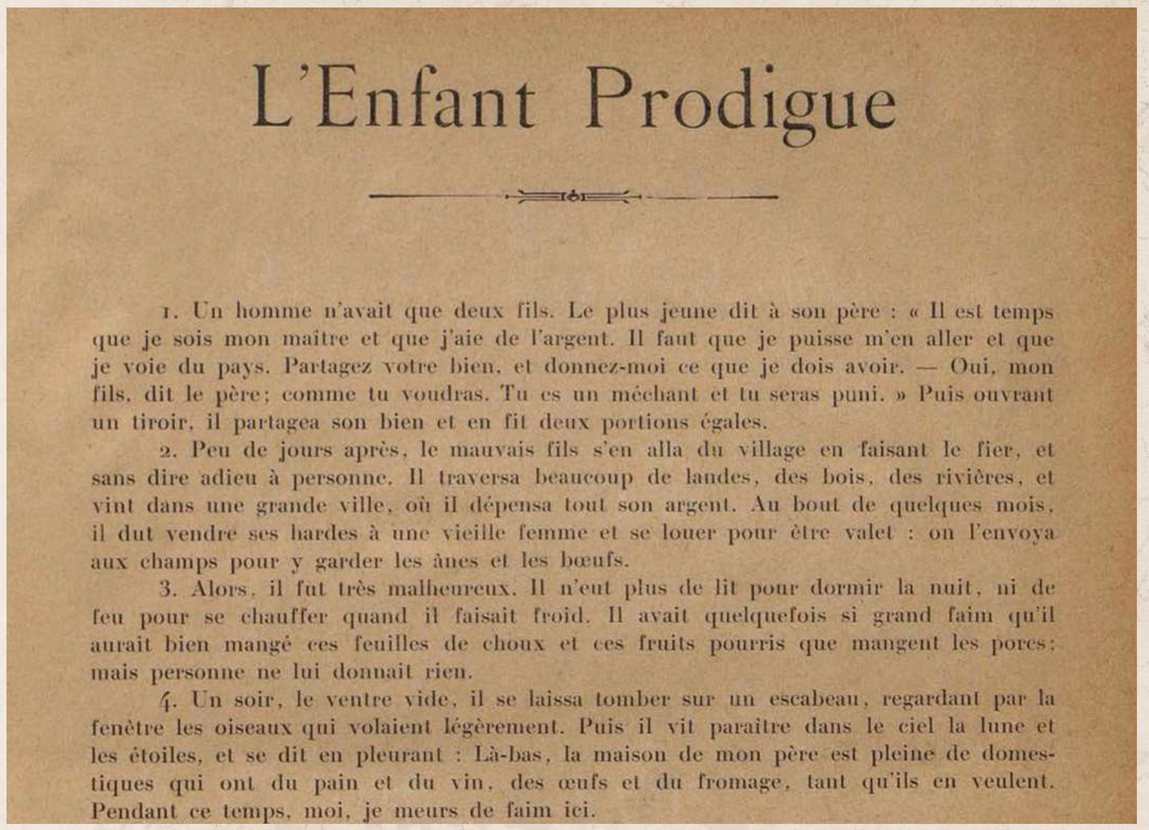
# Analyse statistique d'une enquête linguistique : Projet Bourciez

ABARKAN Suhaïla, MOUCHRIF Dounia,
ROMAN Karina

# Sommaire



# 1. Contextualisation - Enquête



L'Enfant Prodigue

1. Un homme n'avait que deux fils. Le plus jeune dit à son père : « Il est temps que je sois mon maître et que j'aie de l'argent. Il faut que je puisse m'en aller et que je voie du pays. Partagez votre bien, et donnez-moi ce que je dois avoir. — Oui, mon fils, dit le père; comme tu voudras. Tu es un méchant et tu seras puni. » Puis ouvrant un tiroir, il partagea son bien et en fit deux portions égales.

2. Peu de jours après, le mauvais fils s'en alla du village en faisant le fier, et sans dire adieu à personne. Il traversa beaucoup de landes, des bois, des rivières, et vint dans une grande ville, où il dépensa tout son argent. Au bout de quelques mois, il dut vendre ses hardes à une vieille femme et se louer pour être valet : on l'envoya aux champs pour y garder les ânes et les bœufs.

3. Alors, il fut très malheureux. Il n'eut plus de lit pour dormir la nuit, ni de feu pour se chauffer quand il faisait froid. Il avait quelquefois si grand faim qu'il aurait bien mangé ces feuilles de choux et ces fruits pourris que mangent les porcs; mais personne ne lui donnait rien.

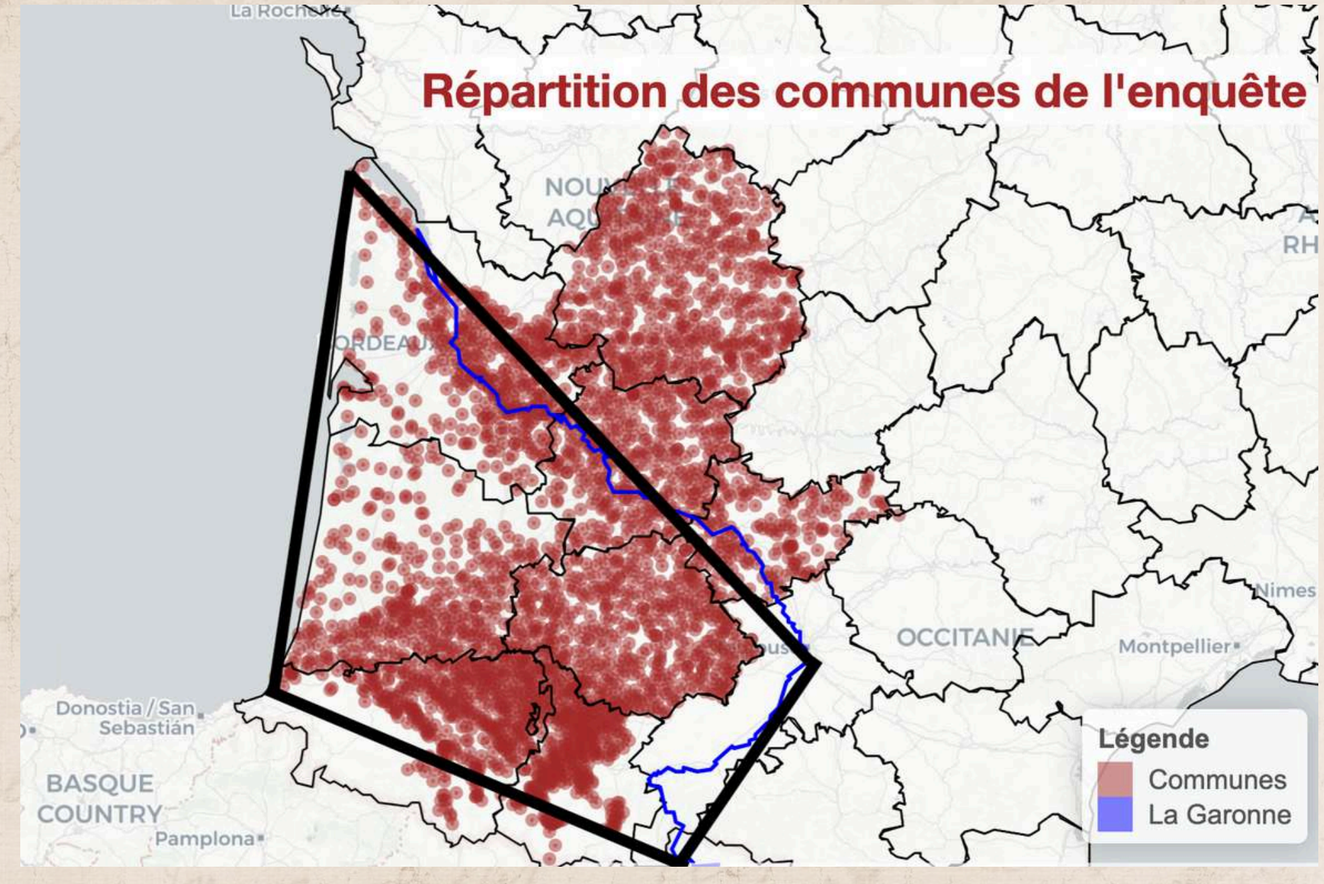
4. Un soir, le ventre vide, il se laissa tomber sur un escabeau, regardant par la fenêtre les oiseaux qui volaient légèrement. Puis il vit paraître dans le ciel la lune et les étoiles, et se dit en pleurant : Là-bas, la maison de mon père est pleine de domestiques qui ont du pain et du vin, des œufs et du fromage, tant qu'ils en veulent. Pendant ce temps, moi, je meurs de faim ici.

- <u>Enquête linguistique :</u>
  - Menée par Edouard Bourciez (Professeur à l'Université de Bordeaux) en 1894-1895
  - Académies de Bordeaux et Toulouse
- <u>Méthode :</u> Traduction de la parabole de "l'Enfant Prodigue" par les instituteurs de la commune où ils enseignent
- <u>Localisation :</u> 10 départements du sud-ouest de la France, englobant la région historique de la Gascogne
- <u>But de l'enquête :</u> Visualisation des variations diatopiques (à travers l'espace) de ces idiomes (français et occitan (langues romanes), basque)

# 1. Contextualisation - Géographie

- La Gascogne :
  - Régions françaises (Nouvelle-Aquitaine et Occitanie)
  - Définie par la Garonne, les Pyrénées et l'océan Atlantique

- L'occitan (incluant le gascon) :
  - Tiers sud de la France
  - Catalogne dans le nord est de l'Espagne
  - Vallées occitanes en Italie



# 1. Données

| | département | canton | commune | x | y | UN_Mot | HOMME_Mot | NE_Mot | AVAIT_Mot | PAS_Mot |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gironde | Canton d'Audenge | Andernos-les-Bains | -1.09251840 | 44.74769 | ün | homme | n | aouè | *NA* |
| 2 | Gironde | Canton d'Audenge | Arès | -1.13799280 | 44.76764 | ün | homme | n | aouait | *NA* |
| 3 | Gironde | Canton d'Audenge | Audenge-bis | -1.01549929 | 44.69498 | un | homme | n | aoué | *NA* |
| 4 | Gironde | Canton d'Audenge | Audenge | -1.01985680 | 44.68836 | ün | homme | n | aoué | *NA* |
| 5 | Gironde | Canton d'Audenge | Biganos | -0.96962710 | 44.64409 | ün | home | n | aoué | pas |

**Déroulement :**

➙ 3061 réponses à l'enquête

➙ Numérisées à l'université de Montaigne

➙ Transcriptions et segmentation informatique, puis création de la base de données (par M. Genadot et son équipe)

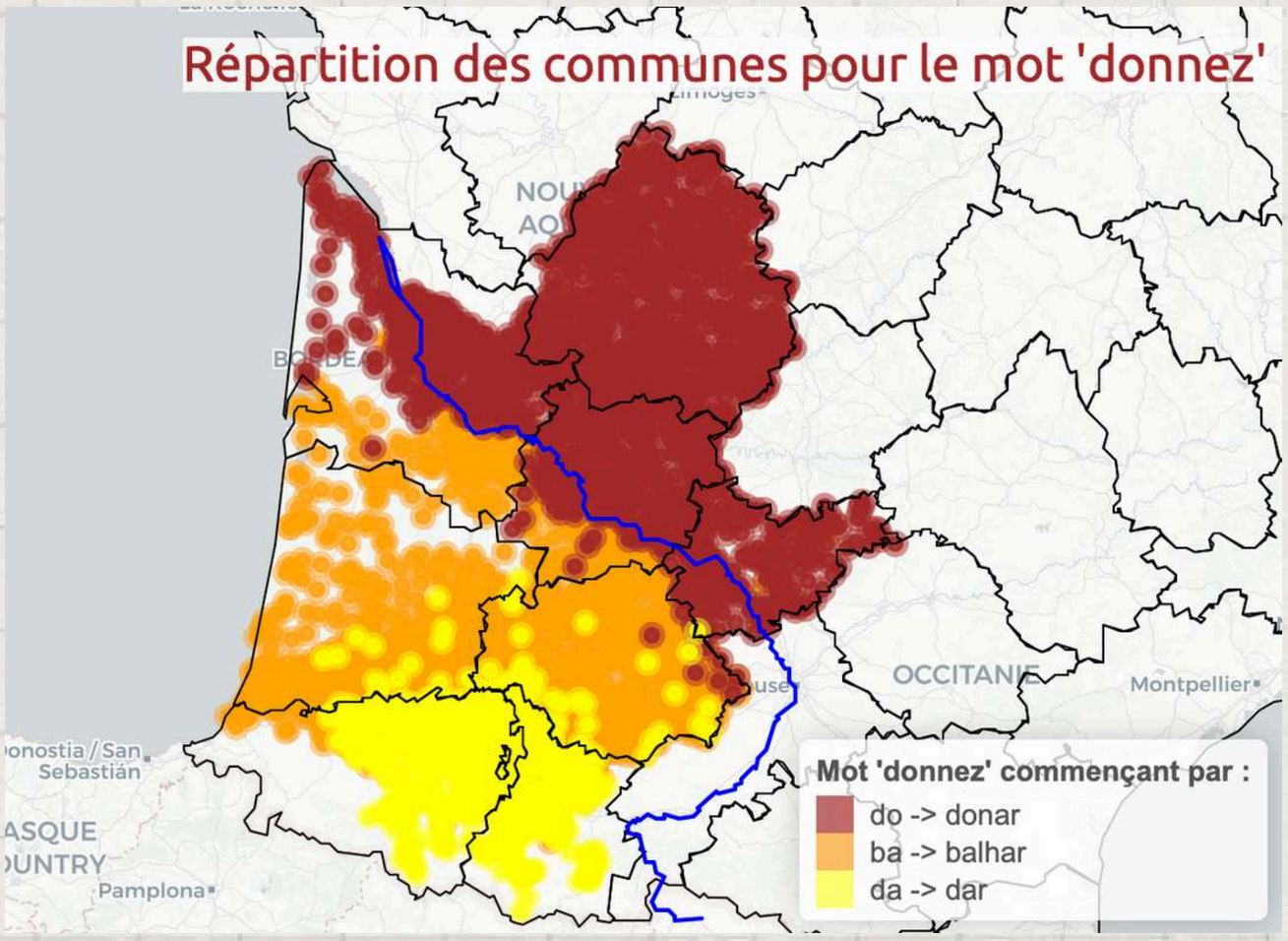
**Données qualitatives, textuelles :**

- en lignes : les 3061 communes (1 commune par ligne)
- en colonnes : 5 colonnes d'indications géographiques (département, canton, commune et coordonnées géographiques)
- 101 colonnes de traductions de chaque mot (1 mot par colonne)
- 3% de données manquantes

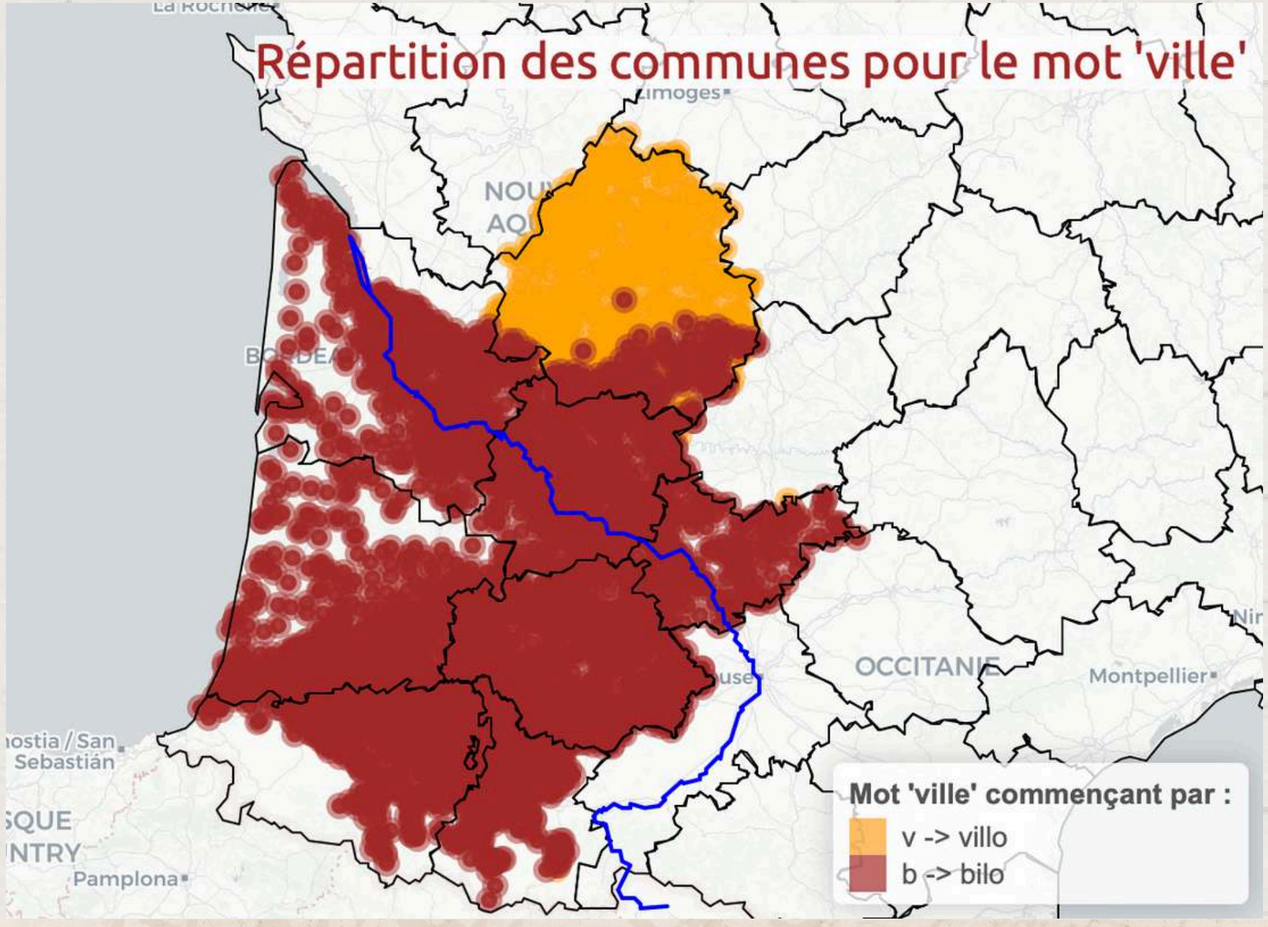
# 2. Cartes des caractéristiques linguistiques - Variation lexicale

Mot “donner” qui possède 3 lemmes :

- <donar> : cluster rouge
- <balhar> : cluster orange
- <dar> : cluster jaune




# 2. Cartes des caractéristiques linguistiques - Variation phonologique




**Mots commençant par “V” en latin qui possèdent des prononciations distinctes :**

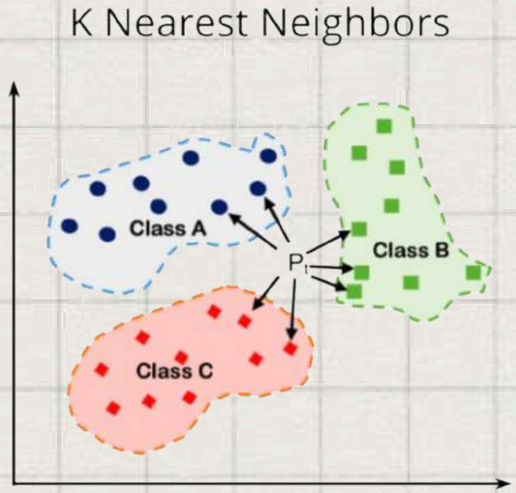
- **/v/ : cluster orange**
- **/b/ : cluster rouge**

# 3. Clustering avec la méthode des KNN

**KNN (Méthode des K plus proches voisins) :** Méthode de classification supervisée utilisée pour attribuer une classe à un nouvel échantillon en se basant sur les classes des échantillons voisins dans l'espace des caractéristiques et donc sur les distances entre chaque donnée obtenue grâce à l'ACM

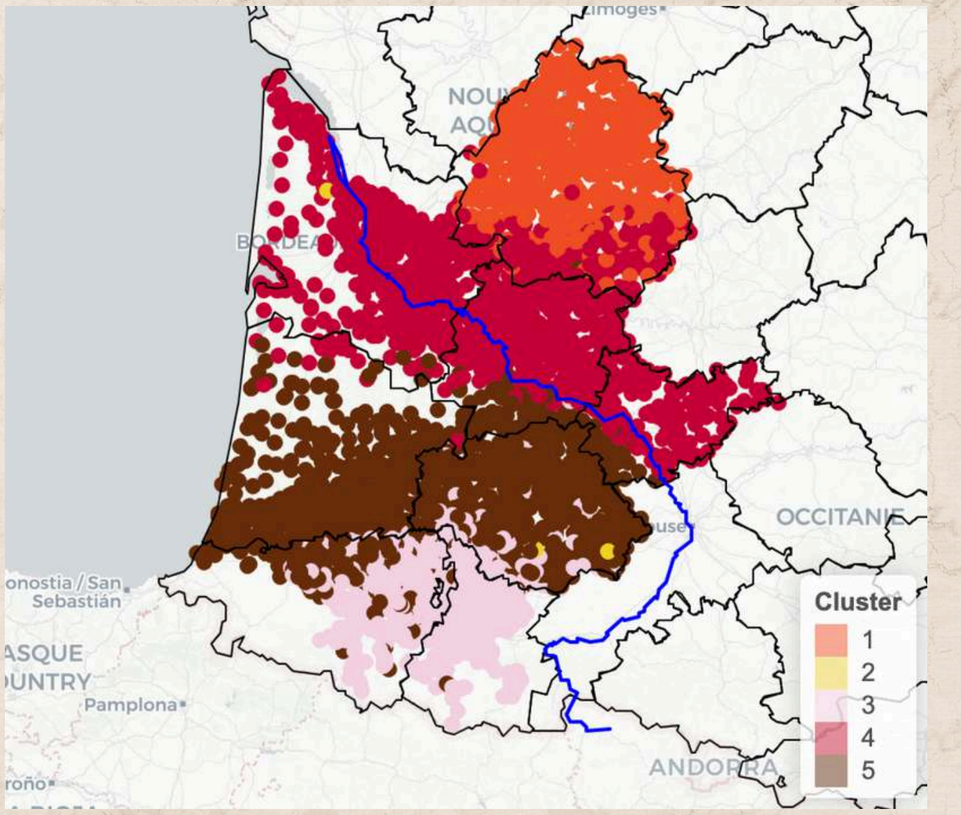
**Données qualitatives → ACM (Analyse des Correspondances Multiples)**

**Exemple**



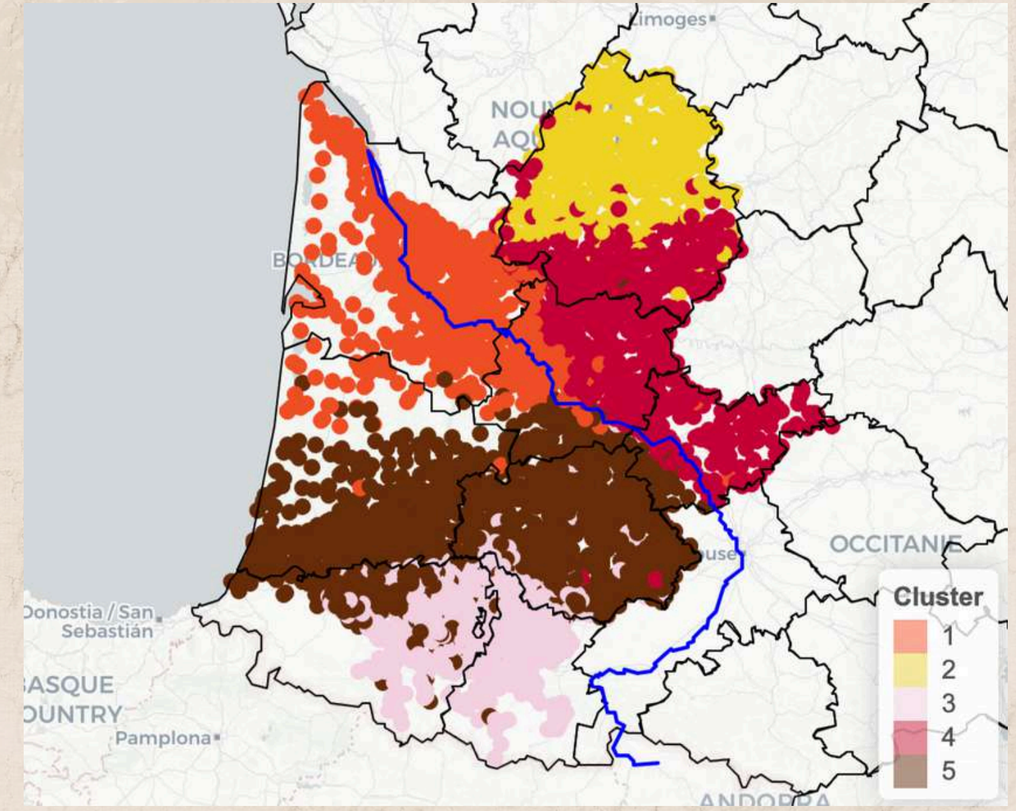
# 3. Clustering avec la méthode des KNN

En prenant compte des accents




| Numéro de cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Nombre de communes | 387 | 4 | 697 | 1024 | 949 |

En ne prenant pas compte des accents




| Numéro de cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Nombre de communes | 548 | 298 | 673 | 581 | 961 |

# 3. Explication des clusters

**Caractérisation des clusters obtenus avec la méthode des KNN en 5 classes à partir de la caractéristique “finissant par “os”, “es”, “as” ou NA”**

En prenant compte des accents

```
$`1`
                        Cla/Mod   Mod/Cla    Global      p.value     v.test
Finit_par_osesas=os    34.814815 71.816638 39.692911 1.626523e-69  17.623490
Finit_par_osesas=autre 27.408994 21.731749 15.256452 2.784738e-06   4.686090
Finit_par_osesas=as     4.575163  2.376910  9.996733 5.906734e-15  -7.805924
Finit_par_osesas=es     2.236719  4.074703 35.053904 1.714327e-87 -19.827782

$`2`
                         Cla/Mod    Mod/Cla    Global      p.value    v.test
Finit_par_osesas=as    47.05882353 48.3221477  9.996733 1.723744e-76  18.50973
Finit_par_osesas=autre 30.83511777 48.3221477 15.256452 3.203433e-47  14.43310
Finit_par_osesas=os     0.74074074  3.0201342 39.692911 5.490094e-55 -15.61800
Finit_par_osesas=es     0.09319664  0.3355705 35.053904 5.311648e-58 -16.05456
```

En ne prenant pas compte des accents

```
$`1`
                        Cla/Mod    Mod/Cla    Global       p.value     v.test
Finit_par_osesas=es    46.6915191 94.5283019 35.053904 1.246826e-226  32.13928
Finit_par_osesas=autre  4.2826552  3.7735849 15.256452  5.742054e-20  -9.14908
Finit_par_osesas=as     0.6535948  0.3773585  9.996733  4.794444e-24 -10.11389
Finit_par_osesas=os     0.5761317  1.3207547 39.692911 1.456529e-117 -23.05043

$`2`
                        Cla/Mod   Mod/Cla    Global      p.value    v.test
Finit_par_osesas=as    50.9803922 49.523810  9.996733 2.121551e-86  19.70082
Finit_par_osesas=autre 31.4775161 46.666667 15.256452 6.875028e-46  14.22010
Finit_par_osesas=es     0.2795899  0.952381 35.053904 8.812079e-58 -16.02312
Finit_par_osesas=os     0.7407407  2.857143 39.692911 5.079593e-59 -16.19955
```

# 4. Classification ascendante hiérarchique

- **Comparaison des idiomes** avec des distances linguistiques

➙ Distance de **Jaccard** : $\frac{MOT_1 \cap MOT_2}{MOT_1 \cup MOT_2}$

*Exemple : DJ(aimer , manger) = 4\7 = 0,57*

➙ Distance de **Levenshtein** : nombre minimum d'opérations nécessaires pour passer d'une chaîne de caractère à l'autre

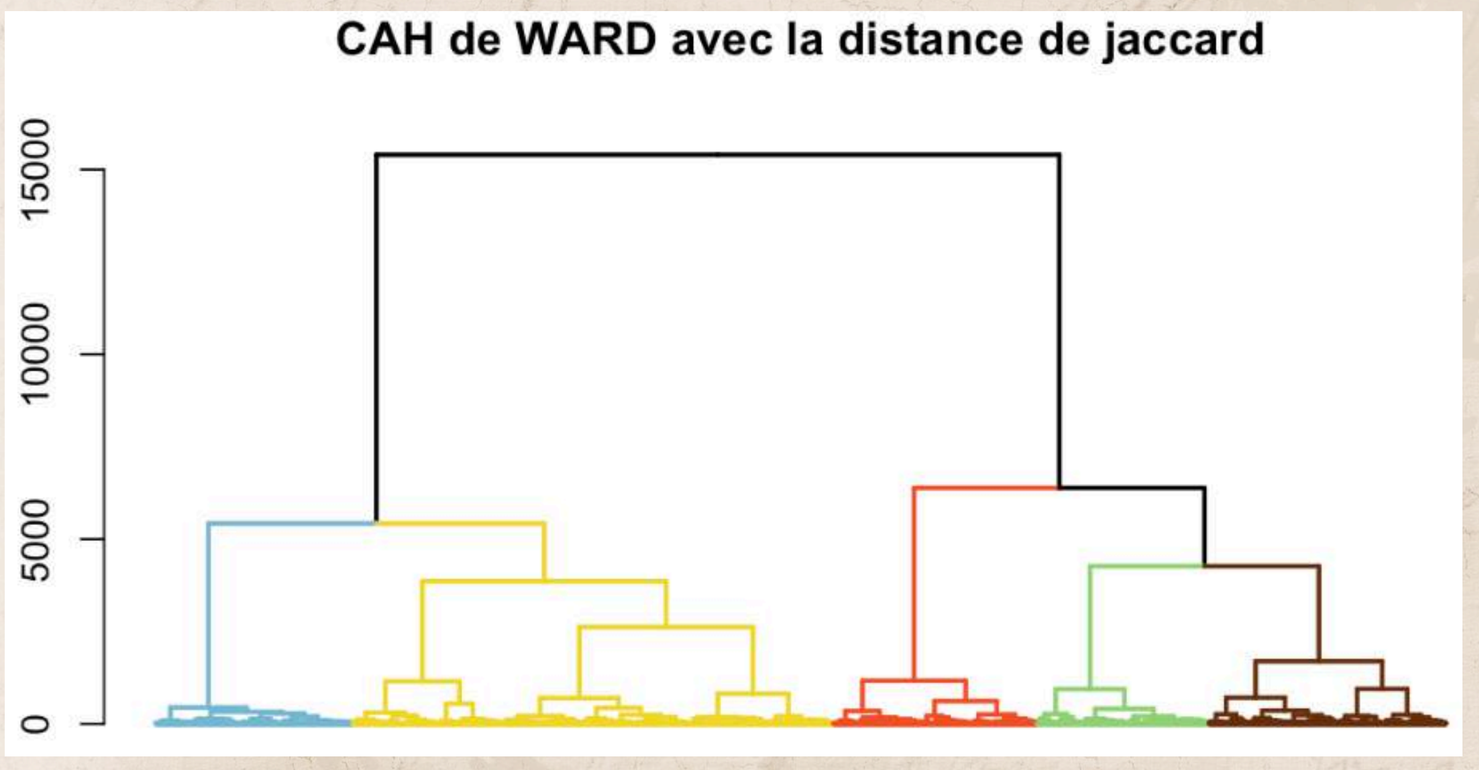
*Exemple : DL(party , park) = 2 (1 remplacement + 1 suppression)*

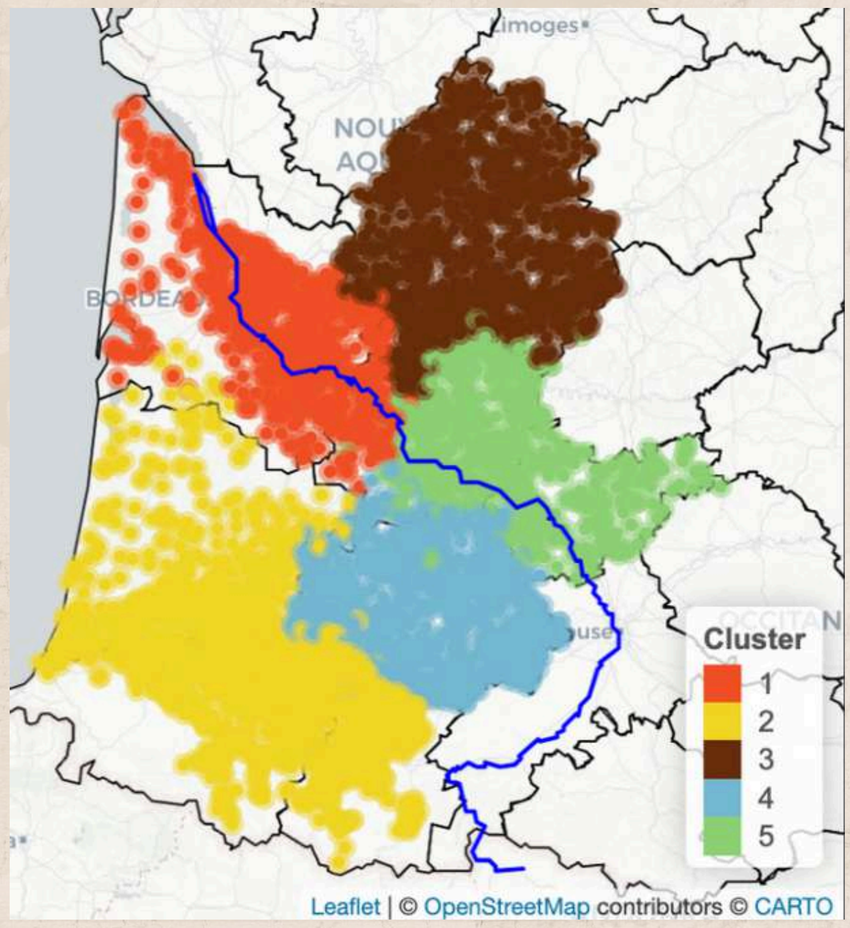
- Réalisation de **matrice de dissimilarités** à partir de ces distances pour construire des arbres de classification

- **Méthode de classification :** WARD et COMPLETE
  ➙ Structures de cohérance differentes

# 4. Dendrogramme et carte de clustering

## Visualistaion de la partition obtenue

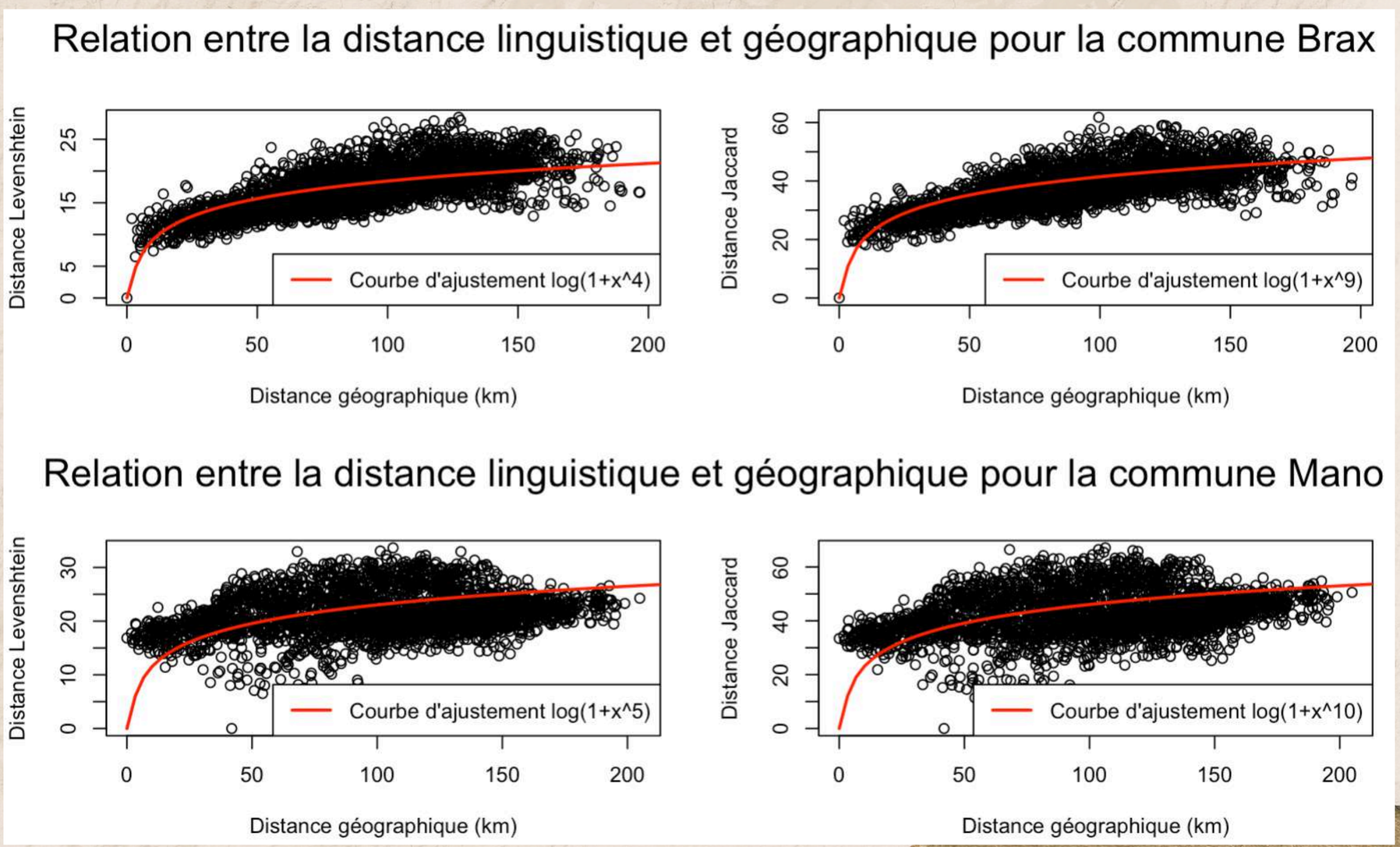


# 5. Distance linguistique et géographique

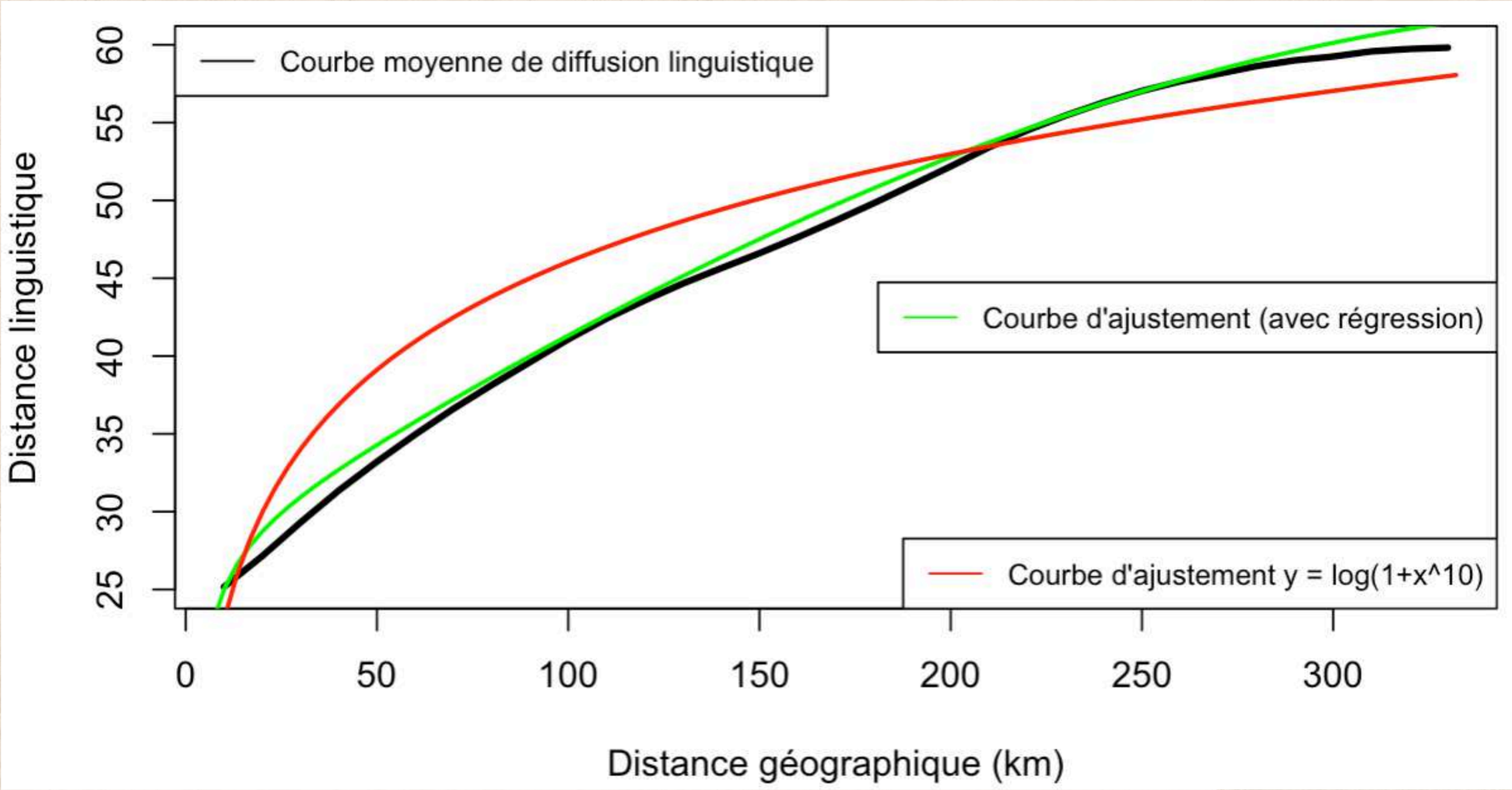
## Corrélation entre les distances

| | Distance géographique |
|---|---|
| Distance de Levenshtein | 0.794122 |
| Distance de Jaccard | 0.7895414 |



# 5. Distance linguistique et géographique

## Courbe moyenne de diffusion linguistique



# 6. Conclusion

- **Application d’outils statistiques simples pour aider les linguistes à l’analyse des données**

- **Réalisation d’une application interactive Shiny regroupant tous nos résultats**

# Bibliographie


https://www.auch-tourisme.com/loccitan-et-le-gascon/

https://www.univ-montp3.fr/uoh/occitan/une_langue/co/module_L_occitan_une%20langue_24.html

https://escolagastonfebus.com/langue/le-gascon-culture-et-langue-indissociables/

https://fr.wikipedia.org/wiki/Gascon

https://shiny.posit.co

Measuring the diffusion of linguistic change, John Nerbonne,
https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2010.0048

*Classification automatique - clustering*, cours de Mme Marie Chavent

https://irem.univ-lille.fr/~site/IMG/pdf/142_distance_entre_les_mots.pdf

https://lrouviere.github.io/TUTO_VISU/shiny.html

# Merci pour votre attention !

# Temps pour des questions!