes et des logiciels dédiés aux données textuelles [2, 4].
Nous décrivons dans ce travail l'intérêt de l'Analyse Géométrique Textuelle, et en particulier de la visualisation des résultats, ainsi que l'intérêt de l'Analyse Spécifique de Classe pour l'étude de groupes d'individus.
Nous présentons également une méthodologie de traitements informatiques conçue en 2017 pour ces approches statistiques
Un exemple de cette méthodologie est mise en œuvre sur des verbatim provenant d'enquêtes initiées par le CEVIPOF (centre de recherches politiques de Sciences Po Paris), nous en donnerons les principaux résultats.

2. Analyse Géométrique Textuelle

2.1 Analyse Géométrique

Donner un cadre géométrique en analyse et synthèse de données [5] permet de démontrer des propriétés plus élégamment, mais aussi de s'affranchir d'artefacts intrinsèques au contexte exclusivement quantitatifs ou qualitatifs.

Dans le cadre d'enquêtes qui contiennent à la fois des questions fermées et des questions ouvertes, on construit divers tableaux de données :

- en juxtaposant les questions fermées aux variables textuelles,
- en croisant dans des tableaux lexicaux de contingence mots, segments, mots clés,
- en faisant varier le choix des éléments variables actifs et supplémentaires des analyses

etc.

Les axes et plans principaux d'Analyse des Correspondances (AC) de ces tableaux sont ensuite calculés, visualisés et interprétés. On détermine ensuite des profils de répondants ou de mots à l'aide de classifications euclidiennes (avec la métrique de l'AC) et on visualise les classes d'individus ainsi obtenues sur les plans principaux, avec leurs ellipses de concentration.
On étudie également des groupes d'individus spécifiques (une classe d'âge,…) en référence à l'ensemble global. C'est l'analyse spécifique de classe (CSA) qui détermine les axes et variables principales du groupe en ne changeant ni les distances entre individus ni leurs poids, en bref, on analyse un sous-nuage du nuage global d'individus. Pour ce faire, nous avons établi pour l'AC des formules analogues à celles pour la CSA dans le cadre de l'ACM [3].

2.2 Spécificité des données textuelles

Dans les corpus textuels, comme dans toute analyse de données, le choix de descripteurs doit être fait en fonction de la problématique: les choix dans un but de détection de plagiat ne seront pas les même que ceux faits dans un but purement sémiotique, de pragmatique linguistique, sémantique etc. De plus, la phase de recodage est essentielle; elle implique un travail de lemmatisation adaptée aux questions d'intérêt :

- inclusion ou non des connecteurs (indispensables par exemple en pragmatique linguistique mais exclus parce que qualifiés de 'mots vides' dans d'autres contextes)
- intégration des taxonomies a priori lexicales de mots, segments, etc., du domaine pour faire face aux variations des signifiants en rapport aux référents et aux signifiés et de leurs diverses déclinaisons.
- prise en charge d'ontologie des domaines pour labelliser au mieux les éléments initiaux
- traque des poids faibles parasites des analyses géométriques
- repérage et séparation éventuelles de signes identiques selon leurs fonctions grammaticales

etc.

La phase de prétraitement en analyse de données textuelles est tout à fait cruciale: elle va du nettoyage orthographique à la lemmatisation, et permet de sélectionner les données qui seront analysées et visualisées.

3. Étude d'une question verbatim extraite du « Baromètre de la confiance politique » du CEVIPOF (Sciences Po Paris)

3.1 Les données

Les données analysées ici proviennent d'enquêtes[1] initiées par le CEVIPOF[2] (centre de recherches politiques de Sciences Po Paris), qui portent sur la confiance des Français en ce qui concerne le monde politique (les institutions, les acteurs politiques, les entreprises, etc.) ainsi que la confiance dans les relations sociales, etc. Ces enquêtes ont eu lieu chaque année depuis 2009 [6] . Les échantillons (d'environ 1500 à 2000 personnes) sont représentatifs des électeurs Français; ils ont été constitués par la méthode des quotas, au regard des critères de sexe, d'âge et de catégorie socio-professionnelle, après stratification par région de résidence et de taille de la commune.

On se propose dans cette communication d'analyser les réponses à une question libre sur les raisons de la perte de confiance en François Hollande parmi ceux qui ont répondu positivement à la question : « Vous aviez confiance en lui, mais vous l'avez perdue» dans les enquêtes de 2015 et 2016.

Le tableau de données, contient ainsi 875 individus et six questions fermées (le sexe, l'âge, le diplôme, la proximité partisane, les votes aux deux tours de l'élection présidentielle 2012) et une question ouverte sur les raisons de la perte de confiance en François Hollande.

3.2 Lemmatisation

La lemmatisation est opérée à l'aide du logiciel SPAD (version 9 de 2017). Cet outil [7] présente l'intérêt d'une lemmatisation riche en paramétrages automatiques et également ouverte à une semi automatisation, ce qui permet une adaptation contextuelle et experte des choix effectués.

Le nombre de « mots » distincts est de 1327, on en retiendra après une première épuration automatique, 647 pour la lemmatisation.
La recherche automatique de sous ensemble de mots contigus répétés avec certains seuils de longueur et de fréquence permet de détecter des segments fréquents: augmentation des impôts et du chômage (50 occurrences), promesses non tenues (104), état de la France (15), politique étrangère de la France (8), politique de droite (19), etc.
On a retenu douze taxons généralisant des sous-ensembles de mots qu'on ré-labellise: promesses non tenues, politique de droite, mollesse, mariage pour tous, loi travail, livre président ne devrait pas dire ça, insécurité, Europe, déchéance de nationalité, menteur, baisse du pouvoir d'achat, augmentation du chômage et des impôts.

3.3 Les traitements statistiques

Après étude des statistiques élémentaires, on a procédé à la construction d'un tableau individu-variables ou les variables sont de deux types : d'une part les six questions fermées et d'autre part les mots segments ré-labellisés issus de la question ouverte.

On procède alors à des analyses de correspondance, suivies de classifications, d'où la visualisation des segments lexicaux retenus, des classes de la classification, des parangons de classes sur les plans principaux.

---

[1] Les enquêtes ont été réalisées par l'institut Opinion Way en utilisant le système CAWI (Computer Assisted Web Interview).

[2] Cf. le site web http://www.cevipof.com/fr/le-barometre-de-la-confiance-politique-du-cevipof/ pour une présentation du contexte de l'étude et de ses principaux résultats d'analyse

En utilisant l'analyse spécifique de classe (CSA), on positionne ensuite des sous-groupes d'intérêt (catégorie socioprofessionnelle, préférence partisane, etc.) sur ces mêmes plans.
On met en individus actifs la vague de l'année 2016, la vague 2015 est donc supplémentaires dans l'analyse. Les variables actives sont les douze taxons, les variables supplémentaires sont les questions questions fermées et l'ensemble des mots retenus après lemmatisation.

3.4 Les résultats de l'AC de la vague 8, 2016

On ne présente dans ce texte qu'un premier graphique, celui du premier plan principal de l'analyse des correspondances. Cette AC du tableau croisant le sous-ensemble des individus de l'enquête ayant répondu à la question ouverte et les taxons précités (en éléments supplémentaires les six variables catégorisées, les questions fermées, et les mots initiaux, avant segmentation et après épuration) conduit aux résultats suivants.

Le premier axe oppose deux types de perte de confiance en François Hollande :

- à gauche du premier axe, le taxon 'mollesse'. On y trouve aussi les modalités suivantes : âge supérieur à 65 ans, retraités, études supérieures, proximité partisane UDI,
- sur la droite du premier axe les thèmes de 'loi travail' et 'mariage pour tous', puis plus à l'extrémité, 'insécurité', 'Europe', 'baisse du pouvoir d'achat', 'augmentation du chômage et des impôts'. On y trouve aussi les modalités suivantes: front national, classe d'âge 35-50 ans, employés, diplômes CAP-BEP
- Le taxon 'menteur' est proche du centre de gravité, marque plutôt un consensus parmi les répondants et non un clivage

Sur le deuxième axe, on trouve une opposition entre:

- en haut de l'axe, les concepts 'A mené une politique de droite', 'Promesses non tenues', avec pour modalités supplémentaires 'abstention aux dernières élections, aux deux tours' et préférence partisane 'parti communiste', hommes.
- contre en bas de l'axe les événements 'Déchéance de nationalité' et, à l'extrémité , parution du livre 'Un président ne devrait pas dire ça', avec en modalité supplémentaire, femmes

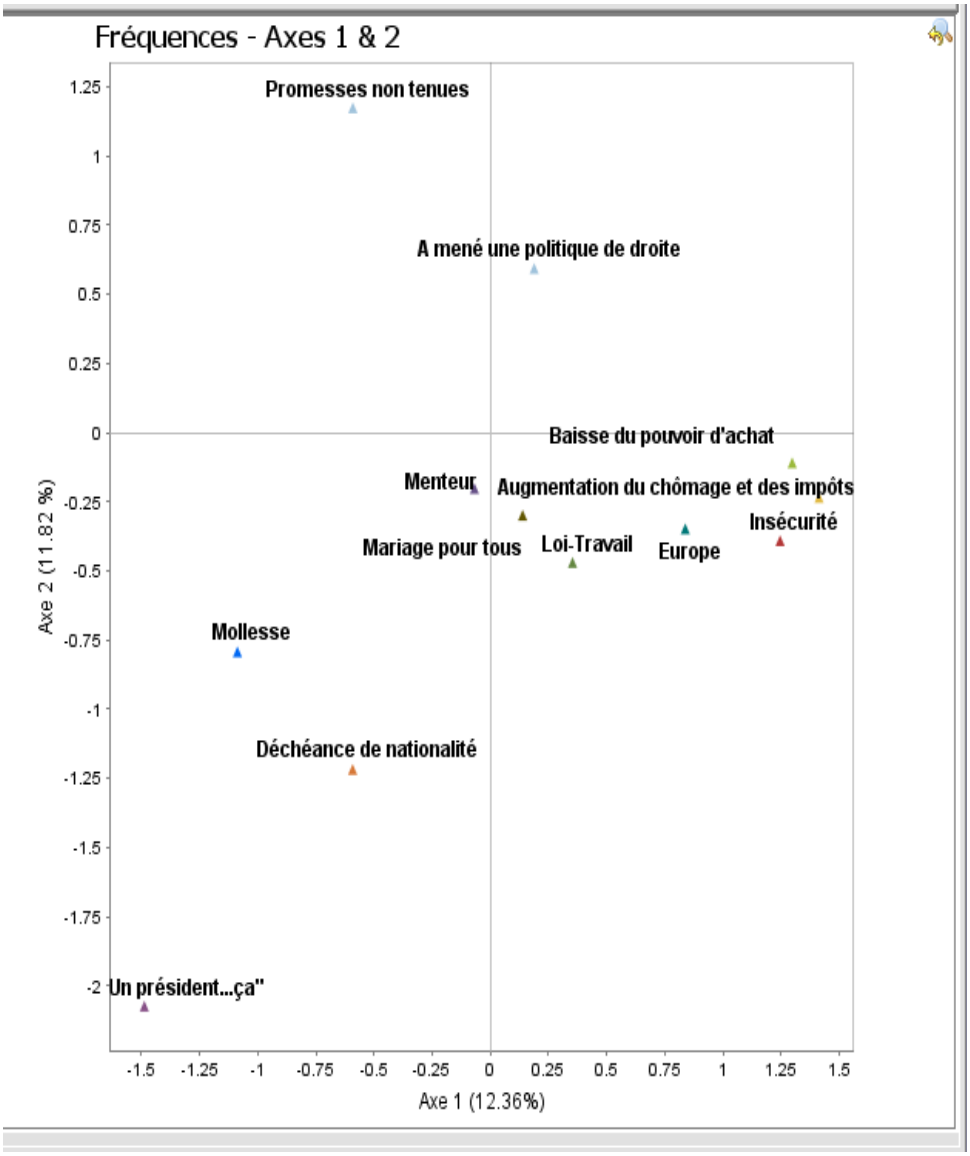

Figure 1

Premier plan principal Analyse des Correspondances fréquences actives Vague 8

On présentera en exposé in extenso l'ensemble des traitements et des résultats, classifications et analyse spécifiques incluses.

Références

[1] Baromètre de la Confiance (2016) http://www.cevipof.com/fr/le-barometre-de-la-confiancepolitique-du-cevipof/

[2] Benzécri J.-P. (1989). Essai d'analyse des notes attribuées par un ensemble de sujets aux mots d'une liste, Les Cahiers de l'Analyse des Données, vol. XIV, 1, 73-88.

[3] Chiche J. et Le Roux B., Développements récents en analyse des correspondances multiples. Revue MODULAD, n°42 (p.110-117).

[4] Lebart L. and Salem A. Statistique textuelle, Dunod, Paris (1994)

[5] Le Roux B., Analyse géométrique des données multidimensionnelles Dunod Paris (2014)

[6] Le Roux, B and Perrineau P. (2011). Les différents types d'électeurs au regard de différents types de confiance, *Les Cahiers du CEVIPOF*, http://www.cevipof.com/fr/les-publications/les-cahiers-du-cevipof/, 54, 5-35.

[7] SPAD9-Coheris (2017), www.coheris.com/analytics/logiciel-data-mining/