Devoir : L'analyse des corrélations et la régression linéaire, le cas des résultats électoraux à l'élection présidentielle de 2007 dans le Val-de-Marne (94)

TOUREILLE Etienne (FS 2007 CIST – UMR Géographie Cités)

Pour citer cette feuille : Toureille E., 2019, « l'analyse des corrélations et la régression linéaire, le cas des résultats électoraux à l'élection présidentielle de 2007 dans le Val-de-Marne (94) », *Feuilles de géographie*, 2019-1, 11 p.

Type de Feuille	Exemple : Feuille d'évaluation
Niveau	L2-3
Durée	Exemple : 1 séance de 2h
Objectifs	L'objectif de cette feuille est d'évaluer l'acquisition des connaissances relatives à l'analyse des corrélations et de la régression linéaire. Après un retour sur quelques connaissances en statistiques descriptives, cette feuille a pour objectif de vérifier tant des connaissances théoriques (principe de la modélisation statistique, différence entre corrélation et causalité) que la capacité de l'étudiant.e à lire, analyser et interpréter différents résultats (coefficient de Bravais-Pearson, coefficient de détermination, paramètres du modèle de régression linéaire, interprétation graphique de la relation et de ses résidus). L'exercice est appliqué à un cas d'étude particulier : les résultats électoraux à l'élection présidentielle de 2007 dans le département du Val-de-Marne. Au-delà d'un exercice statistique, c'est donc plus généralement le formalisme et le cadre d'analyse propre à l'analyse statistique de deux variables quantitatives qui est proposé pour répondre à une problématique et des hypothèses explicitement formulées s'inspirant de la littérature ou de problèmes de société (question du vote d'extrême droite périurbain). Le choix du Val-de-Marne est quant à lui plus anecdotique : il présentait surtout l'intérêt d'être un

CC BY NC SA

	espace d'étude francilien (région de l'Université dans laquelle cette évaluation fut proposée) comprenant une large palette de bureaux de vote localisés dans des espaces urbain et périurbains. Cet exercice prend le soin de revenir aux fondamentaux de la méthode, il est donc adapté à une évaluation intermédiaire comme à un contrôle terminal, à un exercice de remise à niveau ou une évaluation diagnostic (niveau L2 ou L3, selon le niveau). L'exercice ne nécessite pas d'ordinateur et peut être réalisé sans calculatrice. Ce dernier choix est lié à la volonté d'éviter à l'étudiant.e de gérer de front les problèmes liés à la manipulation de l'outil informatique (parfois facteur de stress supplémentaire dans un contexte d'examen) et la mise en place d'un raisonnement statistique appliqué à un thème relié à des problématiques plus générales de géographie humaine (ici la géographie politique). Le choix d'un examen sur document papier permet également de proposer une plage de temps sanctuarisée pour traiter de manière approfondie de questions statistiques sans prendre le risque d'en perdre du fait de difficultés techniques (manipulation logiciel, problème lié au parc informatique, par exemple). Un court retour sur expérience, présentant le contexte d'application de l'exercice et sa réception par les étudiant.e.s est disponible à la fin de cette feuille, afin de proposer quelques propositions d'adaptations possibles aux utilisateurs.
Mots-clés	Analyse des corrélations, régression linéaire, statistique bivariée, géographie électorale, France, Ile-de-France.
Remarques sur la réception auprès des étudiants (optionnel)	Voir RETOUR SUR EXPERIENCE

CC BY NC SA

NOM : …… Prénom : …….

Séance terminale : devoir sur table

Durée : 2h. L'usage de l'ordinateur n'est pas autorisé. Le barème est indicatif. Les réponses aux questions ouvertes doivent être ***rédigées.*** *Des points pourront être retirés si le soin et la qualité de la rédaction (grammaire, orthographe, etc.) laissent à désirer.*

Exercice 1 : Le clivage gauche-droite dans le Val-de-Marne lors des élections présidentielles de 2007 *(7 points)*

L'élection de 2007 marque un tournant dans l'histoire de l'élection présidentielle française, avec un haut score d'un candidat centriste (F. Bayrou pour le MODEM – 18,57% des voix) venant talonner les deux finalistes (S. Royal pour le PS – 25,87% et N. Sarkozy pour l'UMP – 31,18%). Bien qu'originaire du centre droit (UDF), il s'afficha durant la campagne comme une alternative au « clivage gauche droite » et refusa de donner une consigne de vote en faveur de l'un ou l'autre des vainqueurs du premier tour. Dans un contexte historique de bipolarisation de la vie politique française depuis le passage de l'élection présidentielle au suffrage universel direct (1962), cette élection remet en question l'idée d'une dualité de l'échiquier politique, centré d'un côté sur les partis de gauche historiques – SFIO puis Parti Socialiste, PCF et PRG et ceux de la droite gaulliste (UDR, RPR, UMP) ou centriste (UDF). Pourtant, à la lueur des profils des bureaux de vote dans un territoire localisé – ici le Val-de-Marne – le « clivage gauche-droite » a-t-il vraiment cessé d'être pertinent pour l'analyse les comportements électoraux ?

Pour répondre à cette question, on dispose d'une base de données électorale (données du Ministère de l'Intérieur compilées par l'ANR CARTELEC[1]), qui regroupe les résultats aux élections collectées au niveau des bureaux de vote de l'ensemble des communes du Val-de-Marne (94), un département qui a l'avantage de présenter une grande diversité de types d'espaces (urbains et périurbains) à proximité de Paris. On dispose des données pour 748 bureaux de votes pour les résultats au premier tour de l'élection. Dans certains cas, on a procédé à des agrégations (par exemple entre les différents partis d'extrême gauche). Les variables se présentent sous la forme de variables quantitatives de taux exprimées en pourcentages (voir **Tab. 1**). Elles sont présentées dans le tableau **Tab. 2**.

Répondez aux questions suivantes pour évaluer la pertinence du clivage gauche-droite dans le Val-de-Marne.

Tab. 1 : Présentation du tableau de données (Extrait des trois premières lignes)

Code	COM	COM_NOM	BV	ExtGau	Ecolo	Socia	MODEM	UMPDD	ExDro
94002_001	94002	Alfortville	01	5,9	2,9	30,2	24,3	31,4	5,2
94002_002	94002	Alfortville	02	6,7	2,7	26,3	24,1	32,7	7,6
94002_003	94002	Alfortville	03	8,6	3,4	31,2	18,1	30,5	8,1

Tab. 2 : Présentation des variables du tableau

Intitulé	Description
Code	Code bureau de vote
COM	Code commune (code INSEE)
COM_NOM	Nom complet de la commune
BV	Numéro du bureau de vote
ExtGau	% de votes pour l'extrême Gauche (Besancenot, Laguiller, Buffet et Chivardi)
Ecolo	% de votes écologistes (Voynet et Bové)
Socia	% de votes socialistes (Royal)
MODEM	% de votes MODEM (Bayrou)
UMPDD	% de votes UMP (aujourd'hui Les Républicains) et Divers Droite (Sarkozy, Nihous, de Villiers)
ExDro	% de votes FN et Ext. Droite (Le Pen)

[1] Source : Colange, Beaugitte et Freire-Diaz, 2013, Base de données socio-électorales Cartelec (2007-2010). Disponible sur le site de HALshs [https://halshs.archives-ouvertes.fr/halshs-00839899/document], dernière consultation le 23.04.2019.

1) On a calculé différents paramètres univariés (de position et de dispersion) que l'on a retranscrits dans le **Tab. 3**. *(3,5 points)*

Tab.3 : Résumé statistique du tableau de données

	ExtGau	Ecolo	Socia	MODEM	UMPDD	ExDro
Min	0,8	0,5	11,9	9,5	15,1	2,4
Moyenne	8,0	2,7	29,5	19,5	32,9	7,4
Médiane	7,2	2,6	29,1	19,7	32,0	7,2
Ecart-type	4,5	0,8	7,5	3,7	8,9	2,2
Max	30,2	5,6	53,8	29,2	63,3	16,4
C.V.	0,56	0,29	0,26	0,19	0,27	0,30

a. Quelle est l'orientation dominante du vote dans le Val-de-Marne ? *(0,5 point)*

……
……

b. Quel parti a enregistré le plus petit score ? Le plus grand score ? *(0,5 point)*

……
……

c. Définissez ce qu'est un coefficient de variation (noté C.V. dans le **Tab. 3**). *(0,5 point)*

……
……
……

d. Interprétez ces coefficients pour l'ensemble des votes représentés. Quel est le candidat qui obtient les résultats les plus inégaux dans le Val-de-Marne ? *(1 point)*

……
……
……
……
……
……

e. Au regard de ce résumé statistique, quel est probablement le profil de distribution des variables observées ? Ce profil est-il adapté pour chercher l'existence de corrélations linéaires entre variables ? *(1 point)*

……
……
……
……
……
……

CC BY NC SA

2) Le tableau suivant (**Tab. 4**) est ce que l'on appelle une matrice des corrélations. Il présente les coefficients de corrélations (ici de Bravais-Pearson) entre chaque couple de variables. A partir de ce tableau, répondez aux questions suivantes sur les liens entre types de vote. *(3,5 points)*

Tab. 4 : Matrice des corrélations des pourcentages de votes à l'élection présidentielle de 2007

	ExtGau	Ecolo	Socia	MODEM	UMPD	ExDro
ExtGau	*1,00*	0,15	0,54	-0,70	-0,76	0,33
Ecolo	0,15	*1,00*	0,09	0,05	-0,24	-0,11
Socia	0,54	0,09	*1,00*	-0,60	-0,89	0,05
MODEM	-0,70	0,05	-0,60	*1,00*	0,55	-0,43
UMPD	-0,76	-0,24	-0,89	0,55	*1,00*	-0,27
ExDro	0,33	-0,11	0,05	-0,43	-0,27	*1,00*

a. Qu'est-ce que le coefficient de corrélation de Bravais-Pearson ? Donner les principes de son calcul (si possible noter la formule) ? En termes statistiques, que permet-t-il d'analyser ? *(1 point)*

……
……
……
……

b. Interprétez le coefficient de Bravais-Pearson entre le vote UMP-divers droite (UMPDD) et le vote d'extrême gauche (ExtGau). *(1 point)*

……
……
……
……
……
……

c. Dans le tableau, entourez les coefficients qui identifient des corrélations fortes. *(0,5 point)*

d. A partir de votre lecture du **Tab. 3**, peut-on valider l'hypothèse selon laquelle il existe un clivage droite-gauche lors de l'élection présidentielle de 2007 ? *(1 point)*

……
……
……
……
……
……
……
……
……
……

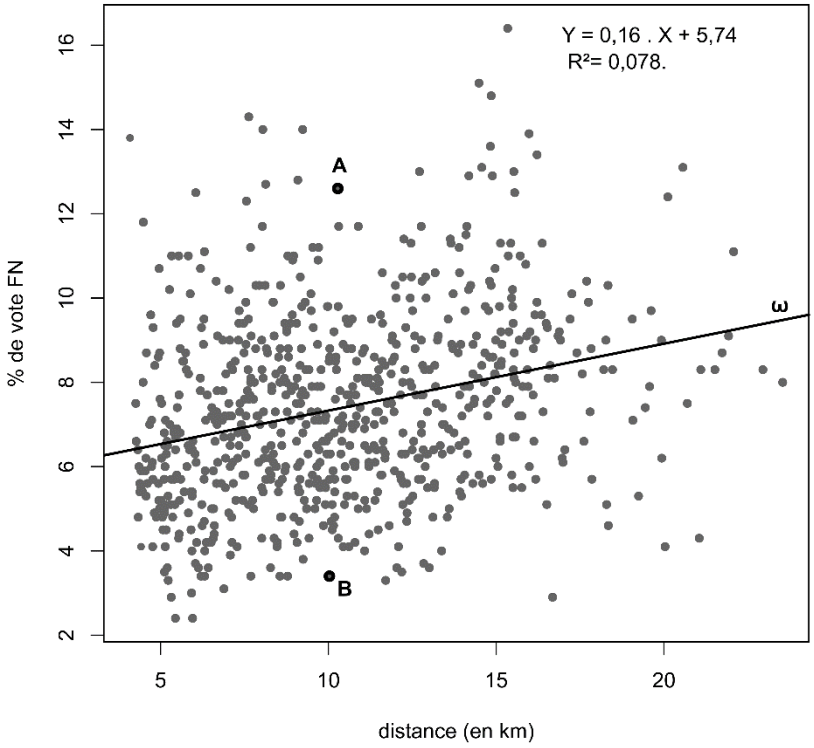
BY NC SA

Exercice 2 : Vote d'extrême droite et distance au centre *(13 points*)

Depuis la fin des années 2010, certains géographes se sont fait connaître par leurs analyses sur le lien supposé entre le vote Front National et le fait de résider en milieu périurbain ou dans les espaces ruraux. Ces thèses furent l'objet d'intenses polémiques dans les milieux de la géographie politique et de la géographie urbaine, parfois relayées par les médias nationaux (par les tribunes d'E. Charmes et de J. Lévy[2], dans le quotidien *Libération*).

Sans entrer dans le détail de la polémique, on se propose de vérifier l'existence d'une relation entre le taux de vote Front National (ExDro) et une nouvelle variable (Dist), qui figure la distance euclidienne (« à vol d'oiseau ») de chaque bureau de vote au parvis de Notre Dame de Paris, exprimée en nombre de kilomètres. Les résultats de l'analyse sont retranscrits sur la figure suivante (Fig. 1).

Fig. 1 : Relation entre la distance des bureaux de vote du Val-de-Marne au centre de Paris et le pourcentage de votes en faveur du Front National lors de l'élection présidentielle de 2007



1) En utilisant le vocabulaire adéquat, décrivez le nuage de point représenté **Fig. 1**. *(1,5 point)*

……
……
……
……
……
……
……

[2] E. Charmes, et *al.*, (2013), « Le périurbain, France du repli », *La vie des idées*, [ISSN : 2105-3030, URL: http://www.laviedesidees.fr/Le-periurbain-France-du-repli.html], dernière consultation le 11.06.2018 ; ainsi que les différentes tribunes parues dans le quotidien *Libération* auxquelles cet article fait référence.

2) Complétez l'équation suivante avec les termes « **Dist »**, représentant la distance (en km) et le terme « **ExDro** » pour qu'elle corresponde à l'hypothèse d'analyse proposée. *(0,5 point)*

……. = 0,16 x ……. + 5,74

3) Sur l'équation précédente : **entourez** le nombre figurant la *pente* et **soulignez** le nombre figurant la *constante*. *(0,5 point)*

4) Que représente la droite notée **ω** dans la **Fig. 1** ? *(0,5 point)*

……
……

5) A partir de ces différents éléments (pente, constante et R^2), interprétez l'efficacité de cette régression linéaire. *(1 point)*

……
……
……
……
……
……

6) Le résultat du coefficient de Bravais est de R = 0,279. A l'aide de la table de Bravais-Pearson disponible en Annexe de ce devoir, interprétez la relation entre les deux variables pour un risque d'erreur α d'une valeur usuelle en sciences sociales. Ecrire l'opération effectuée. Ce résultat est-il en adéquation avec les résultats précédents ? Justifiez. *(1,5 points)*

……
……
……
……
……
……
……
……

7) A partir de cette analyse, peut-on conclure à l'existence d'une relation entre la distance et le vote front national ? Pourquoi ? *(1,5 points)*

……
……
……
……
……
……
……
……

CC BY NC SA

……
……
……
……

8) On a identifié les bureaux de vote **A** et **B** sur le graphique. Quelles devraient être les valeurs théoriques du vote Front National à l'élection présidentielle dans ces bureaux de vote si cela devait déprendre uniquement de la distance (selon le modèle proposé) ? *(1 point)*

……
……
……

9) Comment appelle-t-on la mesure de l'écart entre cette valeur théorique et celle observée sur le graphique ? Qu'est-ce qu'ils permettent d'observer ? *(1 point)*

……
……
……
……
……
……

10) A partir de vos connaissances générales en géographie urbaine et en géographie politique, proposez des éléments d'explications de ces écarts. *(1,5 point)*

……
……
……
……
……
……
……

11) Quelles autres variables disponibles dans la statistique publique et d'usage courant pourraient être pertinentes à analyser pour étudier le vote Front National ? Proposez une hypothèse de relation précise, en identifiant clairement les variables utilisées. *(1,5 point)*

……
……
……
……
……
……
……
……
……
……
……

BY NC SA

12) <u>En général</u>, peut-on conclure à l'existence d'une relation de causalité à partir de l'observation d'une corrélation ou d'une régression linéaire ? Quelles sont les limites posées par cette méthode ? *(1 point)*

……
……
……
……
……
……
……
……
……
……
……

Annexe : Table du R de Bravais-Pearson

	α		
v	**0,1000**	**0,0500**	**0,0200**
1	0,9877	0,9969	0,9995
2	0,9000	0,9500	0,9800
3	0,8054	0,8183	0,9343
4	0,1293	0,8149	0,8822
5	0,6694	0,8115	0,8329
6	0,6215	0,7545	0,7887
7	0,5822	0,7067	0,7498
8	0,5494	0,6664	0,7155
9	0,5214	0,6319	0,6851
10	0,4973	0,6021	0,6581
20	0,3598	0,4227	0,4921
50	0,2306	0,2732	0,3218
100	0,1638	1,1946	0,2301

cc BY NC SA

RETOUR SUR EXPERIENCE :

Retour réflexif sur la réception de l'exercice par les étudiants, le contexte dans lequel il a été proposé et suggestions pour son adaptation éventuelle.

Ce sujet d'examen fut proposé en seconde session d'un cours de statistiques bivariées de niveau L3, à destination d'étudiant.e.s en parcours d'« études urbaines » à l'Université de Cergy-Pontoise. Comme souvent lors des secondes sessions –et qui plus est avec un petit effectif en TD (une vingtaine d'étudiant.e.s)– peu d'étudiant.e.s (3 exactement) furent présent.e.s le jour de l'examen, ce qui rend difficile l'appréciation de la difficulté de l'exercice. Fait important : le TD faisait également office de CM, prenant ainsi la forme d'un cours intégré de 18 heures (ou 12 séances d'1h30) associant la dimension théorique des méthodes bivariées à leur application, notamment sur ordinateur (via l'utilisation des options de visualisation et de calcul d'un tableur). Compte tenu des lacunes rencontrées lors des premières séances, les premiers cours et exercices ont consistés en un certain nombre de rappel de statistiques univariées (définition des centres, des paramètres de dispersion, des types de variables et analyses graphiques).

On peut cependant comparer sa réception à un devoir similaire présenté en première session, qui associait à un exercice portant sur l'analyse d'un tableau de corrélation sur des variables connues des étudiant.e.s (tableau sur des données démographiques du PNUD, année 2014), un second inspiré du cas d'étude d'une autre *Feuille de Géographie*[3]. Il révéla – sans surprise – que le second exercice était beaucoup plus discriminant que le premier. L'hétérogénéité des notes, allant de 3 à 15,5 (moyenne de 8,6 et médiane de 7,2), s'est surtout manifestée à travers une opposition entre un groupe d'étudiant.e.s maîtrisant dans l'ensemble les deux exercices et ceux mis en difficulté par la régression linéaire, notamment dans son interprétation et ses finalités (apprentissage scolaire).

L'intérêt d'un premier exercice présentant des questions parfois élémentaires (voir question 2.b) réside dans sa capacité d'adaptation à l'hétérogénéité des publics et ne pas complètement décourager les étudiant.e.s rencontrant le plus de difficultés. Il va de soi que le format, le temps imparti et le barème appliqué pour l'examen doivent être adaptés au public et au contexte d'enseignement, s'il devait être reproduit pour un autre public ou dans un autre contexte.

A l'origine proposé pour une séance d'examen d'une heure et demie, les étudiant.e.s les plus avancé.e.s ont mentionné des difficultés à achever l'exercice (manque de temps pour répondre aux deux dernières questions). Ainsi, et notamment à la suite des suggestions des évaluateurs[4] de cette feuille, proposant l'ajout des questions 1 et 6 de l'Exercice 2, un format de deux heures semblerait plus adapté pour traiter de l'ensemble de l'exercice s'il devait être présenté au même public.

L'intérêt de proposer un exercice « sur papier » plutôt que sur machine (objet d'un examen intermédiaire à mi-semestre et d'un « Devoir Maison ») permet de limiter les effets cumulatifs posés par la manipulation conjointe de la méthode (appliquée ici à des éléments de géographie électorale) et de la manipulation de l'outil informatique, parfois fortement discriminant en situation d'examen – notamment en temps limité, et de se focaliser sur l'analyse réfléchie des méthodes statistiques sans prendre le risque d'être débordé par d'éventuels problèmes techniques (parc informatique défectueux ou imprévisible, difficulté à accéder aux données sur l'ENT, oubli de leurs identifiants par les étudiant.e.s, par exemple). De nouveau, à l'évaluateur de juger la place impartie à la manipulation logicielle dans ses objectifs pédagogiques, les exercices pouvant

[3] Commenges H., 2017, « Analyse de données et cartographie – Devoir sut table. Etude de la ségrégation urbaine au Cap (Afrique du Sud) », *Feuilles de géographie*, Feuille 2017-2, 6 pages [https://feuilles-de-geographie.parisnanterre.fr/2018/01/04/analyse-de-donnees-et-cartographie-devoir-sur-table-etude-de-la-segregation-urbaine-au-cap-afrique-du-sud/], dernière consultation le 23.04.2019.

[4] Que nous remercions pour leurs remarques et suggestions.

être adaptés sur machine compte tenu de l'accessibilité des données sur le site de l'ANR Cartelec[5]. Dans le cadre d'un examen terminal de cours intégré, on a ici privilégié la dimension analytique et réflexive des étudiant.e.s, en cherchant notamment à appliquer la connaissance des outils et méthodes de la statistique avec une thématique de géographie humaine et plus généralement à une question d'ordre sociétale (débat public autour de l'analyse du vote à l'élection présidentielle).

Enfin, l'analyse de la base de données de l'ANR Cartelec fut proposée en cours durant les séances précédentes dans le cadre d'exercices d'application portant, entre autre, sur l'analyse de la régression linéaire). Les étudiant.e.s connaissaient ainsi la construction, l'objectif et le contexte de réalisation de la base (présentation d'un texte introductif tiré du rapport de l'ANR[6]) facilitant l'entrée dans l'exercice ainsi que l'appréhension de la partie thématique du sujet.

L'une des principales difficultés présentées par l'exercice est sans doute liée au caractère relativement ancien des données présentées (élections de 2007). On peut, en effet, supposer que la référence à des partis et personnalités politiques parfois anciennes peuvent déconcerter les étudiant.e.s par leur exotisme. Cependant, l'élection de 2007, avec la poussée observée du Modem constitue – jusqu'à l'élection présidentielle de 2017 – un cas particulièrement original de remise en question – au moins dans les discours – dudit « clivage gauche-droite » par les dirigeants d'un parti obtenant au final près d'un cinquième des votes exprimés. Ensuite, les données de l'ANR Cartelec ont également comme originalité d'être uniformisées pour proposer le niveau d'analyse le plus fin des comportements électoraux (bureaux de vote) dans les aires métropolitaines (voir présentation de la base par les auteurs[7]). Enfin, même si elles peuvent paraître moins accrocheuses pour une population jeune, la convocation de bases de données utilisées depuis plusieurs années permet de traiter de manière plus distanciée et informée des objets potentiellement polémiques et documentés. On peut également penser que l'élection de 2007 puisse être réarticulée avec les connaissances acquises lors de leur formation en Histoire Géographie lors de leur passage au collège et au lycée (programme de terminale, notamment). Ainsi, une perspective possible pourrait être d'adapter l'exercice à d'autres élections, mais aussi à d'autres espaces d'études que le Val-de-Marne.

Dans la mesure où ce cours n'intégrait pas d'enseignements en cartographie statistique, on a volontairement laissé de côté ici la dimension cartographique. Dans le cas contraire, une carte des résidus pourrait être produite en fin de devoir pour être confrontée au conclusions portant sur l'analyse de la distance au centre.

[5] Sur le site de l'Université de Rouen [http://cartelec.univ-rouen.fr/?page_id=3609], dernière consultation le 23.04.2019.

[6] Source : Beauguitte, L., et Colange, Céline, (2003), *Analyser les comportements électoraux à l'échelle du bureau de vote, ANR Cartelec : mémoire scientifique*, p. 3 [https://halshs.archives-ouvertes.fr/halshs-00839899/document], dernière consultation le 23.04.2019).

[7] *Ibid.*

CC BY NC SA