IAE NANTES
ÉCONOMIE & MANAGEMENT

UNIVERSITÉ DE NANTES

Econométrie des variables qualitatives

Les déterminants du vote Front National par commune en 2017

Election présidentielle au premier tour

Darlène NIBOGORA

Ariinui TERIITEHAU

Master 1, Économétrie et Statistiques appliquées (EKAP) 2019-2020

Abstract :

This paper consists in building a model to explain the outcome of the first round of the French presidential elections by municipality with the rise of « Front National » and the emergence of the new political party by E. Macron, « En Marche ! ». The study will highlight the role of socio-economic and political factors, but also the geographic location of the municipalities for explain the « frontist » trend. We will use two different models to analyse the choice of vote : a binary logit model and an ordered mutinomial model.

Résumé :

Le but de ce dossier est de construire un modèle permettant d'expliquer les résultats du premier tour de l'élection présidentielle française en 2017 par commune avec la montée du Front National (FN) de Marine Le Pen et l'émergence du nouveau partie politique d'Emmanuel Macron, « En Marche ! » (EM). L'étude mettra en évidence le rôle des facteurs socio-économique et politique, mais aussi l'emplacement géographique des commune afin d'expliquer la tendance frontiste. On utilisera le deux modélisations différentes pour analyser ce choix de vote : une modèle binaire logit et un modèle multinomial ordonnées.

Mots-clés : Economie du vote, Front National, élection présidentielle française, modèle binaire logit, multinomial ordonnée.

Key-words : Economic voting, Front National, French presidential elections, a binary logit model, an ordered multinomial model.

Sommaire

I. Introduction

Le Rassemblement national (RN), dénommé Front national (FN) jusqu'en 2018, est un parti politique français fondé en 1972 à l'initiative d'ordre nouveau[1] sous la dénomination officielle «Front national pour l'unité française ». Il est présidé par Jean-Marie Le Pen de sa création à 2011, puis par sa fille Marine Le Pen jusqu'aujourd'hui.

Le Front National rentre véritablement dans la vie politique française dans les années 1980 où il a obtenu notamment, à l'issue des élections législatives de 1986, un groupe parlementaire constitué de 35 députés. Cependant, il avait participé aux élections législatives de 1973, où il avait obtenu «1,32 % d'électeurs à l'échelle nationale»[2]. Candidat à l'élection présidentielle à cinq reprises, Jean-Marie Le Pen parvient à se qualifier au second tour du scrutin présidentiel de 2002 face à Jacques Chirac. Par la suite, le FN obtient d'importants succès électoraux, terminant notamment en première position aux élections européennes de 2014 et au premier tour des régionales de 2015. Marine Le Pen se qualifie au second tour de l'élection présidentielle de 2017, à l'issue duquel elle obtient 33,90 %[3] des voix face à Emmanuel Macron. Il est à rappeler qu'il avait obtenu une deuxième place au premier tour des mêmes élections avec «21.4% des suffrages exprimés»[4].

Son objet :"le Rassemblement National est une formation politique qui concourt à l'expression du suffrage dans le cadre des institutions de la République française et du pluralisme démocratique, conformément à l'article 4 de la Constitution du 4 octobre 1958.

Attaché à l'égalité devant la loi de tous les citoyens français sans distinction d'origine, de race ou de religion, le Rassemblement National défend la souveraineté, l'indépendance et l'identité de la nation.
Il protège le caractère indivisible, laïc, démocratique et social de la République, ainsi que l'intégrité du territoire national, en métropole comme en outre-mer"[5].
Actuellement, le RN (depuis le 01/06/2018) fait parti des partis politiques français les plus puissants, à cause de l'effet médiatique et l'accroissement du taux d'abstention.

« Le RN ne s'est jamais aussi bien porté électoralement que depuis 2012. Il améliore ses scores à chaque scrutin. La forte abstention donne cependant une image tronquée de ses scores : si on les ramène non plus aux suffrages exprimés mais au nombre d'inscrits, on constate qu'au plus haut de sa forme le parti d'extrême droite a séduit "seulement" quinze électeurs sur 100 »[6].

Plusieurs analyses sur le vote de FN ont démontré que ce parti a plus d'adhérents dans les milieux ruraux , dans les régions à problèmes économiques, dans les régions où il y a peu d'immigrés et qu'il séduit plus les classes moyennes.

- Pourquoi le Front National séduit-il de plus en plus d'électeurs ?

Pour répondre à cette question, il convient d'expliquer par le vote FN durant le premier tour de l'élection présidentielle 2017 par différents facteurs sociaux-économique et politiques, par des modélisations de choix binaire logit mais aussi par un modèle multinomial ordonné en expliquant le vote d'autres partis politique (EM, UMP et autres). Lors de la modélisation, on observera le comportement des communes sur le choix des votes et on conclura les différents déterminants pouvant expliquer le choix du vote FN.

Revue littéraire

Nombreuses études ont expliqué par des modèles politico-économétriques électorales et prévisions électorales lors des élections de 2017 en France (J-D. Lafay, F. Facchini et A. Auberger). La variable à expliquer est représenté par une fonction de popularité et de vote pour un candidat et est aussi utilisée comme un instrument de prévision sur une masse d'observations (individus) afin d'améliorer la qualité de la prévision. Ray C. Fair (1971) expliquait le choix binaire d'un président américain sortant républicain ou démocrate par des variables économiques, taux de croissance du PIB, le taux d'inflation et l'indice des prix de PIB entre la période 1916 et 1992 sur des prédictions présidentielles. Dans ce dossier, l'objectif de notre étude est de définir les tendances des communes sur le vote FN.

Présentation de la base de données

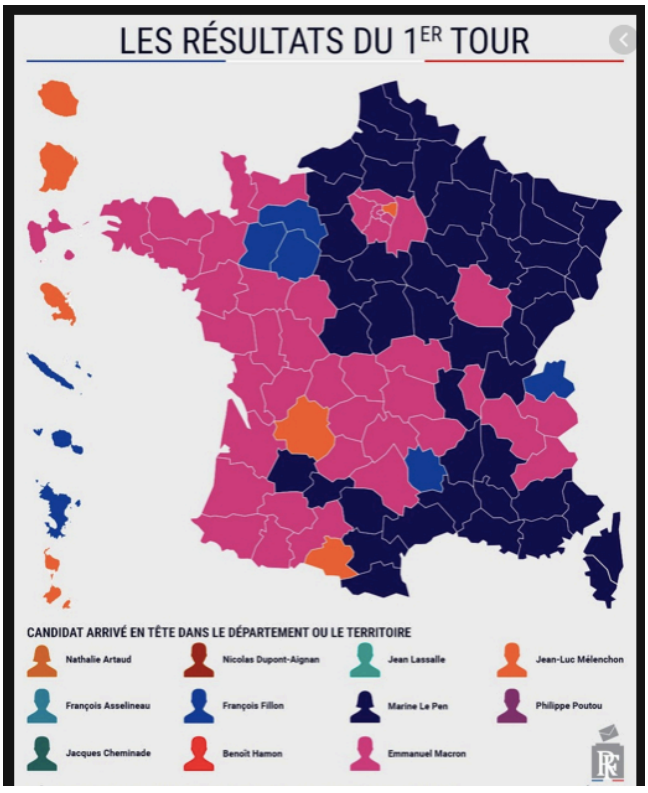
La base de données provient d'un open data de l'INSEE, d'une enquête de recensement de la population française en 2016 (« Enquête emploi en continu, Activité, emploi et chômage en 2016 »). La base comportait plus de 20 millions d'observations d'individus en France. Pour le choix du vote par commune, on a sélectionné tous les français et on les a classé par communes codées par CANTVILLE. On a obtenu une base avec 1771 observations.

Le vote par commune du premier tour des élections présidentielles en 2017 provient aussi d'un open data mis à disposition pour le public par le gouvernement (data.gouv.fr). On se focalise sur les résultats des votes essentiellement sur trois partis : Front National, La République en Marche et Union pour un Mouvement Populaire qui sont également les trois premiers partis ayant gagnés le premier tour. Dans notre base,les communes de la région parisienne et d'Outre-Mer ne sont pas prises en considération car leurs données ne sont pas disponibles.

II - Analyse économique

A - Le vote FN

Figure 1: Les résultats du 1er Tour



Source : elections.interieur.gouv.fr

On cherche un modèle pouvant expliquer le vote du FN et celui de deux autres parties (UMP et EM) par différents facteurs sociaux-économiques en France. Il est logique de choisir comme variable à expliquer le résultat de vote sortant dans la commune. Le FN et l'EM ont remporté les élections présidentielles du premier tour en 2017 avec des votes exprimées respectivement de 21.3% et 24.01%[7]. Les partis traditionnels droite/gauche (UMP et PS) n'ont pas les marches au second tour, alors que le contexte était plus favorable pour le FN avec l'annonce du « BREXIT» par Royaume-Uni, l'élection de D.Trump aux Etats-Unis ou l'insécurité dans le pays par les attaques terroristes subies (attentat à Nice 2016, Charlie Hebdo 2015). C'était un scrutin différent de celui de 2002 qui opposait J. Chirac et J-M Le Pen avec un écart de 62.3% de voix exprimées pour le second tour, c'est une époque où la mobilité anti-FN était très forte. En 2017, l'écart de vote se réduit à 42.1% entre le parti FN et EM pour le second tour. La France fait face à une division de la population sur le fait de voter FN selon la catégorie socio-professionnelle, niveau d'études ou revenus des personnes. Par conséquent, pour le second tour, on retrouve le même scénario pour un vote anti-Le Pen.

B - Le contrôle des frontières (frontiere)

Dans la figure 1, la carte montre un clivage dans la France pour les deux principaux partis FN et EM. Les départements de l'ouest et l'Ile-de-France voient des votes sortants pour l'EM, tandis que le reste de la France est en faveur pour le FN.
Le Sud et la majorité des pays limitrophes de la France ont une tendance pour le FN, ce qui peut être expliqué par la lassitude de la population locale envers les immigrants arrivants. C. Wihtol de Wenden[8] montre en 2002 une émergence du contrôle des frontières en proposant un programme de régulation de flux et la lutte contre l'immigration clandestine.

En effet, le parti FN est connu comme un « grand parti raciste », Marine Le Pen avait toujours proposé des programmes électoraux qui favorisent l'arrêt d'immigration légale et clandestine.

On décide alors de créer une variable binaire qualitative « frontiere » expliquant si la commune se trouve près d'une frontière dans le département ou pas. Dans notre base, on a 492 communes se trouvant près des frontières, ce qui représente 28% des communes[9].

C - Vote sortant des grandes villes (depuiss)

De plus, on remarque aussi que le FN ne possède pas la majorité des votes dans les départements dans lesquels se trouvent des grandes villes comme Lyon, Paris ou Toulouse. Le FN remporte les élections dans les départements qui comportent de nombreuses communes comme dans le département de Pas-de-Calais (38 communes) ou Nord (37 communes) .

Le FN est le premier parti politique qui a remporté les éléctions dans plusieurs communes (817 communes) au premier tour des élections présidentielles en 2017 , puis l' EM avec 630 communes (essentiellement dans les grandes villes avec beaucoup d'habitants) et l'UMP avec 164 communes selon les données de notre base (tableau 1, annexe). Marine Le Pen a pu récolter de nombreuses voix dans les petites et moyennes communes françaises car en 2014 le FN avait gagné une dizaine de municipalités avec 1 546 conseillers municipaux élus (P. Perrineau, « Le vote disruptif »)[10] pour des communes ayant moins de 1000 habitants.

On décide alors de créer une nouvelle variable **depuiss** qui indique si le département possède ou pas une grande ville selon le palmarès des agglomérations de L. Chalard (2006)[11]. L'intérêt de cette variable est de montrer comment le dynamise économique d'une grande agglomération du département peut influencer le choix des résultats de l'élection présidentielle.

En effet, la présence d'une grande ville dans un département encourage une meilleure qualité de vie (emploi, accès à la santé, activité culturelle,…) pour les habitants alors que les communes rurales ou des communes se trouvant loin des grandes villes ont souvent le sentiment d'être délaissé.Il y a de plus en plus du monde dans les grandes villes au détriment des campagnes qui se vident de leurs habitants.

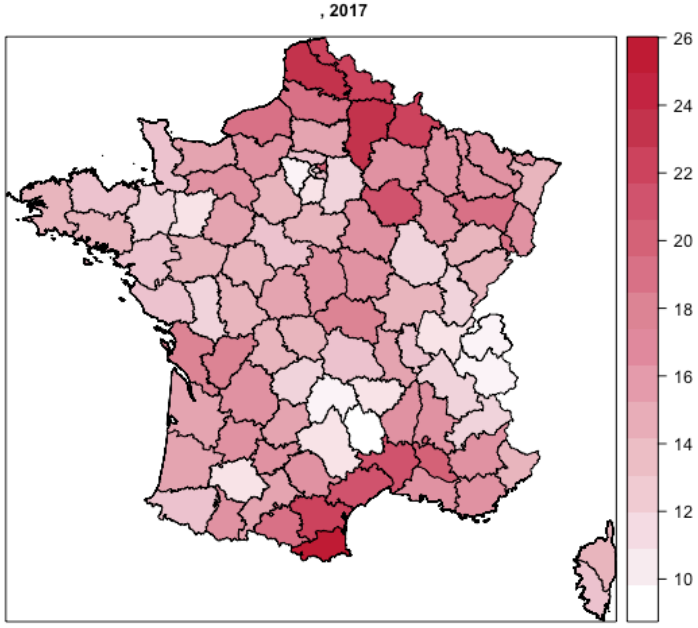
Ainsi, les communes délaissées tendent à se tourner vers le parti FN.«Ce phénomène est clairement lié à la fermeture des services et commerces de proximité.»[12]

Dans notre base, il y a 377 communes qui sont proches de grandes villes dynamiques contre 1394 communes. Ici, on signale qu'on ne prend pas en considération des départements voisins de la grande ville.

D - Le taux de chômage, un facteur important pour le FN (txch)

L'autre variable (**txch**) qui a rendu le FN comme l'un des principaux partis politiques, est le taux de chômage. J-F. Léger[13] (2015) propose une relation entre le chômage et la croissance du vote frontiste après le quinquennat de François Hollande. La figure 2 permet d'avoir une compréhension de la répartition du taux de chômage par département en France. On constate une relation du taux de chômage avec le vote du FN dans les départements, le Nord et le Sud montrent un taux du chômage élevé dans les départements où le vote était majoritairement en faveur de ce parti. Dans ces zones, le FN a su gagner des électeurs entre le 1er trimestre 2012 et le 4ème trimestre 2014 avec une hausse relative du chômage de 27% dans la Mayenne[14] . On s'attend de voir que lorsque le taux de chômage est élevé alors la probabilité de voter FN augmente aussi.

Figure 2: Taux de chômage en France



Source : INSEE 2017, logiciel R

E - Le taux d'immigration,

Une variable opposant la précédente, c'est bien évidemment le taux d'immigration de la population. En 2011, la proportion d'origine étrangère sur trois générations était d'environ de 30% (M. Tribalat, « *Une estimation des populations d'origine étrangère en France* »)[15] où une majorité de ces populations était d'origine africaine et européenne. Dans la ***figure 2*** (annexe), la carte indique la répartition de la population provenant d'immigration en France, on constate une concentration de cette population dans les

grandes villes, aux frontières et dans le Sud de la France. On suppose une relation négative entre cette variable et le fait de voter FN car la majeure partie de population ayant des origines étrangères ne vote pas le FN à cause des propos xénophobes cités par le chef de ce parti.

F - Google, facilité d'accès à l'information (GG)

L'évolution technologique ont amené les partis politiques à exposer aussi leurs programmes électoraux par le biais des médias. Le moteur de recherche Google est devenu aussi un outil indispensable pour la circulation d'information entre les partis politiques et les électeurs. A l'aide de l'application de Google Trends permettant d'identifier les tendances de l'élection présidentielle française selon les mots clés de recherche effectués, Google Trends nous donne une idée des « pronostiques » des résultats d'une élection et permet aussi d'étudier les comportements des internautes sur l'accès à l'information.

On décide d'inclure une nouvelle variable (GGFN) et voir comment elle peut influencer sur le choix des électeurs dans les communes. Cette variable représente les départements dont le pourcentage de recherche est plus élevé pour un parti. Ainsi, elle prend quatre modalités : FN, UMP, EM et autres. On trouve un nombre élevé pour le FN (tableau), on avait dit que le FN a pu conquérir de nombreuses communes de moins de 1000 habitants. On retrouve pratiquement le même pourcentage de répartition entre les communes et les départements. ***Figure 3*** (annexe) indique quel département est majoritaire dans le moteur de recherche du parti FN. On remarque que sur la côte littorale, le Nord et le Centre de la France, la recherche sur le parti FN est plus importante dans ces zones.

Tableau 2 : Google Trend

Partie	FN	EM	UMP	Autres
Nombre de commune	916 (51%)	421(23%)	233 (13%)	204 (11%)
Nombre de département	46 (48%)	24 (25%)	13 (13%)	13 (13%)

Source : Google Trend, recherche vote

G - Variable socio-démographiques

En effet, les variables socio-démographiques influencent beaucoup dans le choix électoral et fournissent des informations importantes sur « les configurations territoriales ».Les variables socio-démographiques choisies dans une commune sont:

- le nombre d'ouvriers (ouvrier),
- le nombre de cadres (cadres),
- le nombre de logements HLM (HLML),
- le nombre d'artisans et de chefs d'entreprises (artis_chef_dentrepri),
- le nombre de personnes qui ont un contrat de travail temporaire(intérim),
- le nombre de retraités (retraite),
- le nombre de personnes possédant un diplôme supérieur au bac et le bac (nivdip),

la population active (actifs),

- le nombre de personnes travaillant dans l'agriculture (agri)
- le nombre de personnes travaillant dans l'industrie (indus)
- le nombre de personnes travaillant dans le BTP (construction)

Plusieurs études ont démontré qu'il existe une relation positive entre le taux de vote pour le FN et le nombre d'ouvriers, par contre le taux de vote pour le FN a une relation négative avec le taux de diplômés du supérieur, de cadres, d'employés, de logement HLM. Cela signifie qu'une forte présence d'ouvriers dans une commune engendre un fort vote de FN et qu'une forte présence de diplômés du supérieur (ou de cadres ou d'employés ou d'étrangers ou de logements HLM) engendre un faible vote de FN .

L'importance du niveau d'éducation

«Le diplôme reste une barrière décisive. Le FN attire certes des intellectuels et des énarques en accord avec ses idées, mais les probabilités de voter pour lui restent d'autant plus élevées que la personne a fait peu d'études. Elles culminent à 45% chez celles qui n'ont qu'une formation technique courte (CAP, brevet professionnel). Un score trois fois plus élevé que celui qu'il atteint chez les personnes ayant suivi au moins un second cycle universitaire.»

En effet, les diplômés du supérieur occupent des hautes fonctions comme cadres, chefs d'entreprises ou métiers libérales.Cette catégorie de population ne vote pas le FN à cause notamment de son programme de sortie de l'UE.

Relation du vote FN et les facteurs socio-démographiques

La majorité considère que c'est une bonne chose d'appartenir à l'Union européenne et qu'il serait inimaginable d'en sortir.Le FN envisage la sortie de l'Union européenne et cela leur est désavantageux car: d'une part, leur zone géographique d'employabilité se rétrécirait et d'autre part, leurs capitaux et leurs entreprises se retrouveraient en grandes difficultés. C'est donc cette incertitude qui leur pousse à ne pas voter pour le FN.

Ils voient aussi de mauvais œil l'idée prônée par le FN de fermer les portes aux migrants puisqu'ils considèrent que ces derniers constituent une main d'oeuvre, de plus en plus rare dans l'Europe dont la population est vieillissante.De plus, la majorité de ces migrants occupent les emplois (non qualifiés ou précaires) que les européens négligent.

Le FN attire beaucoup des personnes issues de classes moyennes qui sont généralement concentrées dans des zones rurales à cause de son programme de sortie de l'Union européenne et la réglementation de l'immigration.
En effet, les habitants des zones rurales ont en général fait peu d'études et occupent des postes moins qualifiés comme agriculteurs, éleveurs, artisans ,petits commerçants, ouvriers, etc.

Cependant, ils voient leurs activités menacées par les multinationales.Ainsi, ils pensent que si le FN gagne, leur situation s'améliorerait grâce notamment à la sortie de l'UE que promet le FN.Ils ne seront donc plus envahis par les produits européens favorisés par le libre change.

De plus, les milieux ruraux sont pour la plupart habités par les français de souche qui ont tendance à considérer la masse des immigrants comme une invasion.En effet, les premiers considèrent que les seconds viennent leur prendre le travail et de profiter du système français(aides, allocations sociales, système de santé, etc;).

Par contre, les retraités ne votent pas le FN non pas pour son programme mais pour sa présidente puisque la plupart considère qu'une femme ne peut pas être leur président, ils considèrent qu'un homme symbolise l'autorité, ce que ne serait, selon eux, représenter une femme (et cette considération est ressentie aussi chez les hommes et les femmes de 60 ans et plus). Pour cela, on suppose une relation négative entre cette variable et le fait de voter FN.

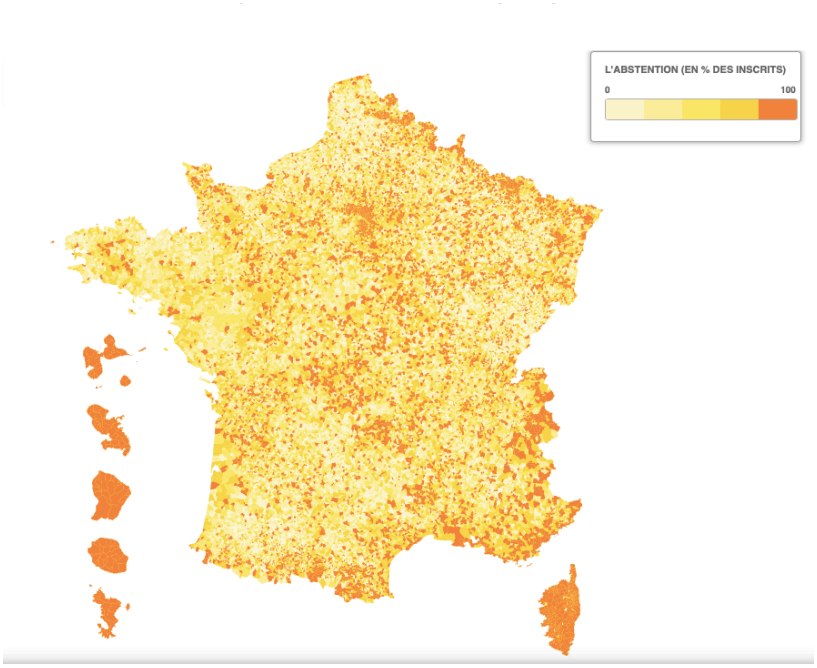
Le FN n'a pas beaucoup de succès auprès des personnes vivant dans des HLM car les HLM sont habités pour la plupart par des français d'origine étrangère et considèrent que le programme du FN leur est désavantageux. Pour cela, on suppose une relation négative entre cette variable et le fait de voter FN.

Le FN séduit moins d'électeurs au sein de la population active par rapport à ses concurrents.Pour cela, on suppose une relation négative entre cette variable et le fait de voter FN.

H - La variable politique (X..Abs.Ins et X..Vot.Ins)

La variable politique qu'on a choisi dans notre étude est le taux d'abstention (X..Abs.Ins) ,il s'agit des gens déçus par les systèmes antérieurs et qui n'attendent rien de leurs gouvernements.La majeure partie de cette population habitent dans des régions délaissées. Cette variable influence positivement le vote du FN. On constate sur ce graphique ***(figure 4)*** que le FN est plus voté dans les régions où le taux de participation est faible.Pour cela, on suppose une relation positive entre cette variable et le fait de voter FN et une relation négative dans le fait de voter (X..Vot.Ins, taux de vote exprimés).

Figure 4 : Carte abstention 1er tour en 2017



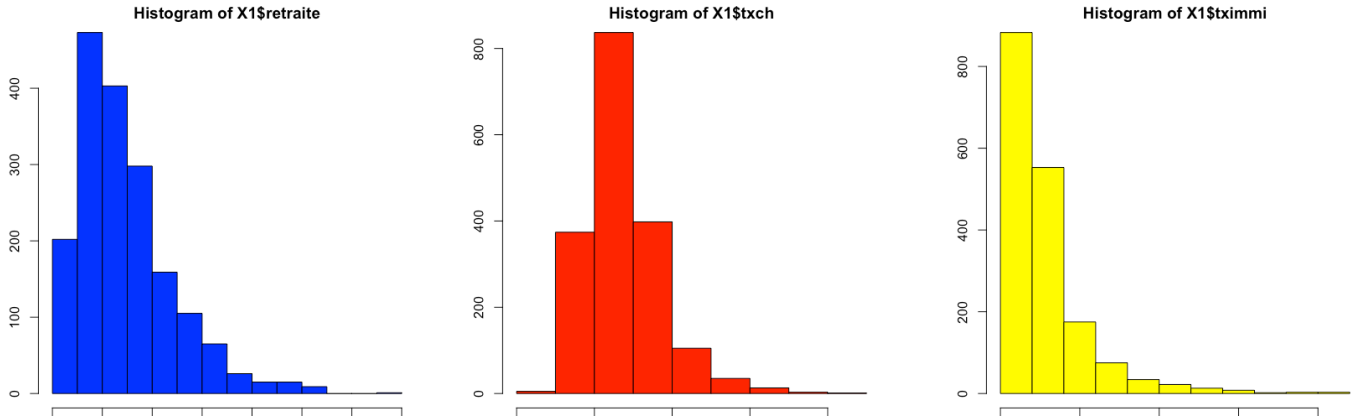
Source : gouvernement, France 2017

III. Analyse descriptive

Dans cette parti, on procède à une analyse des variables explicatives X dans la base de doonnées.

Dans **la figure 5**, on constate que la distribution pour les trois variables (retraite, txch et tximmi) sont différentes. Pour retraite (barre en bleu), on a une distribution qui ne suit pas une loi normale et qu'elle est plus étalée à droite. Txch (barre en rouge) ne suit pas une la normal et est plus étalée à droite, de même pour tximmi (barre en jaune).

Figure 5 : Histogramme retraite, txch et tximmi



Source : INSEE 2017, logiciel R

Sur les deux graphiques ci-dessus, on remarque que :

-la valeur minimale, le premier quartile, la médiane ,le troisième quartile et la valeur maximale sont confondus pour les variables agri et intérim.

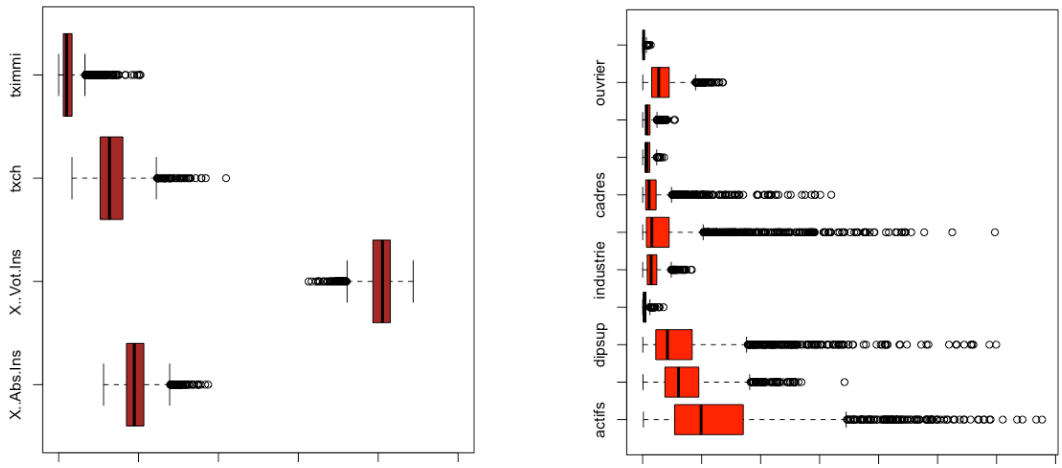
-un léger écart entre la valeur minimale, le premier quartile, la médiane ,le troisième quartile et la valeur maximale pour les variables construct, artisans-chef et HLML

-un moyen écart entre la valeur minimale,le premier quartile, la médiane ,le troisième quartile et la valeur maximale pour les variables ouvriers, cadres, industri, HLML, le taux de chômage, le taux d'abstention, le taux de vote et le taux d'immigration.

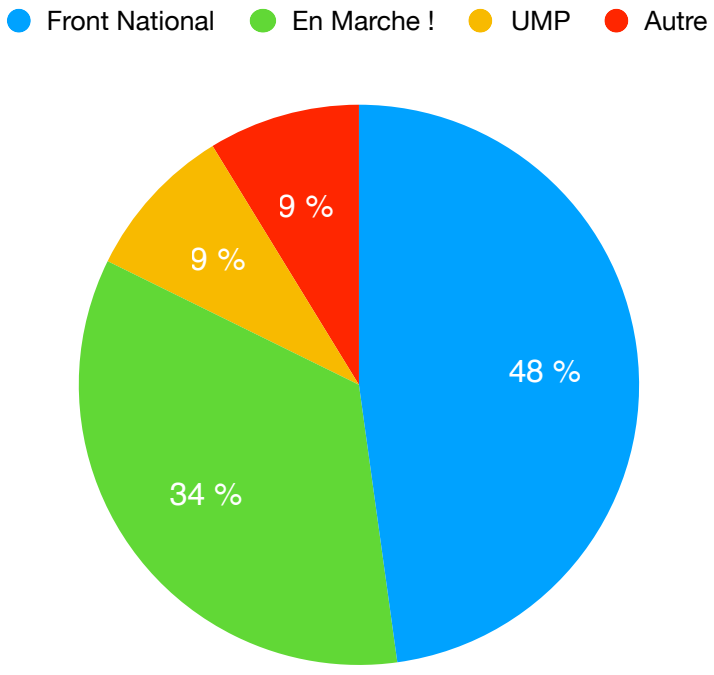
-un grand écart entre la valeur minimale, le premier quartile, la médiane ,le troisième quartile et la valeur maximale pour les variables actifs, retraite, dipsup.

Nous remarquons aussi que dans la distribution des données de nos variables, nous avons une asymétrie à droite pour toutes les variables sauf pour la variable X.Vote.Ins. De plus, on remarque que toutes les variables quantitatives ont plusieurs potentielles valeurs atypiques. Normalement, on devrait les supprimer de notre base de données pour éviter d'avoir les résultats biaisés mais nous décidons de les garder dans notre étude pour éviter de perdre quelques observations qui sont les communes.

Figure 6 : Boite à moustache des variables X



Sur ce graphique ,on remarque que FN a 817 (48%) communes, l'EM a 630 communes, l'UMP a 164 communes.



IV. Analyse économétrique

1. Modèle de choix binaire logit

1.A Méthode Stepwise et cluster par CAH

Après une analyse des variables, on réalise une régression pas à pas « stepwise » afin de réduire les variables X qui pourraient expliquer au mieux la variable binaire Y pour le vote FN. On utilisera une régression de type logit dont les termes d'erreurs suivent une loi logistique.

Régression :

FN~X..Abs.Ins+actifs+retraite+tximmi+X..Vot.Ins+txch+dipsup+agri+industrie+
HLML+cadres+construction+artis_chef_dentrepri+
ouvrier+interim+frontiere+depuiss+
GGFN+GGEM+GGautre

La méthode de Stepwise va être réalisée de 3 façons différentes : « forward » ,
« backward » et « both ». (annexe, ***figure 7)***. Le modèle retenu sont :
txch + tximmi + frontiere + cadres + industrie + dipsup + actifs + HLML + X..Vot.Ins + GGUMP + depuiss + interim + construction + agri + GGFN + construction + X..Abs.Ins

Cependant, on a constaté avec la matrice de corrélation, une dépendance importante entre certaines variables présentant les catégories socio-professionnelles ou le niveau de diplôme dans le modèle retenu précédemment. Ainsi, on décide de résumé l'ensemble des variables quantitative en excluant X..Vot.Ins ,txch et tximmi car leurs corrélations étaient faible avec les autres variables. On utilise la méthode de *classification ascendante hiérarchique* (**CAH**), une méthode de cluster qui catégorise les différentes observations (annexe).

CAH recommande deux catégories. Ainsi, après une étude des résultats, on propose comme thématique de la cat2 : grande commune et pour la cat1 : petite et moyenne commune.

La carte (annexe, ***CAH***) indique la répartition de la cat2 en France et on voit que les communes concernées se trouvent essentiellement dans la région parisienne, ou sont proches de grandes villes (comme Lyon, Toulouse ou Marseille) ou sont au Nord de la France. On s'attend d'avoir une relation positive entre cat1 et le vote pour le FN, puis négative avec la cat2.

Pour vérifier la pertinence de la nouvelle variable cat, on procède de nouveau en « stepwise », avec les variables suivantes : txch + tximmi + frontiere + X..Vot.Ins + GGUMP + depuiss + GGFN+ cat1.

La régression propose de retenir le modèle suivant :
FN~txch + tximmi + frontiere + X..Vot.Ins + cat1 + depuiss (annexe, ***figure 8)***

Vérification d'indépendance entre les variable

Ensuite, on décide de vérifie l'indépendance entre les variables explicative par deux test, celui du Chi Deux et T-Stat. Le test Chi-deux vérifie l'indépendance entre deux variables qualitative, si la p-value est supérieur à 10% alors il y a une indépendance entre elles, de même pour le test T-Stat qui vérifie l'indépendance entre une variable qualitative et une variable quantitative.

Le tableau 3 résume l'indépendance entre les variables, les cellules vertes représentent l'indépendance des variables. On remarque que la variable GGFN est dépendante avec toutes les variables et que la variable frontiere est aussi dépendante sauf pour la variable cat1. Alors il faudrait faire attention lors des estimations des modèles et de l'interprétation des résultats des modèles estimés.

Tableau 3 : test d'indépendance entre les variables

	txch	tximmi	Frontier	Vos.Ins	GGFN	cat1	depuiss
txch							
tximmi							
Frontie							
Vos.Ins							
GGFN							
cat1							
Depuis							

A partir de ce tableau, on peut proposer les modèles avec des variables indépendantes entre elles :

- Modele 1 : FN~txch + tximmi + frontiere + X..Vot.Ins + cat1 + depuiss
- Modele 2 : FN~txch + tximmi + X..Vot.Ins + Depuiss
- Modele 3 : FN~cat1 + frontiere

1.B Régression par logit et vérification des tests

On réalise un modèle logit avec l'ensemble des variables retenues et on doit penser à vérifier la multicolinéarité, l'hypothèse de nullité de l'ensemble des coefficients, l'existence d'observations potentielles influençant de manière significative l'estimation et l'hypothèse d'homoscédasticité. (annexe, ***figure 9***)

Figure 10 : modèle avec fonction glm

```
Call:
glm(formula = (FN ~ txch + tximmi + frontiere + X..Vot.Ins +
    GGFN + cat1 + depuiss), family = binomial(logit), data = X1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8822  -0.9877  -0.3167   1.0566   2.0067

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.71004    1.51861   1.126  0.26014
txch         0.17986    0.01541  11.673  < 2e-16 ***
tximmi      -0.32850    0.03630  -9.049  < 2e-16 ***
frontiere1   1.05321    0.12542   8.397  < 2e-16 ***
X..Vot.Ins  -0.05218    0.01750  -2.981  0.00287 **
GGFN1       -0.29325    0.11260  -2.604  0.00920 **
cat11        0.60070    0.29227   2.055  0.03985 *
depuiss1     0.27807    0.13781   2.018  0.04361 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2444.5  on 1770  degrees of freedom
Residual deviance: 2051.7  on 1763  degrees of freedom
AIC: 2067.7

Number of Fisher Scoring iterations: 5
```

Source : régression avec Logiciel R

D'après la régression du modèle initial 1 (***figure 10***), on remarque que toutes les variables ont un impact significative sur le fait d'une commune dans la majorité de la population votait FN. Ici, on ne tient en compte que le signe du coefficient des variable. Par exemple, si txch augmente alors la probabilité de voter FN augmente aussi. De même, pour les communes ayant une grande ville dans le département (depuiss), la probabilité

augmente aussi. Par contre, lorsque tximmi ou Vot.Ins augmente alors la probabilité de voter FN diminue aussi. Donc, les variables frontiere et cat1 ont une relation positive sur le vote FN alors que GGFN a une relation négative.

Il est important de vérifier la multicolinéarité entre les variables explicatives avec la fonction vif (annexe, ***figure 9***), les valeurs ne sont pas très élevées alors il n'y a pas de colinéarité entre les variables. On n'a aucune instabilité des coefficients.

Dans un autre temps, il faut montrer l'intérêt du modèle initial 1 par le ratio de Vraisemblance indiquant la significativité globale du modèle et on prend en compte le test statistique de Chi deux. Si la valeur du ratio est inférieure à 5% alors on peut retenir le modèle. Dans la ***figure 9*** (annexe), la valeur est inférieure à 5%, on retient donc le modèle initial.

Commentaire des odd-ratios et les effets marginaux

Maintenant, on s'intéresse au calcul des odd-ratios (annexe, ***figure 9***) lequel on calcule l'exponentiel des coefficients de la régression du modèle logit.

On avait dit que toutes les variables étaient significatives ainsi il est possible de les interpréter.

- une commune considéré comme petite ou moyenne (cat1) a 1.82 fois plus de chance que le FN remporte par rapport aux grandes villes.
- une commune proche d'une frontière a 2.86 fois plus de chances que le FN y gagne par rapport à une commune qui est éloignée de la frontière.
- la variable GGFN, si la commune se trouve dans un département où la majorité de recherches google sur le parti FN alors la commune a 1.35 (1/0.74) fois moins de chances qu'on y vote FN par rapport à une commune qui ne se trouve pas dans ce type de département.
- une commune proche de la frontière a 2.86 fois plus de chance que FN soit majoritaire par rapport aux communes qui sont éloignées de la frontière.

Puis, on calcule les différents effets marginaux du modèle initial (annexe, ***figure 9***), on trouve:

- pour la variable Vot.Ins, que l'augmentation de cette variable de 1% dans une commune fera diminuer de 0.01 la probabilité de la majorité du vote FN dans cette commune.

- Pour txch, une augmentation de cette variable de 1% entraîne une augmentation de 0.036 de la probabilité de voter FN.

- Une augmentation de 1% pour tximmi entraîne une baisse de 0.066 la probabilité de voter FN.

Prévision, taux d'erreur et estimation modèle 1bis

Dans cette partie, on se préoccupe des prévisions et du taux d'erreur du modèle estimé. Pour le modèle estimé, on a trouvé un taux d'erreur équivalent à 33% qui est assez élevé. La valeur de la sensibilité vaut 58.7% et la spécificité est égale à 74%.

Un bon modèle devrait avoir des valeurs faibles pour le taux d'erreur et des valeurs élevées pour la sensibilité, précision et la spécificité.

On vérifie ensuite le modèle avec la fonction hitmiss sous R, l'estimation fournit une prédiction de 67%; le pourcentage correct de prédiction dans le cas où la commune vote FN est égal à 58.8% alors que le modèle prédit correctement dans le cas contraire 74% (ne votant pas FN). (annexe, ***figure 10***)

L'existence d'observations potentielles influençant de manière significative l'estimation dans le modèle initiale. La figure (annexe, ***figure 10***) représente les résidus du modèle.

Ainsi tous les points se trouvant dans la partie supérieure de la droite rouge, ce sont toutes les communes votants pour le FN, dans la partie inférieure, c'est le cas contraire.

On remarque qu'il existe des observations influençant significativement l'estimation car on voit des observations qui ne sont pas comprises entre -2 et 2. Les douze communes identifiées sont : Mitry-Mory, Mennecy, Couserans Est, Couserans Ouest, La Haute-Vallée de l'Aude, Rochefort, La Grand-Combe, Le Vigan, Etaples, Thiers, Le Mans-4, Elbeuf.

On décide de ré-estimer un nouveau modèle en enlevant les 12 communes pour voir si il y a une amélioration de la qualité du modèle avec le R2 McFadden.

Pour le modèle 1, on a un R2 McFadden de 16% qui est faible. En ré-estimant le modèle (avec suppression des observations qui influencent la significativité de l'estimation (annexe, ***estimation modèle 1bis***)), on trouve une meilleure R2 qui est égale à 18% mais aussi il y a une amélioration au niveau de la sensibilité du modèle. On constate par contre que l'exponentiel des coefficients et les effets marginaux ne varient pas beaucoup par rapport au modèle initial 1. De plus, on remarque que le nouveau modèle n'a pas de problèmes de multi-colinéarité en utilisant le vif. (Modèle 1 et 1bis résumer dans le tableau 5)

1.C Test d'hétéroscédasticité

Puis, il est important de vérifier l'hypothèse de d'homoscédasticité des erreurs. On effectue une nouvelle estimation du modèle 1 sous le logiciel R avec la fonction hetglm (annexe, ***figure 11***). On teste l'hypothèse d'homoscédasticité en tenant compte de l'hétéroscédasticité des erreurs si c'est nécessaire.

Dans la première estimation(annexe, ***figure 11***), on remarque que les variables txch, tximmi, frontiere, Vot.Ins et cat1 ont un impact significatif au seuil de 5% sur la variance des résidus. Le LR test d'homoscédasticité admet une p-value inférieur à 5% donc on refuse l'hypothèse nulle (homoscédasticité). Il faut aussi vérifier le vif de multicolinéarité entre les variables. On avait des valeurs très élevées qui montrent une forte colinéarité entre les valeurs.

Par conséquent, on réalise une nouvelle estimation du modèle en retirant la variable qui n'était pas significative dans la première estimation (depuiss). On a la variable GGFN qui n'est plus significative et l'hypothèse nulle est aussi refusée au seuil de 5%(car la p-value est inférieure à 5%). On retire de nouveau la variable non significative dans l'estimation suivante (annexe, ***figure 11***), on a toutes les variables qui sont significatives et l'hypothèse nulle qui est refusé mais on a un problème de multicolinéarité avec des valeurs plutôt élevée pour txch et tximmi dans le vif.

On décide alors de faire une dernière estimation en retirant txch (car la valeur du vif est très élevée), ce modèle réunit vérifie toutes les conditions vues précédemment .De plus, les valeurs du vif sont très faibles, cela permet de conclure que le problème de

multicolinéarité est résolu. Les deux variables retenues : txch et frontiere ont un impact significatif et positif au seuil de risque de 5% sur la probabilité de voter FN.

Ensuite, on s'intéresse à la qualité du modèle avec le R2 McFadden (annexe, ***figure 12***), le ratio vaut 17% ce qui est bien meilleur que dans le modèle où les erreurs sont supposées homoscédastiques avec le R2 McFadden de 16%.

On réalise une dernière estimation du modèle à erreurs supposées homoscédastique en utilisant la fonction sous R « lrtest », ainsi c'est un test de log vraisemblance qu'on utilisera. La p-value est inférieur à 5% alors on conclut que le modèle hétéroscédastique est préférable par rapport au modèle homoscédastique. On peut donc retenir le modèle et interpréter les résultats du modèle initial 1.

Tableau 4: résumé des modèles

	Modèle 1			**Modèle 2**			**Modèle 3**			**Modèle 1 bis**		
	Sign	**Exp**	**effma**	**Sign**	**Exp**	**effma**	**Sign**	**Exp**	**effma**	**Sign**	**Exp**	**effma**
Txch	+ (***)	1.19	0.036	+ (***)	1.18	0.036				+ (***)	1.23	0.041
Tximm	- (***)	0.72	-0.066	- (***)	0.74	-0.062				- (***)	0.69	-0.069
frontie	+ (***)	2.86	0.21				+ (***)	2.77	0.23	+ (***)	3.14	0.222
Vot.Ins	- (**)	0.94	-0.01	- (***)	0.94	-0.012				- (**)	0.95	-0.01
GGFN	- (**)	0.74	-0.058							- (**)	0.72	-0.062
cat1	+ (*)	1.82	0.119				+ (***)	2.89	0.24	+ (.)	1.78	0.112
depuis	+ (*)	1.32	0.05	+ (*)	1.37	0.06				+ (*)	1.36	0.059
Err	33 %			33 %			39 %			33.32 %		
Sens	58 %			56 %			34 %			60.24 %		
Spec	74 %			74 %			82 %			74.04 %		
R_Mc	16 %			12 %			5 %			18 %		
Hsct	tximmi,Vot.ins, GGFN, cat1, depuiss			Txch, Vot.ins, depuiss			Frontiere			tximmi, txch ,Vot.ins, GGFN, cat1, depuiss		
R_Mc H	17 %			12.89%			7 %			18.5%		
H	refus			refus			refus			refus		

Sign : significative des variables, **Exp :** exponentielle des coefficients**, effma :** effet marginal
R_Mc : R2 Mc Fadden**, Hsct :** cause de l'homoscédasticité**, R_McH :** R2 Mc Fadden hétérosc**. H :** hypothèse d'homoscédasticité des erreurs**.**
Err : taux d'erreur**, Sens :** taux de sensibilité**, Spec:** taux de spécificité**.**

On avait aussi décidé de réaliser d'autres modèles avec une indépendance absolue des variables (modele 2 et 3). Dans l'annexe, on a réalisé toutes les étapes faites précédemment et on a porté une attention sur les résultats résumés dans le ***tableau 4*** avec le test d'homoscédasticité en incluant le modèle 1bis.

Les coefficients des variables dans quatre modèles ne sont pas différents sauf pour la variable cat1 on a une différence de 1.7 entre le modèle 1 et 3. Dans le modèle 3, on a le taux d'erreur de prédiction qui est plus élevé et une qualité du modèle R2 McFadden la plus faible avec 5% et 13% pour le modèle 2 qui est inférieur à celui du modèle 1 et 1bis.
Les quatre modèles refusent l'hypothèse d'homoscédasticité des erreurs.
On se doit de retenir le modèle 1bis par le critère R2 Mc Fadden qui est le plus élevé parmi les 4 modèles proposés.

2. Modèle multinomial ordonné

2.A Régression et vérification des tests

Dans cette partie, on veut expliquer le choix de vote des communes sur quatre modalités des partis politiques qui sont classés dans l'ordre suivant : FN, EM, UMP et autre.

Dans un premier temps, avant de réaliser les estimations pour pouvoir expliquer la nouvelle variable Y à quatre modalités, on se doit de vérifier la significativité des variables explicatives sur la probabilité de vote un parti politique. Le premier test avec la fonction polr (annexe, ***figure 13***) indique que le découpage entre les catégories FN/EM/UMP/autre n'a pas de sens puisque la p-value est supérieur à 5% et certaines variables n'a aucune impact sur les différentes catégories (vot.Ins, GGFN, cat1, depuiss).

On décide de réaliser comme dans le modèle binaire un stepwise afin de garder les variables qui pourraient expliquer la variable Y. Les résultats de stepwise (annexe, ***figure 14***) montrent qu'il faudrait choisir les trois variables : txch, tximmi et frontiere.

En effet, on avait vu dans le test précédent que les trois variables pouvaient expliquer significativement le Y, on re-vérifie par le test de significativité des variables explicatives (annexe, ***figure 15***) et on remarque que le découpage des quatre catégories ont du sens car les p-values sont inférieurs à 5%. Txch, tximmi et frontiere ont un impact significatif sur la probabilité de voter pour un tel parti.

Lorsque le txch augmente de 1% alors la probabilité de voter pour le FN va aussi augmenter (car on a ordonné dans cette estimation) ou le fait que la commune soit proche d'une frontière. Et on a une relation négative entre le variable tximmi et voter FN, cela signifie que lorsqu'on augmente de 1% la probabilité va augmenter pour les autres partis (EM, UMP, autres).

Le tableau 5 indique les valeurs des odds ratios, une commune qui est à côté d'une frontière a 2.28 fois plus de chances de voir le parti FN vainqueur par rapport aux communes éloignées des frontières.

Tableau 5: résumé du modèle 1

Modèle 1			
	Signe estim	**exp(coef)**	**eff.mar**
Txch	+ (***)	1.12	0.036
Tximmi	- (***)	0.72	-0.066
frontiere	+ (***)	2.28	0.21
Qualité	55.6 %		
R_Mc	7 %		

Il faut vérifier l'hypothèse de nullité de l'ensemble des coefficients des variables explicatives du modèle avec le test de Chi deux. La p-value est inférieure à 5% alors la probabilité d'accepter la nullité de l'ensemble des coefficients des variables est nulle. Donc le modèle estimé a donc un intérêt. On peut donc passer à l'interprétation des exponentiels des coefficients du modèle multinomial ordonné estimé.

2.B Test de l'égalité des pentes

Nn va vérifier l'homoscédasticité des erreurs et l'égalité des pentes au niveau global avec la fonction vglm sous R et on constate que le logiciel affiche un problème lors de l'estimation (annexe, ***figure 16***).

On décide de catégorise par trois les variable txch (cattxch) et tximmi (cattximmi) pour contourner ce problème pour pouvoir créer quatre autres variables pour les deux derniers tiers . On a cattxch et tximmi qui est séparé en trois parties comme dans le tableau 6 suivant :

Tableau 6: nouvelle variable cattxch et catximmi

cattxch	**[3.33, 11.1]**	**[11.1,14.7]**	**[14.7,41.9]**
Nombre de commune	591	590	590

Catximmi	**[0, 1.39]**	**[1.39, 2.76]**	**[2.76,20.5]**
Nombre de commune	591	590	590

Dans l'annexe, on a réaliser le test de significativité des variables sur Y et le test d'hypothèse de nullité qui sont validés. On peut donc procéder au test d'homoscédasticité des erreurs du modèle et le test d'égalité des pentes. En procédant au test de l'égalités des pentes avec chi deux, on a une p-value inférieur à 5% donc on refuse l'hypothèse pour l'ensemble des variables explicatives de ce modèles. La solution subtile est de chercher la ou les variables ne vérifiant pas l'égalité des pentes et de la retirer comme variable explicative du modèle ordonné.

Après avoir procédé à plusieurs test sur chaque variable (annexe, ***figure 16***), il y a une deux variables qui vérifie l'hypothèse de l'égalité des pentes, cattximmi0 et cattximmi1 qui représentent les deux derniers tiers de tximmi (voir tableau). On se doit de ré-estimer sans les variables cattxch0, cattxch01 et frontiere (annexe, ***figure 16 et 17***), on peut valider l'hypothèse alors cattximmi0 et cattximmi1 vérifient l'égalité des pentes (p-value = 0.69) avec pseudo R2 qui vaut 0.2% qui montre une qualité du modèle très faible. Cependant, ***figure 18*** indique que catximmi1 est la seul variable qui est significative au seuil de 1%.

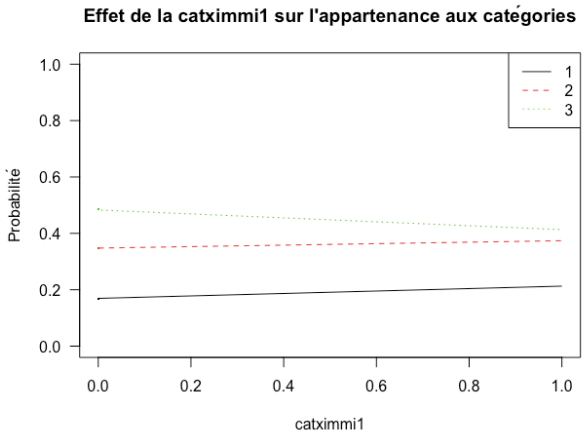
Figure 18 : estimation modele catximmi0 et catximmi1

```
Call:
vglm(formula = vote4 ~ catximmi0 + catximmi1, family = cumulative(parallel = TRUE,
    reverse = TRUE), data = X1, link = "logit")

Pearson residuals:
                      Min      1Q  Median    3Q    Max
logitlink(P[Y>=2]) -2.193  0.2629  0.2646 0.688 0.7501
logitlink(P[Y>=3]) -1.191 -1.0543 -0.3907 1.001 1.1557

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1  1.60770    0.08947  17.969  < 2e-16 ***
(Intercept):2 -0.05070    0.07927  -0.640  0.52247
catximmi0     -0.01813    0.11002  -0.165  0.86908
catximmi1     -0.30037    0.10928  -2.749  0.00598 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])

Residual deviance: 3657.644 on 3538 degrees of freedom

Log-likelihood: -1828.822 on 3538 degrees of freedom

Number of Fisher scoring iterations: 3

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
catximmi0 catximmi1
0.9820286 0.7405423
```

Source : Dossier économétrique, logiciel R

On illustre graphiquement (figure 19) la probabilité de l'effet de la catégorie d'immigration 1 sur le vote d'un partie. Pour la catximmi1, la courbe du vote FN (courbe verte) indique une diminution de la probabilité. La probabilité augmente et diminue pour une commune appartenant à cattximmi0. Les courbes pour le vote « autre » sont parallèles à celle du vote EM donc elles évoluent de la même direction.

Figure 19



Source : Dossier économétrique, logiciel R

2.C Hypothèse d'homoscédasticité des erreurs

Tout d'abord, on réalise de nouvelle estimation avec la fonction oglmx (annexe, ***figure 20***). On suppose que les deux variables retenues sont la cause de l'hétéroscédasticité. Au seuil de 5%, les deux variables cattximmi0 et cattximmi1 n'est pas prise en compte dans l'hétéroscédasticité des erreurs (p-value = 0.69, ***figure 21***), donc aucune des variables n'est coupable alors on accepte l'hypothèse d'homoscédasticité.

Figure 21 : Test de ratio de vraisemblance

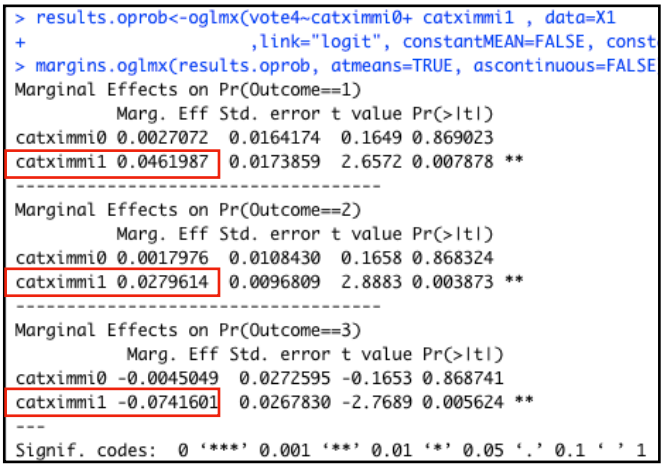
```
Likelihood ratio test

Model 1: vote4 ~ catximmi0 + catximmi1
Model 2: vote4 ~ (catximmi0 + catximmi1 | catximmi0 + catximmi1)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -1828.8
2   6 -1828.5  2 0.7395     0.6909
```

Source : Dossier économétrique, logiciel R

Par conséquent, ayant accepter l'homoscédasticité des erreurs, on doit interpréter l'estimation faite avec la fonction vglm (***figure 18***). La variable catximmi1 étant la seule significative, elle a une relation négative avec le fait de voter pour le FN dans une commune. Pour les communes ayant un taux d'immigration compris entre 2.76% et 20.5%, ont 1.35 (1/0.74) de chances que le parti sortant soit le FN par rapport aux communes qui ont un taux d'immigration inférieur à 2.76%.

Puis, on s'intéresse en effets marginaux (*figure 22*) permettant de mesurer la sensibilité de la probabilité d'un événement par rapport à des variations dans les variables explicatives. Il y a que la variable cattximmi1 qui est significative, ainsi pour un seuil de risque de 5%, une commune ayant un taux d' immigration compris entre 2.76% et 20.5% impact les trois catégorie de vote (FN, EM et autre). En valeur absolue des effets marginaux, l'effet de catximmi1 sera plus forte pour le parti FN et plus faible pour le EM par rapport à la catégorie autres.

Figure 22 : Effets marginaux de l'estimation



```
> results.oprob<-oglmx(vote4~catximmi0+ catximmi1 , data=X1
+                     ,link="logit", constantMEAN=FALSE, const
> margins.oglmx(results.oprob, atmeans=TRUE, ascontinuous=FALSE
Marginal Effects on Pr(Outcome==1)
          Marg. Eff Std. error t value Pr(>|t|)
catximmi0 0.0027072  0.0164174  0.1649 0.869023
catximmi1 0.0461987  0.0173859  2.6572 0.007878 **
-------------------------------------
Marginal Effects on Pr(Outcome==2)
          Marg. Eff Std. error t value Pr(>|t|)
catximmi0 0.0017976  0.0108430  0.1658 0.868324
catximmi1 0.0279614  0.0096809  2.8883 0.003873 **
-------------------------------------
Marginal Effects on Pr(Outcome==3)
           Marg. Eff Std. error t value Pr(>|t|)
catximmi0 -0.0045049  0.0272595 -0.1653 0.868741
catximmi1 -0.0741601  0.0267830 -2.7689 0.005624 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source : Dossier économétrique, logiciel R

V. Conclusion

Déterminer les facteurs du choix des électeurs de chaque commune était la base de cette étude et de proposer un modèle pertinent pouvant expliquer la tendance du vote frontiste. Lors de la modélisation du choix binaire résumé dans le tableau 5, le modèle 1bis a été retenu dont seulement la variable cat1 n'était pas statistiquement significative. Cette variable qualitative binaire a été défini par la méthode CAH afin d'éviter la dépendance entre les variables quantitatives et représente la dimension de la commune si elle est petite ou moyenne. La figure 23 résume le lien entre les hypothèses émis dans l'analyse économique et le résultat du modèle, il y a deux variables qui ne vérifient pas nos hypothèses :

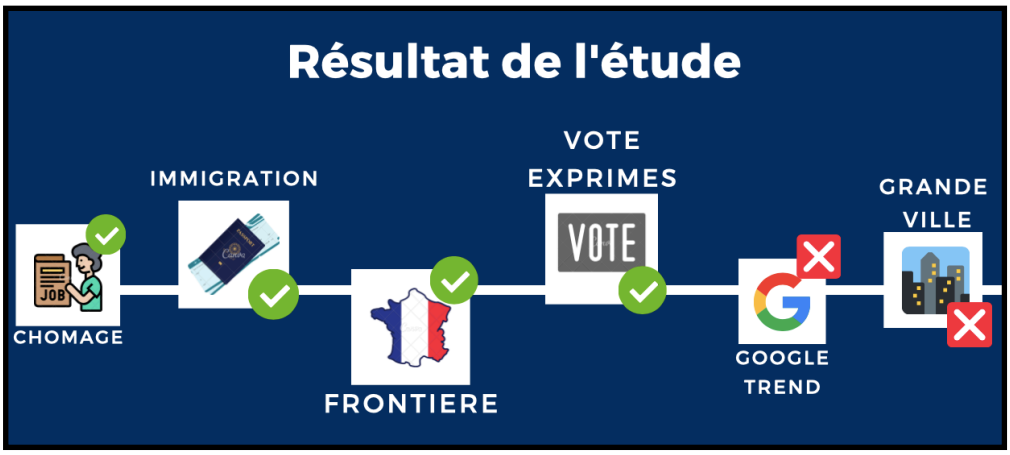
- la variable GGFN qui représentait le comportement des internautes en recherchant des informations sur le moteur de recherche Google du parti FN dans un département, on s'attendait d'avoir une relation positive dans le faite de voter FN en acquérant de l'information sur le programme électoral. Le modèle indique une relation négative, ce qui peut s'expliquer par le fait que les personnes font des recherches par curiosité. Les internautes n'ont pas l'intention de voter pour le FN. De plus, les personnes votant FN ne maitrisent pas son programme électoral[16] ce qui pousse à conclure que c'est même personne ne font pas de recherche sur le parti.

- *Depuiss* est une variable binaire qui indique si la commune appartient à un département ayant une grande ville (comme Paris ou Marseille). C'est une variable mal spécifier, c'est-à-dire que nombreuses petites communes sont pris en compte dans le

département avec une grande ville (exemple : département Nord). Dans l'analyse économique, on avait supposer que les petites et moyennes communes avaient des préférences pour le FN, ce qui fausse notre hypothèse de départ.

Autrement, on a trouvé des résultats cohérents pour les autres variables tels que le taux de chômage, le fait qu'une commune se trouve proche de la frontière ont une relation positive sur le choix de vote FN. Pour une augmentation du taux de chômage et qui réunit les deux autres conditions augmentera la probabilité du FN sortant dans cette commune. Une relation contraire pour les variables du taux de chômage et le taux de vote exprimés, lorsque ces deux taux augmentent dans une commune, alors la probabilité d'avoir dans cette même commune une majorité pour le vote FN diminuera.

Figure 23 : résumé de l'étude



Source : Dossier économétrique

Pour le modèle multinomial ordonnée, on trouvait un modèle avec deux variables (catximmi1 et catximmi0) qui pouvait expliquer le choix des 3 catégories (FN, EM et autre). C'était les seules variables qui vérifier l'égalités des pentes et les termes erreurs homoscédastiques. Les variables catximmi1 et catximmi0 indiquant les deux derniers tiers des catégories dont le taux de d'immigration se trouve entre 2.76% et 20.5% (catximmi1) puis entre 1.39% et 2.76% (catximmi0).

La variable catximmi1 est la seul qui était statistiquement significative. Les résultats montraient que les catégories EM et autre avaient la même relation pour le vote par rapport à catximmi1. Ainsi lorsqu'une commune appartenait à catximmi1, alors la

probabilité d'une commune votant pour EM ou autre augmente et le cas contraire pour le FN.

Le fait d'avoir une commune appartement à cette catégorie de taux de chômage (entre 2.76% et 20.5%) peut avoir une relation directe avec les grandes villes. En effet, en regardant la carte de la répartition de l'immigration dans la France (Annexe, ***figure 2***), on remarque que ce type de population se concentre autour des agglomérations urbaines et dans la ville. Ces grandes villes sont attractive à cause de l'accès facile à l'emploi, logement et une meilleure qualité de vie. On peut dire que la variable catximmi1 pourrait remplacer celle de *depuiss* à cause de cette relation directe.

VI. Limites et discussion

Des solutions à l'étude et proposition d'amélioration

Malgré que certaines variables sont statistiquement significatives, elles ne respectaient pas les hypothèses citées au début du dossier. Il est actuellement possible de trouver des solutions pour l'une des variables comme ***depuiss***.

A cause d'un classement des communes mal spécifier par la variable ***depuiss*** dans notre base, elle a faussé notre hypothèse de départ à cause de l'absence des communes dans la région parisienne. On pourrait indiquer par exemple les communes étant proche d'une grande ville avec une distance connu par des coordonnées géographique (altitude et longitude). Il est possible d'intervenir l'économétrie spatial et modéliser le choix de vote selon ce critère.

Pour la variable de Google donne une information ambiguë, elle donnait le parti sortant de recherche en majorité dans un département, ce qui est intéressant, c'est d'avoir des données à l'échelle communal afin de mieux expliquer par cette variable par les recherches sur le parti FN

Pour améliorer la qualité des modèles, il est tout à fait possible d'ajouter d'autre variable importante pouvant expliquer le choix de FN l'information sur le type de religion pratiqué dans la commune ou les origines ethniques des personnes. Effectivement, le Front National crée une hostilité contre l'Islam[17] et les personnes d'origines africaines. Une

autre variable permettant d'améliorer les modèles, c'est en ajoutant l'indice de popularité mesuré par l'IFOP (variable proposé par Bélanger, Fauvelle-Aymar et Lewis-Beck, 2017).

L'utilité de cette étude dans la vie réelle.

En addition, ce modèle économétrique est une bonne manière de faire des prévision électoral ou proposé des solutions stratégies de communication. Par exemple, un tel parti politique pourrait intervenir à travers des « meetings » dans les communes ayant un taux de chômage élevé ou proches des frontières pour optimiser l'audience et impacter le plus de personne possible. Un modèle de prédiction pourrait se modéliser avec les données de l'INSEE sur la population pour les prochaines élections en 2022 avec une enquête de la même année.

Avec une étude plus poussée en Big Data par le biais d'algorithme sophistiqué (IA) et la récolte de données provenant des réseaux sociaux, il est possible de prédire ou d'influencer le choix des électeurs . Par exemple, en Etats-Unis, l'un des plus grand scandale entre le Géant Facebook et Cambridge Analytica (CA)[18]. En utilisant les données personnelles d'utilisateurs, CA a pu influencer le vote pour le Brexit ou à la campagne électorale de Donald Trump. Si le FN se procurait des services de CA, serait-il possible que Marine Le Pen gagne l'élection de 2017…

La question que l'on peut se poser : **« Aujourd'hui, la démocratie n'est-elle pas en danger avec l'évolution de la technologie et l'IA ? »**

Bibliographie

1. Nicolas Lebourg, « Ordre nouveau, fin des illusions droitières et matrice activiste du premier Front national » Studia Historica. Historia Contemporánea, Vol. 30 (Derecha radical, fascismo y extrema derecha en Europa y América), 2012, 209.
2. Grégoire Kauffmann, « Les origines du Front National », Pouvoirs, 2016/2, PP 5-15. (lien)
3. Gouvernement, Ministère de l'Intérieure, « Résultats de l'élection présidentielle 2017 ». (lien)
4. J. Bouchet-Peterson « Qui sont les 21.4% d'électeurs de Marine Le Pen », Libération, 24/04/17. (lien)
5. Statuts du Rassemblement National, lien.
6. Alternatives économique, « Evolution du vote FN aux principales élections, en % des inscrits », lien.
7. Gouvernement, Ministère de l'Intérieure, résultat présidentiel du premier tour, lien
8. Wihtol de Wenden, Catherine. « Ouverture et fermeture de la France aux étrangers. Un siècle d'évolution », *Vingtième Siècle. Revue d'histoire*, vol. no 73, no. 1, 2002, pp. 27-38.
9. Jaffré, Jérôme. « Chapitre 16 - Un second tour de présidentielle si différent du duel Chirac-Le Pen de 2002 »,
10. Pascal Perrineau éd., *Le vote disruptif. Les élections présidentielle et législatives de 2017.* Presses de Sciences Po, 2017, pp. 269-284.
11. Chalard, Laurent. « France, Le palmarès des agglomérations », *Population & Avenir*, vol. 680, no. 5, 2006, pp. 4-8.
12. Mayer Nonna, « Les constantes du vote FN », Revue Projet, 2016/5 (N° 354), p. 11-14. DOI : 10.3917/pro.354.0011. URL : lien
13. Léger, Jean-François. « Le chômage, terreau du vote Front national ? », *Population & Avenir*, vol. 723, no. 3, 2015, pp. 4-7.
14. Perrineau, Pascal. « Chapitre 15 - Marine Le Pen au premier tour : la puissance d'une dynamique, l'échec d'une ambition », Pascal Perrineau éd., *Le vote disruptif. Les élections présidentielle et législatives de 2017.* Presses de Sciences Po, 2017, pp. 251-268.
15. Michèle Tribalat, « Une estimation des populations d'origine étrangère en France en 2011 », *Espace populations sociétés* [En ligne], 2015/1-2 | 2015.
16. S. Billard, « Qui vote FN ? Pourquoi ? 3 idées reçues sur les électeurs du Front National, Nouvelobs, 22 mars 2017, (lien)
17. « Proposition de résolution de Marion Maréchal-Le Pen et Gilbert Collard, 2015 », https://rassemblementnational.fr/terme/islam/
18. « Facebook says Cambridge Analytica may have gained 37m more users data », Guardian

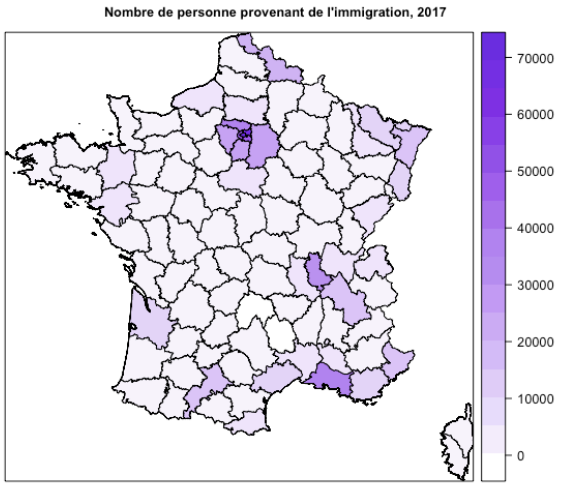
Annexe

Tableau 1 : table croisée entre vote et depuiss

Parti politique/nb de commune pour depuiss	FN	EM	UMP	Autre	Total
0	621	510	130	133	1394
1	196	120	34	27	377
Total	817	630	164	160	1771

Source : gouvernement France, logiciel R

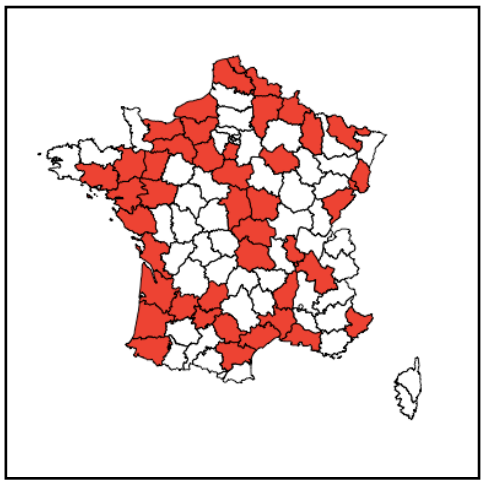
Figure 2 : Carte d'immigration en France



Source : Insee, 2016 avec logiciel R

Retour partie taux d'immigration

Figure 3 : Google Trend, recherche FN



Source : Google Trend, logiciel R

Retour partie google

Figure 7: stepwise

Stepwise forward

```
Step:  AIC=1798.51
FN ~ txch + tximmi + frontiere + cadres + industrie + dipsup +
    actifs + HLML + X..Vot.Ins + ouvrier + depuiss + GGFN + interim +
    construction + agri

                       Df Deviance    AIC
<none>                      1766.5 1798.5
+ X..Abs.Ins            1   1764.8 1798.8
+ artis_chef_dentrepri  1   1765.4 1799.4
+ GGEM                  1   1765.6 1799.6
+ GGautre               1   1766.1 1800.1
+ retraite              1   1766.5 1800.5
```

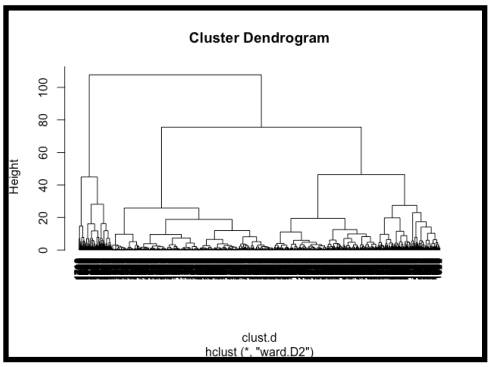
Stepwise backward

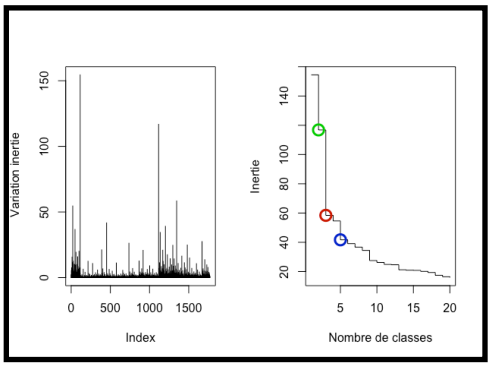
```
Step:  AIC=1797.18
FN ~ actifs + tximmi + X..Vot.Ins + txch + dipsup + agri + industrie +
    HLML + cadres + construction + interim + frontiere + depuiss +
    GGFN + GGUMP

              Df Deviance    AIC
<none>             1765.2 1797.2
- GGUMP         1   1767.9 1797.9
- industrie     1   1767.9 1797.9
- GGFN          1   1768.2 1798.2
- construction  1   1768.4 1798.4
- agri          1   1771.2 1801.2
- cadres        1   1774.0 1804.0
- interim       1   1775.7 1805.7
- X..Vot.Ins    1   1777.1 1807.1
- depuiss       1   1777.2 1807.2
- HLML          1   1788.3 1818.3
- tximmi        1   1794.5 1824.5
- frontiere     1   1843.3 1873.3
- actifs        1   1846.7 1876.7
- dipsup        1   1887.6 1917.6
- txch          1   1929.6 1959.6
```

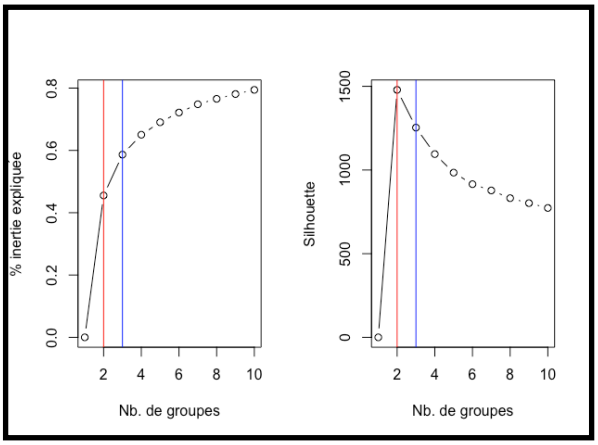
Stepwise both

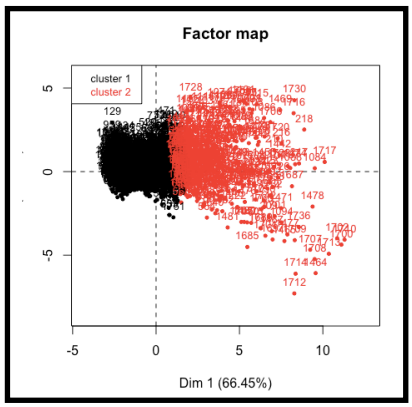
```
Step:  AIC=1797.18
FN ~ txch + tximmi + frontiere + cadres + industrie + dipsup +
    actifs + HLML + X..Vot.Ins + GGUMP + depuiss + interim +
    construction + agri + GGFN

                       Df Deviance    AIC
<none>                      1765.2 1797.2
+ X..Abs.Ins            1   1763.3 1797.3
+ artis_chef_dentrepri  1   1763.7 1797.7
+ ouvrier               1   1763.7 1797.7
- GGUMP                 1   1767.9 1797.9
- industrie             1   1767.9 1797.9
- GGFN                  1   1768.2 1798.2
- construction          1   1768.4 1798.4
+ retraite              1   1765.2 1799.2
+ GGEM                  1   1765.2 1799.2
+ GGautre               1   1765.2 1799.2
- agri                  1   1771.2 1801.2
- cadres                1   1774.0 1804.0
- interim               1   1775.7 1805.7
- X..Vot.Ins            1   1777.1 1807.1
- depuiss               1   1777.2 1807.2
- HLML                  1   1788.3 1818.3
- tximmi                1   1794.5 1824.5
- frontiere             1   1843.3 1873.3
- actifs                1   1846.7 1876.7
- dipsup                1   1887.6 1917.6
- txch                  1   1929.6 1959.6
```

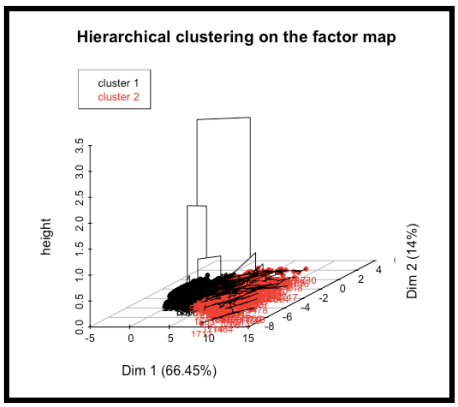
Classification ascendante hiérarchique CAH

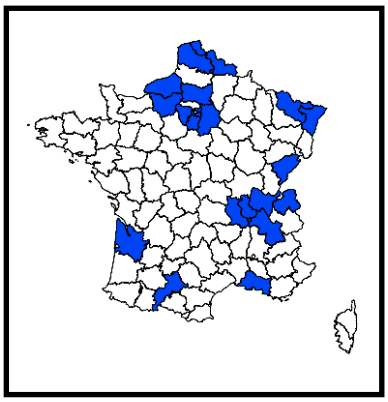












Commentaire du CAH :

La méthode de classification ascendante va résumer l'ensemble des variables afin d'éviter de biaisé les modèles. Elle permet de regrouper toutes les variables qui sont corrélés comme dans l'ACP et catégorise les observations. Le dendogramme (figure) classe toutes les observations dans un arbre et aide sur le choix du nombre de classe à choisir. La variation d'inertie, le pourcentage d'inertie expliquée et la silhouette pousse à choisir 2 groupes. En effet, cela porte sur le même choix des k-means d'après le graphique de la silhouette, on choisi le point maximum indiquer par la droite horizontal rouge.

On représente toutes les observations sur un plan en deux dimension dont la première dimension représente l'inertie la plus important et explique à elle seul 66.45% de l'inertie totale et 14% pour la deuxième dimension. Les observations sont présentées en deux couleurs noir et rouge pour les deux groupes. Il est important de nommer ces deux groupes, en analysant les différentes caractéristiques sur le nombre d'actifs ou le nombre personne ayant fait des études supérieures dans la base selon le groupe, on remarque que les communes représenter en rouge (cat2) dans le plan, ce sont essentiellement des communes se trouvant dans la régions parisiennes ou des communes dans l'agglomération de grandes villes comme Toulouse, Marseille ou Lyon. Par conséquent, on nomme la catégorie rouge de « grande commune » et celle en noir (cat1) de « petite et moyenne commune ». La carte représente la répartition de la catégorie 2 en bleue dans la France, on a 160 communes qui se trouvent la catégorie 2 et 1611 dans la catégorie 1.

Figure 8: Stepwise avec clusters

Stepwise forward

```
Step:  AIC=2067.66
FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss

        Df Deviance    AIC
<none>       2051.7 2067.7
+ GGUMP  1   2050.3 2068.3
```

Stepwise Both

```
Step:  AIC=2067.66
FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss

             Df Deviance    AIC
<none>            2051.7 2067.7
+ GGUMP       1   2050.3 2068.3
- depuiss     1   2055.7 2069.7
- cat1        1   2055.9 2069.9
- GGFN        1   2058.5 2072.5
- X..Vot.Ins  1   2060.7 2074.7
- frontiere   1   2125.3 2139.3
- tximmi      1   2156.8 2170.8
- txch        1   2214.2 2228.2
```

Figure 9 : Vérification des tests

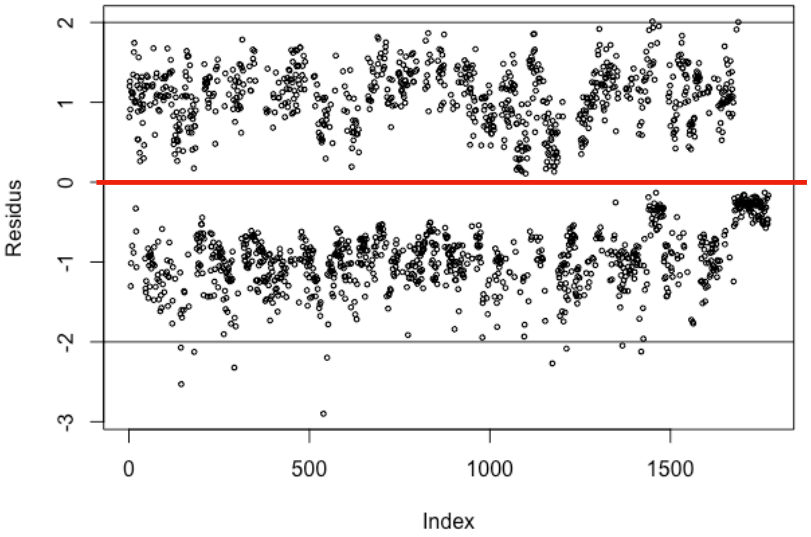
Retour parti 1.C

```
> vif(modele_test)
     txch     tximmi  frontiere X..Vot.Ins       GGFN       cat1    depuiss
 1.296389   1.577365   1.098332   1.348127   1.119110   1.234922   1.125026
> chi2<- (modele_test$null.deviance-modele_test$deviance)
> ddl<-modele_test$df.null-modele_test$df.residual
> pvalue<-pchisq(chi2,ddl,lower.tail=F)
> print(pvalue)
[1] 8.124686e-81
> exp(coef(modele_test))
(Intercept)        txch      tximmi  frontiere1  X..Vot.Ins       GGFN1       cat11    depuiss1
  5.5291632   1.1970465   0.7200035   2.8668295   0.9491573   0.7458381   1.8234002   1.3205767
> #calcul des effets marginaux
> mean(dlogis(predict(modele_test,type="link")))*coef(modele_test)
(Intercept)        txch      tximmi  frontiere1  X..Vot.Ins       GGFN1       cat11    depuiss1
 0.34133743  0.03590099 -0.06557115  0.21022877 -0.01041570 -0.05853447  0.11990529  0.05550477
> #tableau de prévision et pourcentage d'erreur du modeles estime
> pred.proba<-predict(modele_test,type="response")
> pred.moda<-factor(ifelse(pred.proba>0.5,"1","0"))
> mc<-table(X1$FN,pred.moda)
> print(mc)
   pred.moda
      0   1
  0 706 248
  1 337 480
> err<-(mc[2,1]+mc[1,2])/sum(mc)
> print(err)
[1] 0.3303219
> Sensibilite<-mc[2,2]/(mc[2,1]+mc[2,2])
> print(Sensibilite)
[1] 0.5875153
> Specificite<-mc[1,1]/(mc[1,1]+mc[1,2])
> print(Specificite)
[1] 0.7400419
> library(pscl)
```

Source : Dossier économétrie, logiciel R

Figure 10 : Fonction Hitmiss et plot de résidu

```
> hitmiss(modele_test)
Classification Threshold = 0.5
       y=0 y=1
yhat=0 706 337
yhat=1 248 480
Percent Correctly Predicted = 66.97%
Percent Correctly Predicted = 74%, for y = 0
Percent Correctly Predicted = 58.75%  for y = 1
Null Model Correctly Predicts 53.87%
[1] 66.96781 74.00419 58.75153
> R2_Mc_Fadden<-1-(modele_test$deviance/modele_test$null.deviance)
> R2_Mc_Fadden
[1] 0.1607081
```



[Retour parti 1.C](#)

Figure 11 : Hétéroscédasticité du modèle 1

```
Call:
hetglm(formula = FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1
    txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss, data =

Deviance residuals:
    Min      1Q  Median      3Q     Max
-2.2274 -1.0132 -0.1798  1.0442  2.2834

Coefficients (binomial model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0443412  0.0961485   0.461    0.645
txch         0.0025691  0.0053664   0.479    0.632
tximmi      -0.0043751  0.0091193  -0.480    0.631
frontiere1   0.0122785  0.0257054   0.478    0.633
X..Vot.Ins  -0.0009275  0.0019691  -0.471    0.638
GGFN1       -0.0024896  0.0054165  -0.460    0.646
cat11        0.0041474  0.0091795   0.452    0.651
depuiss1     0.0030026  0.0064261   0.467    0.640

Latent scale model coefficients (with log link):
           Estimate Std. Error z value Pr(>|z|)
txch        0.06670    0.01516   4.399 1.09e-05 ***
tximmi     -0.07357    0.03072  -2.395  0.01661 *
frontiere1 -0.68447    0.16515  -4.145 3.40e-05 ***
X..Vot.Ins -0.06487    0.02435  -2.664  0.00772 **
GGFN1      -0.26265    0.14950  -1.757  0.07894 .
cat11       0.55447    0.24051   2.305  0.02114 *
depuiss1    0.11155    0.15199   0.734  0.46301
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -1002 on 15 Df
LR test for homoskedasticity:  47.3 on 7 Df, p-value: 4.88e-08
Dispersion: 1
Number of iterations in nlminb optimization: 37
```

```
Call:
hetglm(formula = FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss |
    txch + frontiere + cat1, data = X1, family = binomial(logit))

Deviance residuals:
    Min      1Q  Median      3Q     Max
-2.1652 -1.0032 -0.1815  1.0309  2.1502

Coefficients (binomial model with logit link):
           Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.99424    6.01437   1.495 0.134795
txch         0.80213    0.24455   3.280 0.001038 **
tximmi      -1.39293    0.40061  -3.477 0.000507 ***
frontiere1   3.78462    1.10532   3.424 0.000617 ***
X..Vot.Ins  -0.22912    0.09049  -2.532 0.011343 *
GGFN1       -0.97637    0.49938  -1.955 0.050564 .
cat11        1.31543    0.90265   1.457 0.145033
depuiss1     1.12636    0.55555   2.027 0.042614 *

Latent scale model coefficients (with log link):
           Estimate Std. Error z value Pr(>|z|)
txch        0.07296    0.01265   5.769 7.99e-09 ***
frontiere1 -0.56848    0.15196  -3.741 0.000183 ***
cat11       0.59025    0.20757   2.844 0.004461 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -1007 on 11 Df
LR test for homoskedasticity: 37.99 on 3 Df, p-value: 2.846e-08
Dispersion: 1
Number of iterations in nlminb optimization: 13
> vif(modeleh)
    txch     tximmi  frontiere X..Vot.Ins       GGFN       cat1    depuiss
19.262476 13.579974  8.434039  2.605752  1.728437  1.792932  1.665720
```

```
Call:
hetglm(formula = FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss |
    txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1, data = X1, family = binomial(logit))

Deviance residuals:
    Min      1Q  Median      3Q     Max
-2.2437 -1.0134 -0.1907  1.0443  2.2526

Coefficients (binomial model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0413424  0.0894693   0.462    0.644
txch         0.0022330  0.0046665   0.479    0.632
tximmi      -0.0038938  0.0081175  -0.480    0.631
frontiere1   0.0106824  0.0223736   0.477    0.633
X..Vot.Ins  -0.0008391  0.0017820  -0.471    0.638
GGFN1       -0.0022054  0.0047916  -0.460    0.645
cat11        0.0036917  0.0081247   0.454    0.650
depuiss1     0.0027851  0.0059179   0.471    0.638

Latent scale model coefficients (with log link):
           Estimate Std. Error z value Pr(>|z|)
txch        0.06614    0.01521   4.349 1.37e-05 ***
tximmi     -0.07320    0.03064  -2.389   0.0169 *
frontiere1 -0.67081    0.16254  -4.127 3.67e-05 ***
X..Vot.Ins -0.06598    0.02429  -2.717   0.0066 **
GGFN1      -0.23136    0.14593  -1.585   0.1129
cat11       0.51933    0.23512   2.209   0.0272 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -1002 on 14 Df
LR test for homoskedasticity: 46.81 on 6 Df, p-value: 2.041e-08
Dispersion: 1
Number of iterations in nlminb optimization: 37
```

```
Call:
hetglm(formula = FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss |
    txch + frontiere, data = X1, family = binomial(logit))

Deviance residuals:
    Min      1Q  Median      3Q     Max
-2.2751 -0.9969 -0.3407  1.0235  2.1542

Coefficients (binomial model with logit link):
           Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.94073    2.97683   1.660  0.09697 .
txch         0.41660    0.07484   5.566 2.60e-08 ***
tximmi      -0.76143    0.14198  -5.363 8.19e-08 ***
frontiere1   1.99030    0.35296   5.639 1.71e-08 ***
X..Vot.Ins  -0.12427    0.03817  -3.256  0.00113 **
GGFN1       -0.50556    0.22007  -2.297  0.02160 *
cat11        0.88594    0.55796   1.588  0.11233
depuiss1     0.65641    0.26879   2.442  0.01460 *

Latent scale model coefficients (with log link):
           Estimate Std. Error z value Pr(>|z|)
txch        0.06337    0.01155   5.486 4.11e-08 ***
frontiere1 -0.53323    0.14466  -3.686 0.000228 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -1011 on 10 Df
LR test for homoskedasticity: 29.02 on 2 Df, p-value: 5e-07
Dispersion: 1
Number of iterations in nlminb optimization: 10
> vif(modeleh)
    txch     tximmi  frontiere X..Vot.Ins       GGFN       cat1    depuiss
6.670415  5.705479  3.189862  1.656775  1.239788  1.267555  1.331488
```


Source : dossier économétrie, logiciel R

Figure 12 : R2 MC Fadden et fonction

```
> (R2McFAdden<-1-(modeleh$loglik/h1c$loglik))
[1] 0.1725784
> h1h <- hetglm(FN~txch + tximmi + frontiere +
+                 X..Vot.Ins + GGFN + cat1 + depuiss| 1
+               ,data=X1,family=binomial(logit) )
> summary(h1h)

Call:
hetglm(formula = FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss |
    1, data = X1, family = binomial(logit))

Deviance residuals:
    Min      1Q  Median      3Q     Max
-2.8822 -0.9877 -0.3168  1.0566  2.0067

Coefficients (binomial model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.71004    1.51861   1.126  0.26014
txch         0.17986    0.01541  11.673  < 2e-16 ***
tximmi      -0.32850    0.03630  -9.049  < 2e-16 ***
frontiere1   1.05321    0.12542   8.397  < 2e-16 ***
X..Vot.Ins  -0.05218    0.01750  -2.981  0.00287 **
GGFN1       -0.29325    0.11260  -2.604  0.00920 **
cat11        0.60070    0.29227   2.055  0.03985 *
depuiss1     0.27807    0.13781   2.018  0.04361 *

> lrtest(h1h,modeleh)
Likelihood ratio test

Model 1: FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss |
    1
Model 2: FN ~ txch + tximmi + frontiere + X..Vot.Ins + GGFN + cat1 + depuiss |
    txch + frontiere
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -1025.8
2  10 -1011.3  2 29.017     5e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source : dossier économétrie, logiciel R

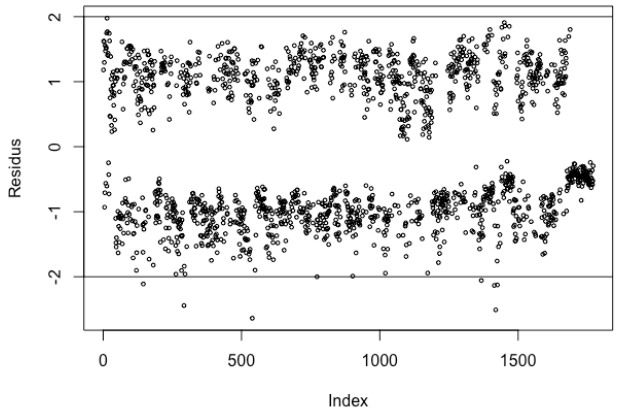
Retour parti 1.C

Estimation modèle 2

```
Call:
glm(formula = (FN ~ txch + tximmi + X..Vot.Ins + depuiss), family = binomial(logit),
    data = X1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6225  -1.0038  -0.4776   1.0932   1.9718

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.86874    1.47516   1.945 0.051811 .
txch         0.17260    0.01457  11.844  < 2e-16 ***
tximmi      -0.29853    0.02902 -10.289  < 2e-16 ***
X..Vot.Ins  -0.05758    0.01715  -3.357 0.000788 ***
depuiss1     0.31486    0.12652   2.488 0.012829 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2444.5  on 1770  degrees of freedom
Residual deviance: 2136.4  on 1766  degrees of freedom
AIC: 2146.4

Number of Fisher Scoring iterations: 4
```

```
> hitmiss(modele_test)
Classification Threshold = 0.5
       y=0 y=1
yhat=0 708 356
yhat=1 246 461
Percent Correctly Predicted = 66.01%
Percent Correctly Predicted = 74.21%, for y = 0
Percent Correctly Predicted = 56.43%  for y = 1
Null Model Correctly Predicts 53.87%
[1] 66.00791 74.21384 56.42595
> R2_Mc_Fadden<-1-(modele_test$deviance/modele_te
> R2_Mc_Fadden
[1] 0.1260369
```

```
> vif(modele_test)
      txch     tximmi X..Vot.Ins    depuiss
  1.253732   1.334102   1.379408   1.008741
> chi2<- (modele_test$null.deviance-modele_test$deviance)
> ddl<-modele_test$df.null-modele_test$df.residual
> pvalue<-pchisq(chi2,ddl,lower.tail=F)
> print(pvalue)
[1] 1.938623e-65
> exp(coef(modele_test))
(Intercept)        txch     tximmi  X..Vot.Ins    depuiss1
 17.6147430   1.1883894   0.7419108   0.9440503   1.3700606
> #calcul des effets marginaux
> mean(dlogis(predict(modele_test,type="link")))*coef(modele_test)
(Intercept)        txch     tximmi  X..Vot.Ins    depuiss1
 0.59991400  0.03609413 -0.06242820 -0.01204034  0.06584291
> #tableau de prévision et pourcentage d'erreur du modeles estime
> pred.proba<-predict(modele_test,type="response")
> pred.moda<-factor(ifelse(pred.proba>0.5,"1","0"))
> mc<-table(X1$FN,pred.moda)
> print(mc)
   pred.moda
      0   1
  0 708 246
  1 356 461
> err<-(mc[2,1]+mc[1,2])/sum(mc)
> print(err)
[1] 0.3399209
> Sensibilite<-mc[2,2]/(mc[2,1]+mc[2,2])
> print(Sensibilite)
[1] 0.5642595
> Specificite<-mc[1,1]/(mc[1,1]+mc[1,2])
> print(Specificite)
[1] 0.7421384
```

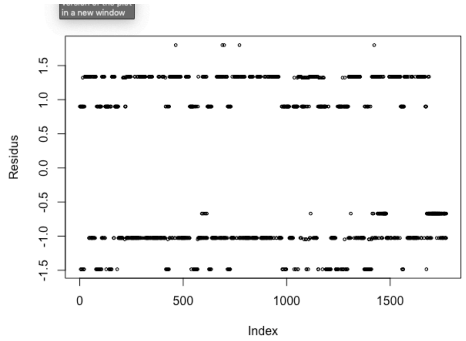


Estimation modèle 3

```
Call:
glm(formula = (FN ~ cat1 + frontiere), family = binomial(logit)
    data = X1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4839  -1.0268  -0.6683   1.3358   1.7938

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3856     0.1888  -7.337 2.18e-13 ***
cat11         1.0205     0.1910   5.343 9.15e-08 ***
frontiere1    1.0617     0.1118   9.498  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2444.5  on 1770  degrees of freedom
Residual deviance: 2322.6  on 1768  degrees of freedom
AIC: 2328.6

Number of Fisher Scoring iterations: 4
```

```
> hitmiss(modele_test)
Classification Threshold = 0.5
       y=0 y=1
yhat=0 790 539
yhat=1 164 278
Percent Correctly Predicted = 60.3%
Percent Correctly Predicted = 82.81%, for y = 0
Percent Correctly Predicted = 34.03%  for y = 1
Null Model Correctly Predicts 53.87%
[1] 60.30491 82.80922 34.02693
> R2_Mc_Fadden<-1-(modele_test$deviance/modele_test$null.deviance)
> R2_Mc_Fadden
[1] 0.04986741
```

```
> library(car)
> vif(modele_test)
     cat1 frontiere
 1.007895  1.007895
> chi2<- (modele_test$null.deviance-modele_test$deviance)
> ddl<-modele_test$df.null-modele_test$df.residual
> pvalue<-pchisq(chi2,ddl,lower.tail=F)
> print(pvalue)
[1] 3.383435e-27
> exp(coef(modele_test))
(Intercept)       cat11  frontiere1
  0.2501849   2.7746311   2.8913182
> #calcul des effets marginaux
> mean(dlogis(predict(modele_test,type="link")))*coef(modele_test)
(Intercept)       cat11  frontiere1
 -0.3212506   0.2366142   0.2461655
> #tableau de prévision et pourcentage d'erreur du modeles estime
> pred.proba<-predict(modele_test,type="response")
> pred.moda<-factor(ifelse(pred.proba>0.5,"1","0"))
> mc<-table(X1$FN,pred.moda)
> print(mc)
   pred.moda
      0   1
  0 790 164
  1 539 278
> err<-(mc[2,1]+mc[1,2])/sum(mc)
> print(err)
[1] 0.3969509
> Sensibilite<-mc[2,2]/(mc[2,1]+mc[2,2])
> print(Sensibilite)
[1] 0.3402693
> Specificite<-mc[1,1]/(mc[1,1]+mc[1,2])
> print(Specificite)
[1] 0.8280922
```



Retour parti 1.C

Estimation modèle 1bis

```
> summary(modele1bis)

Call:
glm(formula = (FN ~ txch + tximmi + frontiere + X..Vot.Ins +
    GGFN + cat1 + depuiss), family = binomial(logit), data = base_resid2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0497  -0.9752  -0.2791   1.0306   2.0095

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.32579    1.54641   0.857  0.39126
txch         0.20928    0.01644  12.727  < 2e-16 ***
tximmi      -0.35871    0.03798  -9.445  < 2e-16 ***
frontiere1   1.14493    0.12910   8.869  < 2e-16 ***
X..Vot.Ins  -0.05103    0.01782  -2.865  0.00418 **
GGFN1       -0.32148    0.11461  -2.805  0.00503 **
cat11        0.57969    0.30319   1.912  0.05588 .
depuiss1     0.30730    0.14076   2.183  0.02903 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2429.0  on 1758  degrees of freedom
Residual deviance: 1988.1  on 1751  degrees of freedom
AIC: 2004.1

Number of Fisher Scoring iterations: 5
```

```
> library(car)
> vif(modele1bis)
      txch     tximmi  frontiere X..Vot.Ins       GGFN       cat1    depuiss
  1.312605   1.619675   1.112346   1.341730   1.118445   1.234602   1.127369
> chi2<- (modele1bis$null.deviance-modele1bis$deviance)
> ddl<-modele1bis$df.null-modele1bis$df.residual
> pvalue<-pchisq(chi2,ddl,lower.tail=F)
> print(pvalue)
[1] 3.979165e-91
> exp(coef(modele1bis))
(Intercept)        txch      tximmi  frontiere1  X..Vot.Ins       GGFN1        cat11    depuiss1
  3.7651703   1.2327883   0.6985786   3.1422286   0.9502474   0.7250757   1.7854813   1.3597510
> #calcul des effets marginaux
> mean(dlogis(predict(modele1bis,type="link")))*coef(modele1bis)
 (Intercept)         txch       tximmi   frontiere1   X..Vot.Ins        GGFN1         cat11
 0.256947705  0.040559590 -0.069519975  0.221895654 -0.009890533 -0.062304856  0.112347473
    depuiss1
 0.059557133
```

```
> print(mc)
   pred.moda
      0   1
  0 699 245
  1 324 491
> err<-(mc[2,1]+mc[1,2])/sum(mc)
> print(err)
[1] 0.3234792
> Sensibilite<-mc[2,2]/(mc[2,1]+mc[2,2])
> print(Sensibilite)
[1] 0.602454
> Specificite<-mc[1,1]/(mc[1,1]+mc[1,2])
> print(Specificite)
[1] 0.7404661
> library(pscl)
> hitmiss(modele1bis)
Classification Threshold = 0.5
       y=0 y=1
yhat=0 699 324
yhat=1 245 491
Percent Correctly Predicted = 67.65%
Percent Correctly Predicted = 74.05%, for y = 0
Percent Correctly Predicted = 60.25%  for y = 1
Null Model Correctly Predicts 53.67%
[1] 67.65208 74.04661 60.24540
> R2_Mc_Fadden<-1-(modele1bis$deviance/modele1bis$null.deviance)
> R2_Mc_Fadden
[1] 0.1815164
```

Figure 15 : estimation du modèle multinomial ordonné

```
> p<-pnorm(abs(ctable[,"t value"]),lower.tail=FALSE)*2
> p2<-round(p,4)
> (ctable<-cbind(ctable,pvalue=p2))
               Value Std. Error    t value pvalue
txch        0.1104950 0.01179940   9.364461      0
tximmi     -0.2478930 0.01909384 -12.982880      0
frontiere1  0.8274323 0.11100068   7.454300      0
0|UMP      -1.5424549 0.16619496  -9.280997      0
UMP|EM     -0.6823502 0.15285550  -4.464022      0
EM|FN       1.1757839 0.15470642   7.600097      0

 Call:
 polr(formula = vote ~ 1, data = X1, method = c("logistic"))

 No coefficients

 Intercepts:
       Value    Std. Error t value
 0|UMP   -2.3106   0.0829   -27.8631
 UMP|EM  -1.4970   0.0615   -24.3534
 EM|FN    0.1550   0.0477     3.2511

 Residual Deviance: 4116.246
 AIC: 4122.246
 > require(lmtest)
 > lrtest(modelord,modelord0)
 Likelihood ratio test

 Model 1: vote ~ txch + tximmi + frontiere
 Model 2: vote ~ 1
   #Df  LogLik Df  Chisq Pr(>Chisq)
 1   6 -1913.6
 2   3 -2058.1 -3 289.11  < 2.2e-16 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
 > exp(coef(modelord))
       txch     tximmi frontiere1
  1.1168308  0.7804435  2.2874378
 > R2_Mc_Fadden<-1-(modelord$deviance/modelord0$deviance)
 > R2_Mc_Fadden
 [1] 0.0702358
 > X1$predict.m <- predict(modelord)
 > (mc<-table(X1$predict.m , X1$vote))

         0 UMP  EM  FN
   0    31   0  25   0
   UMP   0   0   0   0
   EM   32  63 282 144
   FN   97 101 323 673
 > qualite<-((mc[1,1]+mc[2,2]+mc[3,3]+mc[4,4])/sum(mc))*100
 > print(qualite)
 [1] 55.67476
```

Source : Dossier économétrique, logiciel R

Retour parti 2.A

Figure 15 : Problème fonction vglm

```
> fitmodel<-vglm(vote1~ txch + tximmi + frontiere  ,data=X1,link="logit",
+                family=cumulative(parallel=TRUE,reverse=TRUE))
> fitmodel2<-vglm(vote1~ txch + tximmi + frontiere  ,data=X1,link="logit",
+                 family=cumulative(parallel=FALSE,reverse=TRUE))
Error in lm.fit(X.vlm, y = z.vlm, ...) : NA/NaN/Inf dans 'y'
De plus : Warning message:
In Deviance.categorical.data.vgam(mu = mu, y = y, w = w, residuals = residuals,  :
  fitted values close to 0 or 1
> 1-pchisq(deviance(fitmodel)-deviance(fitmodel2),df=df.residual(fitmodel)-
+            df.residual(fitmodel2))
[1] 0
```

Source : Dossier économétrique, logiciel R

Retour parti 2.B

Figure 16 : Vérification de l'égalité des pentes

```
> #verification des pentes selon les variables
> #pour la variable cattxch0
> fitmodel<-vglm(vote4~ cattxch0+cattxch1  + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                family=cumulative(parallel=TRUE,reverse=TRUE))
> fitmodel2<-vglm(vote4~ cattxch0+cattxch1  + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                 family=cumulative(parallel=FALSE~+1+cattxch0,reverse=TRUE))
> 1-pchisq(deviance(fitmodel)-deviance(fitmodel2),df=df.residual(fitmodel)-
+            df.residual(fitmodel2))
[1] 0.03319028
> #pour la variable cattxch1
> fitmodel<-vglm(vote4~ cattxch0+cattxch1  + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                family=cumulative(parallel=TRUE,reverse=TRUE))
> fitmodel2<-vglm(vote4~ cattxch0+cattxch1  + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                 family=cumulative(parallel=FALSE~+1+cattxch1,reverse=TRUE))
> 1-pchisq(deviance(fitmodel)-deviance(fitmodel2),df=df.residual(fitmodel)-
+            df.residual(fitmodel2))
[1] 0
> #pour la variable catimmi0
> fitmodel<-vglm(vote4~ cattxch0+cattxch1 + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                family=cumulative(parallel=TRUE,reverse=TRUE))
> fitmodel2<-vglm(vote4~ cattxch0+cattxch1 + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                 family=cumulative(parallel=FALSE~1+catximmi0,reverse=TRUE))
> 1-pchisq(deviance(fitmodel)-deviance(fitmodel2),df=df.residual(fitmodel)-
+            df.residual(fitmodel2))
[1] 0.7960985
> #pour la variable catimmi1
> fitmodel<-vglm(vote4~ cattxch0+cattxch1 + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                family=cumulative(parallel=TRUE,reverse=TRUE))
> fitmodel2<-vglm(vote4~ cattxch0+cattxch1 + catximmi0+ catximmi1 + frontiere  ,data=X1,link="logit",
+                 family=cumulative(parallel=FALSE~1+catximmi1,reverse=TRUE))
> 1-pchisq(deviance(fitmodel)-deviance(fitmodel2),df=df.residual(fitmodel)-
+            df.residual(fitmodel2))
[1] 0.5730203
> #pour la variable frontiere
> fitmodel<-vglm(vote4~ cattxch0+cattxch1 + cattximmi + frontiere  ,data=X1,link="logit",
+                family=cumulative(parallel=TRUE,reverse=TRUE))
> fitmodel2<-vglm(vote4~ cattxch0+cattxch1 + cattximmi + frontiere  ,data=X1,link="logit",
+                 family=cumulative(parallel=FALSE~1+frontiere,reverse=TRUE))
> 1-pchisq(deviance(fitmodel)-deviance(fitmodel2),df=df.residual(fitmodel)-
+            df.residual(fitmodel2))
[1] 1.160255e-08
> fitmodelbis<-vglm(vote4~   catximmi0+ catximmi1   ,data=X1,link="logit", family=cumulative(parallel=TRU
reverse=TRUE))
> fitmodel2bis<-vglm(vote4~ catximmi0+ catximmi1  ,data=X1,link="logit", family=cumulative(parallel=FALSE
everse=TRUE))
> 1-pchisq(deviance(fitmodelbis)-deviance(fitmodel2bis),df=df.residual(fitmodelbis)- df.residual(fitmodel
is))
[1] 0.6909172
> fit0 <- vglm(vote4~ 1, data=X1, link="logit", family = cumulative(parallel=TRUE, reverse=TRUE))
> print(pseudo_R2 <- 1 - deviance(fitmodelbis) / deviance(fit0))
[1] 0.002588579
```

Figure 17 : Vérification de l'égalité des pentes de catximmi1 et 0

```
Call:
vglm(formula = vote4 ~ catximmi0 + catximmi1, family = cumulative(parallel = FALSE,
    reverse = TRUE), data = X1, link = "logit")

Pearson residuals:
                      Min      1Q  Median     3Q    Max
logitlink(P[Y>=2]) -2.217  0.2608  0.2608 0.6718 0.7777
logitlink(P[Y>=3]) -1.182 -1.0794 -0.3982 1.0084 1.1404

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  1.627825   0.111056  14.658   <2e-16 ***
(Intercept):2 -0.057545   0.082303  -0.699   0.4844
catximmi0:1   -0.002026   0.157070  -0.013   0.9897
catximmi0:2   -0.023855   0.116468  -0.205   0.8377
catximmi1:1   -0.364133   0.148994  -2.444   0.0145 *
catximmi1:2   -0.270797   0.117208  -2.310   0.0209 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])

Residual deviance: 3656.904 on 3536 degrees of freedom

Log-likelihood: -1828.452 on 3536 degrees of freedom

Number of Fisher scoring iterations: 3

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
catximmi0:1 catximmi0:2 catximmi1:1 catximmi1:2
  0.9979757   0.9764269   0.6947991   0.7627716
```

Figure 20 : Test d'homoscédasticité des erreurs

```
Ordered Logit Regression
Log-Likelihood: -1828.822
No. Iterations: 4
McFadden's R2: 0.002588579
AIC: 3665.644
           Estimate Std. error t value Pr(>|t|)
catximmi0 -0.018135   0.109765 -0.1652 0.868774
catximmi1 -0.300372   0.109441 -2.7446 0.006059 **
----- Threshold Parameters -----
                 Estimate Std. error  t value Pr(>|t|)
Threshold (1->2) -1.607701   0.089363 -17.9906   <2e-16 ***
Threshold (2->3)  0.050699   0.079146   0.6406   0.5218
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Heteroskedastic Ordered Logit Regression
Log-Likelihood: -1828.452
No. Iterations: 6
McFadden's R2: 0.002790227
AIC: 3668.904
----- Mean Equation ------
           Estimate Std. error t value Pr(>|t|)
catximmi0 -0.022815   0.113593 -0.2008  0.84082
catximmi1 -0.290046   0.113316 -2.5596  0.01048 *
----- SD Equation ------
           Estimate Std. error t value Pr(>|t|)
catximmi0 -0.012869   0.088567 -0.1453   0.8845
catximmi1  0.056973   0.087438  0.6516   0.5147
----- Threshold Parameters -----
                 Estimate Std. error  t value Pr(>|t|)
Threshold (1->2) -1.627824   0.111057 -14.6576   <2e-16 ***
Threshold (2->3)  0.057545   0.082303   0.6992   0.4844
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```