Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts
Western Switzerland

MSc HES-SO en Business Administration

Orientation :
Management des Systèmes d'information

Analyse et prévision des choix électoraux

le cas du Conseil d'Etat valaisan

Réalisé par
Renzo Scuderi

Sous la direction de
Prof. Florian Evéquoz

Lausanne, août 2019

REMERCIEMENTS

Je tiens à présenter mes remerciements à toutes les personnes qui m'ont aidé et soutenu durant la réalisation de mon travail de Master.

Je souhaite ainsi remercier tout particulièrement :

Dr. Florian Evéquoz, Professeur HES-SO Valais, pour ses précieux conseils, son regard critique, sa grande disponibilité et l'aide apportée dans la structuration des différentes parties de ce travail. Il a su me conseiller et m'orienter dans la bonne direction en me laissant libre de mes choix.

Bernhard Altermatt, Historien (chercheur et enseignant), pour l'aide apportée dans le démarrage du projet.

Vanessa Fresquet, Tania Schaad et Luca Scuderi pour leurs relectures et leurs corrections de mon rapport.

ABREVIATIONS

MSc	Master of sciences
Km^2	Kilomètre carré
PCA	Principal Component Analysis
WSS	Total within-cluster sum of square
PDC	Parti démocrate-chrétien
PLR	Les Libéraux-Radicaux
UDC	Union démocratique du centre
PS	Parti Socialiste
Indé.	Indépendant
PRD	Parti radical-démocratique
PCS	Parti chrétien-sociale
RCV	Rassemblement Citoyen Valais

ABSTRACT

On observe une tendance mondiale à la mise en place d'open. Cette mise à disposition des données issues des administrations publiques rend possible de nouvelle approche centrée sur les données pour divers domaines et notamment les sciences politiques. Plusieurs études ont démontré l'efficacité de l'utilisations de ces sources de données pour analyser ou prédire des situations politiques. En Turquie, plusieurs études ont montré des approches par partitionnement de données qui ont permis d'analyser des situations et de confirmer des aprioris politiques. Dans le domaine de la prédiction, une étude suisse a proposé un modèle capable de prédire un résultat de vote à l'échelle nationale sur la base de résultats d'une seule commune.

Tablant sur ces résultats encourageant, nous avons cherché à mesurer l'impact de la langue et de différents critères sur les comportements électoraux d'une population. Nous avons voulu mesurer la capacité de prédiction d'un modèle basé sur les données dans le cas d'une élection de personnes. Pour réaliser cela, nous nous sommes intéressés au cas particulier de la suisse et plus précisément d'un canton bilingue, le Valais

Nous avons étudié l'élection au Conseil d'Etat valaisan entre les années 2001 et 2017, soit cinq occurrences. Nous avons utilisé une approche par partitionnement des données afin de séparer les différents comportements de vote des électeurs. Nous avons ensuite mesuré l'homogénéité des groupes en fonction des facteurs étudiés que sont la langue, la répartition des âges, la répartition en secteur économique, les partis politique et la densité de populations des communes. A partir de ces annotations, nous avons regardés quels facteurs expliquent de la façon la plus robuste les compositions des groupes. Dans une seconde partie, nous avons testés la capacité de prédiction des communes sur le résultat final.

L'application de notre méthodologie a permis d'obtenir des résultats intéressants dans l'analyse de la situation politiques valaisanne ainsi que dans la prédiction des élus. Tout d'abord, aucun facteur testé n'explique mieux la composition des groupes de comportements de vote que la langue des communes. En effet, les groupes sont composés au minimum à 90% de commune de même langue et deux années sur cinq cette valeur est de 100%. Ensuite, lorsque l'on regarde la différence de comportement entre commune de même langue alors deux facteurs se démarque. Pour les communes francophones, nous avons observés un clivage basé sur la densité de population. Dans le cas des communes alémanique se sont les secteurs d'activités qui diverges. Nos modèles de prédictions ont fait émerger des communes avec des caractéristiques intéressantes. La petite commune de Chippis possède un taux 96% de prédiction lorsque l'on s'intéresse à la liste des personnes élues. Ensuite, La commune de Crans-Montana est la commune dont la répartition des voix est le plus proche du résultat final. Enfin, nous avons montré qu'un apprentissage sur la base de données d'exemple ne fonctionnais pas pour notre jeu de données, principalement à cause du nombre insuffisant d'exemples et du grand nombre d'élus par rapport au nombre de candidat.

Nous avons confirmé que la langue était le facteur explicatif le plus robuste pour les comportements de votes du Conseil d'Etat valaisan. Nous avons montré un clivage ville-campagne pour la région francophone et un vote de classe du coté alémanique. Nos modèles de prédictions fond ressortir des communes avec des caractéristiques intéressantes qui pourraient être exploitées pour des sondages afin de prédire un résultat.

Finalement, nos résultats invitent à continuer à explorer ces approches avec plus de données. Nous pourrions commencer par intégrer d'autres années d'élections dans nos données afin de confirmer les tendances observées. Ensuite, il serait intéressant d'appliquer cette approche à d'autres cantons bilingues et observer les différences, puis de tester la méthode sur un canton non-bilingue afin d'observer quel autre facteur émerge à la place de la langue. Enfin, avec davantage de donnée, des modèles de prédictions basés sur de l'apprentissage pourraient rendre de bon résultat en intégrant en plus, par exemple, les caractéristiques personnelles dans candidats.

Mot clés : data sciences, data mining, sciences politique, Valais, élections, démocratie, analyses, prédictions

Chapitre 1. INTRODUCTION

1.1 CONTEXTE GÉNÉRAL

Ce rapport est mon Travail de Master de fin d'études MSc en Business Administration orientation Management des Systèmes d'Information. Il concerne la data science appliquée aux sciences politiques.

Le 16 avril 2014, la Confédération a approuvé une stratégie nationale d' « Open Government Data » [1]. Celle-ci a pour objectif de mettre à disposition des citoyens des données officielles à travers un portail centralisé (opendata.swiss) afin de gérer la création et la distribution des données. La Confédération, les cantons ou encore les communes peuvent publier leurs données sur ce portail. La stratégie de la Confédération est de créer une culture de « l'open datas » en mettant à disposition tous les outils nécessaires aux administrations publiques pour partager leurs données [1]. En juillet 2019, le portail « opendata.swiss » met à disposition 6921 jeux de données sur des sujets comme l'administration, l'agriculture ou encore la politique [2]. La plateforme décrit sa stratégie comme la promotion de la transparence, de la participation et de l'innovation chez les citoyens [2]. La Suisse n'est pas un cas isolé et l'on observe une tendance mondiale à la publication d'open data par les gouvernements [3]. Nous pouvons notamment citer les États-Unis avec leur plateforme data.gov lancée par Barack Obama [4] et qui vise plus de transparence des pouvoirs publics par là mise à disposition d'open data.

Ces stratégies politiques ont pour effet de mettre à disposition du public une grande quantité de données structurées dans domaines variés. De nouvelles perspectives de data science – au sens exploration de données brutes pour en tirer des informations utiles – deviennent possibles et notamment pour les sciences politiques. Le grand nombre de données et les nombreuses catégories rendent difficile de déterminer ce que cela peut apporter à un domaine précis comme les sciences politiques.

Dans ce rapport, nous allons nous intéresser à déterminer les apports de l'utilisation des méthodes de datas science pour les sciences politiques. Nous proposerons un exemple de mise en œuvre de compétence de datas science à des fins d'analyse et de prédictions dans le domaine de la politique. Nous appliquerons cela à un cas pratique : les élections du Conseil d'État valaisan. Nous décrirons une méthode structurée pour l'analyse des comportements électoraux des citoyens et nous explorerons les possibilités de prédictions de résultats électoraux en exploitant les résultats précédents.

1.2 CONTEXTE POLITIQUE SUISSE

Nous parlerons rapidement dans ce chapitre du système fédéral Suisse afin de montrer le rôle important des cantons dans la politique Suisse. En effet, le cas étudié dans ce rapport est une élection cantonale, et il est nécessaire de discuter de ce système fédéral afin de souligner les enjeux lors d'élection d'un Conseil d'État.

Ensuite, nous discuterons du système politique du canton du Valais ainsi que des partis politiques présents dans ce canton afin de mieux comprendre le contexte des élections qui seront analysées.

1.2.1 Le fédéralisme en Suisse

La Suisse est un État fédéral qui compte quatre langues nationales : l'allemand, l'italien, le français et le romanche. Cette fédération est composée de 26 cantons et 2212 communes [5]. Les pouvoirs sont répartis entre la Confédération, les cantons et les communes. La Confédération a autorité sur des sujets comme la politique extérieure, les douanes, la monnaie ou encore la défense. Cependant, tous les pouvoirs ne sont pas transférés à la confédération. En effet, les cantons ont une autonomie politique sur des sujets comme la finance, l'énergie, le système politique ou la formation [6]. Les communes quant à elles disposent d'un pouvoir de décision sur d'autres thèmes comme la protection sociale ou la fiscalité [6]. Cette répartition des pouvoirs est garantie dans l'article 3 de la Constitution fédérale de la Confédération Suisse [7].

Avec ce système fédéral, la Suisse est un des pays les plus décentralisés [8]. Cette autonomie des cantons sur des sujets centraux donne de l'importance aux élections des gouvernements cantonaux et notamment l'élection du

Conseil d'État valaisan. Il ne s'agit pas d'élire des gouvernements locaux sans pouvoir face aux décisions prises par la Confédération. Ils disposent d'une vraie autonomie définie dans la constitution sur des sujets de société importants. En effet, la Constitution fédérale de la Confédération Suisse définit dans l'article 5a le principe de subsidiarité pour l'attribution et l'accomplissement des tâches étatiques [7]. Cela impose que chaque action publique doit être menée par la plus petite entité capable de l'assumer, et d'autre part que si l'entité n'en est pas capable alors l'entité supérieure doit l'y aider [6].

1.2.2 Le canton du Valais

Ce chapitre a pour but de présenter plus en détail le canton du Valais. Nous parlerons de sa structure géographique, de son organisation politique et de ces spécificités. Nous présenterons ensuite les différentes forces politiques présentes en listant les partis politiques. Nous décrirons plus en détail l'élection du Conseil d'État ainsi que les missions qui lui sont confiées.

Le Valais est le vingtième canton de la Confédération Suisse. Depuis 2017, il est composé de 126 communes faisant partie de 13 districts [9]. La géographie du valais diffère significativement par rapport à celle des autres cantons suisses. En effet, si l'on cumule les surfaces improductives et les surfaces boisées alors celles-ci représentent 77% des 5'224,25 km2 du canton, la moyenne Suisse étant à 50% [10]. Les surfaces construites représentent quant à elles plus de deux fois moins que la moyenne Suisse avec une valeur de 3.5% par rapport à une moyenne de 8.8% dans le pays. Ces différences s'expliquent par un paysage montagneux où il est extrêmement difficile de construire ou d'exploiter des surfaces agricoles.

Le canton est découpé en trois zones géographiques appelées « bas valais », « valais central » et « haut valais ». Les communes des deux premières zones sont francophones alors que celles de la troisième sont alémaniques.

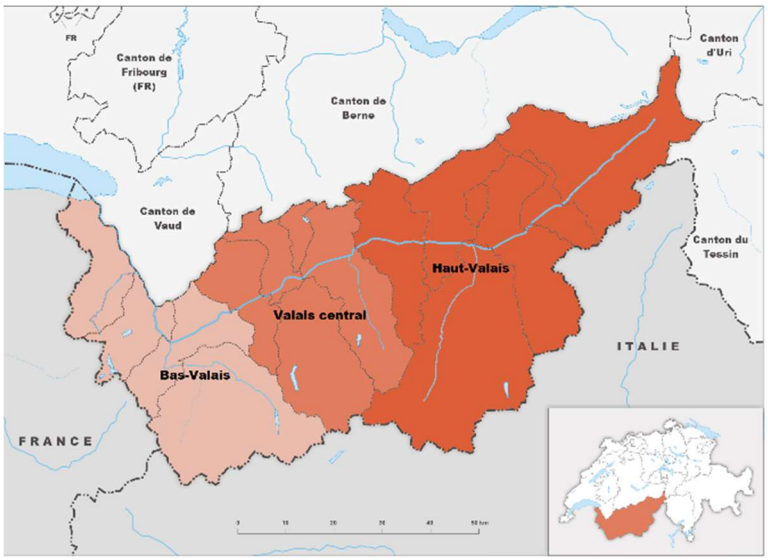


Figure 1 Découpage du canton du Valais en trois zones : bas valais, valais central et haut valais.
Source : https://fr.wikipedia.org/wiki/Bas-Valais#/media/Fichier:R%C3%A9gions_du_Valais.png

En plus de se distinguer par la langue, ces trois zones géographiques disposent de caractéristiques sensiblement différentes. Le Tableau 1 présente la superficie de chacune des zones et la densité de population. Le tableau précise également les valeurs minimums et maximums de densité pour les districts présents.

Tableau 1 Superficie et densité de population par zone géographique [10]. Valeur min et max pour les districts des zones.

	Superficie (km2)	Densité moyenne (hab/km^2)	min, max (hab/km^2)
Bas Valais	1080,37 km^2	114 hab/km^2	24, 180
Valais central	1249,06 km^2	156 hab/km^2	23, 363
Haut Valais	2620,24 km2	32 hab/km2	8, 62

On remarque que la superficie du haut valais est plus de deux fois supérieure aux deux autres et que sa densité de population est très nettement inférieure. Le district de Conches représente la densité minimum du Haut Valais et du canton avec 8 habitants par km^2. Le district de Sierre en représentant 363 habitants par km^2, est le maximum du canton. Nous pouvons noter que même s'il s'agit du même canton, ces différentes zones semblent offrir un cadre de vie différent.

Le Conseil d'État valaisan est une organisation politique chargée d'exercer le pouvoir exécutif et administratif [11]. Ce conseil est composé de cinq membres qui sont élus par le peuple tous les quatre ans. L'élection est un scrutin à deux tours avec un système majoritaire. Le deuxième tour à lieu uniquement si les majorités nécessaires ne sont pas réunies lors du premier tour. La Constitution du canton valaisan impose que parmi les cinq élus, un d'entre eux soit d'origine du bas valais, un d'entre eux du valais central et un du haut valais [11]. Les deux autres sièges peuvent être attribués librement, cependant l'article 52 alinéa 4 de la Constitution cantonale du Valais précise qu'il ne peut pas y avoir plus d'un élu par district [11]. Une fois élus, les cinq membres du Conseil d'état valaisan se répartissent la direction des départements suivants : finances et énergie ; santé, affaire sociale et culture ; économie et formation ; sécurité, institutions et sport ; mobilité, territoire et environnement [12].

La Suisse est un pays avec un nombre de partis politiques au-dessus de la moyenne européenne [13]. Certains partis existent à l'échelle nationale et d'autres uniquement à l'échelle cantonale. Nous allons nous focaliser sur les partis présents dans les aspects législatifs et exécutifs du canton du Valais.

Tableau 2 Partis politiques présents au Grand Conseil valaisan en 2017 [14]

Partis	Nombre de sièges
Les libéraux-Radicaux (PLR)	26
Parti démocrate-chrétien (PDC)	55
Parti socialiste (PS)	13
Union démocratique du centre (UDC)	23
Parti chrétien social (PCS)	4
Parti écologiste Suisse PES	8
Autre	1

Tableau 3 Partis politiques représentés par au moins un élu au Conseil d'État valaisan entre 1981 et 2017 [14]

	1981	1985	1989	1993	1997	2001	2005	2009	2013	2017
PLR	1	1	1	1	1	1	1	1		1
PDC	4	4	4	4	3	3	3	3	3	3
PS					1	1	1	1	1	1
UDC									1	

Le Tableau 2 et le Tableau 3 présentent les partis politiques présents dans le canton du valais. Ces parties politiques disposent soit de sièges au Grand Conseil en 2017 ou d'un élu au Conseil d'État valaisan (ou les deux pour certains partis). À noter que nous n'avons ici pas listé l'ensemble des partis qui se sont présentés aux élections, mais uniquement ceux élus au moins une fois. Nous pouvons observer que les PDC et les PLR sont les deux partis majoritaires dans les organes législatifs et exécutifs du canton du Valais. Le Conseil d'État valaisan était uniquement composé d'élus de ces deux partis de 1981 à 1993. Le parti socialiste est le premier parti politique à réussir à partager le pouvoir exécutif avec le PLR et le PDC. C'est également ce parti qui fera élire la première femme au Conseil d'État valaisan en 2009 [14]. En 2013 le PLR n'est plus représenté au profit d'un candidat UDC.

1.3 OBJECTIFS DE RECHERCHE

Ce rapport vise à montrer les apports de la data science pour les sciences politiques. Nous illustrerons ces apports à travers un cas pratique : les élections au Conseil d'État valaisan. Nous souhaitons répondre aux questions de recherche suivantes :

> « Est-ce que le comportement électoral des citoyens valaisans est marqué par des préférences qui peuvent être liées à la langue des communes ? »

> « Quels autres facteurs explicatifs peuvent être mobilisés pour tenter de comprendre ces phénomènes ? »

> « Peut-on prédire le résultat d'une élection future sur la base d'un échantillonnage de la population ? Comment choisir cette part réduite de la population ? »

En répondant à ces questions de recherches, nous voulons montrer les apports des méthodes de datas science pour de l'analyse de situation politique et de la prédiction de résultats. Il s'agit de montrer que cette approche peut aider les sciences politiques en proposant une méthode pour la compréhension des facteurs influant lors d'une élection.

1.4 PÉRIMÈTRE ET LIMITE DE LA RECHERCHE

Cette recherche est focalisée sur les élections de personnes et non pas sur des votations au sujet de thème précis. Le domaine des votations est un sujet plus traité par la recherche alors que les élections de personne est un aspect de la vie démocratique moins traité en Suisse.

Le périmètre de cette recherche se limite aux élections du Conseil d'État valaisan de 2001 à 2017, soit cinq élections. Nous allons nous focaliser sur le cas du canton du valais, car il présente plusieurs caractéristiques intéressantes : il s'agit d'un canton bilingue avec une géographie variée qui va nous permettre d'analyser l'influence de la langue sur les comportements électoraux. De plus, le canton valaisan met à disposition les résultats des élections au Conseil d'État et cela dans le cadre de la stratégie Suisse d'open data.

Chapitre 2. REVUE DE LITTERATURE

Dans ce chapitre, nous allons nous intéresser aux méthodes et techniques de recherche utilisées en datas science pour les sciences politiques. Nous parlerons des concepts théoriques et nous étudierons les méthodologies utilisées. Nous voulons observer l'utilisation des compétences de data science dans le domaine des sciences politiques. Nous regarderons des projets de recherches qui ont déjà été faits pour la Suisse et dans d'autres pays, puis nous mettrons en lumières les différentes approches pertinentes. Enfin, nous étudierons les résultats de ces recherches ainsi que les applications proposées.

2.1 PARTITIONNEMENT DES DONNÉES ET ANALYSES

Le partitionnement des données est un outil, des datas sciences souvent utilisé et qui permet de résoudre efficacement certains problèmes. Celui-ci a notamment déjà fait ses preuves dans des domaines comme l'analyse de clientèles. Il est utile pour des entreprises d'être capable de segmenter leurs clientèles afin d'adapter leurs stratégies marketing à chacun de ces groupes. Les clients sont alors regroupés par type de comportement semblable de consommation du produit ou du service. D'autres regroupements ont déjà été utilisés, par exemple par « loyauté des clients » [15].

La segmentation de données a déjà été utilisée dans les sciences politiques et notamment en Suisse. En effet, Vincent Etter, Julien Herzen et al. ont publié un article titré « Mining Democracy » dans lequel ils présentent une façon d'utiliser le « data-mining » dans les sciences politiques [16]. Ils montrent qu'il est possible de découvrir des patterns intéressants dans les comportements de vote sans nécessiter de connaissance préalable du domaine [16]. Pour cela, ils utilisent une approche « data-driven » que l'on peut définir par une analyse dirigée par les données. Comme points de départ, ils utilisent les données de votations nationales des communes entre 1981 et 2011. Chaque commune est alors associée à un vecteur qui représente l'ensemble de ses votations, il s'agit d'un vecteur à 245 dimensions correspondant aux nombres de votation dans la période de temps étudiée. Afin de réduire ce nombre de dimensions à deux et ainsi pouvoir visualiser les communes sur un plan, ils utilisent une technique de réduction de dimension : Principal Component Analysis (PCA) [16]. Les communes peuvent ensuite être affichées sur deux axes qui capturent leur comportement de vote. Sur ce plan, une nette séparation apparait entre les communes francophones d'un côté et les alémaniques, italiennes et romanches de l'autre [16]. Cette séparation capture un comportement de vote différent duquel ils déterminent deux classes. À noter que la droite utilisée pour séparer le plan est posée arbitrairement après l'utilisation du PCA.

Ali T. Akarca, Cem Baslevent et al. proposent une autre approche pour partitionner les données [17]. En effet, plutôt que de projeter les données après PCA sur un plan et de chercher visuellement les séparations, ils utilisent un algorithme pour classifier leurs données. Leurs jeux de données sont les résultats d'élections de 81 provinces turques entre 1999 et 2009. Afin de créer des groupes de communes proches dans leur manière de voter, ils proposent d'utiliser un algorithme non supervisé sur des données non annotées afin de classifier les 81 provinces [17]. Ils soulignent qu'il est important que les données ne soient pas annotées – uniquement les résultats des votes – afin de ne pas introduire de biais dans la création des groupes [17]. Deux algorithmes sont présentés par les chercheurs : Hierarchical clustering et K-mean. En se basant sur une étude de Jefferson West [18], ils choisissent d'utiliser l'algorithme K-mean, car celui-ci est plus adapté au domaine des sciences politiques [17]. Leur but est de capturer les principales divisions politiques du pays, c'est-à-dire de créer des groupes – appelé clusters – qui correspondent à une façon de voter.

K-mean est un algorithme itératif qui prend en paramètre une valeur numérique entière appelée k. Cette valeur correspond au nombre de groupe dans lequel on souhaite séparer nos données. Une fois démarré avec un paramètre k, l'algorithme commence par une étape d'initialisation dans laquelle il place aléatoirement k centres appelés « centroïde ». Ces centroïdes sont placés dans le même espace mathématique que les données sur lequel

est appliqué l'algorithme[1]. Après cette initialisation, l'algorithme K-mean rentre dans sa phase itérative composée de deux étapes.

1. Il assigne chaque point du jeu de données au centroïde le plus proche[2]
2. Chaque centroïde calcule la moyenne des distances euclidiennes de chaque point associées au point 1. Cette moyenne devient le nouveau centroïde.

Lors de la fin du point 2, il recommence au point 1. La condition de sortie de l'algorithme est que tous les centroïdes soient égaux à la moyenne des distances euclidiennes de tous leurs points associés. Les outputs de l'algorithme K-mean sont la liste des centroïdes – centre des clusters – et l'attribution de chacun des points du jeu de données à un cluster.

Il est nécessaire de préciser à l'algorithme K-mean le nombre k de clusters à chercher. Pour trouver le nombre k optimal – la valeur de k qui capture l'entier des différences de comportement sans créer des groupes trop proches - Ali T. Akarca et Cem Baslevent appliquent une méthode empirique. Ils essaient successivement plusieurs valeurs pour k afin qu'un pattern apparaisse [17]. Ils existent d'autres manières de procéder afin de déterminer une valeur optimale pour k. On peut notamment citer une approche appelée « Elbow Method » [19]. L'idée est de tester l'algorithme K-mean avec plusieurs valeurs pour k. À chaque itération, il s'agit de mesurer la dispersion intra cluster. Cette dispersion correspond à l'éloignement d'éléments à l'intérieur du cluster : plus ces éléments sont proches, plus cette dispersion est faible. Ces mesures intra cluster sont alors sommées pour donner une mesure appelée « total within-cluster sum of square (WSS)" [19]. Pour déterminer le k le plus pertinent, il faut afficher un graphique avec les valeurs de k testées en abscisses et les WSS en ordonnées. Ensuite, il s'agit d'observer le segment passant par l'ensemble des points du graphe afin de trouver un « coude ». Cette cassure sur ce segment indique qu'il s'agit d'un k avec une valeur pertinente, car elle apporte un gain significatif sur le WSS par rapport à $k-1$ et que $k+1$ apporte un gain négligeable.

Les deux études effectuent des interprétations spatiales des groupes créés, mais leurs méthodes diffèrent sensiblement. Pour l'étude réalisée en Suisse, les auteurs distribuent une palette de couleurs sur le plan utilisé pour afficher les valeurs du PCA afin d'attribuer une nuance de couleur à chaque commune. Ils affichent ensuite ces communes coloriées sur la carte de la Suisse. Ceci leur permet de mettre en évidence un changement clair de ton des couleurs à l'emplacement géographique où la langue des cantons suisse change [16]. Pour le cas de l'étude des provinces turques, les chercheurs attribuent une couleur aux clusters et non pas une nuance aux provinces : toutes les provinces du même cluster ont la même couleur. Ils affichent également les provinces avec leurs couleurs sur la carte de la Turquie. Ces deux méthodes permettent des interprétations géographiques des regroupements de communes par type de comportement de vote. L'approche réalisée en Suisse a pour avantage que toutes les communes peuvent être comparées entre elles, mais elle ne délimite pas de frontière claire entre les groupes.

Afin de tenter d'expliquer ces comportements similaires, Ali T. Akarca et Cem Baslevent annotent chacune des provinces et leurs clusters avec des informations comme le nombre d'habitants par kilomètre carré, le pourcentage de population urbaine, l'âge moyen, la durée de la scolarité et le taux d'immigration. Ils ne précisent pas quelle méthode est utilisée pour sélectionner ces indicateurs. Ils indiquent que ceux-ci sont des valeurs sociales, économiques et démographiques pertinentes [17]. Sur la base de ces valeurs et des différences entre les clusters, une analyse politique est faite sur le « pourquoi » des similarités de comportement de vote des provinces.

[1] Prenons comme exemple un jeu de données correspondant au résultat d'une élection entre cinq candidats. Alors chaque commune disposera d'un vecteur projeté dans un espace à cinq dimensions correspondant aux scores attribués à chaque candidat. La pointe du vecteur correspond à un point dans cette espace et c'est dans le même espace à cinq dimensions que sont placés aléatoirement les centroïdes.

[2] Le plus proche selon une distance euclidienne généralisée à une dimension k.

Les méthodes de partitionnement des données provenant des datas sciences permettent l'analyse de situation politique avec une approche innovante et sans nécessité de connaissances poussées du domaine. Elles permettent de vérifier certaines hypothèses comme l'influence de facteur sur le comportement de vote de citoyens. Une représentation géographique de groupe de personnes partageant un comportement proche ainsi que des clusters annotés constituent des informations précieuses pour de l'analyse politique. Ces approches centrées sur les données sont un bon moyen d'éviter d'insérer des biais de confirmation et de créer de l'information utile à de l'analyse politique.

2.2 ANALYSE PRÉDICTIVE

L'analyse prédictive consiste à extraire de l'information depuis des données brutes afin de trouver des régularités et ainsi être capable de prédire un évènement futur. Cette méthode permet la mise en place de système d'aide à la décision dans divers domaines. Par exemple, prédire ce qu'un client est susceptible d'acheter en fonction de son historique d'achat et de celui d'autres clients. Ou encore, dimensionner une infrastructure informatique en fonction de l'historique de trafics. De manière générale, on cherche à prédire un résultat sur la base d'expériences passées ou sur des caractéristiques particulières.

Dans le domaine des sciences politiques, plusieurs recherches appliquent des méthodes d'analyses prédictives et cela principalement pour chercher à déterminer le résultat de vote ou d'élections. Andranik Tumasjan et al. présentent dans un article une approche pour prédire une élection à partir de message sur le réseau social Twitter [20]. Ils utilisent plus de 100'000 Tweets[3] postés avant l'élection fédérale du parlement allemand en 2009, dont chacun de ces Tweets faisaient référence à l'élection ou à un parti politique. Avec l'ensemble des messages étudiés, ils ont mesuré une forte corrélation entre la quantité de messages qui mentionnaient un parti et son futur résultat à l'élection [20]. Les auteurs ne proposent pas directement de système de prédiction, mais une approche pour utiliser ces messages comme indicateurs pertinents pour les acteurs et observateurs politiques. Toujours dans le domaine d'étude des messages postés sur Twitter, Marko Skoric et al. ont étudié l'élection générale de Singapore de 2011 [21]. Ils ont cherché à tester la capacité de prédictions des Tweets sur le résultat de l'élection. L'ensemble des Tweets ont été collectés durant la campagne et leurs analyses à montrer la même corrélation que l'étude allemande[4].

Les deux études basées sur des données issues des réseaux sociaux présentent des applications intéressantes, mais aucune des deux ne propose un système de prédictions. D'autres études proposent des modèles de prédiction plus aboutie et les mettent à l'épreuve avec des méthodes de contrôles. C'est notamment le cas d'une étude sur les élections présidentielles américaines menée par J. Scott Armstrong et Andreas Graefe [22]. Dans leur article, les deux auteurs décrivent un modèle de prédiction des élections présidentielles américaines. Ce modèle attribue 59 variables biographiques aux candidats et chacune de ces variables ajuste un score global par candidat. C'est le candidat avec le plus haut score qui est prédit comme gagnant. Ils testent ce modèle sur les élections de 1896 à 2008 soit 29 occurrences. Les prédictions sont correctes pour 27 élections sur 29 [22]. Les auteurs déclarent fournir un modèle plus performant que les sondages classiques et indiquent des applications pour les partis politiques lors de leurs choix de candidats [22]. Une partie de l'étude Suisse sur les votations fédérales parle également d'une approche centrée sur les données pour prédire le résultat d'une votation [16]. Pour rappel, ils disposent d'un jeu de données composé de 245 votes pour l'ensemble des communes suisses sur des sujets divers. Ils commencent par séparer leurs données en deux : une première partie d'entrainement composée de 196 résultats et une seconde de tests composée des 49 autres. Ensuite, ils entraînent un algorithme de type « Decision Tree » pour chaque commune avec les 196 votes des données d'entrainements. Les chercheurs testent les classificateurs avec les données de contrôles – les 49 votes restants – pour mesurer leur taux de prédiction.

[3] Message de 140 caractères maximum postés sur le réseau social Twitter

[4] La corrélation est de même nature, mais plus faible dans l'étude de Singapore [21]

Les conclusions présentées sont assez surprenantes : environ 10% des communes prédisent à 90% les données de validations, comme par exemple, la commune de Ebikon qui prédit les données de validations à 97.5%[5] (seulement deux échecs sur les 49 testés) [16]. Pour expliquer les bonnes capacités de prédiction de la commune Ebikon, les auteurs expliquent que celle-ci est située au cœur de la Suisse et représente bien sa diversité. Enfin, ils présentent leur modèle comme utile pour améliorer la qualité des sondages en ciblant des échantillons représentatifs à l'intérieur de ces communes disposant d'une capacité de prédiction.

En conclusion, l'analyse prédictive est une approche utile pour des systèmes d'aide à la décision. Les exemples présentés ci-dessus montrent que cette méthode a des applications pour les sciences politiques. Il est souvent question de prédiction de résultat de vote ou d'élection. Les méthodologies utilisées varient de simples recherches de corrélation à des modèles plus complexes comme des calculs de scores pour candidat pour les présidentielles Américaines ou l'entrainement d'algorithme par des données d'exemples. Il est nécessaire d'adapter les outils utilisés à sa problématique et à la nature des données à disposition. Les modèles de prédictions présentés dans l'étude Suisse sur les votations des communes et celui de l'étude sur les présidentielles Américaines peuvent être utilisés comme de l'aide à la décision pour des acteurs politiques comme les partis ou les instituts de sondages.

2.3 OPEN DATA

Toutes les recherches présentées dans les chapitres précédents, utilisent des techniques d'analyse qui demandent de disposer de données. La taille et la qualité des données sont des facteurs impactant les résultats possibles. Pour des domaines comme la segmentation de clientèle, les entreprises concernées génèrent elles-mêmes les données à travers par exemple des cartes de fidélités. Dans le domaine des sciences politiques, les stratégies d'ouverture des données des états jouent un rôle central [16]. La mise à disposition des données est ce qui permet d'explorer ces nouvelles approches en combinant des techniques de data sciences avec de l'analyse politique. Les recherches utilisant des données libres sont plus transparentes et il est plus facile de le reproduire [23]. En 2006, 50.8% des chercheurs aux États-Unis rapportent que leurs recherches sont impactées négativement par un refus de mettre à dispositions des données [23]. Ils indiquent également perdre du temps à comprendre les conditions d'utilisations variables et compliquées des données [23].

Pour établir une culture de l'innovation dans des domaines inattendus, le gouvernement Suisse dispose de données utiles, de qualités et ils proposent une stratégie de mise à disposition de données officielles [1]. Ils encouragent une coopération de tous les secteurs publics à l'échelle fédérale, cantonale et communale [1]. Pour faciliter l'accessibilité aux données, ils mettent en place plusieurs points : une centralisation des données, des métadata pour décrire les jeux de données, un inventaire des données disponibles et des conditions d'utilisation uniformes et compréhensibles [1].

Les données sont souvent la base de la recherche scientifique [23]. La mise à disposition de données en open data impacte positivement la recherche en facilitant l'accès, la compréhension et l'utilisation des données. La société en générale en profite aussi, car les institutions publiques sont poussées à plus de transparence [23].

Enfin, l'open data est indispensable pour certaines applications des méthodes de data science dans le domaine des sciences politiques. En effet, les approches pilotées par les données sont, par nature, dépendantes de données publiques comme le résultat de votations et d'élections. Les informations sociales, démographiques et les données géographiques sont un prérequis pour analyser les résultats des approches par partitionnement.

[5] Ils introduisent le terme de « commune Oracle » pour parler de ces communes capable de prédire au niveau local des résultats de vote nationaux [16]

Chapitre 3. MÉTHODOLOGIE

Ce chapitre est dédié à la méthodologie utilisée pour notre cas pratique : l'élection du Conseil d'État valaisan. Nous décrirons l'ensemble des étapes et méthodes utilisées pour répondre à nos questions de recherches. Celles-ci concernent l'impact des langues, les facteurs influents et la capacité de prédictions sur des données existantes.

Notre méthodologie est séparée en deux parties : la première concerne l'analyse des comportements électorale alors que la deuxième se concentre sur le modèle de prédiction.

Dans la première partie, nous parlerons des jeux de données, de la récolte et les prétraitements appliqués. Ensuite, nous décrirons l'algorithme de partitionnement utilisé ainsi que les prérequis appliqués aux données. Enfin, il s'agira de décrire les annotations post-classement et les méthodes de visualisation des données utiles pour l'analyse.

La seconde partie est consacrée à la prédiction de résultat de l'élection au Conseil d'État. Nous commencerons par décrire les traitements appliqués aux données. Ensuite, nous étudierons comment déterminer un résultat d'une élection et nous décrirons notre approche de classement des communes par « capacité de prédiction ». Enfin, nous décrirons notre méthode de contrôle pour notre modèle.

3.1 ANALYSE DES COMPORTEMENTS ÉLECTORAUX

Nous parlerons dans cette partie du document de la méthodologie pour la mise en place de l'analyse des comportements électoraux. Nous regarderons les différentes sources de données utilisées et les prétraitements nécessaires. Ensuite, nous détaillerons le partitionnement des données et leurs annotations par d'autres sources de données. Enfin, nous regarderons les méthodes utilisées pour la présentation et la visualisation des résultats.

3.1.1 Sources de données et prétraitements

Pour notre cas d'étude, nous commençons par la récolte des données. Tous les jeux de données utilisés dans notre étude proviennent uniquement de source officielle du secteur public Suisse. Nous utilisons des données mises à disposition par des communes, des cantons ou par la Confédération qui respectent les termes d'usages des open data dans le cadre de la stratégie de la Confédération.

Notre jeu de données principal est les résultats des élections pour le Conseil d'État valaisan pour les années 2001, 2005, 2009, 2013 et 2017. Les données sont mises à disposition par l'Etat du valais à travers leur plateforme web et correspondent aux nombres de voix récoltées de chaque candidat dans l'ensemble des communes valaisannes. Ensuite, nous utilisons d'autres jeux de données afin d'annoter chaque commune avec différentes caractéristiques. Pour commencer, nous attribuons une langue majoritaire à chaque commune. En plus de cette information, nous ajoutons d'autres informations socio-économiques comme la pyramide des âges et la répartition en secteur de l'activité économique. Enfin, nous annotons les communes avec leurs densités de population. Le Tableau 4 liste les jeux de données utilisés pour notre étude de cas.

Tableau 4 Jeux de données utilisées

Jeux de données
Résultats des votations du Conseil D'État valaisan pour 2001, 2003, 2005, 2009 et 2017
Langue majoritaire des communes suisses
Répartition des âges des communes Suisse
Répartition du secteur économique Suisse – secteur primaire, secondaire et tertiaire
Données GIS des cartes du valais

Maintenant que nous avons précisé la récolte des données, nous devons préparer les différents jeux de données afin qu'ils puissent être exploitables par la suite. Pour cela, nous appliquons des prétraitements aux données. Les résultats des élections doivent être normalisés afin d'être utilisés par des méthodes de partitionnement. Le nombre

de bulletins de vote n'étant pas égal dans chaque commune, les valeurs de nombre de votes reçus par les candidats ne sont pas directement comparables entre communes. Pour créer des vecteurs – représentant la répartition des bulletins entre les candidats – comparables entre les communes, nous devons normaliser cette répartition. Pour cela, nous choisissons de ramener cette valeur numérique au pourcentage de bulletins attribués par candidat.

Tableau 5 Normalisations appliquées aux données d'élections du Conseil d'État valaisan

Avant normalisation :

Nombre de bulletin[6]
2826

Candidat 1	Candidat 2	….	Candidat N
840	712	…	…

Après normalisation :

Candidat 1	Candidat 2	….	Candidat N
0.297239915	0.251946214	..	…

Les autres jeux de données utilisés pour les annotations ne demandent pas de normalisation. Les traitements qui doivent être effectués sur ces jeux sont plus de l'ordre du nettoyage de données. Les noms des communes peuvent varier entre les différentes sources, car celles-ci peuvent être écrites dans différentes langues ou avec des précisions[7]. Nous devons uniformiser ces informations dans l'ensemble de nos jeux de données afin de pouvoir utiliser les noms de communes comme des clés uniques lors de nos traitements.

Le dernier prétraitement de données est la gestion des fusions des communes. En effet, notre jeu de données des élections contient des valeurs pour les années 2001 à 2017 et durant ces années, plusieurs fusions de communes ont eu lieu dans le canton du Valais. L'étude réalisée sur les votations populaires des communes suisses par Vincent Etter, Julien Herzen et al. a traité cette problématique en supprimant du jeu de données les communes résultantes des fusions ainsi que celles fusionnées [16]. Nous ne suivons pas la même méthode, car notre jeu de données est plus petit, nous traitons les communes valaisannes – ils traitent l'ensemble des communes Suisses – et les données de cinq élections – ils traitent 245 votations populaires. Nous ne voulons pas réduire davantage notre jeu de données et pour cela nous allons appliquer une méthode pour gérer ces fusions. L'idée de notre approche, est de partir de l'état des communes valaisannes en 2017 et de reproduire cet état dans le passé. Pour cela, nous fusionnerons les communes dans le passé afin de disposer des mêmes communes pour toutes les années. Ces fusions à reproduire pour l'ensemble des années étudiées impactent l'ensemble de nos jeux de données. Nous devons sommer les résultats, le nombre de bulletins et le nombre de participants des communes fusionnantes – avant la normalisation – et créer une nouvelle entrée pour la commune fusionnée. Cette opération est à faire pour toutes les années. Les données utilisées pour les annotations de communes doivent être regroupées par communes fusionnantes afin de créer les communes fusionnées. En fonction de la nature des données, nous utilisons des moyennes ou des sommes. Le Tableau 6 présente les types de traitements à effectuer sur les jeux de données afin de fusionner les communes pour les années antérieures à 2017.

[6] Chaque bulletin peut distribuer entre 1 et 5 voix

[7] Lorsque des communes portent le même nom dans deux cantons alors celui-ci est précisé : (VS), (VD), etc.

Tableau 6 Traitements à effectuer sur les jeux de données pour la gestion des fusions de communes

Jeux de données	Traitement pour les fusions
Résultats des votations du Conseil d'État valaisan pour 2001, 2003, 2005, 2009 et 2017	Sommes des résultats, bulletins et participations
Répartition des langues pour les communes Suisse	Aucun traitement
Répartition des âges des communes Suisse	Moyenne des communes fusionnées
Répartition du secteur économique Suisse – secteur primaire, secondaire et tertiaire	Moyenne des communes fusionnées

Nous disposons de données pour chacune des années d'élections, nous pourrons comparer les résultats entre les années. Mais, nous ne disposons pas de jeux de données qui résume l'entier des votes des communes pour les cinq années. Nous devons réunir l'ensemble de nos jeux de données afin de créer ce super espace qui résume l'entier des votes des communes. Ainsi, nous pourrons également appliquer nos traitements sur ce nouvel espace. Les résultats de partitionnement obtenu sur ce super espace représenterons tous les comportements de votes entre 2001 et 2017.

Afin de créer ce super espace, nous devons réunir nos jeux de données. Il s'agit de créer une matrice $Communes\ X\ Candidats$. Les lignes des communes ne sont pas dupliquées, le super espace contient une ligne par commune. Pour les colonnes candidats, nous gardons chacune des candidatures. Par exemple, un candidat qui c'est présenter en 2001 et 2005 sera présent dans deux colonnes du super espace : la première « candidat_A (2001 » et « candidat_A (2005) ». Avec cette structure, nous disposons de l'entiers des votes réalisés entre 2001 et 2005 pour commune.

3.1.2 Partitionnement des données

Nos données sont prêtes à être utilisées par des méthodes de partitionnement. Nous voulons regrouper les communes valaisannes par type de comportement de vote similaire. Cette séparation en groupes – appelés clusters – est faite à l'aide d'un algorithme. Nous utilisons l'algorithme « K-mean » déjà utilisé dans des recherches sur les comportements électoraux en Turquie par Jefferson West en 2005 [18] et en 2011 par Ali T. Akarca et Cem Baslevent [17]. Nous avons vu dans notre revue de littérature que l'algorithme « K-mean » était préféré aux « Hierarchical clustering » pour le domaine de l'analyse de situation politique.

Nous appliquons cet algorithme aux cinq élections du Conseil d'État valaisan comprisses entre 2001 et 2017 ainsi qu'à notre super-espace représentant les cinq élections.

Dans notre revue de littérature, nous avons indiqué qu'il était nécessaire de configurer l'algorithme K-mean avec une valeur pour la variable k et que celle-ci détermine le nombre de clusters final. Ce nombre de clusters correspond au nombre de types de comportements différents détectés dans le jeu de données, il doit donc être choisi avec précaution. Nous utilisons la méthode « Elbow method » présentée dans notre revue de littérature. Nous commençons par calculer les scores WSS pour un k allant de 1 à 9. Nous représentons les valeurs WSS sur un plan, l'abscisse correspondant aux valeurs de k et l'ordonnée à celui des scores WSS.

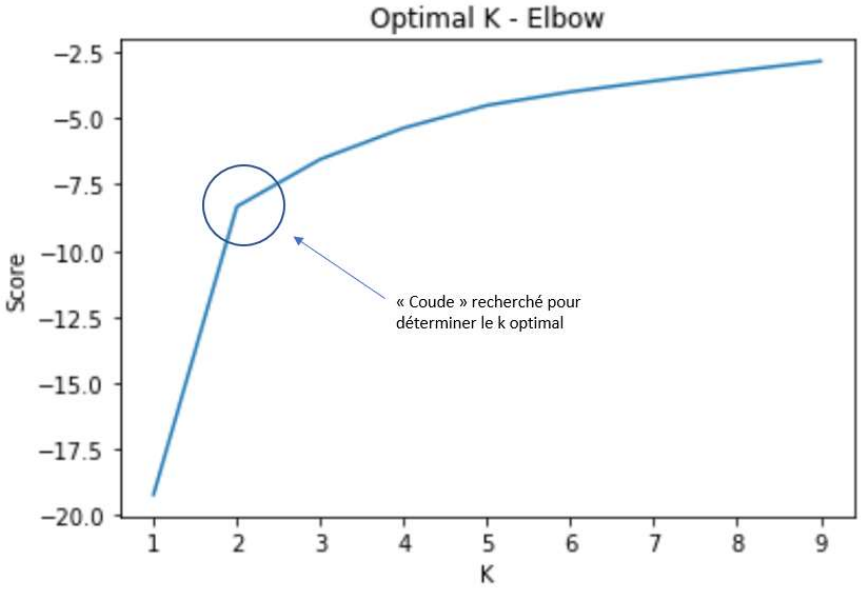


Figure 2 Exemples de forme recherchée pour déterminer une valeur optimale de k

Pour chaque année d'élection ainsi que pour le super-espace, nous étudierons un graphique du même type que la Figure 2 afin de trouver la ou les valeurs pertinentes pour la variable k. Une fois k trouvé, nous exécutons l'algorithme K-mean sur nos jeux de données afin de créer les clusters.

3.1.3 Annotations des données

Maintenant que nos données sont réparties dans des groupes sur la base des comportements de vote des communes, nous passons à la phase d'annotations des données. L'objectif est de tenter d'expliquer les comportements similaires avec des facteurs externes à l'algorithme de partitionnement. L'ensemble des annotations est répété pour les cinq années d'élections ainsi que pour le super-espace.

Pour commencer, nous indiquons la langue majoritaire pour toutes les communes classées dans nos clusters en utilisant une seconde source de données : la liste des langues parlées majoritairement pour les communes suisses. Nous calculons ensuite le pourcentage que représentent les langues officielles Suisses à l'intérieur de chacun de nos clusters. Avec ces pourcentages, nous pouvons déterminer la langue majoritaire de chaque cluster et les communes mal classées – qui ne sont pas de mêmes langues que celle majoritaire dans leur cluster – seront listées. En résumé, nous disposons d'un ensemble de clusters avec leur langue majoritaire et une liste de communes mal placées.

Ensuite, nous ajoutons à chaque commune les informations sociales et démographiques suivantes : la pyramide des âges, la répartition en secteurs d'activités économiques et la densité de population. Ces valeurs sont également prises depuis nos sources de données décrites plus haut dans le document. Avec ces nouvelles annotations, nous calculons la moyenne pour les âges et le secteur économique et la médiane pour la densité afin de pouvoir annoter les clusters avec ces informations. La médiane est utilisée pour la densité, car les communes valaisannes présentent une forte dispersion ainsi que des valeurs extrêmes [8]. Nous pourrons également analyser nos clusters

[8] Valeur min. à 8 et max. à 363 hab/km2

à travers ces nouvelles annotations et déterminer les communes mal classées du point de vue de la densité, de la pyramide des âges ou de la distribution des secteurs économiques.

Finalement, la dernière information ajoutée à nos données d'élections est le parti des candidats. Ces informations sont manuellement ajoutées à chaque candidat. Les sources de données varient en fonction des personnes. L'information est parfois récupérée sur la page Wikipédia du candidat et parfois sur son site internet personnel. Nous ne constatons pas de candidat qui se serait présenté à deux élections avec deux partis différents. Avec les informations du parti attachées aux candidats, nous pouvons calculer et ajouter la répartition des voix par partis pour les communes et les clusters.

3.1.4 Visualisation des données

Notre dernière étape pour l'analyse des comportements électoraux est la visualisation des données. Les jeux de données ont été classés, croisés entre eux et annotés. Nous pouvons mettre en place les visualisations des données afin de rendre plus facile l'analyse des résultats.

Pour rappel, voici la liste des données à notre disposition après les étapes décrites ci-dessus :

- Les communes avec la répartition du nombre voix par candidat en pour cent de bulletins total :
    - La langue majoritaire de la commune
    - La densité de population
    - La répartition des âges
    - La répartition en secteur d'activité
    - La répartition des voix par partis
- Des clusters de commune avec des comportements de votes similaires
    - La langue nationale majoritaire
    - La répartition en pourcentage des langues nationales
    - La moyenne de la répartition des âges
    - La moyenne de la répartition en secteur d'activité
    - La médiane de la densité de population
    - La somme des répartitions des voix par partis

Nous allons commencer par parler des visualisations des données des communes. Les premières données que nous souhaitons visualiser sont la répartition des voix par candidats. Il s'agit de pouvoir représenter le vote des communes sur un plan à deux dimensions. Le nombre de candidats des élections varie entre 7 et 13. Afin de visualiser ces vecteurs de 7 à 13 dimensions sur deux axes, nous utilisons la méthode de réductions de dimension discutée dans notre revue de littérature : le principale component analysis (PCA). Celle-ci va nous permettre d'afficher le vecteur de votes des communes dans un plan. Nous affichons toutes les communes sur un graphique de nuage de points, chaque point représente le vote de la commune ramené à deux dimensions.

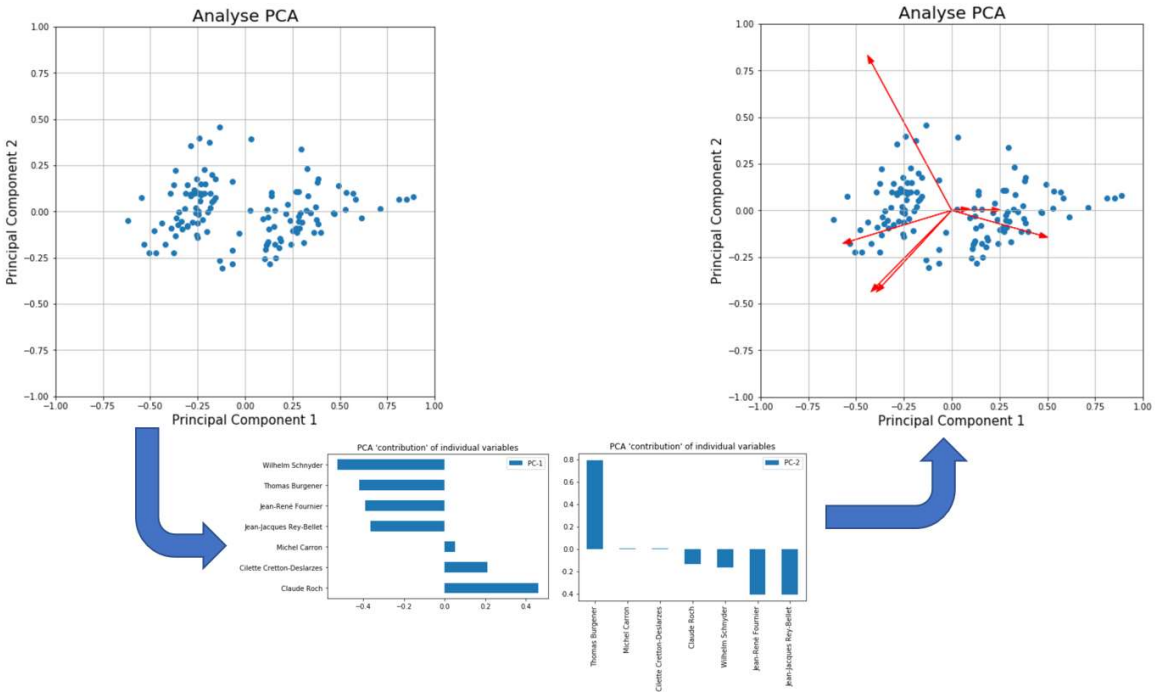


Figure 3 Étapes pour la création d'un graphique créé par PCA et complété avec les contributions des variables

La Figure 3 représente la création d'un graphique pour l'analyse des comportements de votes des communes. Sur cet exemple, il s'agit d'une élection à 7 candidats et chaque point correspond au vote d'une commune pour ces 7 candidats ramenés sur deux dimensions. Afin de donner une signification aux axes « principal component 1 » (PC1) et « principal component 2 » (PC2), nous affichons la contribution de chacune des 7 variables dans la création de PC1 et PC2. Ces contributions sont les deux graphiques situés en bas de la Figure 3. Enfin, nous calculons un vecteur de contribution pour les 7 variables que nous affichons sur le graphique de base. Le résultat peut être observé dans la partie droite de la Figure 3. Chaque vecteur peut alors être annoté avec le nom du candidat ainsi que son parti politique. Avec cette représentation graphique, nous pouvons comparer les comportements des communes les unes avec les autres et observer la contribution des partis politiques et des candidats dans la distribution des points.

Nous avons décrit la méthode utilisée pour visualiser les comportements de vote des communes et nous allons maintenant parler des méthodes de visualisation pour les clusters. Pour commencer, nous allons parler des visualisations de la répartition de la langue. La première visualisation est la répartition en pourcentage des langues présentes dans le cluster.

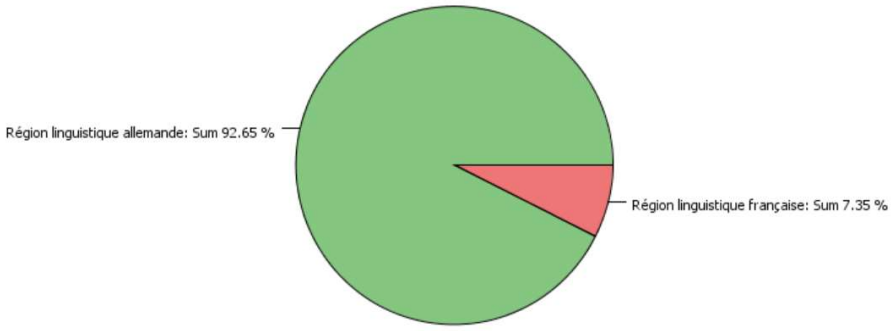


Figure 4 Visualisations de la répartition des régions linguistique à l'intérieur d'un cluster

La Figure 4 présente le type de graphique utilisé afin de visualiser la répartition intra cluster des régions linguistiques. Pour rappel, les communes qui ne parlent pas la même langue que celle parlée par la majorité dans leur cluster sont notées sur une liste afin de les identifier. Toujours pour visualiser cette répartition des langues dans les clusters, nous utilisons un deuxième type de visualisation. Il s'agit d'un graphique similaire à celui réalisé avec le PCA, mais sans les vecteurs de contribution et annoté avec les informations de cluster et de langues parlées pour les communes.

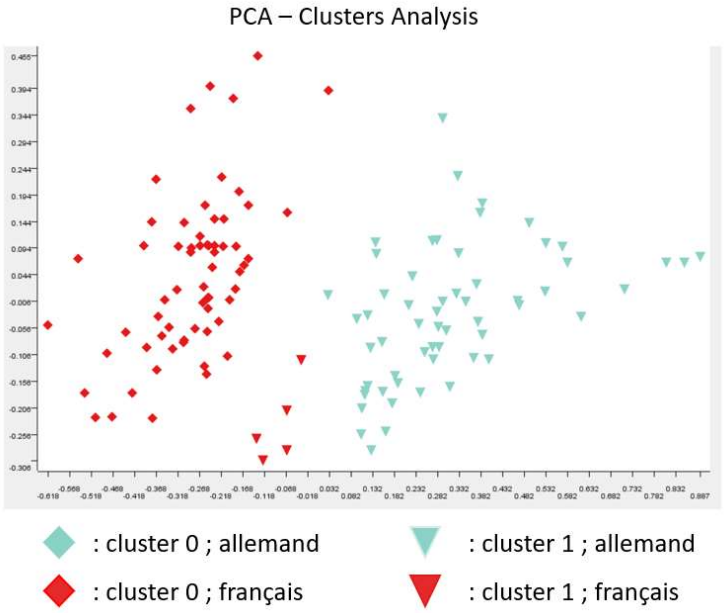


Figure 5 Visualisations des votes par PCA avec annotations de la répartition des langues et des clusters

La Figure 5 est aussi une visualisation des comportements de votes des communes par PCA. Les différentes formes représentent les clusters alors que les couleurs indiquent la langue majoritaire de la commune. Avec ce graphique, nous pouvons visualiser l'emplacement sur le PCA des communes mal classées et il permet également d'observer la dispersion intra cluster et les distances inter cluster. Il s'agit de la répartition idéologique des clusters.

La seconde représentation visuelle des clusters est l'affichage de chaque commune avec une couleur associés à son cluster sur la carte du Valais.

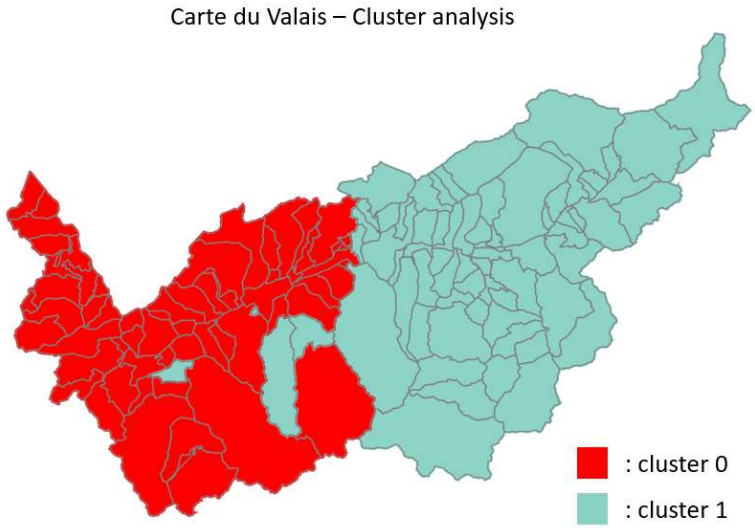


Figure 6 Représentation des communes valaisannes avec une couleur associée à leur cluster

La Figure 6 nous permet d'observer la répartition géographique des clusters obtenus avec l'algorithme K-mean. Cette visualisation met en évidence les communes « mal classées » d'un point de vue géographique. À noter que quand une commune est mal classée d'un point de vue de la langue alors elle l'est aussi géographiquement, mais cela est propre à la situation du Valais. La carte que nous utilisons est générée depuis les données officielles des frontières suisses, il s'agit de la carte du Valais en 2017. Nous utilisons la carte de 2017, car nous avons appliqué les fusions des communes pour les années antérieures et cela nous permet d'utiliser la même carte pour l'ensemble de nos années d'élections ainsi que pour le super-espace.

Le dernier type de visualisation pour les clusters est dédié à l'analyse des facteurs sociaux économiques que nous avons à annoter sur les communes et les clusters. Premièrement, nous voulons afficher la pyramide des âges par clusters. Pour cela, nous utilisons un graphique en barres.

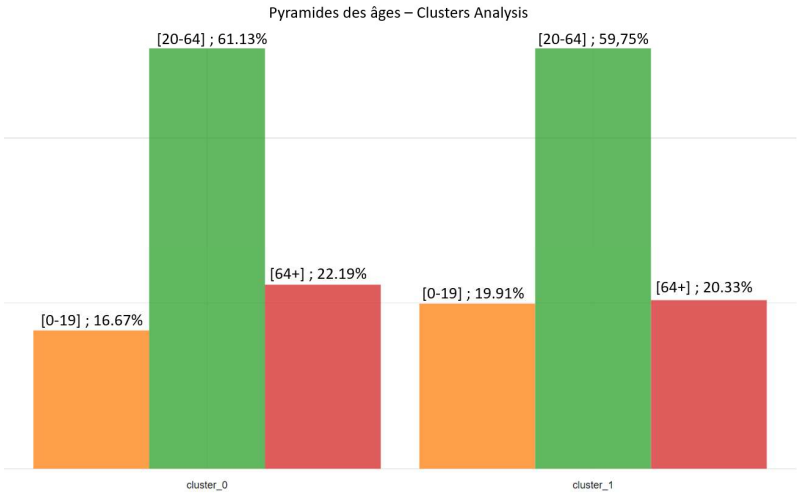


Figure 7 Méthodes de visualisation de la pyramide des âges des clusters

La Figure 7 présente la visualisation de la pyramide des âges. Les âges sont regroupés en trois classes : de 0 à 19 ans, de 20 à 64 ans et 64 ans et plus. Le graphique affiche la répartition des âges dans les communes du cluster. Nous pouvons ainsi comparer les clusters entre eux d'un point de vue des âges. Cette comparaison est faite uniquement à l'échelle des clusters et non pas commune par commune.

Pour la comparaison de la densité de population, nous utilisons la même approche. Il s'agit d'un graphique en barres avec les clusters sur l'axe des abscisses.

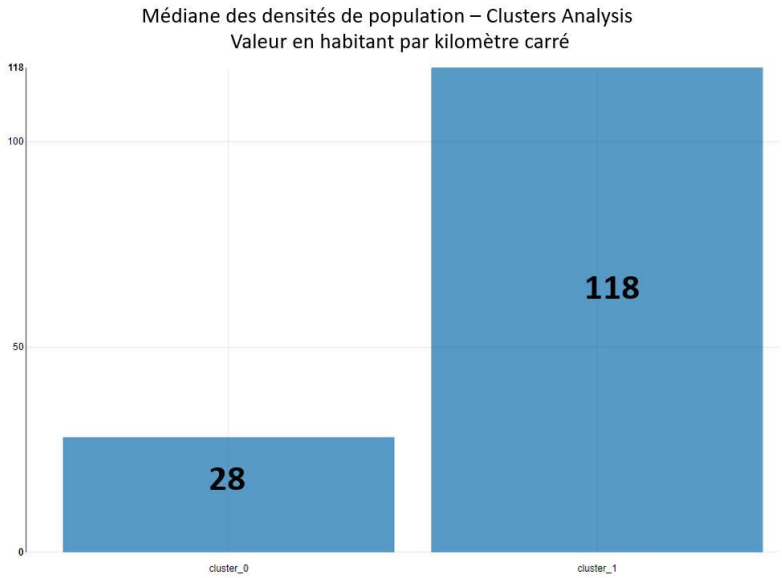


Figure 8 Méthodes pour comparer les densités de population entre les clusters

Comme le montre la Figure 8, chaque cluster dispose d'une barre qui représente la médiane des densités de population des communes du cluster. Cette densité est mesurée en habitant par kilomètre carré.

Enfin, nous voulons visualiser de la même manière la répartition en secteur d'activité des communes à l'intérieur des clusters. Pour rappel, il s'agit d'afficher le pourcentage que représente le secteur primaire, secondaire et tertiaire pour les communes dans les clusters.

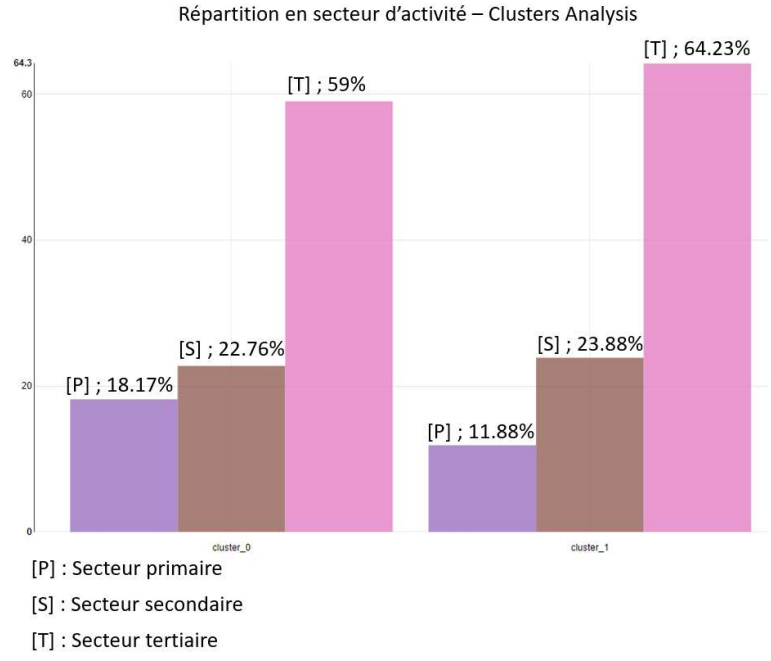


Figure 9 Méthode pour comparer la répartition en secteur d'activité des clusters

La Figure 9 reprend les mêmes principes que les deux précédentes. Nous cherchons toujours à visualiser les différences entre les clusters. La répartition en secteur d'activité est exprimée en pourcentage sur les moyennes des communes.

Nous allons faire une répartition de nos données en clusters pour chaque année. Afin de visualiser la qualité de nos regroupements et de comparer les années, nous allons représenter sur un graphique la dispersion inter et intra clusters. La distance intra cluster représente l'écartement moyen des points à l'intérieur des clusters alors que la distance inter cluster est l'écartement moyen des centres de cluster. Plus la distance intra cluster est faible plus les éléments du cluster sont proches les uns des autres. Une distance inter cluster faible indique des clusters proches entre eux alors qu'une valeur haute indique des clusters bien séparés

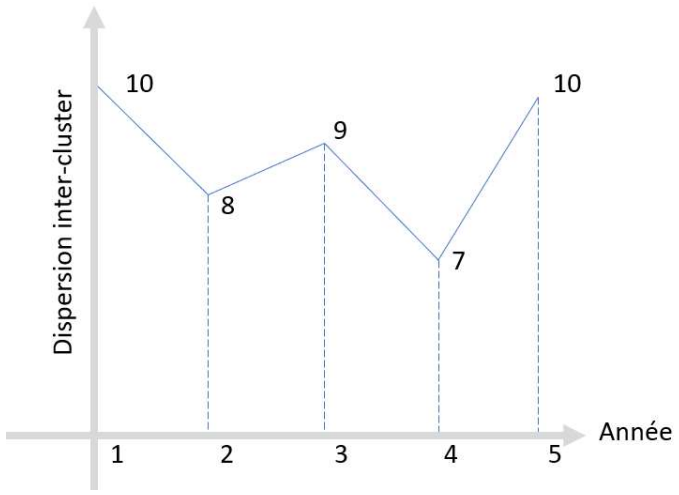


Figure 10 Représentations des dispersions intra et inter clusters pour les cinq années d'élections

La Figure 10 indiques la visualisation utilisée pour afficher l'évolution dans le temps des dispersions inter et intra clusters.

Avec ce dernier exemple, nous avons vu l'ensemble des méthodes de visualisations utilisées pour l'analyse des comportements électoraux.

3.2 ANALYSE PRÉDICTIVE

Dans cette partie du document, nous allons parler de notre méthodologie pour la création d'un modèle de prédiction des résultats de l'élection du Conseil d'État valaisan. Nous allons décrire les étapes pour tenter d'appliquer le modèle de commun oracle présenté par l'étude réalisée à EPFL [16]. Nous voulons regarder s'il est possible d'appliquer cette approche et quelles sont les modifications que l'on doit apporter pour adapter la méthode à notre cas d'étude et à nos données.

Pour notre analyse prédictive, nous utilisons comme principale source de données les résultats des élections par communes pour les années 2001 à 2011. Pour chacune des cinq élections, nous créons un vecteur représentant les candidats élus. Pour rappel, l'élection au Conseil d'État valaisan élit cinq personnes lors des votations, nous avons alors des vecteurs listant les 5 candidats de manière ordrée : du candidat arrivé en tête à celui en cinquième position. Ces données nous serviront de résultat pour les élections.

3.2.1 Approche par arbre de décisions

Pour la détection de commune oracle, l'étude sur les votations populaires suisses utilise un arbre de décision qu'il entraine avec 80% de leur jeu de donnée [16]. Les données d'entrée sont le sujet des votations et pourcentage de « oui » de la commune ainsi que le résultat final à l'échelle nationale. Il existe un arbre par commune.

L'arbre de décision est alors entrainé avec la structure de données suivantes :

Tableau 7 Structure de données pour l'entrainement d'un arbre de décision

Thème	Pourcentage de oui	Résultat national
Économie	54%	Non
Immigration	33%	Oui
…	…	…

Colonnes de calculs

Colonne de résultats

L'algorithme va alors créer un arbre de décision à l'aide du jeu de données d'entrainement. Cet arbre est une succession de choix binaire basée sur les deux premières colonnes de calculs du Tableau 7. Le dernier niveau de chaque embranchement est le résultat prédit.

La Figure 11 est un exemple d'arbre de décision possible basé sur la structure de donnée du Tableau 7. Il présente une succession de chemins qui terminent tous par une prédiction du résultat de la votation à l'échelle nationale.

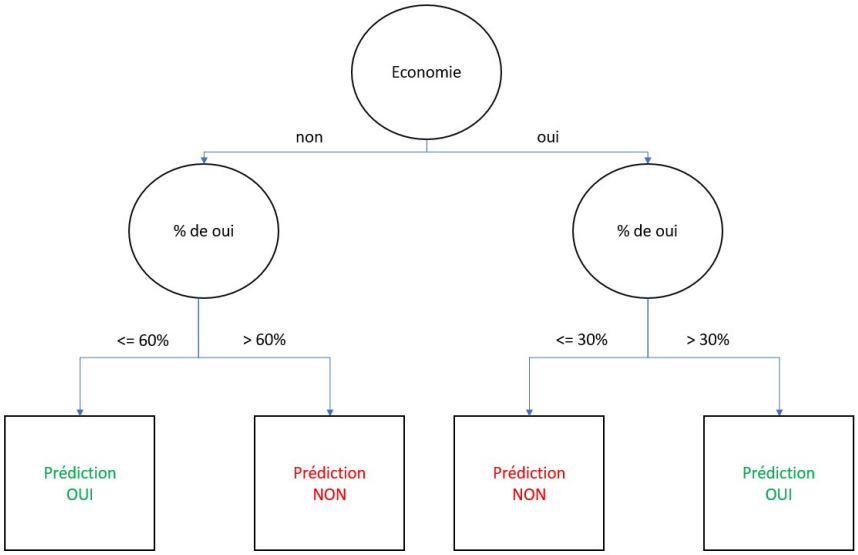


Figure 11 Exemples simplifiés d'un arbre de décision binaire

Afin de pouvoir être utilisé, il faut renseigner à l'arbre de décision une valeur pour chacune des colonnes de calculs. Ces valeurs sont prisées du jeu de donnée test. Dans ce cas, il s'agit des colonnes « Thème » et « pourcentage de oui ». Nous allons maintenant regarder une prédiction avec les valeurs présentées dans le Tableau 8.

Tableau 8 Valeurs de tests pour la prédiction

Thème	Pourcentage de oui
Économie	22%
Immigration	60%

La Figure 12 représente l'utilisation de l'arbre de décision de la Figure 11 pour les données de tests du Tableau 8. Les objets de couleurs vertes indiquent le flux de données. La prédiction faite par l'arbre de décision se trouve sur le dernier niveau de celui-ci.

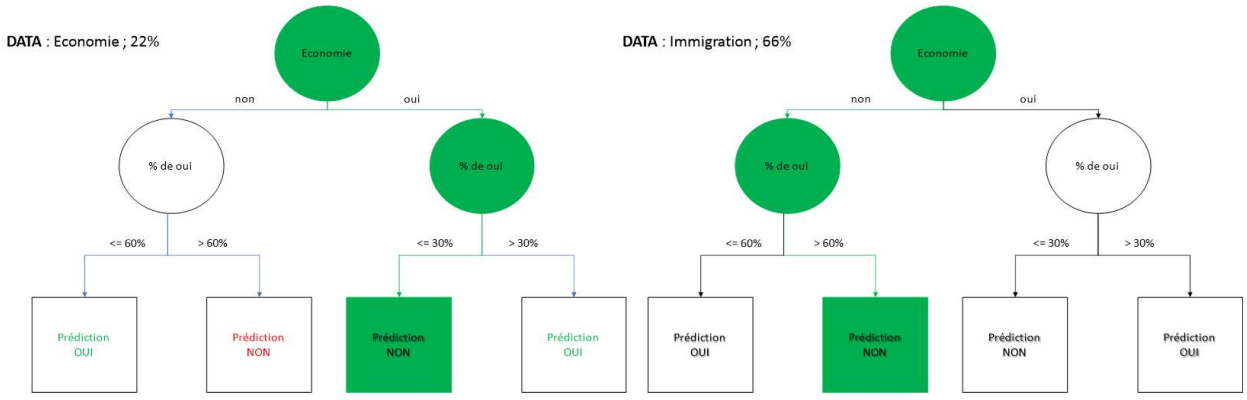


Figure 12 Prédiction à l'aide d'un arbre de décisions

La dernière étape est de comparer le résultat prédit par l'arbre avec la valeur qui est dans le jeu de test et ainsi mesurer la capacité de prédiction de l'arbre.

Avec notre jeu de données, nous ne pouvons pas appliquer cette méthode pour déterminer la capacité de prédiction de nos communes.

Tout d'abord, notre jeu de données varie durant les différentes élections. Dans l'exemple des votations populaires, la structure des colonnes de calculs est toujours la même : le thème et le pourcentage de « oui ». Pour les élections du Conseil d'État valaisan, le nombre de candidats varie entre les années. Nous disposons une fois de sept colonnes et une autre année de douze colonnes. Nous ne disposons pas de données fixes pour toutes les années pouvant être utilisées comme colonne de calcul, car le nombre de candidats est variable.

Ensuite, le résultat que l'on souhaite prédire est un vecteur de dimension égale au nombre de candidats (qui est variable). Cette prédiction d'un vecteur de résultat n'est pas adaptée aux arbres de décision.

Enfin, nous disposons d'un volume de données réduit en comparaison avec l'étude sur les votations populaires. Si nous appliquons la méthode des 80% des données pour l'entrainement et 20% pour le test sur nos 5 années d'élections alors l'algorithme s'entrainera sur seulement 4 années et sera testée sur une seule année. Pour rappel, il est question d'un arbre par commune et un jeu d'entrainement de quatre ans implique seulement quatre exemples pour l'algorithme.

Bien que la prédiction de résultat – c'est-à-dire les cinq personnes élues – ne soit pas possible avec cette méthode, nous pouvons réécrire le problème. Il nous faut stabiliser les colonnes de calculs et fixer un ensemble fini de prédiction. Nous proposons alors le problème suivant :

« Sachant le pourcentage de voix attribué à un candidat par une commune, est-il possible de prédire si celui-ci sera élu ? »

Avec cette formulation, nous ne cherchons plus à prédire les cinq élus, mais les résultats individuels. Notre colonne de calcul est fixée - le pourcentage de voix attribuées par la commune au candidat – et le résultat – élu, pas élu- est binaire. Avec cette approche, nous disposons d'une structure de données qui peut être utilisée pour entrainer un arbre de décision. La quantité de données utilisable augmente légèrement, car le nombre d'années est multiplié par le nombre de candidats. À noter que cette quantité reste faible pour le domaine de l'apprentissage machine. Nous séparons nos données entre entrainement et résultat en respectant le ratio 80-20.

Tableau 9 Structure de données utilisées pour entrainer les arbres de décision de chaque commune

Pourcentage obtenu par le candidat	Est élu
60	Oui
30	Non
…	…

Le Tableau 9 montre la structure de données que nous utilisons pour l'entrainement de nos arbres de décisions. Nous pourrions rajouter d'autres colonnes de calculs, par exemple des caractéristiques des candidats comme le sexe, l'âge, candidat sortant ou encore la langue parlée. Nous désirons de ne pas le faire dans un premier temps afin de tester uniquement la capacité de prédiction de la commune.

3.2.2 Approche par Arrangement

Une autre approche est de classer les communes par un score déterminé sur leur capacité à prédire un arrangement correct des candidats élus. Notre vecteur de résultat est un arrangement de candidat et nous pouvons le comparer avec le classement des candidats dans chaque commune. Le score de la commune est compris entre zéro et cinq : zéro correspond à aucune valeur de l'arrangement correct et cinq correspond à un arrangement similaire.

Donnons un exemple pour illustrer cela en imaginant le vecteur de résultat suivant :

$$V_{resultat} = \{Paul, Luc, jean, Henry\ , Bernard\}\ ^{9}$$

Ensuite, prenons le résultat de l'élection pour la commune A :

$$V_A = \{Paul, jean, Luc, Henry\ , Bernard\}\ ^{10}$$

La commune A obtient un score de $\mathbf{3/5}$ car les place 1, 4 et 5 ont été correctement prédites et que les places 2 et 3 sont fausses.

Nous appliquons ce calcul à chaque commune pour les cinq élections étudiées. Nous disposons alors d'une liste de communes par années d'élection classées par score de prédiction. Finalement, nous observons les communes arrivées en tête sur les cinq années afin de déterminer si un pattern se dessine.

3.2.3 Approche par Combinaison

L'approche par capacité à prédire un arrangement, impose aux communes de déterminer les élus dans le bon ordre. Cependant, une commune qui a classé les bonnes personnes dans un mauvais ordre à tous de même prédit les cinq bons élus. Afin d'intégrer cela dans le modèle, nous testons une approche par combinaison. En effet, une combinaison est un ensemble d'éléments dont l'ordre ne compte pas. Nous pouvons alors attribuer un système de points à chaque commune de la même manière que pour les arrangements, mais cette fois nous testons l'égalité entre deux combinaisons.

Reprenons comme exemple notre vecteur de résultats précédent :

$$V_{resultat} = \{Paul, Luc, jean, Henry\ , Bernard\}$$

Ainsi que notre commune A d'exemple :

$$V_A = \{Paul, jean, Luc, Henry\ , Bernard\}\ ^{11}$$

Le score obtenu par la commune $\boldsymbol{A}$ est de $\mathbf{5/5}$ car l'ordre des éléments ne compte pas dans la comparaison. Chaque commune obtient un score de la même manière et nous les classons du plus élevé au plus faible. Nous observons ensuite l'évolution des communes bien placées pour chacune des années d'élections. Nous observerons également la quantité d'égalités que cette solution pourrait générer par rapport à la solution par comparaison d'arrangement. En effet, plusieurs classements différents donnant la même combinaison, il risque d'y avoir un nombre conséquent de communes qui donnent le bon résultat et donc des égalités.

[9] Pour rappel, il s'agit d'un arrangement, l'ordre des valeurs est donc important.

[10] Le vecteur correspond aux cinq candidats arrivés en tête dans la commune A (du premier au cinquième).

[11] Le vecteur correspond aux cinq candidats arrivés en tête dans la commune A (du premier au cinquième).

3.2.4 Approche par distance vectorielle

Enfin, notre dernière approche pour fournir un modèle de prédiction est d'utiliser une distance entre une représentation vectorielle des votes des communes et celui du résultat final. Pour cela, nous devons créer un vecteur représentant le résultat de l'élection. Nous projetons ce résultat dans un espace de dimension égale aux nombres de candidats et nous projetons ensuite les résultats des communes dans le même espace. Nous calculons les distances entre le résultat et chacune des communes et sur la base de cette valeur nous pouvons faire un classement. Le Figure 13 est l'illustration de la méthode utilisée pour calculer toutes les distances entre les communes et le résultat final. Celle-ci est présentée dans une situation en deux dimensions afin de pouvoir la visualiser, mais l'approche est généralisable à n'importe quel nombre entier positif de dimensions.

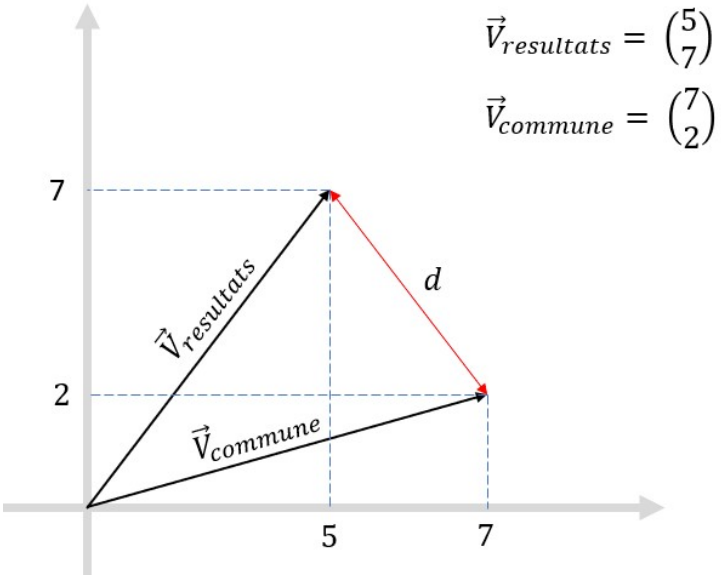


Figure 13 Exemples de calcul de distance euclidienne entre deux vecteurs pour un espace à deux dimensions

Une fois ces calculs effectués, nous disposons à nouveau d'une liste de communes ordrée de la plus proche du résultat à la plus éloignée. Nous répétons ces opérations pour les cinq élections étudiées et observons les classements afin de trouver des communes régulièrement présentes en tête de liste. L'objectif est de trouver une ou plusieurs communes fréquemment plus proches du résultat lors des élections de 2001 à 2017.

Chapitre 4. RÉSULTATS ET ANALYSES

Ce chapitre est consacré à la mise en place de notre méthodologie pour les élections du Conseil d'État valaisan. Nous présenterons nos résultats ainsi que les analyses et interprétations qui leur sont liées. Les solutions techniques utilisées pour appliquer notre méthodologie seront aussi décrites.

La première partie du chapitre est dédiée aux résultats de l'analyse de comportements électoraux dans laquelle nous présenterons l'ensemble de nos résultats, observation, analyse et interprétation. La seconde concerne les résultats de l'application des quatre méthodes d'analyse prédictive présentée dans notre méthodologie.

4.1 ANALYSE DES COMPORTEMENTS ELECTORAUX

Nous présenterons dans cette partie l'ensemble des résultats concernant l'analyse des comportements électoraux. Pour chaque partie, nous commencerons par décrire les flux des données et traitement appliqués. Ensuite, nous décrirons la mise en place technique avant de présenter les résultats. Enfin, nous commenterons les résultats présentés ainsi que leur interprétation et analyses.

4.1.1 Prétraitements des données

Nous commençons par appliquer les traitements nécessaires pour obtenir les « dataframe » exploitable pour la suite de nos traitements. Figure 14 présente le flux de traitement appliqué à nos données sources afin d'obtenir des dataframes correctement structurés.

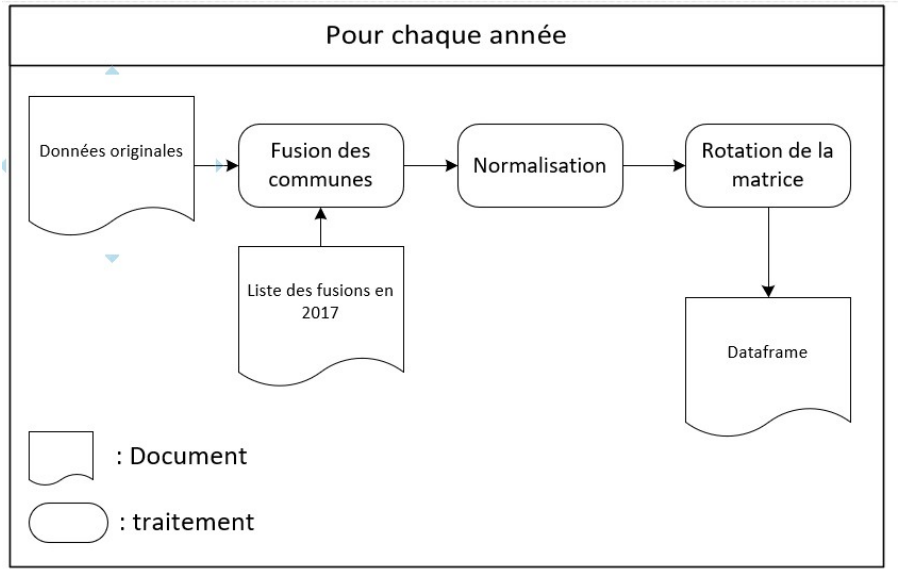


Figure 14 Flux des traitements entre les données originales et les dataframes

Les données utilisées sont les résultats par communes des élections au Conseil d'État valaisan pour les années 2001, 2005, 2009, 2013 et 2017. Pour chaque année, nous disposons d'un fichier en matrice où les lignes sont les candidats et les informations sur l'élection et les colonnes sont les communes. Le Tableau 10 présente la structure de base des fichiers sources. Chaque bulletin valable peut distribuer entre une et cinq voix aux candidats.

Tableau 10 Structure de données des fichiers sources

	Commune 1	Commune 2
Bulletins valables	…	…
Candidat 1	…	…
Candidat 2	…	…

Nous voulons avoir les mêmes communes pour chaque année. Comme précisé dans notre méthodologie, nous appliquons l'état de 2017 à toutes les autres années. Le premier traitement dans le flux de données est la gestion des fusions pour toutes les années d'élections. Ce traitement se déroule comme ceci :

Pour chaque année et pour chaque fusion :

1. Sommer les scores des candidats
2. Sommer les bulletins valables
3. Créer une nouvelle entrée dans le fichier pour la commune fusionnée
4. Supprimer les communes fusionnantes du fichier

Ces simples sommes nous permettent de créer dans le passé les communes fusionnées. À noter que le système doit être capable de déterminer si la fusion est nécessaire. Par exemple, une fusion qui a lieu en 2003 doit être appliquée pour le jeu de données de 2001, mais ignorée pour les jeux de données plus récents que 2003. Ce traitement est réalisé sur la plateforme d'analyse Suisse KNIME.

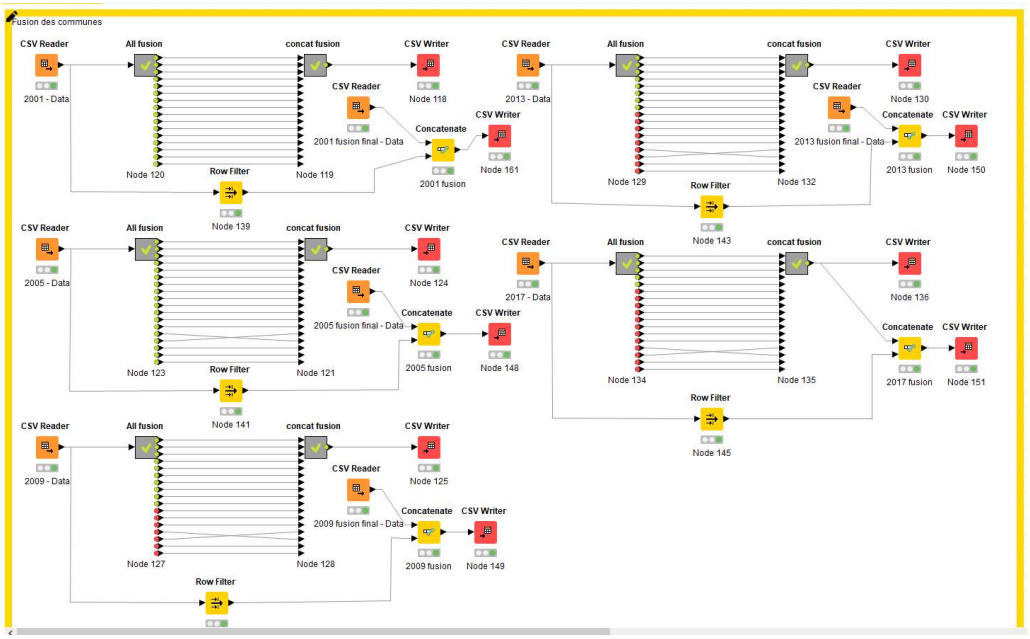


Figure 15 Modules de gestion des fusions sur le logiciel KNIME

La Figure 15 présente le module de gestion des fusions. On distingue cinq parties pour les cinq élections. Chaque partie démarre du fichier de base et lui applique les fusions de communes. Les points verts indiquent que la fusion était utile et que celle-ci a réussi. Les points rouges indiquent que la fusion était inutile et que le système l'a ignorée. Les outputs sont les fichiers sources avec les fusions appliquées.

Maintenant que nos fusions sont appliquées, nous pouvons normaliser nos données. En effet, le nombre de voix que peuvent obtenir les candidats est directement lié au nombre de bulletins valable, et cela crée une différence entre les petites et grandes communes. Une commune avec plus d'habitants – donc plus de bulletins valables – aura des candidats avec sensiblement plus de voix que dans une petite commune. Afin de comparer les communes malgré un nombre de bulletins valables variables, nous normalisons les scores des candidats. Cette normalisation consiste à calculer le pourcentage que représente chaque voix attribuée aux candidats par rapport aux bulletins inscrits.

Le dernier traitement à faire sur nos données d'élections, est d'effectuer une rotation de la matrice $Candidats\ X\ Communes$ afin d'obtenir les communes en lignes et les candidats en colonnes. La ligne correspondant au bulletin valable n'est pas gardée. Cette opération est faite avec un script en Python.

Nous disposons alors d'une matrice $Communes\ X\ Candidats$ par année avec pour chaque commune des données numériques normalisées correspondant aux voix reçues par les candidats. Toutes nos années ont des lignes de communes identiques grâce aux fusions appliquées dans le passé. Les lignes des communes avec les valeurs numériques attribuées aux candidats correspondent à nos vecteurs de votation. Le Tableau 11 présente la structure finale de nos dataframes après avoir appliqué l'ensemble de traitement aux données de bases. La Figure 16 montre comme exemple les premières lignes du dataframes de l'élection de 2001.

Tableau 11 Structure de nos dataframes pour chacune des années d'élections

		Candidat 1	**Candidat 2**	**…**	**Candidat N**
Vecteur de votation $\rightarrow_{V_1}$	**Commune 1**	Valeur numérique normalisée	…	…	…
Vecteur de votation $\rightarrow_{V_2}$	**Commune 2**	…	…	…	…
…	**…**	…	…	…	…
Vecteur de votation $\rightarrow_{V_n}$	**Commune N**	…	…	…	…

Row ID	Thomas Burgener	Claude Roch	Wilh…	Jea…	Je…	M…	C…
Agarn	0.4	0.025	0.721	0.683	0.664	0.025	0.043
Agettes	0.204	0.106	0.627	0.662	0.662	0.042	0.155
Albinen	0.439	0.007	0.633	0.583	0.59	0.022	0.165
Arbaz	0.34	0.126	0.52	0.579	0.55	0.083	0.252
Ardon	0.168	0.308	0.546	0.539	0.555	0.057	0.25
Ausserberg	0.415	0.003	0.668	0.653	0.662	0.012	0.074
Ausserbinn	0.933	0.133	0.933	0.8	0.8	0.133	0.6

Figure 16 Premières lignes du dataframes de l'élection de 2001

Nous pouvons avec nos cinq dataframes créer le super-espace. Celui-ci regroupe pour chaque commune et en un seul vecteur de votation, l'ensemble de toutes les élections. Les étapes pour créer ce nouveau document sont présentées dans la Figure 17.

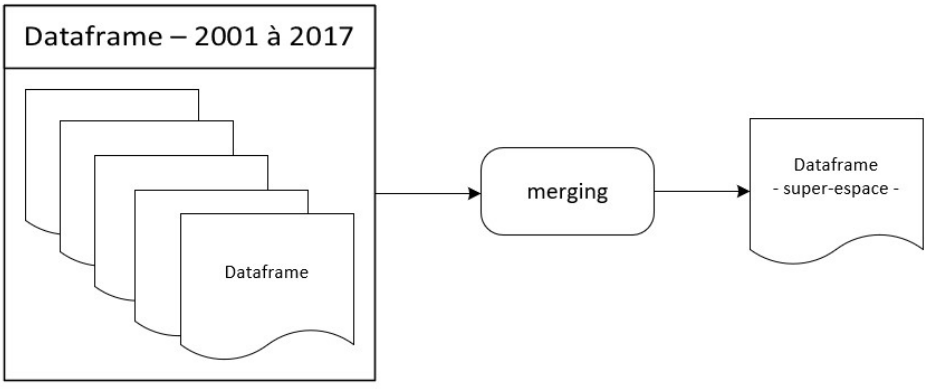
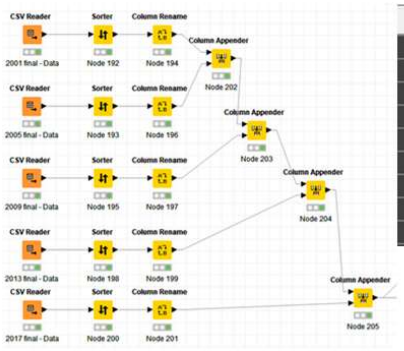


Figure 17 Flux de données pour la création du super-espace

Nous partons des cinq fichiers sources transformés en dataframes et appliquons un traitement de fusion des données. Cette fusion consiste à fusionner l'ensemble des lignes – groupement par commune - et de dupliquer les colonnes. La subtilité dans cette fusion est que lorsqu'un même candidat s'est présenté à plusieurs élections alors, il faut garder deux colonnes dans notre super-espace. Pour cela, nous annotons les colonnes « candidats » avec l'année à laquelle ils se sont présentés. La Figure 18 affiche à gauche le flux créé sur KNIME et à droite les premières lignes et colonne du résultat. Nous pouvons observer l'annotation effectuée sur les candidats afin d'éviter que deux colonnes soient fusionnées. Quant aux lignes, elles sont fusionnées pour n'en garder qu'une seule par colonne.



Row ID	Thomas Burgener (2001)	...	...	...	...	...	...	Claude Roch (2005)
Agarn	0.4	0.025	0....	0....	0....	0....	...	0.055
Albinen	0.439	0.007	0....	0....	0.59	0....	...	0.054
Anniviers	0.297	0.183	0....	0....	0....	0....	...	0.322
Arbaz	0.34	0.126	0.52	0....	0.55	0....	...	0.197
Ardon	0.168	0.308	0....	0....	0....	0....	...	0.386
Ausserberg	0.415	0.003	0....	0....	0....	0....	...	0.037
Ayent	0.351	0.14	0....	0....	0....	0....	...	0.199
Bagnes	0.156	0.204	0....	0....	0....	0.08	...	0.39
Baltschieder	0.733	0.086	0....	0....	0....	0....	...	0.166

Figure 18 Modules KNIME pour la création du super-espace et première ligne du nouveau dataframe créé pour le super-espace

Nous obtenons alors une matrice de 127 lignes et 44 colonnes. Celle-ci capture, dans un seul vecteur de votation par commune, toutes les distributions de voix des communes aux candidats pour toutes les élections de 2001 à 2017.

Cette étape marque la fin des prétraitements sur nos données de base. Nos cinq dataframes d'élections ainsi que notre super-espace sont prêts à être utilisés dans nos prochains traitements.

4.1.2 Partitionnement des données

Maintenant que nos données sont prêtes, nous passons à la partie partitionnement des données et analyse des résultats. Nous commençons par mettre en place le flux de traitements pour choisir le nombre de clusters pour notre algorithme K-mean. Nous choisirons une ou plusieurs valeurs pour l'ensemble de nos dataframes. La Figure 19 présente le flux logique de nos traitements pour définir le bon nombre de clusters. Comme présentés dans notre méthodologie, nous utilisons l'approche Elbow Method. Le flux présenté nous permet de créer l'ensemble des graphiques afin de rechercher les « coudes ». Ces traitements sont réalisés avec une script Python.

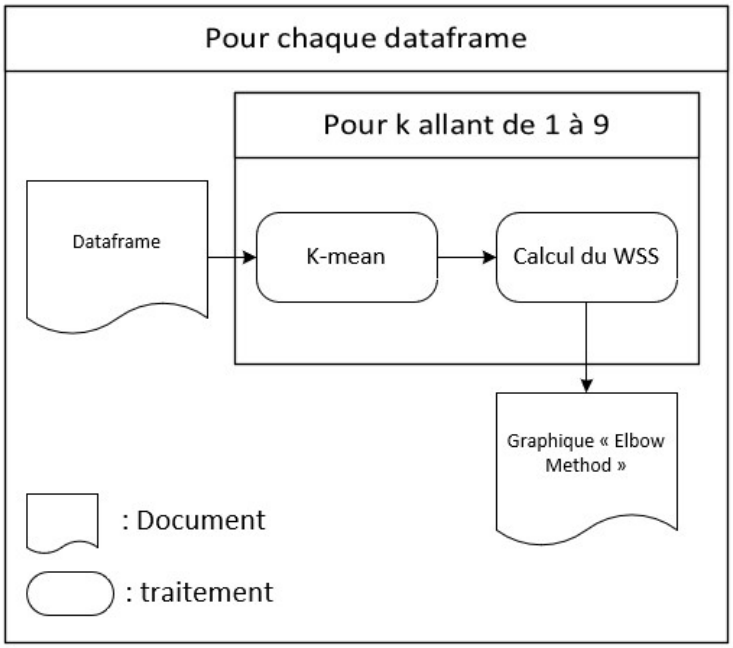


Figure 19 Flux de traitements pour choisir le bon nombre de clusters

Après l'exécution du script, nous disposons d'un graphique par dataframe que nous allons pouvoir étudier.

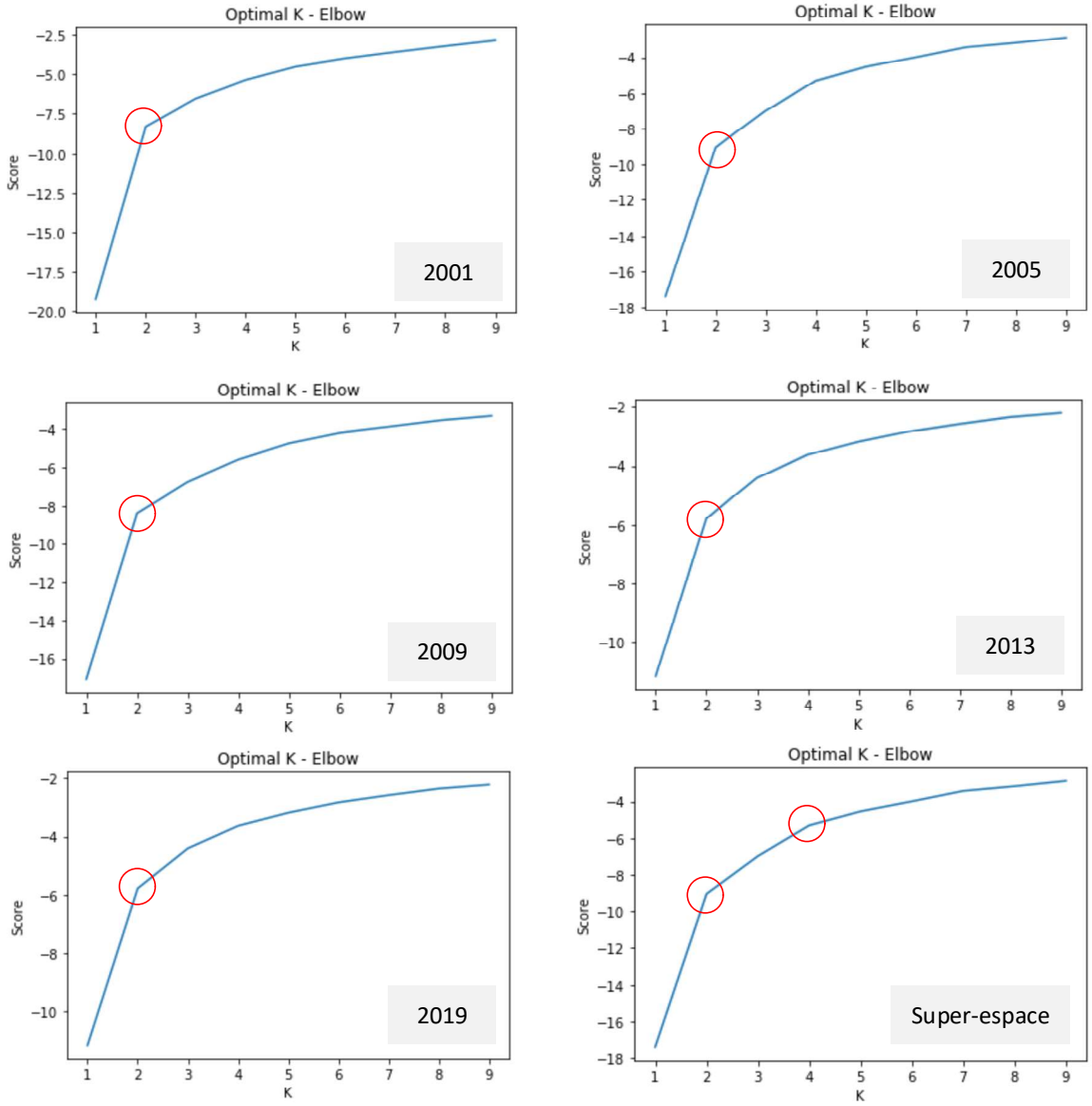


Figure 20 Graphique Elbow Method pour trouver le k optimal. Un graphique par dataframe

Nous pouvons observer sur la Figure 20 que tous nos dataframe présentent une cassure pour un $k = 2$. Celui-ci est très net pour tous les graphiques. L'angle semble toutefois plus fermé en 2001 et 2009 par rapport aux autres années. Le super-espace semble également présenter une légère cassure pour un $k = 4$. Nous pouvons alors choisir une valeur de 2 pour l'ensemble de nos dataframes, car celle-ci semble capturer le plus efficacement les différences. Nous testerons quand même une valeur de 4 pour le super-espace afin d'analyser aussi cette répartition.

Nous avons nos valeurs pour le nombre de clusters à déterminer à l'aide de K-mean. Nous pouvons alors lancer l'algorithme de partitionnement et analyser nos clusters. Le flux suivant présente les traitements effectués.

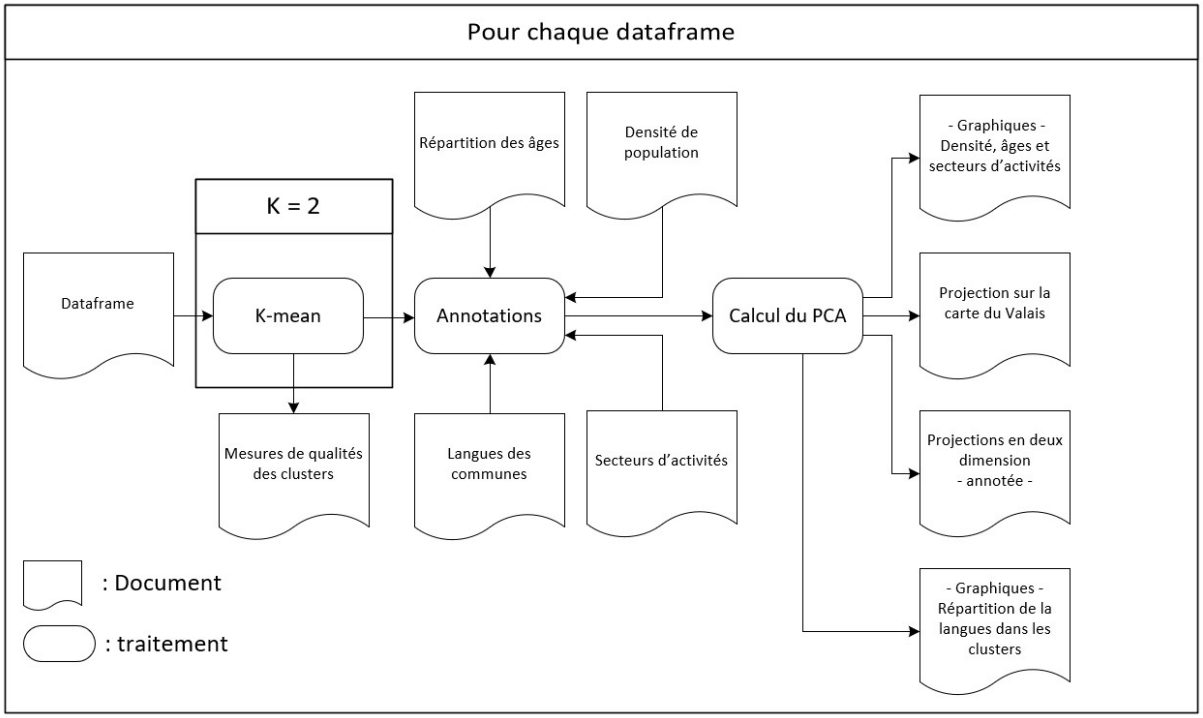


Figure 21 Flux des traitements pour le partitionnement des données et la génération des graphiques et mesure nécessaire à l'analyse.

Prenons les éléments de la Figure 21 de droite à gauche. Le flux commence par le dataframe, celui-ci est alors utilisé par l'algorithme K-mean afin de créer deux clusters. Le traitement par l 'algorithme produit deux outputs : le dataframe annoté avec le cluster attribué à chaque commune et des mesures de qualités des clusters comme les distances intra et inter cluster. Le flux continu par une étape d'annotations des données par les valeurs que nous avons décrite dans notre méthodologie. La dernière étape est la création des graphiques pour l'analyse. Nous utilisons la méthode du PCA afin de pouvoir projeter les valeurs sur deux dimensions. L'ensemble de ces traitements a été réalisé sur KNIME avec le module présenté dans la Figure 22. À noter que la génération de la carte et la projection des clusters a été produite en Python sur la base des données de frontières fournies par la confédération (swissBOUNDARIES3D).

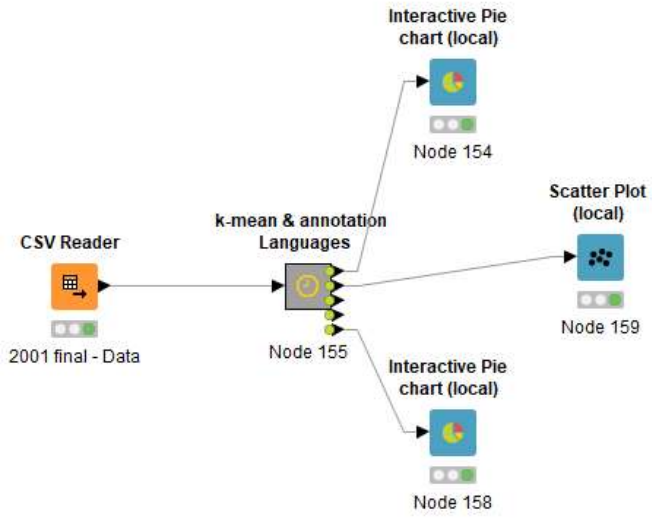


Figure 22 Modules KNIME de partitionnement des données

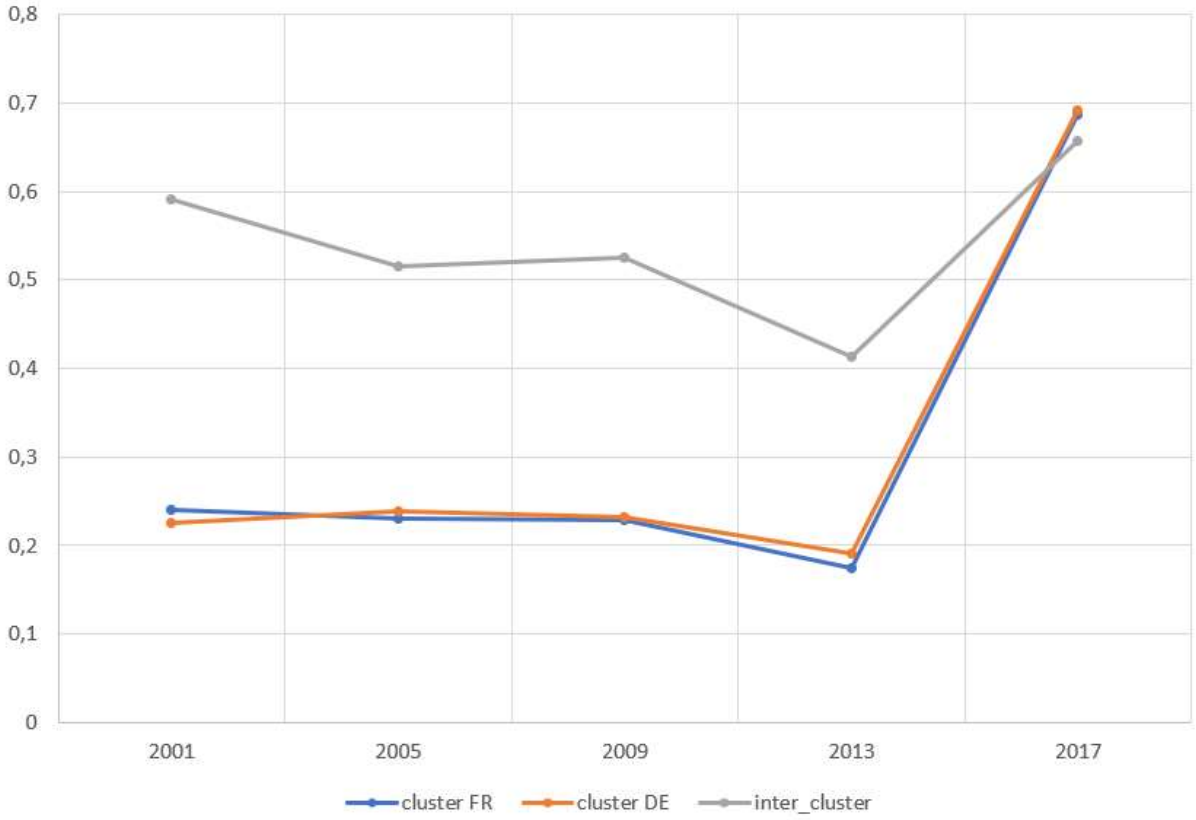


Figure 23 Mesures de dispersions inter et intra clusters pour les années 2001 à 2017

Nous pouvons observer sur la Figure 23 l'évolution des mesures de dispersion inter et intra clusters. Les mesures intra clusters pour le groupe FR et DE sont fortement corrélées et restes stables jusqu'en 2013. Nous constatons une augmentation de la dispersion intra cluster en 2017. La mesure de distance entre les deux clusters est également corrélée à la mesure intra cluster. Cette valeur inter cluster augmente en 2017. Nous avons des données en 2017 plus séparées dans les clusters, mais en deux groupes plus éloignés. En 2013, les données se contractent dans les clusters, mais ceux-ci se rapprochent.

Regardons maintenant l'ensemble des outputs afin d'analyser la répartition des communes par comportement de vote.

4.1.2.1 L'influence de la langue

Dans un premier temps, nous allons observer la projection des clusters sur la carte du Valais. Pour rappel, nous utilisons la situation des communes de 2017 comme base pour toutes les années. Nous utiliserons également l'annotation des langues faites sur les communes afin de déterminer la répartition à l'intérieur des clusters. À noter que les annotations ne sont pas prises en compte lors de la création des clusters par l'algorithme K-mean. Seules les valeurs des votes sont interprétées pour créer les clusters.

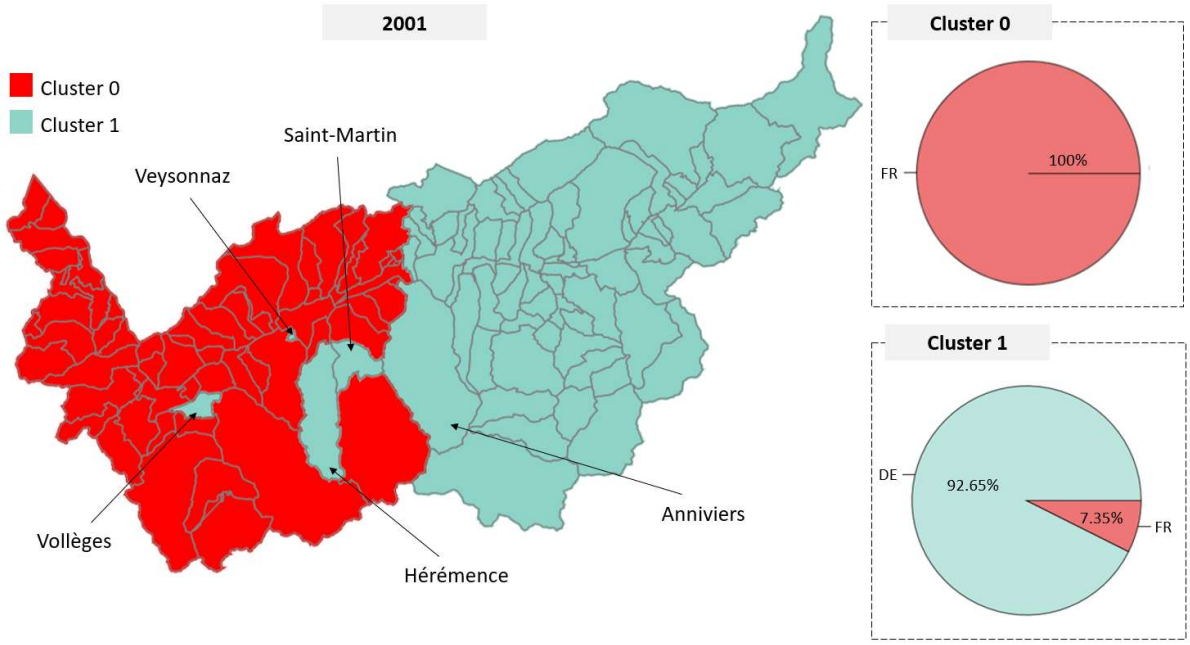


Figure 24 Projection des clusters sur la carte du valais et répartition des langues dans les clusters pour l'année 2001

La Figure 24 affiche la répartition en cluster – avec une valeur de 2 pour k – des comportements de votes pour l'année 2001. Nous pouvons constater une séparation géographique claire sur la frontière entre le Valais central et le haut valais présenté sur la Figure 1. Cette frontière correspond à la frontière entre les communes francophones et les communes alémaniques. Les éléments à droite de la Figure 24 indiquent la répartition des langues dans les deux clusters. Le cluster 0 est entièrement composé de communes francophones, il n'y a aucune commune mal placée dans ce cluster. En revanche, le cluster 1 est composé de 92.65% de communes alémaniques. En effet, 7.35% des communes du cluster 1 sont de langue française. Cette valeur correspond à cinq communes affichées par des flèches sur la carte. Nous constatons que Vollèges, Veysonnaz, Saint-Martin, Hérémence et Anniviers ont un comportement de vote plus proche des communes alémanique que francophone. La commune d'Anniviers à un contact direct avec la frontière valais central et haut valais, les autres communes sont entourées de commune francophone.

Observons maintenant l'évolution de ces comportements durant les quatre autres années.

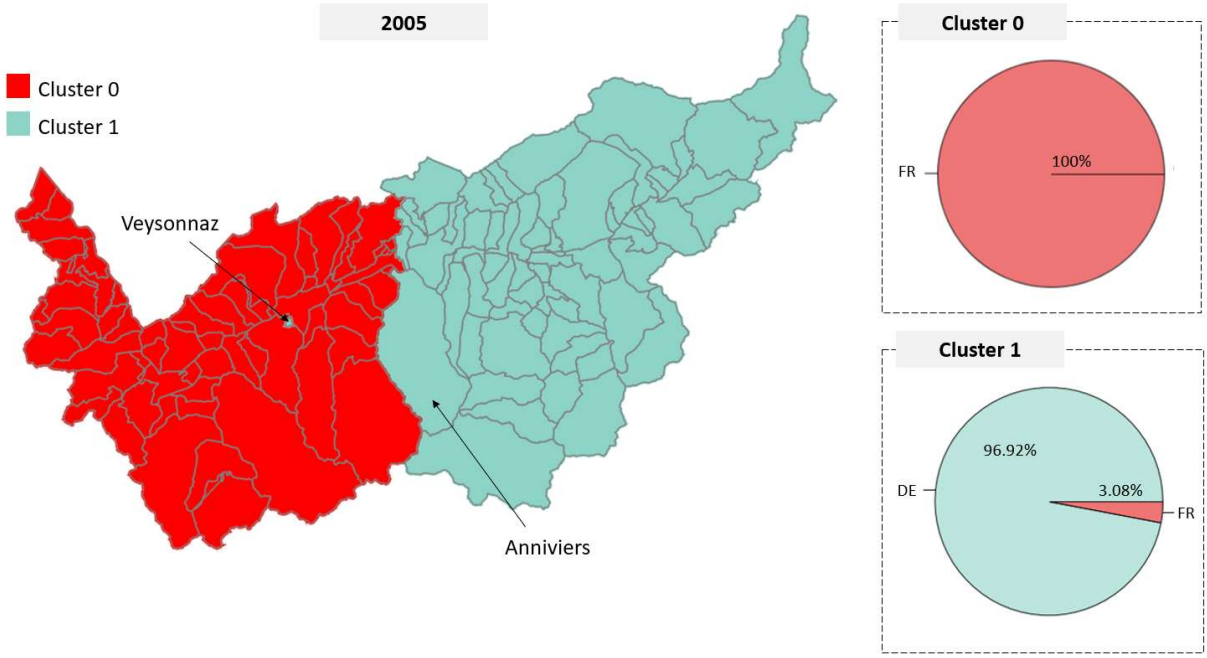


Figure 25 Projection des clusters sur la carte du valais et répartition des langues dans les clusters pour l'année 2005

La Figure 25 affiche la situation pour l'année 2005. Sur les cinq communes mal classées de 2001, seul Veysonnaz et Anniviers sont encore attribués par leurs comportements de vote au cluster à majorité alémanique.

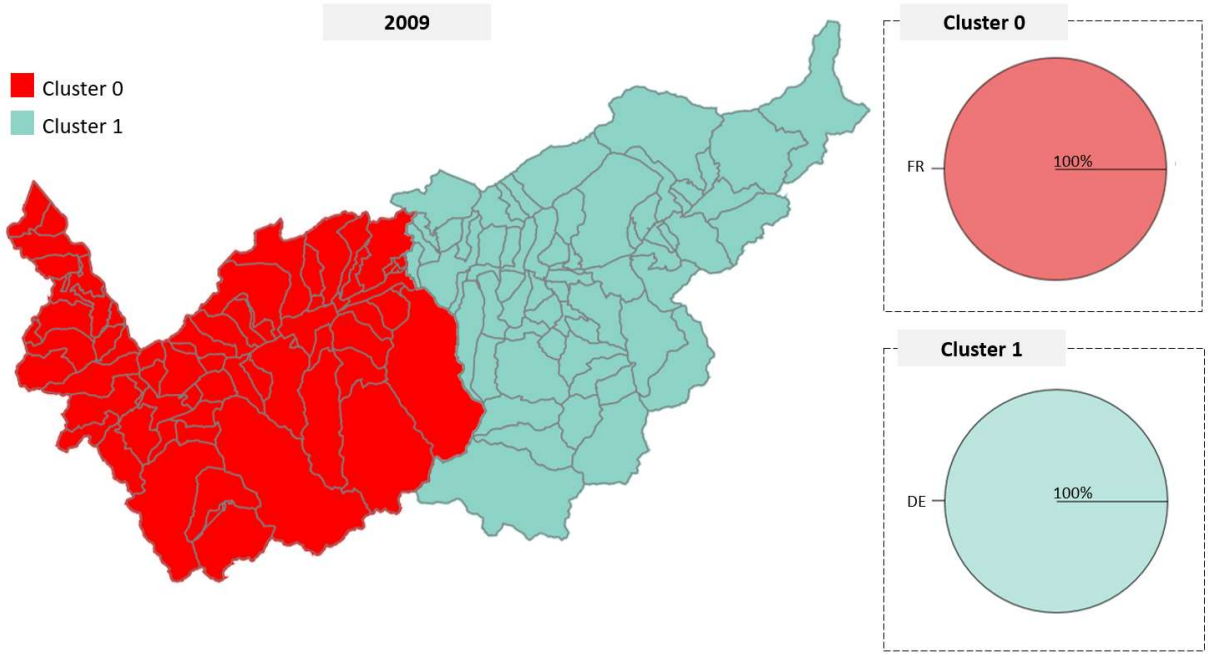


Figure 26 Projection des clusters sur la carte du valais et répartition des langues dans les clusters pour l'année 2009

Les clusters et la répartition des langues pour l'année 2009 sont présentés sur la Figure 26. Il s'agit de la première année de notre jeu de donnée avec un classement parfait. En effet, toutes les communes de chaque cluster parlent la même langue. Les communes mal classées des années précédentes sont en 2009 dans leur bon cluster.

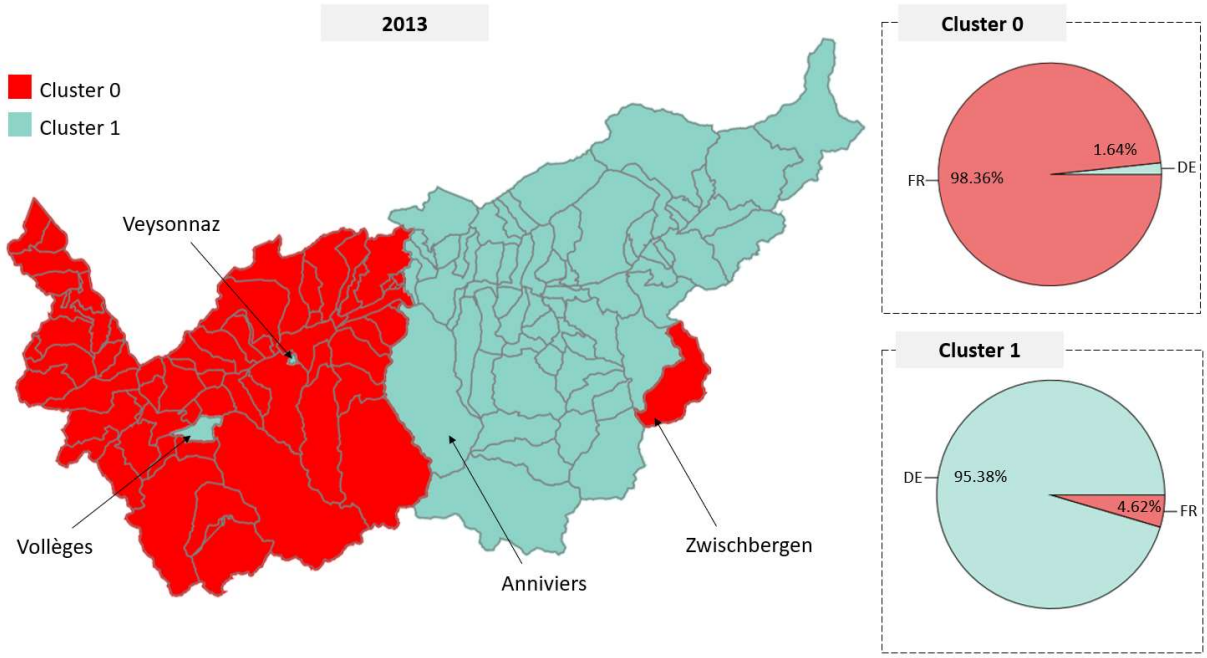


Figure 27 Projection des clusters sur la carte du valais pour l'année 2013 et répartition des langues dans les clusters

L'élection de 2013 présente une situation encore différente. Bien que la frontière des langues soit toujours dessinée, la Figure 27 nous permet de constater que pour la première fois le cluster francophone n'est pas homogène. En effet, la commune de Zwischbergen rejoint la liste des communes mal classées, car celle-ci à un comportement de vote plus proche des communes francophones avec une langue majoritaire allemande. Zwischbergen n'est pourtant pas une commune de la frontière linguistique du Valais et nous observons également le retour dans le cluster alémanique des communes Vollèges, Veysonnaz et Anniviers.

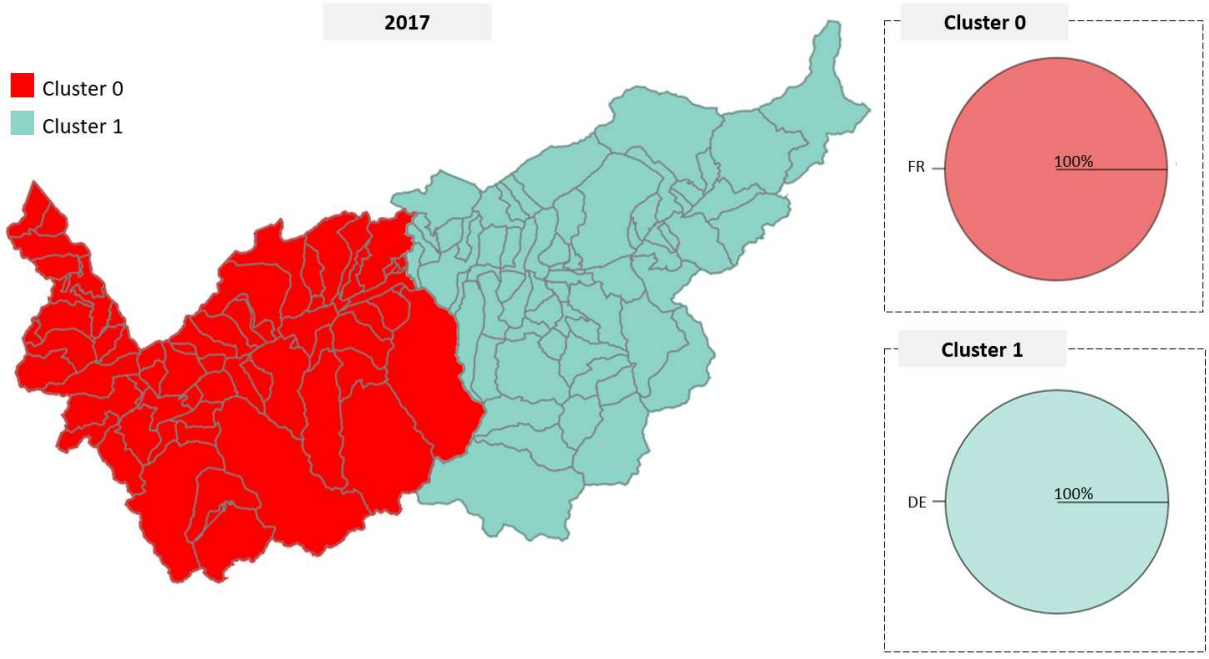


Figure 28 Projection des clusters sur la carte du valais et répartition des langues dans les clusters pour l'année 2017

Pour l'année 2017, nous retrouvons la situation de 2009. C'est-à-dire que l'ensemble des clusters sont homogènes par rapport à la langue des communes. La Figure 28 présente cette situation. Il s'agit de la deuxième année avec un classement parfait de notre jeu de données.

Nous avons vu la répartition de la langue dans les clusters pour les cinq années d'élections. Les deux années 2009 et 2017 présentent un classement parfait alors que les années 2001 et 2005 ont un cluster alémanique composé de communes francophones. L'année 2013 présente encore une spécificité en proposant une situation dans laquelle les deux clusters s'échangent des communes. Nous pouvons maintenant regarder quelle situation l'emporte lorsque nous traitons toutes les élections en une fois à l'aide de notre super-espace. Pour rappel, le super-espace reprend tous les votes faits par toutes les communes entre 2001 et 2017.

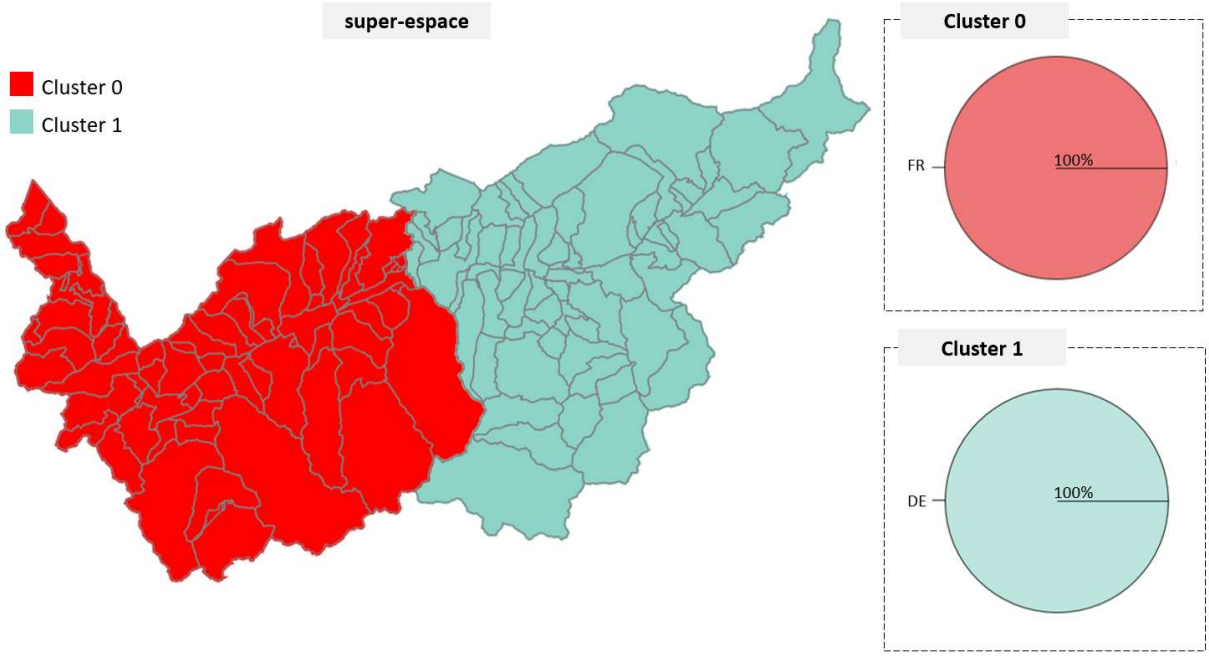


Figure 29 Projection des clusters sur la carte du valais et répartition des langues dans les clusters pour le super-espace

La Figure 29 montre à nouveau une répartition parfaite pour le super-espace. Nous constatons que même si les situations avec des communes mal placées étaient majoritaires avec 3 cas sur 5, c'est une répartition parfaite qui est obtenue lors de l'analyse de l'ensemble des élections. Cela peut être expliqué par le fait que les clusters ne sont jamais descendus en dessous des 90% de communes parlant la même langue, et qu'uniquement Veysonnaz et Anniviers sont mal classées sur l'ensemble des trois situations atypiques.

Nous avons observé, à l'aide de l'approche « Elbow method », que les comportements de votes pouvaient être séparés en deux catégories. Une fois cette séparation faite avec l'algorithme K-mean, nous observons une corrélation forte entre les groupes crées sur la base des comportements électoraux et la langue parlée des communes. En effet, deux années sur cinq présente une correspondance parfaite, et pour les autres, celle-ci se chiffre à plus de 90%. Le super-espace confirme encore cela en présentant lui aussi une corrélation parfaite. Avec notre méthodologie, nous avons confirmé que la langue est un facteur qui peut expliquer la différence de comportement électoral des communes valaisannes. De plus, celle-ci semble stable dans le temps et nous n'observons pas de tendances particulières entre les années 2001 et 2017. Enfin, nous pouvons souligner qu'à l'exception de la commune d'Anniviers, les communes frontières ne sont pas plus susceptibles de passer d'un cluster à l'autre par rapport au reste des communes.

4.1.2.2 Les facteurs sociaux économiques

Nous avons confirmé que la langue peut être un facteur explicatif pour la répartition obtenue avec l'algorithme K-mean. Nous allons maintenant regarder la répartition dans nos clusters des communes par rapport aux autres annotations que nous avons ajoutées à nos données. L'idée est de regarder si un autre clivage apparait en plus de la langue des communes. Pour ce faire, nous avons annoté nos données avec les valeurs suivantes : densité de population, pyramide des âges et secteurs d'activités. Nous ne disposons pas de données pour chacune des années pour l'ensemble des communes, alors nous étudierons ces nouvelles valeurs directement sur le super-espace. Observons à l'aide de la Figure 30 la répartition de ces valeurs à l'intérieur des clusters de notre super-espace.

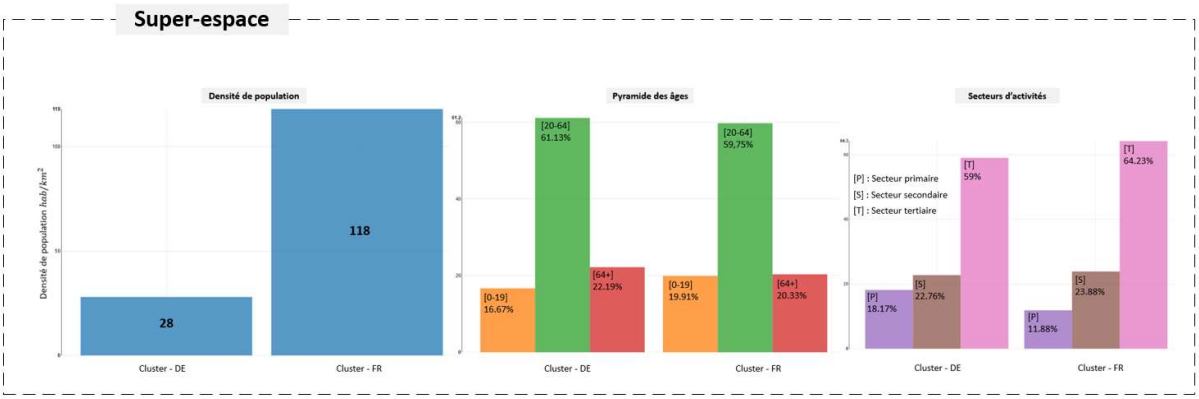


Figure 30 Densité de population, pyramide des âges et secteurs d'activités pour les clusters du super-espace

Le premier graphique situé à gauche affiche la médiane de la densité de population des deux clusters. La notation « DE » représente le cluster dans lequel les communes sont majoritairement alémaniques alors que « FR » représente les francophones. Nous constatons une nette différence entre les deux clusters. Le cluster FR à une médiane plus de quatre fois supérieure à celle du cluster DE. Si nous observons nos clusters en prenant la densité comme critère de classement alors nous pouvons de la même manière que pour les langues lister les communes mal-classées. Nous commençons par rechercher les communes de faible densité dans le cluster FR. Six communes se retrouvent en dessous de la valeur 28 représentant la médiane de cluster DE. Il s'agit des communes suivantes :

- Bourg-Saint-Pierre : 2 hab/km^2
- Trient : 5 hab/km^2
- Evolène : 8 hab/km^2
- Anniviers : 11 hab/km^2
- Orsières 19 hab/km^2
- Saint-Martin : 22 hab/km^2

Les communes d'Anniviers et Saint-Martin étant déjà des communes mal classées par rapport à la langue. Le cluster alémanique est légèrement plus homogène, car nous trouvons cinq communes dépassant les 118 de densités :

- Visp : 598 hab/km^2
- Lalden : 515 hab/km^2
- Brig-Glis : 348 hab/km^2
- Bitsch : 169 hab/km^2
- Salgesch : 133 hab/km^2

Au vu du graphique, la densité est clairement un facteur qui divise nos deux clusters et qui peut être un candidat pour l'interprétation de ce classement en cluster. Cependant, le nombre de communes mal classées est plus important qu'avec le critère de la langue. Nous avons onze communes pour la densité contre cinq au maximum pour la langue. À noter que lorsque nous regardons uniquement le super-espace, alors la langue classe parfaitement les communes.

Passons maintenant au deuxième graphique de la Figure 30. Celle-ci représente la distribution des âges des communes dans trois classes : moins de 20 ans, entre 20 et 64 ans et plus de 64 ans. La forme de ce graphique est sans surprise en pyramide. Les deux clusters sont composés de communes avec une population majoritairement dans la classe du milieu. En plus de présenter la même forme, la répartition est sensiblement la même dans les deux clusters. Nous constatons une légère différence avec un peu plus de personnes âgées et moins de jeunes dans le cluster alémanique, mais rien de significatif. Nous pouvons alors affirmer que la différence de comportement de vote montrée par les deux clusters s'explique très peu par la distribution des âges dans les communes. Ce critère est à éliminer totalement.

Regardons le dernier graphique sur la répartition en secteur d'activité de la Figure 30. Celui-ci va nous permettre de comparer nos clusters en fonction de la répartition du secteur primaire, secondaire et tertiaire. Nous pouvons observer que les deux graphiques ont la même forme en escalier. L'activité économique des deux clusters est principalement concentrée dans le secteur tertiaire. Cependant, le secteur primaire approche les 20% pour le cluster alémanique alors qu'il est plus proche des 10% dans le cluster francophone. Le secteur secondaire est lui sensiblement le même. Pour tenter de comprendre nos comportements de votes à travers le critère du secteur d'activité, nous devons nous intéresser au secteur primaire. Afin de trouver les communes mal placées, nous allons rechercher les communes francophones avec un secteur d'activités primaire s'approchant des valeurs du cluster alémanique. Voici la liste des communes :

- Liddes : 35%
- Fully : 32%
- Miège : 32%
- Chamoson : 32 %
- Isérable : 30 %
- Collonges : 27 %
- Evolène : 22 %
- Venthône : 22 %
- Saillon : 18%

Nous chercherons l'inverse dans le cluster alémanique :

- Zermatt : 0.58%
- Visp : 0.7%
- Saas-Fee : 0.7%
- Brig-Glis : 0.8%
- Fiesch : 1.2%
- Leukerbad : 2.1%
- Bettmeralp : 3.7%
- Steg-Hohtenn : 4.0%
- Tasch : 4.8%
- Bitsch : 6.4%
- Bellwald : 6.4%
- Mörel-Filet : 7.1 %
- Grächen : 7.1 %
- Saas-Grund : 8.6%
- Naters : 8.7%
- St. Niklaus :9%
- Lalden : 10%
- Saas-Almagell : 10%
- Stalden : 10%
- Riederalp 11%
- Leuk : 11%

Nous constatons cette fois bien plus de communes mal-classées. De plus, il y en a davantage dans le cluster alémanique, avec 21 communes contre 9 dans le francophone. Bien que des différences puissent être constatées sur les graphiques, les secteurs d'activités, comme la densité, expliquent d'une moins bonne manière que la langue la création de nos deux clusters.

Nous confirmons, d'une autre manière, l'influence de la langue dans la création des clusters en soulignant l'impact moins important des autres critères tels que l densité, l'âge et les secteurs économiques dans la composition de deux clusters.

4.1.2.3 Analyses des clivages intracluster

Bien que la langue ressorte comme le facteur qui explique le mieux la composition de nos clusters, les critères, densités et secteurs économiques se sont révélés intéressants. En effet, il existe une différence notable constatée sur les graphiques. De plus, l'analyse des communes des clusters à travers ces critères laisse imaginer des sous-groupes à nos deux clusters.

Afin d'étudier cet aspect plus en détail, nous allons lancer notre algorithme de classement sur notre super-espace avec un $k = 4$. D'une part, nous voulons observer si la langue reste homogène dans les quatre clusters et nous voulons aussi étudier si des séparations autour de la densité et des secteurs économiques apparaissent entre deux clusters de même langue.

Nous commençons par projeter nos données sur la carte du valais afin de confirmer que les langues des clusters restent homogènes.

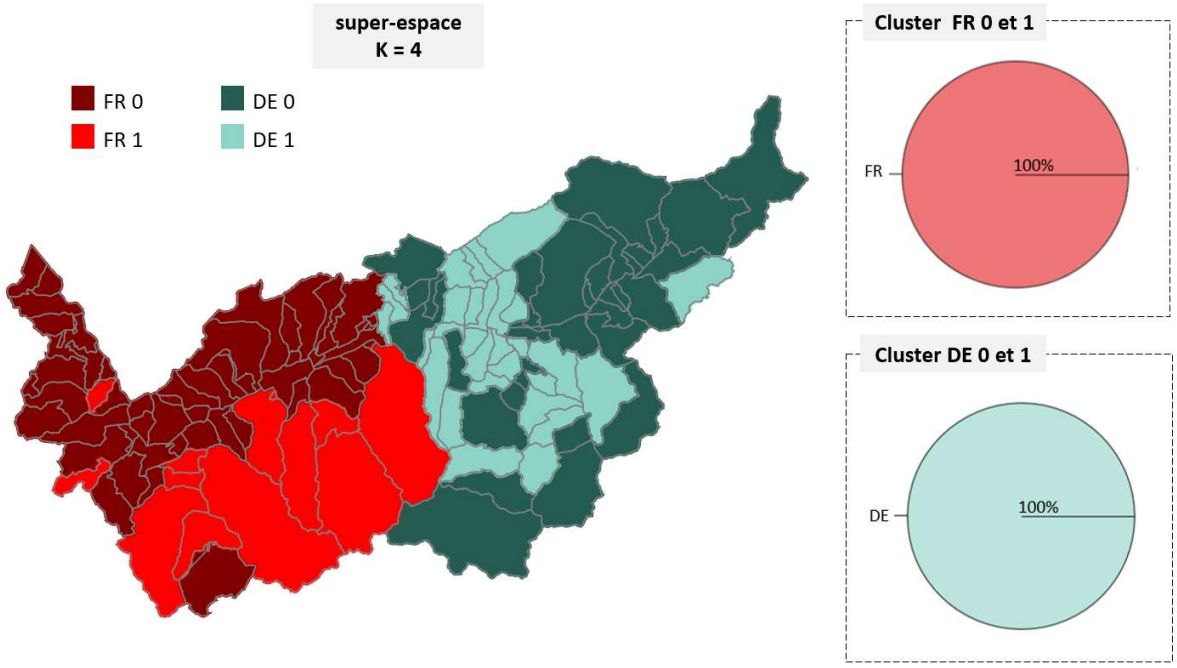


Figure 31 Projection des clusters avec un k =4 sur la carte du Valais et répartition des langues

La Figure 31 nous montre que la séparation en 4 groupes de comportements de votes conserve la séparation parfaite des langues pour les communes. Nous pouvons remarquer une séparation géographique très nette entre les deux clusters francophones. Pour ce qui concerne les deux clusters alémaniques, nous ne constatons pas de séparation géographique claire. Nous pouvons observer qu'il y a deux comportements de vote distinct dans les communes francophones et deux comportements distincts pour les communes alémaniques.

Maintenant que nous avons confirmé que la langue demeure un facteur dans la séparation de nos clusters avec une valeur de 4, nous pouvons rechercher quels facteurs justifient la séparation interne aux communes francophones et alémaniques. Pour cela, nous répétons l'opération faite pour la séparation en deux clusters. Nous comparons cette fois nos quatre clusters avec nos données annotées.

Commençons par observer la répartition des densités de population dans nos quatre clusters.

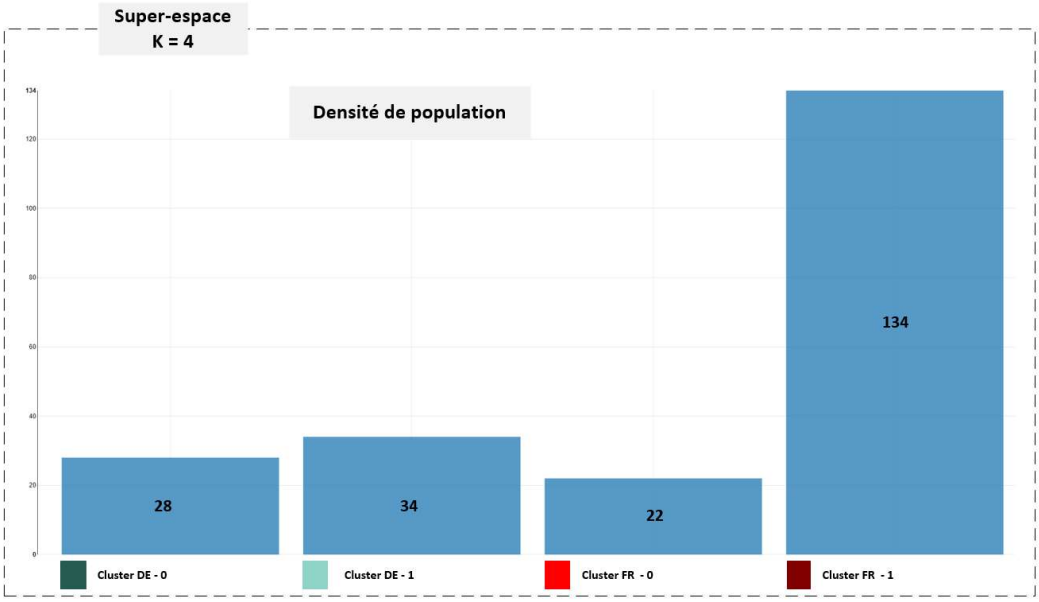


Figure 32 Répartition de la densité de population dans les quatre clusters

La Figure 32 présente la répartition de la densité dans nos quatre clusters. Ce graphique est intéressant pour plusieurs raisons. Tout d'abord, nous pouvons constater que le cluster avec la plus petite densité est un cluster francophone et ceci confirme que, malgré le graphique de la Figure 30 affichant une différence lors d'une séparation en deux clusters, la langue est le meilleur facteur pour expliquer cette première séparation en deux groupes. Ensuite, nous constatons que la densité dans les deux clusters alémaniques est repartie de façon homogène avec des valeurs proches et ce n'est donc pas un bon facteur explicatif des deux comportements de vote différents. Enfin, la séparation dans le comportement de vote pour les deux clusters francophones est explicable avec le facteur de la densité d'habitation avec une très nette différence entre les deux groupes.

Étudions maintenant nos clusters au travers des facteurs de secteurs économiques. Cette nouvelle répartition est affichée sur la Figure 33.

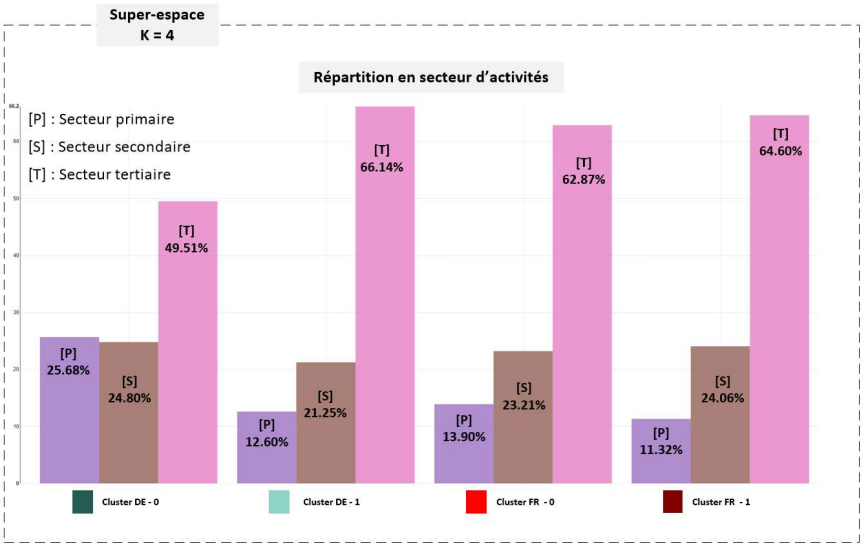


Figure 33 Répartitions en secteur d'activités des quatre clusters

Plusieurs points sont à relever dans ce nouveau graphique. En premier lieu, il n'existe pas de forte différence entre les deux clusters francophones. Ceux-ci disposent d'une répartition en secteur économique similaire, la séparation de ces deux clusters ne peut pas être expliquée en invoquant ce facteur. Ensuite, nous avons un cluster alémanique avec une forme atypique. En effet, le cluster 0 n'est pas en forme d'escalier, car le secteur primaire dépasse le secteur secondaire. Le second cluster alémanique est lui de la même forme que les deux autres. Nous pouvons alors indiquer que les secteurs économiques peuvent expliquer la distinction dans les clusters alémaniques.

En conclusion, les analyses effectuées sur les clusters obtenus avec une séparation en quatre groupes nous ont permis d'analyser les comportements de votes plus en détail. Premièrement, nous constatons que la première séparation la plus importante est faite autour du critère de la langue, nous avons les communes francophones et les communes alémaniques séparées. Puis, à l'intérieur des communes francophones, nous pouvons constater deux comportements distincts qui séparent les communes à forte densité de population de celle à faible densité. Cette séparation par la densité de population n'est pas applicable aux communes alémaniques. Cependant, deux comportements de vote différents existent également du côté alémanique. Mais ceux-ci s'apparentent plus à un vote de classe, car il divise les communes en deux groupes avec d'un côté les communes avec un secteur primaire dépassant le secteur secondaire et de l'autre les communes avec un secteur primaire faible.

4.1.2.4 Influence des partis politiques

Dans cette partie, nous allons observer et discuter la répartition des communes sur les deux axes du PCA. Nous regarderons la contribution de chaque variable au PCA et annoterons ensuite des vecteurs de partis politiques sur le graphique du PCA afin de visualiser les forces politiques qui s'opposent dans le comportement de votes. Nous parcourrons les résultats des cinq années.

Pour commencer, nous prenons la liste des candidats pour l'ensemble des années d'élections et nous annotons chaque personnalité avec le parti politique auquel il est lié. Ceci va nous permettre de créer les vecteurs de partis politiques dans nos représentations en deux dimensions des votes des communes.

Nom	Partis politiques	Année(s) de candidature(s)
Jean-René Fournier	PDC	2001, 2005
Wilhelm Schnyder	PDC	2001
Jean-Jacques Rey-Bellet	PDC	2001, 2005
Thomas Burgener	PS	2001, 2005
Claude Roch	PRD – PLR	2001, 2005, 2009
Cilette Cretton-Deslarzes	PRD – PLR	2001
Michel Carron	Indépendant	2001, 2005
Jean-Michel Cina	PDC	2005, 2009, 2013
Georges Darbellay	Vert	2005
Ignace Rey	Indépendant	2005
Maurice Tornay	PDC	2009, 2013
Jacques Melly	PDC	2009, 2013, 2017
Esther Waeber Kalbermatten	PS	2009, 2013, 2017
Franz Ruppen	UDC	2009
Marylène Volpi Fournier	Vert	2009
Graziella Walker Salzmann	Chrétienne-Sociale	2009
Eric Felley	Indépendant	2009
Oskar FREYSINGER	UDC	2013, 2017
Christian VARONE	PLR	2013
Christophe CLIVAZ	Vert	2013
Christophe Darbellay	PDC	2017
Roberto Schmidt	PCS	2017
Stéphane Rossini	PS	2017
Nicolas Voide	PDC	2017

Jean-Michel Bonvin	PCS	2017
Frédéric Favre	PLR	2017
Sigrid Fischer-Willa	UDC	2017
Thierry Largey	Vert	2017
Jean-Marie Bornet	RCV	2017
Claude Pottier	PLR	2017

Les graphiques ci-dessous présentent en deux dimensions l'ensemble des votes des communes pour les cinq années d'élections

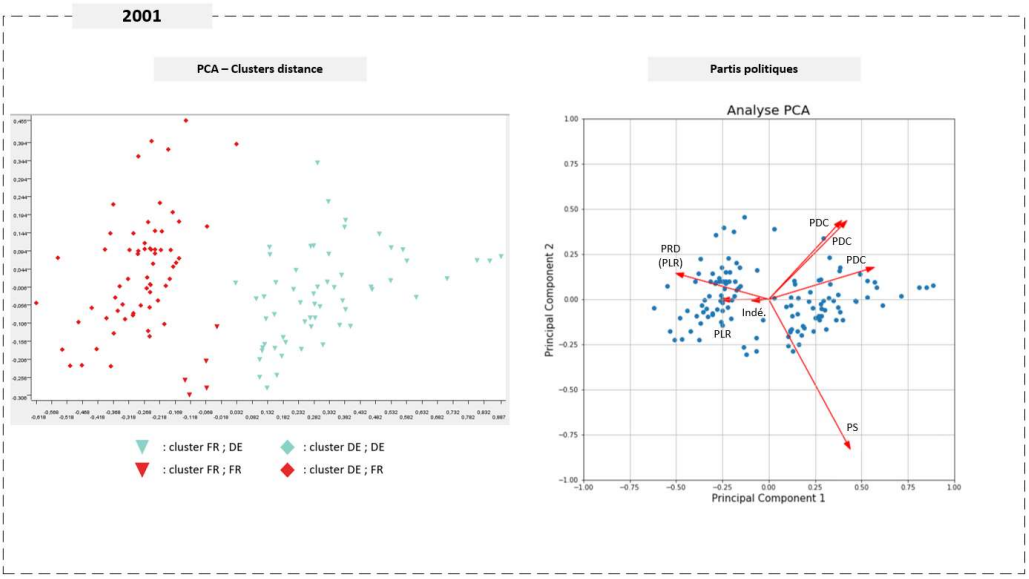


Figure 34 Vecteurs des partis politiques pour les comportements de vote de 2001

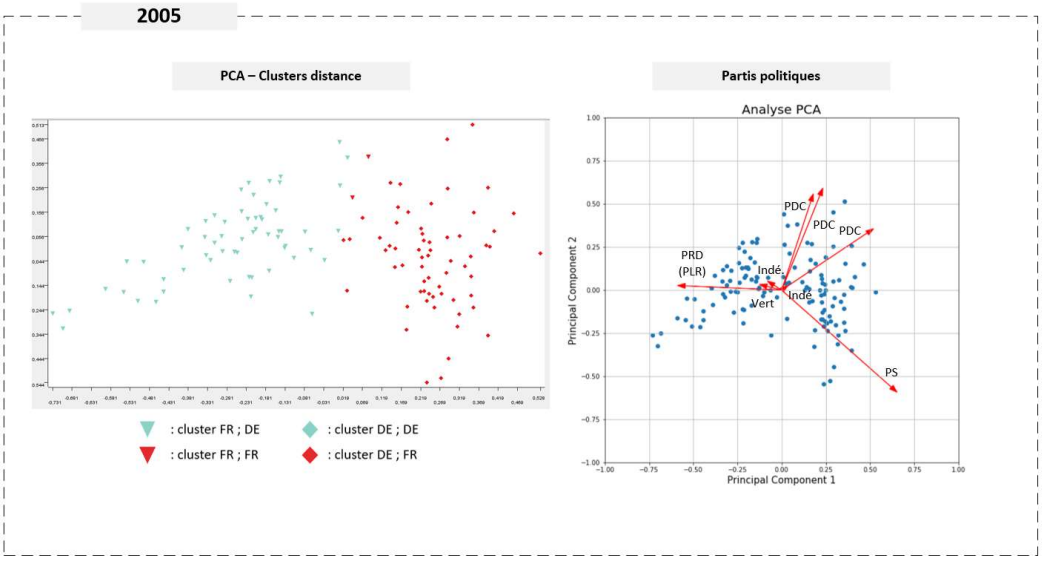


Figure 35 Vecteurs des partis politiques pour les comportements de vote de 2005

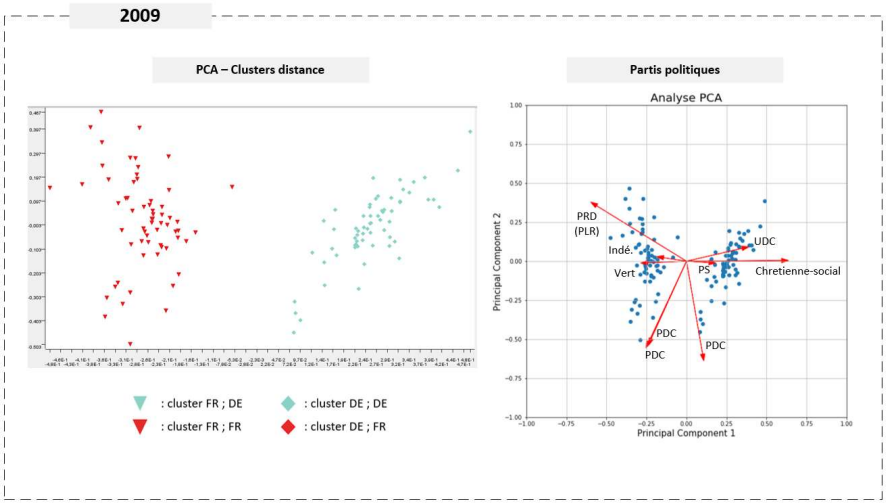


Figure 36 Vecteurs des partis politiques pour les comportements de vote de 2009

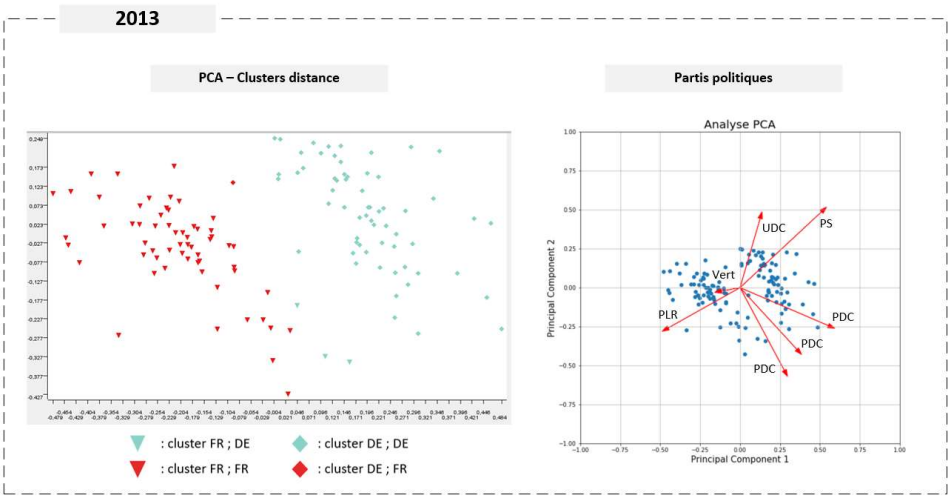


Figure 37 Vecteurs des partis politiques pour les comportements de vote de 2013

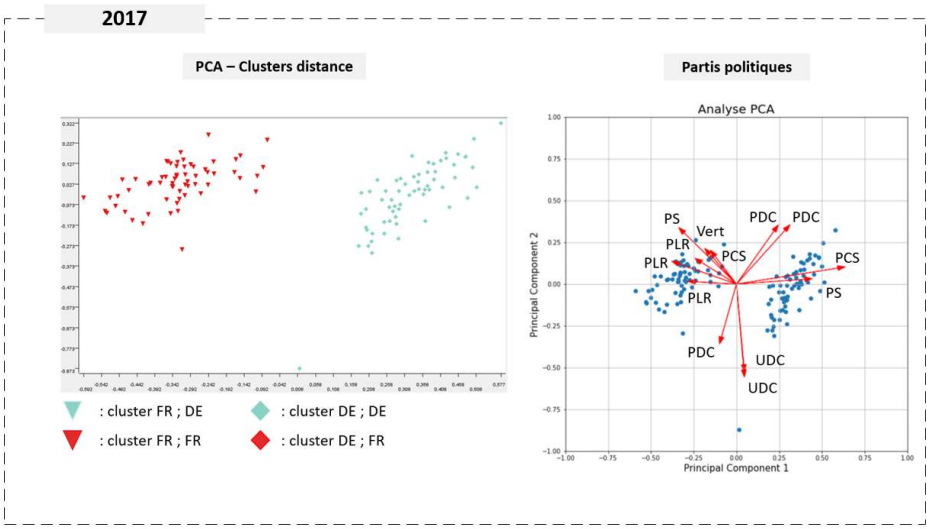


Figure 38 Vecteurs des partis politiques pour les comportements de vote de 2017

Le graphique de droite affiche les vecteurs de contribution dans les deux axes. La longueur et la direction indiquent l'intensité et le sens de cette contribution. Nous avons annoté pour chaque vecteur – qui correspond à un candidat – le parti politique associé.

En observant les graphiques, nous constatons que les vecteurs de mêmes partis ont la même direction. Nous pouvons en déduire que les partis politiques jouent également un rôle dans les comportements de votes des communes. Nous pouvons également observer que sur l'ensemble des graphiques il existe une opposition entre PDC et PLR. Le parti politique PLR influence la position des points dans la direction des communes francophones alors que le PDC influence dans la direction des communes alémaniques. Ceci nous laisse supposer que dans le comportement de votes des communes alémaniques plus de voix sont attribuées à des candidats PDC par rapport aux francophones. Pour les communes francophones, le comportement de votes indique plus de voix données au PLR par rapport aux Alémaniques. L'UDC influence également plus du côté alémanique alors que le vecteur du parti vert à une direction orientée vers les francophones. Étonnamment, nous pouvons aussi constater que le PS est orienté régulièrement dans la même direction que l'UDC vers les communes alémaniques.

Nous avons regardé l'influence de partis dans les comportements de votes et afin de confirmer les résultats relevés nous allons étudier la distribution des voix pour chaque parti politique lors des élections. Pour cela, nous utilisons notre super-espace afin de résumer l'ensemble des cinq élections et calculons pour chaque vecteur de parti politique le pourcentage de bulletin reçu. Nous affichons sur la Figure 39 les résultats par clusters.

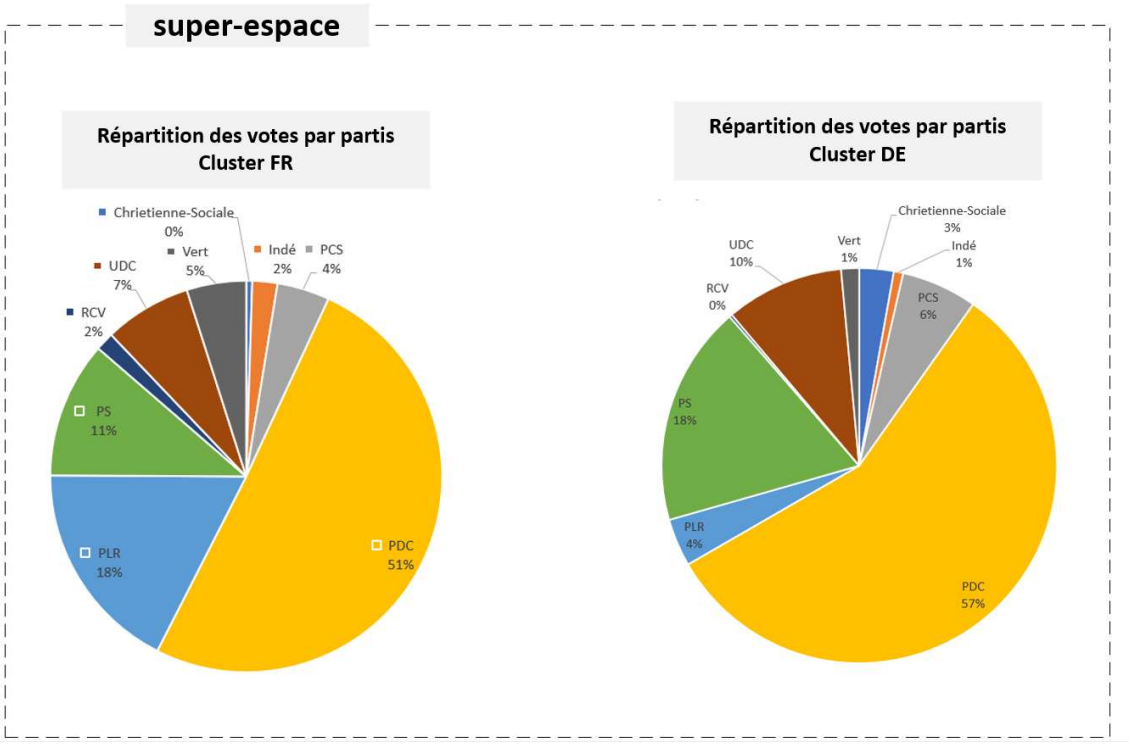


Figure 39 Répartition des bulletins de vote par partis politiques pour l'ensemble des années et pour les deux clusters

Nous pouvons confirmer que le PLR obtient de meilleurs résultats dans les communes francophones alors que le PDC est meilleur dans les communes alémaniques. Notons également que même si les PLR performent mieux dans le cluster FR, le PDC reste le parti en tête de ce classement dans les deux clusters. Nous pouvons également observer la différence pour les partis verts, PS et UDC. Le premier étant plus haut chez les francophones alors que les deux autres obtiennent de meilleurs résultats cotés alémaniques.

Bien que l'influence des partis ne soit pas aussi lisible que la langue pour expliquer la différence entre nos deux clusters, nous constatons tout de même des différences qui peuvent expliquer en partie la séparation en deux

comportements de vote. Les deux clusters votent majoritairement pour le PDC, mais les communes francophones votent plus pour le PLR ainsi que pour les verts alors que les communes alémaniques votent, après le PDC, pour le PS et l'UDC.

4.2 ANALYSE PREDICTIVE

Dans cette dernière partie du chapitre sur les résultats et analyses, nous allons parler de la mise en place des quatre approches pour appliquer de l'analyse prédictive sur nos données des élections au Conseil d'État valaisan. Nous décrirons la mise en place technique, discuterons des résultats et nous prendrons du recul afin de mesurer la réelle capacité de prédictions de nos modèles. Nous commencerons avec l'approche par distance vectorielle puis nous regarderons les classements de communes par arrangement et combinaisons. Enfin, nous finirons le chapitre par l'entrainement et l'évaluation de nos arbres de décisions.

4.2.1 Distance vectorielle

L'ensemble des votes des communes ainsi que le résultat de l'élection peuvent être représentés par un vecteur dans un nombre de dimensions égales au nombre de candidats. Dans cette partie, nous allons chercher à calculer la distance de chaque vecteur de commune par rapport au vecteur de résultat. Nous utiliserons la distance euclidienne comme méthode pour mesurer la séparation entre nos vecteurs. Pour calculer le vecteur de résultat, nous utilisons la même technique que pour les vecteurs des communes avec une normalisation des données avec le nombre total de bulletins. L'ensemble des vecteurs représente la distribution entre candidats des bulletins de vote.

Le flux global des traitements effectués est visible sur la Figure 40 :

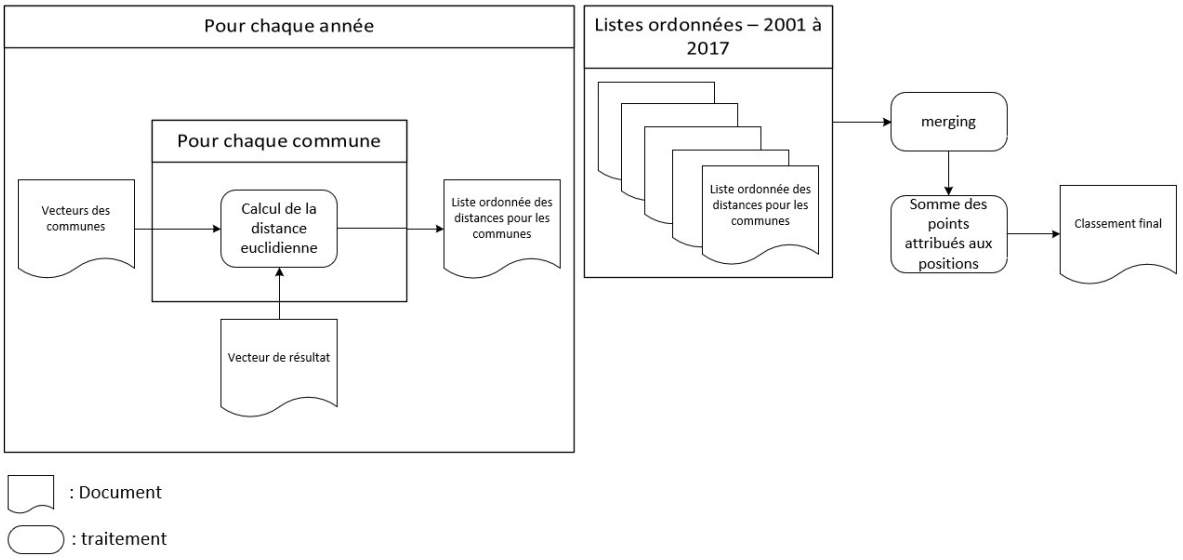


Figure 40 Flux globaux des traitements pour l'approche par distance vectorielle

Nous commençons par calculer la distance euclidienne des vecteurs de communes avec le vecteur de résultat pour nos cinq années d'élections. Pour cela, nous utilisons un script python avec comme output les listes ordrées des distances de chaque commune pour les cinq années. La Figure 41 affiche les dix premières lignes de ces listes pour l'ensemble des années avec la position dans la liste de chaque commune.

	2001			2005			2009			2013			2017	
1	Grimisuat	0,0832825	1	Veyras	0,0520308	1	Lens	0,1402947	1	Vionnaz	0,0794475	1	Crans-Montana	0,139968
2	Savièse	0,1018507	2	Mont-Noble	0,0755536	2	Savièse	0,143203	2	Ayent	0,084979	2	Sierre	0,146021
3	Saint-Léonard	0,1070631	3	Crans-Montana	0,0800105	3	Vétroz	0,1650545	3	Sierre	0,0937907	3	Anniviers	0,1682687
4	Mont-Noble	0,1074472	4	Evionnaz	0,1076597	4	Saillon	0,165752	4	Mont-Noble	0,0976825	4	Chippis	0,175228
5	Nendaz	0,1118983	5	Saint-Léonard	0,1092506	5	Fully	0,1686157	5	Massongex	0,1013598	5	Saint-Léonard	0,1784428
6	Sion	0,1336892	6	Nendaz	0,1138378	6	Crans-Montana	0,1724719	6	Saillon	0,1046445	6	Sembrancher	0,1825318
7	Arbaz	0,1434414	7	Sierre	0,1199933	7	Sierre	0,1746159	7	Vernayaz	0,1048851	7	Savièse	0,1855502
8	Ayent	0,1476786	8	Ayent	0,1309528	8	Vex	0,1830715	8	Grône	0,1095241	8	Ardon	0,1870531
9	Crans-Montana	0,1545353	9	Sion	0,1320709	9	Chamoson	0,1862635	9	Crans-Montana	0,1109875	9	Lens	0,188178
10	Evionnaz	0,1614851	10	Grimisuat	0,1347948	10	Ardon	0,1864494	10	Fully	0,1147546	10	Fully	0,1943514

Figure 41 Listes ordonnée des distances vectorielles pour les cinq années d'élection

Ensuite, nous pouvons passer à la deuxième partie de nos traitements affichés sur la partie de droite de la Figure 40. Il s'agit de réunir ces cinq listes et de calculer le score de chaque commune. Le score d'une commune est la somme de ces positions. Par exemple, une commune arrivée 1ère, 2ème, 5ème, 10ème et 3ème aura un score de 21. Plus le score est faible, plus la commune est globalement proche du vecteur de résultat sur les cinq années. Nous devons alors classer cette dernière liste du plus petit score au plus grand afin d'obtenir notre classement final. Voici les dix communes arrivées en tête de notre classement par distance vectorielle.

Commune	Score
Crans-Montana	28
Sierre	38
Saint-Léonard	55
Mont-Noble	60
Vernayaz	67
Ardon	72
Chamoson	75
Saillon	78
Fully	81
Bovernier	83

Figure 42 Classements finaux des distances vectorielles des communes

Sur la Figure 42, nous observons que Crans-Montana arrive en tête avec un score de 28. Sierre arrive en seconde place avec un delta de 10 par rapport à la première place. Afin de mesurer la signification du chiffre 28, nous pouvons analyser les classements par années et ainsi déterminer les cinq positions de la commune de Crans-Montana.

	2001	2005	2009	2013	2017
1	Grimisuat	Veyras	Lens	Vionnaz	Crans-Montana
2	Savièse	Mont-Noble	Savièse	Ayent	Sierre
3	Saint-Léonard	Crans-Montana	Vétroz	Sierre	Anniviers
4	Mont-Noble	Evionnaz	Saillon	Mont-Noble	Chippis
5	Nendaz	Saint-Léonard	Fully	Massongex	Saint-Léonard
6	Sion	Nendaz	Crans-Montana	Saillon	Sembrancher
7	Arbaz	Sierre	Sierre	Vernayaz	Savièse
8	Ayent	Ayent	Vex	Grône	Ardon
9	Crans-Montana	Sion	Chamoson	Crans-Montana	Lens
10	Evionnaz	Grimisuat	Ardon	Fully	Fully

Figure 43 Positions dans les listes par année de la commune de Crans-Montana

Nous constatons sur la Figure 43 que le chiffre 28 de la commune de Crans-Montana est composé des positions suivantes : 9, 3, 5, 9 et 1. Nous pouvons affirmer les points suivants pour la commune de Crans-Montana. Premièrement, celle-ci est toujours dans les dix premières communes. Ensuite, elle arrive deux fois sur le podium avec une 3ème place en 2005 et une 1ère place en 2017. De plus, la position la plus mauvaise pour Crans-Montana est une 9ème place. Enfin, il s'agit de la commune avec le plus petit score global.

En conclusion, notre approche par distance vectorielle fait ressortir la commune de Crans-Montana comme un bon indicateur de la répartition des bulletins entre les candidats à l'échelle cantonale. En effet, le vecteur de Crans-Montana est en moyenne le plus proche du vecteur final. De plus, les positions de la commune varient faiblement, avec un maximum de 9 et un minimum de 1. À noter que cette tendance est observée uniquement sur cinq élections.

4.2.2 Arrangement et combinaison

L'approche par distance vectorielle nous donne de bonnes indications sur la distribution des bulletins, mais n'apporte pas directement d'information sur les personnes élues. Nous allons regarder les approches par arrangement et combinaison afin de mesurer la capacité des communes à prédire un ensemble d'élus. Pour cela, nous devons faire plusieurs traitements sur nos données. Afin de les visualiser, la Figure 44 présente le flux global des actions réalisées.

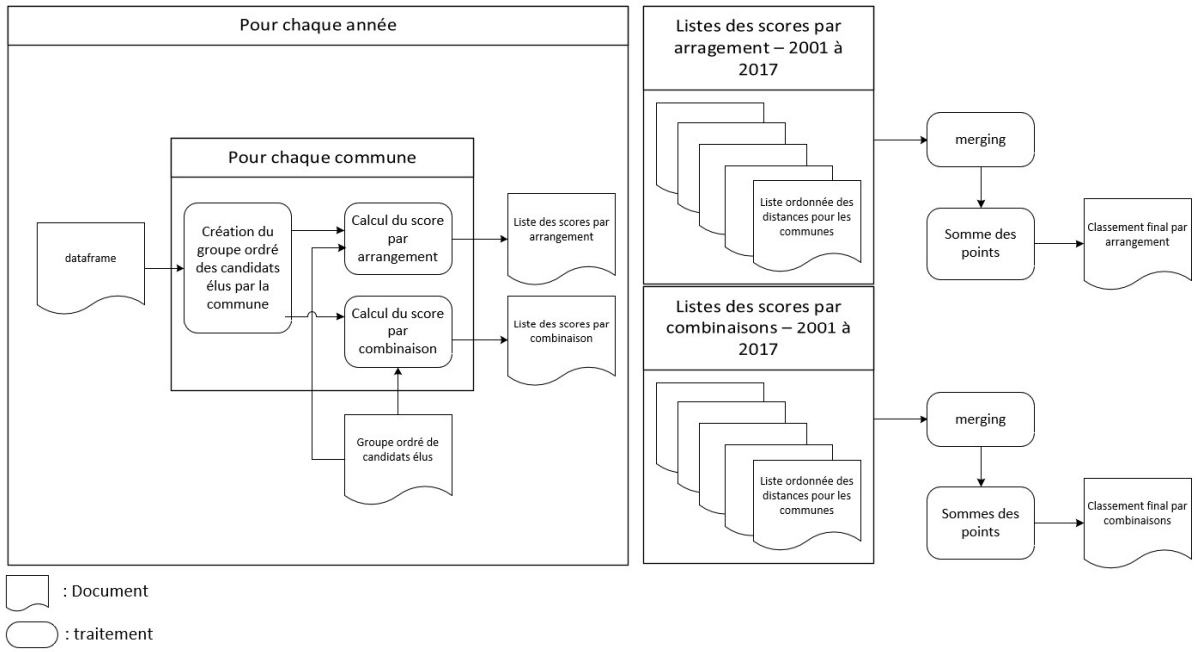


Figure 44 Flux des traitements pour l'approche par arrangement et combinaison

La première étape consiste à créer un groupe ordré des candidats élus pour chaque année. Pour cela, nous regardons le résultat final de l'élection et créons une liste ordonnée des candidats élus. La Figure 45 présente la liste des ensembles ainsi créés pour les cinq années.

```
V_result_2001 = ['Jean-René Fournier', 'Wilhelm Schnyder', 'Jean-Jacques Rey-Bellet', 'Thomas Burgener', 'Claude Roch'];
V_result_2001 = [3, 2, 4, 0, 1];
V_result_2005 = ['Jean-René Fournier', 'Jean-Jacques Rey-Bellet', 'Jean-Michel Cina', 'Thomas Burgener',  'Claude Roch'];
V_result_2005 = [2, 3, 4, 5, 0];
V_result_2009 = ['Jean-Michel Cina', 'Maurice Tornay', 'Jacques Melly', 'Claude Roch', 'Esther Waeber Kalbermatten'];
V_result_2009 = [2, 4, 3, 0, 8];
V_result_2013 = ['Oskar FREYSINGER', 'Esther WAEBER-KALBERMATTEN', 'Jean-Michel CINA', 'Jacques MELLY', 'Maurice TORNAY'];
V_result_2013 = [6, 5, 0, 1, 2];
V_result_2017 = ['Roberto SCHMIDT', 'Jacques MELLY', 'Christophe DARBELLAY', 'Esther WAEBER-KALBERMATTEN', 'Frédéric FAVRE'];
V_result_2017 = [12, 10, 11, 7, 4];
```

Figure 45 Groupes ordrés de candidats élus pour les cinq années

Ensuite, nous devons créer pour chaque commune le classement des candidats par rapport au nombre de bulletins reçus. Avec cette opération, nous pouvons comparer nos ensembles des communes avec celui du résultat pour les cinq années. L'attribution de points pour les communes est faite de deux manières suivantes.

Pour l'approche par combinaison, nous attribuons un point lorsqu'un des éléments dans l'ensemble de la commune est présent (sans se soucier de l'ordre) dans le groupe du résultat. Une commune peut avoir entre 0 et 5 points par année. Autrement expliqué, nous regardons les candidats élu à l'échelle de la commune et attribuons un point par candidats réellement élus lors du résultat final.

Pour l'approche par arrangement, nous attribuons aussi entre 0 et 5 points, mais uniquement si le candidat est à la même place que celui du résultat. En somme, nous regardons les candidats élus à l'échelle de la commune et attribuons un point uniquement si le candidat arrive à la même place que lors du résultat final.

Finalement, pour les deux approches, nous regroupons toutes les années et classons les résultats par ordre décroissant. La commune avec le plus de points est celle le plus capable de prédire la combinaison du résultat ou l'arrangement du résultat.

L'ensemble de ces opérations est effectué à l'aide d'un script Python et de fichiers Excel pour l'affichage ordré des listes.

Regardons maintenant les résultats obtenus avec ces deux approches. La Figure 46 affiche la liste des dix communes avec le score le plus élevé pour une approche par arrangement.

Commune	Points
Wiler	15
Bettmeralp	13
Obergoms	13
Saillon	13
Chamoson	12
Ferden	12
Fieschertal	12
Gampel-Bratsch	12
Anniviers	11
Arbaz	11

Figure 46 Liste des dix premières communes avec une approche par arrangement

Nous constatons que la commune de Wiler arrive en tête avec un score de 15. Regardons maintenant la répartition des points de chaque année pour arriver à ce score.

- 2001 : 3 points sur 5
- 2005 : 2 points sur 5
- 2009 : 3 points sur 5
- 2013 : 3 points sur 5
- 2017 : 4 points sur 5

En arrivant en tête, la commune de Wiler ne propose pas de très bon score pour l'ensemble des années. En 2005, elle produit un résultat plus faible que 50% de prédiction. À noter qu'elle ne prédit pas systématiquement les mêmes places, il est donc impossible d'attribuer un degré de confiance pour les emplacements. Avec un taux de prédiction proche des 50%, l'approche par arrangement ne nous apporte pas d'information.

Nous pouvons maintenant regarder les résultats d'une approche par combinaison. Cette approche vise à nous donner des informations sur les personnes élues, peu importe si le candidat est élu en première position ou en cinquième position. Regardons à nouveau la liste des dix premières communes, mais cette fois par combinaison.

Commune	Points
Chippis	24
Conthey	23
Eischoll	23
Mont-Noble	23
Oberems	23
Saas-Almagell	23
Saas-Fee	23
Saas-Grund	23
Vex	23
Visperterminen	23

Figure 47 Liste des dix premières communes avec une approche par combinaisons

Nous pouvons constater sur la Figure 47 que la petite commune de Chippis arrive en tête. Le score de 24 est supérieur au score obtenu par la meilleure commune lors d'une approche par arrangement et cela laisse présupposer une meilleure capacité de prédiction. À noter également que les neuf autres communes partagent une deuxième place avec un score de 23. Afin de mesurer la qualité du score de la commune de Chippis, nous regardons l'ensemble de ces prédictions pour les cinq années.

- 2001 : 5 points sur 5
- 2005 : 5 points sur 5
- 2009 : 5 points sur 5
- 2013 : 5 points sur 5
- 2017 : 4 points sur 5

Le score de 24 n'est pas loin d'une prédiction parfaite. En effet, lors de quatre années sur cinq, la commune de Chippis nous permet de prédire à 100% les candidats élus. Lors de l'année 2017, elle prédit quatre candidats sur les cinq élus.

En abandonnant l'ordre du classement pour notre modèle de prédiction, nous sommes capables de trouver des communes avec des taux de prédiction beaucoup plus intéressants. Notamment la commune de Chippis qui s'en sort avec un taux de 96% de prédiction.

Sur la base des données à notre disposition, et sachant que les communes les moins efficaces prédisent la composition des élus à 64% avec un score de 16, il est plus intéressant de faire un sondage sur la commune de Chippis que sur l'ensemble des communes pour déterminer le résultat final des personnes élues.

4.2.3 Arbre de décisions

La dernière partie de ce chapitre est dédiée à l'approche par apprentissage. Nous allons regarder les résultats que nous pouvons obtenir avec la création et l'entrainement d'arbre de décision. L'idée est de fournir des exemples à un algorithme afin qu'il crée un arbre de décision basé sur les liens des valeurs numériques des exemples et leurs outputs. Comme présenté dans notre méthodologie, nous voulons créer des arbres capables de répondre à la question suivante « Sachant le pourcentage de voix obtenu par un candidat, est-ce que celui-ci sera élu ? ». Nous allons devoir entrainer un arbre par commune. Comme pour les autres approches présentées, commençons par regarder le flux logique des traitements affiché sur la Figure 48.

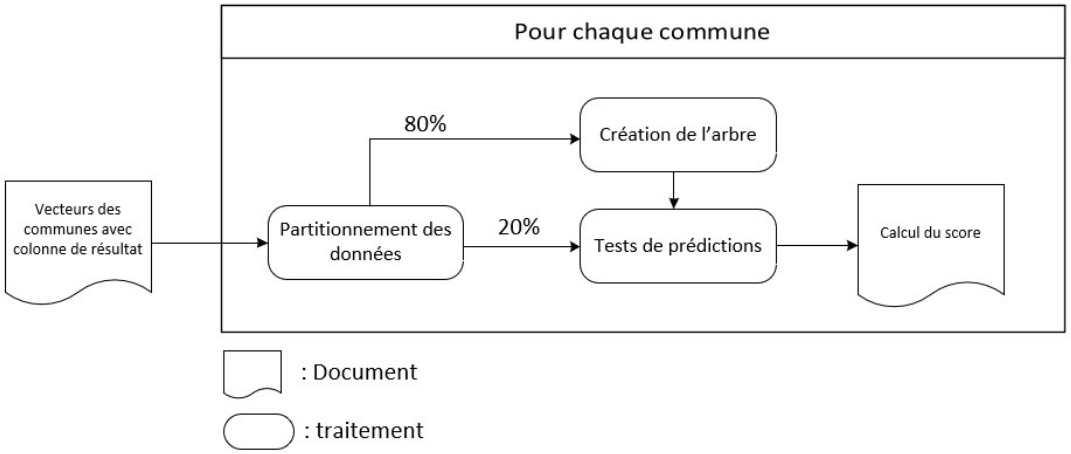


Figure 48 flux des traitements pour la création des arbres de décision

La première étape est de créer la source de données capable d'entrainer nos arbres. Pour cela, nous devons avoir un vecteur par commune qui reprend l'ensemble des bulletins obtenu par le candidat et l'associé avec le résultat « élut » « non élu ».

row ID	Elu	Agarn	Albinen	Anniviers
Thomas Burgener (2001)	1	0,39954338	0,43884892	0,296938776
Claude Roch (2001)	1	0,02511416	0,00719425	0,182653061
Wilhelm Schnyder (2001)	1	0,72146119	0,63309353	0,626530612
Jean-René Fournier (2001)	1	0,6826484	0,58273381	0,659183673
Jean-Jacques Rey-Bellet (2001)	1	0,66438356	0,58992806	0,62244898
Michel Carron (2001)	0	0,02511416	0,02158273	0,092857143
Cilette Cretton-Deslarzes (2001)	0	0,043379	0,16546763	0,28877551
Claude Roch (2005)	1	0,05497382	0,05405405	0,321618743
Georges Darbellay (2005)	0	0,06544503	0,05405405	0,104366347
Jean-René Fournier (2005)	1	0,59947644	0,71621622	0,71884984

Figure 49 Première ligne et colonnes de la matrice créée pour les arbres de décision

Sur la Figure 49, nous pouvons observer la première colonne qui correspond à l'ensemble des candidats qui se sont présentés durant les cinq élections. Si un candidat s'est présenté plusieurs fois alors la matrice contiendra une ligne pour chacune de ces candidatures. Ensuite, la deuxième colonne correspond à l'output de la candidature : 1 pour élu et 0 pour non-élu. Puis, les colonnes suivantes sont les résultats obtenus par le candidat dans chaque commune. La colonne deux est utilisée par tous les arbres en association avec un vecteur de commune. La matrice est composée alors de 44 lignes de résultats pour les 127 communes. Après partitionnement, nous disposons de 35 exemples et 9 tests de validations pour chacun de nos arbres.

Pour la création des arbres, nous avons utilisé le logiciel KNIME et créé le flux présenté sur la Figure 50.

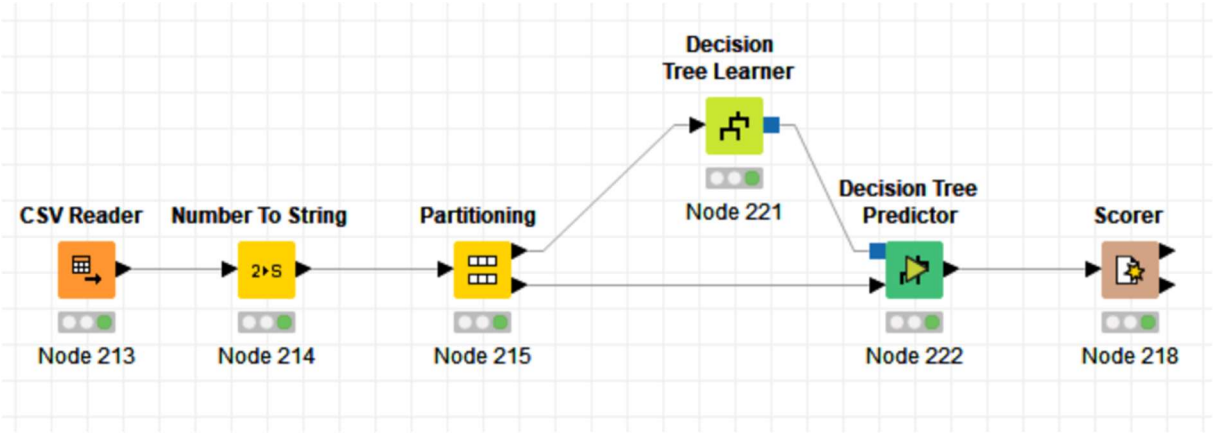


Figure 50 Flux de création des arbres de décisions sur le logiciel KNIME

Avec ces éléments, nous pouvons commencer à créer nos arbres de décisions pour ensuite tester leurs capacités de prédiction sur la partie test de nos données.

Nous avons testé cette approche avec l'ensemble des données. Tous les arbres crées sur la base des exemples des communes, décident de fournir « élu » comme output dans 100% des cas. La Figure 51 présente les résultats des prédictions de l'arbre de décisions entrainées sur les données de la commune de Crans-Montana.

Row ID	S Elu	S Crans-...	S Prediction (Elu)
Jean-RenÃ© Fournier (2001)	1	0,496107573	1
Jean-Jacques Rey-Bellet (2001)	1	0,44621373	1
Michel Carron (2001)	0	0,108987969	1
Ignace Rey (2005)	0	0,058043118	1
Eric Felley (2009)	0	0,115057915	1
Jacques Melly (2013)	1	0,362059621	1
Christian Varone (2013)	0	0,246883469	1
Thierry Largey (2017)	0	0,150672085	1
Christophe Darbellay (2017)	1	0,386058143	1

Figure 51 Prédictions sur le jeu de tests pour la commune de Crans-Montana

En d'autres termes, l'ensemble des arbres, sur la base des 35 exemples, ont déterminé que la meilleure façon de prédire une élection d'un candidat selon sont pourcentage dans une commune et de le prédire tout le temps comme élu. Il n'y a donc aucun apprentissage intéressant effectué par les arbres sur la base des exemples fournis. Ce résultat peut être expliqué par plusieurs facteurs. Premièrement, nous disposons uniquement de 35 exemples à présenter à l'algorithme et visiblement ce chiffre est trop faible pour déterminer des patterns intéressants. De plus, dans nos exemples, le nombre de candidats élus est très important, car pour chaque élection il y a cinq personnes élues. L'élection de 2001 présente uniquement sept candidats pour cinq places.

En conclusion, l'approche par arbre de décision ne nous permet pas de créer un modèle capable de fournir des informations utiles sur la base des données à notre disposition.

Chapitre 5. CONCLUSION

5.1 CONCLUSION

Nous avons commencé par souligner la tendance de nombreux gouvernements à lancer des projets d'open datas. Puis, nous avons discuté de la stratégie dans ce domaine de la Suisse. Cette mise à disposition au public de données, nous a amené à nous intéresser aux apports de la data science et en particulier pour les sciences politiques. Plus précisément, il s'agissait de déterminer les apports des méthodes d'analyses et de prédictions de la datas science dans le domaine des sciences politiques.

Pour cela, nous avons démarré par un état de l'art sur l'analyse de données ainsi que de la prédiction pour des cas de votations et d'élections. Nous avons décrit les méthodologies utilisées pour ces projets ainsi que les différentes techniques employées. De plus, nous nous sommes intéressés aux résultats obtenus par chacune de ces études. Nous avons également souligné l'importance des open data pour améliorer la recherche et les projets dans le domaine des datas science.

Sur la base de cette revue de littérature, nous avons proposé une méthodologie pour l'analyse et la prédiction de situation politique. Pour l'analyse des comportements électoraux, nous avons choisi une approche par classification de données brutes et nous avons proposé une méthode pour sélectionner le bon nombre de groupe. Une fois les groupements faits sur la base uniquement des votes, la méthode proposée a été d'annoter ces groupes afin de tenter d'expliquer cette différence dans les comportements. Enfin, nous avons montré différentes visualisations utilisables pour faciliter l'analyse des situations politiques.

La suite du document a été consacrée aux méthodes de prédictions. Dans notre revue de littérature, nous avons parlé de différentes recherches utilisant de l'analyse prédictive dans le domaine des sciences politiques. Sur cette base, nous avons proposé trois approches utilisables pour de la prédiction d'élection de personnes.

Après la description de cette méthodologie, nous avons sélectionné l'élection au Conseil d'État valaisan afin de la mettre en pratique. Pour cela, il nous a fallu passer par plusieurs étapes. Premièrement, nous avons récolté les données des cinq dernières élections. Sur ces données, nous avons appliqué les prétraitements nécessaires à l'application des techniques décrites dans notre méthodologie, notamment avec une approche pour la gestion des fusions entre communes. Une fois cela fait, nous avons appliqué nos algorithmes de partitionnement, annoté nos données et commenté les résultats. Finalement, nous avons expérimenté et commenté nos approches de prédictions sur ce cas pratique.

Lors de l'étude de cas du Conseil d'État valaisan, nous pouvons souligner les résultats intéressants suivants.

Tout d'abord, aucun des autres facteurs sélectionnés, y compris les partis politiques, n'explique mieux la séparation des comportements que la différence de langue. En effet, la langue reste le facteur qui annote le mieux les groupes créés avec des taux variant entre 92% et 100% des communes correctement placées.

Ensuite, nous avons montré, avec une séparation en quatre clusters, que la densité peut expliquer une différence de comportement à l'intérieur du cluster francophone. Et pour la séparation dans le cluster alémanique, celle-ci est explicable par la répartition des secteurs économiques.

De plus, nous avons montré avec l'analyse vectorielle que la commune de Crans-Montana était en moyenne plus proche de la répartition des bulletins à l'échelle du canton que les autres communes. Pour la prédiction de l'élection d'un candidat, nous avons pu montrer que l'approche par arrangement ne permettait pas d'obtenir plus d'information des données. Cependant, lorsque nous tenons plus compte de l'ordre des données et que l'on s'intéresse à la combinaison alors la petite commune de Chippis permet une prédiction à 96% du résultat final. Cette valeur est intéressante, car d'autres communes prédisent ce résultat avec des scores bien plus faibles. Il devient alors intéressant de s'intéresser à cette commune plutôt qu'à l'ensemble des communes pour prédire les cinq élus au Conseil d'État. Nous sommes conscients, qu'avec le peu de données utilisées, il faille confirmer cette tendance sur plus de données afin de consolider ces approches.

Finalement, nous avons démontré qu'il n'était pas possible d'appliquer sur nos données la même approche que les communes « oracles » de l'étude de Suisse sur les votations populaires. En effet, les arbres de décisions créés ne capturent aucune information et prédisent très mal le résultat des élections. La quantité de données et le rapport entre le nombre d'élus et le nombre de candidats rendent ces techniques inefficaces pour apporter de nouvelles informations.

À la fin de notre cas d'étude et sur la base des résultats, nous sommes capables de répondre à nos questions de recherches.

> « Est-ce que le comportement électoral des citoyens valaisans est marqué par des préférences qui peuvent être liées à la langue des communes ? »

Les comportements des électeurs du valais peuvent être expliqués par la langue des communes. Cette explication est plus robuste qu'une explication basée sur les partis politiques, la densité de population, l'âge des électeurs ou sur leur secteur d'activités économiques.

> « Quels autres facteurs explicatifs peuvent être mobilisés pour tenter de comprendre ces phénomènes ? »

Les partis politiques jouent également un rôle dans la répartition francophone-alémanique. La densité de population peut être utilisée pour tenter de comprendre les différents comportements dans les communes francophones alors que pour les communes alémaniques, il s'agit de regarder la répartition en secteurs d'activités.

> « Peut-on prédire le résultat d'une élection future sur la base d'un échantillonnage de la population ? Comment choisir cette part réduite de la population ? »

Sur la base des données à notre disposition, Crans-Montana est la commune qui prédit le mieux la répartition des bulletins de vote. La commune de Chippis est la meilleure commune pour prédire les cinq personnes élues avec un taux de réussite à 96%.

5.2 AMELIORATIONS ET SUITE

En appliquant notre méthodologie au cas de l'élection du Conseil d'État valaisan, nous avons pu obtenir des résultats intéressants pour l'analyse des comportements de vote ainsi que pour la prédiction de résultat. Cependant, il reste des points que nous pourrions approfondir ou d'autres que nous pourrions traiter.

Tout d'abord, nous pourrions tester notre méthode d'analyse à d'autres cantons. Notamment au canton de Fribourg afin d'observer les similarités et différences entre deux cantons bilingues. Il serait intéressant de sélectionner aussi un autre canton non bilingue afin de voir si d'autres facteurs explicatifs du comportement électoral émergent à la place de la langue. Nous pourrions observer si ce nouveau facteur est commun au canton non bilingue ou si celui-ci peut varier.

De plus, notre analyse est faite avec les données de cinq élections sur une période de vingt ans. Nous pourrions intégrer plus de données dans nos modèles afin de renforcer nos résultats ou peut-être en observer de nouveaux.

Ensuite, nos approches de prédiction sont centrées sur les communes et nous pourrions tester des approches centrées sur les candidats. L'étude présentée dans notre revue de littérature sur les élections aux États-Unis serait un bon point de départ. Nous pourrions choisir un certain nombre de caractéristiques pertinentes pour annoter nos candidats et tenter de prédire le résultat « élu » - « non-élus » sur la base de ces caractéristiques. L'approche par candidat tenterait de créer un modèle de prédiction capable de répondre à la question suivante : « Sachant les valeurs des caractéristiques suivantes d'un candidat, est-ce que celui-ci sera élu ? » Des annotations comme la

langue des candidats, si c'est un candidat sortant ou encore le sexe semble être des valeurs qui pourraient être utilisées pour de l'analyse prédictive. Nous pourrions également mélanger le concept des caractéristiques des candidats avec des annotations comme le score sur des communes clés et ainsi mélanger l'approche par commune avec l'approche par candidat. Intégrer les adversaires et leur nombre dans les modèles de prédictions d'élections semble également une piste intéressante.

Enfin, la mise à disposition de toujours plus de données avec les politiques d'open data et les avancées dans les techniques de data sciences offres toujours plus d'opportunités pour les sciences politiques et d'autres applications sont à envisager pour l'avenir.

ATTESTATION

Je déclare sur l'honneur, que j'ai effectué ce Travail de Master seul, sans autre aide que celles dûment signalées dans les références, et que je n'ai utilisé que les sources expressément mentionnées. Je ne donnerai aucune copie de ce rapport à un tiers sans l'autorisation conjointe du Responsable de l'Orientation et du Professeur chargé du suivi du Travail de Master et de l'institution ou entreprise pour laquelle ce travail a été effectué.

Lausanne, 20 août 2019.

Renzo Scuderi

REFERENCES BIBLIOGRAPHIQUES

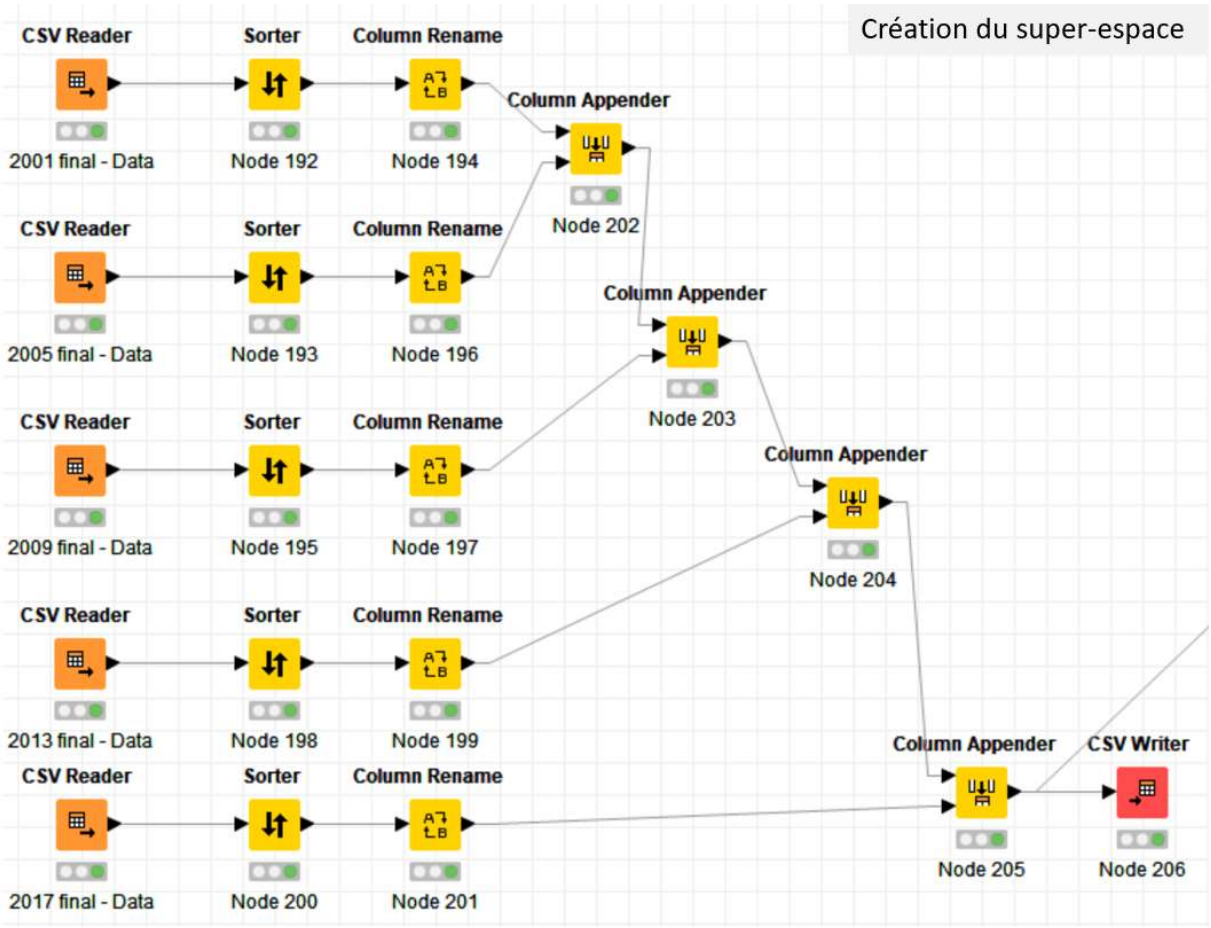
[1] 1. Federal Council: Open Government Data Strategy for Switzerland, (2014)
[2] 2. About – opendata.swiss, https://opendata.swiss/en/about/
[3] 3. Lathrop, D., Ruma, L. eds: Open government: collaboration, transparency, and participation in practice. O'Reilly, Beijing ; Cambridge [Mass.] (2010)
[4] 4. Transparency and Open Government, https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government
[5] 5. statistique, O. fédéral de la: Les 2212 communes de la Suisse au 1.1.2019 (Communes) | Carte, https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/cartes.assetdetail.7008568.html
[6] 6. Le fédéralisme suisse - www.ch.ch, https://www.ch.ch/fr/democratie/federalisme/le-federalisme-suisse/
[7] 7. RS 101 Constitution fédérale de la Confédération suisse du 18 avril 1999, https://www.admin.ch/opc/fr/classified-compilation/19995395/
[8] 8. Linder, W., Vatter, A.: Institutions and outcomes of Swiss federalism: The role of the cantons in Swiss politics. West Eur. Polit. 24, 95–122 (2001). doi:10.1080/01402380108425435
[9] 9. Application des communes suisses | Application des communes suisses, https://www.agvchapp.bfs.admin.ch/fr/home
[10] 10. Le Valais en chiffres, (2017)
[11] 11. Constitution Cantonale du Valais. (2017)
[12] 12. Gouvernement - vs.ch, https://www.vs.ch/web/gouvernement
[13] 13. Ladner, A.: Swiss political parties: Between persistence and change. West Eur. Polit. 24, 123–144 (2001). doi:10.1080/01402380108425436
[14] 14. statistique, O. fédéral de la: Canton du Valais: élections nationales et cantonales depuis 1919 - 1919-2017 | Tableau, https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/tableaux.assetdetail.2100105.html
[15] 15. Hosseini, S.M.S., Maleki, A., Gholamian, M.R.: Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Syst. Appl. 37, 5259–5264 (2010). doi:10.1016/j.eswa.2009.12.070
[16] 16. Etter, V., Herzen, J., Grossglauser, M., Thiran, P.: Mining democracy. In: Proceedings of the second edition of the ACM conference on Online social networks - COSN '14. pp. 1–12. ACM Press, Dublin, Ireland (2014)
[17] 17. Akarca, A.T., Başlevent, C.: Persistence in regional voting patterns in Turkey during a period of major political realignment. Eur. Urban Reg. Stud. 18, 184–202 (2011). doi:10.1177/0969776411399342
[18] 18. Jefferson West, W.: Regional cleavages in Turkish politics: An electoral geography of the 1999 and 2002 national elections. Polit. Geogr. 24, 499–523 (2005). doi:10.1016/j.polgeo.2005.01.003
[19] 19. Syakur, M.A., Khotimah, B.K., Rochman, E.M.S., Satoto, B.D.: Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conf. Ser. Mater. Sci. Eng. 336, 012017 (2018). doi:10.1088/1757-899X/336/1/012017

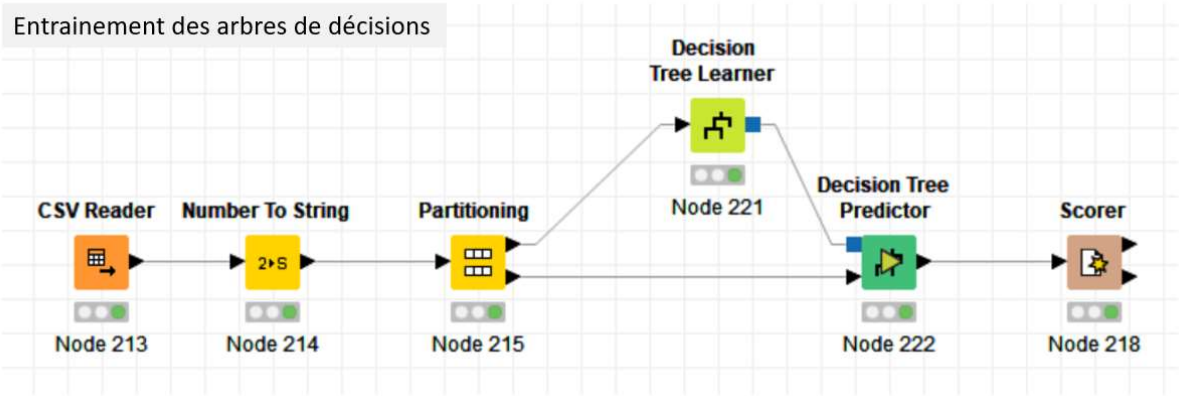
[20] 20. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. 8
[21] 21. Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., Jiang, J.: Tweets and Votes: A Study of the 2011 Singapore General Election. In: 2012 45th Hawaii International Conference on System Sciences. pp. 2583–2591. IEEE, Maui, HI, USA (2012)
[22] 22. Armstrong, J.S., Graefe, A.: Predicting elections from biographical information about candidates: A test of the index method. J. Bus. Res. 64, 699–706 (2011). doi:10.1016/j.jbusres.2010.08.005
[23] 23. Molloy, J.C.: The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biol. 9, e1001195 (2011). doi:10.1371/journal.pbio.1001195

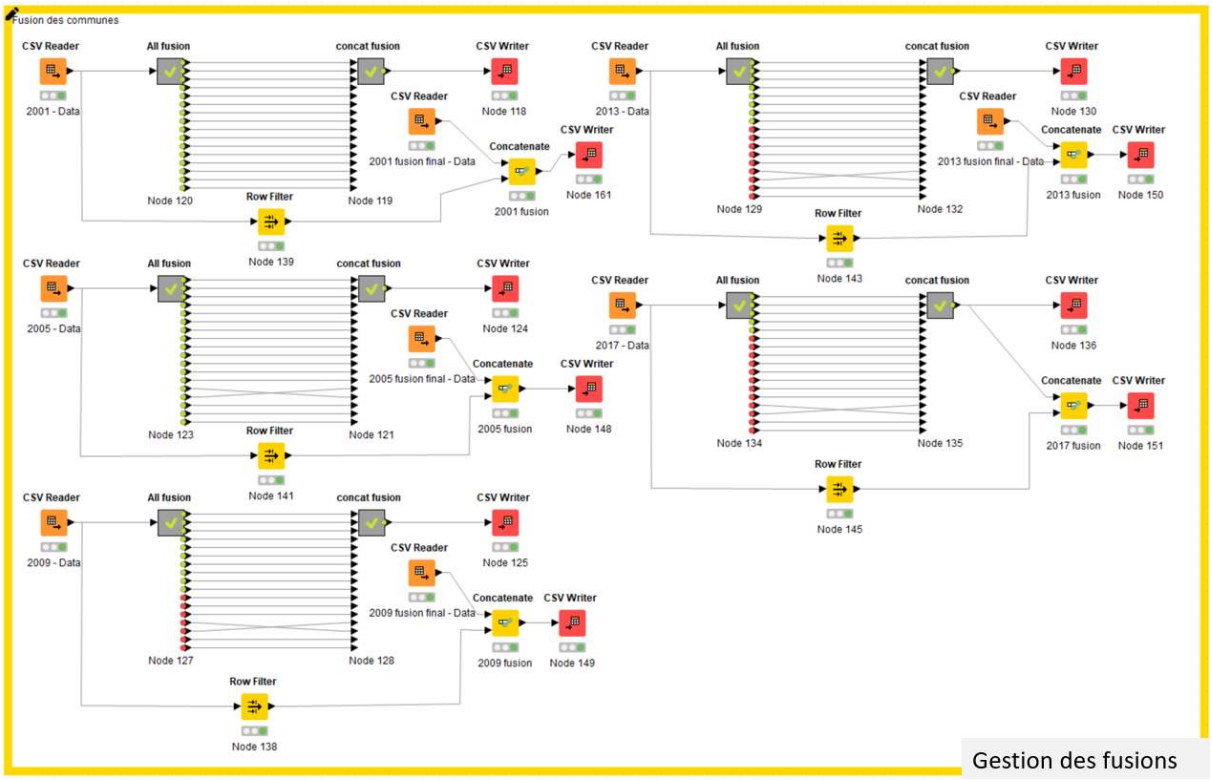
TABLE DES ILLUSTRATIONS

TABLEAUX

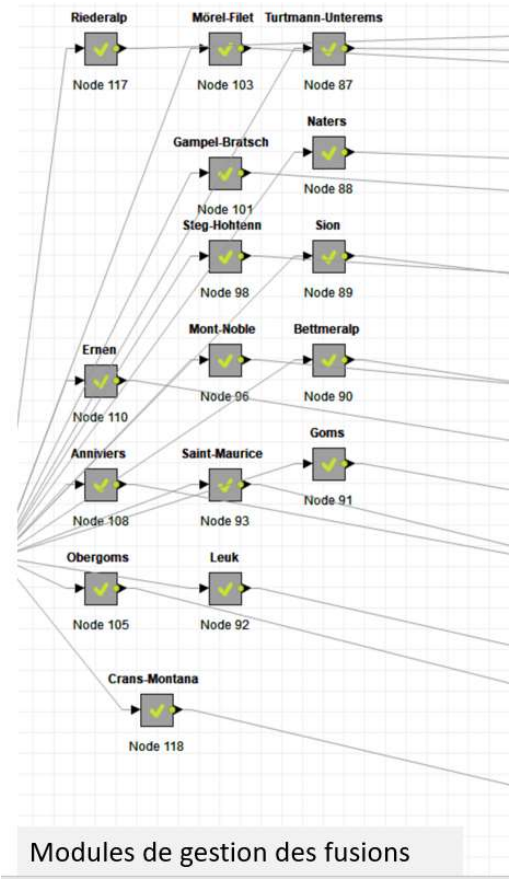
ANNEXES


Création du super-espace
CSV Reader
Sorter
Column Rename
2001 final - Data
Node 192
Node 194
Column Appender
Node 202
CSV Reader
Sorter
Column Rename
2005 final - Data
Node 193
Node 196
Column Appender
Node 203
CSV Reader
Sorter
Column Rename
2009 final - Data
Node 195
Node 197
Column Appender
Node 204
CSV Reader
Sorter
Column Rename
2013 final - Data
Node 198
Node 199
CSV Reader
Sorter
Column Rename
2017 final - Data
Node 200
Node 201
Column Appender
CSV Writer
Node 205
Node 206

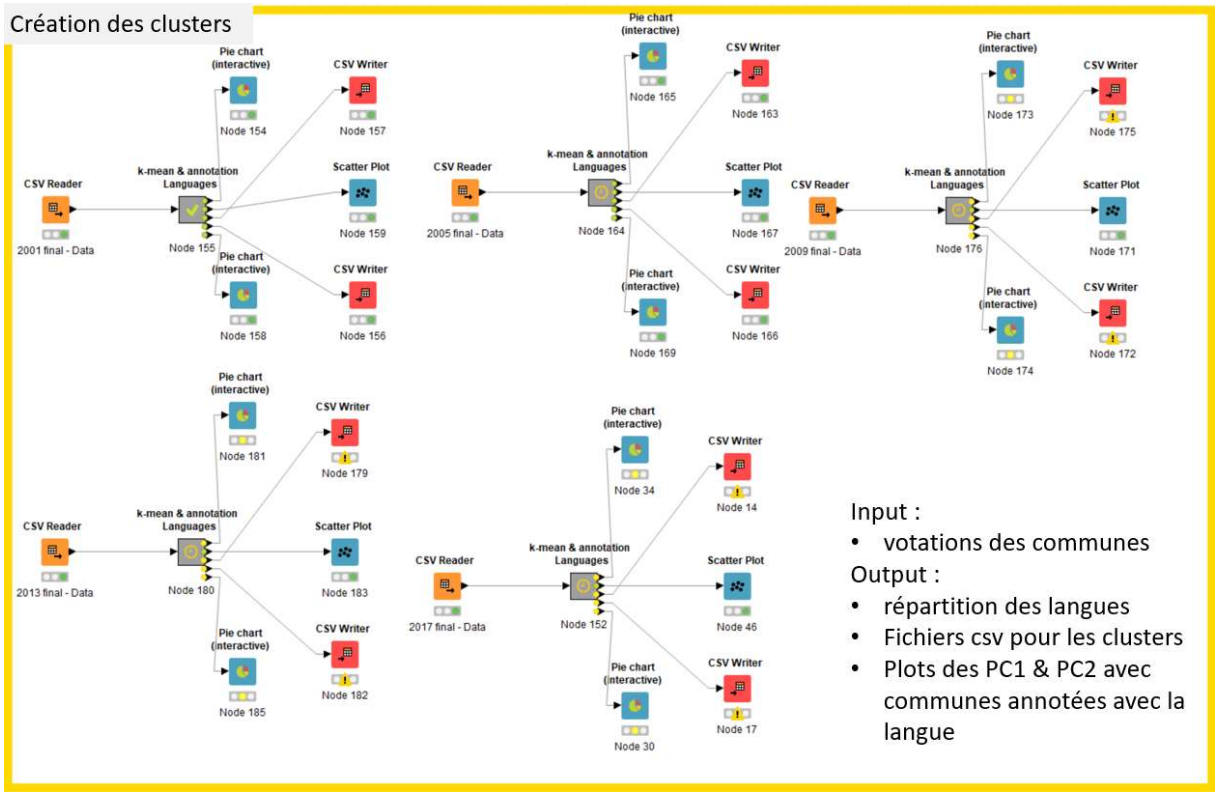

Entrainement des arbres de décisions
Decision
Tree Learner
Node 221
CSV Reader
Number To String
Partitioning
Decision Tree
Predictor
Scorer
Node 213
Node 214
Node 215
Node 222
Node 218

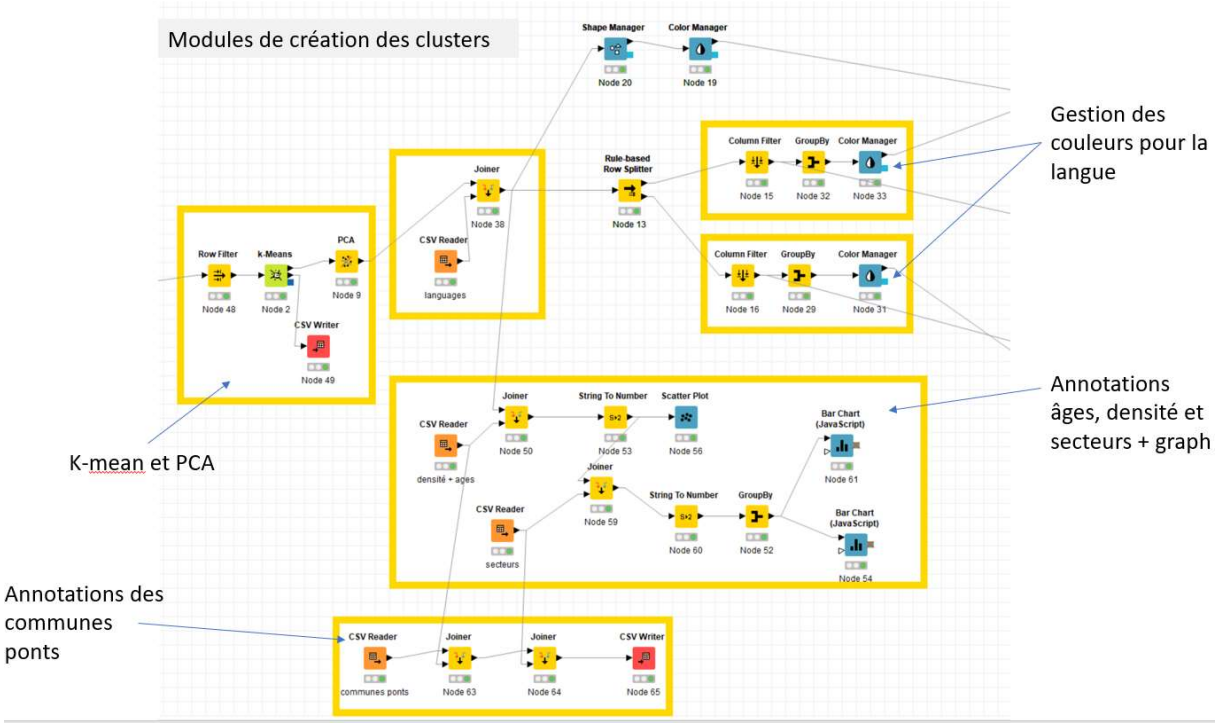


Gestion des fusions



Modules de gestion des fusions


Création des clusters
CSV Reader
2001 final - Data
k-mean & annotation Languages
Node 155
Pie chart (interactive)
Node 154
CSV Writer
Node 157
Scatter Plot
Node 159
Pie chart (interactive)
Node 158
CSV Writer
Node 156
CSV Reader
2005 final - Data
k-mean & annotation Languages
Node 164
Pie chart (interactive)
Node 165
CSV Writer
Node 163
Scatter Plot
Node 167
Pie chart (interactive)
Node 169
CSV Writer
Node 166
CSV Reader
2009 final - Data
k-mean & annotation Languages
Node 176
Pie chart (interactive)
Node 173
CSV Writer
Node 175
Scatter Plot
Node 171
Pie chart (interactive)
Node 174
CSV Writer
Node 172
CSV Reader
2013 final - Data
k-mean & annotation Languages
Node 180
Pie chart (interactive)
Node 181
CSV Writer
Node 179
Scatter Plot
Node 183
Pie chart (interactive)
Node 185
CSV Writer
Node 182
CSV Reader
2017 final - Data
k-mean & annotation Languages
Node 152
Pie chart (interactive)
Node 34
CSV Writer
Node 14
Scatter Plot
Node 46
Pie chart (interactive)
Node 30
CSV Writer
Node 17
Input :
votations des communes
Output :
répartition des langues
Fichiers csv pour les clusters
Plots des PC1 & PC2 avec communes annotées avec la langue


Modules de création des clusters
Shape Manager
Node 20
Color Manager
Node 19
Gestion des couleurs pour la langue
Column Filter
Node 15
GroupBy
Node 32
Color Manager
Node 33
Column Filter
Node 16
GroupBy
Node 29
Color Manager
Node 31
Rule-based Row Splitter
Node 13
Joiner
Node 38
CSV Reader
languages
Row Filter
Node 48
k-Means
Node 2
PCA
Node 9
CSV Writer
Node 49
K-mean et PCA
Annotations âges, densité et secteurs + graph
CSV Reader
densité + ages
Joiner
Node 50
String To Number
Node 53
Scatter Plot
Node 56
Bar Chart (JavaScript)
Node 61
CSV Reader
secteurs
Joiner
Node 59
String To Number
Node 60
GroupBy
Node 52
Bar Chart (JavaScript)
Node 54
Annotations des communes ponts
CSV Reader
communes ponts
Joiner
Node 63
Joiner
Node 64
CSV Writer
Node 65