Quelle modélisation de l'espace politique français sur Twitter ?

(Version longue et annexes)

Nicolas Hervé
Institut National de l'Audiovisuel - Ina
nherve@ina.fr
http://www.herve.name

1er décembre 2020

Ce document [1] est une version longue de l'article publié à la conférence EGC2021 (Extraction et Gestion des Connaissances, Montpellier, 25 - 29 janvier 2021). Il s'agit d'une version remaniée et enrichie, notamment avec de nombreux graphiques en annexe. Nous y présentons plusieurs modélisations de l'espace politisé français sur Twitter dans l'optique de fixer un cadre d'analyse pour y étudier la propagation des contenus. Quatre modélisations sont proposées et comparées pour leurs caractéristiques respectives et ce qu'elles permettent de percevoir de l'espace partisan sur ce réseau social. Dans un second temps, une analyse des 13 000 principaux *hashtags* utilisés entre juillet 2018 et juillet 2020 est réalisée sur une base de 3.8 milliards de *tweets*. Cette analyse de leur polarisation contribue à éclairer la discussion sur la validité et l'interprétation possible des modélisations présentées.

1 Introduction

Dans le cadre général de l'analyse de la propagation des informations sur Internet et dans les médias, il est utile de considérer la dimension politique comme un des critères à observer et quantifier. La polarisation politique est en effet un phénomène connu et largement documenté. Il nous semble important de le prendre en compte dans le cadre d'une analyse visant à mieux comprendre les comportements des producteurs et consommateurs d'informations dans l'écosystème médiatique [Barberá et al., 2015], [Conover et al., 2011b], [Garimella and Weber, 2017]. Nous cherchons par exemple à répondre aux questions suivantes : les choix éditoriaux des principaux médias d'information sont ils en partie guidés par des considérations économiques, journalistiques ou politiques ? Comment les discussions concernant l'actualité sur Twitter évoluent, dans quelles communautés ? Quelle est l'importance de la polarisation politique des individus dans ces échanges ? En quoi ces discussions influent sur les médias [Cagé et al., 2020] ? Y a-t-il

1. Voir la page dédiée à ce travail : http://www.herve.name/twitter-polarization

certains contenus fortement connotés politiquement ? Quel lien entre une forte polarisation politique et la propagation de *fake news* [Tucker et al., 2018] ?

Nous présentons ici la première étape de notre démarche avec la construction et la validation de l'espace politique latent sur Twitter. Notre objectif est la création de cet espace politisé en tant que grille d'analyse des contenus qui circulent sur Twitter. Nous nous distinguons en cela des auteurs, principalement les politologues, qui étudient l'espace politisé pour lui-même, notamment pour analyser le positionnement relatifs des *élites* et de la population. De nombreuses publications ont déjà élaboré des approches pour y parvenir. Elles se situent toutefois majoritairement dans le cadre des USA pour lequel l'espace politique bipolaire républicains / démocrates peut sembler, du point de vue européen, plus simple à modéliser. De plus la validation de cette modélisation par des données externes est également facilitée par leur disponibilité accrue outre-atlantique du fait d'une réglementation beaucoup moins protectrice pour les individus. Enfin, les comportements des individus sur un réseau social comme Twitter répondent à des contraintes universelles (les fonctionnalités mises à disposition par Twitter) mais également à des codes sociaux qui peuvent être différents d'un pays à l'autre et qui peuvent évoluer dans le temps. Pour toutes ces raisons, il nous semble donc important de pouvoir établir une modélisation spécifique pour l'espace francophone / français de Twitter. Nous reviendrons sur cette distinction entre la langue et le pays plus loin. La mise en place de cette modélisation de l'espace politisé sur le Twitter francophone est également l'occasion de questionner et approfondir certains choix faits dans les publications de l'état de l'art. Pour faciliter la lecture et la compréhension, nous conserverons les termes anglais du vocabulaire de Twitter sans les franciser. Après un état de l'art dans la partie 2, nous présentons les modélisations que nous étudions et les données utilisées dans la partie 3. Une première analyse de la polarisation de contenus est effectuée sur des *hashtags* dans la partie 4 avant de conclure. Davantage de graphiques sont présentés en annexes (page 27).

2 État de l'art

Twitter est un réseau social de microblogging. Les utilisateurs y postent de courts messages (*tweets*) qui sont visibles publiquement [2]. Par défaut, chaque utilisateur voit sur son fil les *tweets* des comptes auxquels il s'est préalablement abonné (*friends*) et diffuse ses propres *tweets* aux comptes qui le suivent (*followers*). Les utilisateurs peuvent interagir entre eux en repostant certains messages (*retweet*), en les commentant (*quote*) ou en y répondant (*reply*). Il peuvent également mentionner (*mention*) un compte dans un *tweet*. Le contenu des *tweets* peut être constitué de texte, d'images, de vidéos, d'URLs et de *hashtags*. Ces derniers permettent d'annoter les messages avec une métadonnée, généralement standardisée au sein d'une communauté, pour rapidement identifier le sujet du *tweet*.

Twitter est un des réseaux sociaux les plus étudiés du fait du caractère public des messages échangés et de sa politique d'accès aux données qui permet de facilement créer

2. Twitter offre la possibilité d'avoir des comptes privés, mais cette pratique est très limitée et, par nature, impossible à analyser.

des corpus. Il est donc naturel que les chercheurs en sciences humaines se soient rapidement intéressés aux comportements de ses utilisateurs. Parmi les multiples axes d'analyse, l'étude des opinions politiques et de la circulation de contenus fortement polarisés a concentré de nombreux efforts à cause, notamment, des enjeux démocratiques sous-jacents. On peut avoir un premier aperçu des approches possible en se référant à des revues de l'état de l'art récentes : [Severo and Lamarche-Perrin, 2018] et [Tucker et al., 2018].

Dans l'étude des groupes sociaux, le principe d'homophilie statue que les gens s'associent généralement à des groupes de personnes leur ressemblant. Ce phénomène a également été observé dans de nombreuses communautés en ligne telles que les blogs. Une des principales question de recherche concernant les réseaux sociaux est de savoir si la facilité avec laquelle il est maintenant possible de communiquer tend à enfermer les utilisateurs dans des bulles homogènes (*echo chambers*) au sein desquelles les opinions sont partagées et se renforcent mutuellement ou si, au contraire, le fait de pouvoir être potentiellement exposés à des arguments plus variés permet un meilleur débat public et des prises de position plus éclairées sur tous les débats de société. Halberstam and Knight [2016], par exemple, étudient ce phénomène. De plus, l'influence des algorithmes des plateformes qui choisissent les contenus auxquels les utilisateurs sont potentiellement exposés ne sont également pas neutres (*filter bubbles*).

2.1 Modélisation de l'espace politisé

Dans une des premières publications sur le sujet, Yardi and Boyd [2010] étudient les discussions sur Twitter concernant la fusillade et la mort du Dr Tiller liées à la polémique entre les pro et anti avortement aux USA. Ils utilisent un petit corpus de *tweets* et catégorisent manuellement les comptes Twitter. Ils expliquent que des échanges ont bien lieu entre les deux communautés mais qu'ils conduisent toutefois difficilement à des discussions fructueuses.

Conover et al. [2011b] étudient quant à eux Twitter pendant la campagne pour les élections de mi-mandat aux USA en 2010. Le corpus est constitué de 250 000 *tweets* produits par 45 000 comptes Twitter à partir desquels les réseaux de *retweets* et de *mentions* sont extraits. L'obtention des *tweets* est basée sur une liste de *hashtags* définie semi-automatiquement et directement liés aux campagnes des républicains et des démocrates. L'analyse sur les graphes est ensuite conduite avec un algorithme de détection de communautés. Afin d'analyser les communautés extraites et leur cohérence, les comptes Twitter sont caractérisés d'une part avec des vecteurs *one-hot* des *hashtags* qu'ils utilisent et d'autre part certains comptes sont manuellement annotés selon trois catégories politiques (droite, gauche, indécidable). Les auteurs mettent en évidence que le réseau formé par les *retweets* fait clairement apparaître les deux communautés, contrairement à celui se basant sur les *mentions*. De plus, ils expliquent que la stratégie d'utilisation des *hashtags* par chacune de ces communautés vise à permettre aux *tweets* d'atteindre la communauté adverse, ce qui est mis en évidence en calculant la valence des *hashtags*. Dans une seconde publication, Conover et al. [2011a] tentent de classifier automatiquement la polarisation politique des comptes Twitter en utilisant le même jeu de données et en encodant des

caractéristiques telles que le contenu de leurs *tweets* ou celles du réseau en utilisant les clusters précédemment détectés. Toutefois, les très bons scores de classification obtenus sont à relativiser. C'est notamment ce qu'indiquent Cohen and Ruths [2013] en soulignant deux problèmes importants du jeu de données : la faible représentativité de la base de comptes Twitter utilisés (généralement des profils hautement politisés puisque cela fait partie des critères pour la constitution du corpus) et le manque d'évaluation de la généralisation des modèles entraînés à d'autres corpus.

La papier principal de Barberá [2015] décrit l'approche qui est maintenant considérée comme standard. Il se base sur la matrice de contingence de *followers* indiquant si l'utilisateur i suit le compte politique j. Il suppose l'existence d'un espace latent dans lequel on peut mesurer la distance entre les utilisateurs et les comptes politiques. Il introduit deux variables supplémentaires à son modèle mesurant d'une part l'intérêt politique pour chaque utilisateur et la popularité des comptes politiques d'autre part. Le modèle est alors défini par :

$$P(\mathbf{Y}_{ij} = 1|\alpha_i, \beta_j, d_{ij}) = logit(\alpha_i + \beta_j - d_{ij})$$

Ce modèle bayésien est estimé avec une approche MCMC. Cette méthodologie est exploitée sur des comptes Twitter aux USA et dans cinq pays européens (mais pas la France). L'avantage de cette approche est que les utilisateurs et les comptes politiques sont plongés dans le même espace latent. Trois types de validation sont utilisés. Il montre que le positionnement des comptes politiques est équivalent à celui obtenu par les approches traditionnelles en sciences politiques, que la distribution des utilisateurs anonymes sur un axe droite-gauche est cohérente et enfin, il identifie certains des comptes anonymes avec des informations publiques de dons financiers aux campagnes politiques. Il note toutefois que les utilisateurs de Twitter ne sont pas représentatifs de la population en âge de voter. Barberá insiste sur le fait que s'abonner à un compte politique sur Twitter est une action hautement informative sur le positionnement idéologique d'un utilisateur. Tout l'enjeu pour lui est de correctement choisir les comptes politiques qu'il faut considérer pour établir le modèle en tenant compte de leur représentativité mais sans chercher l'exhaustivité qui rendrait les temps de calcul impraticable. Dans un second papier Barberá et al. [2015] utilisent la mesure mise en place pour étudier 12 évènements au prisme des discussions polarisées sur Twitter. Dans les annexes en ligne, l'utilisation de l'analyse factorielle de correspondances [Benzécri, 1973] est abordée pour remplacer l'estimation bayésienne, beaucoup plus rapide et globalement équivalente.

Dans [Darmon et al., 2015], même s'il n'est pas explicitement question de polarisation politique, les auteurs indiquent que ne tenir compte que du réseau de *followers* est réducteur pour identifier les communautés sur Twitter.

Enfin, dans [Garimella and Weber, 2017], la polarisation est vue plus simplement comme la probabilité de suivre un compte de gauche (α comptes suivis) ou de droite (β comptes suivis). Dans le monde bipolaire des USA, la polarisation est ensuite un simple ratio :

$$p = 2.|0.5 - \frac{\alpha}{\alpha + \beta}|$$

Cette mesure est utilisée pour analyser la polarisation de *hashtags* et l'évolution de cette polarisation sur le long-terme.

2.2 Espace politisé français

On trouve quelques travaux spécifiques à la France. Briatte and Gallic [2015] appliquent la même approche que Barberá pour des comptes français. Dans des travaux plus récents[3], ils proposent l'utilisation de *detrended CA* pour limiter les effets d'éloignement des extrêmes lié à l'AFC. Ramaciotti Morales et al. [2020] utilisent également l'AFC et proposent deux méthodes basées sur les graphes pour propager la polarisation à un nombre étendu de comptes Twitter. Chavalarias et al. [2019] se basent sur les communautés détectées à partir des réseaux de *retweets* dans le cadre de la campagne présidentielle de 2017 pour analyser les stratégies de communication sur Twitter des partis politiques et des candidats. Enfin, Cardon et al. [2019] intègrent les *retweets* de quelques comptes politiques pour propager une polarisation à des comptes de médias. Nous savons qu'en France, comme dans les autres pays, les personnes actives sur Twitter ne sont pas parfaitement représentatives de la population. Un sondage annuel réalisé par le Reuters Institute[4] interroge les habitants de plusieurs pays sur leurs habitudes de consommation des informations selon différents canaux. Si on regarde les chiffres de l'année 2018 pour la France, réalisé auprès de 2 000 personnes, on observe que Twitter est utilisé par 16% d'entre eux. Cette population utilisant Twitter est plus masculine, plus jeune, plus francilienne, avec de meilleurs revenus, un plus haut niveau d'études que la moyenne. Elle est également plus politisée et légèrement plus à gauche (voir la figure 1).

3 Propositions de modélisation

Il existe deux approches principales pour modéliser globalement l'espace politisé sur Twitter. Elles se basent sur une liste pré-établie $\mathcal{P}$ de comptes politiques de référence pour lesquels le positionnement politique est connu. Ce positionnement politique est propagé aux comptes Twitter anonymes en utilisant les liens de *follow*. Intuitivement, plus un compte Twitter va suivre de comptes politiques d'un même parti, plus on va considérer qu'il en est proche idéologiquement. Une approche naïve comme [Garimella and Weber, 2017] utilise une simple combinaison linéaire pondérée alors que celle de Barberá [2015], plus évoluée, ne nécessite aucun *a priori* sur les partis politiques et infère automatiquement leurs positionnements relatifs et l'importance de chacun des comptes politiques de référence. Ces deux modélisations peuvent également être appliquées en utilisant les informations de *retweet*. Il n'y a en effet aucune raison pour que l'action de suivre un compte politique particulier soit plus porteuse d'information sur le positionnement idéologique d'un utilisateur que lorsqu'il *retweete* ses messages. Nous avons donc 4 modélisations possibles que nous allons appliquer à l'espace français de Twitter.

3. https://github.com/briatte/epsa2019
4. http://www.digitalnewsreport.org/

	Pop.	Homme	-35 ans	Île de France	Haut revenus	Haut niveau éducation	Pos. pol. indéterminé
Étude	2006	47.4%	25.6%	18.1%	21.9%	31.0%	26.4%
Facebook	63.4%	45.3%	28.2%	17.5%	20.3%	31.1%	27.1%
Twitter	16.2%	57.2%	35.7%	22.5%	25.8%	36.6%	16.6%

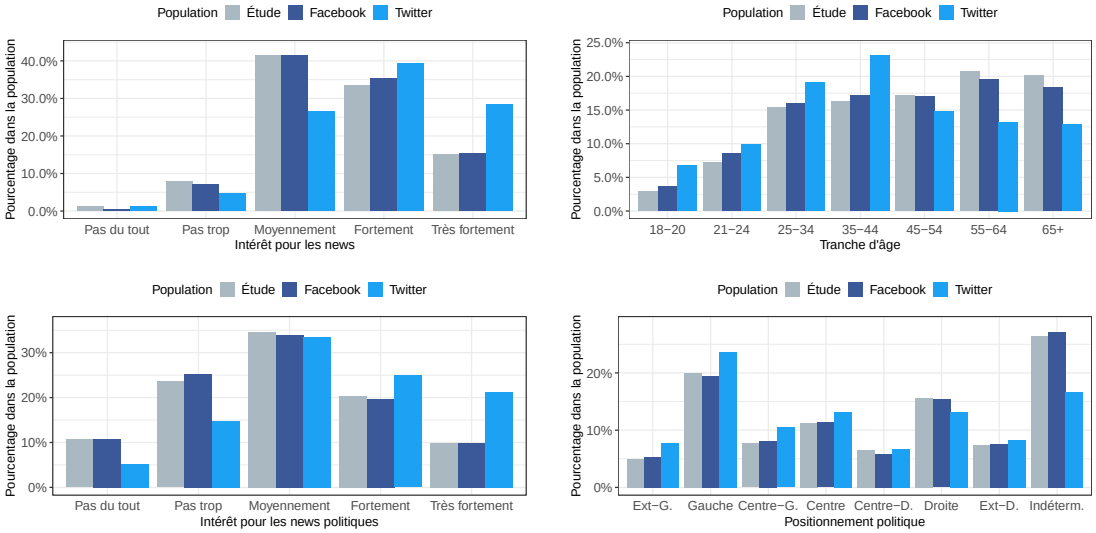

FIGURE 1 – Caractéristiques des utilisateurs de réseaux sociaux en France (Reuters Institute, 2018)

On considère que l'on a n comptes Twitter C_i pour lesquels on dispose d'une information les reliant à m comptes particuliers P_j (des comptes Twitter pour lesquels indiquer une appartenance à une catégorie particulière, potentiellement propre à définir une classification des individus). Dans cette étude on s'intéresse au spectre politique. On peut également envisager de regarder les sphères médiatiques, que ce soit là encore sous un angle politique ou bien selon une segmentation thématique (information générale vs. presse spécialisée) ou encore géographique (local vs. national).

Sous l'hypothèse d'homophilie du réseau, on va chercher à propager la polarisation des comptes particuliers P_j aux comptes Twitter C_i. La labellisation des comptes P_j est supposée être plus simple puisque concernant un nombre très restreint d'individus (au regard de la base d'utilisateurs de Twitter) et que l'on peut facilement catégoriser, généralement du fait de leur notoriété au regard, justement, de l'axe de catégorisation qui nous intéresse.

Afin de propager cette catégorisation, on va chercher à représenter les comptes C_j grâce à des propriétés les reliant aux comptes P_j. Ces informations peuvent être de deux natures différentes :

— structurelle : le compte C_i suit le compte P_j

— activité : le compte C_i a retweeté, quoté ou mentionné au moins t fois le compte P_j sur une période de temps donnée

3.1 Données Twitter

Les données que nous utilisons sont obtenues via les API mises à disposition par Twitter. [5]. Nous pouvons d'une part capter les données de la structure du réseau social : informations sur les comptes Twitter et les relations de *follow* / *friend*. D'autre part, nous pouvons capter les *tweets*. Pour cela nous utilisons l'approche proposée par Mazoyer et al. [2018] pour les *tweets* émis en français. Elle permet en effet de collecter un ensemble représentatif de *tweets* en se basant sur l'utilisation de requêtes formées avec des mots neutres de la langue française. Puisque l'objectif est de pouvoir analyser sur le long terme l'ensemble des sujets de discussion sur ce réseau social, il n'est ni possible ni souhaitable de définir au préalable une liste de mots clés ou de *hashtags*. De plus, selon les auteurs, cette approche assure que les *tweets* captés sont statistiquement similaires à la distribution globale des *tweets* en français et qu'il est possible de capter environ 70% des *tweets* émis en français sur la plateforme. La captation couvre une période de deux ans, de juillet 2018 à juillet 2020 avec en moyenne 5 millions de *tweets* par jour. Au total, le corpus $\mathcal{T}$ contient 3 898 250 863 *tweets* en français (en comptant les *retweet*). La captation est régulière (figure 18) avec en moyenne 5 millions de tweets en français par jour. Les quelques rares artefacts sont liés à des indisponibilités de serveurs qui peuvent causer un délai dans le stockage des tweets. On remarque la légère augmentation de la volumétrie sur la récente période du confinement au début de l'année 2020. Sur cette période plusieurs événements médiatiques d'importance ont eu lieu. On peut en citer quelques-un : victoire de la France lors de la coupe du monde de football, affaire Benalla, démission de Nicolas Hulot, mouvement des gilets jaunes, attentat de Strasbourg, Greta Thunberg et les marches pour le climat, affaire de Rugy, grèves dans les hôpitaux et contre la réforme des retraites, incendies de Notre-Dame et de Lubrizol, élections européennes et municipales, début de la pandémie de Covid-19, ... De plus, une captation aléatoire de 1% des tweets mondiaux est effectuée grâce à l'API *sample* de Twitter. Au total, le corpus contient 6 343 941 893 tweets dont 3 898 250 863 en français (en comptant les *retweet*).

On peut alors établir une liste $\mathcal{C}$ de comptes Twitter actifs sur cette période. En s'inspirant des travaux des auteurs précédemment cités, nous fixons les critères suivants pour qu'un compte soit considéré comme actif : avoir au moins 200 *tweets* captés sur la période dont au moins 50% sont en français (selon la métadonnée de langue fournie par Twitter). Nous avons alors $|\mathcal{C}| = 1\ 220\ 504$. La définition de compte Twitter actif répond à notre contrainte de créer un espace pour ensuite pouvoir y plonger des contenus à analyser. Contrairement aux auteurs dont la finalité est l'analyse de l'espace politisé, qui peuvent alors ne pas tenir compte de ce critère et considérer tous les comptes Twitter qui suivent des comptes politiques, nous devons pouvoir propager la polarisation politique aux contenus analysés. Cette étape se fait via les comptes qui diffusent effectivement

5. https://developer.twitter.com/en/docs/twitter-api/getting-started/guide

du contenu. Le seuil fixé à 100 tweets captés par an pour qu'un compte soit considéré comme actif nous semble raisonnable (10% des utilisateurs en France). Ce point pourra être approfondi dans de futurs travaux. Nous n'effectuons aucun filtrage supplémentaire, notamment pour éliminer les robots, puisque l'objet de nos travaux est d'observer les échanges de tous les contenus. Nous savons que des réseaux de faux comptes Twitter sont parfois utilisés à des fins politiques, ils peuvent tout à fait entrer dans le cadre de notre étude. Nous n'avons pas non plus cherché à catégoriser géographiquement les comptes Twitter. Certaines métadonnées sont disponibles pour cela mais elles sont soit très peu utilisées (localisation GPS des *tweets*), soit fournies sous forme de texte libre (localisation déclarée par les utilisateurs). Nous avons donc dans notre base d'utilisateurs des comptes de tous les pays francophones.

3.2 Comptes politiques de référence

Pour cette première version, les comptes politiques sont obtenus en fusionnant automatiquement des listes de personnalités politiques créées manuellement provenant de différentes sources. Ces personnalités ne sont pas les seuls comptes à même de nous renseigner sur l'idéologie potentiellement véhiculée. Des militants, des associations, des médias, voire des individus ont parfois un positionnement politique clairement affirmé. Nous les considérerons dans de futurs travaux. Nous utilisons les données issues du site de l'Assemblée Nationale pour les parlementaires élus en 2012 et 2017, les membres du gouvernement, une liste de comptes politiques constituée par le journaliste Adrien Sénécat, les listes du compte Twitter @topolitiq et les données de Briatte and Gallic [2015]. Suite à cette fusion, nous ne conservons que les comptes Twitter encore actifs, soit 7 392 comptes. Nous proposons une approche simple permettant la labellisation automatique des comptes politiques en utilisant un vote majoritaire sur une liste de termes normalisés associés aux différents partis politiques. Ces termes peuvent être présents dans le nom ou la description d'un compte Twitter ou bien dans les métadonnées existantes (groupe politique pour les députés par exemple). Cela nous permet notamment d'étiqueter les comptes des personnalités politiques ayant récemment changé de parti pour aller à LREM. Suite à cette opération, il reste $|\mathcal{P}| = 4\ 824$ comptes pour lesquels nous pouvons déterminer le parti. L'étiquetage des comptes politiques de référence est donc réalisé uniquement sur la base de listes d'autorité constituées manuellement et complété par un vote sur des mots clés caractéristiques. L'activité de ces comptes de référence (tweets et/ou retweets) n'est pas utilisée à ce stade.

Une première difficulté provient du fait que l'identifiant unique d'un compte sur Twitter (*id*) est un entier alors que la plupart du temps, pour des raisons pratiques évidentes, le nom du compte est utilisé (*screenName*). Il est toutefois possible de modifier ce nom. Cette pratique est notamment utilisée par certaines personnalités politiques. Par exemple le compte d'Édouard Philippe (identifiant 1110890216) était @ephilippepm lorsqu'il était premier ministre, il est maintenant renommé en @ephilippe_lh depuis qu'il est maire du Havre. Cet historique des noms de compte n'est pas disponible sur Twitter, seule une captation régulière peut permettre de refaire les associations. La seconde difficulté est due à la nature cumulative et statique des données que nous manipulons alors que l'affiliation

politique est parfois changeante. Typiquement, dans les dernières années la création du parti En Marche a provoqué des migrations d'élus de gauche et de droite vers ce nouveau mouvement. Dans ce contexte, quelle affiliation choisir pour un compte Twitter existant depuis plusieurs années ? L'objectif étant la création d'un espace politisé pour analyser du contenu récent, nous pensons qu'il est souhaitable d'avoir un étiquetage au plus près de la situation politique actuelle. Le détail est présenté dans le tableau 1.

total	ind	lfi	pcf	eelv	dvg	ps	prg	lrem	mod	udi	dvd	lr	dlf	rn
7 392	2 568	325	141	665	40	795	191	418	375	374	46	1 061	33	360
		ext-gauche		eelv	gauche			centre		droite			ext-droite	
		466		665	1026			793		1481			393	

TABLE 1 – Répartition des comptes politiques issus de la fusion de listes manuelles.

Selon la modélisation choisie, on ne conserve que les comptes Twitter qui sont liés à au moins 5 comptes politiques et les comptes politiques étant liés à au moins 25 comptes Twitter. Après filtrage avec les informations de *follow*, pour construire la matrice de contingence $\mathcal{M}^{flw}$, nous avons $|\mathcal{P}^{flw}| = 4\ 377$ comptes politiques et $|\mathcal{C}^{flw}| = 184\ 229$ comptes politisés . Ils ont posté $|\mathcal{T}^{flw}| = 487\ 074\ 392$ *tweets*, ce qui représente 12.5% de $\mathcal{T}$. Pour $\mathcal{M}^{rt}$, nous avons un corpus plus restreint avec $|\mathcal{P}^{rt}| = 1\ 438$ et $|\mathcal{C}^{rt}| = 168\ 393$ mais couvrant davantage de *tweets* : $|\mathcal{T}^{rt}| = 889\ 810\ 466$ soit 22.8% de $\mathcal{T}$. Puisque nous imposons aux comptes d'être liés à au moins 5 comptes politiques français, nous pouvons raisonnablement penser que la très grande majorité des comptes restants après le filtrage sont en France et que les comptes d'autres pays francophones ont été écartés. Cette distinction entre les comptes francophone et les comptes en France, n'est qu'un éclairage sur la composition de notre corpus et non une contrainte. La détermination de la polarisation des contenus sur Twitter peut être faite en tenant compte des utilisateurs présents à l'étranger à partir du moment où ils tweetent en français et qu'ils s'intéressent un minimum à l'actualité politique française (lien avec les comptes politiques de référence).

Nous avons de plus $|\mathcal{C}^{flw} \cap \mathcal{C}^{rt}| = 92\ 949$. Les ensembles de comptes Twitter utilisés sont donc assez dissemblables. Il peut y avoir plusieurs explications à ces différences. L'action de suivre un compte est généralement pérenne dans le temps : on peut avoir décider de suivre un compte pour une raison politique il y a 5 ans, ne pas s'être désabonné mais ne plus spécialement être intéressé par ce qu'il poste depuis. La différence de taille entre $\mathcal{P}^{flw}$ et $\mathcal{P}^{rt}$ est également liée au fait que nous ne captons pas l'intégralité des *tweets*. Même si nous en avons une proportion significative, des *retweets* peuvent être absents de notre corpus et contribuer à réduire les interactions que nous mesurons. Enfin, il peut y avoir diverses raisons poussant à suivre un compte politique dans le but de s'informer mais sans forcément être en accord avec le contenu qu'il diffuse et encore moins souhaiter le *retweeter* à son tour. Nous pensons par exemple aux journalistes qui doivent majoritairement avoir ce type de profil. Une première vérification grossière [6] semble le

6. Nous identifions automatiquement 67 000 comptes comme étant potentiellement des journalistes car contenant dans leur description un des métiers de la profession ou un nom de média. Cette approche

confirmer, nous indiquant une présence d'environ 3% de comptes de journalistes dans $\mathcal{C}^{rt}$ et 8% dans $\mathcal{C}^{flw}$. Le tableau 2 résume les statistiques de ce corpus.

total	lfi	pcf	eelv	dvg	ps	prg	lrem	mod	udi	dvd	lr	dlf	rn
4 377	294	125	575	33	746	139	415	299	360	42	996	31	322
	ext-gauche		eelv	gauche			centre		droite			ext-droite	
	419		575	918			714		1 398			353	

TABLE 2 – Répartition des comptes politiques de référence suite au filtrage.

3.3 Modélisations des comptes politisés

Nous définissons le degré de politisation d'un compte c_i selon les comptes qu'il suit (ses *friends*). On peut y mesurer la proportion de comptes politiques ou à l'inverse quelle proportion de tous les comptes politiques cela représente. Ces deux mesures sont complémentaires et peuvent se combiner :

$$DP^{flw}(c_i) = \frac{1}{2}\left(\frac{|friends(c_i) \cap \mathcal{P}^{flw}|}{|friends(c_i)|} + \frac{|friends(c_i) \cap \mathcal{P}^{flw}|}{|\mathcal{P}^{flw}|}\right)$$

Cette mesure est évidemment incomplète puisque nous ne prétendons pas que les 4 377 comptes couvrent tous les comptes politiques français de Twitter. Ils sont toutefois, par construction, représentatifs et a priori les principales personnalités politiques au sein des partis.

Pour définir un critère de polarisation politique sur un axe classique droite-gauche, nous adoptons une approche similaire à Garimella and Weber [2017] à l'aide d'une combinaison linéaire pondérée du nombre de comptes de chaque parti politique suivi. Un poids est donc attribué à chaque parti, de façon subjective, selon notre propre perception du spectre politique français (tableau 3). À l'extrême-gauche du spectre, il manque dans cette version des comptes représentant les partis LO et NPA qui auraient eu un poids de 0. Une version plus élaborée pourra être envisagée par la suite en tenant compte de travaux de politologues, de sondages ou encore de votes des parlementaires au niveau européen sur des textes caractéristiques. On pourrait également tenir compte de la variance propre à chaque parti.

$$Pol^{flw}_{naif}(c_i) = 2 * (-0.5 + \frac{1}{|friends(c_i) \cap \mathcal{P}^{flw}|} \sum_{p \in friends(c_i) \cap \mathcal{P}^{flw}} weight(p))$$

De manière symétrique, on peut définir DP^{rt} et Pol^{rt}_{naif} en considérant les liens de *retweet*.

Pour la définition de la polarisation selon Barberá, nous utilisons l'approche utilisant une AFC. Elle nécessite une décomposition en valeurs singulières d'une très grande

nécessite d'être affinée avant de pouvoir pleinement être exploitée.

lfi	pcf	eelv	dvg	ps	prg	lrem	modem	udi	dvd	lr	dlf	rn
0.1	0.1	0.25	0.4	0.4	0.4	0.5	0.5	0.6	0.6	0.7	0.9	1.0

TABLE 3 – Poids attribués aux partis politiques pour la version naïve de la polarisation.

matrice. Plutôt qu'une version exacte, il est intéressant d'utiliser une décomposition approximative puisque seules les premières valeurs propres nous intéressent. L'approche de Baglama and Reichel [2005] le permet. Nous la mettons en oeuvre dans notre implémentation de l'AFC en R [7]. Nous conservons les 5 premières dimensions ainsi obtenues (figure 2).

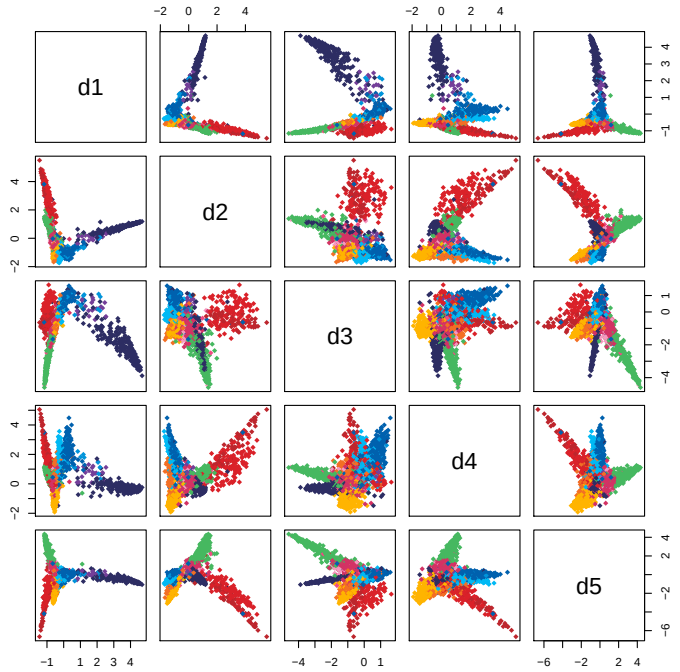

FIGURE 2 – Approche Barberá : 5 premières dimensions de l'espace latent obtenu avec l'AFC sur la matrice de contingence des liens de *follow* pour les 2 235 comptes politiques que nous pouvons projeter.

Dans toutes les représentations des espaces politiques latents nous avons les projections des comptes d'utilisateurs anonymes actifs. Toutefois, il se trouve que certains d'entre eux ne sont pas si anonymes que cela puisqu'ils font également partie de notre ensemble de comptes politiques de référence. Nous utiliserons leur rattachement politique pour mieux visualiser les différentes zones des espaces que nous étudions. Le sens

7. https://github.com/nrv/myCA

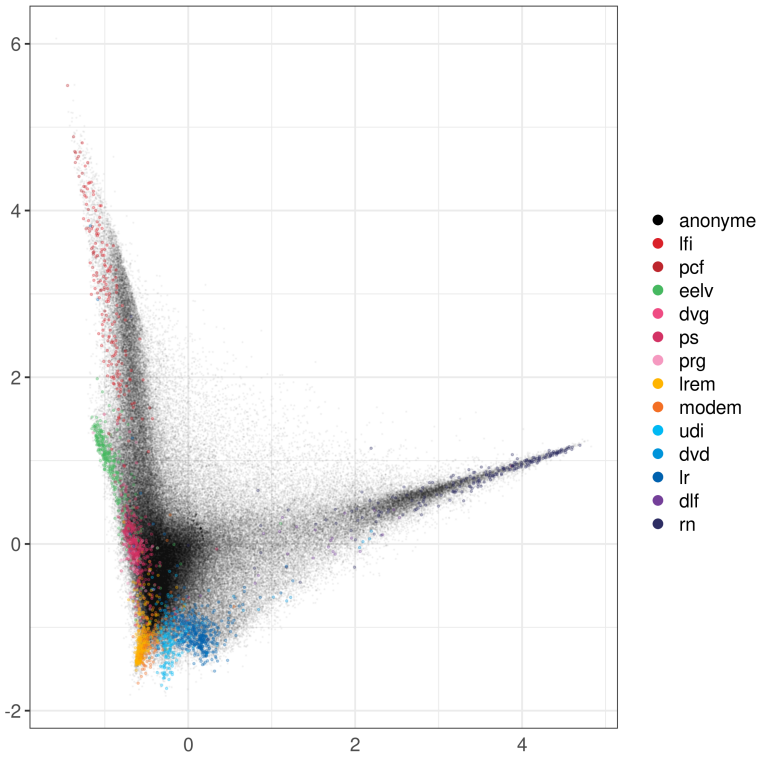

FIGURE 3 – Approche Barberá : 2 premières dimensions de l'espace latent obtenu avec l'AFC sur la matrice de contingence des liens de *follow*.

de la polarité droite-gauche n'est pas trouvé par l'AFC, seul le positionnement relatif des partis est inféré. Pour une meilleure lisibilité des graphiques, nous avons manuellement inversé les axes lorsque cela était nécessaire. Sur les figures 3 et 4 on ne représente que les deux premières, et principales, dimensions des espaces. On y projette tous les comptes politisés. La forme en fer à cheval est assez caractéristique de l'AFC. Pour l'espace basé sur les *followers*, on remarque que la position des comptes politiques est légèrement décentrée. On voit immédiatement que des clusters relativement cohérents sont formés. Il est en revanche plus difficile d'interpréter la signification des axes déterminés par l'AFC. Il semble toutefois bien que les données soient distribuées sur un *manifold* de dimension intrinsèque plus faible. La seule constatation que l'on peut faire à ce stade est qu'on observe un classement des partis politiques selon l'axe traditionnel droite-gauche qui semble préservé en partant de LFI/PC en haut à gauche, en passant par un bloc central proche de l'origine et en finissant par le RN à droite. Briatte and Gallic [2015] font un constat similaire, notamment dans leur mise à jour récente[3]. On retrouve le même ordonnancement des partis sur l'espace basé sur les *retweets*, mais les comptes anonymes y sont plus uniformément répartis et l'espace est clairement tripolaire.

La particularité de l'AFC est de pouvoir plonger les individus et les caractéristiques

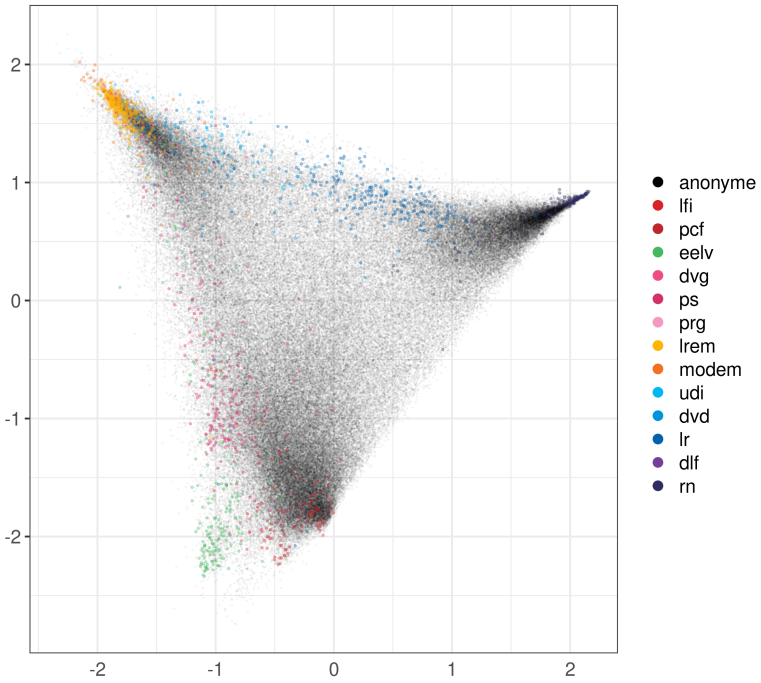

FIGURE 4 – Approche Barberá : 2 premières dimensions de l'espace latent obtenu avec l'AFC sur la matrice de contingence des liens de *retweet*.

dans le même espace latent. Il se trouve toutefois ici qu'avec l'approche de Barberá les caractéristiques que nous utilisons sont elles mêmes directement rattachées à des individus. Ainsi pour les 2 235 comptes politiques qui sont également présents dans notre base d'utilisateurs actifs, nous bénéficions de deux positions dans l'espace latent : une position en tant que caractéristique globale et une en tant qu'individu. Les profils ligne et colonne représentent donc le positionnement des comptes politiques selon qui les suit (colonne) ou selon qui ils suivent eux-mêmes (ligne). Nous représentons ces couples de coordonnées sur la figure 20 et pouvons observer la relative stabilité puisque seuls de rares points y ont un fort déplacement. On remarque toutefois une plus grande cohérence de l'espace basé sur les *follow* que celui des *retweets*.

3.4 Comparaison des différents espaces politisés

Pour mieux visualiser ces espaces latents, on peut utiliser des approches non-linéaires de réduction de dimensions pour les "déplier" en deux dimensions. De nombreuses approches de ce type sont basées sur la distance entre les points dans l'espace de départ qu'elles cherchent à préserver en construisant un espace de représentation de dimension plus faible. On peut notamment citer tSNE et UMAP parmi ces approches les plus

récentes (figure 19). Elles fournissent des représentations avec les partis politiques clairement regroupés mais l'interprétation des formes obtenues s'apparente plutôt à un test de Rorschach.

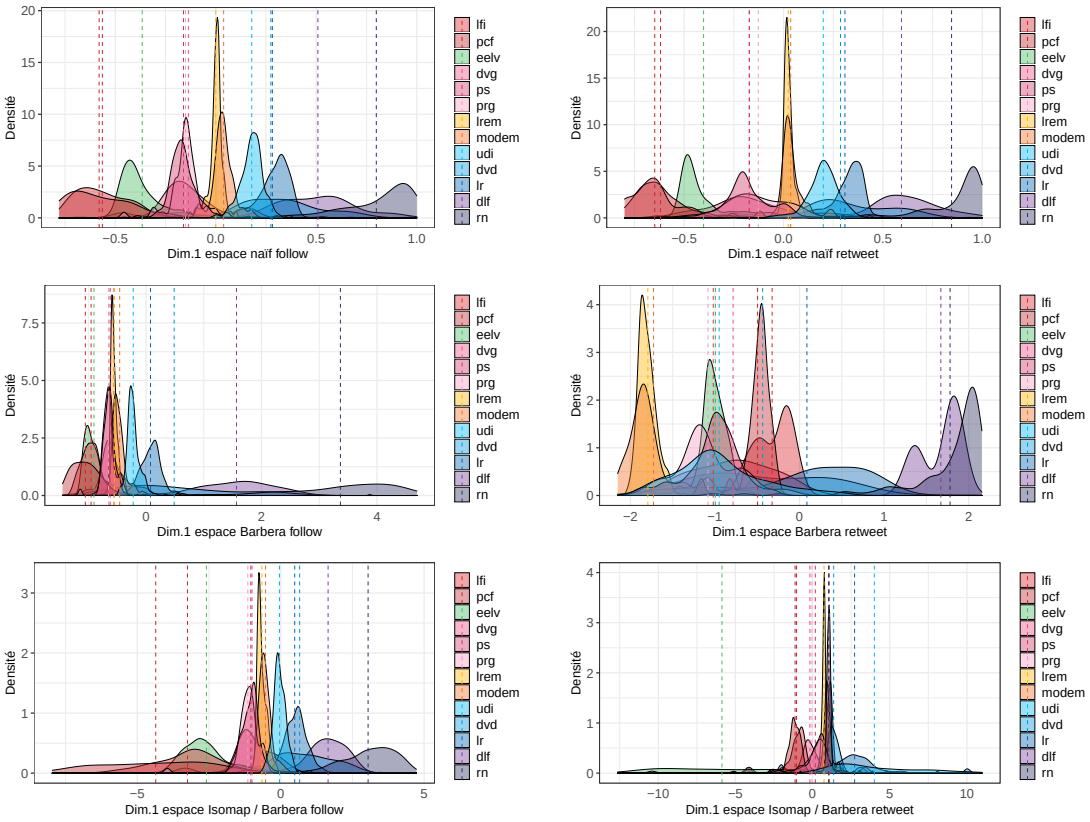

FIGURE 5 – Distribution des comptes politiques selon l'axe droite-gauche pour chaque espace latent calculé.

Nous allons donc plutôt utiliser Isomap [Tenenbaum et al., 2000] pour essayer de retrouver un axe politique droite-gauche plus clairement établi sur une seule dimension. Nous l'appliquons sur les 5 premières dimensions issues de l'AFC. Les distributions des comptes politiques de les figures 5 et 6 font apparaître un positionnement sur l'axe droite-gauche très cohérent pour les deux espaces naïfs. Leur construction a toutefois été légèrement supervisée avec l'utilisation de poids définis manuellement. Pour l'espace Barberá *follow*, on retrouve la projection de la figure 3 avec un fort étalement de l'extrême-droite et une concentration des autres partis. Ils sont toutefois correctement positionnés les uns par rapport aux autres. L'utilisation d'Isomap permet bien de mieux répartir les comptes sur la première dimension de l'espace intrinsèque. En revanche, la structure en forme de triangle de l'espace Barberá *retweet* ne permet pas d'avoir une projection correcte sur une seule dimension.

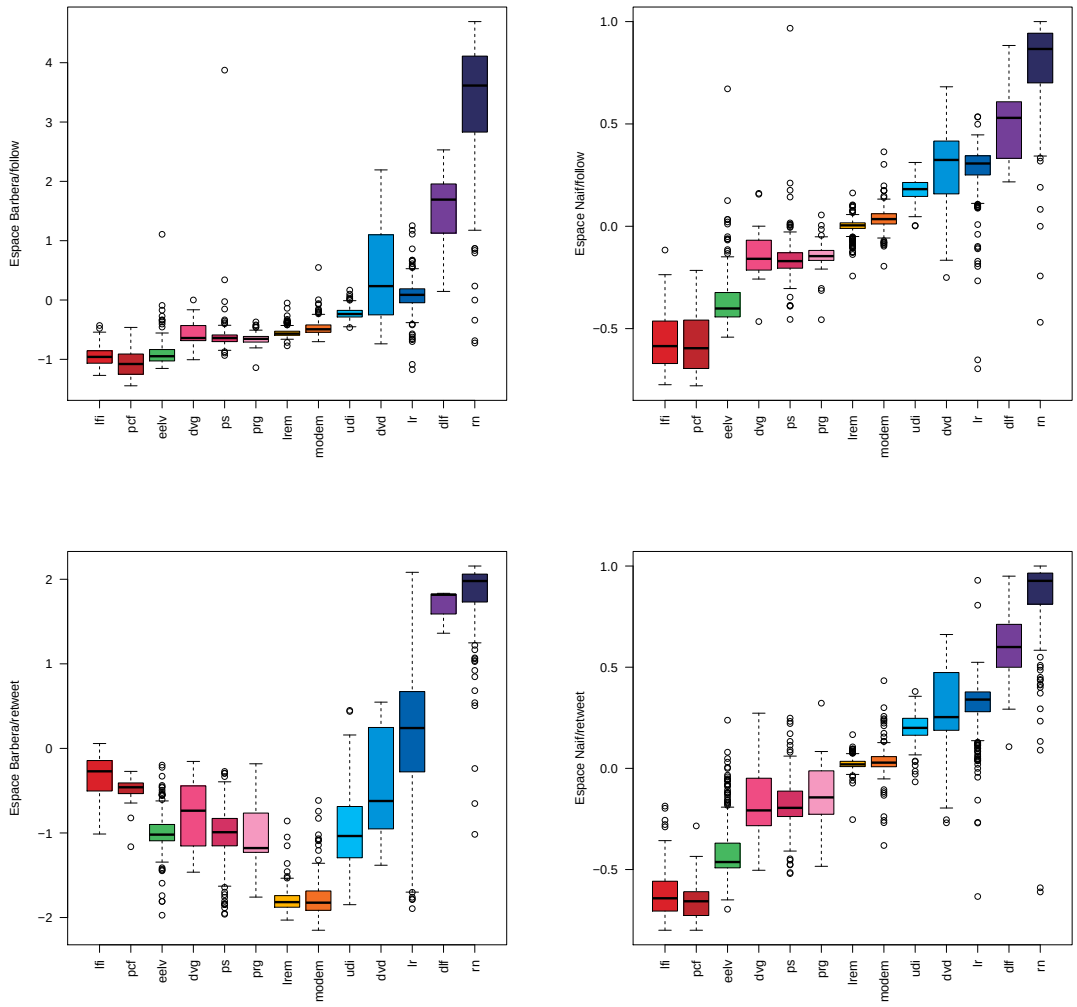

FIGURE 6 – Distribution des comptes politiques selon l'axe droite-gauche pour chaque espace latent calculé.

On compare maintenant les distributions des 92 949 comptes anonymes communs à tous les espaces sur leur dimension droite-gauche (figure 7). Le plus forte corrélation est entre les espaces naïf et Isomap basés sur la matrice de *followers*. Les espaces naïfs *follow* et *retweet* sont également fortement corrélés. Comme pour la distribution des partis, on voit que les espaces Barberá et Isomap issus des *retweets* sont moins en phase avec les autres. Enfin, à titre de validation, on observe que la distribution des comptes sur Pol^{rt}_{naif} est très similaire aux données du sondage Reuters Institute pour les utilisateur français de Twitter ($\rho = 0.940$). Elle l'est moins pour Pol^{flw}_{naif} ($\rho = 0.909$).

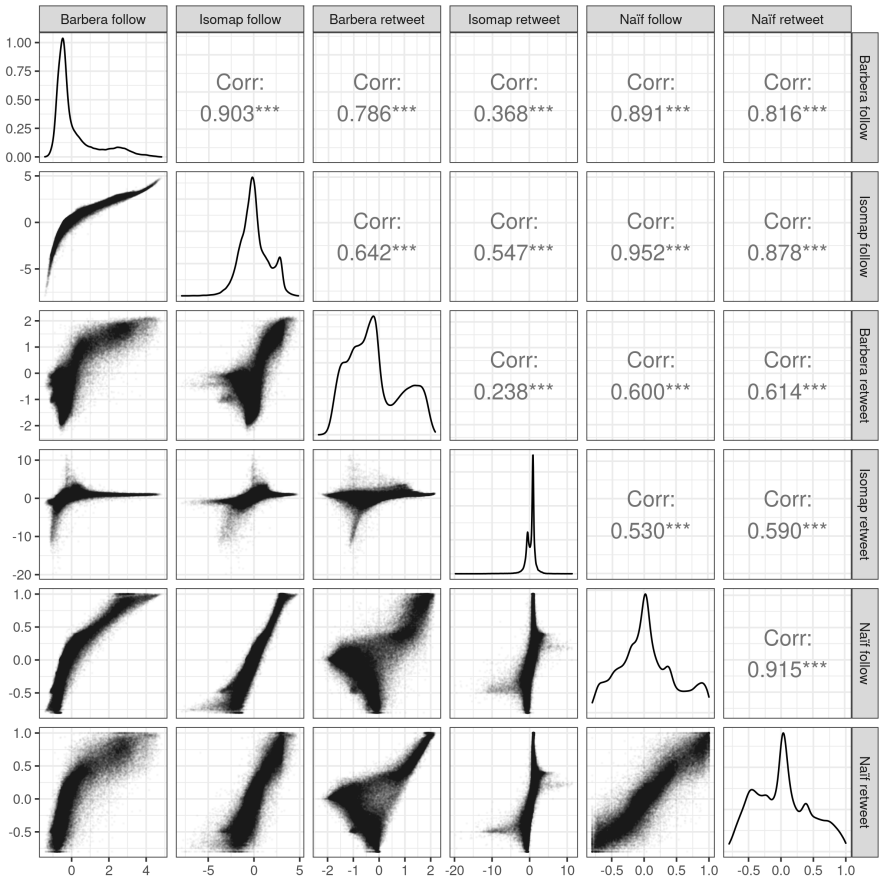

FIGURE 7 – Comparaison de la première dimension des espaces naïfs, Barberá et Isomap pour les 92 949 comptes anonymes communs.

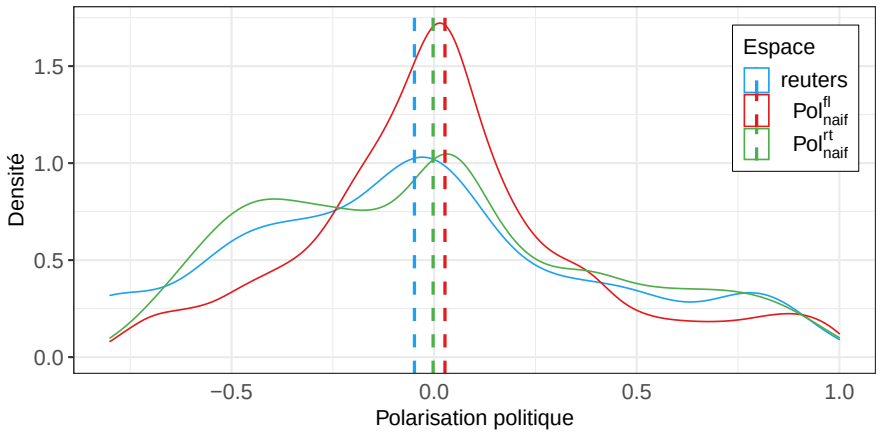

FIGURE 8 – Comparaison des distributions de tous les comptes des deux espaces naïfs avec l'estimation de la population Twitter française selon le Reuters Institute.

3.5 Visualisation des espaces Barberá

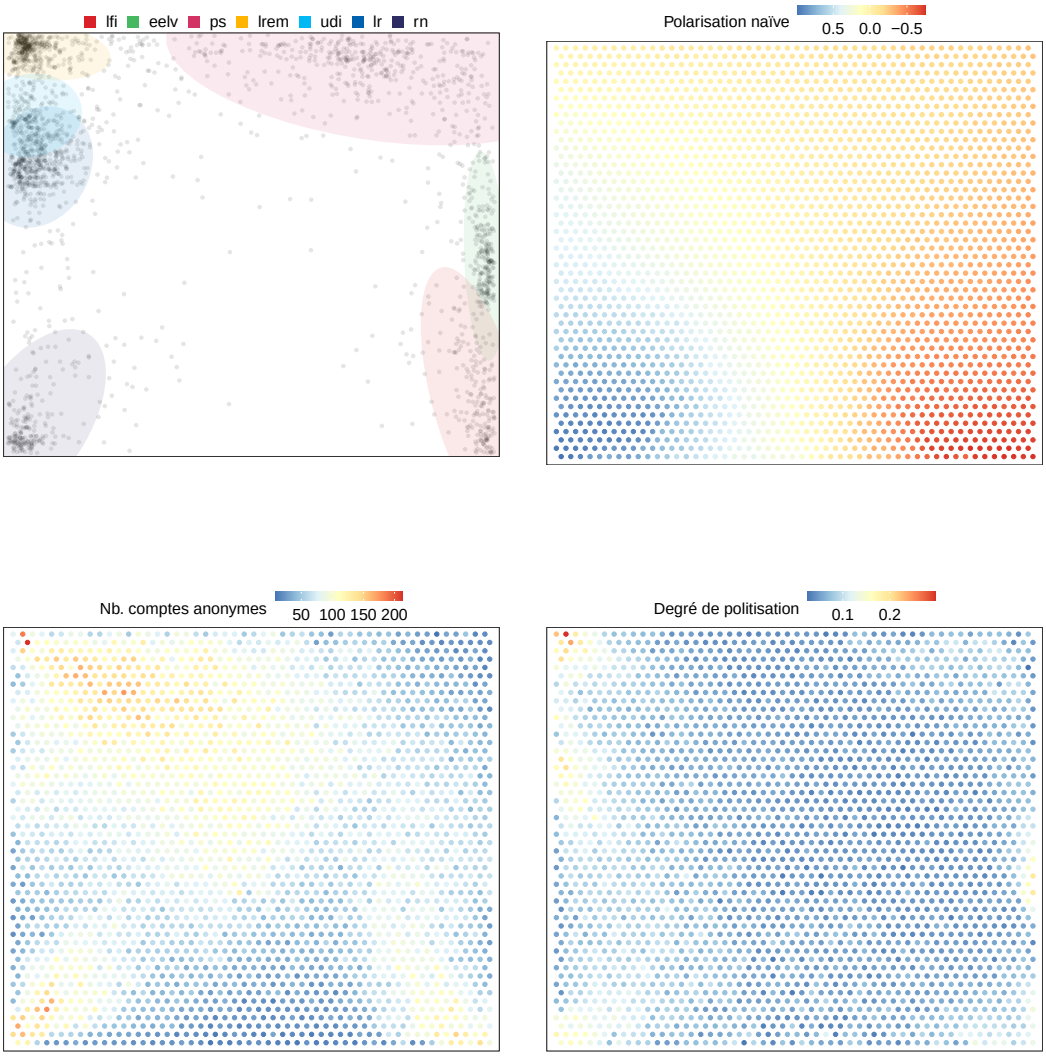

FIGURE 9 – Cartes de Kohonen de l'espace de Barberá *follow*.

On a vu que les espaces de Barberá ne peuvent être complètement ramenés à une seule dimension, notamment celui construit sur les relations de *retweet*. Dans ces espaces latents les données sont assez ramassées et ne permettent pas de tirer pleinement partie de tout l'espace offert pour une visualisation en 2D. De plus, certaines zones sont particulièrement denses. C'est pourquoi nous allons plutôt utiliser une visualisation basée sur les cartes de Kohonen [Kohonen, 1997] qui, dans la même logique que les approches précédentes

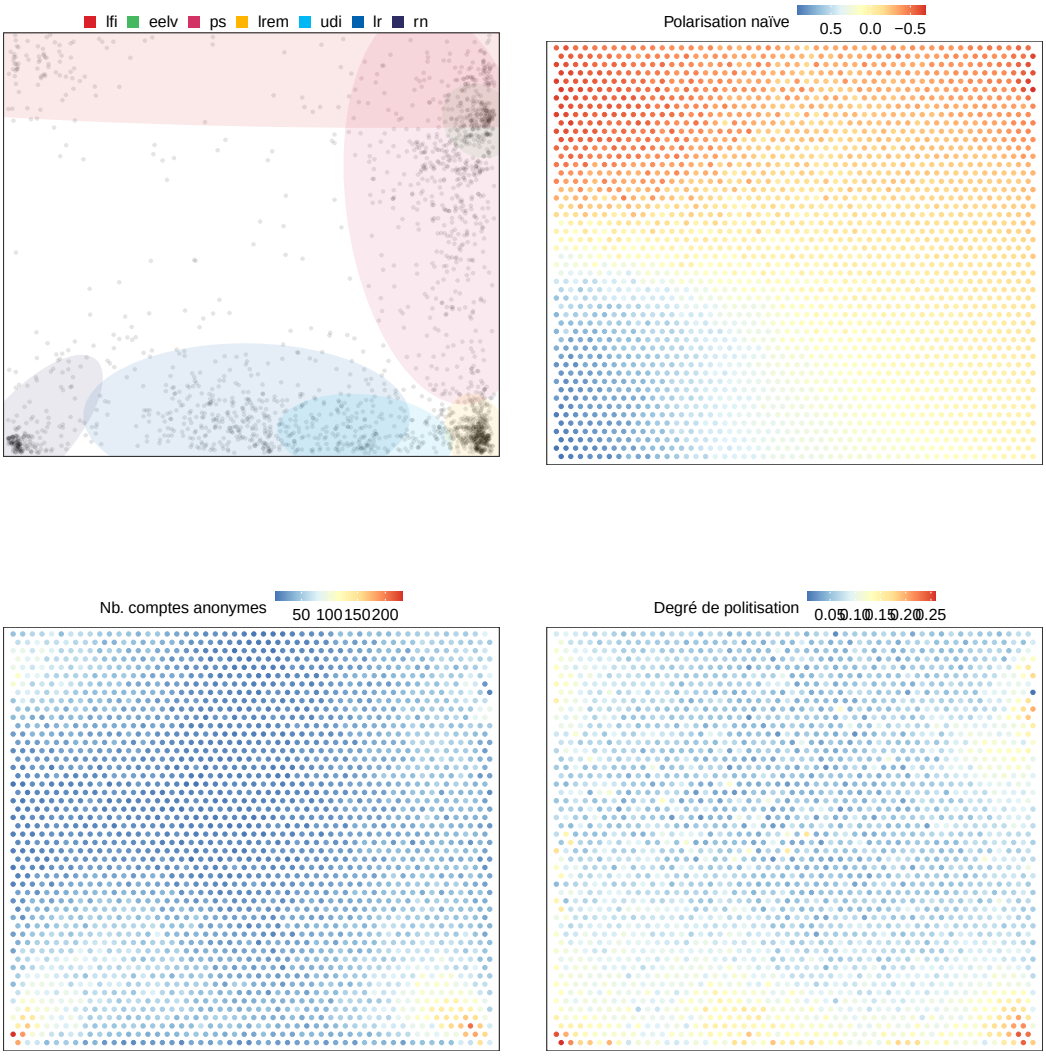

FIGURE 10 – Cartes de Kohonen de l'espace de Barberá *retweet* (bas).

de réduction de dimension basées sur les distances entre points, ajoute une contrainte de couverture de l'espace 2D. Ces cartes permettent de répartir des données complètement sur une grille en 2D. La visualisation des données est ainsi simplifiée. Nous utilisons une grille de 2 500 cellules. Puisque l'espace est déformé de façon non-homogène on se gardera toutefois d'interpréter les distances entre les points. Pour une meilleure lisibilité, un léger *jitter* est appliqué afin de différencier les comptes attribués à la même cellule de la carte. Afin de rendre cette visualisation plus complète, on fait figurer une estimation par une gaussienne de la zone correspondant aux principaux partis politiques, basée sur les

comptes pour lesquels nous disposons de l'information. On retrouve l'ordonnancement des partis politiques sur les pourtours des deux cartes qui concentrent les comptes politiques de $\mathcal{P}^{flw}$ et $\mathcal{P}^{rt}$ et, plus globalement, les comptes avec un fort degré de politisation DP^{flw} et DP^{rt}. On remarque que c'est le sommet LFI/EELV du triangle de l'espace Barberá *retweet* qui a été étirée par la carte de Kohonen. Les centres des cartes sont constitués de comptes anonymes ayant une polarisation centrale.

4 Visualisation de la polarisation des *hashtags*

Le corpus dont nous disposons permet d'analyser la propagation de tous les types de contenus qui sont échangés sur Twitter. On associe à chaque *tweet* la polarisation politique du compte qui l'a posté. Nous nous focalisons sur les *hashtags*. Ils permettent aux utilisateurs d'encoder dans un token unique une signification, potentiellement partisane, qui est partagée par une communauté. L'interprétation de leur propagation est ainsi simplifiée et correspond à l'usage que nous souhaitons en faire ici : une première validation de l'espace politisé et un cadre d'analyse dans cet espace et dans le temps. Pour chaque espace, la représentation d'un *hashtag* est alors simplement la moyenne des *tweets* dans lesquels il apparaît.

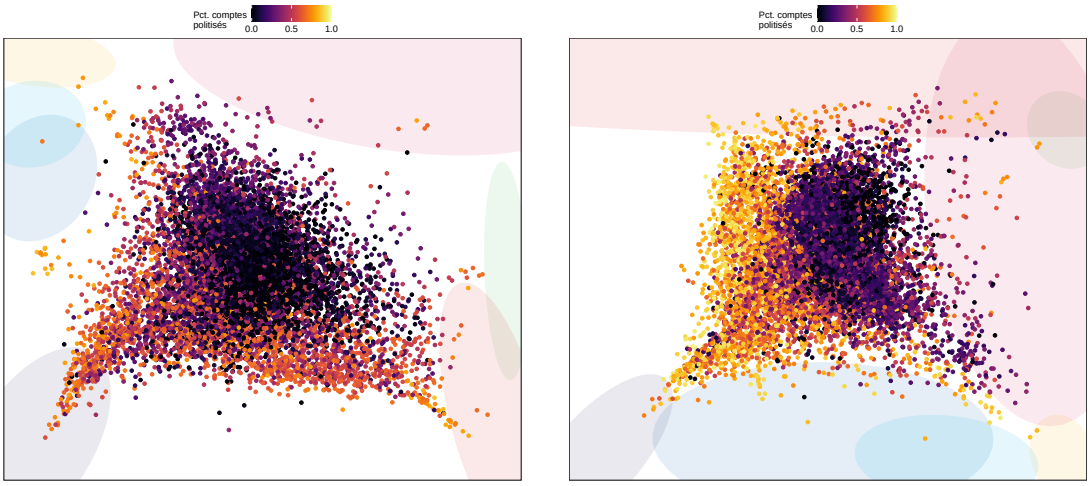

FIGURE 11 – Principaux *hashtags* sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

Certains *hashtag* sont propres à un événement particulier et ne sont donc utilisés que quelques jours avant de disparaître de Twitter. D'autres, au contraire, sont plus génériques et, même s'ils peuvent connaître des pics d'activité, sont régulièrement *tweetés*.

Pour qu'un *hashtag* soit dans notre corpus, il doit satisfaire une des deux conditions suivantes : être dans le top 100 des *hashtags* captés sur une journée ou être dans le top 1 000 des *hashtags* sur l'ensemble de la période. Cela représente un ensemble de 13 377 *hashtags* qui sont détectés dans 13.15% des *tweets*. Les *hashtags* ayant le plus de *tweets* captés dans notre corpus sont en annexe (page 29).

En plus de son positionnement dans un espace latent, nous avons pour chaque *hashtag* la distribution temporelle des *tweets* l'utilisant ainsi que la proportion de *tweets* issus de comptes politisés par rapport à l'ensemble des comptes Twitter. On commence par regarder globalement les *hashtags*. Nous ne conservons que les 12 267 ayant été *tweetés* par au moins 10 comptes politisés. On les représente sur les deux cartes de Kohonen issues des espaces de Barberá (figure 11). On observe dans les deux cas que les *hashtags* qui sont *tweetés* par une proportion importante de comptes de $\mathcal{C}^{flw}$ et $\mathcal{C}^{rt}$ sont sur un axe LFI/RN et dans une moindre mesure sur le second axe RN/LREM. Au centre on retrouve la plupart des *hashtags* pour lesquels la grande majorité des comptes qui les utilisent ne sont pas dans $\mathcal{C}^{flw}$ et $\mathcal{C}^{rt}$, c'est-à-dire des comptes peu actifs ou bien peu liés à nos comptes politiques de référence. On peut raisonnablement en déduire que ces *hashtags* ne sont pas connotés politiquement. On remarque enfin une plus forte proportion de comptes utilisant ces *hashtags* dans $\mathcal{C}^{rt}$ que dans $\mathcal{C}^{flw}$.

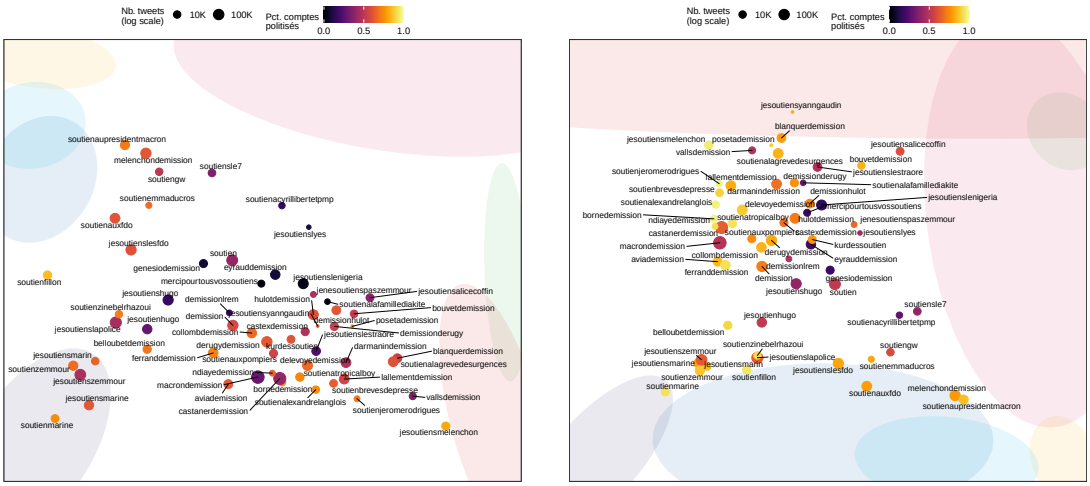

FIGURE 12 – *Hashtags* contenant 'soutien' ou 'demission' sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

Nous regardons maintenant certains *hashtags* qui sont *a priori* partisans et que l'on s'attend à retrouver dans des régions des espaces latents éloignées du centre. Un premier ensemble est constitué de tous les *hashtags* contenant 'soutien' ou 'demission', ils sont régulièrement utilisés pour marquer son soutien ou son opposition à une personne. Un

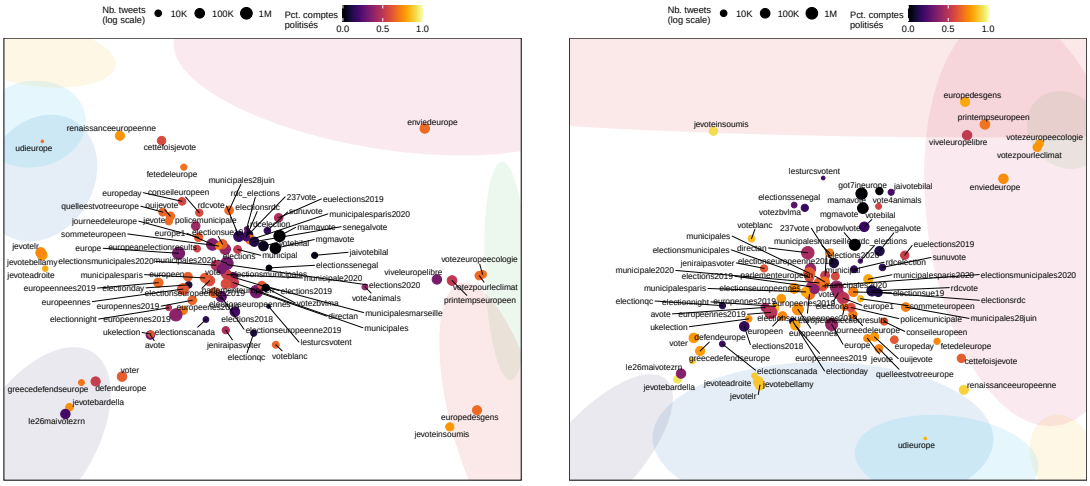

FIGURE 13 – *Hashtags* contenant 'vote', 'election', 'europe', 'municipale' ou 'directan' sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

second ensemble est lié au processus électoral, par nature partisan, avec les termes 'vote', 'europe', 'election', et 'municipal' afin de couvrir les deux élections qui se sont tenues sur la période de notre corpus ainsi que le *hashtag* #directan qui concerne l'actualité des débats à l'Assemblée Nationale. Parmi les *hashtags* très proches des zones identifiées des partis, on retrouve naturellement ceux qui ont été utilisés pour les campagnes électorales (#envideurope, #renaissanceeuropeenne, #votezpourleclimat, #jevoteinsoumis, ...) ou liés à certaines personnalités (#soutienfillon, #soutienmarine, ...). Plus au centre, on a des *hashtags* ayant suscité de nombreux *tweets* mais ne concernant que très peu de comptes politisés tels que #got7ineurope (saison 7 de la série TV *Game of Thrones*) ou #jesoutienslenigeria. On a également les *hashtags* dont la distribution peut être multimodale, ce qu'une simple moyenne ne peut faire apparaître. À titre d'exemple, nous introduisons la visualisation temporelle de #directan dans la figure **??**. Nous rappelons que la distribution des comptes sur l'axe droite-gauche n'est pas uniforme (voir figure **??**), ce qui apparaît bien ici selon que l'on observe simplement le nombre de *tweets* ou bien le pourcentage de comptes politisés ayant utilisé #directan chaque jour. La dynamique temporelle d'utilisation du *hashtag* apparaît, ainsi que la distribution sur l'échelle de polarisation politique.

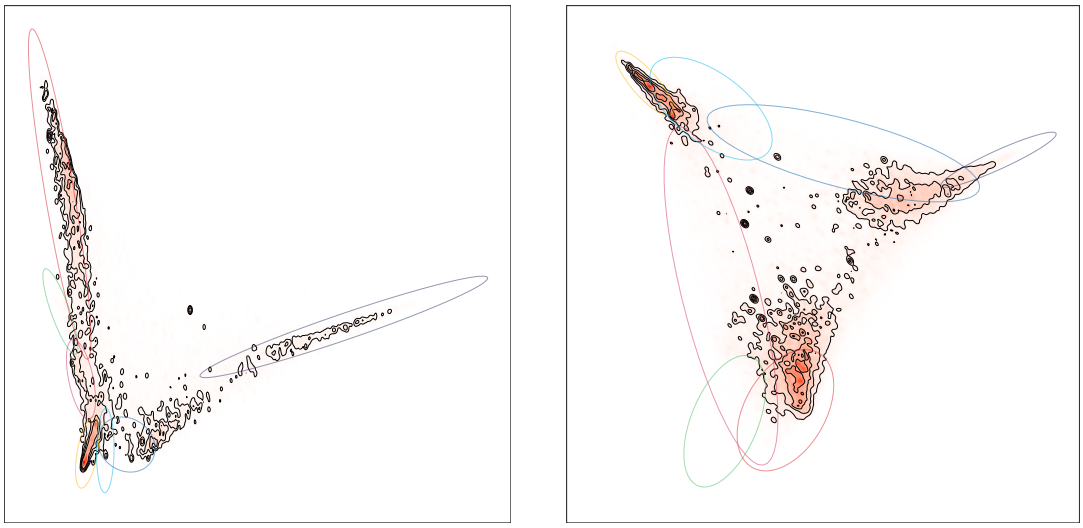

FIGURE 14 – Densité de tweets utilisant le *hashtag* #directan sur les espaces de Barberá *follow* (gauche) et *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

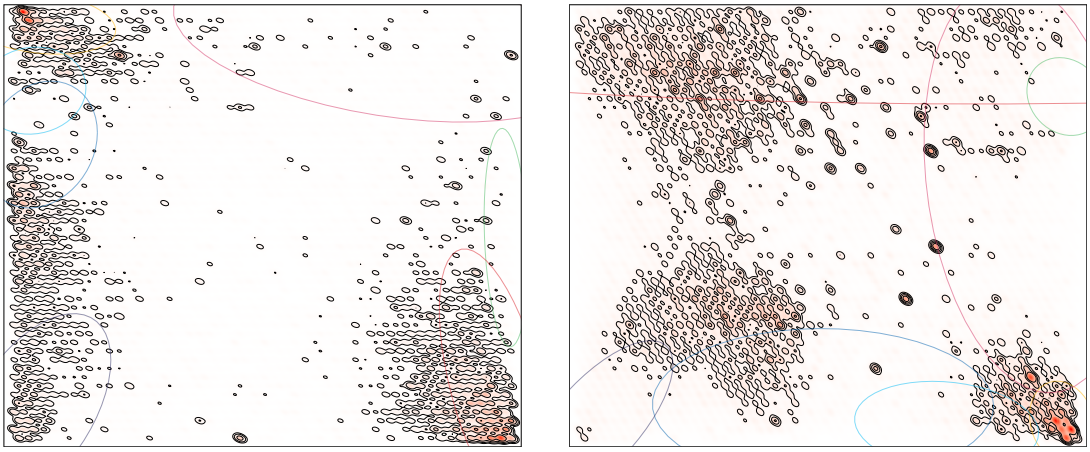

FIGURE 15 – Densité de tweets utilisant le *hashtag* #directan sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

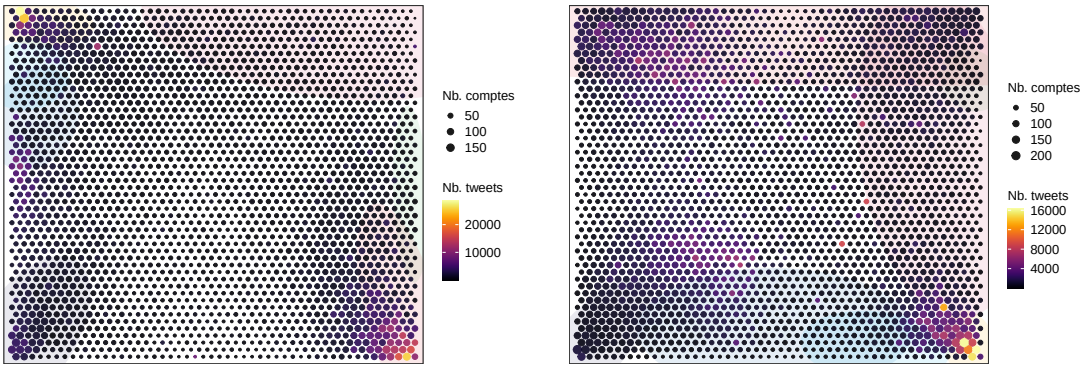

FIGURE 16 – Évolution spatiale des caractéristiques de #directan sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

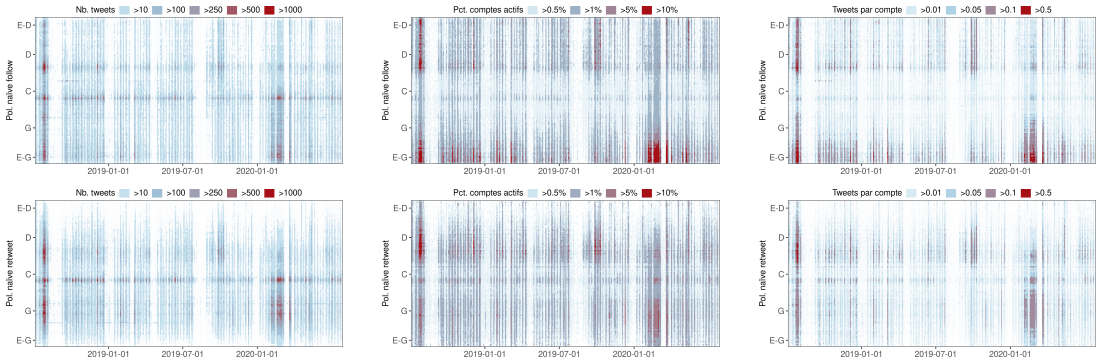

FIGURE 17 – Évolution temporelle des caractéristiques de #directan sur les espaces naïfs *follow* (haut) et *retweet* (bas).

Remerciements

Nous tenons fortement à remercier Béatrice Mazoyer pour avoir conçu et développé la solution de captation de tweets, Julia Cagé pour l'accès aux données du Reuters Institute, ainsi que François Briatte et Ewen Gallic pour les échanges fructueux que nous avons eus et enfin Sabrina pour sa relecture attentive.

5 Conclusion

Nous avons modélisé l'espace politisé de Twitter avec l'approche de Barberá et une pondération naïve nous basant sur les liens de *follow* ou de *retweet* entre les comptes Twitter et des comptes politiques de référence. L'utilisation des *retweets* n'a, à notre connaissance, jamais été utilisée pour déterminer un espace latent. La validation formelle de la polarisation politique des comptes Twitter n'est pas directement accessible faute de données nominatives disponibles (et c'est heureux). Aussi, comme les travaux pour d'autres pays, nous ne pouvons procéder qu'à une validation partielle. Nous avons deux approches inédites : corrélation avec des données issues d'un sondage auprès d'utilisateurs Twitter français et analyse qualitative de la polarisation de hashtags que l'on sait être très partisans.

Si on ne s'intéresse qu'à une modélisation sur une dimension unique de l'axe droite-gauche, alors l'utilisation de Barberá, éventuellement complétée d'Isomap, permet d'avoir un bon résultat avec les liens de *follow*. Toutefois, l'approche naïve fournit des résultats équivalents pour une mise en œuvre plus simple et plus souple. Nous avons vu que la distribution des comptes obtenue selon cette polarisation correspond à celle d'un sondage pour les utilisateurs français de Twitter. Si on souhaite plutôt baser nos analyses sur les liens de *retweet*, alors il conviendra de conserver au moins deux dimensions de l'espace latent. Au delà de l'interprétation des représentations, nous avons également vu que pour une même base de comptes politiques l'utilisation des informations de *follow* ou *retweet* conduit à caractériser des ensembles de comptes Twitter différents et donc à pouvoir observer des contenus différents. Il est important de bien cerner ce périmètre avant d'interpréter les polarisations induites pour les contenus.

Nous confirmons que l'espace politique est tripolaire sur l'espace francophone. Cela apparaît clairement en utilisant les *retweet*, mais c'est également bien visible avec les *follow*. Ces trois pôles sont l'extrême-droite, le centre et l'extrême-gauche. Les portions de l'espace politique qui correspondent à la droite et la gauche classiques (PS, LR) tendent à être moins représentés et moins actifs.

Un contenu à analyser se ramène finalement à un simple ensemble de *tweets* judicieusement agrégés. Nous avons présenté quelques prémices de résultats et de visualisations en considérant les *hashtags* et avons observé un positionnement cohérent dans les espaces latents pour certains d'entre eux, *a priori* partisans. On peut appliquer la même méthodologie à des contenus différents : événements médiatiques, URLs d'article de presse, contenus multimédia, *mentions* de certains comptes, thématiques définies par des mots clés, ... Ce sera l'objet de futurs travaux.

Références

[Baglama and Reichel 2005] Baglama, James ; Reichel, Lothar : Augmented implicitly restarted Lanczos bidiagonalization methods. In : *SIAM Journal on Scientific Computing* 27 (2005), Nr. 1

[Barberá 2015] Barberá, Pablo : Birds of the Same Feather Tweet Together : Bayesian Ideal Point Estimation Using Twitter Data. In : *Political Analysis* 23 (2015), Nr. 1

[Barberá et al. 2015] Barberá, Pablo ; Jost, John T. ; Nagler, Jonathan ; Tucker, Joshua A. ; Bonneau, Richard : Tweeting From Left to Right : Is Online Political Communication More Than an Echo Chamber ? In : *Psychological Science* 26 (2015), Nr. 10

[Benzécri 1973] Benzécri, Jean-Paul : L'Analyse des Correspondances. In : *L'analyse des données* Bd. 2. Dunod Paris, 1973

[Briatte and Gallic 2015] Briatte, François ; Gallic, Ewen : Recovering the French Party Space from Twitter Data. In : *Science Po Quanti.* Paris, France, Mai 2015

[Cagé et al. 2020] Cagé, Julia ; Hervé, Nicolas ; Mazoyer, Beatrice : Social Media and Newsroom Production Decisions / Social Science Research Network. Rochester, NY, Juli 2020 (ID 3663899). – SSRN Scholarly Paper

[Cardon et al. 2019] Cardon, Dominique ; Cointet, Jean-Philippe ; Ooghe, Benjamin ; Plique, Guillaume : Unfolding the multi-layered structure of the French mediascape. (2019)

[Chavalarias et al. 2019] Chavalarias, David ; Gaumont, Noé ; Panahi, Maziyar : Hostilité et prosélytisme des communautés politiques. Le militantisme politique à l'ère des réseaux sociaux. In : *Réseaux* 214-215 (2019), Nr. 2-3. – ISSN 9782348043581

[Cohen and Ruths 2013] Cohen, Raviv ; Ruths, Derek : Classifying political orientation on Twitter : It's not easy ! In : *Seventh international AAAI conference on weblogs and social media*, 2013

[Conover et al. 2011a] Conover, Michael D. ; Gonçalves, Bruno ; Ratkiewicz, Jacob ; Flammini, Alessandro ; Menczer, Filippo : Predicting the political alignment of twitter users. In : *IEEE third international conference on social computing*, 2011

[Conover et al. 2011b] Conover, Michael D. ; Ratkiewicz, Jacob ; Francisco, Matthew R. ; Gonçalves, Bruno ; Menczer, Filippo ; Flammini, Alessandro : Political Polarization on Twitter. In : *Fifth International Conference on Weblogs and Social Media*, , 2011

[Darmon et al. 2015] Darmon, David ; Omodei, Elisa ; Garland, Joshua : Followers are not enough : A multifaceted approach to community detection in online social networks. In : *PloS one* 10 (2015), Nr. 8

[Garimella and Weber 2017] Garimella, Venkata Rama K.; Weber, Ingmar : A long-term analysis of polarization on Twitter. In : *11th International Conference on Web and Social Media, ICWSM 2017*, 2017

[Halberstam and Knight 2016] Halberstam, Yosh; Knight, Brian : Homophily, group size, and the diffusion of political information in social networks : Evidence from Twitter. In : *Journal of public economics* 143 (2016), S. 73–88. – Publisher : Elsevier

[Kohonen 1997] Kohonen, Teuvo : Exploration of very large databases by self-organizing maps. In : *International conference on neural networks (icnn'97)* Bd. 1, 1997

[Mazoyer et al. 2018] Mazoyer, Béatrice; Cagé, Julia; Hudelot, Céline; Viaud, Marie-Luce : Real-time collection of reliable and representative tweets datasets related to news events. In : *BroDyn 2018, co-located with ECIR 2018*, 2018

[Ramaciotti Morales et al. 2020] Ramaciotti Morales, Pedro; Cointet, Jean-Philippe; Laborde, Julio : Your most telling friends : Propagating latent ideological features on Twitter using neighborhood coherence. In : *ASONAM 2020 - IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2020

[Severo and Lamarche-Perrin 2018] Severo, Marta; Lamarche-Perrin, Robin : L'analyse des opinions politiques sur Twitter. In : *Revue française de sociologie* 59 (2018), Nr. 3

[Tenenbaum et al. 2000] Tenenbaum, Joshua B.; De Silva, Vin; Langford, John C. : A global geometric framework for nonlinear dimensionality reduction. In : *Science* 290 (2000), Nr. 5500

[Tucker et al. 2018] Tucker, Joshua A.; Guess, Andrew; Barberá, Pablo; Vaccari, Cristian; Siegel, Alexandra; Sanovich, Sergey; Stukal, Denis; Nyhan, Brendan : Social media, political polarization, and political disinformation : A review of the scientific literature. (2018)

[Yardi and Boyd 2010] Yardi, Sarita; Boyd, Danah : Dynamic Debates : An Analysis of Group Polarization Over Time on Twitter. In : *Bulletin of Science, Technology & Society* 30 (2010), Oktober, Nr. 5. – ISSN 0270-4676

6 Annexes

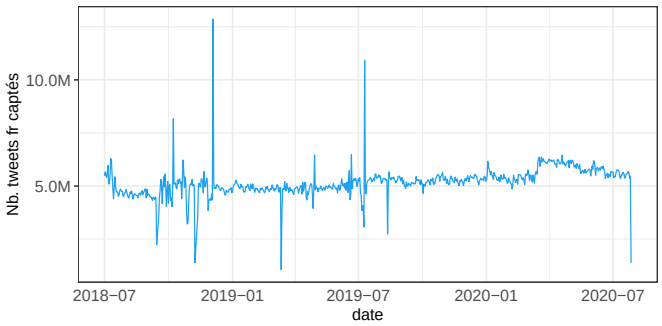

FIGURE 18 – Nombre de tweets captés par jour sur la période.

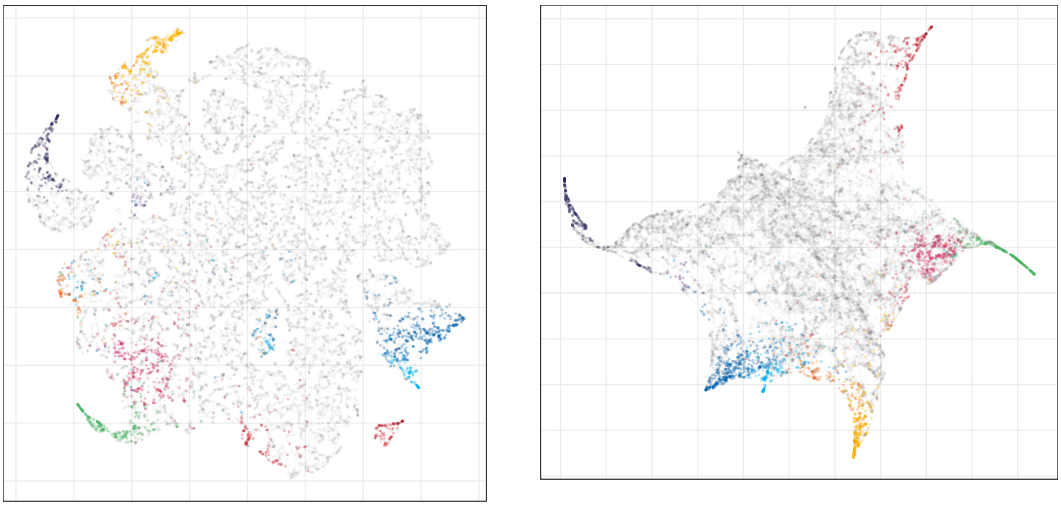

FIGURE 19 – Réduction de dimensions de l'espace Barberá *follow* en utilisant tSNE (gauche) et UMAP (droite). Les 2 235 comptes politiques que nous pouvons projeter sont présentés avec leurs couleurs respectives, un échantillon de comptes anonymes est également positionné.

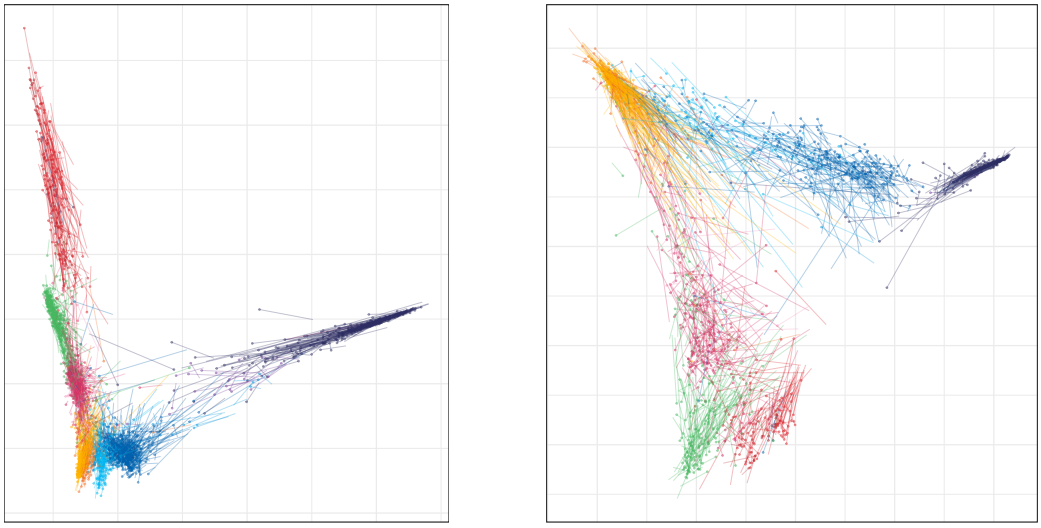

FIGURE 20 – Approche Barberá : profils ligne et colonne sous forme de segments des comptes politiques actifs sur les 2 premières dimensions des espaces latents obtenus avec l'AFC sur les matrices de contingence des liens de *follow* (gauche) et *retweet* (droite).

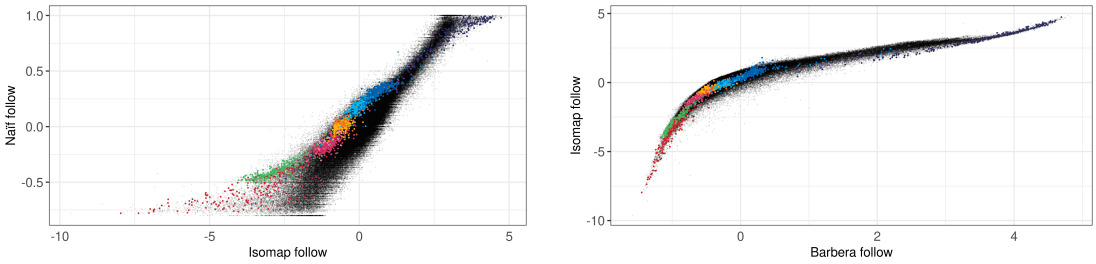

FIGURE 21 – Comparaison de la première dimension des espaces *follow* pour les 92 949 comptes anonymes communs, les comptes politiques sont indiqués avec la couleur de leur parti.

6.1 Premiers *hashtags* de notre corpus

hashtag	Tweets	Comptes	*hashtag*	Tweets	Comptes
#giletsjaunes	33 888 079	1 061 476	#marseille	2 217 396	299 103
#concours	28 767 920	1 144 322	#europeennes2019	2 175 902	167 642
#macron	23 185 019	691 407	#toulouse	2 141 507	264 766
#rt	16 153 404	1 519 832	#nantes	2 140 770	266 365
#covid19	14 176 642	1 294 039	#fiersdetrebleus	2 097 634	441 671
#coronavirus	13 179 868	1 190 881	#lyon	2 095 755	369 760
#paris	9 560 999	1 031 015	#petitpapatopachat	2 074 239	191 512
#france	7 510 839	830 002	#trump	2 046 075	286 167
#lrem	6 143 602	218 267	#fortnite	1 868 121	295 694
#bts	5 698 517	675 092	#quotidien	1 790 390	430 388
#benalla	5 565 348	218 006	#cm2018	1 750 490	334 242
#rdc	4 757 743	178 449	#bordeaux	1 715 924	299 434
#follow	4 228 410	587 042	#wwiii	1 688 513	356 882
#afp	4 148 642	636 707	#bfmtv	1 662 125	166 828
#retraites	4 026 598	215 032	#rn	1 661 679	116 400
#covid_19	3 991 231	684 002	#blacklivesmatter	1 661 213	412 602
#psg	3 975 739	266 772	#migrants	1 648 701	141 551
#kohlanta	3 904 478	415 363	#frabel	1 645 570	397 230
#teamom	3 863 223	229 867	#ggrmc	1 643 227	180 609
#violencespolicieres	3 313 501	336 652	#reformedesretraites	1 635 198	132 183
#directan	3 026 540	282 905	#strasbourg	1 622 428	288 629
#castaner	2 924 697	287 463	#rouen	1 617 907	262 057
#confinement	2 844 697	580 041	#giletjaune	1 616 096	194 350
#teamparieur	2 542 886	140 835	#sante	1 598 739	280 515
#polqc	2 483 144	71 282	#europe	1 575 410	224 967
#jk	2 462 049	486 046	#mercato	1 540 864	157 792
#giveaway	2 429 361	412 744	#fakenews	1 520 509	201 785
#municipales2020	2 353 818	226 938	#twitch	1 488 928	85 074
#tpmp	2 341 248	335 251	#grevegenerale	1 482 919	185 877
#notredame	2 329 454	512 860	#lci	1 468 294	142 051
#zemmour	2 278 133	199 350	#kebetu	1 462 794	92 449
#exo	2 267 597	277 020	#chine	1 461 136	338 366
#jeuconcours	2 265 944	260 212	#cesoirtribunal	1 458 541	190 436
#police	2 242 702	367 430	#rediff	1 439 339	348 468
#algerie	2 223 254	278 572	#macron20h	1 387 578	366 433

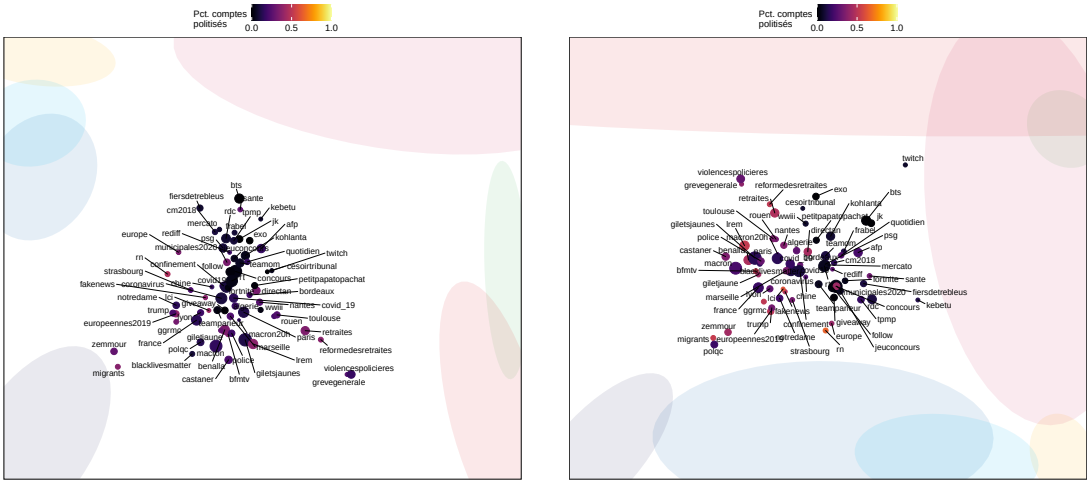

FIGURE 22 – Position moyenne des *hashtags* ayant le plus de tweets captés dans notre corpus sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

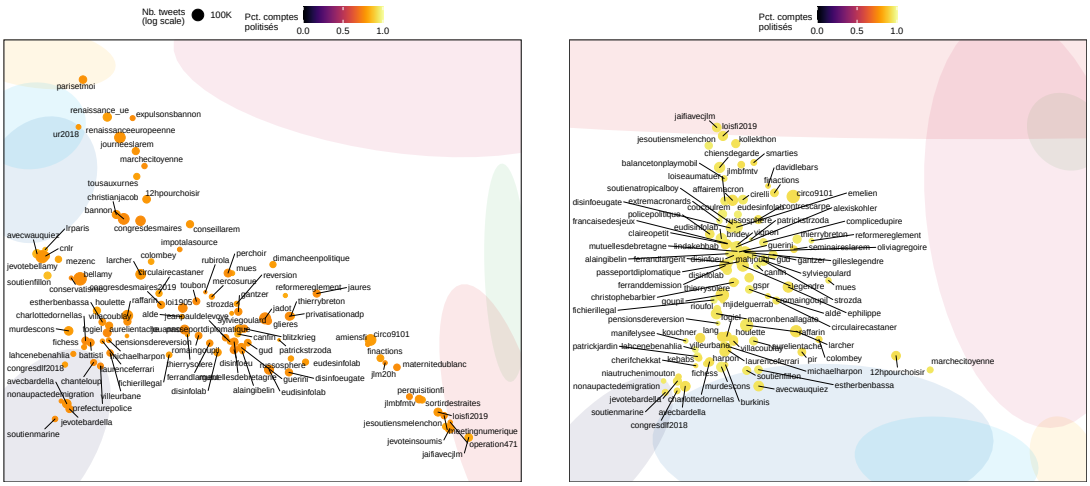

FIGURE 23 – Position moyenne des 100 *hashtags* ayant le plus haut pourcentage de comptes polarisés et au minimum 10 000 tweets captés dans notre corpus sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

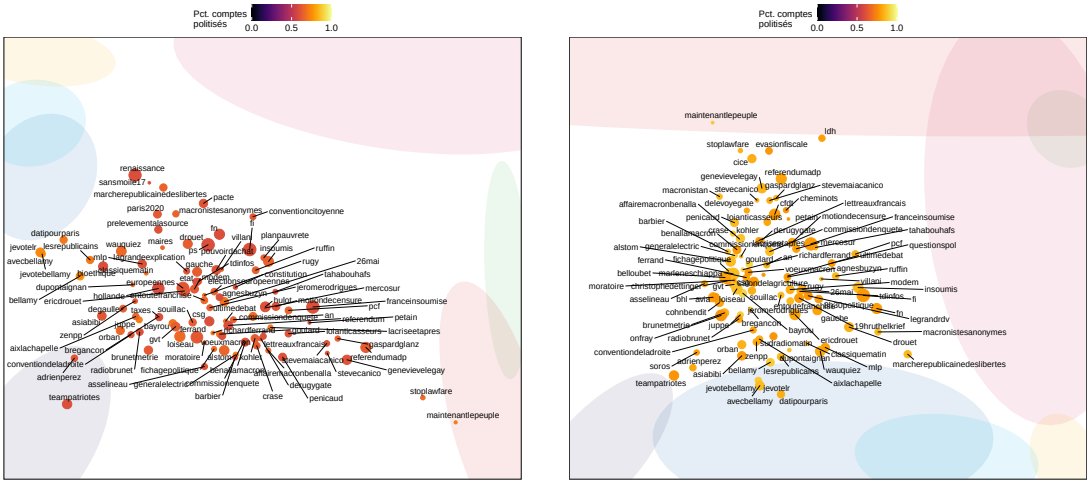

FIGURE 24 – Position moyenne des 100 *hashtags* ayant le plus haut pourcentage de comptes polarisés et au minimum 100 000 tweets captés dans notre corpus sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

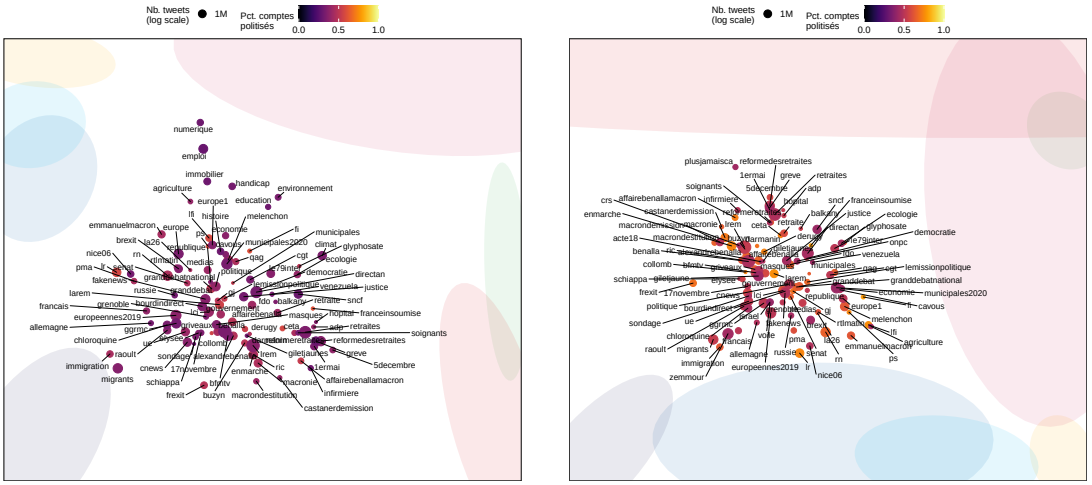

FIGURE 25 – Position moyenne des 100 *hashtags* ayant le plus haut pourcentage de comptes polarisés et au minimum 500 000 tweets captés dans notre corpus sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.2 Visualisation des *hashtags* liés à la police

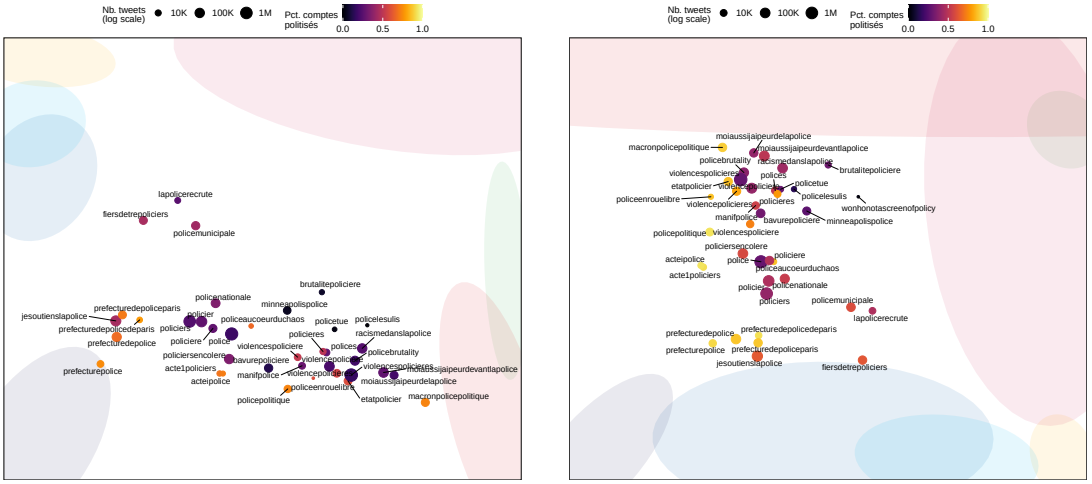

FIGURE 26 – Position moyenne des *hashtags* contenant 'polic' sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.3 Visualisation des *hashtags* liés à l'affaire Benalla

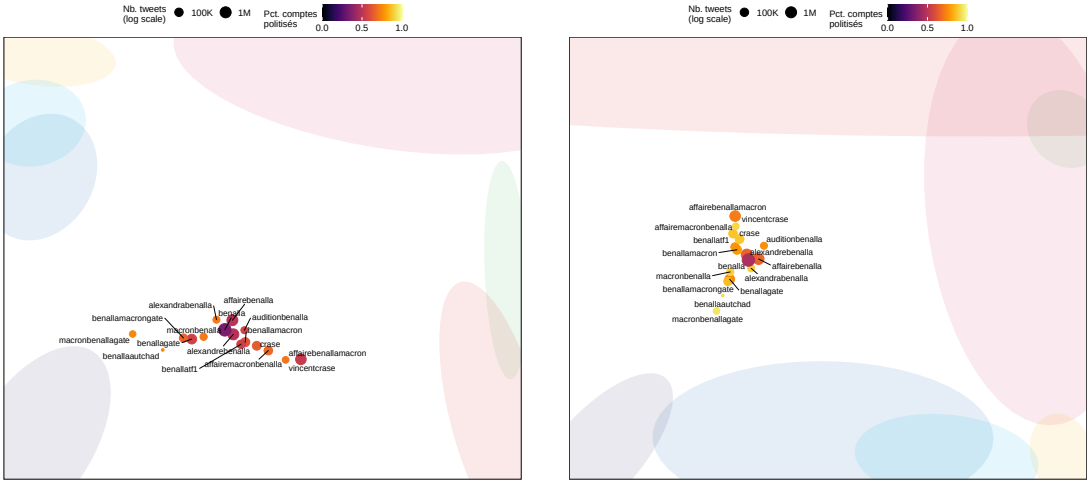

FIGURE 27 – Position moyenne des *hashtags* liés à l'affaire Benalla sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

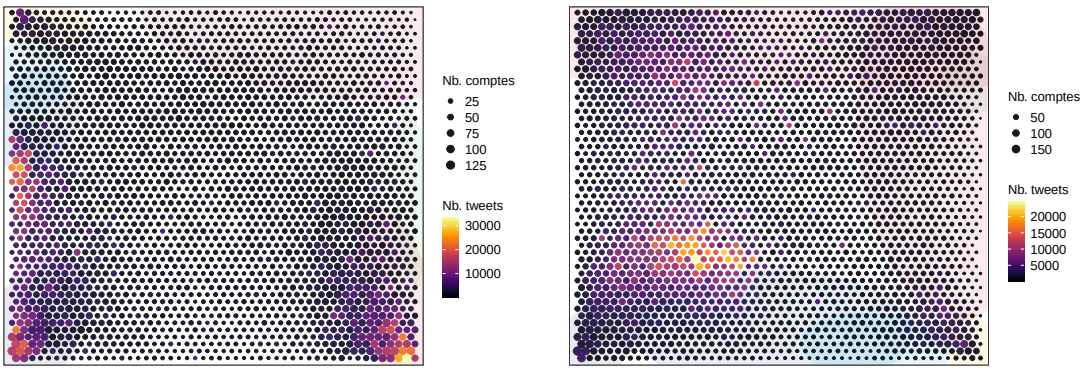

FIGURE 28 – *Hashtag* #benalla : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

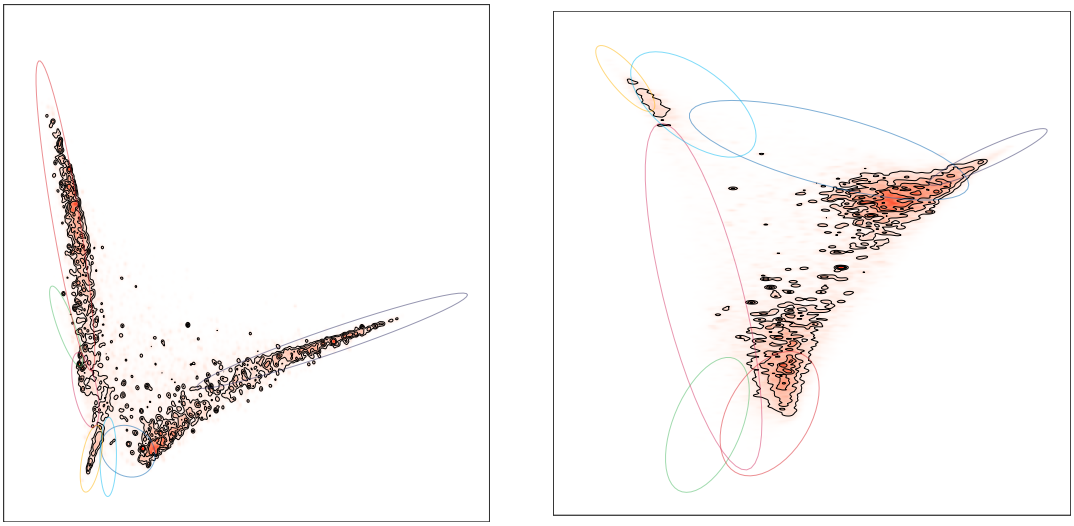

FIGURE 29 – Densité de tweets utilisant le *hashtag* #benalla sur les espaces de Barberá *follow* (gauche) et *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

6.4 Visualisation des *hashtags* liés à la mort d'Adama Traoré

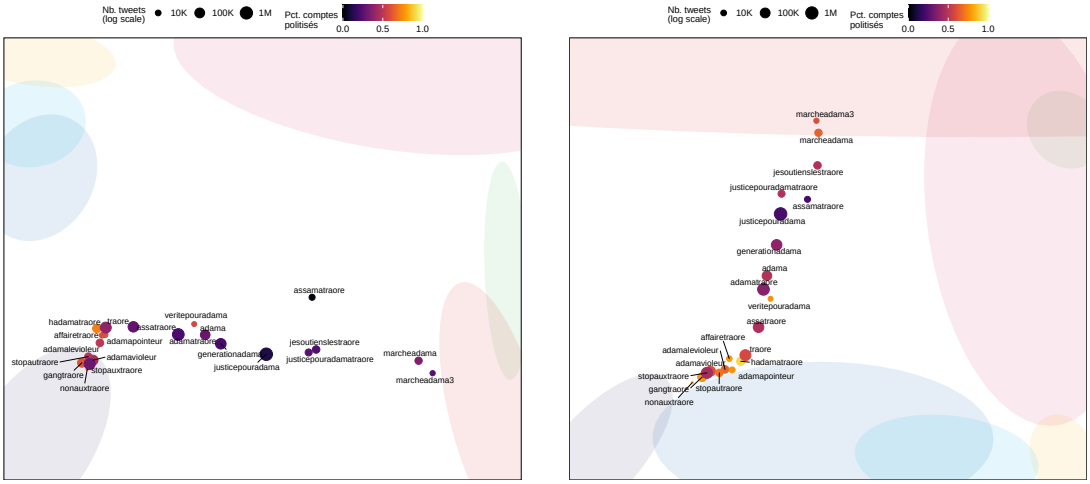

FIGURE 30 – Position moyenne des *hashtags* liés à la mort d'Adama Traoré sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.5 Visualisation des *hashtags* liés à Jean-Luc Mélenchon

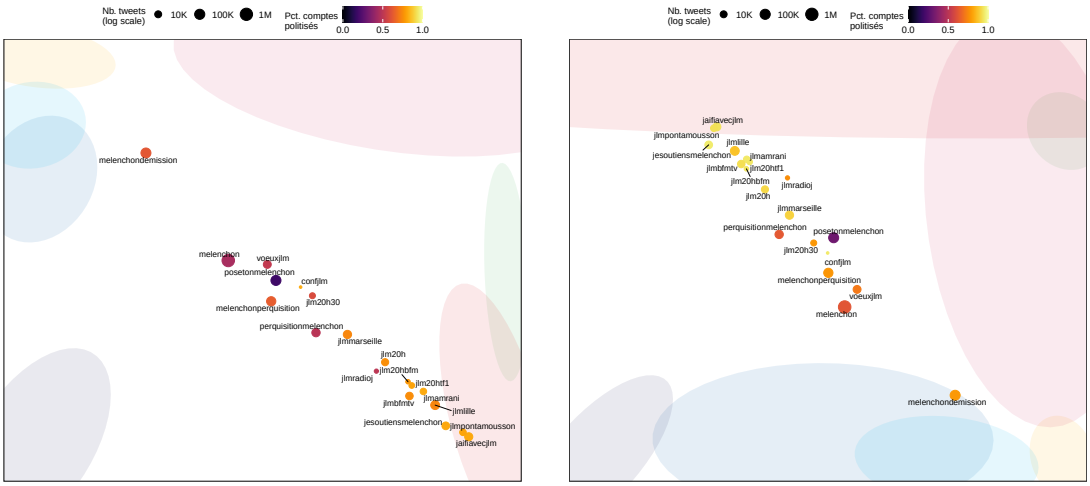

FIGURE 31 – Position moyenne des *hashtags* contenant 'melenchon' ou 'jlm' sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.6 Visualisation des *hashtags* liés à Emmanuel Macron

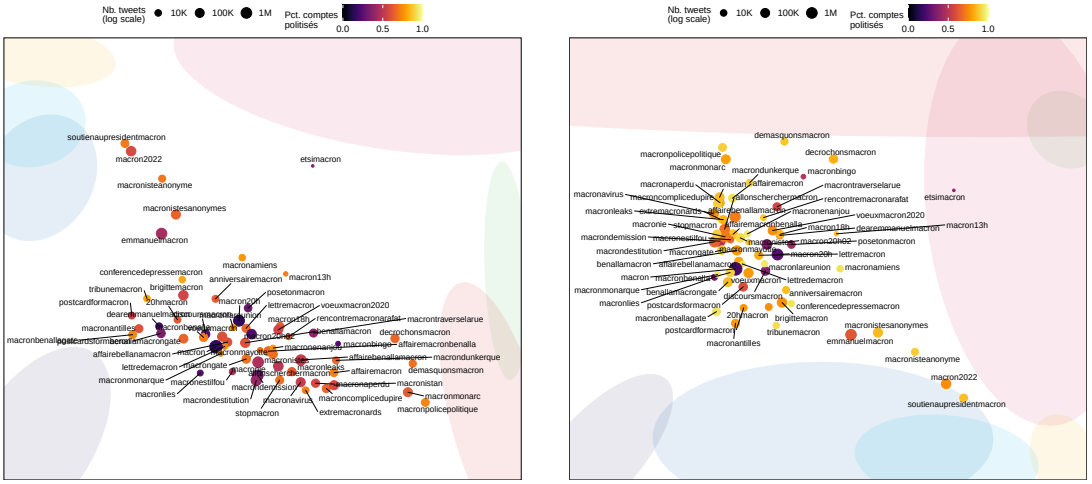

FIGURE 32 – Position moyenne des *hashtags* contenant 'macron' sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.7 Visualisation des *hashtags* liés à #metoo

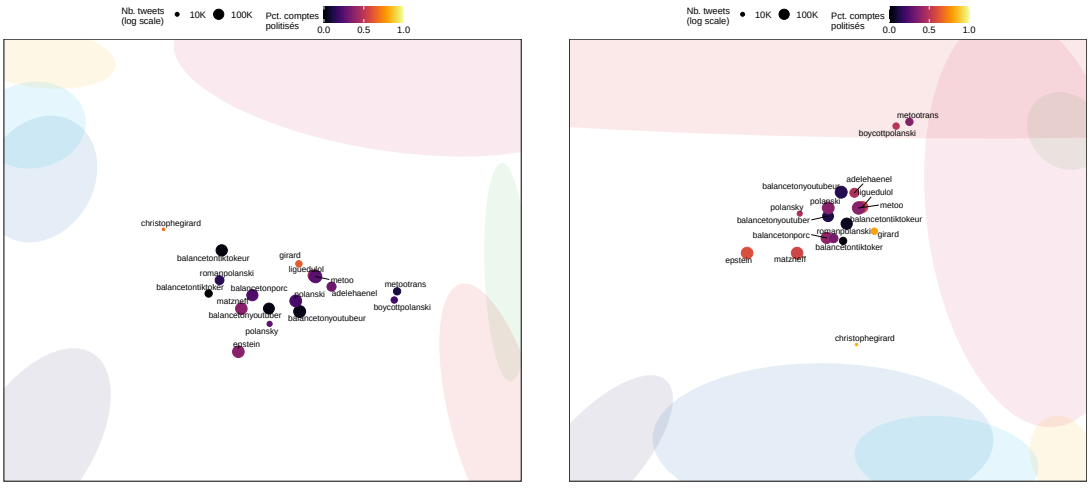

FIGURE 33 – Position moyenne des *hashtags* liés à #metoo et ses suites sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.8 Visualisation des *hashtags* liés au mouvement Black lives matter

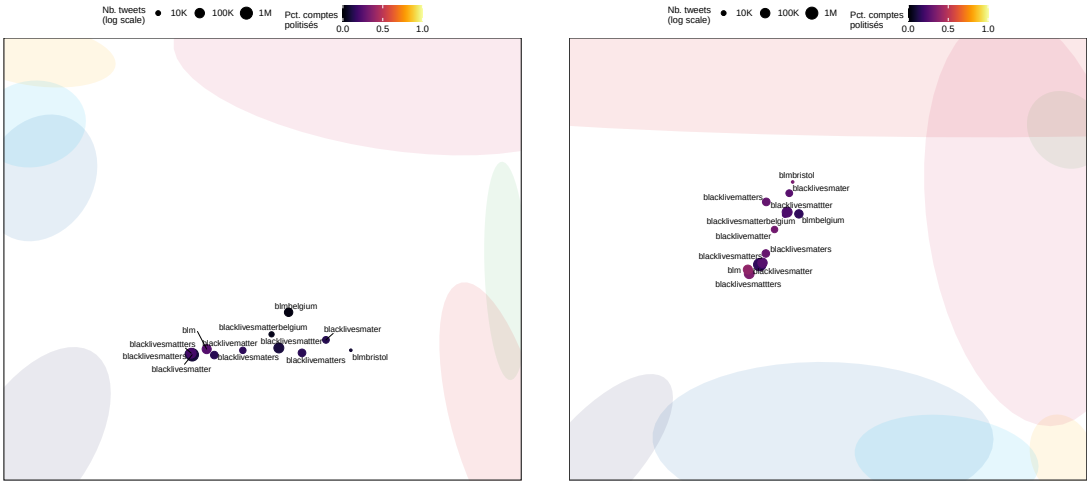

FIGURE 34 – Position moyenne des *hashtags* liés au mouvement blacklivesmatter sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

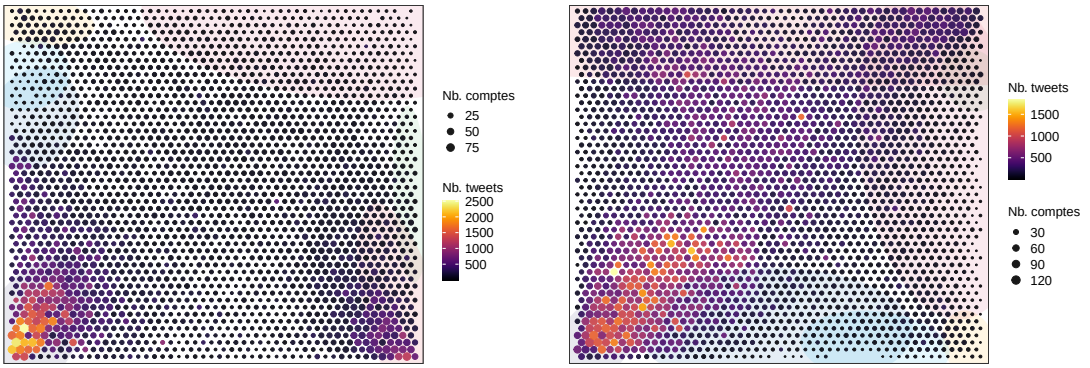

FIGURE 35 – *Hashtag* #blacklivesmatter : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

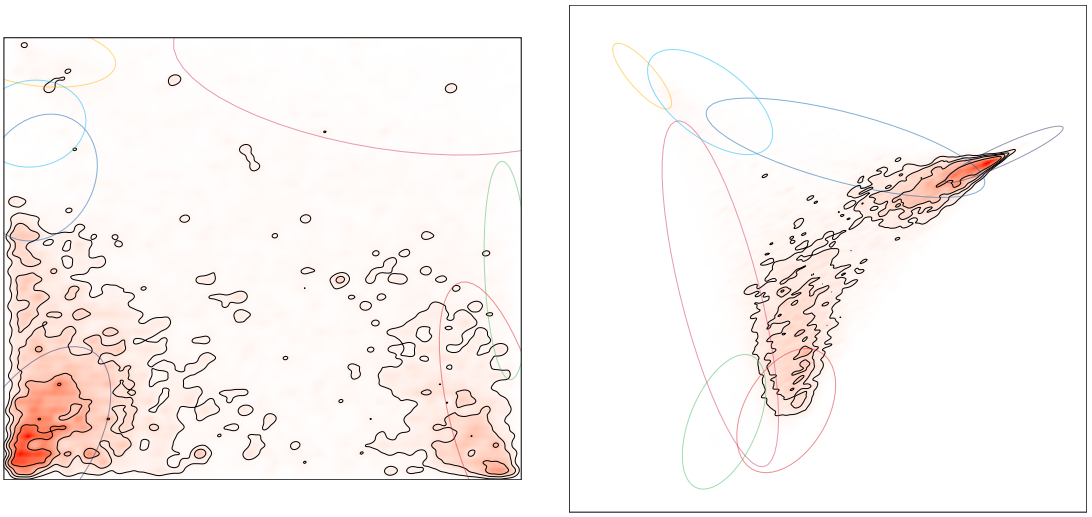

FIGURE 36 – Densité de tweets utilisant le *hashtag* #blacklivesmatter sur la carte de Kohonen de Barberá *follow* (gauche) et sur l'espace Barberá *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

6.9 Visualisation du *hashtag* #climat

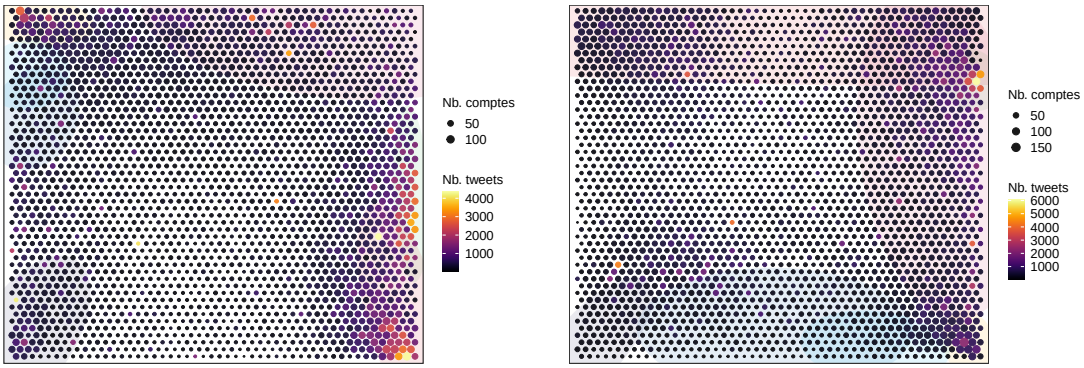

FIGURE 37 – *Hashtag* #climat : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.10 Visualisation du *hashtag* #cnps

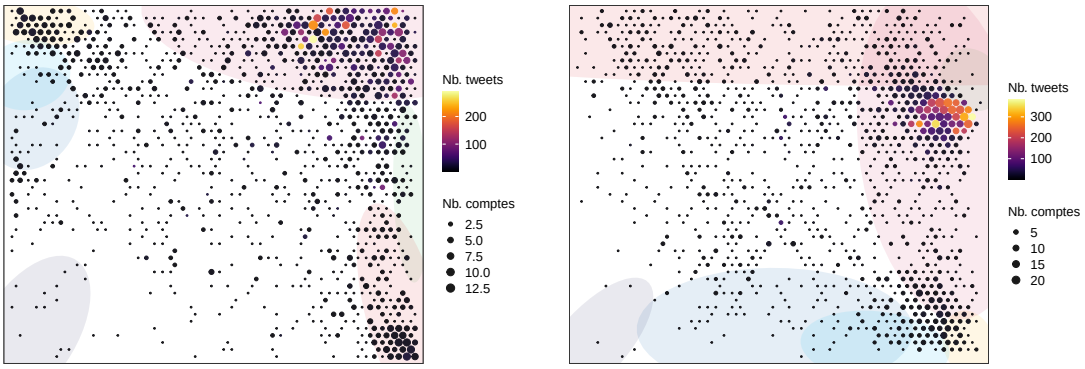

FIGURE 38 – *Hashtag* #cnps : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.11 Visualisation du *hashtag* #notredame

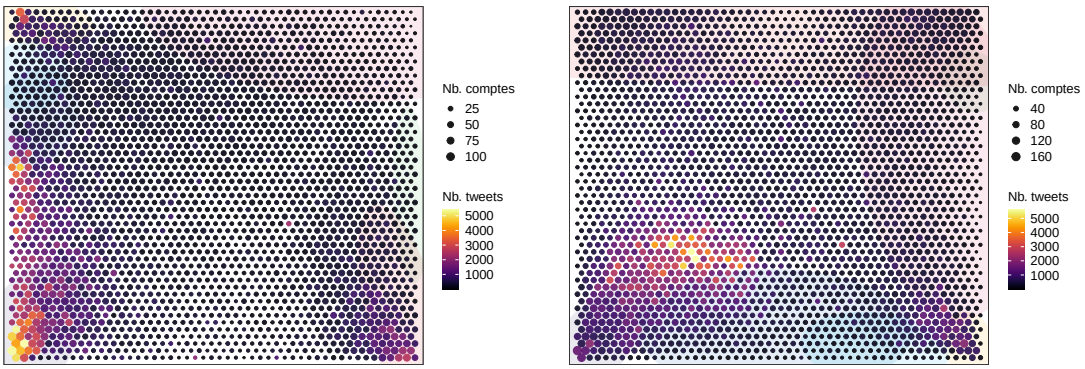

FIGURE 39 – *Hashtag* #notredame : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

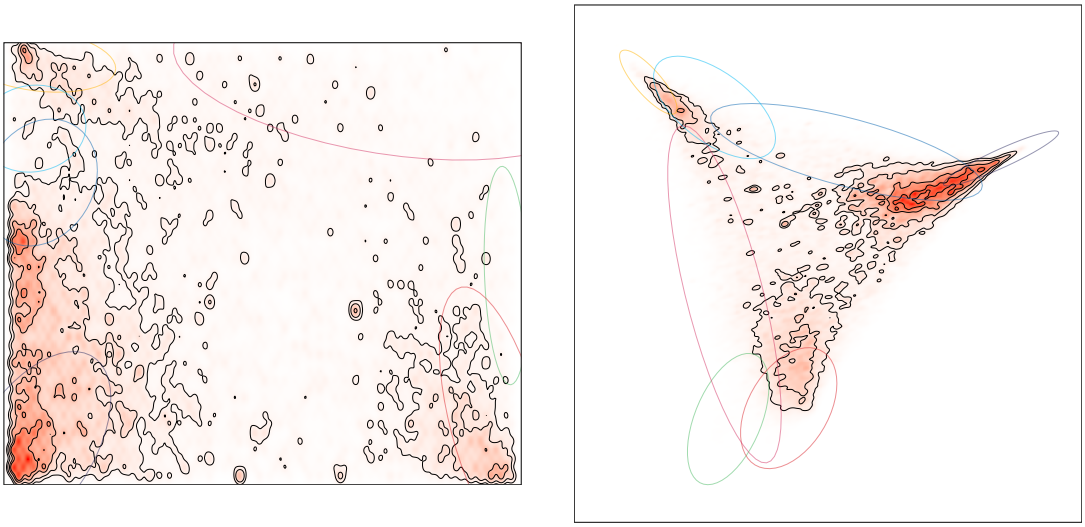

FIGURE 40 – Densité de tweets utilisant le *hashtag* #notredame sur la carte de Kohonen de Barberá *follow* (gauche) et sur l'espace Barberá *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

6.12 Visualisation du *hashtag* #kohlanta

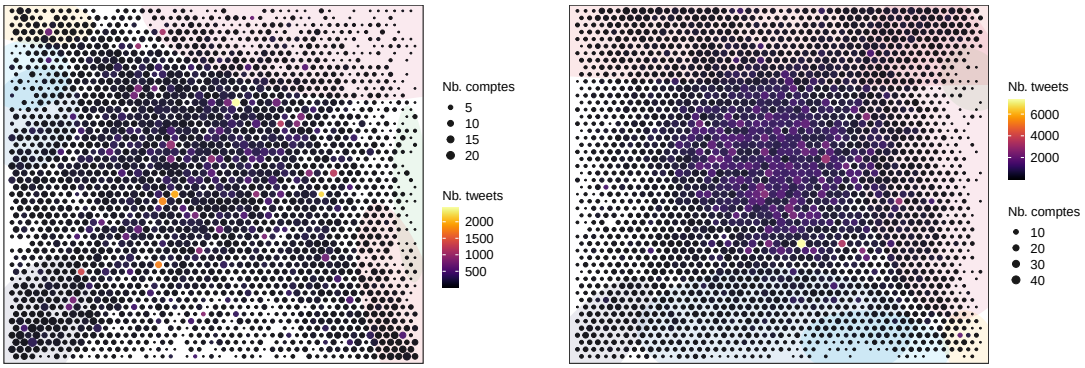

FIGURE 41 – *Hashtag* #kohlanta : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

6.13 Visualisation du *hashtag* #afp

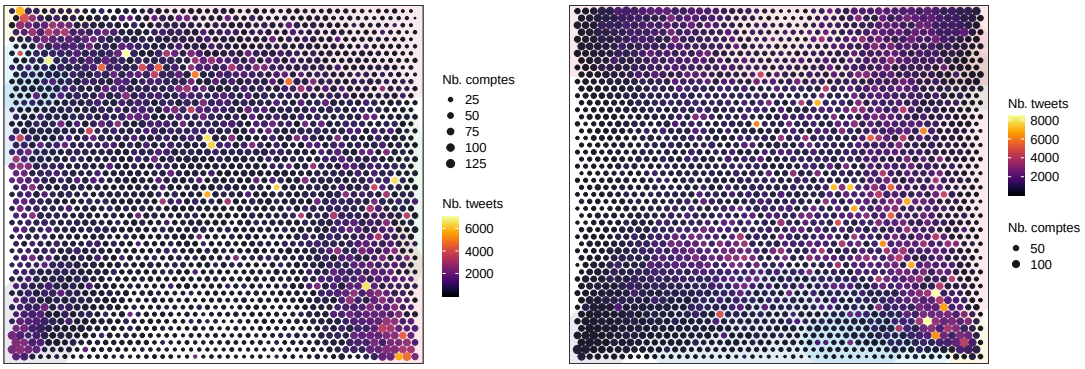

FIGURE 42 – *Hashtag* #afp : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

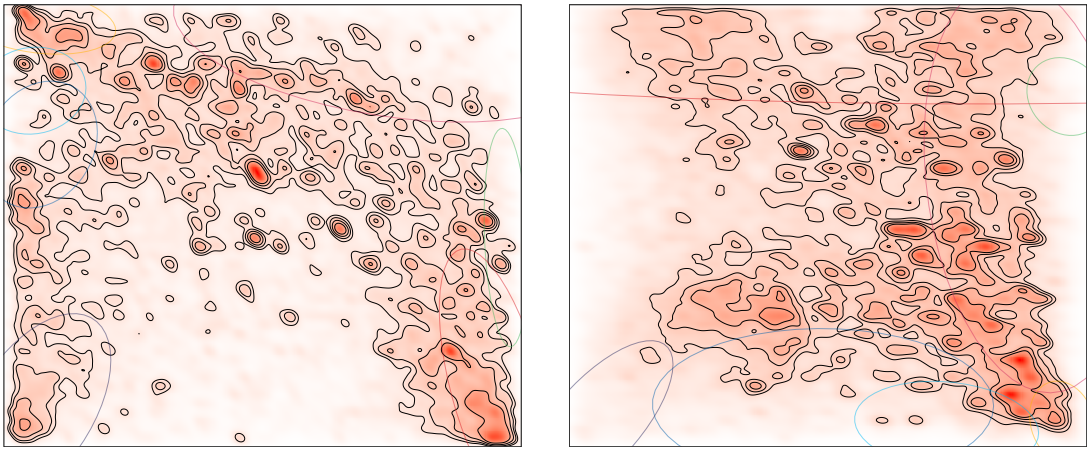

FIGURE 43 – Densité de tweets utilisant le *hashtag* #afp sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

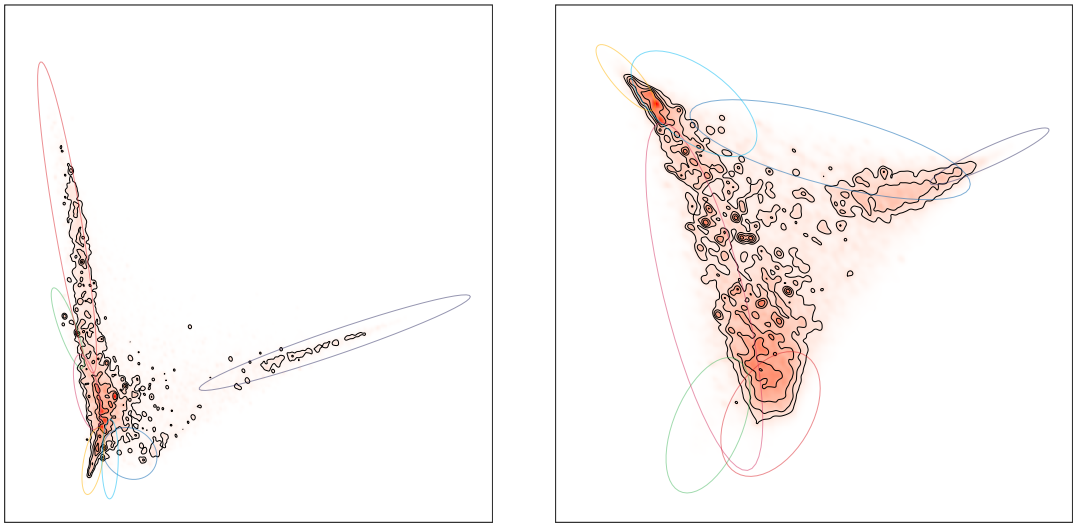

FIGURE 44 – Densité de tweets utilisant le *hashtag* #afp sur les espaces de Barberá *follow* (gauche) et *retweet* (droite). Courbes de niveaux à 10%, 25%, 50% et 75%.

6.14 Visualisation des *hashtags* de concours

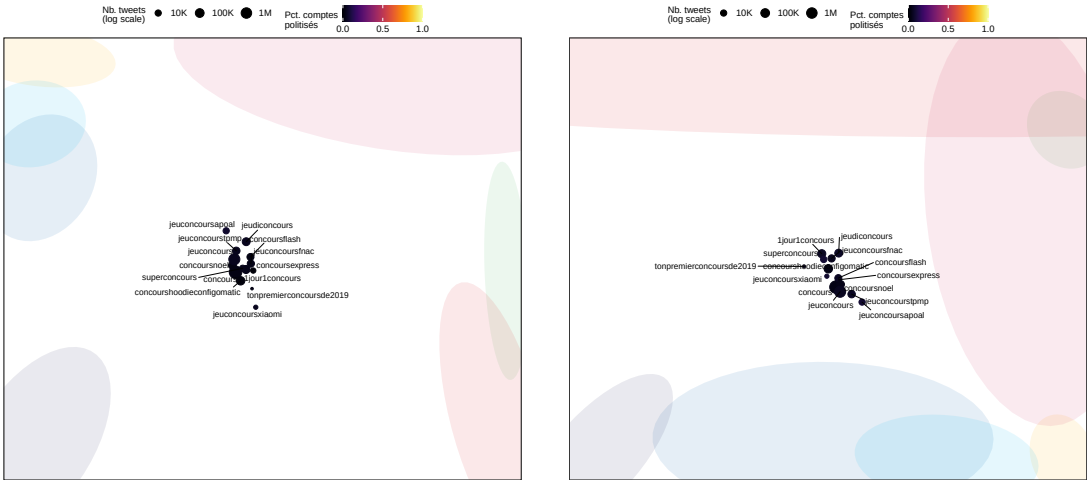

FIGURE 45 – Position moyenne des *hashtags* liés aux concours sur les cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).

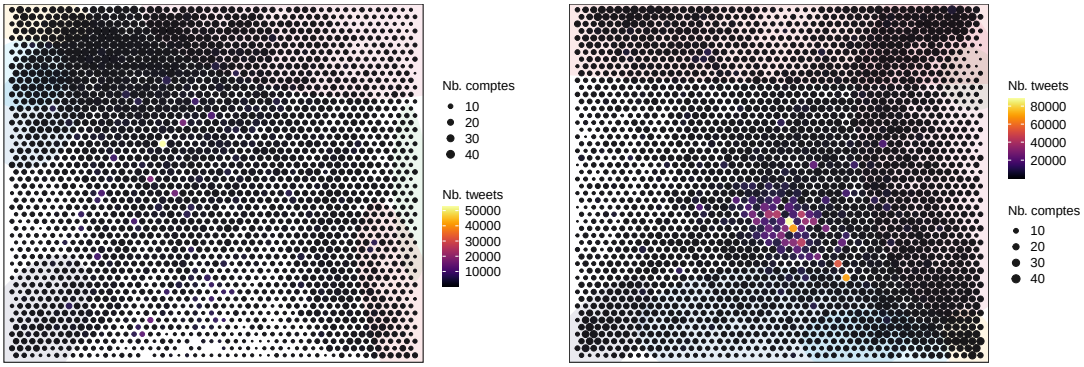

FIGURE 46 – *Hashtag* #concours : nombre de tweets et de comptes pour chaque cellule des cartes de Kohonen de Barberá *follow* (gauche) et *retweet* (droite).