Classification d'aires de dispersion à l'aide d'un facteur géographique

Application à la dialectologie

Clément Chagnaud[1,3], **Philippe Garat**[2], **Paule-Annick Davoine**[1,3], **Guylaine Brun-Trigaud**[4]

1. *Université Grenoble Alpes, CNRS, Grenoble INP, LIG*
   clement.chagnaud@univ-grenoble-alpes.fr
2. *Université Grenoble Alpes, CNRS, Grenoble INP, LJK*
   philippe.garat@univ-grenoble-alpes.fr
3. *Université Grenoble Alpes, CNRS, Grenoble INP, PACTE*
   paule-annick.davoine@univ-grenoble-alpes.fr
4. *Université Côte d'Azur, CNRS, BCL*
   guylaine.brun-trigaud@univ-cotedazur.fr

RÉSUMÉ. *Nous proposons une procédure d'analyse statistique multidimensionnelle couplant des méthodes de projection et de classification pour identifier des ensembles cohérents au sein d'un corpus d'entités géographiques surfaciques que l'on appelle aires de dispersion. La méthodologie intègre un facteur géographique dans la construction de l'espace de représentation pour la projection des données. En appliquant ces méthodes sur des données géolinguistiques, nous pouvons identifier et expliquer de nouvelles structures spatiales au sein d'un corpus d'aires de dispersion de traits linguistiques.*

ABSTRACT. *We propose a multidimensional statistical analysis procedure using projection and classification methods, in order to identify coherent clusters into a set of surficial entities called dispersion areas. The methodology includes a geographical factor to build the representation space for the projection of the data. By applying this method on geolinguistic data, we are able to identify and explain new spatial patterns among a set of dispersion areas of linguistic features.*

MOTS-CLÉS : *géolinguistique, classification, analyse spatiale, statistiques, humanités numériques.*

KEYWORDS: *geolinguistics, geomatics, clustering, spatial analysis, statistics, digital humanities.*

DOI: 10.3166/rig.2020.00107

1. Introduction

La dialectologie s'intéresse à l'étude des traits linguistiques caractéristiques des langues à tradition orale comme les parlers locaux, appelés patois ou encore dialectes. Ces traits linguistiques peuvent être de nature très différente – phonétique, morphosyntaxique, lexicale, sémantique ou prosodique – et évoluent dans un espace géographique donné, dans le temps et au contact de la société. Pour étudier ces parlers locaux, la dialectologie s'est spécialisée dans la constitution de corpus de données descriptives, collectées via une méthodologie d'enquête qui repose sur des questionnaires, sur le choix de réseaux de points linguistiques localisés et d'informateurs (Contini, 2003). Le traitement et l'analyse des données de terrain se fait au moyen de supports cartographiques, sur lesquels sont portées, pour chaque notion linguistique, les localités enquêtées et les formes linguistiques associées en transcription phonétique. À chaque notion ou concept, exprimée par une entrée lexicale sous la forme d'un titre de carte en français, est associée une et une seule carte sur laquelle figurent, en implantations ponctuelles, toutes les formes dialectales transcrites phonétiquement désignant le concept en question (figure 1).

À partir des données des atlas linguistiques, les géolinguistes sont confrontés à la construction de *cartes interprétatives* élaborées à partir de l'étude de la variation dialectale, c'est-à-dire l'ensemble des réalisations phoniques, morphologiques,

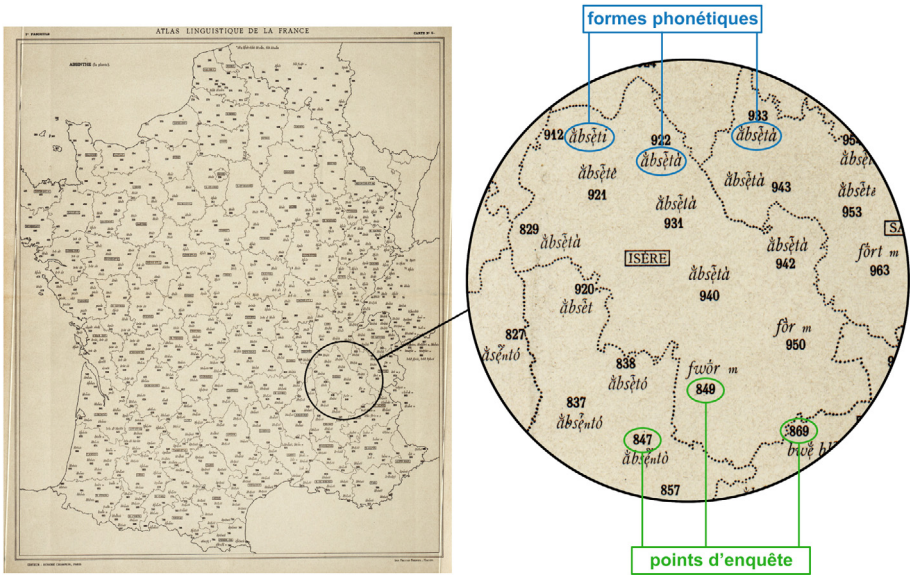


*Figure 1. Exemple de carte extraite de l'*Atlas Linguistique de la France *(Gilliéron et Edmont, 1902-1910) représentant la notion* 'absinthe'. *Cette carte présente les formes phonétiques, correspondant aux réponses collectées à chaque point d'enquête. Les limites administratives (frontières et noms des départements) constituent les seuls éléments de contexte*

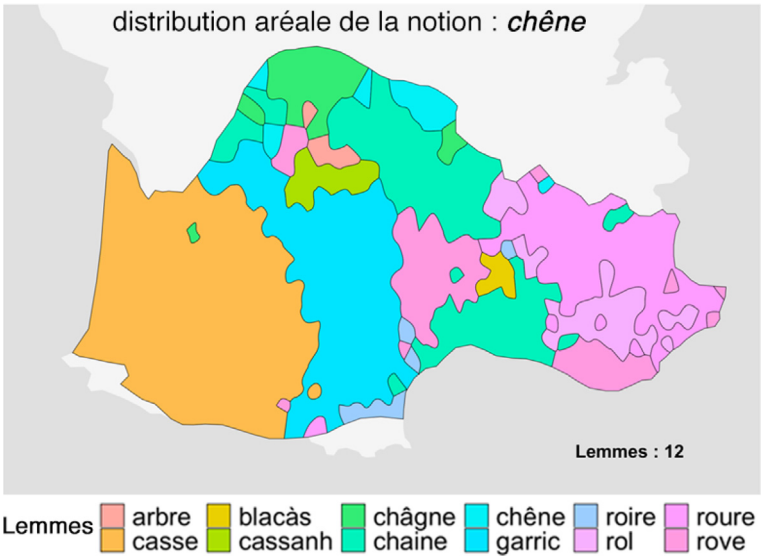


Figure 2. Distribution aréale des lemmes de la notion « chêne » sur le domaine occitan. Données issues du THESOC

syntaxiques et lexicales variant en fonction de la localisation géographique. Ces cartes sont basées sur le concept de lignes isoglosses, c'est-à-dire des « *lignes imaginaires séparant deux zones géographiques se distinguant par un trait linguistique particulier, qui peut être de nature lexicale, sémantique, phonologique, phonétique ou d'un autre type* » (Chambers et Trudgill, 1998). Ce type de cartographie permet de mettre en évidence la *distribution aréale* des variantes dialectales d'une notion linguistique et d'identifier ce qu'on appelle des *aires de dispersion* (délimitées par les lignes isoglosses). Par exemple, la figure 2, illustre la distribution aréale des différentes variantes lexicales (les *lemmes*) de la notion « *chêne* » observables sur le domaine occitan (tiers Sud de la France). Les aires de dispersion des lemmes ont été élaborées à partir de données issues du THESOC[1].

À travers l'analyse de la distribution aréale d'un corpus de plusieurs dizaines de cartes, les géolinguistes cherchent à identifier des motifs spatiaux récurrents et à élaborer des typologies d'aires de dispersion pour certains traits linguistiques (Léonard, 2001 ; Brun-Trigaud *et al.*, 2005). Aujourd'hui, l'élaboration des cartes interprétatives et des typologies s'appuient encore sur des approches empiriques basées sur la connaissance experte et intuitive des dialectologues. Force est de constater que l'outillage logiciel destiné à l'exploitation et à l'analyse spatiale des données géolinguistiques est très en retard : le recours aux systèmes d'information géographique (SIG), largement reconnus pour leur capacité cartographique et les outils d'analyse spatiale qu'ils proposent, est

1. http://thesaurus.unice.fr

encore peu développé dans le domaine de la géolinguistique (Hoch et Hayes, 2010 ; Silber *et al.*, 2012).

La dialectométrie (Séguy, 1973) offre des méthodes de classification permettant de mettre en évidence des structures spatio-linguistiques (Heeringa, 2004) à partir desquelles certaines typologies peuvent se retrouver. Cependant les méthodes mobilisées s'appuient sur les données brutes des points d'enquête et non pas sur les aires de dispersion des traits linguistiques. Par ailleurs, ni l'analyse typologique empirique des géolinguistes, ni l'approche statistique des outils dialectométriques n'étudient les liens entre les caractéristiques linguistiques des structures identifiées et les éléments contextuels (géographiques, géohistoriques, socio-économiques, etc.) susceptibles de constituer des facteurs explicatifs.

Cet article, organisé en six sections, décrit une méthode de classification pour identifier automatiquement des typologies d'aires de dispersion à l'aide d'un facteur géographique ainsi que son usage à travers un environnement d'analyse exploratoire. Faisant suite à la présente introduction, la section 2 développe davantage la problématique de recherche. La section 3 décrit les différentes étapes de la méthode puis la section 4 son implémentation. Enfin, la section 5 propose un cas d'étude appliqué aux données du THESOC puis la section 6 conclut et ouvre une discussion sur les limites de notre proposition.

2. Problématique

Comme cela est défini dans le *Dictionnaire général des sciences humaines* (Thines et Lempereur, 1975), dans le cadre de l'étude de l'occupation d'une espèce en bioécologie : l'aire de dispersion est la « *surface totale, continue ou non, à l'intérieur de laquelle l'espèce est représentée, avec une fréquence variable selon la zone partielle considérée* ». C'est le sens que l'on cherche à transposer dans le cadre de l'étude de l'occupation spatiale des différentes formes d'une notion linguistique. L'aire de dispersion d'un trait linguistique représente la surface à l'intérieur de laquelle ce trait est représenté : elle peut être continue ou non, éclatée ou fragmentée en plusieurs lieux géographiques, regroupée au sein d'une zone géographique particulière ou en périphérie d'un territoire. L'étude des aires de dispersion s'intéresse à la dimension synchronique des phénomènes, c'est-à-dire à l'analyse de leur distribution géographique à un instant donné (Lafkioui, 2015).

Très tôt les dialectologues ont cherché à réaliser des typologies des aires lexicales. Ces analyses typologiques consistent à regrouper des aires de dispersion circonscrites à des espaces particuliers et présentant des similarités quant à leur forme, leur extension et leur localisation spatiales (Brun-Trigaud, 2012). Sur la figure 3 on peut observer que les aires de dispersion en rouge occupent sensiblement le même espace. La forme et l'occupation de cette zone constituent un motif spatial récurrent en ce sens que, à travers un corpus de plusieurs cartes issues d'analyses des variations dialectales, on retrouve de façon plus ou moins récurrente des aires de dispersion occupant cet espace en particulier. Ces analyses permettent de révéler des structures spatiales jusqu'alors

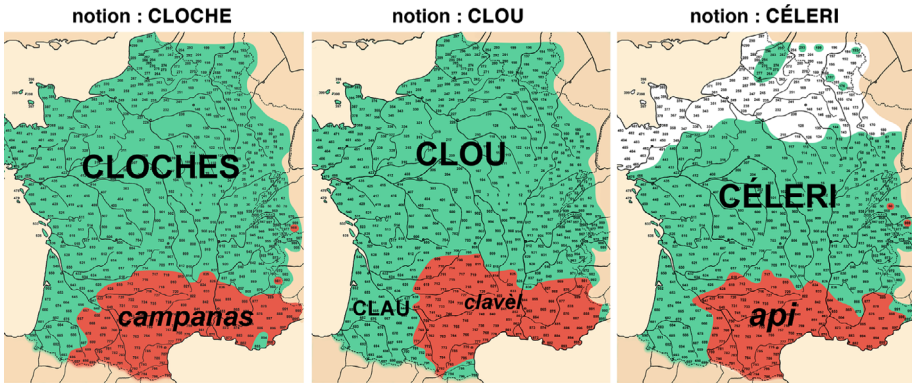


Figure 3. Chacune de ces cartes montrent la distribution aréale des lemmes de 3 notions linguistiques (cloche, clou et céleri). Les aires de dispersion des lemmes campanas, clavel *et* api *(en rouge) présentent des similarités géographiques et constituent un motif spatial récurrent. Cartes interprétées de l'*Atlas Linguistiques de la France *(Brun-Trigaud* et al.*, 2005)*

invisibles modifiant ou ajustant les divisions des aires linguistiques traditionnellement définies (Brun-Trigaud et Malfatto, 2013). Ces études s'accompagnent généralement d'une volonté d'expliquer ces phénomènes par des facteurs socio-historiques ou géographiques (Saussure, 1971 ; Dalbera, 2013) pour comprendre l'émergence de ces motifs spatiaux linguistiques.

Les premières analyses typologiques ont été réalisées de manière empirique en étudiant la distribution aréale de plusieurs cartes d'un domaine d'étude et en se basant sur des connaissances d'experts. Des structures spatio-linguistiques peuvent aussi être observées à travers le résultats des méthodes statistiques de classification offertes par la dialectométrie (Goebl, 2002). À partir de données issues de plusieurs centaines de cartes d'atlas, ces méthodes classifient les points d'enquête (lieux géographiques visités lors des enquêtes de terrain) en fonction de leurs similarités linguistiques (Goebl, 1981). En utilisant la distance de Levenshtein entre deux chaînes de caractères phonétiques (Kessler, 1995), il est possible d'établir une distance linguistique quantitative entre points d'enquête (Nerbonne *et al.*, 1996) qui sont ensuite regroupés avec des méthodes classiques de classification (souvent ascendantes hiérarchiques) (Everitt *et al.*, 2011). Il existe des outils, comme Gabmap[2] (Leinonen *et al.*, 2016), qui utilisent ces approches dialectométriques pour produire des cartes de synthèse représentant des régions à l'intérieur desquelles les points d'enquête sont linguistiquement proches (c'est à dire que les réponses récoltées sont souvent similaires). Ces méthodes permettent une approche géolinguistique au niveau global mais elles ne sont pas adaptées à l'analyse locale des aires de dispersion ni à l'identification de typologies.

2. https://gabmap.nl/

Notre proposition de classification consiste à considérer l'ensemble des aires de dispersion contenues dans un important corpus de cartes interprétatives comme les individus statistiques à analyser, et de les classifier en fonction de leur similarité spatiale, afin d'identifier des classes cohérentes au niveau des co-occurrences spatiales. La difficulté réside dans le fait que les aires de dispersion sont très hétérogènes en taille et en forme : leurs limites ne sont jamais strictement les mêmes.

De plus, pour comprendre les typologies identifiées, notre approche s'appuie sur l'intégration d'un *facteur géographique* pour lequel nous pouvons mesurer l'impact sur ces structures. À ce stade, le facteur géographique dont nous parlons est une partition de l'espace en plusieurs entités géographiques représentant des sous-ensembles territoriaux. Afin de simplifier la lecture nous appellerons ces sous-ensemble territoriaux des *régions*, ce qui, dans notre cas, ne comporte pas de connotation administrative. Nous émettons l'hypothèse que des facteurs géographiques tels que l'organisation des bassins versants ou les limites d'anciennes provinces peuvent avoir une influence sur la diffusion des traits linguistiques dans l'espace au cours du temps. L'intérêt de la méthode est de croiser les données géolinguistiques initiales avec des données spatiales de différente nature, afin d'établir des liens entre des phénomènes linguistiques et diverses réalités géographiques, historiques, sociologiques, etc.

3. Méthode

Nous cherchons donc à classifier un ensemble hétérogène d'aires de dispersion en groupes cohérents et en intégrant comme facteur de regroupement un découpage géographique. Les aires de dispersion se présentent sous la forme d'objets géographiques surfaciques représentés initialement dans l'espace à deux dimensions d'une carte.

On souhaite que cette méthode puisse être capable, sans a priori, de discriminer les aires de dispersion qui, topologiquement, se situeraient soit strictement à l'intérieur d'une ou plusieurs régions du facteur géographique soit recouvrant des frontières entre plusieurs régions. Pour cela nous devons être en mesure de caractériser chaque aire de dispersion selon des critères spatiaux induits par le facteur géographique choisi. Nous transformons les aires de dispersion en points dans un espace de représentation multidimensionnel propre au facteur géographique.

Une approche classique consisterait à effectuer une analyse factorielle croisant les aires de dispersion directement avec les régions du facteur géographique. Cette approche permettrait de placer les aires de dispersion comme des points dans un espace de représentation dirigé par des axes factoriels qui rassemblent un maximum d'inertie. Cependant, dans une perspective d'exploration des données, nous souhaitons rendre la structure de cet espace indépendante du corpus d'aires de dispersion à analyser. En effet, nous souhaitons pouvoir facilement modifier le corpus en ajoutant, supprimant ou fusionnant certaines aires par exemple sans impacter la structure de l'espace de représentation. Nous proposons donc de passer par l'intermédiaire d'un maillage neutre

qui va permettre de discrétiser notre espace et ainsi placer les aires de dispersion dans l'espace de représentation en tant que barycentres des mailles de discrétisation.

Notre corpus d'aires de dispersion devient alors un ensemble de points dans un espace de représentation qui peut maintenant faire l'objet d'une classification. Nous proposons de nous appuyer sur deux méthodes de classification non supervisée : la *Classification Ascendante Hiérarchique* (CAH) et les nuées dynamiques (*k-means*) (Nakache et Confais, 2004).

Les étapes de notre méthode sont les suivantes : 1) création d'un espace de représentation adapté, 2) projection des objets dans cet espace de représentation et 3) leur classification.

3.1. Espaces de représentation

Nous choisissons dans un premier temps de discrétiser l'espace étudié en n unités spatiales de forme hexagonale (maillage plus ou moins fin selon le niveau de discrétisation souhaité) ; cela constitue un espace de représentation initial E (de dimension n) où toute aire de dispersion est discrétisée sous la forme de n coordonnées surfaciques : la i-ème coordonnée est la part de surface occupée dans la maille numéro i. La classification dans un tel espace serait peu efficace car les objets seraient trop dispersés (fléau de la dimension).

L'idée est désormais de construire un espace de représentation F de dimension p plus petite. La dimension p devra être raisonnablement faible afin d'assurer une classification efficace par la suite. Nous utilisons le facteur géographique pour construire cet espace F, en procédant comme suit :

1) calcul de la *matrice des aires* A de dimension $n \times p$ en croisant toutes les unités spatiales avec toutes les régions du facteur géographique : la cellule (i, j) de la matrice A contient l'aire de l'intersection entre la $i^{ème}$ unité spatiale et la $j^{ème}$ région du facteur géographique (figure 4a). Avec cette matrice des aires A nous pouvons obtenir le profil de chaque unité spatiale en fonction du découpage géographique choisi en calculant les pourcentages lignes de A. Inversement nous pouvons exprimer les régions du découpage en fonction des unités spatiales en calculant les pourcentages colonnes de la matrice A.

2) mise en oeuvre d'une *analyse des correspondances* (CA) (Rencher, 2002) sur la matrice A. Cette méthode d'analyse statistique multidimensionnelle va efficacement représenter les unités spatiales comme un ensemble de n points dans l'espace vectoriel F (figure 4b).

3.2. Projection barycentrique

Les unités spatiales (hexagonales) sont positionnées dans l'espace de représentation F propre au facteur géographique choisi ; nous projetons les aires de dispersion dans ce

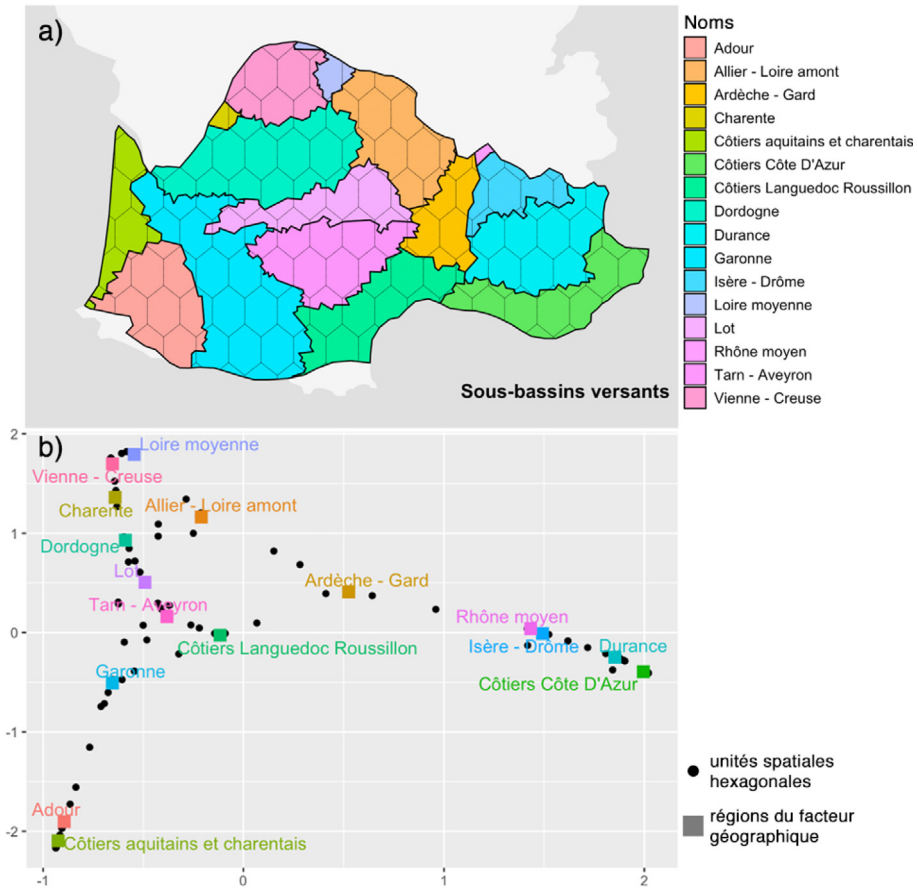


Figure 4. Analyse des correspondances : (a) superposition des régions du facteur géographique et du maillage hexagonal (b) représentation simultanée (semi-barycentrique) des p *régions du facteur géographique et des* n *unités spatiales hexagonales, selon les deux premiers axes factoriels*

même espace. De la même façon que pour la première étape décrite dans la section 3.1 nous calculons une matrice des aires A' entre les aires de dispersion et les unités spatiales. La cellule (i, j) de la matrice A' contient l'aire de l'intersection entre la $i^{ème}$ aire de dispersion et la $j^{ème}$ unité spatiale. On calcule ensuite les pourcentages lignes de cette matrice afin d'obtenir le profil de chaque aire de dispersion en fonction des unités spatiales. Nous utilisons ces profils pour projeter les aires de dispersion dans l'espace de représentation en tant que barycentres des unités spatiales qui les composent. La figure 5 illustre la projection de l'ensemble des aires de dispersion du corpus d'étude, et met en exergue l'exemple de l'aire de dispersion du lemme *châgne* ainsi que les numéros des unités spatiales qui la décrivent.

3.3. Classification

La dernière étape de notre analyse consiste en une procédure de classification non supervisée des aires de dispersion qui sont désormais représentées par des points (en bleu dans la figure 5b) dans l'espace de représentation *F*. Plusieurs possibilités existent compte tenu du grand nombre de méthodes de classification disponibles dans la littérature (Everitt *et al.*, 2011). À ce stade nous procédons de la manière suivante : tout d'abord nous utilisons l'algorithme de Classification Ascendante Hiérarchique (CAH) avec différents critères d'agrégation tels que celui de Ward, du saut minimal, du saut maximal ou de la distance moyenne. Le dendrogramme produit par la CAH peut être coupé (élagué) à n'importe quel niveau en fonction du nombre de classes souhaité. Nous proposons ensuite la possibilité de consolider les classes avec l'algorithme *k-means* dont les points de départ sont initialisés avec les centroïdes des classes déjà établies précédemment par la CAH (Nakache et Confais, 2004). Afin de mesurer le pouvoir classificatoire du facteur géographique, nous proposons deux indicateurs (R^2_E et R^2_F) qui représentent la part de l'inertie inter-classes sur l'inertie totale de l'ensemble des données, lorsque celles-ci sont représentées respectivement dans l'espace *E* et dans l'espace *F*. Plus ces indicateurs sont proches de 1 plus les classes sont compactes ce qui signifie que la classification est efficace. Il est ainsi possible pour les praticiens de comparer différents facteurs géographiques et d'évaluer lesquels sont les plus pertinents pour leur jeu de données.

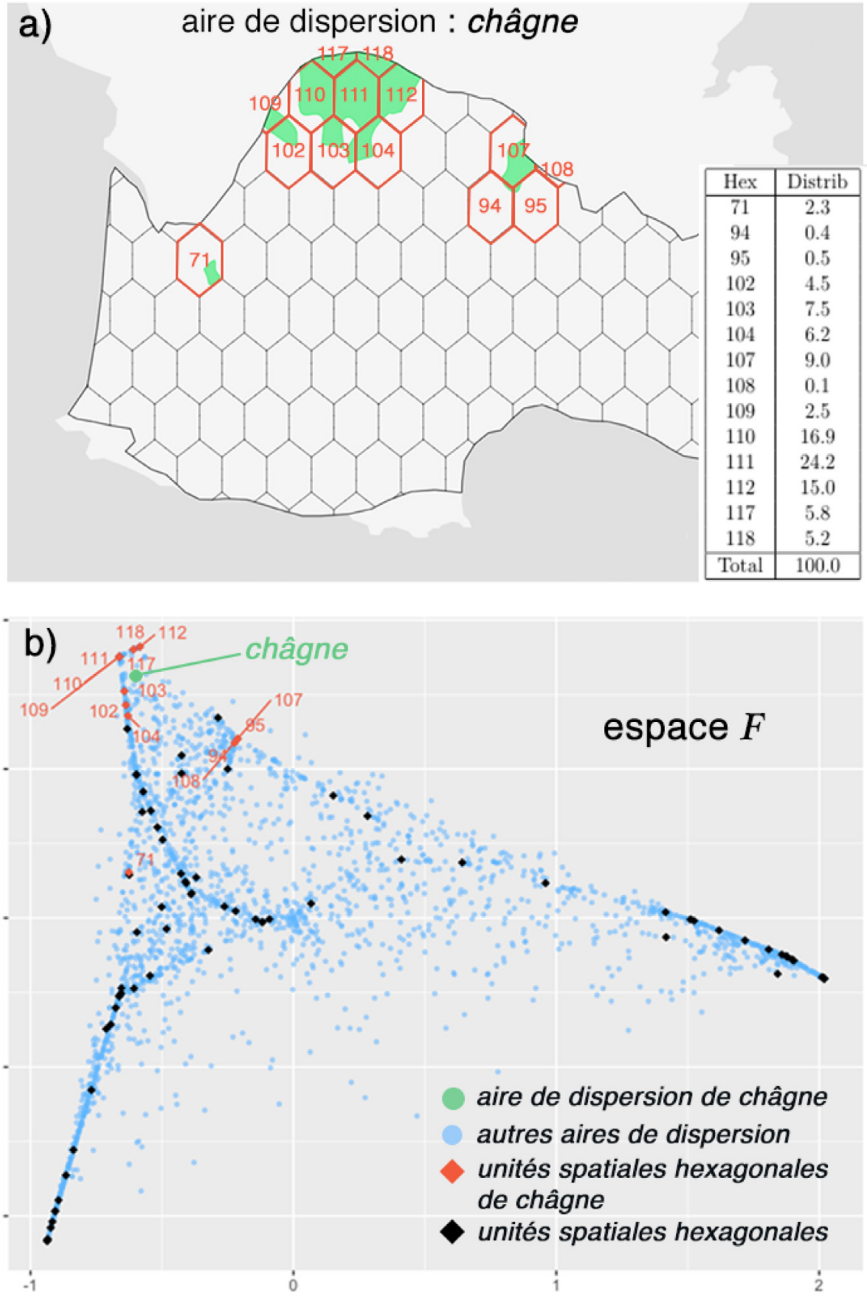
4. Implémentation de la méthode

La méthode décrite ci-dessus a été implémentée dans un environnement d'analyse exploratoire de données géographiques développé avec le langage R[3]. Il offre à l'utilisateur la possibilité d'explorer le processus de classification en agissant de façon interactive sur les paramètres de classification et de visualiser graphiquement et cartographiquement les résultats. La visualisation et la classification étant liées dynamiquement, l'utilisateur peut explorer chaque classe pour faciliter l'interprétation des regroupements.

4.1. Exploration visuelle

La représentation cartographique a pour objectif de mettre en évidence le profil de concentration de chaque classe constituée. Pour chaque classe, nous identifions les zones recouvertes localement par un nombre *n* d'aires de dispersion de la classe. Un indice (ou score de concentration) n/N_{max}, est alors créé, N_{max} étant le plus grand nombre d'aires superposées identifié pour cette classe. Plus la valeur de n/N_{max} est proche de 1 plus la concentration en aires sur la zone est forte. Ainsi on visualise l'épicentre de la classe qui peut être interprété comme l'origine géographique de diffusion d'un phénomène (figure 6).

3. https://cran.r-project.org



Hex	Distrib
71	2.3
94	0.4
95	0.5
102	4.5
103	7.5
104	6.2
107	9.0
108	0.1
109	2.5
110	16.9
111	24.2
112	15.0
117	5.8
118	5.2
Total	100.0

Figure 5. Projection barycentrique des aires de dispersion. (a) le lemme « châgne » (en vert) occupe seulement quelques unités spatiales (en rouge) dans l'espace géographique. (b) l'espace de représentation F est ici visualisé selon ses deux premiers axes factoriels. Chaque aire de dispersion

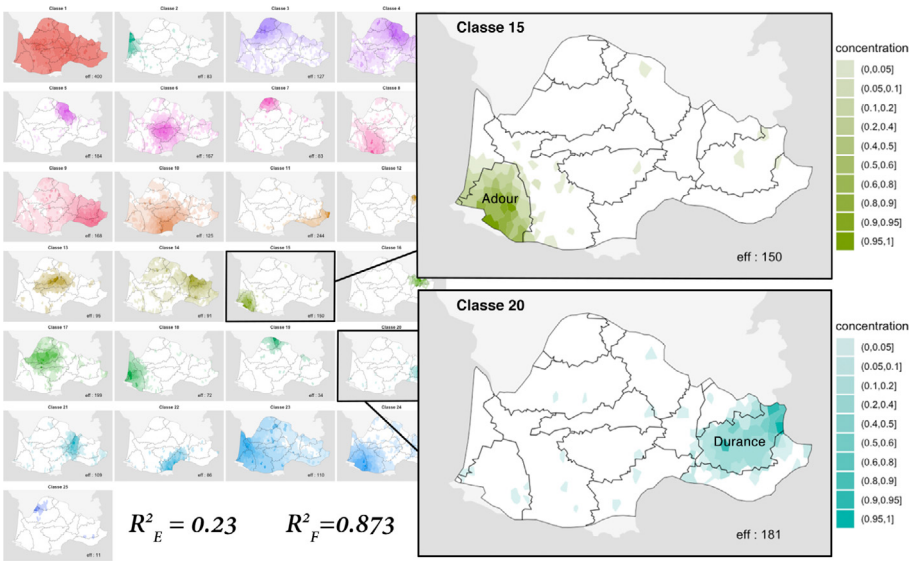


Figure 6. Classification en 25 classes à l'aide du facteur géographique des sous-bassins versants. Agrégation avec le critère Ward et consolidée avec k-means

L'environnement offre également la possibilité d'explorer chaque classe, notamment à travers une analyse des caractéristiques géographiques (contour, extension, localisation) et thématiques des aires de dispersion qui la composent. L'analyse géographique vise à dégager le meilleur représentant de la classe (parangon) en relation avec le profil de concentration. C'est-à-dire l'aire de dispersion qui occupe le plus d'espace au sein de la classe (un exemple de parangon est donné en figure 7). Les autres individus de la classe sont triés par ordre décroissant de représentativité.

4.2. Caractérisation des classes

Afin de mieux caractériser chaque classe, nous procédons à une analyse thématique à l'aide de variables illustratives : la distribution des thèmes et sous-thème de chaque classe est comparée à celle de l'ensemble du corpus. Sur le plan graphique, les diagrammes *radar* mettent en évidence les écarts de distribution comme l'illustrent les figures 8 et 9 et dont l'interprétation est présentée en section 5.

5. Cas d'étude

Nous avons appliqué notre méthode sur un corpus de 235 cartes géolinguistiques issues du *Thesaurus Occitan* ou *THESOC* (Dalbera *et al.*, 2012) qui rassemble les atlas linguistiques régionaux du domaine occitan de la France. Pour rappel, chaque carte décrit la distribution spatiale des lemmes utilisés pour désigner une notion linguistique

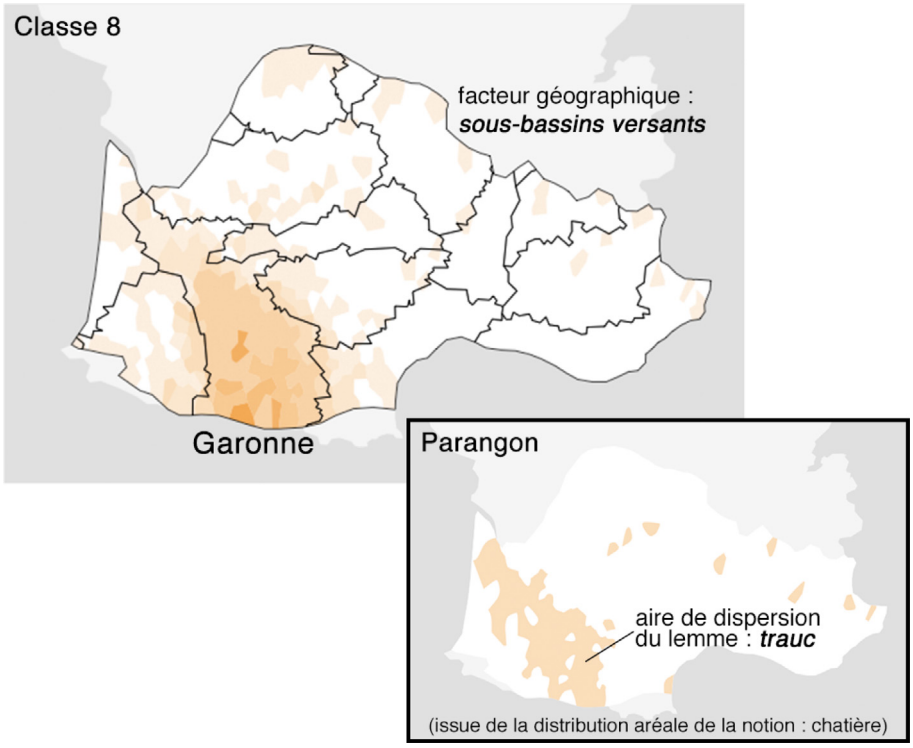


Figure 7. L'aire de dispersion du lemme « trauc *» issue de la distribution aréale de la notion «* chatière *» est le parangon de la classe 8 (Garonne)*

collectée sur 645 points d'enquête répartis dans la région étudiée. Pour chacune des notions linguistiques, les aires de dispersion propres à chaque lemme ont été définies au moyen de méthodes d'interpolation appliquées à des données qualitatives (Chagnaud *et al.*, 2017). Un corpus de 3 437 aires lexicales est ainsi créé, à partir duquel des co-occurrences spatiales et les facteurs géographiques associés doivent être identifiés. Nous proposons d'appliquer différents types de facteurs géographiques à notre étude de cas : limites environnementales (les sous-bassins versants (figure 4 a) ou les hydro-écorégions[4]), administratives (les départements) ou historiques (les Généralités en 1 789 ou les provinces gauloises en 450).

En utilisant le facteur géographique des sous-bassins versants, l'analyse montre qu'il est préférable de travailler avec au moins 25 classes pour voir se dégager des regroupements intéressants. La figure 6 montre que certains sous-bassins versants permettent de regrouper un grand nombre de lemmes qui sont bien confinés à l'intérieur de leurs limites, c'est le cas par exemple du bassin *Côtiers Côte d'Azur* avec 244

4. https://www.irstea.fr/fr/les-hydroecoregions-de-france-metropolitaine

Classe 15

Thèmes	Effectif classe	Effectif total	Pourcentage classe	Pourcentage total	Écart
1-CULTURES	35	527	23.33	15.33	8.00
2-ELEVAGE	25	670	16.67	19.49	-2.83
3-HABITAT, VIE QUOTIDIENNE	41	956	27.33	27.81	-0.48
4-HOMME	17	344	11.33	10.01	1.32
5-NATURE	32	940	21.33	27.35	-6.02

Figure 8. Analyse thématique de la classe 15 (Adour) en cinq thèmes : tableau des fréquences et diagramme radar. Sur le diagramme, le trait en pointillé représente le profil moyen, le trait plein représente le profil de la classe

lemmes, de celui de la *Durance* avec 181 lemmes ou encore de l'*Adour* avec 150 lemmes. Si les cas d'une cohérence parfaite entre une région du facteur géographique et une aire lexicale sont assez rares, on peut néanmoins trouver une bonne adéquation comme par exemple dans le cas de l'aire de *dragon* (lemme de la notion « *faux à blé* ») avec le bassin de l'Adour ou celle de l'aire du lemme *estraçaire* (de la notion « *chiffonnier* ») sur la Côte d'Azur. Ces phénomènes de répartition trouvent peut-être leurs origines dans les mouvements migratoires de ces métiers saisonniers. Ainsi dans le cas de la classe décrivant le bassin de l'Adour, le diagramme radar de la figure 8 nous montre une prédominance d'aires appartenant au vocabulaire de la culture (auquel appartient le terme *dragon*). Quant au diagramme radar de la figure 9 qui rend compte des lemmes regroupés dans le bassin de la Durance, il indique par la présence des sous-thèmes que les notions rattachées à la flore sont très présentes dans ce groupe. Ces analyses permettent de trouver s'il existe, par exemple, une interprétation par les thèmes qui pourrait unir les aires lexicales regroupées par la classification.

En changeant de facteur géographique, de nouveaux regroupements apparaissent. Chaque critère géographique a sa pertinence et génère des regroupements différents, montrant bien que les parlers ne sont pas seulement influencés par le fait d'appartenir à une aire linguistique, mais que d'autres facteurs entrent en jeu et donnent leur propre cohérence. Ils conduisent néanmoins à des classifications plus ou moins efficaces (voir tableau 1). Cette table nous montre que le découpage administratif des provinces de Gaule en 450 semble être le facteur géographique qui donne les classes les moins compactes alors que le découpage des départements donne une classification plus efficace.

6. Discussion et conclusion

En proposant une méthode automatique de classification d'objets géographiques surfaciques représentant des aires de dispersion de phénomènes linguistiques, il nous a été possible d'analyser et identifier les co-occurrences spatiales au sein d'un corpus de plusieurs milliers d'entrées lexicales. La méthode est implémentée selon une approche d'analyse exploratoire de données, afin d'offrir aux linguistes l'opportunité d'identifier plus facilement les liens existants entre phénomènes linguistiques et facteurs

Classe 20

Sous-thèmes	Effectif classe	Effectif total	Pourcentage classe	Pourcentage total	Écart
1-Céréales	10	217	5.52	6.31	-0.79
1-Jardin	14	203	7.73	5.91	1.83
1-Labours	4	107	2.21	3.11	-0.90
2-Basse-cour	3	104	1.66	3.03	-1.37
2-Bétail	22	399	12.15	11.61	0.55
2-Bêtes de somme	1	37	0.55	1.08	-0.52
2-Chiens	1	52	0.55	1.51	-0.96
2-Ruches	4	78	2.21	2.27	-0.06
3-Linge	13	148	7.18	4.31	2.88
3-Maison	20	317	11.05	9.22	1.83
3-Mobilier	29	491	16.02	14.29	1.74
4-Corps humain	11	171	6.08	4.98	1.10
4-Nourriture	5	173	2.76	5.03	-2.27
5-Animaux sauvages	13	356	7.18	10.36	-3.18
5-Flore	21	252	11.60	7.33	4.27
5-Météo	3	213	1.66	6.20	-4.54
5-Relief	7	119	3.87	3.46	0.41

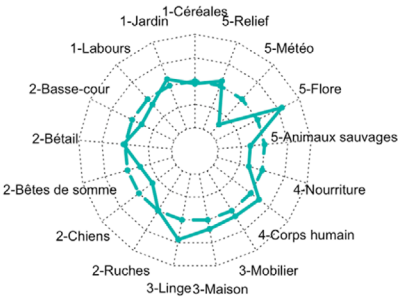


Figure 9. Analyse thématique de la classe 20 (Durance) en dix-sept sous-thèmes : tableau des fréquences et diagramme radar. Sur le diagramme, le trait en pointillé représente le profil moyen, le trait plein représente le profil de la classe

géographiques (ceux-ci étant jusqu'à maintenant essentiellement basés sur leur connaissance du territoire). La force de ces liens est quantifiée grâce aux indicateurs de pouvoir classificatoire (R^2_E et R^2_F). Cependant de futurs travaux devront ajuster ces indicateurs pour tenir compte du nombre de régions des facteurs géographiques car il influence les valeurs de ces indicateurs. Parmi les éléments de la procédure qui peuvent faire l'objet d'un paramétrage de la part de l'utilisateur, on retrouve le maillage par les unités spatiales. En effet, le maillage qui discrétise notre espace pourrait être soit plus dense soit plus large et de formes variables : carrés, triangles, hexagones, etc. Rendre le maillage plus dense entraînerait l'ajout de plus d'éléments structurels dans l'espace F tandis que les aires de dispersion seraient décrites topologiquement avec une meilleure résolution. Cependant, la position barycentrique des aires de dispersion dans l'espace F resterait sensiblement la même. Il faut donc trouver le bon compromis entre un temps de calcul trop élevé induit par un maillage dense et une précision insuffisante induit par un maillage trop large.

Notre environnement d'analyse est interactif, en ce sens qu'il permet d'étudier facilement plusieurs classifications en faisant varier les critères d'agrégation, il offre également la possibilité de caractériser les classes ainsi établies en proposant aux linguistes de tester des clés d'analyse de leur choix (analyse thématique). Il faut garder à l'esprit que notre méthode, au travers d'un facteur géographique, engendre des classes spatialement compactes, mais que ces classes ne sont pas toutes pertinentes d'un point de vue linguistique. Déterminer la validité des classes reste du ressort de l'expert.

Tableau 1. Comparaison du pouvoir classificatoire des différents facteurs géographiques à l'aide des indicateurs R^2_E et R^2_F. Méthode d'agrégation : Ward. Nombre de classes : 25. K-means : oui

Facteur géographique	R^2_E	R^2_F
Aucun	NA	0,374
Généralités en 1789	0,233	0,900
Provinces Gaule en 450	0,191	0,950
Sous-bassins versants	0,230	0,873
Départements	0,308	0,722
Hydro-écorégions	0,222	0,897

La validation complète de la méthode nécessiterait de l'appliquer à une diversité de jeux de données géolinguistiques, mais aussi dans le cadre d'autres thématiques mobilisant des données similaires. Avec le développement des corpus en ligne et les chantiers actuels de numérisation des atlas linguistiques français et étrangers, nous obtiendrons certainement dans un futur proche la matière à de nouvelles applications en dialectologie et géographie linguistique.

Remerciements

Nous remercions Elisabetta Carpitelli et Carole Chauvin, du laboratoire GIPSA-Lab de Grenoble (équipe Voix Systèmes Linguistiques et Dialectologie*, département* Paroles et Cognition*) et impliquées dans le projet ANR ECLATS soutenu par l'Agence nationale de recherche ANR-15-CE-380002, pour leur aide à la validation des propositions cartographiques.*

Bibliographie

Brun-Trigaud G. (2012). Essai de typologie des aires lexicales dans l'Atlas Linguistique du Centre. *Annales de Normandie*, vol. 62, n° 2, p. 77-93.

Brun-Trigaud G., Le Dû J., Le Berre Y. (2005). *Lectures de l'Atlas Linguistique de la France de J. Gilléron et E. Edmont : du temps dans l'espace*. CTHS.

Brun-Trigaud G., Malfatto A. (2013). Limites dialectales vs limites lexicales dans le domaine occitan : un impossible accord ?In *Current Approaches to Limits and Areas in Dialectology*, p. 293-310. Cambridge Scholars Publishing.

Chagnaud C., Garat P., Davoine P.-A., Carpitelli E., Vincent A. (2017). Shinydialect: A cartographic tool for spatial interpolation of geolinguistic data.In *Proceedings of the 1st ACM SIGSPATIAL workshop on Geospatial Humanities*, p. 23-30. Association for Computing Machinery.

Chambers J. K., Trudgill P. (1998). *Dialectology* (2e éd.). Cambridge Univ. Press.

Contini M. (2003). Quel avenir pour la dialectologie ? *I encontro de Estudios Dialectologicos*, p. 17-46.

Dalbera J.-P. (2013). La trajectoire de la dialectologie au sein des sciences du langage. De la reconstruction des systèmes dialectaux à la sémantique lexicale et à l'étymologie. *Corpus*, vol. Dialectologie : corpus, atlas, analyses, n° 12, p. 173-200.

Dalbera J.-P., Ranucci J.-C., Oliviéri M., Brun-Trigaud G. (2012). La base de données linguistique occitane Thesoc. Trésor patrimonial et instrument de recherche scientifique. *Estudis Romànis*, vol. 34, p. 367-387.

Everitt B. S., Landau S., Leese M., Stahl D. (2011). *Cluster Analysis* (5th éd.). John Wiley, UK.

Gilliéron J., Edmont E. (1902-1910). *Atlas Linguistique de la France*. Honoré Champion, Paris.

Goebl H. (1981). Eléments d'analyse dialectométrique (avec application à l'ais). *Revue de Linguistique Romane*, vol. 45, p. 349-420.

Goebl H. (2002). Analyse dialectométrique des structures de profondeur de l'alf. *Revue de Linguistique Romane*, vol. 66, p. 5-63.

Heeringa W. J. (2004). *Measuring dialect pronunciation differences using levenshtein distance*. Thèse de doctorat non publiée, University of Groningen.

Hoch S., Hayes J. (2010). Geolinguistics: The incorporation of geographic information systems and science. *The Geographical Bulletin*, vol. 51, n° 1, p. 23-36.

Kessler B. (1995). Computational dialectology in irish gaelic. *Proceedings of the European ACL*, p. 60-67.

Lafkioui M.B. (2015). Méthodologie de recherche en géolinguistique. *Corpus*, vol. 14, p. 139-154.

Leinonen T., Çoltekin Ç., Nerbonne J. (2016). Using Gabmap. *Lingua*, vol. 178, p. 71-83.

Léonard J.-L. (2001). Aréologie dialectale et modularité des réseaux dialectaux : étagement spatial et structural des processus (morpho)phonologiques dans le réseau dialectal basque.In *Actes du XVe Congrès international de l'Académie basque, 17-19 septembre 2001*, p. 141-168.

Nakache J., Confais J. (2004). *Approche pragmatique de la classification : arbres hiérarchiques, partitionnements*. Editions Technip, Paris.

Nerbonne J., Heeringa W., Hout E. van den, Kooi P. van de, Otten S., Vis W. van de. (1996). Phonetic distance between dutch dialects.*CLIN VI: Proceedings of the Sixth CLIN Meeting*, p. 185-202.

Rencher A.C. (2002). *Methods of Multivariate Analysis* (2nd éd.). Wiley-Interscience, Hoboken, NJ, USA.

Saussure F. de. (1971). *Cours de linguistique générale*. Payot, Paris.

Silber P., Weibel R., Glaser E., Bart G. (2012). Cartographic visualization in support of dialectology.In *Proceedings of the 2012 AutoCarto International Symposium on Automated Cartography, 2012 sept. 16-18, Colombus, Ohio, USA*.

Séguy J. (1973). Dialectométrie dans l'atlas linguistique de la gascogne. *Revue de linguistique romane*, vol. 37, p. 1-24.

Thines G., Lempereur A. (1975). *Dictionnaire général des sciences humaines*. Editions Universitaires, Paris.