

Multicolinéarité et régression PLS : TP : Élections Présidentielles 2017

L'objectif de ce TP est de construire des modèles sur des données à forte multicolinéarité entre les prédicteurs.

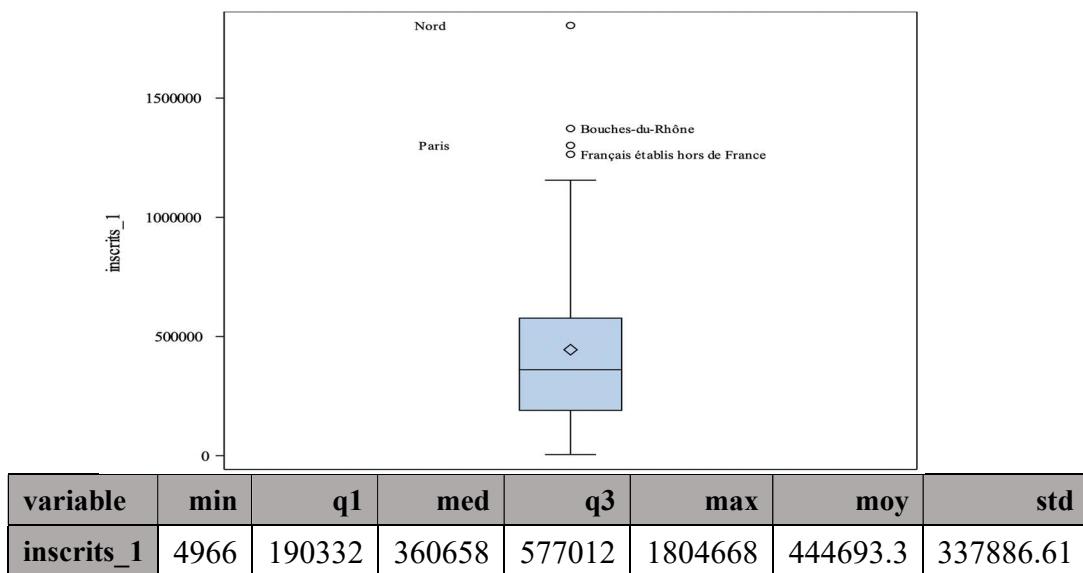
Les données sont issues des résultats des deux tours des élections présidentielles de 2017. L'objectif final est de modéliser les résultats du second tour, de cerner les grandes tendances, de détecter des départements atypiques ou mal reconstitués, etc.

Le jeu de données contient 107 individus qui correspondent aux départements de la métropole, de l'outre-mer et des français à l'étranger. Les variables correspondent aux nombre d'inscrits, de votants, d'abstentions, de suffrages exprimés, de bulletins blancs ou nuls pour les deux tours, de suffrages remportés par chacun des 11 candidats du premier tour et des deux candidats du second tour.

Question 1 :

Analyse univariée :

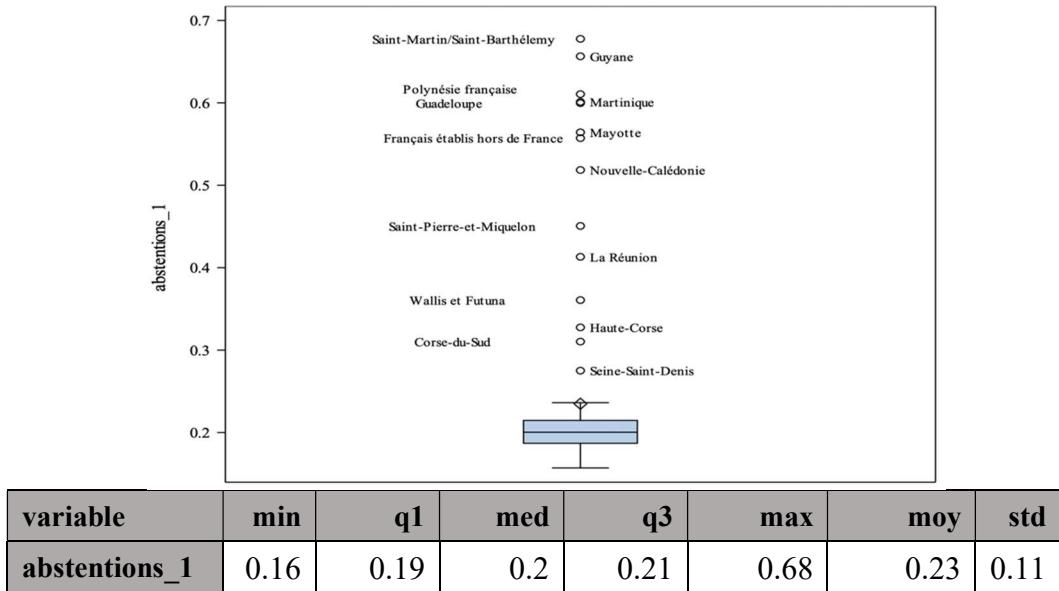
La variable inscrits_1 :



Nous pouvons constater d'après la distribution de la variables inscrits_1, que le nombre d'inscrits au premier tour est plus élevé dans le Nord, Bouches-du-Rhône, à Paris et chez les Français établis hors de France. Cela peut être expliqué par le fait qu'il y a une forte densité de population dans ces départements.

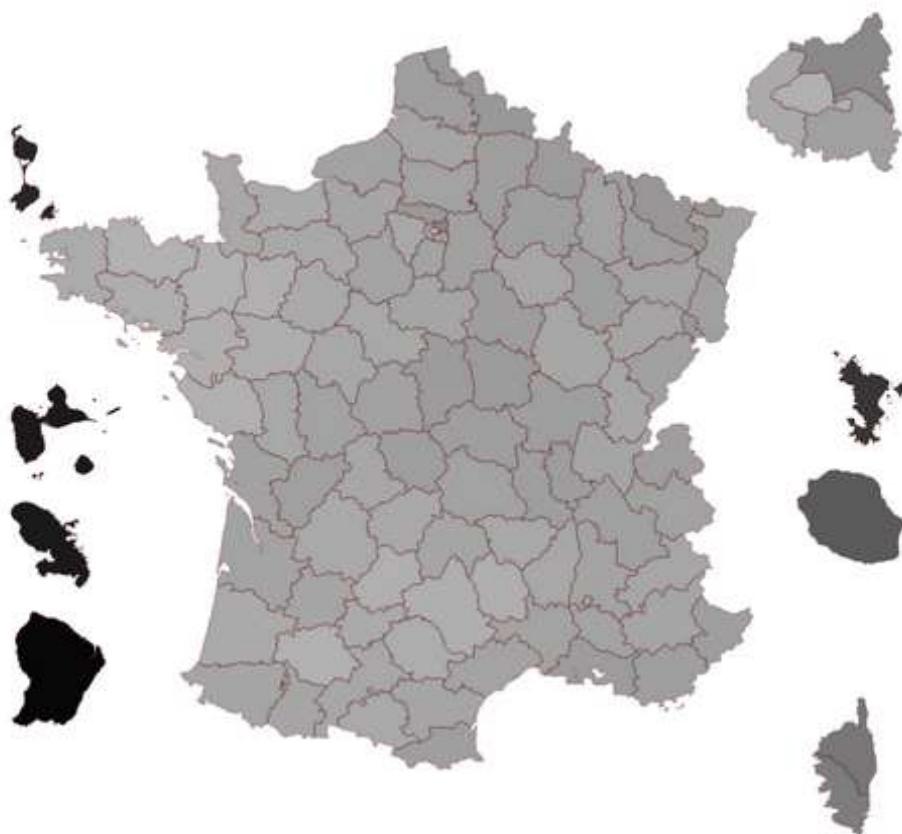
Par conséquent, il y a manifestement un effet de taille générer par les départements les plus peuplés qu'il convient de corriger en prenant le rapport du nombre de votant sur le nombre d'inscrit pour chaque candidat afin d'obtenir une proportion.

La variable abstentions_1 :

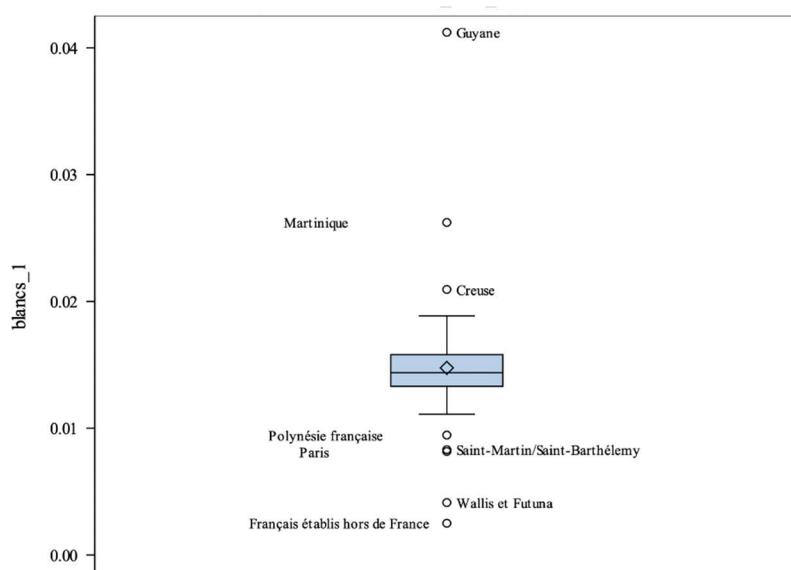


Nous remarquons que la proportion d'abstention au premier tour est très importante pour tous les départements d'outre-mer, chez les Français établis hors de France mais également en Seine-Saint-Denis. Historiquement, le vote de l'Outre-mer se distingue à chaque élection présidentielle par une forte abstention.

abstentions_1



La variable blancs_1 :

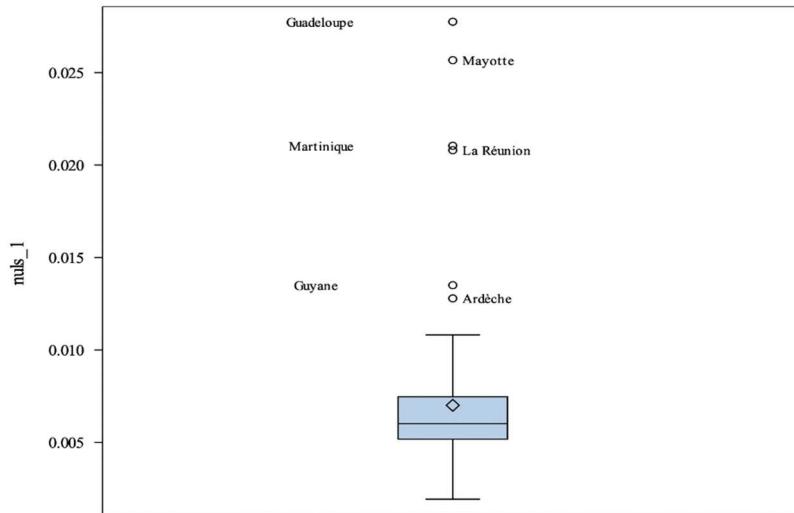


variable	min	q1	med	q3	max	moy	std
blancs_1	0	0.01	0.01	0.02	0.04	0.01	0

Nous remarquons que la proportion de votes blancs au premier tour est largement élevée dans la Guyane suivie de près par la Martinique et la Creuse. Les deux premiers départements sont

éloignés de la France métropolitaine ce qui peut expliquer cette proportion. Contrairement à la Polynésie Française, Paris, Saint Martin / Saint-Barthélemy, Wallis et Futuna et les Français établis hors de France qui sont des départements qui relèvent un pourcentage beaucoup plus faible de votes blancs.

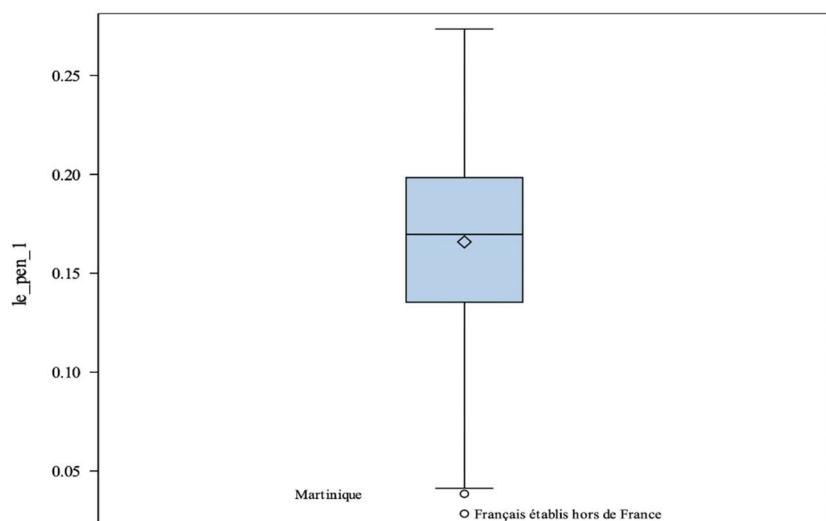
La variable nuls_1 :



variable	min	q1	med	q3	max	moy	std
nuls_1	0	0.01	0.01	0.01	0.03	0.01	0

Nous pouvons déduire que l'écrasante majorité des départements qui ont une proportion élevée de vote nuls au premier tour se situe dans les outre-mer. Par conséquent et selon les autres analyses nous constatons que certains départements situés dans les outre-mer moins engagés dans les élections présidentielles que les autres départements.

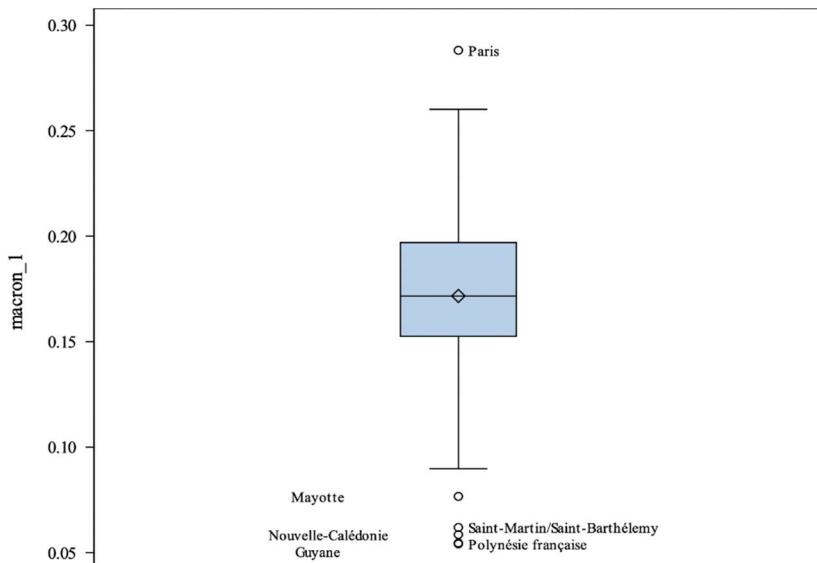
La variable le_pen_1 :



variable	min	q1	med	q3	max	moy	std
Le_pen_1	0.03	0.14	0.17	0.2	0.27	0.17	0.05

Nous constatons une faible proportion de vote pour la candidate Le Pen en Martinique et chez les Français établis hors de France. Comparer aux autres valeurs, les proportions restent sensiblement proches du reste de la distribution.

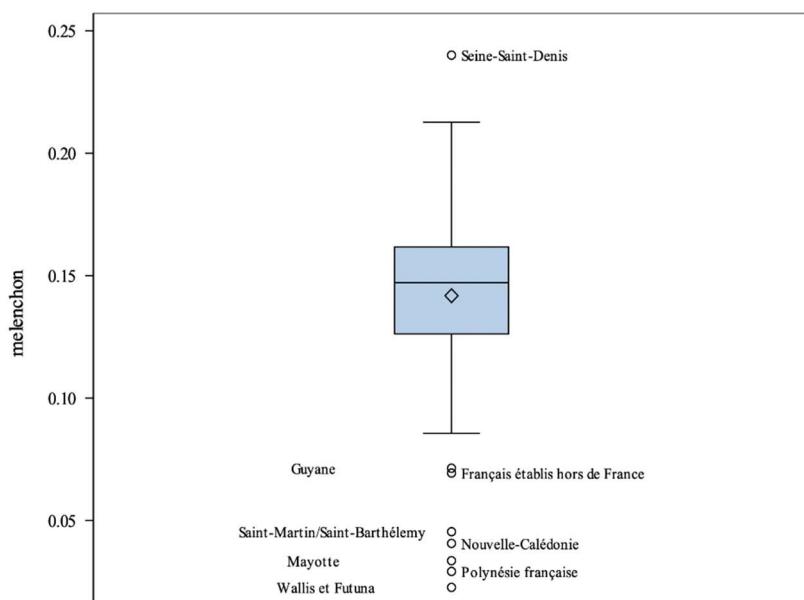
La variable macron_1 :



variable	min	q1	med	q3	max	moy	std
macron_1	0.05	0.15	0.17	0.20	0.29	0.17	0.04

Pour le candidat Macron, nous remarquons une proportion de vote à Paris supérieure aux autres départements, Mais légèrement plus basse dans les départements d'outre-mer. Mais ces valeurs restent raisonnables par rapport au reste de la distribution.

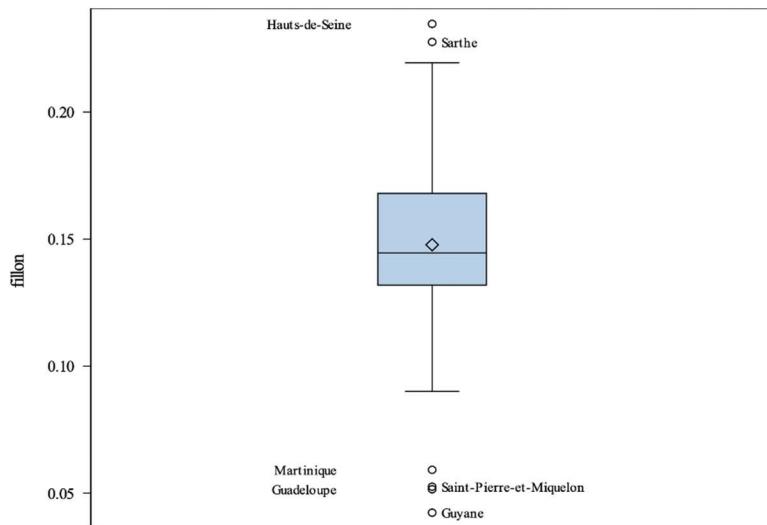
La variable melenchon :



variable	min	q1	med	q3	max	moy	std
melenchon	0.02	0.13	0.15	0.16	0.24	0.14	0.04

Nous remarquons une faible proportion de votes chez plusieurs départements d'outre-mer pour le candidat Melenchon. A l'opposé, la Seine-Saint-Denis représente l'unique département avec le plus de vote pour le même candidat.

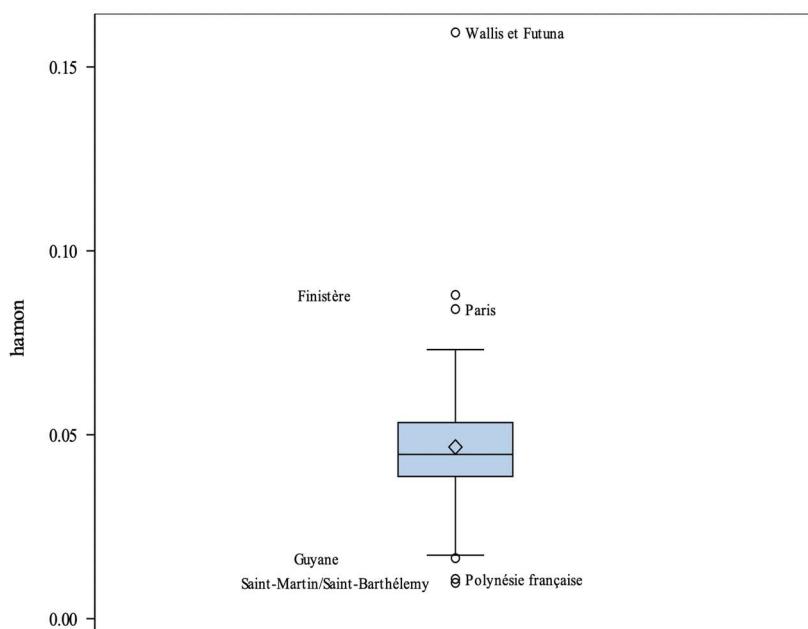
La variable fillon:



variable	min	q1	med	q3	max	moy	std
fillon	0.04	0.13	0.14	0.17	0.23	0.15	0.03

Comme constaté pour les candidats Macron et Melenchon, certain départements d'outre-mer se caractérisent par une proportion de vote très faible aussi pour le candidat Fillon. Par contre, la Sarthe et les Hauts-de-Seine représentent les deux départements avec une proportion de vote élevée pour Fillon.

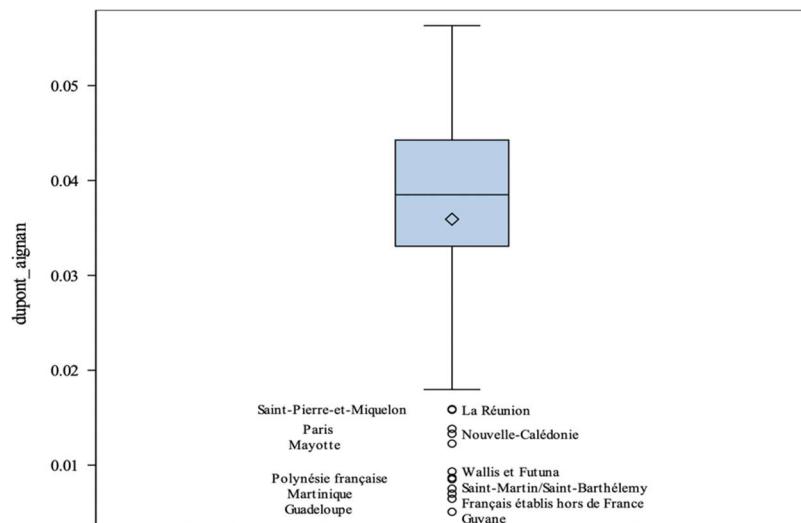
La variable hamon:



variable	min	q1	med	q3	max	moy	std
hamon	0.01	0.04	0.04	0.05	0.16	0.05	0.02

Le candidat Hamon semble principalement soutenu par le département Wallis et Futuna vu que la proportion de vote est largement élevée dans ce département, suivi de loin par Finistère et Paris. En outre, nous remarquons une proportion légèrement faible dans les départements de Guyane, la Polynésie française et Saint-Martin / Saint-Barthélemy.

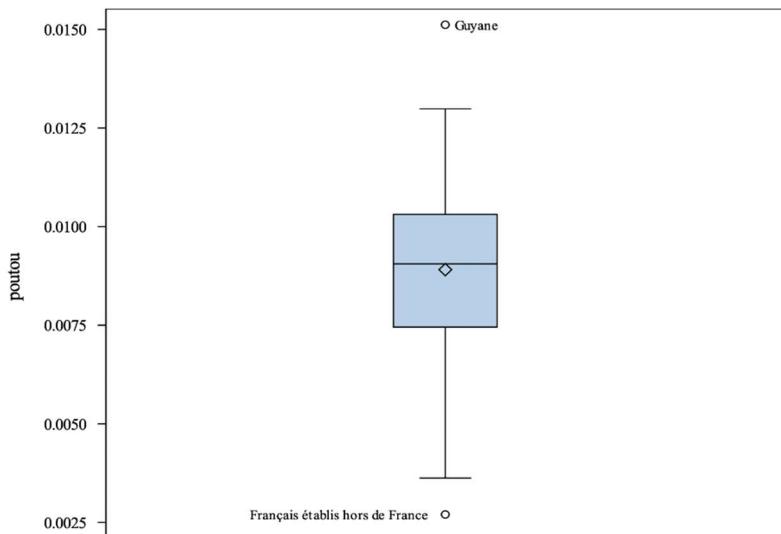
La variable dupont_aignan :



variable	min	q1	med	q3	max	moy	std
dupont_aignan	0.01	0.03	0.04	0.04	0.06	0.04	0.01

Contrairement aux autres candidats, le candidat Dupont Aignan enregistre une proportion de vote très faible dans la majorité des départements et territoires d'outre-mer et chez les Français établis hors de France ainsi que dans Paris.

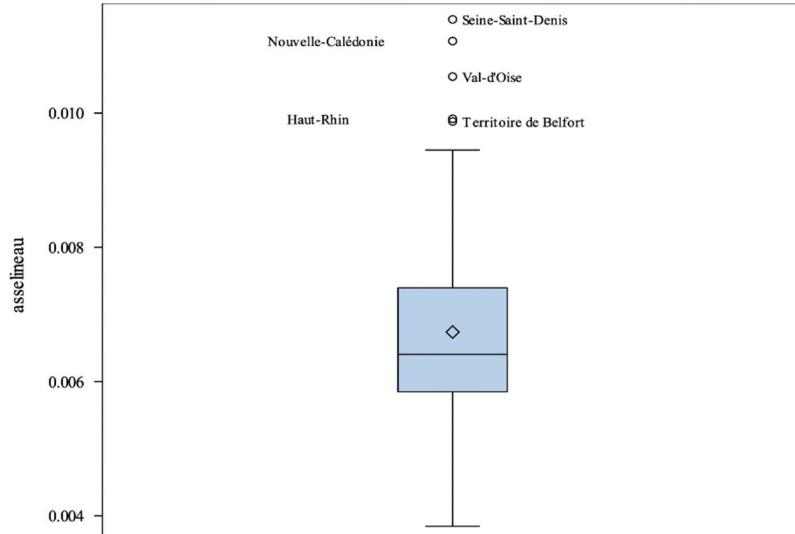
La variable poutou :



variable	min	q1	med	q3	max	moy	std
poutou	0	0.01	0.01	0.01	0.02	0.01	0

Le candidat Poutou semble ne pas avoir de grands écarts de votes. Néanmoins, une proportion de votes plus élevée est constatée en Guyane, et plus basse chez les Français établis hors de France.

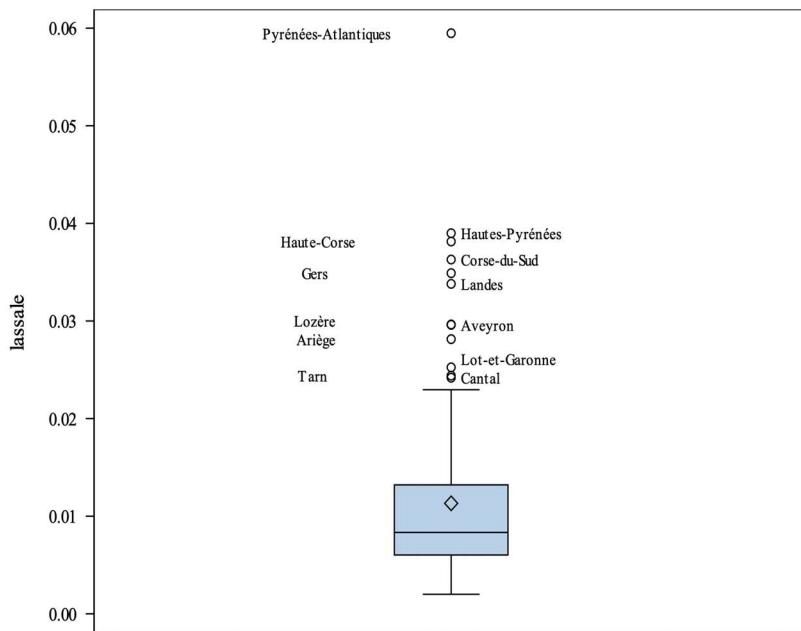
La variable asselineau:



variable	min	q1	med	q3	max	moy	std
asselineau	0	0.01	0.01	0.01	0.01	0.01	0

Le candidat Asselineau est principalement supportée par les départements de la Seine-Saint-Denis en Nouvelle-Calédonie, dans le Val-d’Oise, dans le Haut-Rhin ainsi que dans le Territoire de Belfort.

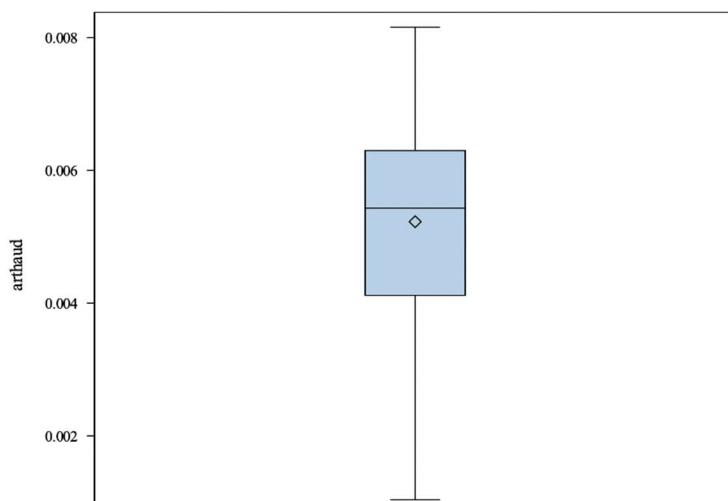
La variable lassale:



variable	min	q1	med	q3	max	moy	std
lassale	0	0.01	0.01	0.01	0.06	0.01	0.01

Nous remarquons la proportion de vote dans Pyrénées-Atlantique est largement supérieure aux autres départements pour le candidat Lassalle, cela s'explique par le fait que ce candidat est très impliqué dans la vie politique de ce département sachant que c'est son département de naissance. De plus, Lassalle a obtenu une proportion de votes élevée dans des départements majoritairement ruraux tels que les Hautes-Pyrénées, la Corse, le Gers, les Landes, la Lozère, l'Aveyron, l'Ariège, le Lot-et-Garonne, le Tarn ou encore le Cantal.

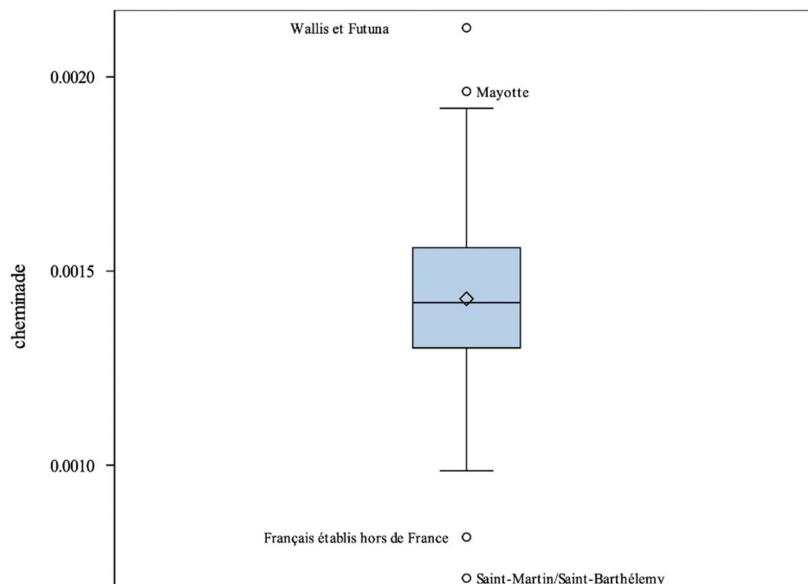
La variable arthaud:



variable	min	q1	med	q3	max	moy	std
arthaud	0	0	0.01	0.01	0.01	0.01	0

Le candidat Arthaud ne semble pas avoir des écarts de proportions de vote atypiques.

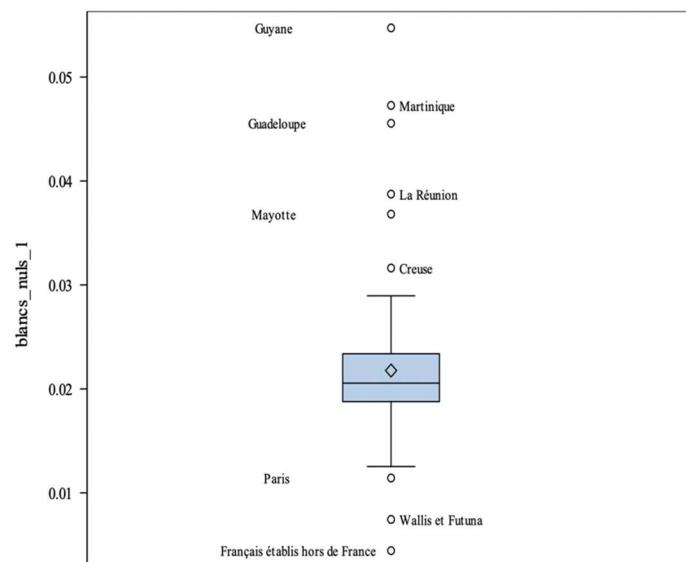
La variable cheminade :



variable	min	q1	med	q3	max	moy	std
cheminade	0	0 .001	0 .001	0.001	0.002	0,001	0

Pour le candidat Cheminade, nous observons une proportion de votes plus importante dans le département Wallis et Futuna suivi de Mayotte. Par contre, les Français établis hors de France et à Saint Martin / Saint-Barthélemy ont une plus faible proportion de votes par rapport aux reste de la distribution.

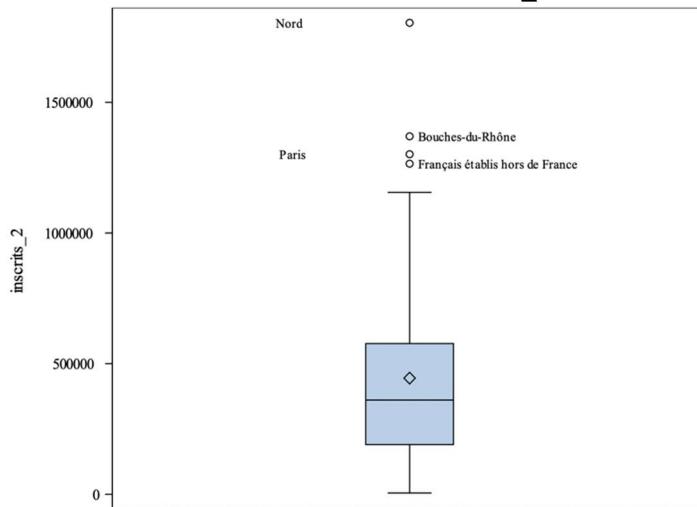
La variable blances_nuls_1:



variable	min	q1	med	q3	max	moy	std
blances_nuls_1	0	0.02	0.02	0.02	0.05	0.02	0.01

Nous pouvons constater encore une fois une proportion de votes blancs et de votes nuls assez élevée dans certains départements d'outre-mer mais plus faible chez les Français établis hors de France et dans les départements de Wallis et Futuna et Paris.

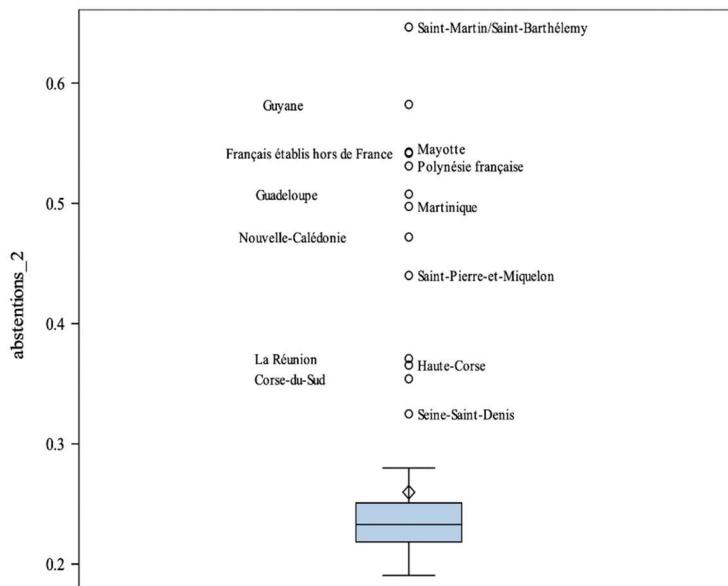
La variable inscrits_2:



variable	min	q1	med	q3	max	moy	std
inscrits_2	4963	190316	360646	576899	1804404	444567.22	337769.32

Nous pouvons constater encore une fois une proportion de votes blancs et de votes nuls assez élevée dans certains départements d'outre-mer mais plus faible chez les Français établis hors de France et dans les départements de Wallis et Futuna et Paris.

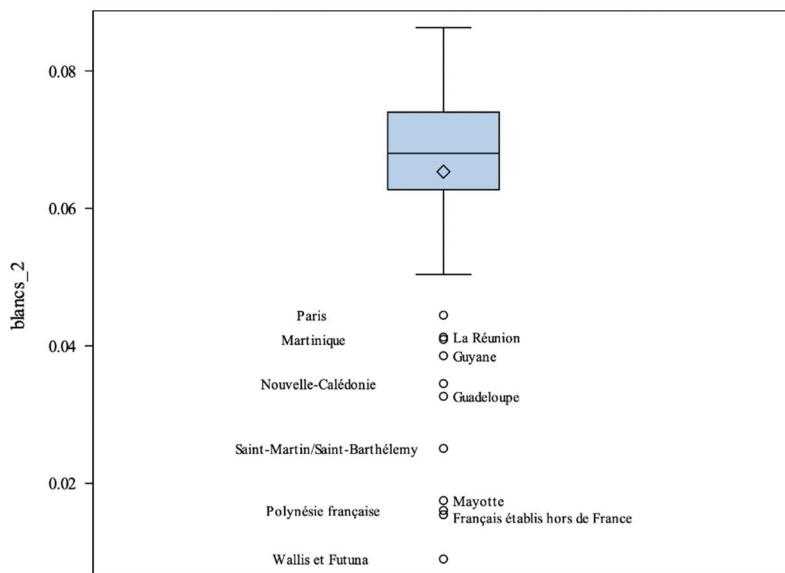
La variable abstentions_2:



variable	min	q1	med	q3	max	moy	std
abstentions_2	0.19	0.22	0.23	0.25	0.65	0.26	0.09

La proportion d'abstention reste toujours aussi importante pour la plupart des départements et territoires d'outre-mer, ainsi qu'en Corse.

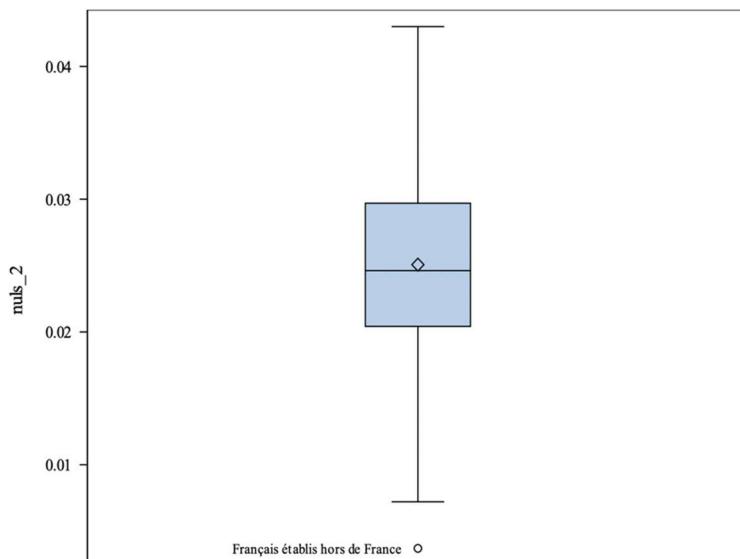
La variable blances_2:



variable	min	q1	med	q3	max	moy	std
blances_2	0.01	0.06	0.07	0.07	0.09	0.07	0.01

Pour les votes blancs du second tour, nous remarquons une proportion très faible dans le département de Wallis et Futuna suivi par plusieurs départements d'outre-mer ainsi que pour les Français établis hors de France et pour Paris. Nous pouvons noter une baisse de la proportion de votes blancs en Guyane et en Martinique, entre le premier et le second tour.

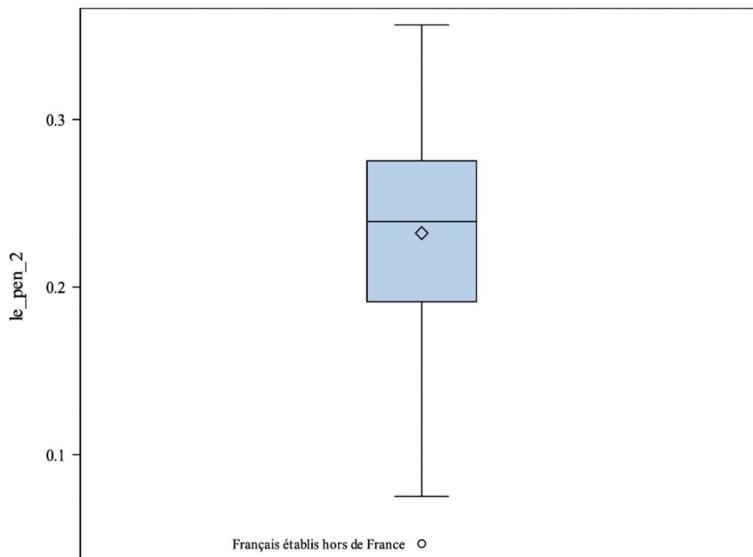
La variable nuls_2:



variable	min	q1	med	q3	max	moy	std
nuls_2	0	0.02	0.02	0.03	0.04	0.03	0.01

Les Français établis hors de France représente le seul département où les votes nul sont très faible.

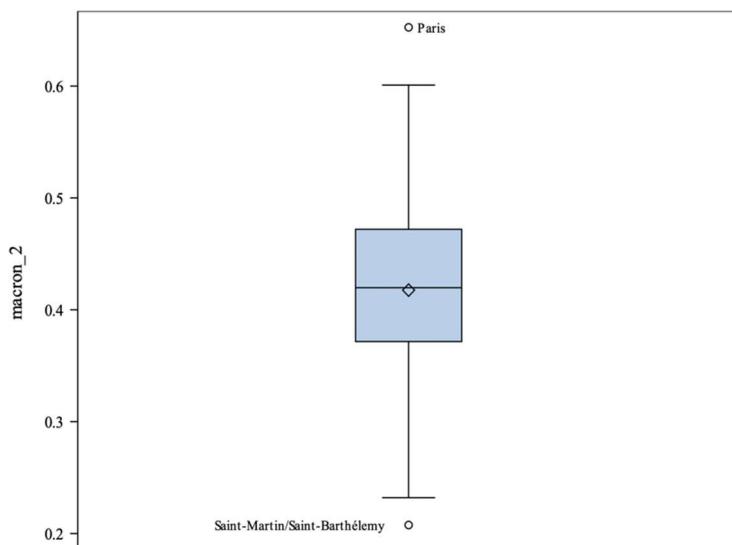
La variable le_pen_2:



variable	min	q1	med	q3	max	moy	std
le_pen_2	0.05	0.19	0.24	0.28	0.36	0.23	0.06

Comparé au premier tour, la proportion de vote pour Le Pen semble être homogène à l'unique exception des Français établis hors de France qui représente la proportion la plus faible des votes.

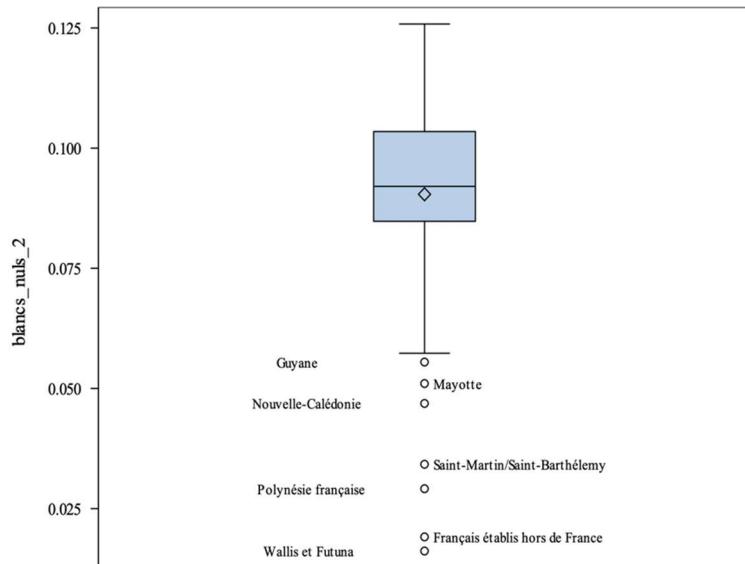
La variable macron_2:



variable	min	q1	med	q3	max	moy	std
macron_2	0.21	0.37	0.42	0.47	0.65	0.42	0.08

Le département de Paris est le plus engagé dans les votes pour Macron, tout comme le premier tour. A contrario de Saint-Martin / Saint-Barthélemy où la proportion des votes est légèrement faible. Nous remarquons aussi une amélioration du taux de vote pour ce candidat dans certains départements et territoires d'outre-mer.

La variable blances_nuls_2:

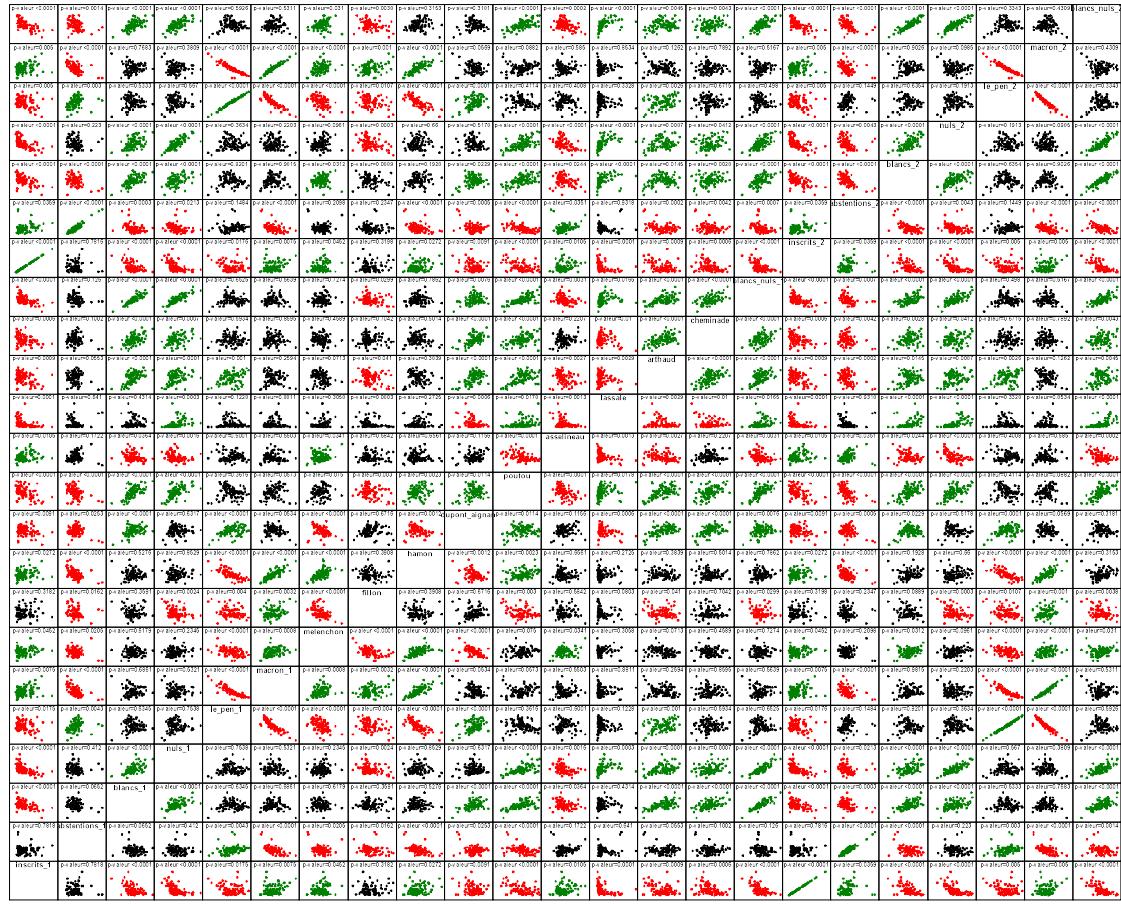


variable	min	q1	med	q3	max	moy	std
blances_nuls_2	0.02	0.08	0.09	0.1	0.13	0.09	0.02

Nous constatons que contrairement au premier tour la proportion de votes blancs et nuls du deuxième tour est plus faible dans certains départements et territoires d'outre-mer.
Nous retrouvons toujours les mêmes départements atypiques qui sont pour la majorité des départements et territoires d'outre-mer. Ces départements peuvent avoir un effet très important dans la modélisation et doivent être traiter.

Analyse bivariée :

L'analyse univariée ne suffit pas à statuer sur les données atypiques et ne permet pas de détecter les liens entre les variables pour notamment détecter les problèmes de colinéarité



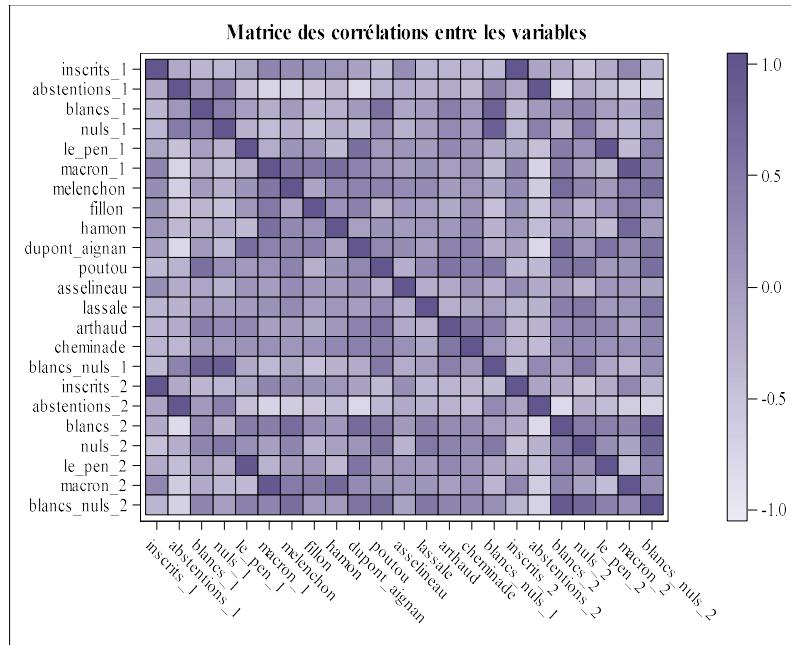
La figure ci-dessus (%SCATTERPLOT) nous permet de visualiser les croisements graphiques entre chaque paire de variables de l'étude, elle affiche pour chaque graphique la force de liaison à l'aide de la p-valeur obtenue par rapport à un seuil alpha fixé (nous prendrons ici = 0,05) et le signe de la relation (en rouge si le sens décroissant et en vert s'il est croissant, en noir si la p-valeur est supérieure à alpha).

La figure nous montre tout d'abord des points atypiques qui se démarque de façon systématique sur quelques distributions. On retrouve ainsi les départements des DOM-TOM qui présentait principalement la majorité des valeurs atypique dans ce jeu de données, pour éviter les effets néfastes que peuvent engendrer des points atypiques nous allons les enlever pour la partie modélisation.

De plus, nous pouvons remarquer que certaines variables semblent fortement liées linéairement. De façon intuitive, nous relevons une relation linéaire positive entre les résultats d'un même candidat entre le premier et le second tour, soit une corrélation positive entre le_pen_1 et le_pen_2, entre macron_1 et macron_2 et entre abstentions_1 et abstentions_2. De même, de façon logique, nous observons une relation linéaire négative entre le_pen_1 et macron_2, entre le pen_2 et macron_1 et entre le pen_2 et macron_2.

Nous pouvons également distinguer une relation linéaire positive entre macron_1 et hamon, et entre macron_2 et hamon.

Pour confirmer cela, nous allons maintenant observer la matrice des coefficients de corrélation simple de Pearson.



Dans l'ensemble, nous observons un grand nombre de corrélations significatives entre les variables.

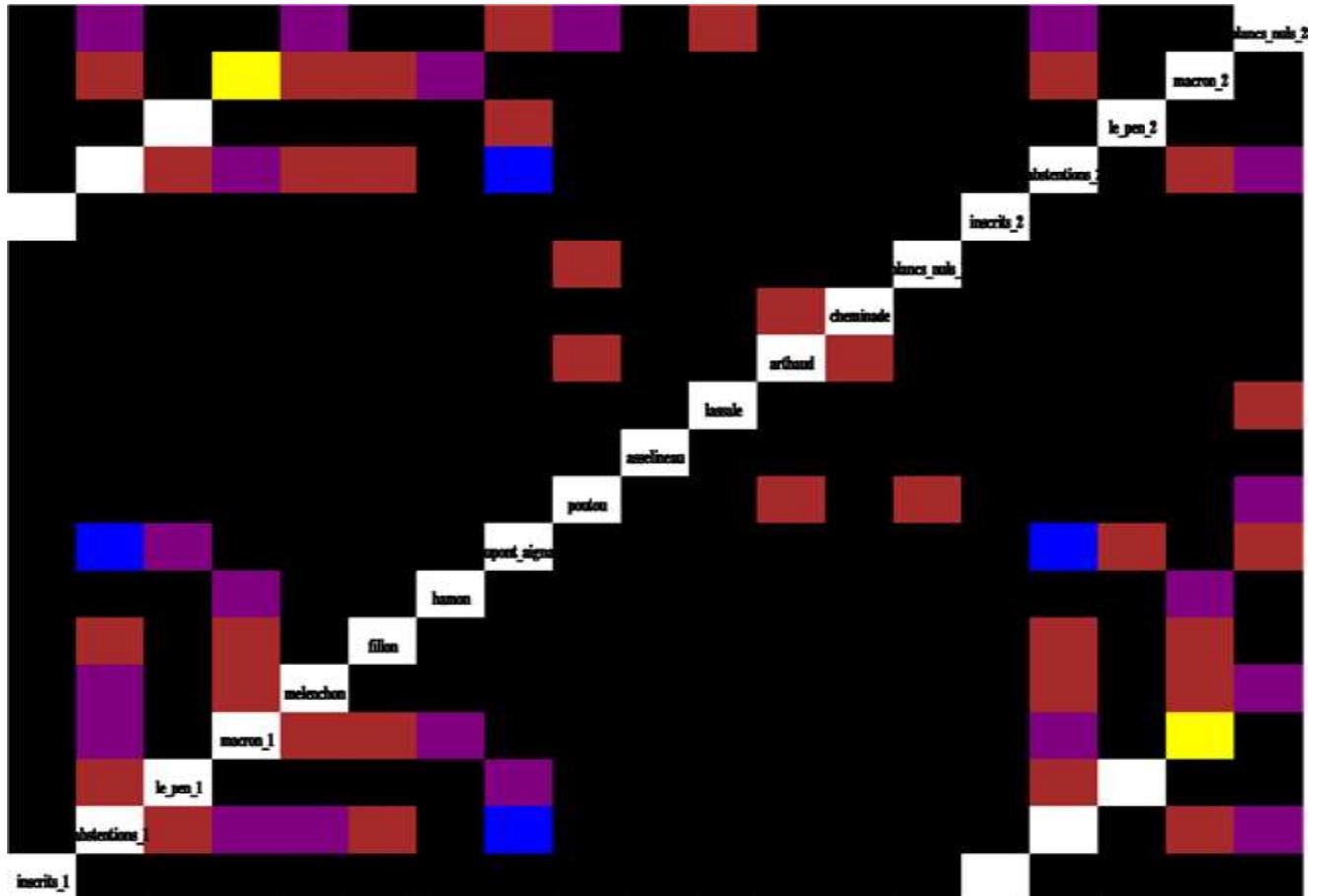
Tout d'abord, de façon intuitive nous constatons une très forte corrélation positive et significative entre les proportions de votes pour un même candidat ou pour un même type de bulletin entre le premier et le deuxième tour. Ainsi, les français semblent avoir été cohérent entre leur vote du premier et du second tour.

En outre, la variable abstentions_1 est fortement négativement corrélée avec plusieurs autres variables telles que macron_1, melenchon, fillon, dupont_aignan, blanc, macron_2 et blances_nuls_2. Elle est également liée plus faiblement et négativement à la variable le_pen_1 et le_pen_2, et positivement à la variable nulls_1.

De même, de façon significative, la variable abstentions_2 est fortement et négativement liée aux variables macron_1, melenchon et fillon, et dans de plus faible proportion à la variable le_pen_1 et hamon.

De plus, la variable blances_nuls_2 est également corrélées avec un grand nombre d'autres variables. Nous remarquons aussi une forte corrélation négative et significative entre blances_nuls_2 et abstentions_2.

Nous pouvons également fournir en complément à ces analyses, une matrice de dissimilarité de liaison entre ces variables. Cette matrice est en fonction du niveau de dissimilitarité, plus la couleur est claire, plus la dissimilitarité est faible.



Enfin, nous observons une forte corrélation entre le couple de variables inscrits_1 et inscrits_2, abstentions_1 et abstention_2, le_pen_1 et le_pen_2, macron_1 et macron_2 qui sont tous en blanc ce qui s'explique par une très faible dissimilarité, et donc une forte corrélation.

Question 2 :

L'augmentation de l'espace de représentation des données pose des problèmes de comparaison et d'interprétation des écarts entre ces données. En effet, l'augmentation de la dimension à tendance à rendre les données plus éparses et donc à fausser les manières traditionnelles d'analyser les données. Toutes les méthodes statistiques qui nécessitent le principe de significativité statistique sont impactées par le manque de densité des données dans l'espace.

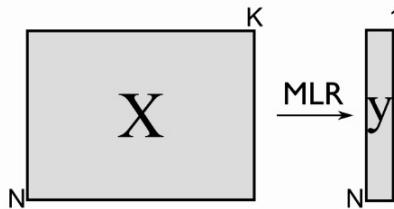
Ainsi, le fléau de la dimension nécessite des techniques de réduction de dimension afin de pouvoir représenter les données dans un espace adéquat et plus facilement interprétable, cela peut être résolu soit en utilisant les méthodes de sélection de variables, mais cette solution ne permet pas de pallier le problème de multicolinéarité, soit en gardant toutes les variables, et dans ce cas, les méthodes sont nombreuses, et souvent très naturelles, pour contourner le problème de multicolinéarité, tels que :

1 - La régression sur composantes principales :

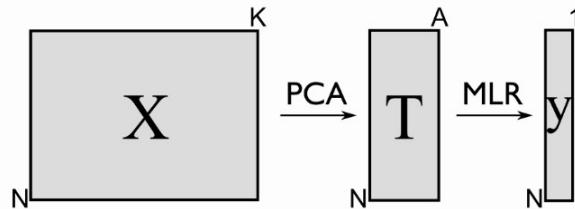
La régression en composantes principales (PCR) est une technique d'analyse de régression basée sur l'analyse en composantes principales (ACP). Plus précisément, l'ACP découle du

fait que chaque modèle de régression linéaire peut être reformulé en termes d'un ensemble de variables explicatives orthogonales. Ces nouvelles variables sont obtenues sous forme de combinaisons linéaires des variables explicatives d'origine. Ils sont appelés les principaux composants.

Multiple linear regression



Principal component regression



Une utilisation majeure de la PCR consiste à surmonter le problème de multicolinéarité qui se pose lorsque deux ou plusieurs des variables explicatives sont proches d'être colinéaires. La PCR peut convenablement faire face à de telles situations en excluant certaines des composantes principales à faible variance dans l'étape de régression.

De plus, en régressant habituellement sur seulement un sous-ensemble de toutes les composantes principales, la PCR peut entraîner une réduction de dimension en abaissant considérablement le nombre effectif de paramètres caractérisant le modèle sous-jacent. Cela peut être particulièrement utile dans les paramètres avec des covariables de grande dimension. De plus, grâce à une sélection appropriée des principaux composants à utiliser pour la régression, la PCR peut conduire à une prédiction efficace des résultats sur la base du modèle supposé.

Inconvénients :

- Les 1ères composantes principales ne sont pas forcément celles qui expliqueront le mieux le Y.
- Le choix de k n'est pas évident.

Il vaudrait donc mieux choisir les composantes principales les plus corrélés avec les Y et cela revient à du Pas à Pas. Cette remarque justifie les développements de la régression **PLS**.

2 - La régression sur premières composantes principales (RFPC) :

La RFPC peut être considérer en tant qu'une PCR spécifique, elle consiste à réaliser une classification hiérarchique descendante et repose sur la recherche de composantes obliques en analyse factorielle exploratoire (AF). Chaque groupe de variables sera normalement unidimensionnel. Cela se traduira par une première valeur propre très importante, par rapport aux autres, et donc suffira à résumer le groupe de variables concerné par la première composante principale. Enfin, une régression sera appliquée sur la réponse en fonction des premières composantes principales de chaque groupe.

3 - Régression Ridge :

En cas de multicolinéarité, l'estimation par les moindres carrés n'est pas faisable car la matrice $(X^T X)$ n'est pas inversible.

Pour réduire la multicolinéarité, nous pouvons utiliser la régularisation, c'est-à-dire conserver toutes les caractéristiques, mais réduire l'amplitude des coefficients du modèle. C'est une bonne solution lorsque chaque variable explicative contribue à prédire la variable dépendante.

La régression Ridge est une technique originale des statistiques permettant de manipuler la colinéarité en régression. Elle se base sur une pénalisation de type norme L^2 et elle a la particularité de ne pas annuler les coefficients β mais plus de les réduire et de les faire tendre vers 0. On parle de « shrinkage » des coefficients.

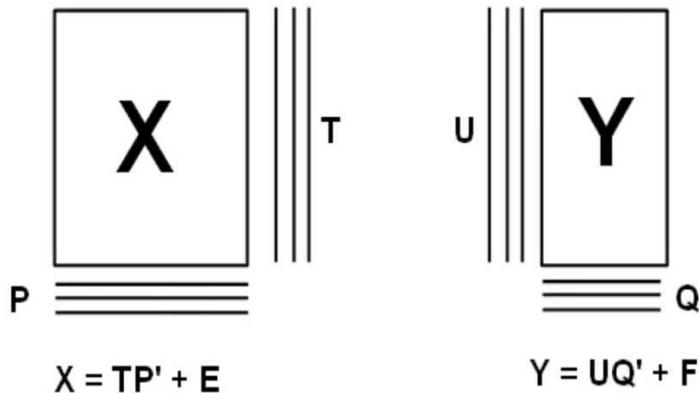
$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \|y - XB\|_2^2 + \lambda \|B\|_2^2$$

Ce type d'approche s'attaque directement à l'estimateur, pour faire face au problème de multicolinéarité en le contraignant sur l'espace des paramètres et non plus sur celui des variables. Mais le biais est encore présent dans ce type d'approche, par contre la variance des erreurs est plus faible.

4 - Régression PLS :

La régression PLS est une méthode appropriée pour l'ensembles de données qui ne correspondent pas aux attentes traditionnelles exigées par la régression ordinaire. Lorsque la taille des données est faible, quand les données souffrent de multicolinéarité, de valeurs manquantes ou lorsque la distribution est inconnue, PLS permet de minimiser les effets néfastes de ces conditions. En utilisant un algorithme itératif basé sur des composantes "NIPALS" et des régressions linéaires entre ces composantes.

Le principal avantage de la régression PLS et ce qui lui donne une forte supériorité face à la régression linéaire multiple c'est qu'elle permet de traiter le problème de la multicolinéarité en sélectionnant un petit nombre de combinaisons linéaires des variables originales qui expliquent le maximum de covariance entre les X et Y.



On distingue souvent la PLS1 de la méthode PLS2. La PLS1 concerne le cas où il y a une seule variable dépendante, la PLS2 celui où il y a plusieurs variables dépendantes.

Question 3 :

Commençons par 3 modèles univariés pour estimer le nombre de voix des abstentionnistes, des bulletins blancs ou nuls, des suffrages de chacun des deux candidats retenus, à l'aide des variables du premier tour.

Pour cela commençons par réaliser un modèle de régression avec la méthode des moindres carrés ordinaires (MCO).

Modèle MCO :

Commençons par l'abstention au second tour :

Root MSE	0.0067 7	R carré	0.999 3
Moyenne dépendante	0.2331 9	R car. ajust.	0.999 2
Coeff Var	2.9050 2		

La méthode des MCO nous donnera toujours des résidus très faibles, en témoigne le **R carré > 99%** néanmoins, notre besoin de pouvoir interpréter le résultat nous oblige à nous pencher sur la significativité des coefficients de notre régression

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > t	Tolérance	Inflation de variance
abstentions_1	1	1.01775	0.02804	36.30	<.0001	0.01505	66.45551
le_pen_1	1	-0.02194	0.02787	-0.79	0.4334	0.01878	53.25626
macron_1	1	-0.08346	0.06025	-1.39	0.1697	0.00394	253.49186
melenchon	1	0.30557	0.05048	6.05	<.0001	0.00817	122.37395
fillon	1	0.21759	0.03737	5.82	<.0001	0.01413	70.79305
hamon	1	0.11851	0.15145	0.78	0.4361	0.00875	114.26952
dupont_aignan	1	0.16471	0.16119	1.02	0.3098	0.01172	85.29875
poutou	1	-2.35717	0.79332	-2.97	0.0039	0.00885	112.93878
asselineau	1	-1.18244	0.81336	-1.45	0.1498	0.01498	66.73959
lassale	1	0.01748	0.11503	0.15	0.8796	0.15054	6.64291
arthaud	1	0.43142	1.06866	0.40	0.6875	0.01399	71.49786
cheminade	1	0.70144	6.20955	0.11	0.9103	0.00588	170.17665
blancs_nuls_1	1	-0.75271	0.36818	-2.04	0.0441	0.00777	128.68852

A l'aide d'un test de Student, il est d'ores et déjà évident que les variables de la plupart des candidats qui ne sont pas significatifs mais ce n'est pas le cas pour Fillon, Mélenchon, Poutou et Asselineau ce qui peut paraître curieux quand on sait qu'il existe une forte colinéarité

entre par exemple Macron et Fillon. De plus, on remarque que les coefficients entre Macron et Fillon sont de signe opposés (-1.39 pour macron, 5.82 pour Fillon) alors que ces variables sont positivement corrélées.

Attardons-nous ensuite le lien entre les composantes principales et les variables explicatives de notre modèle :

Diagnostics de colinéarité							
Effectif	Valeur propre	Index de condition	Proportion de variation				
			abstentions_1	le_pen_1	macron_1	melenchon	fillon
1	12.22617	1.00000	0.00009804	0.00011722	0.00002566	0.00005327	0.00009067
2	0.39750	5.54597	0.00002301	0.00077434	0.00000380	0.00000148	0.00021830
3	0.14623	9.14379	0.00133	0.03938	0.00418	0.00341	0.00139
4	0.09694	11.23052	0.01820	0.01074	0.00008304	0.00038619	0.00892
5	0.05002	15.63430	0.00316	0.01021	0.00352	0.03830	0.15521
6	0.02710	21.24134	0.09960	0.07003	0.00118	0.00518	0.02134
7	0.01870	25.57019	0.14324	0.27392	0.00318	0.00085351	0.00000197
8	0.01127	32.93587	0.28727	0.06782	0.00131	0.13115	0.06778
9	0.00773	39.77424	0.29387	0.00091914	0.00028970	0.01636	0.04478
10	0.00659	43.06887	0.08372	0.04337	0.01721	0.06549	0.07685
11	0.00481	50.42494	0.03571	0.01277	0.02624	0.00944	0.08327
12	0.00436	52.98215	0.01406	0.20879	0.02194	0.53238	0.05703
13	0.00260	68.61626	0.01972	0.26116	0.92085	0.19701	0.48311

Diagnostics de colinéarité							
Effectif	Valeur propre	Index de condition	Proportion de variation				
			dupont_aignan	poutou	asselineau	lassale	arthaud
1	12.22617	1.00000	0.00007591	0.00005789	0.00009571	0.00065065	0.00008817
2	0.39750	5.54597	0.00084170	0.00001453	0.00068439	0.33296	0.00136
3	0.14623	9.14379	0.00314	0.00004201	0.00159	0.02593	0.00672
4	0.09694	11.23052	0.00003983	0.01578	0.03685	0.00157	0.04049
5	0.05002	15.63430	0.00870	0.00085878	0.01931	0.00855	0.00004080
6	0.02710	21.24134	0.13491	0.00327	0.14204	0.05243	0.00665
7	0.01870	25.57019	0.06805	0.00160	0.01481	0.01839	0.00453
8	0.01127	32.93587	0.03083	0.04000	0.02632	0.12298	0.20806
9	0.00773	39.77424	0.09638	0.36586	0.16333	0.21621	0.47210
10	0.00659	43.06887	0.43977	0.26962	0.28985	0.01948	0.03561
11	0.00481	50.42494	0.00173	0.24937	0.16346	0.09680	0.00002727
12	0.00436	52.98215	0.04304	0.02826	0.01220	0.09851	0.21601
13	0.00260	68.61626	0.17251	0.02525	0.12946	0.00555	0.00831

Détection de la multicolinéarité :

Ce tableau nous renseigne sur la proportion de variance des chaque variable se reflétant dans la variance des composantes principales.

Critère de condition des indices de conditionnement :

Il est également indiqué le rapport entre la première valeur propre et les autres valeurs propres dans la colonne « Indexe de condition », **la dernière valeur propre est 68 fois plus petite que la première valeur propre**.

Il est traditionnellement admis dans la littérature **qu'une valeur supérieure à 30 indique une multicolinéarité**. Une valeur supérieure à 100 indique une très forte multicolinéarité.

Critère du facteur d'inflation de la variance :

L'approche la plus classique consiste à examiner les *facteurs d'inflation de la variance (FIV)* (*VIF*) en anglais. Les FIV estiment de combien la variance d'un coefficient est augmentée en raison d'une relation linéaire avec d'autres prédicteurs. Ainsi, un FIV de 1,8 nous dit que la variance de ce coefficient particulier est supérieure de 80 % à la variance que l'on aurait dû observer si ce facteur n'est absolument pas corrélé aux autres prédicteurs.

En général il est recommandé de s'inquiéter d'une forte colinéarité lorsque des variables présentent une inflation de la variance supérieure à 5.

Voici les valeurs trouvées pour nos variables explicatives :

Variable	Inflation de variance
abstentions_1	66.45551
le_pen_1	53.25626
macron_1	253.49186
melenchon	122.37395
fillon	70.79305
hamon	114.26952
dupont_aignan	85.29875
poutou	112.93878
asselineau	66.73959
lassale	6.64291

Il est évident que l'on a affaire à un problème de multi colinéarité sévère justifiant une approche par d'autre modèle qu'une simple régression linéaire.

On retrouve des résultats similaires avec les autres régressions :

Blancs ou nul au second tour :

Root MSE	0.0053 2	R carré	0.997 3
Moyenne dépendante	0.0949 1	R car. ajust.	0.996 9
Coeff Var	5.6080 8		

Paramètres estimés								
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > t	Tolérance	Inflation de variance
abstentions_1	abstentions_1	1	-0.08436	0.02203	-3.83	0.0002	0.01505	66.45551
le_pen_1	le_pen_1	1	0.09734	0.02190	4.45	<.0001	0.01878	53.25626
macron_1	macron_1	1	-0.12352	0.04734	-2.61	0.0108	0.00394	253.49186
melenchon	melenchon	1	0.18085	0.03966	4.56	<.0001	0.00817	122.37395
fillon	fillon	1	0.10199	0.02937	3.47	0.0008	0.01413	70.79305
hamon	hamon	1	0.29648	0.11900	2.49	0.0147	0.00875	114.26952
dupont_aignan	dupont_aignan	1	0.17772	0.12665	1.40	0.1643	0.01172	85.29875
poutou	poutou	1	1.12934	0.62334	1.81	0.0736	0.00885	112.93878
asselineau	asselineau	1	-1.70966	0.63908	-2.68	0.0090	0.01498	66.73959
lassale	lassale	1	0.45475	0.09038	5.03	<.0001	0.15054	6.64291
arthaud	arthaud	1	-2.28082	0.83968	-2.72	0.0080	0.01399	71.49786
cheminade	cheminade	1	11.16530	4.87903	2.29	0.0247	0.00588	170.17665
blancs_nuls_1	blancs_nuls_1	1	2.13788	0.28929	7.39	<.0001	0.00777	128.68852

Cette fois-ci **seul Dupont Aignan et Poutou sont considérés comme non significative**.

On retrouve un **R Carré toujours aussi élevé** et ce même problème de multicolinéarité sévère en regardant l'inflation de variance.

Variable Macron 2 :

R carré	0.999 8
R car. ajust.	0.999 7

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > t	Tolérance	Inflation de variance
abstentions_1	1	0.02651	0.02927	0.91	0.3676	0.01505	66.45551
le_pen_1	1	-0.29505	0.02908	-10.14	<.0001	0.01878	53.25626
macron_1	1	1.32352	0.06288	21.05	<.0001	0.00394	253.49186
melenchon	1	0.47691	0.05269	9.05	<.0001	0.00817	122.37395
fillon	1	0.53731	0.03901	13.77	<.0001	0.01413	70.79305
hamon	1	0.47993	0.15806	3.04	0.0032	0.00875	114.26952
dupont_aignan	1	0.52915	0.16823	3.15	0.0023	0.01172	85.29875
poutou	1	1.49382	0.82798	1.80	0.0748	0.00885	112.93878
asselineau	1	4.92490	0.84889	5.80	<.0001	0.01498	66.73959
lassale	1	0.30638	0.12006	2.55	0.0125	0.15054	6.64291
arithaud	1	4.39965	1.11535	3.94	0.0002	0.01399	71.49786
cheminade	1	-12.01750	6.48085	-1.85	0.0672	0.00588	170.17665
blancs_nuls_1	1	-0.75002	0.38426	-1.95	0.0543	0.00777	128.68852

Variable Le pen 2 :

Root MSE	0.00442	R carré	0.9997
Moyenne dépendante	0.24203	R car. ajust.	0.9997
Coeff Var	1.82761		

Paramètres estimés

Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > t	Tolérance	Inflation de variance
abstentions_1	abstention	1	0.04010	0.0183	2.19	0.0313	0.01505	66.45551
le_pen_1	le_pen_1	1	1.21964	0.0182	67.03	<.0001	0.01878	53.25626
macron_1	macron_1	1	-0.11654	0.0393	-2.96	0.0040	0.00394	253.49186
melenchon	melenchon	1	0.03668	0.0329	1.11	0.2690	0.00817	122.37395
fillon	fillon	1	0.14312	0.0244	5.86	<.0001	0.01413	70.79305
hamon	hamon	1	0.10508	0.0988	1.06	0.2910	0.00875	114.26952
dupont_aignan	dupont_aignan	1	0.12842	0.1052	1.22	0.2259	0.01172	85.29875
poutou	poutou	1	0.73401	0.5180	1.42	0.1602	0.00885	112.93878
asselineau	asselineau	1	-1.03280	0.5310	-1.94	0.0552	0.01498	66.73959

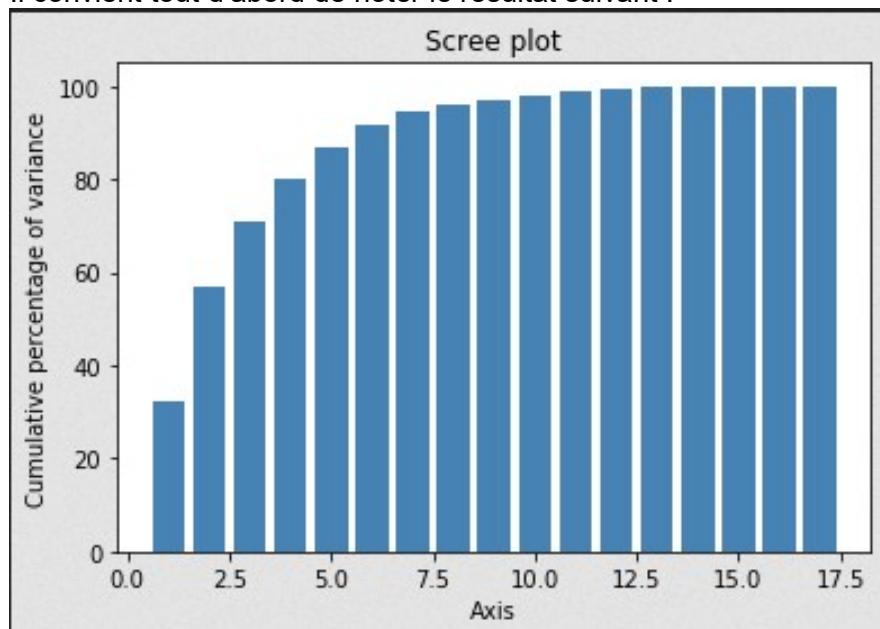
lassale	lassale	1	0.22139	0.0751		2.95	0.0042	0.15054	6.64291
arthaud	arthaud	1	-1.55026	0.6977		-2.22	0.0290	0.01399	71.49786
cheminade	cheminade	1	1.15075	4.0545		0.28	0.7773	0.00588	170.17665
blancs_nuls_1	blancs_nuls	1	0.36486	0.2404		1.52	0.1329	0.00777	128.68852

Dans les 4 modèles, on est en mesure de détecter une multi colinéarité très gênante pour l'interprétation des résultats

Pour espérer pallier à ce problème de multi colinéarité, nous décidons de passer par les composantes principales. En effet celle-ci sont par définition immunisé au problème de multi colinéarité car les directions de ces composantes sont par décomposition, orthogonales.

Modèle PCR.

Il convient tout d'abord de noter le résultat suivant :



Nombre de facteurs extraits	Variation de pourcentage expliquée par Composantes principales			
	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	27.4536	27.453	5.8291	5.8291
2	24.9922	52.445	48.5698	54.398
3	15.1032	67.549	7.3784	61.777
4	11.9879	79.536	1.8019	63.579
5	6.3759	85.912	5.2372	68.816
6	5.0819	90.994	20.9010	89.717

Variation de pourcentage expliquée par Composantes principales				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
7	2.8941	93.888	2.1394	91.856
8	2.1577	96.046	0.3882	92.245
9	1.3516	97.398	0.1772	92.422
10	1.1898	98.588	0.8079	93.230
11	0.7908	99.378	0.5100	93.740
12	0.6212	100.00	0.9930	94.733
13	0.0000	100.00	0.0000	94.733

On remarque que 3 composantes suffisent à capturer presque 70% de la variance et 6 composante 90% de la variance

Ainsi nous nous intéresserons au modèle PCR avec ces 3 composantes.

Abstention_2 :

Poids factoriels des effets du modèle									
Nombre de facteurs extraits	abstentions _1	le_pen_1	macron_1	melencho n	fillon	hamon	dupont_aignan	poutou	asselineau
1	0.220283	0.467788	-0.443528	-0.328219	-0.056982	-0.461466	0.307848	-0.0181	-0.00089
2	-0.306107	-0.063326	0.166847	0.084213	-0.107885	0.181907	0.243143	0.49477	-0.17347
3	-0.269735	-0.130258	0.261067	-0.267300	0.517243	0.008028	0.317481	-0.1875	0.227130
4	0.084522	0.040434	-0.071177	0.469370	-0.456291	0.101525	0.109967	-0.0620	0.603856
5	0.298553	-0.331610	-0.001643	-0.113381	0.218877	-0.156634	0.028329	0.03005	0.402763
6	-0.585740	0.250794	-0.102572	0.063256	-0.022402	-0.056470	0.428776	0.04122	0.243557
7	0.388288	-0.148692	0.192832	-0.391180	-0.290026	0.295173	0.428461	0.20016	0.047297
8	0.137127	-0.203502	0.062297	0.003870	0.032115	0.005536	0.251438	-0.1524	0.188818
9	-0.030058	0.085659	-0.231775	-0.140675	0.292655	0.085233	-0.387949	0.57753	0.462393
10	-0.176795	0.204138	0.070628	-0.321154	-0.131271	0.336412	-0.286853	-0.5239	0.236823
11	0.015586	-0.295278	-0.017207	0.391835	0.176684	-0.387258	0.084027	-0.1519	0.007839
12	0.101813	-0.099173	-0.644005	0.185298	0.313549	0.576665	0.225879	-0.1346	-0.15472
13	0	0	0	0	0	0	0	0	0

Poids factoriels des effets du modèle				
Nombre de facteurs extraits	lassale	arthaud	cheminade	blancs_nuls_1
1	-0.1293	0.24529	0.164772	0.113389
2	0.00616	0.41901	0.369031	0.420508
3	-0.4787	0.04405	0.195707	-0.214564
4	-0.3386	0.01658	0.207201	-0.093947
5	0.40239	-0.2988	0.419688	0.352610
6	0.44229	-0.2886	-0.218554	-0.064906
7	0.21456	0.10372	-0.110518	-0.413260
8	-0.1464	0.18185	-0.663154	0.564282
9	-0.0159	0.22946	-0.237147	-0.135397
10	0.28890	0.40054	0.130403	0.121377
11	0.33525	0.57619	-0.019723	-0.320259

Poids factoriels des effets du modèle				
Nombre de facteurs extraits	lassale	arthaud	cheminade	blancs_nuls_1
12	0.02726	0.01008	0.076828	-0.005308
13	0	0	0	0

Il s'agit des poids des variables sur les composantes factorielles.

Pour chaque variable quelles proportions de sa variance est expliqué par la composante principale :

pour 3 composantes :

Poids factoriels des effets du modèle										
Nombre de facteurs extraits	abstentions_1	le_pen_1	macron_1	melenchon	fillon	hamon	dupont_aignan	poutou	asselineau	
1	0.220283	0.46778	-0.44352	-0.328219	-0.0569	-0.4614	0.307848	-0.0181	-0.00089	
2	-0.306107	-0.0633	0.166847	0.084213	-0.1078	0.18190	0.243143	0.49477	-0.17347	
3	-0.269735	-0.1302	0.261067	-0.267300	0.51724	0.00802	0.317481	-0.1875	0.227130	

Poids factoriels des effets du modèle				
Nombre de facteurs extraits	lassale	arthaud	cheminade	blancs_nuls_1
1	-0.1293	0.2452	0.164772	0.113389
2	0.00616	0.4190	0.369031	0.420508
3	-0.4787	0.0440	0.195707	-0.214564

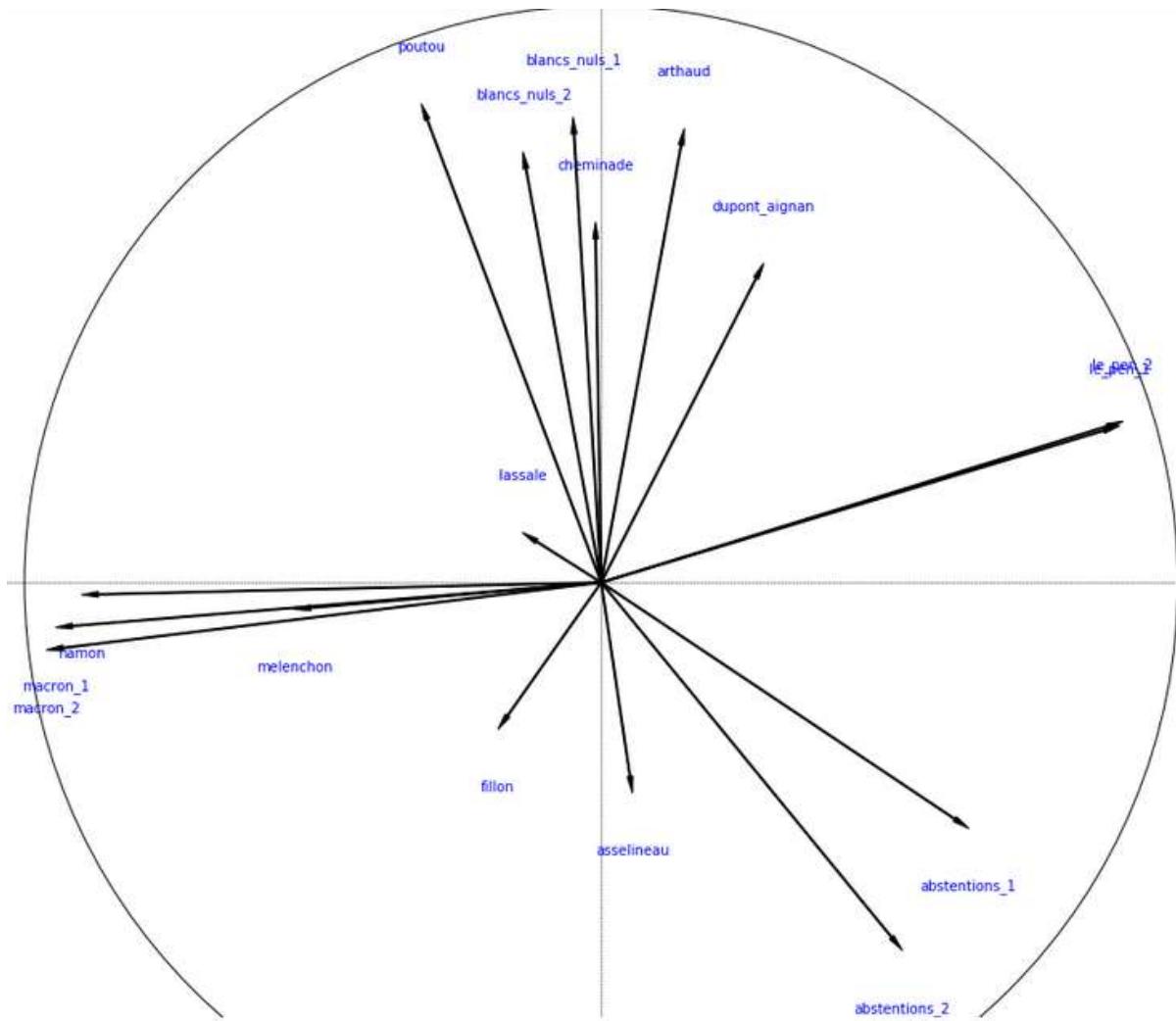
Poids des effets du modèle										
Nombre de facteurs extraits	abstentions_1	le_pen_1	macron_1	melenchon	fillon	hamon	dupont_aignan	poutou	asselineau	
1	0.220283	0.46778	-0.4435	-0.328219	-0.0569	-0.4614	0.307848	-0.0181	-0.0008	
2	-0.306107	-0.0633	0.166847	0.084213	-0.1078	0.18190	0.243143	0.49477	-0.17347	
3	-0.269735	-0.1302	0.261067	-0.267300	0.51724	0.00802	0.317481	-0.1875	0.227130	

Poids des effets du modèle				
Nombre de facteurs extraits	lassale	arthaud	cheminade	blances_nuls_1
1	-0.1293	0.2452	0.164772	0.113389
2	0.00616	0.4190	0.369031	0.420508
3	-0.4787	0.0440	0.195707	-0.214564

Pour 2 composantes :

Coefficients de régression codés pour 2 facteurs extraits	
	abstentions_2
abstentions_1	0.1465057799
le_pen_1	0.0842675963
macron_1	-.1211928507
melenchon	-.0745065235
fillon	0.0344305562
hamon	-.1293081395
dupont_aignan	-.0546664373
poutou	-.1936208716
asselineau	0.0669562575
lassale	-.0189130318
arthaud	-.1306589591
cheminade	-.1216251721
blances_nuls_1	-.1480950154

Dans ce dernier cas, attardons-nous sur le signe des coefficients. Nous allons les comparer avec le cercles de corrélation suivant
 On remarque que, macron_1, Hamon, Dupont Aignan, Poutou, Cheminade et Arthaud ont une contribution de signe négatif à l'abstention_2



Avec le cercle de corrélation des 2 premières composantes principales ci-dessus (et en excluant Fillon, Asselineau, Lassale, Mélenchon à cause de leur faible poids).

On voit des résultats cohérents avec ce qui est visible.

Macron 1, Hamon, Dupont Aignan, Poutou, Cheminade et Arthaud sont effectivement négativement correlés à l'abstention 2

On remarque également 2 clusters variables autour desquels on pourrait effectuer une RFCP dû à une suspicion d'unidimensionalité :

Groupe 1 : Macron, Hamon, Mélenchon,

Groupe 2 : Blanc_nuls, poutou, arthaud, cheminade, dupont-aignan

Root MSE	0.0173 3	R carré	0.617 8
Moyenne dépendante	0.2331 9	R car. ajust.	0.605 3
Coeff Var	7.4333 0		

On constate également un R carré inférieur à celui de la MCO qui est évident puisqu'il y a moins d'information recueilli mais cela reste significatif.

Comparons maintenant les coefficient MCO et PCR :

$Y = \text{abstentions_2}$

Variable	Coefficients	p-value	Corrélation avec Y calculée dans Q1	p-value
Intercept	-0.76651	0.0030		
abstentions_1	1.66481	<.0001	0.98391	<.0001
le_pen_1	0.82268	0.0013	-0.46097	<.0001
macron_1	0.84800	0.0015	-0.72044	<.0001
melenchon	1.11957	<.0001	-0.62127	<.0001
fillon	0.95948	0.0004	-0.51442	<.0001
hamon	0.29767	0.2154	-0.41669	<.0001
dupont_aignan	0.83649	0.0085	-0.76783	<.0001
poutou	0.77392	0.4253	-0.33497	0.0004
asselineau	-0.25708	0.8014	-0.14950	0.1243
lassale	0.66982	0.0338	-0.26202	0.0064
arthaud	-2.60098	0.0617	-0.28579	0.0028
cheminade	6.66799	0.2935	-0.38165	<.0001
blancs_nuls_1	0	.		

On se rend compte que les coefficients issus des 2 méthodes sont de signes différents, il convient donc de rejeter le modèle PCR comme candidat à l'explication de abstention_2

Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0			
	t_1	t_2	t_3
abstentions_1	0.41615	-0.55176	-0.37796
abstentions_1	<.0001	<.0001	0.0001
le_pen_1	0.88373	-0.11414	-0.18252
le_pen_1	<.0001	0.2681	0.0751
macron_1	-0.83790	0.30074	0.36581
macron_1	<.0001	0.0029	0.0002
melenchon	-0.62006	0.15179	-0.37455
melenchon	<.0001	0.1398	0.0002
fillon	-0.10765	-0.19446	0.72477
fillon	0.2965	0.0576	<.0001
hamon	-0.87179	0.32789	0.01125
hamon	<.0001	0.0011	0.9134
dupont_aignan	0.58158	0.43826	0.44486
dupont_aignan	<.0001	<.0001	<.0001
poutou	-0.03432	0.89182	-0.26287
poutou	0.7400	<.0001	0.0097
asselineau	-0.00170	-0.31268	0.31826
asselineau	0.9869	0.0019	0.0016
lassale	-0.24435	0.01111	-0.67085
lassale	0.0164	0.9145	<.0001
arthaud	0.46340	0.75527	0.06172
arthaud	<.0001	<.0001	0.5502
cheminade	0.31128	0.66518	0.27423
cheminade	0.0020	<.0001	0.0069
blancs_nuls_1	0.21421	0.75796	-0.30065
blancs_nuls_1	0.0361	<.0001	0.0029

On se rend compte de certains problèmes de significativités avec les 3 premières CP qui laissent à penser que le modèle ne serait pas capable de reconstituer toutes l'information.

Enfin terminons par l'importance relative des variables dans la prédiction de l'abstention_2 :

variable	vip
poutou	1.599
blancs_nuls_1	1.376
arthaud	1.368

variable	vip
cheminade	1.218
abstentions_1	1.063
dupont_aignan	0.936
macron_1	0.795
hamon	0.774
fillon	0.734
asselineau	0.623
lassale	0.614
le_pen_1	0.579
melenchon	0.562

Nous dégageons donc 5 variables vraiment utiles à la prédiction. Néanmoins on voit par exemple un souci au niveau de la significativité de Poutou dans la première composante principale alors que la variable est candidate à l'explication ce qui est problématique.

Blancs nul 2 :

Paramètres estimés	
	blances_nuls_2
Intercept	0.061574513
abstentions_1	-0.005824226
le_pen_1	0.008017996
macron_1	-0.015390085
melenchon	0.063269957
fillon	-0.102668493
hamon	0.062926485
dupont_aignan	-0.036792497
poutou	1.681210040
asselineau	-1.323676680
lassale	0.225740796
arthaud	1.196007840
cheminade	4.658656215
blances_nuls_1	0.880913601

On se rend compte que seul cheminade, arthaud, poutou et vote blanc sont vraiment explicatif du vote blanc_nuls au 2è tour mais curieusement pas l'abstention ce qui rend compte des limites du modèle.

variable	vip
poutou	1.599
blances_nuls_1	1.376
arthaud	1.368
cheminade	1.218
abstentions_1	1.063
dupont_aignan	0.936
macron_1	0.795
hamon	0.774
fillon	0.734
asselineau	0.623
lassale	0.614
le_pen_1	0.579
melenchon	0.562

Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0			
	t_1	t_2	t_3
abstentions_1	0.41615 <.0001	-0.55176 <.0001	-0.37796 0.0001
le_pen_1	0.88373 <.0001	-0.11414 0.2681	-0.18252 0.0751
macron_1	-0.83790 <.0001	0.30074 0.0029	0.36581 0.0002
melenchon	-0.62006 <.0001	0.15179 0.1398	-0.37455 0.0002
fillon	-0.10765 0.2965	-0.19446 0.0576	0.72477 <.0001
hamon	-0.87179 <.0001	0.32789 0.0011	0.01125 0.9134
dupont_aignan	0.58158 <.0001	0.43826 <.0001	0.44486 <.0001
poutou	-0.03432 0.7400	0.89182 <.0001	-0.26287 0.0097
asselineau	-0.00170 0.9869	-0.31268 0.0019	0.31826 0.0016
lassale	-0.24435 0.0164	0.01111 0.9145	-0.67085 <.0001
arthaud	0.46340 <.0001	0.75527 <.0001	0.06172 0.5502
cheminade	0.31128 0.0020	0.66518 <.0001	0.27423 0.0069
blances_nuls_1	0.21421 0.0361	0.75796 <.0001	-0.30065 0.0029

Quant à la significativité des composantes, on voit qu'il y a un problème avec la plupart des candidats bien qu'on ait capturé 70% de la variance ce qui est assez inquiétant.

Macron 2 :

Coefficients de régression codés pour 3 facteurs extraits	
	macron_2
abstentions_1	-0.2209411930
le_pen_1	-0.2579128729
macron_1	0.2987673506
melenchon	0.0877530368
fillon	0.1526180663
hamon	0.2391851358
dupont_aignan	-0.0150561897
poutou	0.0312639676
asselineau	0.0369756356
lassale	-0.0728177532
arthaud	-0.0356830740
cheminade	0.0352467198
blancs_nuls_1	-0.0471416771

On voit que Macron est très corrélé positivement avec lui-même, Hamon et Fillon ce qui est cohérent avec les résultats précédents. Cela avait été identifié lors d'une analyse précédente. Mais aussi très corrélé négativement avec l'abstention, Lepen, Mélenchon, Dupont Aignan.

variable	Variable importance
le_pen_1	1.489
hamon	1.468
macron_1	1.458
melenchon	1.107
dupont_aignan	1.099
arthaud	0.883
abstentions_1	0.852
lassale	0.796
fillon	0.767
cheminade	0.697
blancs_nuls_1	0.633

variable	Variable importance
poutou	0.567
asselineau	0.368

Lepen_2 :

Coefficients de régression codés pour 3 facteurs extraits	
	le_pen_2
abstentions_1	0.1635157601
le_pen_1	0.2427279142
macron_1	-.2575515479
melenchon	-.1202063722
fillon	-.0955689563
hamon	-.2297444834
dupont_aignan	0.0812891685
poutou	-.0141417877
asselineau	-.0221881586
lassale	0.0093838368
arthaud	0.0803809671
cheminade	0.0236737792
blancs_nuls_1	0.0564882103

variable	vip
le_pen_1	1.687
hamon	1.664
macron_1	1.599
melenchon	1.183
dupont_aignan	1.110
arthaud	0.884
abstentions_1	0.794
cheminade	0.594
lassale	0.466
blancs_nuls_1	0.409
fillon	0.205

variable	vip
poutou	0.065
asselineau	0.003

Lepen_1, hamon, macron_1, melenchon et dupont_aignan sont les variables qui expliquent le mieux lepen_2

On a ainsi pu obtenir une bien meilleur explicativité des variables à l'aide de la PCR néanmoins, il reste quelques soucis de significativité nous poussant vers un modèle qui maximisent la corrélation entre les composantes principales et les variables explicatives.

L'approche PLS est plus intéressante pour cela.

Modèle PLS :

L'approche PLS a été développé pour pallier aux problèmes suivants :

- Forte multicolinéarité entre les variables
- Beaucoup de variable explicatives comparés aux nombre d'individus
- Besoin d'interpréter les coefficients de régression en adéquation avec la réalité pratique

L'Évaluation de la qualité du modèle se fait par validation croisée.

Lepen_2 :

Racine carrée du PRESS moyen min.	0.267 8
Réduction du nombre de facteurs	9
Le plus petit nombre de facteurs avec p > 0.1	5

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
0	1.010526	7.52996	<.0001
1	0.602784	9.285646	<.0001
2	0.40816	12.57141	<.0001
3	0.322238	7.007849	0.0010
4	0.302776	4.579156	0.0240
5	0.28284	1.481129	0.2250

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
6	0.277386	2.247714	0.1320
7	0.272697	1.230632	0.2970
8	0.269495	0.690907	0.4220
9	0.267755	0	1.0000
10	0.26783	0.015659	0.9140
11	0.267866	0.027869	0.8780
12	0.267828	0.012192	0.9170
13	0.267828	0.012192	0.9170

On voit que seule les 5 premiers facteurs sont significatifs. Attardons-nous ceux-ci

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	24.3407	24.3407	73.2139	73.2139
2	10.8456	35.1863	15.7522	88.9661
3	16.0073	51.1936	3.5205	92.4866
4	20.2568	71.4504	0.7142	93.2008
5	6.7708	78.2213	0.7359	93.9367

On voit qu'ils expliquent 78% de la variance des variables explicatives. On remarque également que le premier facteur suffit à expliquer 73% de la variable dépendante c'est tout de suite beaucoup plus encourageant. 93% de la variable dépendante est expliqués par les 5 facteurs projeté sur l'axe PLS

Paramètres estimés	
	abstentions_2
Intercept	0.201891886
abstentions_1	0.810205561
le_pen_1	-0.156274414
macron_1	-0.135571129
melenchon	-0.019402071
fillon	-0.052457320
hamon	-0.166899451
dupont_aignan	-0.387902989
poutou	-2.129299545
asselineau	-0.714043764
lassale	-0.230112944
arthaud	-0.113225191
cheminade	-2.204146976
blancs_nuls_1	-0.644381561

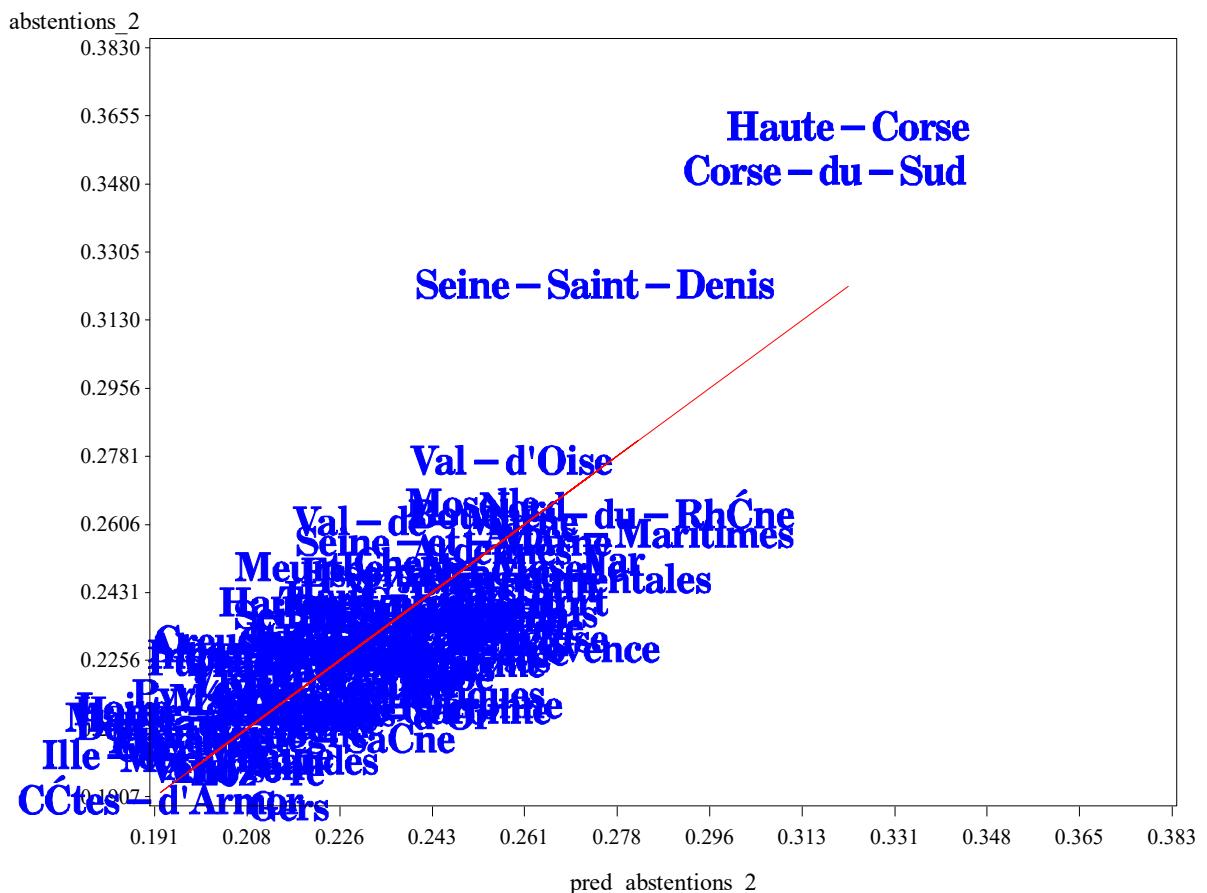
Regardons l'importance des variables :

variable	vip
abstentions_1	2.374
poutou	1.438
macron_1	1.232
hamon	1.060
arthaud	0.951
dupont_aignan	0.889
blancs_nuls_1	0.869
cheminade	0.739
asselineau	0.549
le_pen_1	0.379
melenchon	0.329
fillon	0.312
lassale	0.023

Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0	
	t_1
abstentions_2	0.85565
abstentions_2	<.0001
abstentions_1	0.81196
abstentions_1	<.0001
le_pen_1	0.39743
le_pen_1	<.0001
macron_1	-0.63039
macron_1	<.0001
melenchon	-0.23867
melenchon	0.0192
fillon	-0.04905
fillon	0.6351
hamon	-0.58751
hamon	<.0001
dupont_aignan	-0.31000
dupont_aignan	0.0021
poutou	-0.76485
poutou	<.0001
asselineau	0.28006
asselineau	0.0057
lassale	-0.00010
lassale	0.9992
arthaud	-0.52332
arthaud	<.0001
cheminade	-0.48244
cheminade	<.0001
blancs_nuls_1	-0.52819
blancs_nuls_1	<.0001

On voit que ce sont les mêmes 4 variables candidates à l'explication de abstention_2 que cité précédemment et elles sont significativement corrélés à l'axe PLS.

Valeurs observées vs valeurs prédictives de abstentions_2 avec 1 composante PLS



En dehors de quelques valeurs extrêmes, les données sont plutôt bien reconstitués

Blanc nul 2 :

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
0	1.010526	22.30442	<.0001
1	0.623223	12.35519	0.0010
2	0.527452	4.338987	0.0390
3	0.494878	2.834729	0.0880
4	0.469528	0.497269	0.5640
5	0.464202	0.315145	0.5920
6	0.470317	2.283621	0.1390
7	0.466522	2.059264	0.1360

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
8	0.460384	6.539169	0.0070
9	0.456204	0	1.0000
10	0.456792	1.666956	0.2290
11	0.45669	0.883037	0.4060
12	0.456707	0.917144	0.3990
13	0.456707	0.917144	0.3990

Racine carrée du PRESS moyen min.	0.456 2
Réduction du nombre de facteurs	9
Le plus petit nombre de facteurs avec p > 0.1	4

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	22.8415	22.841 5	68.1149	68.114 9
2	14.4838	37.325 3	10.5633	78.678 3
3	7.6713	44.996 5	4.9020	83.580 3
4	11.5112	56.507 7	0.7903	84.370 5

On voit que seulement 56% de la variance est expliqué avec 4 facteur ce qui laisse présager des performances mitigés

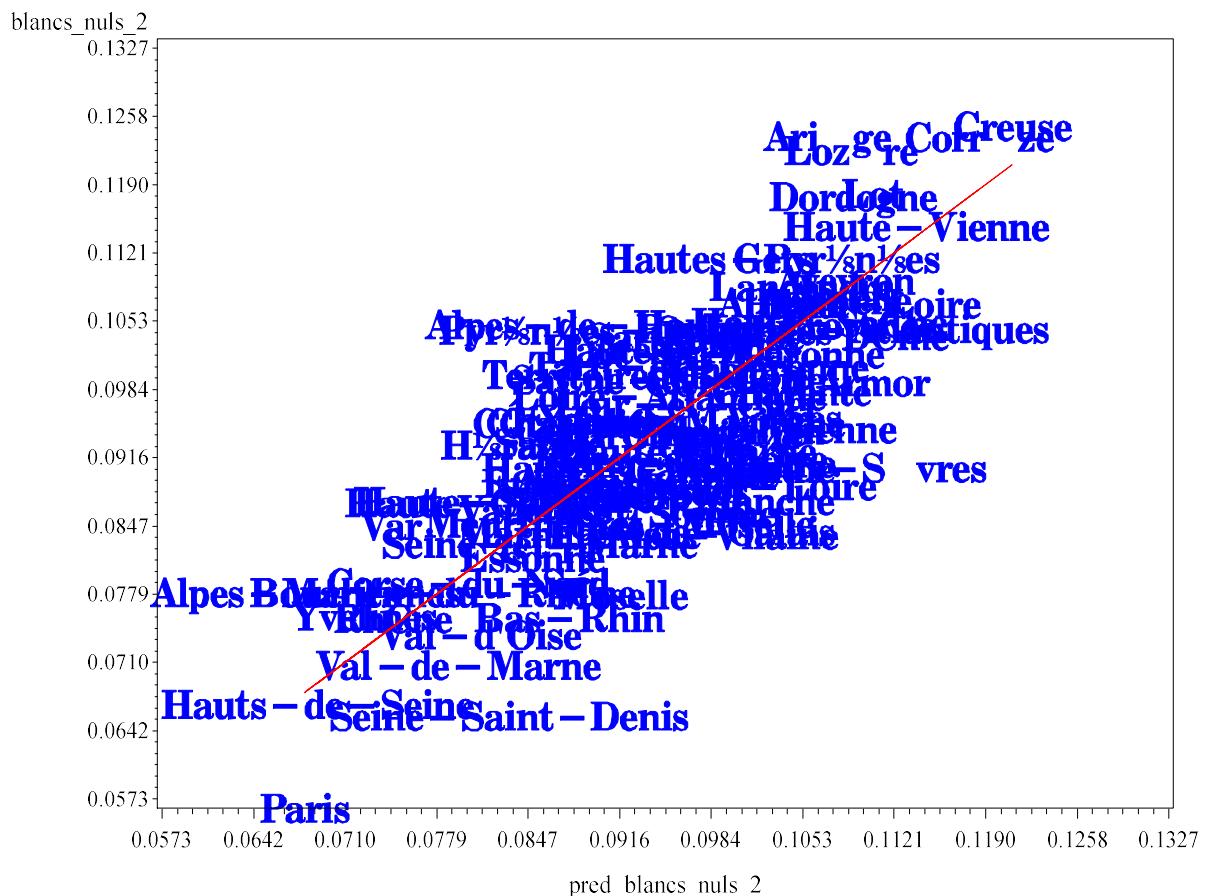
Paramètres estimés	
	blancs_nuls_2
Intercept	0.065603923
abstentions_1	-0.164316519
le_pen_1	0.033354337
macron_1	-0.094583406
melenchon	0.120906082
fillon	0.002292398
hamon	-0.015293259
dupont_aignan	0.025800274
poutou	1.276253541
asselineau	-0.849341610
lassale	0.489153045
arthaud	-1.520063028
cheminade	8.314251008
blancs_nuls_1	1.841766618

variable	vip
blancs_nuls_1	2.064
poutou	1.853
lassale	1.452
asselineau	1.009
abstentions_1	0.886
fillon	0.803
cheminade	0.794
arthaud	0.790
melenchon	0.605
hamon	0.284
dupont_aignan	0.283
macron_1	0.178
le_pen_1	0.152

Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0	
	t_1
blancs_nuls_2	0.82532
blancs_nuls_2	<.0001
abstentions_1	-0.37035
abstentions_1	0.0002
le_pen_1	-0.00508
le_pen_1	0.9608
macron_1	0.09020
macron_1	0.3821
melenchon	0.21248
melenchon	0.0377
fillon	-0.39187
fillon	<.0001
hamon	0.22815
hamon	0.0254
dupont_aignan	0.23062
dupont_aignan	0.0238
poutou	0.91405
poutou	<.0001
asselineau	-0.45880
asselineau	<.0001
lassale	0.38051
lassale	0.0001
arthaud	0.61798
arthaud	<.0001
cheminade	0.47425
cheminade	<.0001
blancs_nuls_1	0.85002
blancs_nuls_1	<.0001

Les variables les plus importante sont significatives c'est très bien !

Valeurs observées vs valeurs prédictes de blancs_nuls_2 avec 1 composantes PLS



Il y a une plus grande erreur de prédiction qui s'explique par une partie de la variance qui n'a malheureusement pas été capturé. Ce qui explique le PRESS de 0.5 qui est relativement élevé et indique un modèle fragile.

Macron 2 :

Croisement Validation du nombre de facteurs extraits		
Nombre de facteurs extraits	Racine carrée du PRESS moyen	Prob > PRESS
0	1.010526	<.0001
1	0.281945	<.0001
2	0.18387	0.0010
3	0.169559	0.0020
4	0.157661	0.0080
5	0.151952	0.0150

Croisement Validation du nombre de facteurs extraits		
Nombre de facteurs extraits	Racine carrée du PRESS moyen	Prob > PRESS
6	0.138031	0.0210
7	0.132215	0.0390
8	0.12954	0.1660
9	0.12818	0.1180
10	0.127476	1.0000
11	0.127694	0.2550
12	0.127698	0.2480
13	0.127698	0.2480

Racine carrée du PRESS moyen min.	0.1275
Réduction du nombre de facteurs	10
Le plus petit nombre de facteurs avec $p > 0.1$	8

8 Composante significative !!

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	26.5911	26.5911	93.7620	93.7620
2	15.9301	42.5213	3.5780	97.3400
3	15.9917	58.5129	0.5868	97.9268
4	13.2680	71.7809	0.4590	98.3857
5	4.2369	76.0178	0.3274	98.7131
6	6.7275	82.7452	0.1277	98.8408
7	6.8369	89.5821	0.0859	98.9268
8	4.4057	93.9879	0.0242	98.9510

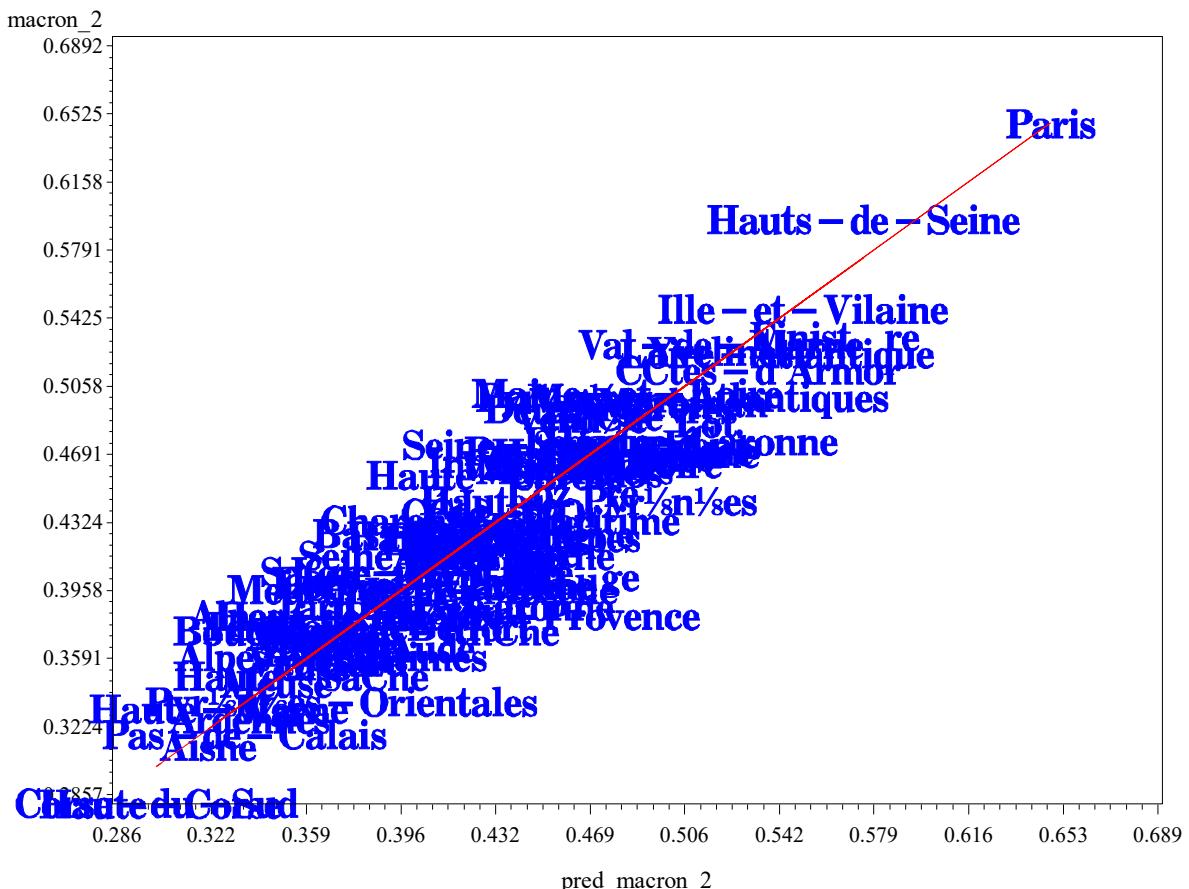
On capture avec ces 8 premières composantes 93% de la variance des variables explicatives et 98% de la variable dépendante. On peut espérer d'excellent résultat

Paramètres estimés	
	macron_2
Intercept	0.28113747
abstentions_1	-0.25069723
le_pen_1	-0.58791351
macron_1	1.01739225
melenchon	0.22100521
fillon	0.27173023
hamon	0.21336658
dupont_aignan	0.29099117
poutou	0.99320661
asselineau	4.36637728
lassale	0.01076176
arthaud	4.27231614
cheminade	-12.91958124
blancs_nuls_1	-0.96466805

variable	vip
macron_1	1.988
le_pen_1	1.875
hamon	1.683
abstentions_1	1.197
melenchon	0.811
fillon	0.678
dupont_aignan	0.399
poutou	0.358
arthaud	0.322
blancs_nuls_1	0.137
asselineau	0.115
cheminade	0.057
lassale	0.039

Coefficients de corrélation de Pearson, N = 96	
Proba > r sous H0: Rho=0	
	t_1
macron_2	0.96831
macron_2	<.0001
abstentions_1	-0.61321
abstentions_1	<.0001
le_pen_1	-0.90915
le_pen_1	<.0001
macron_1	0.95192
macron_1	<.0001
melenchon	0.52420
melenchon	<.0001
fillon	0.22343
fillon	0.0287
hamon	0.90294
hamon	<.0001
dupont_aignan	-0.32217
dupont_aignan	0.0014
poutou	0.21365
poutou	0.0366
asselineau	-0.00840
asselineau	0.9353
lassale	0.08330
lassale	0.4197
arthaud	-0.21605
arthaud	0.0345
cheminade	-0.04161
cheminade	0.6873
blances_nuls_1	-0.05713
blances_nuls_1	0.5803

Valeurs observées vs valeurs prédites de macron_2 avec 1 composantes PLS



On voit ainsi une excellent adéquation du modèle !!

Le pen_2 :

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
0	1.010526	30.33287	<.0001
1	0.368728	27.82044	<.0001
2	0.263616	8.435024	<.0001
3	0.166237	15.34147	<.0001
4	0.143852	18.33574	<.0001
5	0.117399	9.319508	0.0010
6	0.102294	4.775808	0.0150
7	0.092772	1.405739	0.3040
8	0.089622	1.412862	0.2820
9	0.088057	0.111182	0.7640
10	0.087872	0	1.0000

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
11	0.087934	0.881056	0.3510
12	0.087942	0.907933	0.3340
13	0.087942	0.907933	0.3340

Racine carrée du PRESS moyen min.	0.087 9
Réduction du nombre de facteurs	10
Le plus petit nombre de facteurs avec p > 0.1	7

On garde les 7 premiers facteurs

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	27.0768	27.0768	88.2868	88.2868
2	10.3702	37.4470	7.8777	96.1645
3	12.2969	49.7439	1.9397	98.1043
4	20.4151	70.1590	0.3574	98.4617
5	6.0994	76.2584	0.6436	99.1052
6	4.2912	80.5496	0.2074	99.3127
7	7.1258	87.6754	0.0703	99.3830

Le modèle explique extrêmement bien la variance de la variable dépendante et 87% de la variance des variables explicatives. Les variables les plus importantes sont le_pen_1, macron_1, hamon.

variable	vip
le_pen_1	2.086
macron_1	1.865
hamon	1.650
melenchon	0.957
dupont_aignan	0.800
arthaud	0.638

variable	vip
abstentions_1	0.629
fillon	0.545
lassale	0.210
asselineau	0.182
poutou	0.178
blancs_nuls_1	0.147
cheminade	0.092

Paramètres estimés	
	le_pen_2
Intercept	0.472952316
abstentions_1	-0.434285291
le_pen_1	0.733343256
macron_1	-0.602110958
melenchon	-0.392728893
fillon	-0.324000620
hamon	-0.428280580
dupont_aignan	-0.397183125
poutou	0.130865542
asselineau	-1.393146012
lassale	-0.200440616
arthaud	-0.971433839
cheminade	4.466208117
blancs_nuls_1	-0.526761138

Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0	
	t_1
le_pen_2	0.93961
le_pen_2	<.0001
abstentions_1	0.47938
abstentions_1	<.0001
le_pen_1	0.93569
le_pen_1	<.0001
macron_1	-0.90989
macron_1	<.0001
melenchon	-0.57109
melenchon	<.0001
fillon	-0.18278
fillon	0.0747
hamon	-0.89121
hamon	<.0001
dupont_aignan	0.46944
dupont_aignan	<.0001
poutou	-0.11010
poutou	0.2856
asselineau	-0.02255
asselineau	0.8274
lassale	-0.16407
lassale	0.1102
arthaud	0.36057
arthaud	0.0003
cheminade	0.15943
cheminade	0.1208
blancs_nuls_1	0.13268
blancs_nuls_1	0.1975

Elles sont toutes significativement corrélées à l'axe PLS.

Avec ces 4 modèles de l'approche PLS nous obtenons des performances significativement meilleures qu'avec l'approche PCR, ce qui s'explique par des composantes principales souvent plus significatives.

Question 4 :

Dans cette partie, il s'agit de faire des préditions sur plusieurs variables en même temps et confronter l'approche PCR et PLS

Modèle PCR :

Variation de pourcentage expliquée par Composantes principales				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	27.4536	27.4536	39.4525	39.4525
2	24.9922	52.4459	24.0615	63.5140
3	15.1032	67.5490	12.3583	75.8723
4	11.9879	79.5369	1.1678	77.0401
5	6.3759	85.9128	4.0095	81.0497
6	5.0819	90.9947	8.7803	89.8299
7	2.8941	93.8889	2.5186	92.3486
8	2.1577	96.0466	0.5221	92.8707
9	1.3516	97.3982	0.1114	92.9821
10	1.1898	98.5880	0.3129	93.2949
11	0.7908	99.3788	0.6376	93.9325
12	0.6212	100.0000	0.7829	94.7154
13	0.0000	100.0000	0.0000	94.7154

Avec seulement 3 composante on parvient toujours à capturer 67% de la variance sur les variables explicatives, 75% de la variance sur les variables dépendantes !!

C'est le nombre de facteurs que l'on choisit de garder par la suite car il est évidemment optimal.

Coefficients de régression codés pour 3 facteurs extraits				
	abstentions_2	blancs_nuls_2	macron_2	le_pen_2
abstentions_1	0.1987953078	-.0117042915	-.2209411930	0.1635157601
le_pen_1	0.1095188260	0.0270658265	-.2579128729	0.2427279142
macron_1	-.1718020270	-.0356261030	0.2987673506	-.2575515479
melenchon	-.0226890891	0.1169225548	0.0877530368	-.1202063722
fillon	-.0658394364	-.2139953076	0.1526180663	-.0955689563
hamon	-.1308644820	0.0555861969	0.2391851358	-.2297444834
dupont_aignan	-.1162117285	-.0211438154	-.0150561897	0.0812891685
poutou	-.1572539176	0.2346149494	0.0312639676	-.0141417877
asselineau	0.0229259614	-.1372873939	0.0369756356	-.0221881586
lassale	0.0738967897	0.1639315154	-.0728177532	0.0093838368
arthaud	-.1391983546	0.1321219229	-.0356830740	0.0803809671
cheminade	-.1595639140	0.0624167196	0.0352467198	0.0236737792
blancs_nuls_1	-.1065007448	0.2194343640	-.0471416771	0.0564882103

Paramètres estimés				
	abstentions_2	blancs_nuls_2	macron_2	le_pen_2
Intercept	0.18016429	0.05966823	0.28382230	0.47634517
abstentions_1	0.83758073	-0.14402871 0.25731032	-	-0.43624170
le_pen_1	-0.20210054	0.03767325 0.57886999	-	0.74329728
macron_1	-0.26362225	-0.18318527	1.03969563	-0.59288811
melenchon	0.12540385	0.12117765	0.19308568	-0.43966717
fillon	0.03742496	0.04231963	0.25348408	-0.33322868
hamon	-0.06165152	0.23680782	0.19610787	-0.37126417
dupont_aignan	-0.01545659	0.11805139	0.24532777	-0.34792256
poutou	-2.53733717	1.06967364	1.21000089	0.25766269
asselineau	-1.36260502	-1.76932494	4.64107971	-1.50914979
lassale	-0.16268695	0.39508603	0.02255501	-0.25495411
arthaud	0.25125788	-2.34048342	4.11582715	-2.02660169

Paramètres estimés				
	abstentions_2	blancs_nuls_2	macron_2	le_pen_2
cheminade	0.52127916	11.10563457	- 12.3013200 7	0.67440627
blancs_nuls_1	-0.93287762	2.07820895	- 1.03384371	-0.11148760

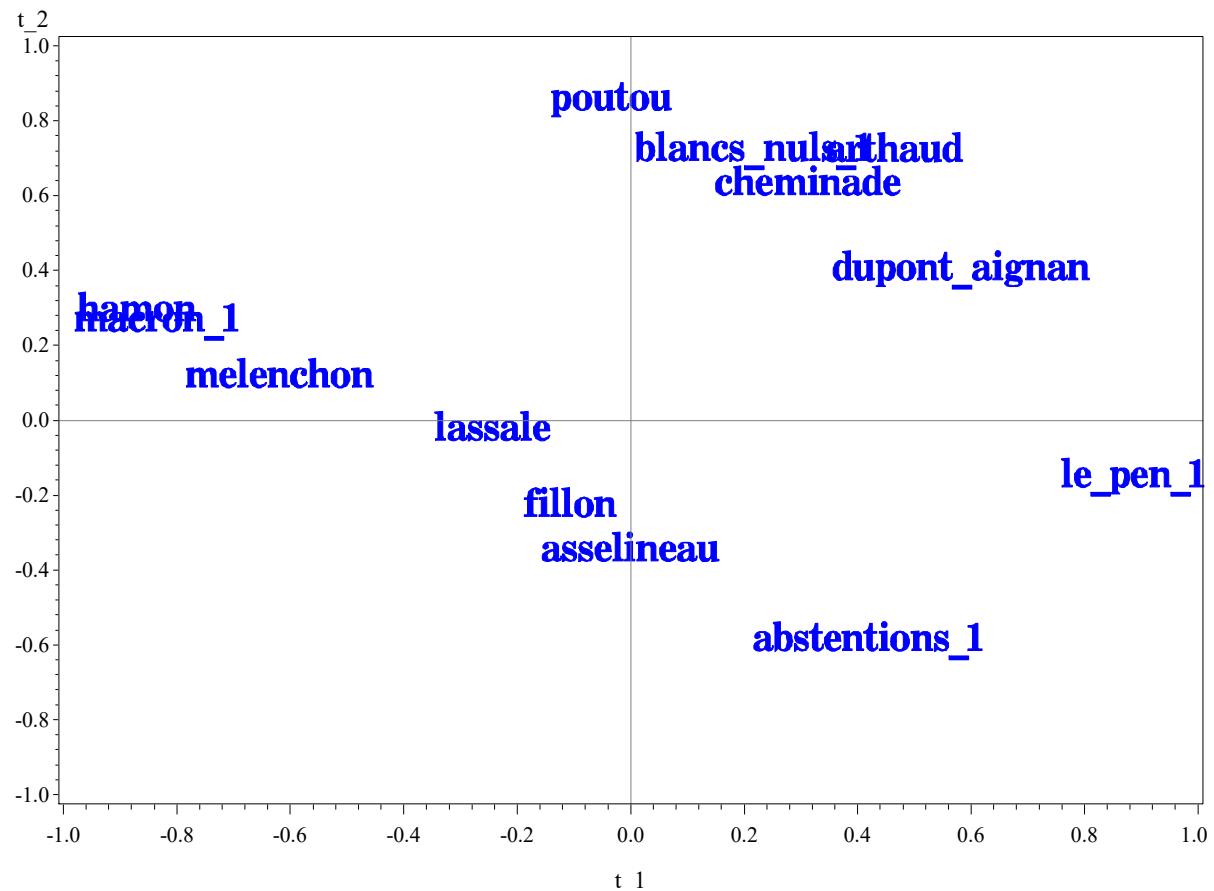
variable	vip
hamon	1.609
macron_1	1.458
melenchon	1.164
dupont_aignan	1.108
abstentions_1	1.063
arthaud	0.878
cheminade	0.620
lassale	0.594
blancs_nuls_1	0.475
fillon	0.466
le_pen_1	0.341
poutou	0.287
asselineau	0.202

On obtient 5 variables candidates à l'explication du modèle

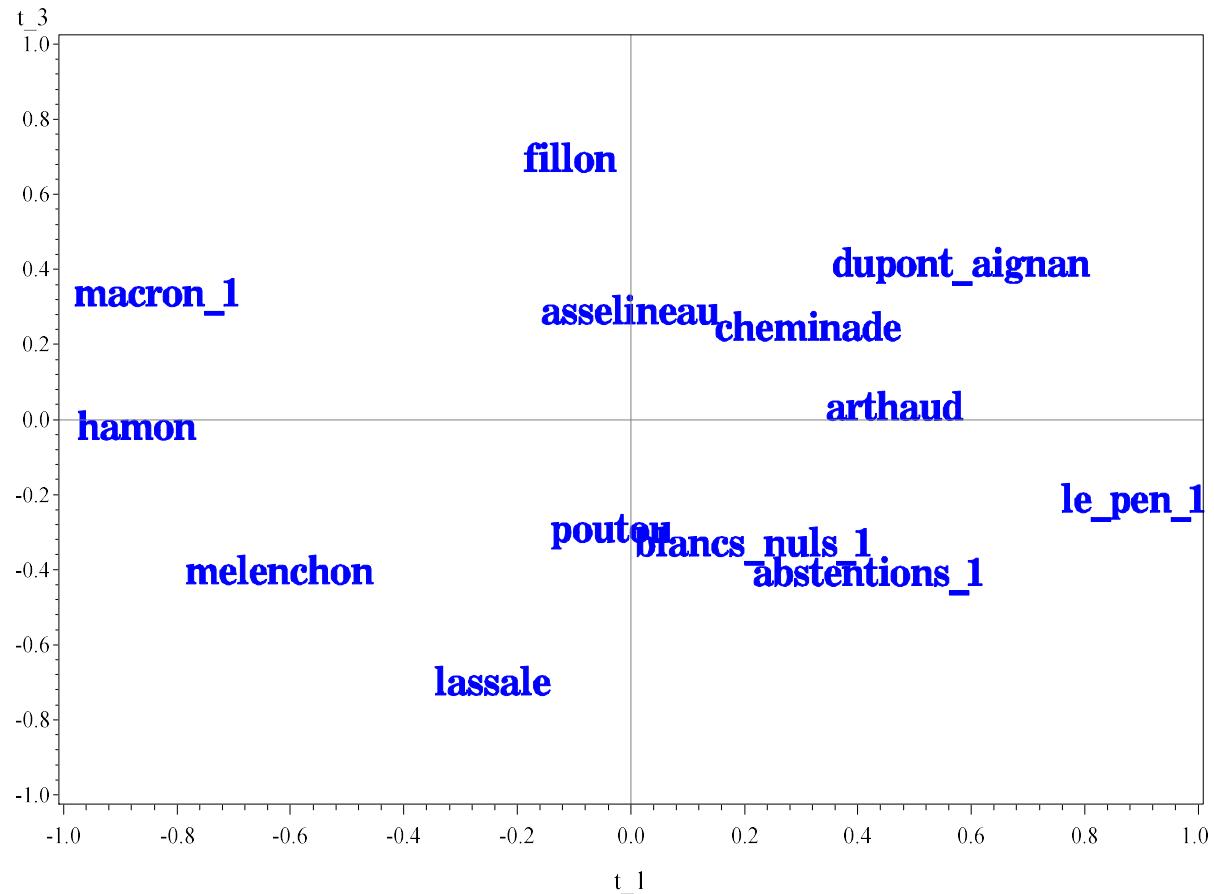
Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0			
	t_1	t_2	t_3
abstentions_1	0.41615	-0.55176	-0.37796
abstentions_1	<.0001	<.0001	0.0001
le_pen_1	0.88373	-0.11414	-0.18252
le_pen_1	<.0001	0.2681	0.0751
macron_1	-0.83790	0.30074	0.36581
macron_1	<.0001	0.0029	0.0002
melenchon	-0.62006	0.15179	-0.37455
melenchon	<.0001	0.1398	0.0002
fillon	-0.10765	-0.19446	0.72477
fillon	0.2965	0.0576	<.0001
hamon	-0.87179	0.32789	0.01125
hamon	<.0001	0.0011	0.9134
dupont_aignan	0.58158	0.43826	0.44486
dupont_aignan	<.0001	<.0001	<.0001
poutou	-0.03432	0.89182	-0.26287
poutou	0.7400	<.0001	0.0097
asselineau	-0.00170	-0.31268	0.31826
asselineau	0.9869	0.0019	0.0016
lassale	-0.24435	0.01111	-0.67085
lassale	0.0164	0.9145	<.0001
arthaud	0.46340	0.75527	0.06172
arthaud	<.0001	<.0001	0.5502
cheminade	0.31128	0.66518	0.27423
cheminade	0.0020	<.0001	0.0069
blancs_nuls_1	0.21421	0.75796	-0.30065
blancs_nuls_1	0.0361	<.0001	0.0029

Cette fois on remarque de bien meilleur résultat en terme de significativité que lorsque les variables dépendantes étaient séparées. On note tout de même un souci pour Macron_1 selon la 2^e composante et Hamon selon la 3^e alors que ces 2 variables sont censés être les variables pertinentes quant à l'explication du modèle.

Carte des variables sur les deux premières composantes PCR : t_1 et t_2

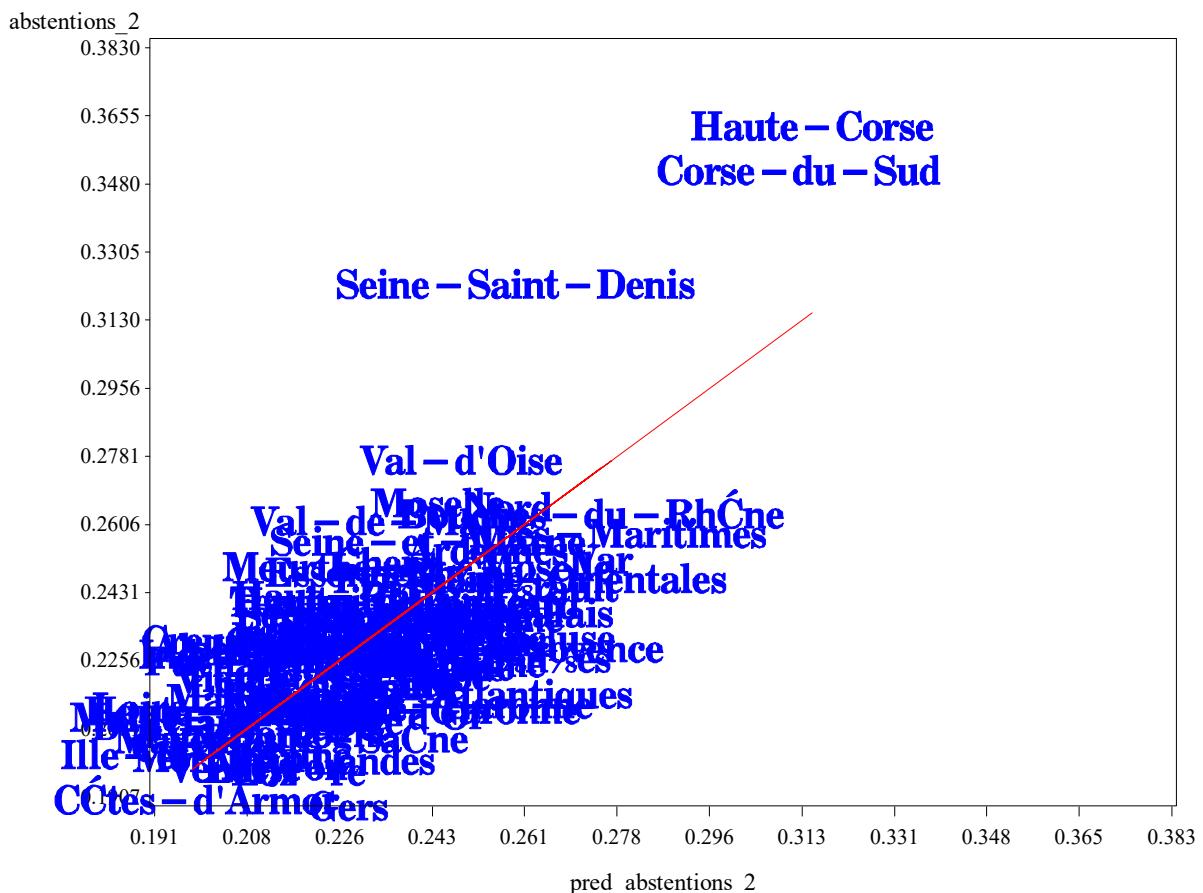


Carte des variables sur la deuxième et la troisième composantes PCR : t_1 et t_3



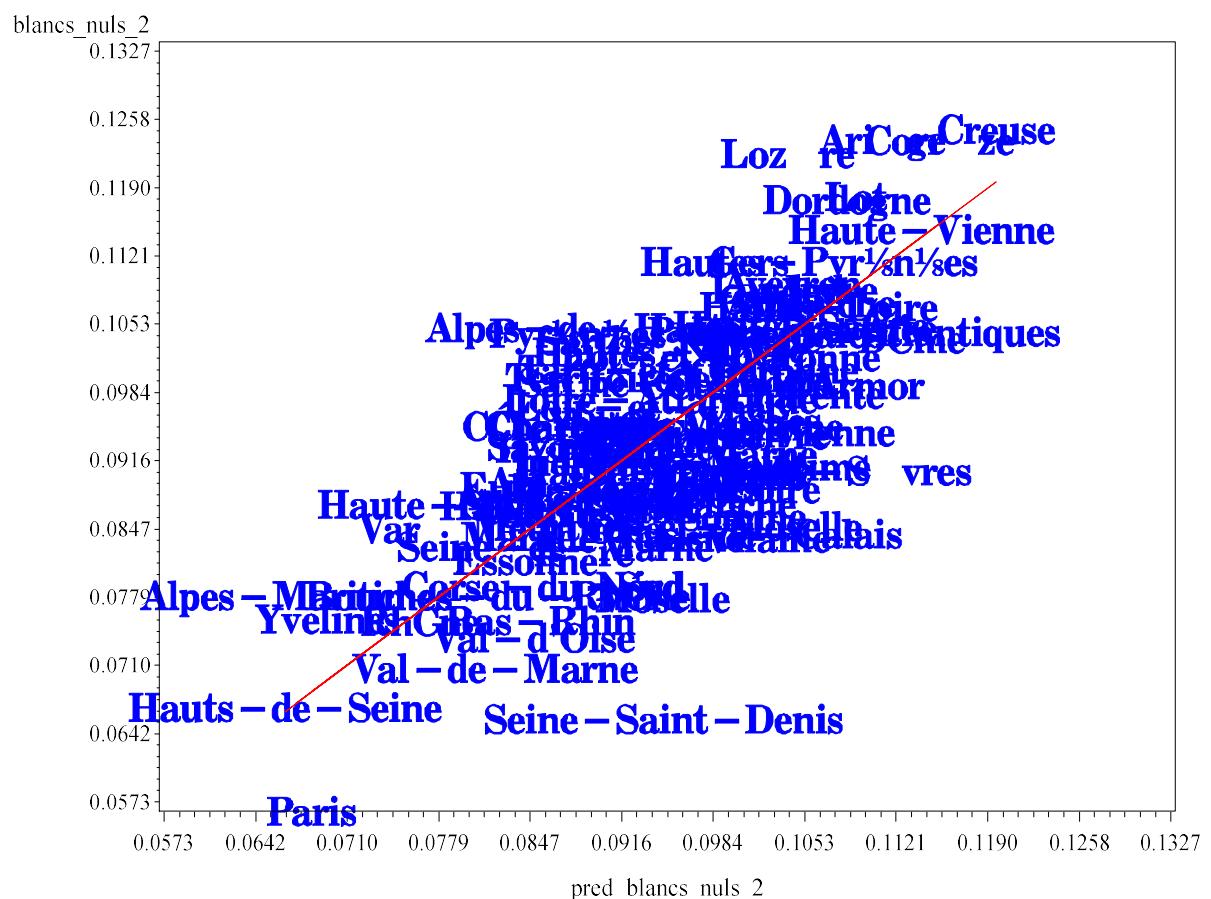
On voit effectivement que mélenchon semble en retrait quant à sa projection sur les 3 axes comparés aux autres variables pertinentes à l'explication. Il convient de noter la disparition de arthaud, cheminade, lassale, blanc_nul_1, fillon, le_pen, poutou, asselineau lors de la résolution des problèmes de multicolinéarité. Il nous suffit lors d'une sélection des variables, des résultats de 5 candidats pour mesure l'ensemble des quantités sous-jacentes nécessaire à l'établissement d'un modèle performant.

Valeurs observées vs valeurs prédictives de abstentions_2 avec 3 composantes PCR



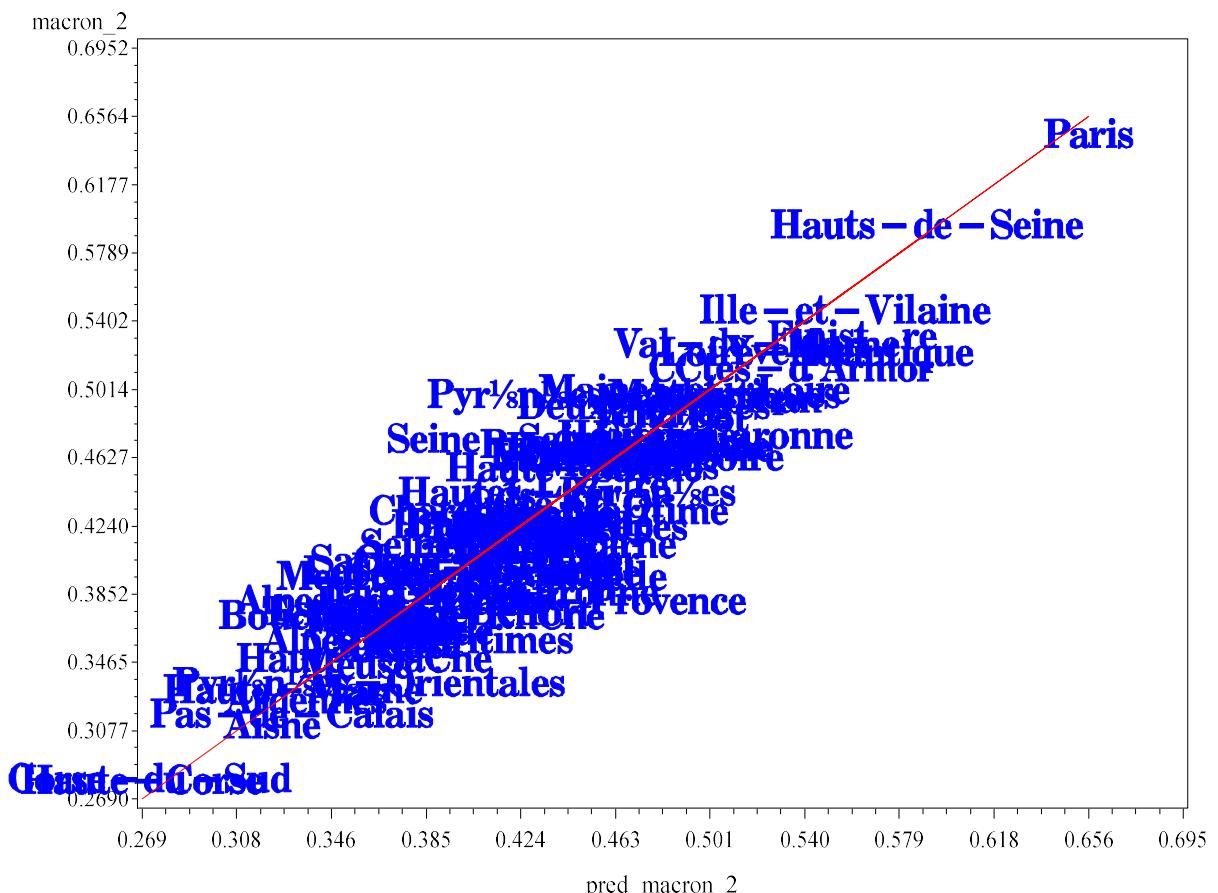
On observe encore l'impact des 3 valeurs extrêmes Seine-Saint-Denis, Haute-Corse et Corse-du-Sud qui nuisent à l'explication. Avec le recul il serait peut-être pertinent de les éliminer avec les autres DOM-TOM.

Valeurs observées vs valeurs prédictes de blancs_nuls_2 avec 3 composantes PCR

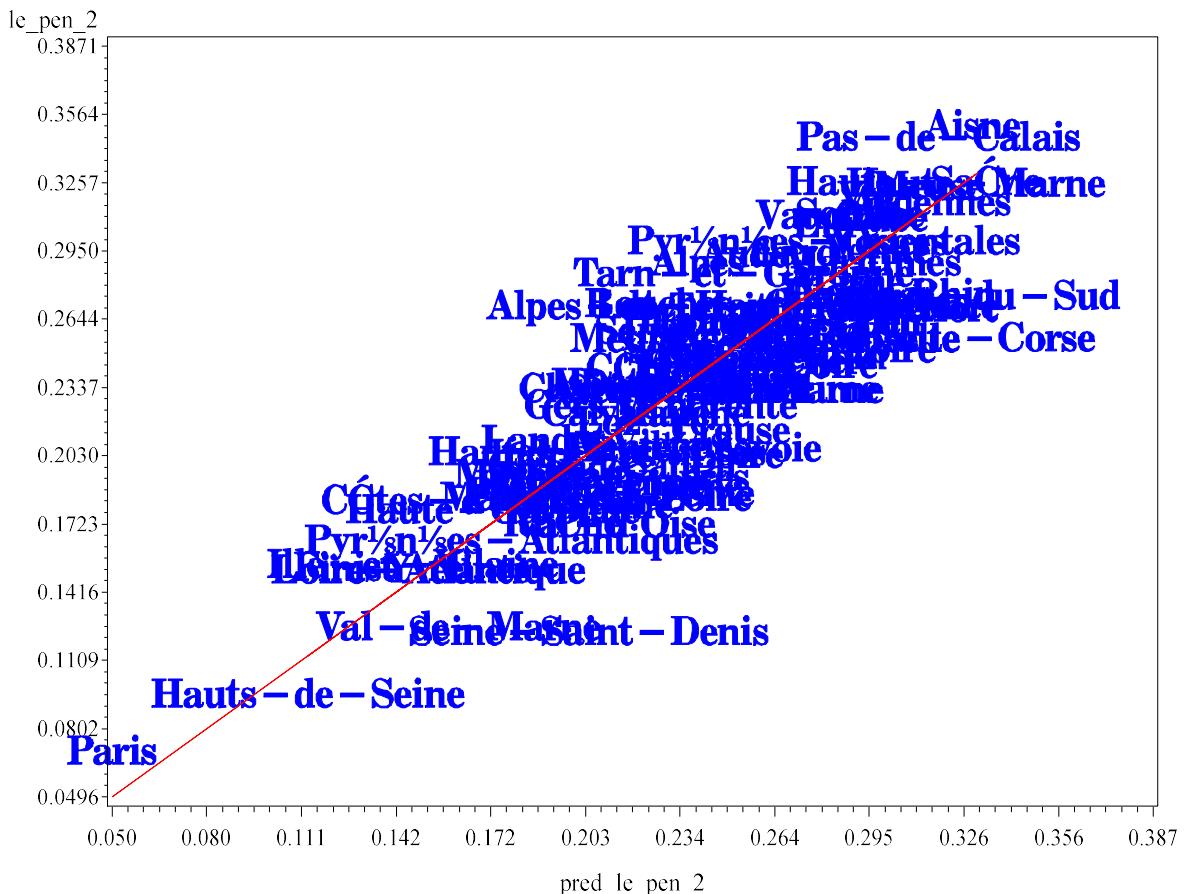


On observe une erreur relativement importante qui est dû au fait que Cheminade est la variable apportant la contribution la plus significative à la prédition et celui-ci n'a pas une corrélation très forte avec les 3 composantes principales. Ce qui se traduit par une variance plus marqué ci-dessus.

Valeurs observées vs valeurs prédictes de macron_2 avec 3 composantes PCR



Valeurs observées vs valeurs prédictes de le_pen_2 avec 3 composantes PCR



Le modèle PCR n'a pas su apporter une bonne prédition pour la variable blanc_nuls_2 mais a des résultats probant sur 3 autres variables dépendantes.

Il y a encore quelques soucis de significativité et surtout le regret de ne pas avoir pu capturer plus de variances en un minimum de facteur.

Modèle PLS2 :

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
0	1.010526	37.32456	<.0001
1	0.742401	22.96201	<.0001
2	0.552198	37.78409	<.0001
3	0.471687	33.00883	<.0001
4	0.371705	24.92216	<.0001

Croisement Validation du nombre de facteurs extraits			
Nombre de facteurs extraits	Racine carrée du PRESS moyen	T**2	Prob > T**2
5	0.33903	18.71956	<.0001
6	0.310354	19.48605	<.0001
7	0.301352	18.62606	<.0001
8	0.289617	14.72609	0.0010
9	0.28413	8.159919	0.0520
10	0.277988	3.317639	0.5510
11	0.276602	7.010542	0.0790
12	0.275841	0	1.0000
13	0.275841	0	<.0001

Racine carrée du PRESS moyen min.	0.275 8
Réduction du nombre de facteurs	12
Le plus petit nombre de facteurs avec p > 0.1	10

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels				
Nombre de facteurs extraits	Effets du modèle		Variables dépendantes	
	En cours	Total	En cours	Total
1	26.6449	26.6449	50.2880	50.2880
2	24.2569	50.9017	24.0644	74.3524
3	15.5021	66.4038	8.0667	82.4191
4	6.6866	73.0904	7.8453	90.2645
5	7.2571	80.3475	1.9320	92.1965
6	9.5634	89.9109	0.8316	93.0281
7	3.5096	93.4205	0.5836	93.6117
8	1.6953	95.1159	0.5464	94.1581
9	1.3410	96.4569	0.3762	94.5343
10	1.2801	97.7370	0.1437	94.6780

Avec 10 facteurs significatif gardés pour l'explication, on peut capturer 97,7% de la variance des variables explicatives et avec 2 facteurs extrait, 74% de la variance des variables dépendantes !

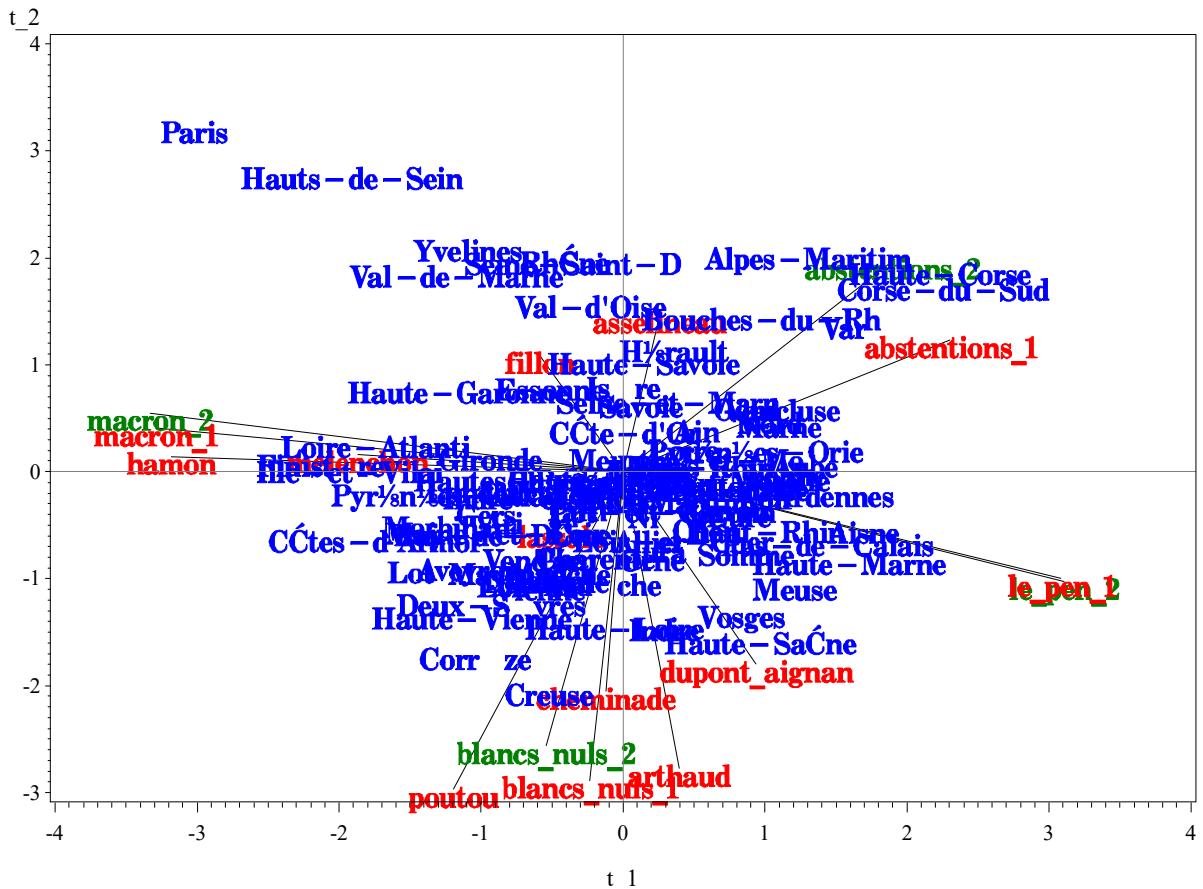
variable	vip
macron_1	1.896
hamon	1.598
abstentions_1	1.329
le_pen_1	0.968
melenchon	0.823
poutou	0.812
blancs_nuls_1	0.577
fillon	0.562
dupont_aignan	0.420
arthaud	0.415
cheminade	0.326
asselineau	0.306
lassale	0.299

La PLS nous indique 3 variables essentiellement projetés : macron, hamon, et abstentions_1

Elles sont au moins chacune significativement corrélés à l'un des 2 axes de la PLS.

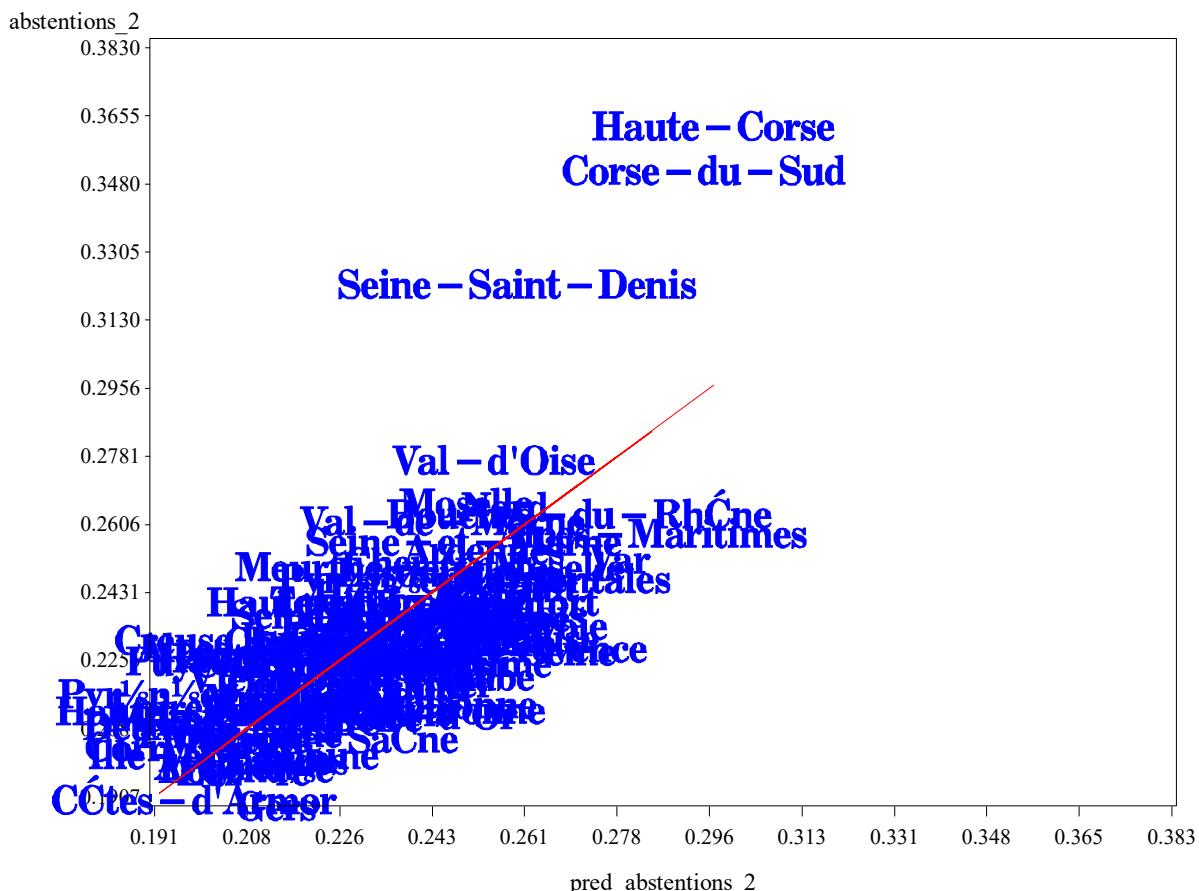
Coefficients de corrélation de Pearson, N = 96 Proba > r sous H0: Rho=0		
	t_1	t_2
abstentions_2	0.54177	0.56202
abstentions_2	<.0001	<.0001
blancs_nuls_2	-0.15533	-0.73241
blancs_nuls_2	0.1308	<.0001
macron_2	-0.95172	0.15737
macron_2	<.0001	0.1257
le_pen_2	0.88775	-0.29245
le_pen_2	<.0001	0.0038
abstentions_1	0.65812	0.35113
abstentions_1	<.0001	0.0005
le_pen_1	0.88158	-0.28461
le_pen_1	<.0001	0.0049
macron_1	-0.94083	0.11591
macron_1	<.0001	0.2608
merlenchon	-0.53406	0.04660
merlenchon	<.0001	0.6521
fillon	-0.16725	0.31040
fillon	0.1034	0.0021
hamon	-0.90936	0.04075
hamon	<.0001	0.6935
dupont_aignan	0.26846	-0.51310
dupont_aignan	0.0082	<.0001
poutou	-0.34139	-0.84901
poutou	0.0007	<.0001
asselineau	0.07347	0.41856
asselineau	0.4769	<.0001
lassale	-0.12563	-0.15343
lassale	0.2226	0.1356
arthaud	0.11280	-0.79265
arthaud	0.2739	<.0001
cheminade	-0.03471	-0.58623
cheminade	0.7371	<.0001
blancs_nuls_1	-0.06710	-0.82505
blancs_nuls_1	0.5160	<.0001

Carte simultanée des variables et des individus sur le premier plan : Biplots



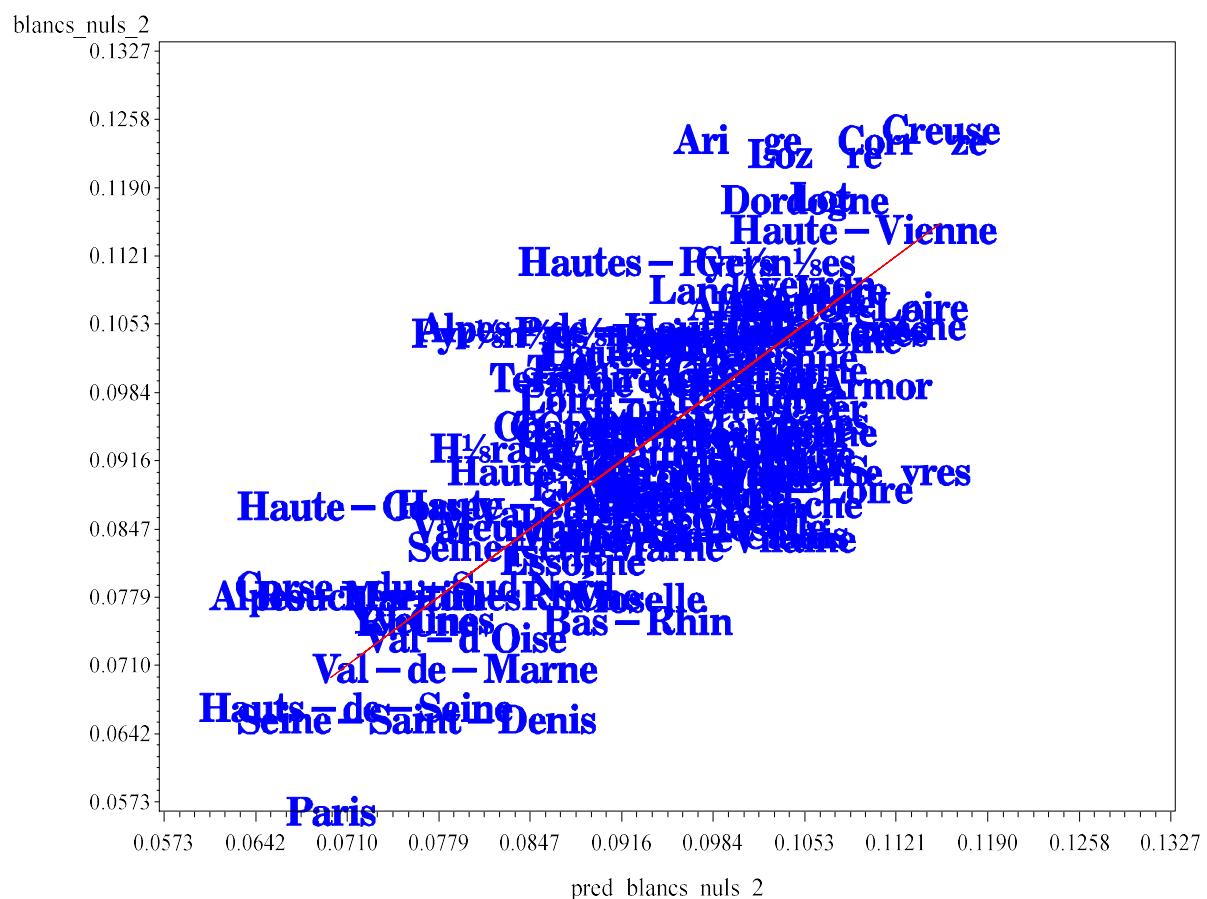
Sur ce biplot, on identifie les 2 mêmes clusters et d'autres variables isolés. On voit bien qu'aucune des variables n'est indépendante des 2 axes PLS ce qui est encourageant !

Valeurs observées vs valeurs prédictes de abstentions_2 avec 2 composantes PLS



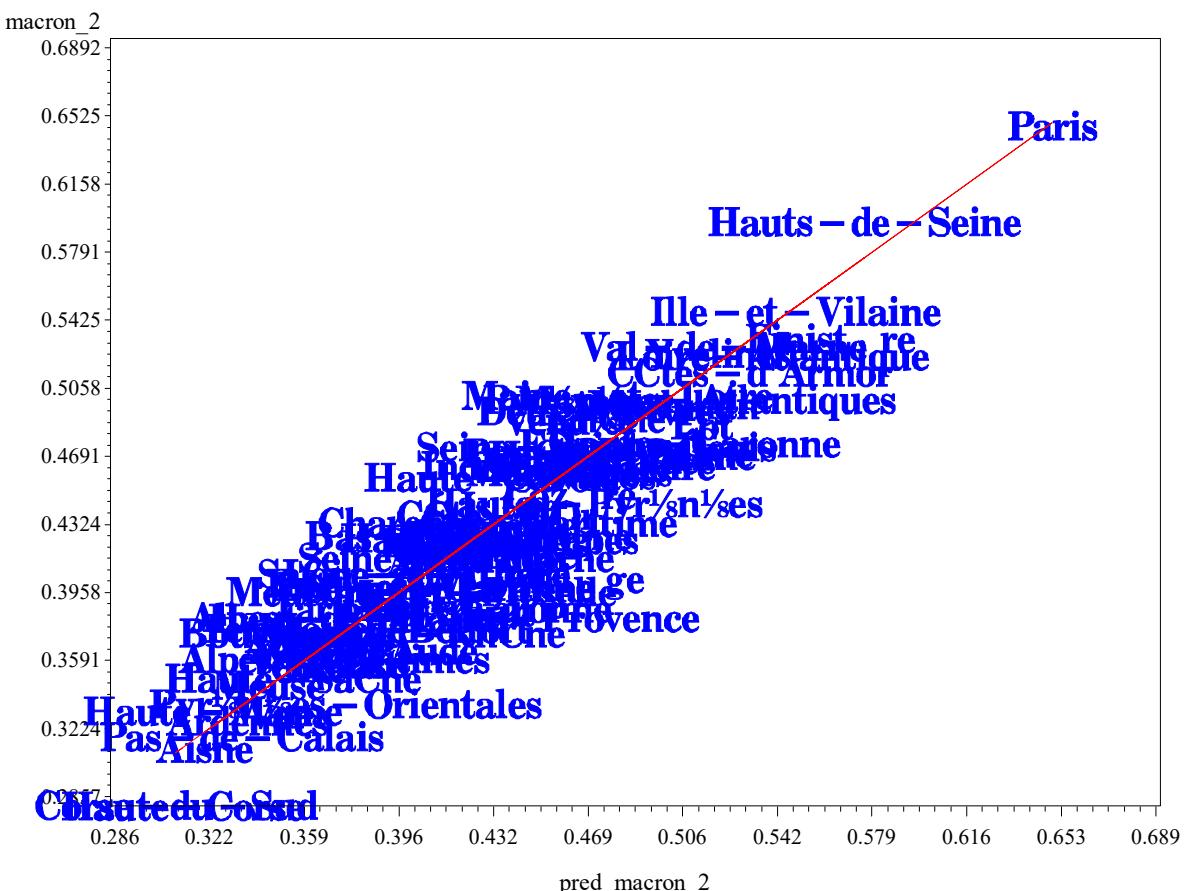
On remarque encore le rôle terrible des 3 département mais en dehors de cela on remarque une assez bonne performance de l'axe de prédiction, on constate une meilleure prédiction qu'avec l'approche PCR.

Valeurs observées vs valeurs prédictes de blancs_nuls_2 avec 2 composantes PLS

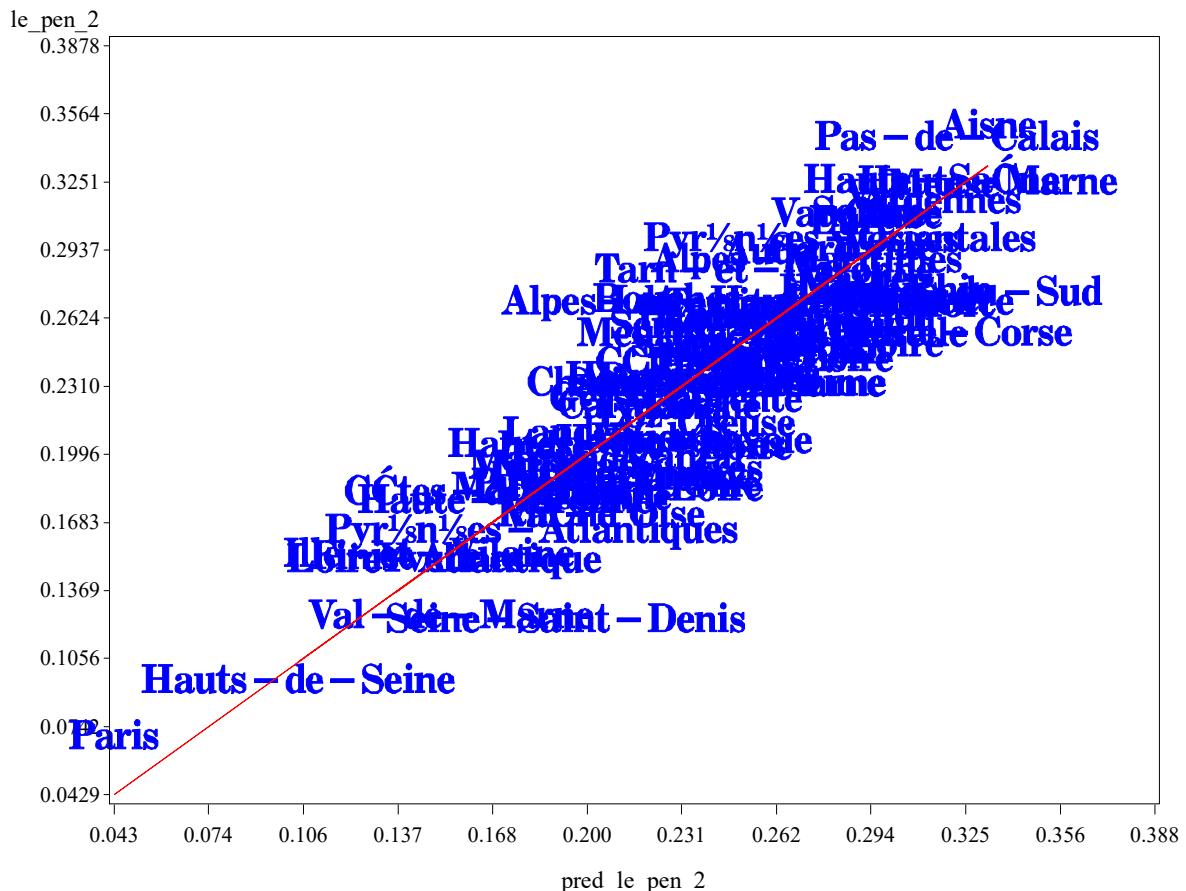


On remarque une erreur résiduelle marquée pour l'abstention et les votes blancs-nul mais toujours bien meilleur qu'avec l'approche PCR.

Valeurs observées vs valeurs prédites de macron_2 avec 2 composantes PLS



Valeurs observées vs valeurs prédites de le_pen_2 avec 2 composantes PLS



On constate également une bien meilleure performance de prédiction pour les variables macron 2 et le pen 2

Conclusion :

- Valeurs atypique : Dom-Tom, Seine-saint-Denis, Corse-du-Sud, Haute-corse
 - Difficultés à prédire l'abstention et blanc_nuls comparé au score des candidats
 - On remarque 2 grand groupes de clusteurs : (macron/hamon/melenchon et pouton/cheminade/artaud/blancs_nul) les autres variables étant relativement peu corrélées entre elles.
 - L'approche PLS s'est avérée supérieure à l'approche PCR dans le cas d'une prédiction et de prédictions multiples