Analyse de questionnaire avec questions ouvertes et facteurs structurants sur le ressenti de la révolution tunisienne de 2011 dans le milieu universitaire

Sadika Rjiba[1], Mireille Gettler Summa[2], Saloua Benammou[3]

[1] Faculté des Sciences Economiques et de Gestion de Sousse – rjibasadika@yahoo.fr
[2] CEREMADE, Université Paris Dauphine – summa@ceremade.dauphine.fr
[3] Faculté des Sciences Economiques et de Gestion de Sousse – Saloua.benammou@yahoo.fr

Abstract

We present in this work the contribution of both Correspondence Analysis and Clustering to the analysis of a questionnaire containing several open-ended questions together with a set of closed questions, in the context of Structured Data. These questionnaires have been administered to a sample of students of a Tunisian University. They deal with their feelings about the Tunisian revolution of 2011 and their views on the economic situation after the revolution. The approach contrasts the supervised and non-supervised points of views, and insists on the available validation procedures involving Bootstrap techniques.

Résumé

Dans le cadre d'une démarche complète d'Analyse des Données, Analyse des Correspondances et Classification Automatique, d'un questionnaire, nous présentons dans ce travail les apports d'une Analyse de Données Structurées dans le contexte textuel. La démarche est principalement exploratoire mais dans la mesure où l'on discriminera des sous-nuages définis par des variables de structure, nous pouvons considérer qu'une partie de l'étude est supervisée. Nous illustrons les méthodes sur un questionnaire avec des réponses ouvertes administrées auprès d'universitaires tunisiens, avec trois facteurs structurants : le ressenti de la révolution tunisienne de 2011, l'identité tunisienne et les points de vue sur la situation économique après la révolution.

Mots-clés : Questions ouvertes, Analyse des correspondances, *Bootstrap*

1. Introduction

L'analyse des données textuelles telle qu'elle s'est développée depuis un demi-siècle (Benzécri et al., 1981) implique que l'on fasse l'hypothèse de la pertinence de plusieurs dimensions pour résumer le *sac de mots* afin de rendre compte de la complexité textuelle. L'analyse factorielle de tableaux de contingence permet une visualisation par cartes factorielles et une synthèse par axes factoriels retenus, les profils lexicaux par exemple. La classification automatique révèle des regroupements qui peuvent faire sens pour la problématique étudiée et que l'expert labellisera selon son domaine linguistique en s'appuyant par exemple sur l'interprétation des mailles d'une grille de Kohonen (carte auto-organisée) (Kohonen, 1984). La visualisation (Le Roux, 2004) est un point fort des approches exploratoires. On peut ainsi associer classification et représentation plane, par projection sur un plan factoriel des classes d'une partition. Les résultats théoriques sur le positionnement d'éléments supplémentaires permettent d'une part de proposer des interprétations pour les résultats de la démarche d'exploration systématique des données, mais ils permettent aussi l'amorce d'une démarche modélisatrice vers l'inférence statistique.

On étudie par exemple par des tests statistiques la significativité de la position de ces éléments illustratifs sur des axes factoriels dont le sens peut faire alors l'objet d'hypothèses à

confirmer. Cette validation externe peut aussi être une étape pour discriminer des sous-nuages, représentés par leur barycentre, projections des modalités d'une variable de l'étude. Cette démarche 'explicative' peut alors bénéficier des outils de validation (Efron et Tibshirani, 1993) (ré-échantillonnage, perturbations, etc.) des démarches supervisées. Dans notre recherche, nous transposerons l'approche de l'Analyse supervisée pour des facteurs structurants (Le Roux, 2004), que l'on peut considérer comme les facteurs d'un plan en analyse de la variance, au contexte des données textuelles. Nous mettrons enfin en œuvre ces nouvelles perspectives pour l'étude d'un questionnaire présenté ci-dessous, et qui comporte des questions ouvertes.

2. Contexte de l'application

Dans ce travail nous présentons des résultats de l'analyse des données textuelles appliquée à une problématique de Sciences Economiques induite par l'évènement que constitue la *révolution tunisienne* en 2011. Les données ont été collectées lors d'une enquête par questionnaire auprès d'un échantillon de 541 personnes. Il s'agit d'étudiants ou de jeunes professionnels: hommes et femmes d'origines très diverses, mais tous inscrits aux facultés (institut, école, etc.) des universités de Sousse et de Monastir. Le terrain ne respecte pas exactement les quotas signalétiques, en particulier en ce qui concerne les divers lieux d'étude proposés au Sahel, la sélection des répondants n'est pas non plus faite au hasard *stricto sensu*. Cependant, même si nous n'avons pas remarqué de biais avéré dans l'échantillon, nous utiliserons principalement les données dans un but d'application des méthodes d'analyse textuelles quantitatives proposées, et non pas pour les conclusions politico-sociologiques éventuelles de l'enquête. Cette enquête a été réalisée au cours de l'année universitaire 2012-2013, par interviews directes en face à face. La durée d'une interview est en moyenne d'une heure. L'objectif des questions fermées est d'identifier l'individu statistique en termes d'âge, de sexe, de niveau d'étude, d'état civil, etc. mais aussi de le caractériser par des réponses sur le sujet même de l'étude. Quatre questions ouvertes sont posées, les deux premières le sont sous forme de commentaires se rapportant à deux questions fermées codées par échelle de Lickert, en particulier relativement au sentiment de fierté du répondant, à son expérience de la révolution et sur sa citoyenneté tunisienne. La troisième question ouverte explicite la question fermée classant les causes majeures de déclenchement de la révolution tunisienne. Finalement, une dernière question ouverte concerne l'avis du répondant sur la situation économique du pays après la révolution.

3. Analyse Spécifique de facteurs structurants de données textuelles

3.1. Les tableaux construits pour l'analyse

Soit un ensemble de mots $\{M_1,\dots , M_l\}$, obtenus à partir des réponses aux questions ouvertes collectées sur le terrain tunisien. Ces mots sont caractérisés par plusieurs catégories telles que : sexe, âge, etc. du répondant. Lorsque l'on connait, parmi les variables un sous-ensemble que nous nommerons ici facteurs structurants $\{F\}$, choisis de façon experte, il est efficace d'analyser l'ensemble des mots recueillis en interaction avec ces variables plutôt qu'avec toutes les variables de l'étude. La notion de facteurs structurants est par exemple classique en sociologie dans les approches quantitatives du champ social selon le point de vue de (Lebaron et Le Roux, 2004).

Notons p le nombre de facteurs structurants, dont la valeur maximale admissible dépend de plusieurs critères comme par exemple le volume lexical du texte. Nous construisons dans ce contexte un tableau de données ayant, en abscisse, k mots $\{M_i\}$, obtenus à partir du corpus de

texte en cours d'analyse (ici les réponses aux questions ouvertes) et, en ordonnée, *p* facteurs structurants, comportant chacun m_{kp} modalités. A l'intersection d'une ligne M_i avec une colonne F_j figure la fréquence d'apparition du mot M_i avec une modalité $m_{j,r}$ du facteur F_j. Ce tableau de contingence qui croise mots et facteurs structurants étend à plusieurs variables catégorisées l'approche étudiée dans (Lebart et al., 2007) des tableaux de contingence ; on définit en particulier pour l'analyse des correspondances les profils en abscisses et les profils en ordonnées.

4. Les résultats de l'enquête 2012-2013 sur la révolution tunisienne

4.1. Analyse des questions fermées

Les questions fermées que nous traitons dans notre analyse sont catégorisées, nous procédons donc par analyse de correspondances multiples (ACM).

4.1.1 Identification des variables actives pour une ACM

L'ACM révèle les individus qui se ressemblent suivant la métrique de l'analyse (Benzécri, 1982 ; Le Roux et al., 2011), la sélection de ces individus se fait à partir des catégories respectives des variables actives étudiées. 5 variables actives qui sont au cœur de la problématique de notre étude ont été choisies (Origine (5), Vécu de la révolution (5), Participation (2), Réseaux (2), Situation économique (5)).

4.1.2 Résultats et interprétations

Un premier point à rappeler est que le choix du nombre des axes à interpréter se fait à partir de l'observation des valeurs propres (14 dans notre cas) ainsi qu'à partir du but de l'étude. Deux points de vue simplifient l'interprétation des axes obtenus, les critères statistiques et les critères experts.

4.1.2.1 Point de vue statistique

Statistiquement, on tient compte des 6 premiers axes. Mais en ACM, les taux d'inertie sous-évaluent l'inertie expliquée par plusieurs chercheurs (un recours au critère de (Benzecri, 1979) des taux d'inertie corrigés est préférable). L'interprétation statistique des axes est fondée sur l'observation des contributions des variables actives et de leurs modalités. Dans le tableau ci-dessous, nous examinons les contributions des différentes catégories des variables actives et leurs signes sur les axes. On examinera les modalités ayant les contributions les plus fortes pour les 3 premiers axes uniquement, car cela se révèle suffisant pour cibler notre problématique.

Axes	**modalité ayant la CTR la + élevée (a)**	**CTR(a)(%)**		**Modalité ayant la CTR la+élevée signe inverse**	**CTR (b) (%)**	
1	Ne pas participer	+	27.48	participer	-	17.03
2	très mauvaise	-	24.76	Mauvaise	+	18.84
3	Tunis capitale et alentours	-	17.46	Très bonne	+	13.00

Tableau 1. Récapitulation des contributions des variables actives

Axe 1 Oppose le sous-nuage des non participants à la révolution tunisienne à une CTR (CTR = Contribution relative) de 27.48 et les participants à une CTR de 17.03.

Axe 2 Oppose le sous-nuage des répondants qui voient que la situation économique est 'très mauvaise' à celui des répondants qui la considèrent 'mauvaise'.

Axe 3 Oppose le sous-nuage des répondants d'origine Tunis capitale et alentours à celui des répondants qui voient que la situation économique est très bonne.

4.1.2.2 Point de vue expert

Afin d'expliciter ces résultats une description experte est utile.

Axe1 l'opposition entre ces deux catégories d'une même variable est attendue. En effet, les réponses et les commentaires des répondants qui n'ont pas participé à la révolution et ceux qui y ont participé diffèrent ; exprimer des avis et des opinions concernant un évènement que l'on n'a pas vécu est extrêmement différent de parler de quelque chose que l'on a vécu.

Axe2 La divergence marquante entre deux catégories d'une même variable qui sont proches en termes de sens et d'intensité peut paraître étrange. En effet, considérer la situation économique 'mauvaise' ou 'très mauvaise' pourrait refléter presque le même point de vue, mais on peut dire que, préférer la modalité 'mauvaise' à celle 'très mauvaise' est simplement ne pas être aussi pessimiste envers ce qui se passe dans le pays après la révolution.

Axe 3 Cette opposition confirme que, tendanciellement, les étudiants originaires de Tunis et alentours sont contre le point de vue affirmant que la situation économique est très bonne. D'ailleurs, et sans aucun doute, la situation économique après la révolution se dégrade. Il est possible aussi que les problèmes qui ont émergé après la révolution aient influencé, négativement, le niveau de fierté des répondants.

4.2. Analyse avec les questions ouvertes

Les questions ouvertes donnent lieu à des réponses libres. Selon le sociologue (Lazarsfeld, 1944) le recours aux questions ouvertes est indispensable lors de la mise au point d'un ensemble de réponses pour une question fermée. (Lebart et Salem, 1988, 1994) notent que l'usage des questions ouvertes est intéressant principalement pour trois raisons : diminuer le temps d'interview, recueillir des informations spontanées, et enrichir l'explication et la compréhension d'une réponse à une question fermée (« Pourquoi ? Commenter votre choix », etc.).

Afin d'observer concrètement la différence entre le traitement des réponses aux questions ouvertes indépendamment des questions fermées et celui des réponses aux questions ouvertes qui dépendent des questions fermées, nous procèderons à une première analyse non supervisée puis à une deuxième supervisée (les questions fermées intervenant dans cette dernière analyse sont celles qui précèdent respectivement les questions ouvertes). Les variables catégorisées sont abrégées du fait que leurs libellé sont longs ; pour la variable situation économique, les catégories sont : très mauvaise (ecotrM), fragile (ecofrag), très bonne (ecotrB), mauvaise (ecomauv) et bonne (ecobonn). Pour la variable citoyenneté tunisienne, les catégories sont : Pas fier c (pasFC), Pas du tout fier c (pasdtFC), Très fier c (trèsFC), Bientôt fier c (bientFC), Fier c (FC). La dernière variable a les catégories : Pas fier v (pasFV), Pas du tout fier v (pasdtFV), Très fier v (trèsFV), Bientôt fier v (bientFV), Fier v (FV).

4.2.1. Analyse non supervisée

Nous débutons notre travail par une analyse des correspondances, dont le tableau des données est constitué des individus en lignes et des mots en colonnes. Une telle analyse permet de conclure sur certains éléments tels que les significations statistiques et expertes de chaque axe du plan factoriel. Si nous observons la représentation correspondante (figure 1), nous pouvons dire que l'information portée par le premier axe concerne d'un côté la situation du pays et de l'autre la situation du citoyen. Le deuxième axe oppose l'identité d'un citoyen après la révolution et le comportement du gouvernement. La zone grise centrale correspond à des projections de catégories mal reconstituées dans le premier plan et qui sont donc cachées sur le graphique. Ces interprétations classiques peuvent être confirmées par une perturbation effectuée sur les données de départ. On a recourt aux techniques de ré-échantillonnage.

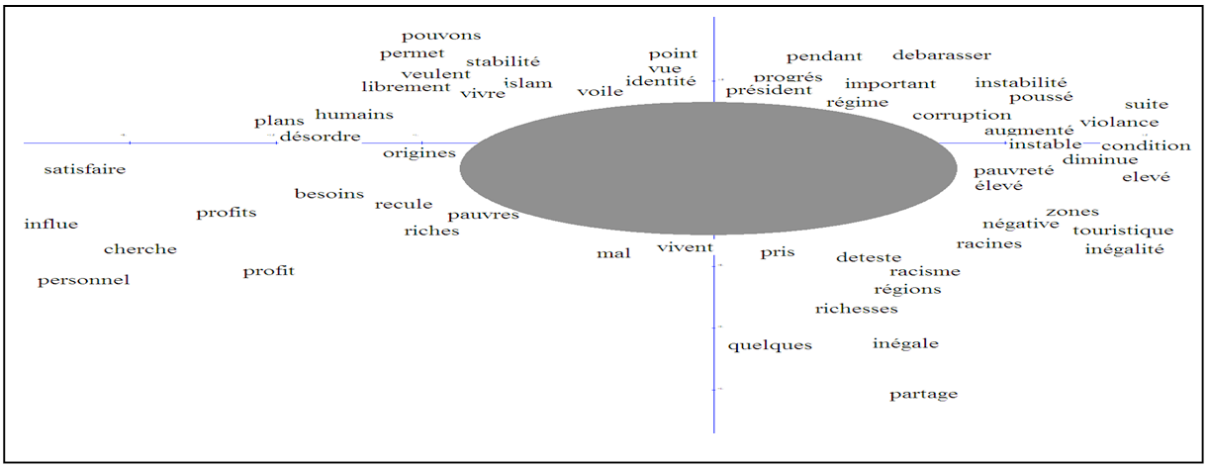

Figure 1. Premier plan factoriel (analyse non supervisée)

4.2.1.1 Validation par bootstrap (ré échantillonnage) dans le cas non supervisé

Une perturbation est faite sur des mots que nous considérons, d'une façon experte, comme étant des mots clés dans notre problématique initiale. Il s'agit donc ici d'une validation qui fait intervenir le processus de lemmatisation.

Nous observons que certains mots qui ont 'pratiquement' le même sens se situent, approximativement dans la même zone, les mots comme : 'dictature' et 'dictateur'. Une explication immédiate est qu'ils sont utilisés dans le même contexte.

Notons bien qu'ils sont au voisinage du mot 'dignité', qui ne peut pas se réaliser dans un système dictatorial. En revanche, si nous examinons les deux termes « pauvre » et « pauvreté », on voit une dissimilitude marquante. Cette dernière est due aux contextes distincts de l'utilisation de ces deux mots.

Une observation des mots positionnés au voisinage des termes en question, permet de dire que la forme graphique 'pauvres' est utilisée dans le but de confirmer la présence de pauvreté après la révolution. Alors que le terme 'pauvreté' est apparu afin d'exprimer un niveau et une évolution au cours du temps.

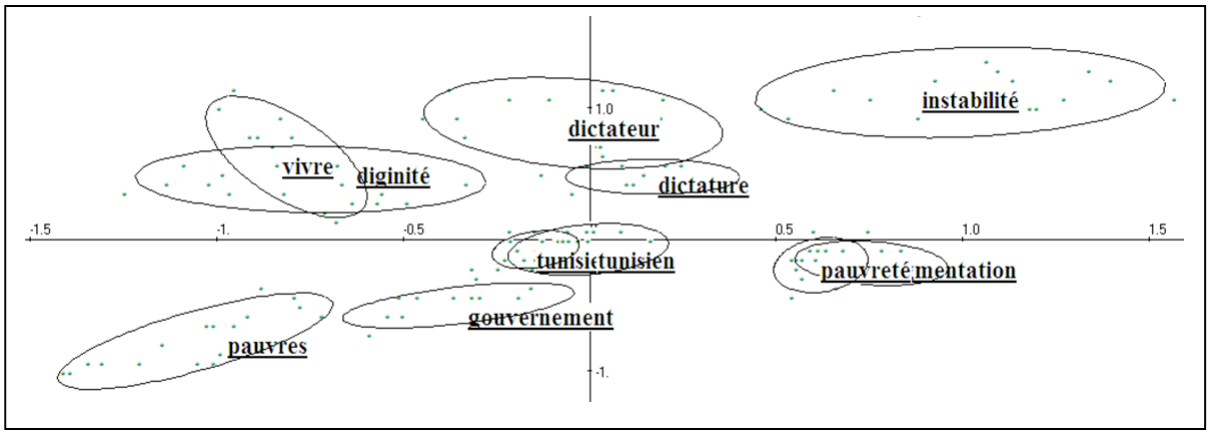

Figure 2. Zones de confiance (bootstrap) non supervisée

Rappelons que l'un des objectifs principaux de ce travail est de représenter les différences existantes entre une analyse non supervisée et une analyse supervisée, appliquées sur des réponses aux questions ouvertes. Nous appellerons ici 'supervisée' une analyse aidée par des facteurs illustratifs supplémentaires qui positionnent le texte en rapport à des facteurs structurants du champ social.

De ce fait, nous relançons donc le ré-échantillonnage en faisant intervenir des variables mises en supplémentaire, qui sont les facteurs structurants de la problématique.

4.2.2. Analyse supervisée

Comme nous l'avons énoncé auparavant, cette analyse est basée sur les trois facteurs structurants suivants : ressenti sur la révolution, identité tunisienne et situation économique.

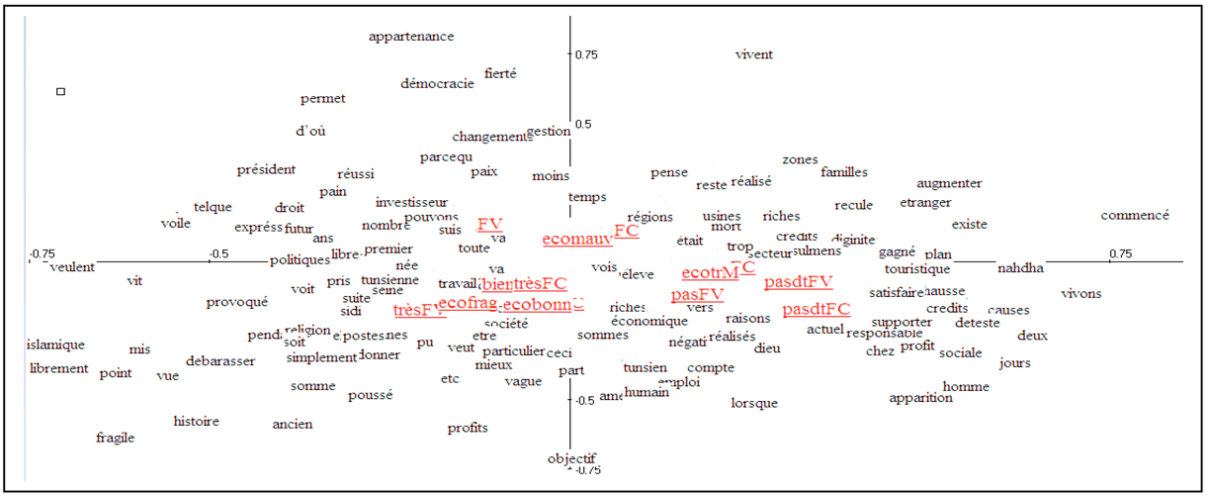

Figure 3. Premier plan factoriel (analyse supervisée)

Une visualisation plane par cartes auto-organisées (cartes de Kohonen) simplifie encore plus.

4.2.2.1 Visualisation par carte de Kohonen

Dans cette représentation plane du plan factoriel, nous exposons cette fois simultanément les mots et les catégories des facteurs structurants. Les cartes de Kohonen (Kohonen, 1989) « permettent d'afficher simultanément les lignes et les colonnes de la table lexicale » (Lebart, 2004).

réalisés jeunes commencé causes assez	valeur temps racisme pousser pasFC parmi historique absence	besoins	vécu usines secteur partout investisseme grave	élevés sont prix	vie pauvreté mauvaise ecotrM donne chomage beaucoup augmentation	mauvaises importante hausse elevé abscence	réaliser non humains commentaire chose
crise changement appartient	société premiers interieurs injustice eux	respect domaine chomeurs	y meilleur encore	problèmes mauvais insupportabl inflation après	sécurité les du	très ma	tunisiens sans revolution manque citoyen
souffre pasdtFV mal etrangères	diplomés claire citoyens cette arabes	toute qui ont	que plusieurs mais economique dans	je il ecomauv des au	tous revolution le la etre est a	plus de	vu sur pas ne j conditions ce cause

Figure 4. Extrait de la carte de Kohonen pour les mots et les catégories des facteurs structurants

Les cases constituant cette représentation, contiennent les formes lexicales les plus fréquentes combinées avec les différentes catégories des 3 facteurs en question. Par exemple, la catégorie situation économique 'très mauvaise' (ecotrM), située dans la 6ème case de la 1ère ligne, est à proximité des mots : pauvreté, mauvaises, élevé, sécurité, etc. Le sens de ce *bag of words* peut être illustré par le fait qu'il se présente au voisinage de la catégorie situation économique 'ecotrM'.

4.2.2.2 Validation par bootstrap dans le cas supervisé

Une validation par *bootstrap*, dans ce cas supervisée (figure 5), illustre les catégories des facteurs structurants et les formes graphiques (figure 6), utilisées lors de la validation faite en analyse non supervisée (figure 2). En effet, ce type d'analyse permet d'explorer les différentes catégories conjointement avec les différentes formes graphiques du texte à étudier ou bien avec des formes graphiques sélectionnées d'une façon experte.

Notons, à titre d'exemple, que les deux catégories 'pasdtFC' et 'pasdtFV' se positionnent au voisinage de la forme graphique 'pauvres', ce qui paraît cohérent avec la réalité vécue en Tunisie.

N'être 'pas du tout fier' de l'identité tunisienne et du vécu de la révolution peut être un résultat de la pauvreté 'aigue' vécue après la révolution, pour la plupart des citoyens. Nous remarquons aussi, que la catégorie 'ecotrB' va de pair avec la plus faible fréquence de la variable 'situation économique', ce qui a entrainé une dégénérescence de l'ellipse de confiance en un segment. Statistiquement, ce résultat est attendu du fait que pour un effectif très faible on ne peut pas garantir assez de réplications pour créer une ellipse.

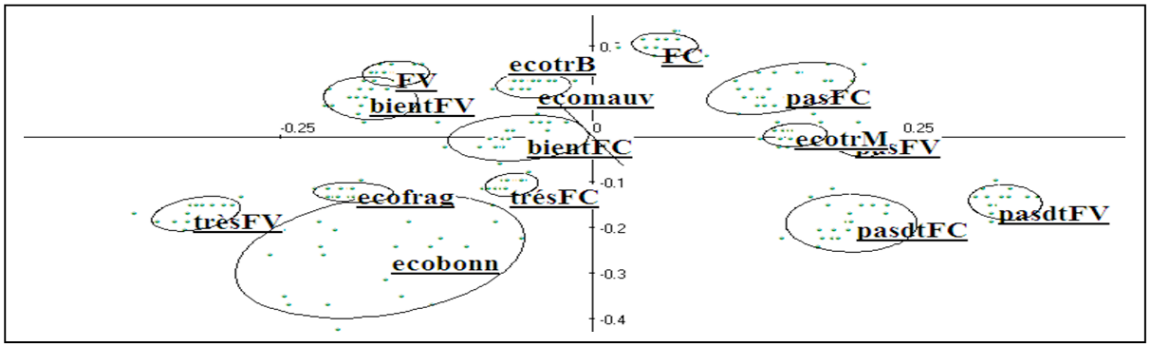

Figure 5. Zones de confiance (bootstrap) des catégories des facteurs structurants

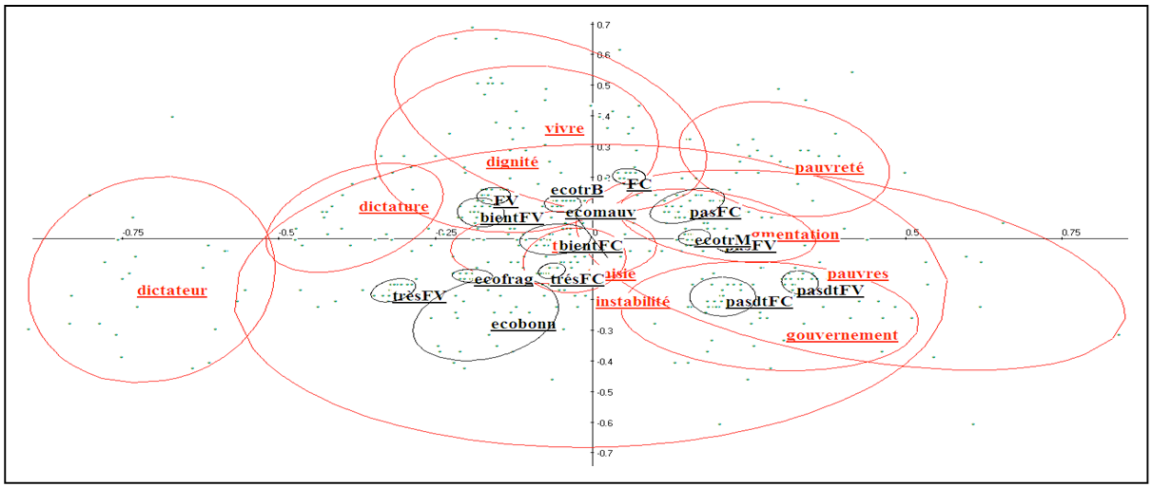

Figure 6. Zones de confiance (bootstrap) des catégories des facteurs structurants et des formes graphiques.

4.3. Comparaison statistique des résultats avant et après lemmatisation

Cette comparaison se limitera aux tableaux des valeurs propres, à la classification par les k-moyennes et la validation par *bootstrap*.

4.3.1. Les valeurs propres

Douze valeurs propres sont retenues à partir de trois variables, à 5 catégories chacune. En effet, la comparaison des pourcentages cumulés des valeurs propres, avant et après lemmatisation, ne permet de constater qu'une faible amélioration, de 21.60 % à 25.37 pour le premier facteur (de même pour les différents autres facteurs).

Nbr	VP nl	VP l	% cumulé nl	% cumulé l
1	.0289	.0132	21.60	25.37
2	.0151	.0055	32.88	36.05
3	.0133	.0051	42.83	45.92
4	.0117	.0049	51.57	55.34
5	.0106	.0044	59.50	63.85
6	.0102	.0036	67.13	70.85

Tableau 2. Valeurs propres des données avant et après lemmatisation

4.3.2. Classification par k-moyennes et lemmatisation

La classification par les k-moyennes (développé par Mac Queen) est considérée l'une des méthodes de classifications les plus rapides et efficaces dans son usage. Elle permet, comme toute méthode de classification, de construire des groupes qui sont homogènes, de telle sorte que plus la variance (variance-inter) entre ces groupes est faible, plus l'hétérogénéité (variance intra) est élevée et vice versa. Douze classes sont fixées a priori (en fait suggérées par une classification ascendante hiérarchique) avant de procéder par la méthode k-moyennes. Nous avons appliqué cette dernière, dans le cas où les données sont à l'état brut (non lemmatisées), puis dans le cas où il y a eu des modifications sur les données (lemmatisées). La détection des disparités entre ces deux cas est engendrée par l'analyse des variances inter-classes et des variances intra-classes.

4.3.2.1 Variance inter-classes

L'examen du tableau 3, comparant les variances inter-classes des données, avant et après lemmatisation, nous permet de remarquer, principalement que le nombre des itérations a diminué après la lemmatisation. Au départ, sept itérations ont été nécessaires pour atteindre la variance inter-classes maximale (0.589612). Ce nombre d'itérations n'est que de six après lemmatisation. Une deuxième remarque est que les variances inter-classes, dans le cas des données 'nl', non lemmatisées, sont plus faibles que dans le cas des données 'l', lemmatisées.

Itération	*Variance nl*	*Variance l*
0	0.537464	.443747
1	0.579583	.480360
2	0.585249	.484078
3	0.587976	.485556
4	0.588833	.486063
5	0.589564	.486234
6	0.589593	.486384
7	0.589612	-

Tableau 3. Variances inter-classes avant et après lemmatisation

A ce niveau, nous pouvons dire que la variance inter-classes, obtenue sur les données 'nl' est plus élevée que celle obtenue sur données 'l'. Ainsi, l'hétérogénéité inter-classes est plus élevée en fin de processus qu'avec les données 'nl', 0.5896 contre 0.4864.

4.3.2.2 Variance intra-classes

Lorsque l'on calcule les variances intra-classes dans les deux cas, 'nl' et 'l', on remarque qu'il y a un nombre de classes plus homogènes dans le cas des données 'l' que dans le cas des données 'nl'. L'observation simultanée des deux figures ci-dessous (figures 7 et 8) confirme qu'avec les données 'nl', on ne peut repérer aisément que deux groupes (6 et 10), alors que les autres groupes sont trop proches et parfois même en superposition partielle. Cette constatation peut s'expliquer par le fait que les groupes ont des tailles plus grandes que dans le cas 'l'. Dans ce dernier cas, nous pouvons observer que même si toutes les classes sont proches, nous pouvons distinguer presque tous les groupes (à part deux d'entre eux encore trop contigus sur le graphique). Cela est dû à ce que les classes sont devenues plus 'propres' et moins 'chargées', après la lemmatisation des données.

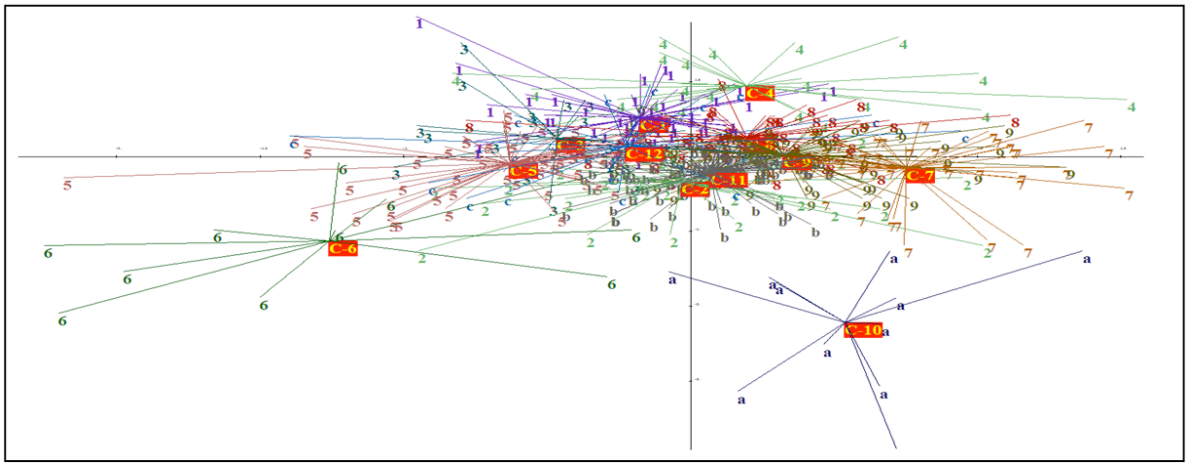

Figure 7. Visualisation par K-moyenne (12 classes) avant lemmatisation des données

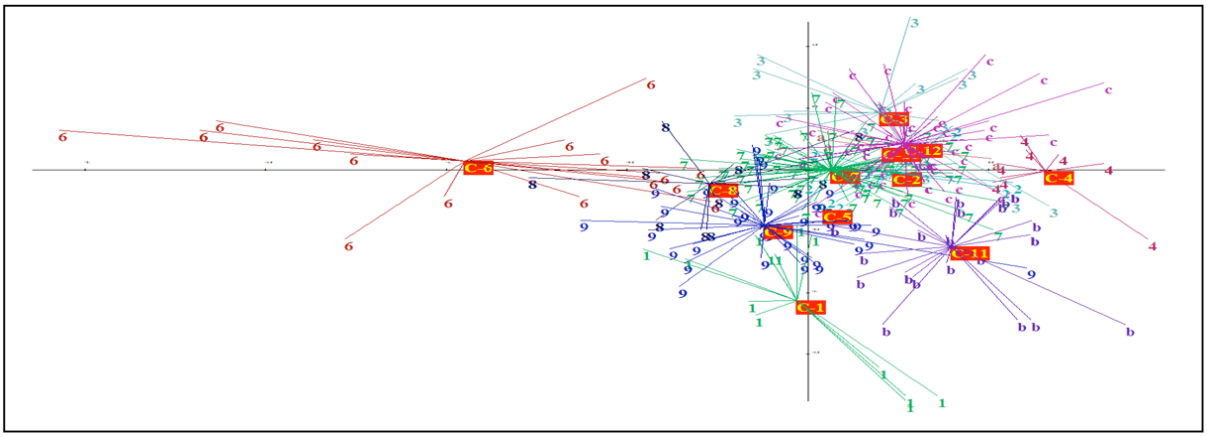

Figure 8. Visualisation par K-moyenne (12 classes) après lemmatisation des données

4.3.3. Validation par bootstrap

La méthode de *bootstrap* (Efron et Tibshirani, 1993) permet de créer des nouveaux échantillons par tirage avec remise, effectué sur l'échantillon de départ. Ce qui provoque une perturbation des données initiales. La validation par *bootstrap* est exécutée une fois qu'on examine les zones de confiances obtenues par cette technique de ré-échantillonnage sur les plans factoriels. A ce niveau, une double comparaison double est possible ; le '*bootstrap*-nl', effectué sur les données 'nl' et le '*bootstrap*-l' effectué sur les données 'l', et dans les deux cas d'analyse : supervisée et non supervisée. Dans le premier cas, comparaison d'une analyse non supervisée sur données avant et après lemmatisation (comparaison des figures 2 et 9), nous pouvons dire qu'après lemmatisation des données les différentes oppositions entre les formes graphiques sont les mêmes. Ainsi, on peut dire que la lemmatisation des données a confirmé les oppositions obtenues lorsque les données sont à l'état brut. On note aussi, que les formes graphiques qui étaient 'presque' non distinctes avant lemmatisation, sont devenues, pour la plupart, vraiment non distinctes, comme par exemple pour les trois formes 'tunisien', 'tunisie' et dignité.

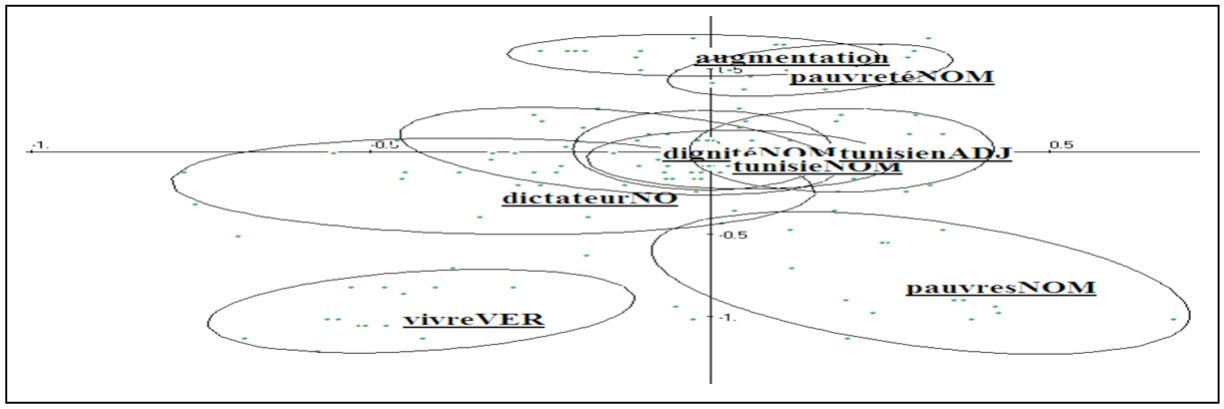

Figure 9. Zones de confiance non supervisée après lemmatisation des données

La comparaison lors d'une analyse supervisée (les 3 facteurs structurants utilisés dans la sous section 4.1.2.2), d'un *bootstrap* appliqué sur des données 'nl' et d'un *bootstrap* sur les mêmes données qui ont été lemmatisées plus tard (figure 10), conduit aux conclusions obtenues dans le cas d'une analyse non supervisée. D'une part, la lemmatisation des données a préservé les mêmes oppositions détectées avant la lemmatisation. D'autre part on observe une diminution

des proximités entre les ellipses de confiances par rapport aux positions de l'analyse non lemmatisée (figures 5 et 10).

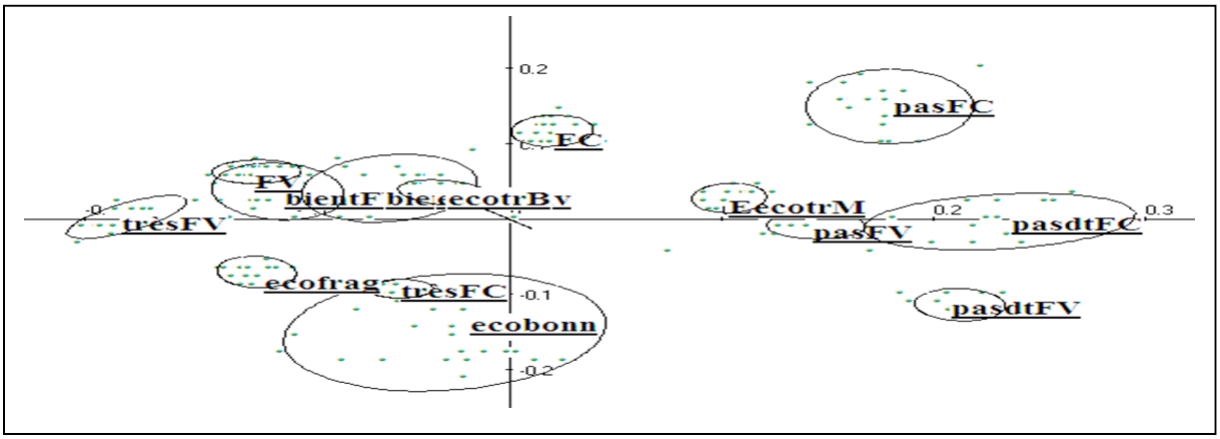

Figure 10. Zones de confiance supervisée après lemmatisation des données

5. Conclusion

L'intervention des questions ouvertes d'une analyse des correspondances multiples sur des variables, reportant dans notre étude les avis des étudiants concernant une révolution vécue pour une première fois, permet d'enrichir les résultats obtenus avec une ACM 'classique'. Ceci se révèle particulièrement après une comparaison effectuée entre une analyse supervisée, entendue ici avec l'intervention de facteurs structurants, et une autre non supervisée sans variables mises en supplémentaire.

Une deuxième comparaison qui paraît pertinente est celle réalisée entre une analyse effectuée sur des données avant et après lemmatisation, alternativement dans le cas supervisé et non supervisé. Il apparait que dans notre étude, la lemmatisation des données a entrainé une amélioration effective des résultats d'un point de vue statistique, notamment lors de la visualisation et de la validation.

Références

Benzécri J.-P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Les cahiers de l'analyse des données*, tome 4: pp 377-378.

Benzécri J.-P. *et al*. (1981). *Pratique de l'Analyse des Données : linguistiques et lexicologie*, tome 3, Dunod, Paris.

Benzécri J.-P. (1982). Histoire et préhistoire de l'Analyse des Données: L'analyse des correspondances. *Les cahiers de l'analyse des données*, tome 2 : pp 9-40.

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).

Efron, B. et Tibshirani, R, J. (1993). *An introduction to the bootstrap*, Chapman and Hall, New York.

Kohonen T. (1989) *Self-Organization and Associative Memory*, Springer

Lazarsfeld P.-E. (1944) the controversy over detailed interviews: An offer for negotiation. *Public opinion quarterly,* vol(8):p38.

Le Roux B. (2004). Structured Data Analysis. , Blasius J. and Greenacre M.. , *Visualization and Verbalization of Data.* CRC Computer Science & Data Analysis, Chapman & Hall.

Le Roux B., Rouanet H., Savage M. et Warde A. (2008). Class and cultural division in the uk. Sociology, *SAGE*, (42):1049–1071.

Le Roux B. Bonnet P. et Lebaron F. (2011). La notion de champ et l'analyse des correspondances multiples (ACM). *Séminaire résidentiel méthodologiqu*e, pp. 3-9.

Lebaron F. et Le Roux B. (2004). *Pratiques culturelles et espace social: la méthodologie de P. Bourdieu en action,* (eds), Dunod, Paris.

Lebart L. et Salem A. (1988). *Analyse Statistique des Données Textuelles, Questions ouvertes et lexicométrie.* Dunod, Paris.

Lebart L. et Salem A. (1994). *Statistique textuelles*, Dunod, Paris.

Lebart L., Piron M. et Steiner J.-F. (2003*). La sémiométrie.* Dunod, Paris.

Lebart L. (2004). Validité des visualisations des données textuelles. Purnelle G., Fairon C., Dister A. (eds). JADT 2004 (*7éme Journées Internationales d'Analyse Statistiques des Données Textuelles*), pp. 708-715.

Lebart L., Piron M. et Morineau A**.** (2007). *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouille de données,* Dunod, Paris.

Lebart L. (2012). L'articulation entre exploration et inférence en analyse statistique de textes. JADT 2012 (*11éme Journées Internationales d'Analyse Statistiques des Données Textuelles)*, pp. 708-715.