Classification d'aires de dispersion à l'aide d'un facteur géographique : application à la dialectologie

Clément Chagnaud, Philippe Garat, Paule-Annick Davoine, Guylaine Brun-Trigaud

Classification d'aires de dispersion à l'aide d'un facteur géographique : application à la dialectologie.

Clément Chagnaud[1], **Philippe Garat**[2], **Paule-Annick Davoine**[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 procedure of a multidimensional statistical analysis using projection and classificatin methods in order to identify coherent clusters into a set of surface entities called dispersion areas. The methodology includes a geographical factor to build the representation space for the projection of the data. By applying this methods 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, géomatique, classification, analyse spatiale, statistiques, humanités numériques*

KEYWORDS: *geolingusitic, géomatic, clustering, spatial analysis, statistics, digital humanities*

1. Contexte et positionnement

La dialectologie s'intéresse à l'étude des traits linguistiques caractéristiques des langues à tradition orale comme les parlers locaux, appelés patois ou 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 (Chambers, Trudgill, 1998).

La géolinguistique s'intéresse à l'étude des phénomènes linguistiques à travers l'analyse de la distribution spatiale des traits linguistiques et à leur variabilité. Pour cela, les géolinguistes cherchent à identifier les frontières délimitant les aires occupées par les différents traits linguistiques. Ces frontières, appelées *isoglosses*, sont définies à partir de données recueillies lors de vastes campagnes d'enquêtes de terrain décrivant, en un lieu donné, les variations phonétiques ou lexicale d'une *notion* linguistique. Ces données brutes ont servi, notamment à élaborer les atlas linguistiques, dont le plus emblématique est l'*Atlas Linguistique de la France* (Gilliéron, Edmont, 1902-1910). Par exemple, la figure 1, illustre les différentes variations lexicales (*lemmes*) de la notion *"chêne"* issue de la base de données du THESOC [1].

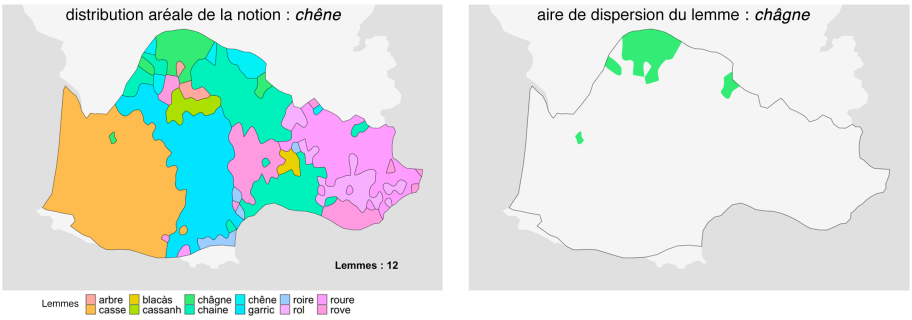


FIGURE *1. Distribution aréale des lemmes de la notion "chêne" sur le domaine occitan et l'aire de dispersion du lemme châgne. Données issues du THESOC.*

Les isoglosses délimitent des *aires de dispersion* qui matérialisent l'occupation d'un trait linguistique dans l'espace. L'étude de ces 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). A travers l'analyse empirique d'un corpus de cartes, les géolinguistes peuvent identifier des motifs spatiaux récurrents (telle que l'occupation d'une zone géographique singulière) dans les aires de dispersion de certains traits linguistiques

1. http://thesaurus.unice.fr

(Léonard, 2001) (Brun-Trigaud *et al.*, 2005). Ils réalisent ainsi des analyses typologiques en regroupant des aires de dispersion qui sont circonscrites à des espaces particuliers et qui sont similaires par leurs formes (Brun-Trigaud, 2012) (voir figure 2). Ces analyses permettent aussi de révéler des phénomènes jusqu'alors invisibles (Brun-Trigaud, Malfatto, 2013) qu'on ne retrouve pas dans les structures des aires linguistiques classiques. Ces études s'accompagnent généralement d'une volonté d'expliquer ces phénomènes par des facteurs sociohistoriques ou géographiques (Saussure, 1971) (Dalbera, 2013).

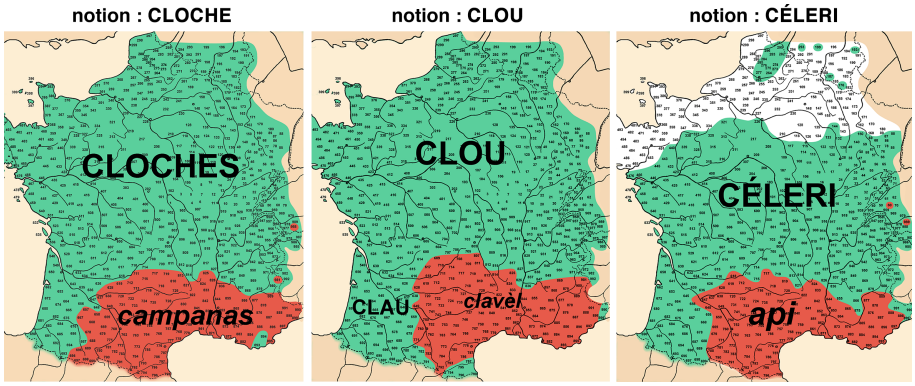


FIGURE *2. Les aires de dispersion des lemmes campanas, clavel et api (en rouge) constituent un motif spatial récurrent. Cartes interprétées de l'Atlas Linguistiques de la France issues du livre de (Brun-Trigaud* et al.*, 2005)*

Si ces analyses typologiques ont d'abord été réalisées de manière intuitive, basées sur des connaissances d'experts, en étudiant la distribution aréale de plusieurs cartes d'un domaine d'étude, elles ont pu être corroborées par des méthodes statistiques de classification appliquées à un plus grand nombre de cartes via une approche dialectométrique (Brun-Trigaud *et al.*, à paraitre). A partir de données issues de plusieurs centaines de carte d'atlas linguistique, les points d'enquête (lieux géographiques visités lors des enquêtes de terrain) sont classifiés en fonction de leurs similarités linguistiques. En utilisant la distance de Levenshtein (Miller *et al.*, 2009) il est possible d'établir une distance linguistique quantitative entre points d'enquête (Heeringa, 2004) ; ils sont ensuite regroupés avec des méthodes classiques de classification (souvent ascendantes hiérarchiques) (Everitt *et al.*, 2011). Il existe des outils (comme Gabmap (Leinonen *et al.*, 2016)) qui utilisent ces approches dialectométriques pour produire des cartographies représentant des régions de cohérences linguistiques autour des points d'enquête.

Les méthodes statistiques de classification permettent en général de retrouver des structures proches des typologies déterminées empiriquement. Toutefois

elles ne traitent pas les aires de dispersion en tant que telles mais plutôt les points d'enquête dont elles sont issues.
Notre proposition de classification consiste à utiliser les aires de dispersion des cartes du corpus comme individus statistiques qui sont classifiées en fonction de leur similarité spatiale. De plus, ni l'analyse typologique empirique des géolinguistes ni l'approche statistique des outils dialectométriques (Nerbonne *et al.*, 2011) n'étudient les liens qui existent entre les caractéristiques linguistiques des structures identifiées et d'autres types de données de contexte (caractéristiques géographiques, géohistoriques, socio-économiques, etc ..) susceptibles de constituer un facteur explicatif.

Dans ce contexte, notre objectif est de proposer une procédure d'analyse spatiale adaptée pour traiter un grand corpus d'aires de dispersion. Nous cherchons en particulier à identifier des ensembles cohérents d'aires de dispersion présentant des co-occurrences spatiales. La difficulté réside dans le fait que les aires de dispersion sont très hétérogènes en taille et en forme, autrement dit leurs limites ne sont jamais les mêmes. Pour comprendre les structures identifiées, notre approche s'appuie sur l'intégration d'un *facteur géographique* pour lequel nous pouvons mesurer l'impact sur ces structures. A ce stade, le facteur géographique dont nous parlons est une partition de l'espace en plusieurs entités géographiques. L'intérêt de la méthode est de croiser des données spatiales multi-sources et multi-formes afin d'établir des liens entre des phénomènes linguistiques et des réalités géographiques, historiques, sociologiques, etc.

Les sections suivantes exposent les étapes successives de notre méthode de classification ainsi que son implémentation. Pour finir, nous illustrons la méthode au travers du cas d'étude sur les données géolinguistiques issues du THESOC.

2. Méthode

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

Classer ces objets nécessite de les comparer entre eux. Pour cela nous devons être en mesure de les caractériser quantitativement selon des critères spatiaux communs. Nous transformons les aires de dispersion en points dans un espace de représentation multi-dimensionnel propre au facteur géographique choisi.

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 supervisée : la *Classification Ascendante Hiérarchique* (CAH) et les nuées dynamiques (k-means) (Nakache, Confais, 2004).

Les étapes de notre méthode sont : la création d'un espace de représentation adapté, puis la projection des objets dans cet espace de représentation et enfin leur classification.

2.1. 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. Sur l'exemple de la figure 4 (a), l'aire *châgne* occupe seulement quelques unités spatiales. 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 3 (a)). Avec cette matrice des aires A nous pouvons obtenir le profil de chaque unités 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 une ensemble de n points dans l'espace vectoriel F (figure 3 (b)).

2.2. Projection barycentrique

Les unités spatiales sont positionnées dans l'espace de représentation F propre au facteur géographique choisi ; nous projetons les aires de dispersion dans ce même espace. De la même façon que pour la première étape décrite dans la section 2.1 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

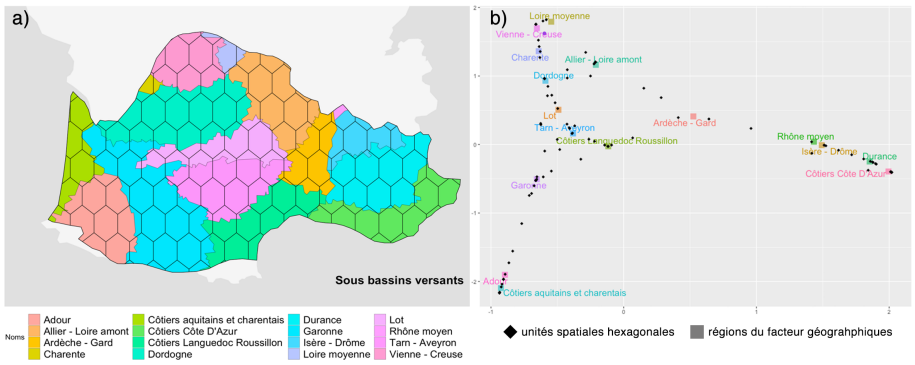
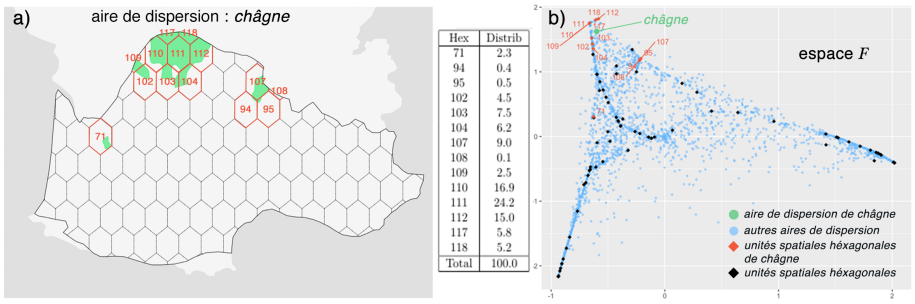


FIGURE *3. Analyse des correspondances : (a) superposition des régions du facteur géographique et du maillage hexagonal (b) représentation simultanée (semi-barycentrique) des p régions du facteur géographique et des n unités spatiales hexagonales, selon les deux premiers axes factoriels.*

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 (figure 4 (b)).



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 *4. Projection barycentrique de l'aire de dispersion du lemme 'châgne' dans l'espace de représentation F ici visualisé selon ses deux premiers axes factoriels.*

2.3. Classification

La dernière étape de notre analyse consiste en une procédure de classification supervisée des points des aires de dispersion 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). A ce stade nous procédons de la manière suivante : tout d'abord nous utilisons un algorithme de Classification Ascendante Hiérarchique (CAH) avec différents critères d'agrégation tels que Ward, saut minimal, saut maximal et distance moyenne. Le dendrogramme produit par la CAH peut être coupé à n'importe quel niveau en fonction du nombre de classes souhaitées. Nous proposons la possibilité de consolider les classes avec un l'algorithme *k-means* dont les points de départ sont initialisés avec les centroïdes des classes déjà établies précédemment par la CAH (Nakache, 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.

3. 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 [2]. 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.

3.1. 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_k d'aires de dispersion de la classe. Un indice (ou score de concentration) n_k/N_{max}, est alors créé, N_{max} étant le plus grand nombre d'aires superposées identifié pour cette classe. Plus la valeur de n_k/N_{max} est proche de 1 plus la concentration en aires

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

sur la k-ème zone est forte. Ainsi on visualise l'épicentre de la classe qui peut être interprété comme l'origine géographique de diffusion d'un phénomène (voir figure 5).

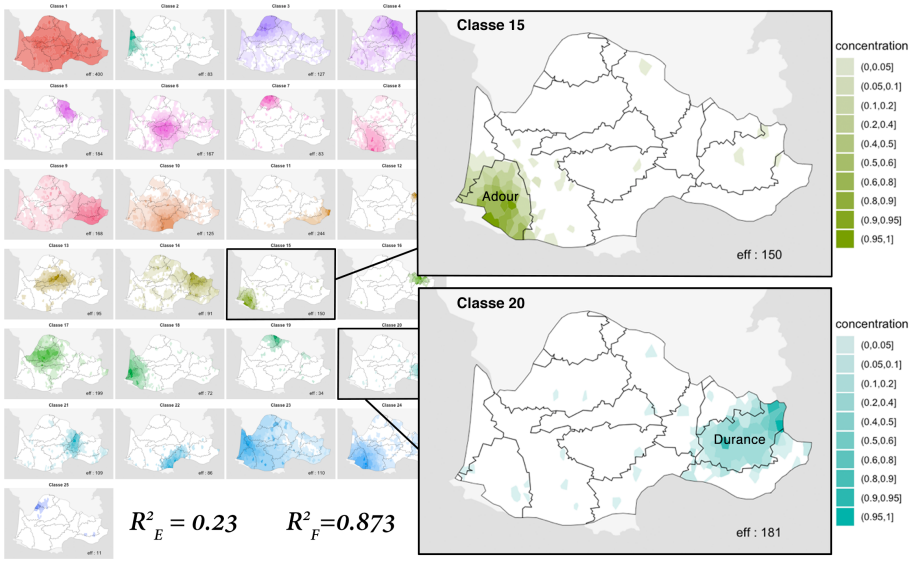


FIGURE 5. *Classification en 25 classes à l'aide du facteur géographique des sous-bassins versants. Agrégation avec le critère Ward et consolidée avec k-means.*

Nous permettons également d'explorer chaque classe, notamment à travers une analyse des caractéristiques géographiques (contour, extension, localisation) et thématiques des aires de dispersion qui la compose. L'analyse géographique vise à dégager le meilleur représentant de la classe (parangon) en relation avec le profil de concentration. Les autres individus de la classe sont triés par ordre décroissant de représentativité. L'analyse thématique quant à elle a pour objectif de caractériser chaque classe à l'aide de variables illustratives : la distribution des thèmes 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 6 et 7.

3.2. Cas d'étude

Nous avons appliqué cette 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 collectées 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 3437 aires lexicales est ainsi créé, à partir duquel, des co-ocurrences 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 3 a)), administratives (les départements) ou historiques (les Généralités en 1789 ou les provinces gauloises en 450)).

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

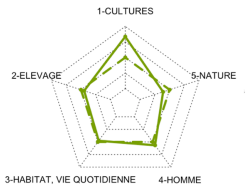


FIGURE *6. Analyse de 5 thèmes de la classe 15 (Adour) : tableau des fréquences et diagramme radar. Sur le diagramme, le trait en pointillé représente le profil moyen, le trait plein représente le profil de la classe.*

L'analyse montre que pour réaliser une classification avec le facteur géographique des sous-bassins versants, nous devons travailler avec au moins 25 classes pour voir se dégager des regroupements intéressants. On voit immédiatement que certains sous-bassins versants regroupent 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 lemmes, de celui de la *Durance* avec 181 lemmes ou encore de l'*Adour* avec 150 lemmes (voir figure 5). 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. On notera que 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.

Notre approche s'inscrit dans une démarche d'analyse exploratoire car elle permet d'étudier facilement plusieurs classification en faisant varier les critères et offre également la possibilité d'analyser les classes ainsi établies. Cette analyse permet de trouver s'il existe, par exemple, une interprétation par les

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

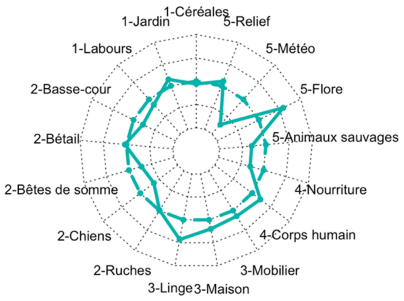


FIGURE *7. Analyse de 17 sous-thèmes de la classe 20 (Durance) : tableau des fréquences et diagramme radar. Sur le diagramme, le trait en pointillé représente le profil moyen, le trait plein représente le profil de la classe.*

thèmes qui pourrait unir les aires lexicales ainsi regroupées. Ainsi dans le cas de la classe décrivant le bassin de l'Adour, le diagramme radar de la figure 6 nous montre une prédominance d'aires appartenant au vocabulaire de la culture (auquel appartient le terme *dragon*). Quant au diagramme radar de la figure 7 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. Ce traitement informatique des données oriente rapidement le linguiste vers des pistes de recherches précises dont il peut définir lui-même les critères qui lui semblent le plus pertinents.

4. Discussion et conclusion

En proposant une méthode de classification d'objets géographiques surfaciques représentant des aires de dispersion de phénomènes géolinguistiques, il devient alors possible d'identifier des co-occurences spatiales et de créer des groupes cohérentes. Cette méthode implémentée selon une approche d'analyse exploratoire de données géographiques offre l'opportunité aux linguistes d'identifier plus facilement les liens existants entre des phénomènes linguistiques et des facteurs géographiques (ceux-ci étant jusqu'à maintenant essentiellement basés sur leur connaissance du territoire). Cependant sa validation nécessite de

l'appliquer à une diversité de jeux données géolinguistiques mais aussi dans le cadre d'autres thématiques mobilisant des données similaires.

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 E. CARRILHO, C. MARGO, X. ALVAREZ (Eds.), *Current Approaches to Limits and Areas in Dialectology*, p. 293-310. Cambridge Scholars Publishing.

Brun-Trigaud G., Malfatto A., Sauzet M. (à paraitre). Essai de typologie des aires lexicales occitanes : regards dialectométriques. In *Fidélités et dissidences. 12e Congrès de l'Association Internationale d'Etudes Occitanes (Albi, 10-15 juil. 2017)*. Albi, France.

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. ACM.

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

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 34*, p. 367-387.

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

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

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

Lafkioui M. B. (2015). Méthodologie de recherche en géolinguistique. *Corpus*, vol. 14, p. 139-164.

Leinonen T., Çoltekin c., 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.

Miller F. P., Vandome A. F., McBrewster J. (2009). *Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau?levenshtein distance, spell checker, hamming distance*. Alpha Press.

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

Nerbonne J., Colen R., Gooskens C., Kleiweg P., Leinonen T. (2011, 01). Gabmap - a web application for dialectology. *Dialectologia*.

Rencher A. C. (2002). *Methods of multivariate analysis* (second edition éd.). Hoboken, NJ, USA, Wiley-Interscience.

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