The knowledge to compete

Sprott Working Paper Series

RESEARCH

Mythes et réalités à propos des données sur l'enrôlement des électeurs au scrutin électoral programmé pour l'année 2011 en République démocratique du Congo: une analyse empirique

Dr. Aaron L. Nsakanda
Associate Professor, Sprott School of Business, Carleton University, Ottawa (ON), Canada

Dr. Moustapha Diaby
Associate Professor, School of Business Administration, University of Connecticut, Storrs (CT), USA

SPROTT SCHOOL OF BUSINESS

Carleton UNIVERSITY

Mythes et réalités à propos des données sur l'enrôlement des électeurs au scrutin électoral programmé pour l'année 2011 en République démocratique du Congo – une analyse empirique -

Aaron L. Nsakanda, PhD, Associate Professor, Sprott School of Business, Carleton University, Ottawa (ON), Canada

Moustapha Diaby, PhD, Associate Professor, School of Business Administration, University of Connecticut, Storrs (CT), USA

Résumé

Cette étude présente les résultats d'une analyse basée sur des méthodes statistiques comparant les données sur l'enrôlement des électeurs observées dans le fichier électoral de 2006 par rapport à celles reportées dans le fichier électoral de 2011. Un modèle de régression linéaire simple a été utilisé afin d'établir l'existence ou pas d'une relation entre nos variables ainsi que la force de cette relation. L'étude recourt à des méthodes de diagnostic basées sur l'analyse graphique et des tests numériques pour identifier la présence des données atypiques dans le fichier électoral de 2011. L'analyse a permis d'identifier treize (13) circonscriptions électorales dont les données dépassent un seuil critique des résidus studentisés et de la statistique DFITs. Le nombre d'enrôlés reportés dans chacune de ces circonscriptions semblent être atypique par rapport au nombre espéré si d'aucuns considèrent une croissance démographique normale de la population.

1. Introduction

La République démocratique du Congo (RDC) est à l'aube de tenir pour la seconde fois de son histoire moderne des élections présidentielle et législatives. Que se passera-t-il dans ce pays au terme de ces élections que d'aucuns souhaitent apaisées, crédibles et inclusives? Cette question est ouverte à des supputations diverses compte tenu de l'expérience de la démocratie électorale en Afrique qui somme toute fait encore l'objet d'un bilan mitigé. Certes, des progrès significatifs ont été réalisés dans certains pays tels que le Ghana ou l'Afrique du Sud où les élections ont contribué à consolider les institutions démocratiques et améliorer les perspectives aussi bien économiques que politiques. Des élections crédibles ont aussi ouvert la voie à une stabilité politique et à la réconciliation nationale après plusieurs années de guerre civile à certains pays tels que le Liberia et la Sierra Leone. D'autres pays tels que le Mali ou le Benin ont connu une émergence de démocratie à la suite de l'organisation des élections. Cependant, il y a encore des pays qui ont connu à la suite des élections douteuses, selon l'avis de certains, la violence et/ou des conflits armés postélectoraux, dont notamment le Kenya en 2008, le Zimbabwe en 2008, le Gabon en 2009, le Togo en 2010, la Côte d'Ivoire en 2011 et le Nigeria en 2011. Le spectrum de ce qui attend la RDC à la conclusion de ses élections est donc très variable. L'une des institutions qui peut grandement contribuer à la réduction de la variabilité de ce spectrum est la Commission Électorale Nationale Indépendante (CENI).

Dans le cadre de ses attributions, et ce, à travers le Journal Officiel de la RDC, la CENI a rendu public les données sur le nombre des circonscriptions électorales, leur répartition par province et le nombre d'électeurs enrôlés par circonscription. Ces données sont critiques sur le plan économique puisqu'une bonne partie de la comptabilité électorale et les coûts afférents y sont directement reliés. Sur le plan politique, la répartition des sièges par circonscription électorale est basée sur ces données. Dans la foulée des changements récents apportés à la constitution de la RDC où l'élection présidentielle se déroulera désormais en un seul tour, la présence de quelques données aberrantes (due par exemple aux erreurs de saisie) peut favoriser certains candidats par rapport à d'autres. Le déploiement des observateurs et témoins électoraux dans tous les bureaux de votes est un défi majeur compte tenu non seulement de leur nombre élevé (63 865) et des coûts impliqués mais aussi des problèmes logistiques et sécuritaires qui accompagnent un tel déploiement. Le recours aux méthodes d'échantillonnage constitue alors une avenue attrayante à explorer pour assurer un équilibre entre d'une part, les coûts de déploiement des observateurs et témoins, et d'autre part, les bénéfices d'un tel déploiement. Les données sur l'enrôlement des électeurs sont à la base d'une telle analyse coûts-bénéfices.

C'est donc sans étonnement que la publication de ces données par la CENI a soulevé diverses analyses et conclusions.

2. Objectifs de l'étude, questions de recherche et contributions

L'objectif de cette étude est de contribuer au débat en y apportant quelques éclairages sur ce que nous pouvons partiellement apprendre des données publiées par la CENI de la RDC sur le nombre d'électeurs enrôlés par circonscription pour le scrutin électoral programmé pour l'année 2011. Contrairement aux publications existantes, nous appuyons notre étude en recourant aux méthodes statistiques. L'ASBL Aprodec a publié un rapport qui présente des tableaux comparatifs du taux d'accroissement moyen du corps électoral entre le nombre d'électeurs enrôlés lors de la tenue du scrutin électoral de 2006 et le nombre d'électeurs reportés pour le scrutin électoral programmé pour l'année 2011 (Aprodec, 2011). L'objectif visé dans ce rapport est d'identifier les circonscriptions électorales présentant des anomalies en termes du taux d'accroissement moyen sur une période de cinq années entre le corps électoral et la population générale. Dans cette étude, les questions auxquelles nous cherchons à répondre peuvent être formulées comme suit :

(a) Peut-on établir statistiquement une relation entre le nombre d'électeurs enrôlés par circonscription électorale lors de la tenue du scrutin électoral de 2006 et le nombre d'électeurs inscrits pour le scrutin électoral programmé pour l'année 2011? Cette relation peut-elle être représentée par une forme linéaire?

(b) Si une telle relation existe, est-ce qu'il y a des données rapportées pour le scrutin électoral programmé pour 2011 qui s'écartent de manière statistiquement significative de

celles qu'on aurait obtenues dans une circonscription donnée en considérant une croissance « normale » de la population?

Les résultats de cette étude peuvent servir à des fins diverses. Par exemple, la CENI peut y recourir pour cibler les circonscriptions où une étude d'audit est requise pour y déceler des erreurs éventuelles. Les diverses parties prenantes ayant des candidats aux élections peuvent y recourir pour s'accorder sur les circonscriptions nécessitant une attention particulière. Les diverses parties prenantes ayant des observateurs et témoins aux élections peuvent y recourir pour déployer stratégiquement les ressources limitées à leur disposition. Somme toute, nous espérons que cette étude propose une approche qui pourrait contribuer à assoir plus de transparence dans le processus électoral non seulement en RDC mais aussi dans plusieurs autres pays d'Afrique en proie à des violences et/ou conflits armés postélectoraux.

3. Brève description de la méthodologie

Notre approche d'analyse consiste à recourir à un modèle de régression linéaire simple que nous formulons comme suit :

$$Y2011_i = a_o + b_o \times Y2006_i + \varepsilon_i \qquad (1)$$

où

$Y2006_i$	est la variable indépendante représentant le nombre d'électeurs enrôlés dans la circonscription électorale *i* lors de la tenue du scrutin électoral de 2006.
$Y2011_i$	est la variable dépendante représentant le nombre d'électeurs enrôlés dans la circonscription électorale *i* pour le scrutin électoral programmé pour l'année 2011.
ε_i	représente l'erreur dans l'approximation de la variable dépendante $Y2011_i$.
a_o, b_o	représentent les coefficients de régression.

Le choix de ce modèle se justifie par le fait qu'il permet d'établir l'existence ou pas d'une relation entre nos variables, la forme de cette relation ainsi que sa puissance. De plus, ce modèle permet de conduire des analyses appropriées pour identifier les données atypiques, la question principale que cette étude vise à répondre. Ainsi, le recours à un modèle de régression linéaire simple est limité dans le cadre de ce travail uniquement à la détection, si elles existent, des données sur le nombre d'électeurs enrôlés par circonscription électorale *i* pour le scrutin électoral programmé pour l'année 2011 qui s'écartent de manière statistiquement significative de celles qu'on aurait obtenues dans une circonscription donnée en considérant une croissance « normale » de la population. Il s'agit ainsi des données atypiques par rapport à la variable dépendante. Pour les identifier, nous recourons aux approches d'analyse graphique et aux tests numériques dont l'examen des résidus studentisés qui sont un type de résidus standardisés pour

la détermination des données qui sortent de l'ordinaire et la statistique DFITs qui est une mesure qui permet de déterminer si les données atypiques identifiées ont aussi une certaine influence sur les valeurs estimées à partir de l'équation de régression.

Notre approche méthodologique se situe à deux niveaux d'analyse. Nous procédons au premier niveau à la vérification de l'adéquation entre nos données et le modèle de régression linéaire simple ainsi que à la vérification des hypothèses de ce modèle. Nous procédons aussi à ce niveau à l'identification des données atypiques. Dans tous les cas, nous recourons à diverses techniques informelles (e.g., analyse graphique) et formelles (e.g., tests numériques). L'une des principales causes de violation des hypothèses d'un modèle de régression linéaire étant la présence des données atypiques, le deuxième niveau d'analyse vise à vérifier le rapprochement de notre modèle à ces hypothèses lorsque les données extrêmes identifiées avec le premier niveau d'analyse sont ignorées. Pour ce faire, nous répétons les analyses similaires à celles effectuées avec le premier niveau d'analyse mais en ne considérant pas les données atypiques.

4. Résultats et Discussion

Nous rapportons dans cette section les résultats de nos analyses. Les données ayant servi à cette analyse proviennent des rapports officiels de la RDC (Journal Officiel de la RDC, 2006 et 2011). Ces données sont constituées des listes de personnes enrôlées pour les élections présidentielles et législatives de 2006 et de 2011 pour toute l'étendue du territoire national de la RDC. Une paire appariée du t-test a été réalisée afin de déterminer si la différence était significative. Les résultats obtenus montrent que la différence est significative au niveau de signification de 95%. Par conséquent, nous rejetons l'hypothèse nulle et concluons qu'il existe une différence significative entre les données rapportées durant les deux périodes d'observations et que le nombre d'enrôlés de 2011 est significativement plus élevé que celui observé en 2006. Les résultats du test apparié de Student sont présentés dans le tableau 1.

Tableau 1. Échantillon apparié de test de Student

	Différences appariées				Taille de l'échantillon
	Moyenne	Écart-type	Intervalle de confiance de la différence à 95%		
			Limite inférieure	Limite supérieure	
Periode 2011 – Période 2006	37442	32292.44	32523.40	42360.34	169

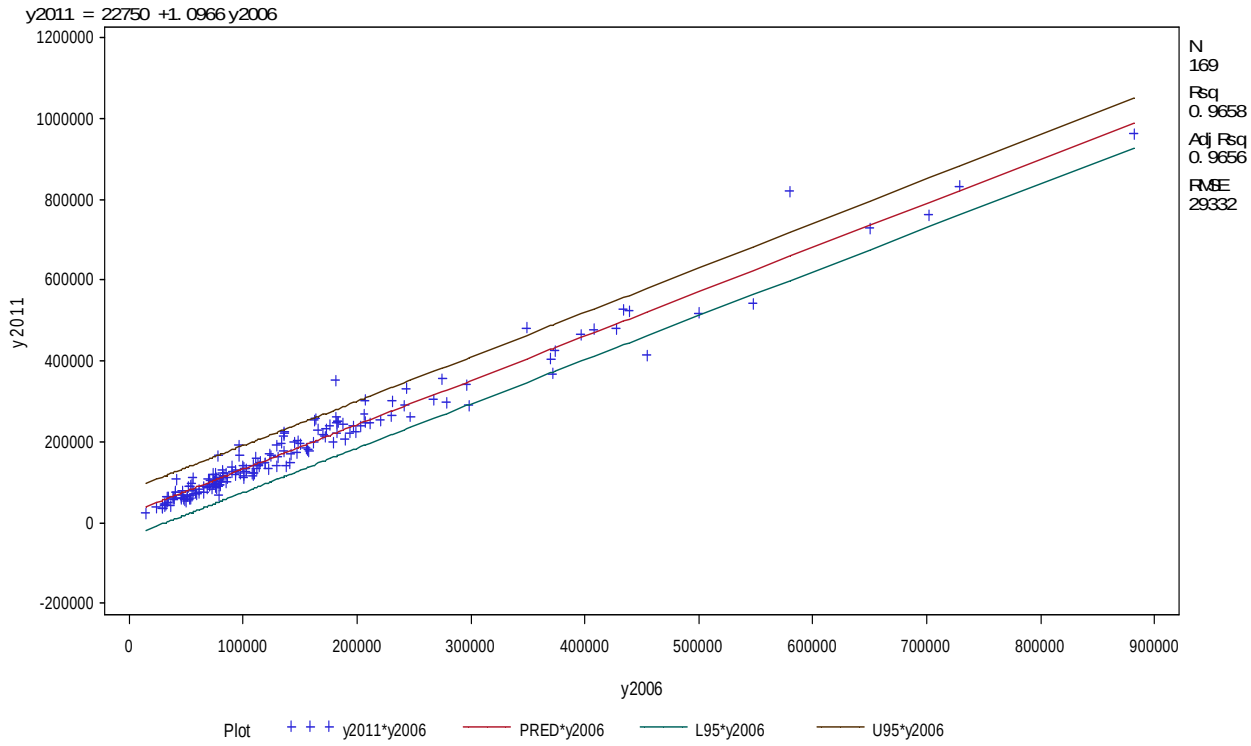
Les résultats sommaires de l'analyse de régression obtenus avec la procédure REG du logiciel SAS sont présentés dans le tableau 2.

Tableau 2 : Résultats de l'analyse de régression

Variables	Modèle	
Constante	Paramètre estimé	22750
	Erreur standard	3314
	t	6.86
	Prob > t	0.0001
Nombre d'électeurs enrôlés pour le scrutin électoral de 2006 (y2006)	Paramètre estimé	1.096
	Erreur standard	0.01597
	t	68.68
	Prob > t	0.0001
F		4717.34
Prob > F		0.0001
R^2		0.9658
R^2 Ajusté		0.9656
Taille de l'échantillon		169

Les résultats obtenus montrent que le modèle de régression linéaire s'ajuste correctement et qu'il est globalement significatif au seuil de signification de 99% (α=0.01). Par conséquent, nous rejetons l'hypothèse nulle et concluons qu'il semble exister une forte adéquation entre le nombre d'électeurs enrôlés par circonscription électorale lors de la tenue du scrutin électoral de 2006 et le nombre d'électeurs inscrits pour le scrutin électoral programmé pour l'année 2011. De plus, le modèle a un pouvoir explicatif très élevé puisque l'examen du coefficient de régression montre que 96.5% de la variation des données sur le nombre d'électeurs rapportés pour le scrutin électoral programmé pour l'année 2011 pourraient être expliqués par le modèle.

La droite de régression linéaire à 95% d'intervalle de confiance est illustrée dans le graphique 1. La projection dans ce graphique des données réelles transcrites pour l'année 2006 et l'année 2011 indique clairement la présence de plusieurs données à l'extérieur ou sur les limites inférieures et supérieures de l'intervalle de confiance, illustrant ainsi la présence des observations extrêmes (i.e des données atypiques) dans notre échantillon.



Graphique 1: Droite de régression linéaire à 95% d'intervalle de confiance

Le tableau 3 présente les résultats du test de White sur l'homoscédasticité des résidus. Les résultats montrent que la variance des résidus est juste à la limite d'être homogène au seuil de signification de 99% (α=0.01).

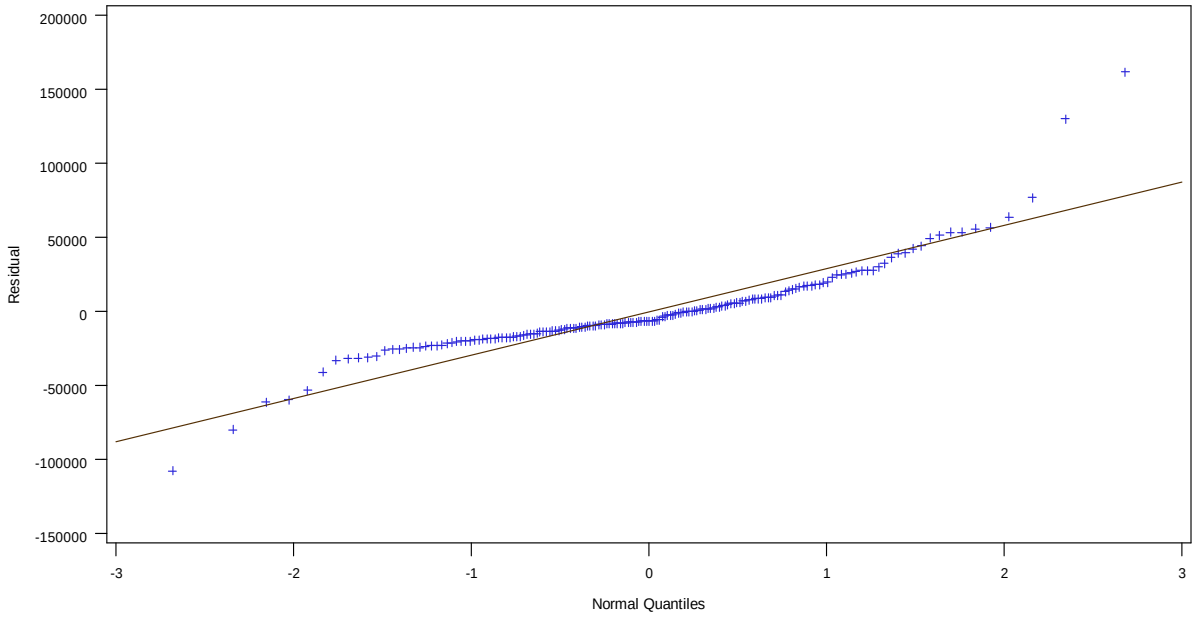
Tableau 3. Résultat du test de White sur l'erreur de variance non constante

dl	Chi-Carré	Pr > Chi-Carré
2	9.30	0.0096

Le tableau 4 présente les résultats du test de Shapiro-Wilk sur l'hypothèse de normalité. La valeur de *p* du test de Shapiro-Wilk est très faible (0.0001), indiquant par conséquent le rejet de l'hypothèse selon laquelle les erreurs sont normalement distribuées. Nous montrons dans le graphique 2le tracé "Normal QQ-Plot" des résidus qui confirme graphiquement les résultats de Shapiro-Wilk sur l'inadéquation à la loi normale des erreurs. De plus, tel que préalablement illustré avec le graphique 1, ce tracé confirme aussi l'existence des données atypiques dans notre échantillon dont la présence contribue à la violation de l'hypothèse de normalité, une hypothèse nécessaire pour effectuer les *t*-tests sur les coefficients.

Tableau 4: Résultats du test de Shapiro-Wilk

Test	Statistique		Valeur de *p*	
Shapiro-Wilk	W	0.862864	Pr < W	<0.0001



Graphique 2 : Tracé "Normal QQ-Plot"

L'objectif de ce travail étant essentiellement l'identification des données atypiques, les analyses qui précèdent permettent de confirmer leur présence. Afin de les identifier, nous procédons à l'analyse des résidus studentisés. La règle de décision générale consiste à accorder une attention particulière aux observations dont les résidus studentisés sont supérieurs à +2 ou -2. Le tableau 5 présente la liste des circonscriptions dont le nombre d'enrôlés en 2011 présente une différence nette significative par rapport au nombre espéré en considérant une croissance normale de la population. Ces circonscriptions présentent des données atypiques exigeant une attention particulière.

Le tableau 6 présente les circonscriptions dont la statistique DFITs dépasse un seuil critique. Cette statistique permet de détecter des données non seulement atypiques mais qui exercent aussi une grande influence (i.e., un grand effet) sur le modèle de régression. Elle indique pour une observation donnée, le changement normalisé dans la valeur prédite, avec et sans ladite observation. La règle de décision générale consiste à accorder une attention particulière aux observations dont la statistique DFITs est supérieure à $2\times\sqrt{k/n}$ où k est le nombre de variables dépendantes i.e., k=1) and n est la taille de l'échantillon (n=169). Ce tableau révèle que toutes les observations (circonscriptions) identifiées au Tableau 5 ont aussi une influence sur le modèle

de régression. De plus, la statistique DFITs a permis d'identifier cinq (5) observations additionnelles atypiques (Beni, Feshi, Kinshasa-2, Kinshasa-4, et Kamonia). Elles n'étaient pas identifiées avec les résidus studentisés, puisqu'il s'agit essentiellement des observations atypiques par rapport à la variable dépendante.

Tableau 5 : Liste des circonscriptions dont les résidus studentisés dépassent le seuil critique

Circonscription	Valeur du résidu studentisé	Nombre d'électeurs enrôlés	
		Année 2006 (y2006)	Année 2011 (y2011)
Bulungu	-3.87582	454198	413455
Mbuji-Mayi-ville	-2.84436	547461	543557
Masimanimba	-2.12221	371544	369213
Kananga-ville	-2.04734	298045	290471
Dimbelenge	2.21605	96578	192693
Idiofa	2.69655	348938	482375
Goma-ville	4.72731	180955	351353
Lubumbashi-ville	6.33443	579941	820857

Tableau 6 : Liste des circonscriptions dont la statistique DFITs dépasse le seuil critique

Circonscription	Statistique DFITs	Nombre d'électeurs enrôlés	
		Année 2006 (y2006)	Année 2011 (y2011)
Bulungu	-0.71562	454198	413455
Mbuji-Mayi-ville	-0.66781	547461	543557
Kinshasa-4	-0.42990	882463	964442
Beni	-0.38657	500357	518735
Kinshasa-2	-0.36552	702023	761256
Masimanimba	-0.30465	371544	369213
Kananga-ville	-0.22783	298045	290471
Feshi	0.17104	78070	165226
Kamonia	0.18384	433823	528981
Dimbelenge	0.18376	96578	192693
Idiofa	0.35887	348938	482375
Goma-ville	0.37232	180955	351353
Lubumbashi-ville	1.60273	579941	820857

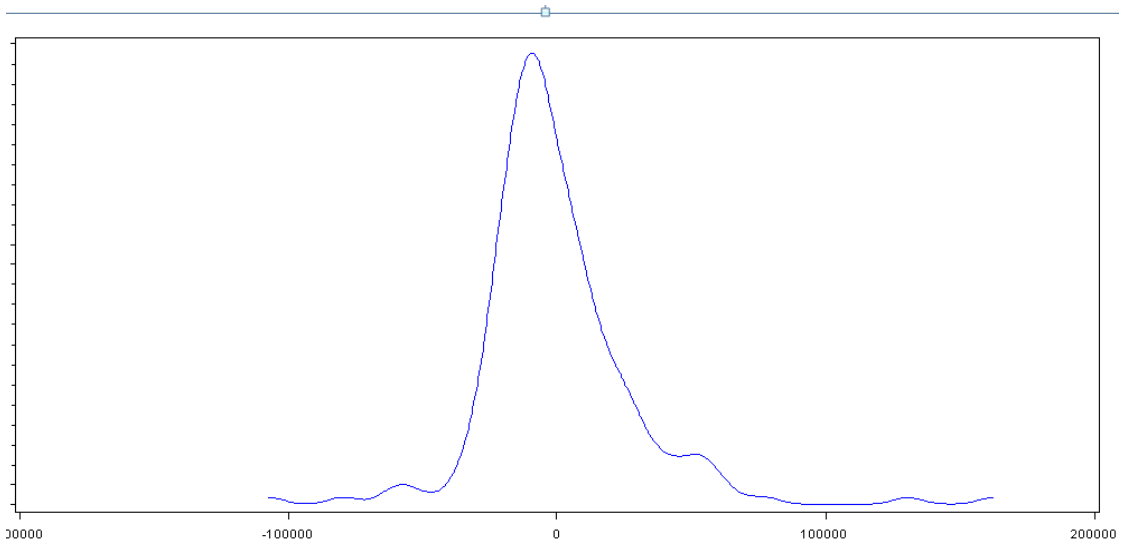
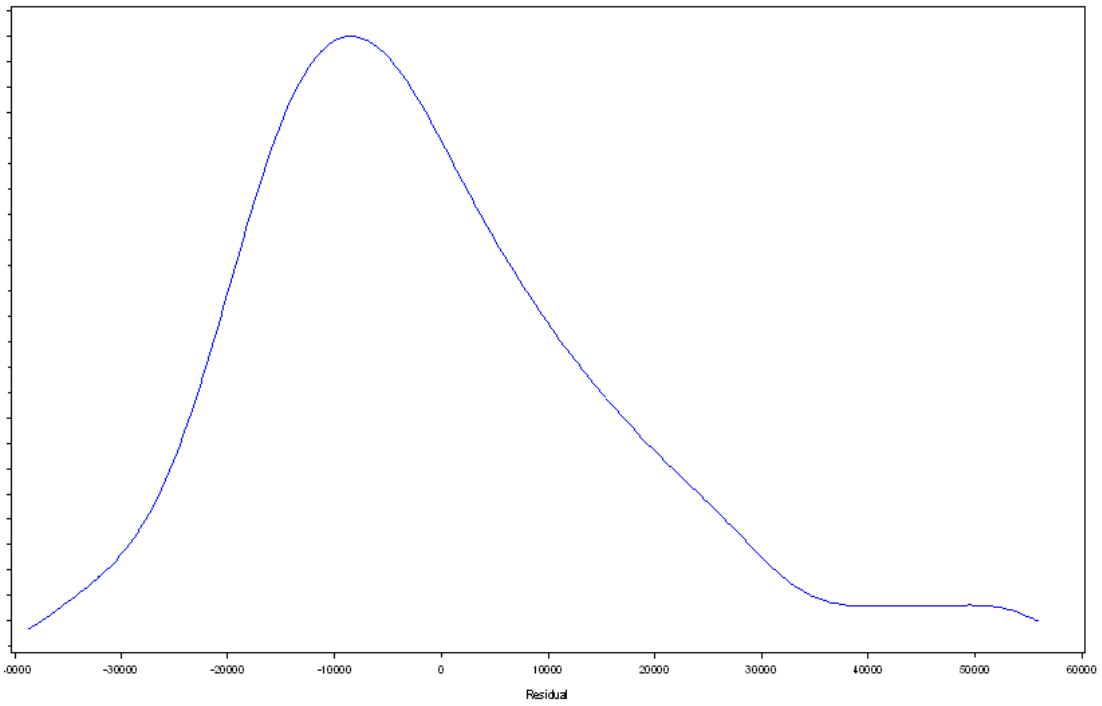
Finalement, afin de tester l'impact de la présence de ces données atypiques sur les hypothèses du modèle de régression, nous présentons aux tableaux 7 et 8 les résultats d'analyse avec un échantillon réduit dans lequel ces données n'ont pas été prises en compte. Ces tableaux révèlent une forte amélioration dans les statistiques observées lorsque les données atypiques sont ignorées

dans notre analyse. Nous observons une réduction du coefficient de dissymétrie ainsi que du coefficient d'aplatissement (kurtosis). Le coefficient de corrélation (R^2) s'est apprécié stipulant ainsi que le pouvoir explicatif du modèle a augmenté (de 96.5% à 97.5%). Le modèle de régression linéaire est demeuré globalement significatif au seuil de signification de 99% (α=0.01) ainsi que les coefficients de régression. Les changements significatifs dans nos conclusions se situent au niveau des hypothèses du modèle de régression. À la suite de l'identification des données atypiques et leur élimination de notre échantillon, le test de White sur l'homoscédasticité des résidus que la variance des résidus est homogène au seuil de signification de 99% (α=0.01). L'analyse graphique des résidus montre qu'ils se rapprochent de plus en plus de la forme gaussienne.

Tableau 7 : comparaison des résultats du modèle de régression avec ou sans les données atypiques

Variables	Échantillon complet (i.e., comprenant les données atypiques)	Échantillon tronqué (i.e., ne comprenant pas les données atypiques)
Coefficient de dissymétrie (skewness)	1.4196	0.8631
Coefficient d'aplatissement (Kurtosis)	8.0981	0.7276
Pr > Chi-Carré (test d'hétéroscédasticité)	0.0096	0.019
F	4717.34	6038
Prob > F	0.0001	0.0001
R^2	0.9658	0.9751
R^2 Ajusté	0.9656	0.9750
Taille de l'échantillon (*n*)	169	156

Tableau 8: Forme gaussienne des résidus du modèle de régression avec ou sans les données atypiques

Échantillon complet (comprenant les données atypiques)	
Échantillon tronqué (ne comprenant pas les données atypiques)	

5. Conclusions

Nous avons présenté dans cette étude les résultats d'une analyse basée sur des méthodes statistiques des données sur l'enrôlement des électeurs observés dans le fichier électoral de 2006 par rapport à celles rapportées dans le fichier électoral de 2011. Un modèle de régression linéaire simple a été utilisé pour établir statistiquement la relation entre le nombre d'électeurs enrôlés par circonscription électorale lors de la tenue du scrutin électoral de 2006 et le nombre d'électeurs enrôlés pour le scrutin électoral programmé pour l'année 2011. L'objectif de l'étude est d'identifier des données rapportées pour le scrutin électoral programmé pour 2011 qui s'écartent de manière statistiquement significative de celles qu'on aurait obtenues dans une circonscription donnée en considérant une croissance « normale » de la population. À cet effet, une approche méthodologique à deux niveaux d'analyse a été utilisée pour détecter la présence des données atypiques. L'analyse graphique ainsi que des tests numériques (e.g., des résidus studentisés, statistique DFITs) ont permis de détecter la présence des données atypiques ainsi que leur identification.

Il est de notre avis que le recours à des analyses statistiques constitue une avenue prometteuse pour apporter un éclairage sur comment les informations contenues dans des fichiers électoraux pourraient être explorées et exploitées davantage aussi bien sur le plan aussi économique que politique pour assurer une plus grande transparence du processus électoral. Cette étude est un pas dans cette direction et nous espérons qu'elle pourrait contribuer non seulement à apporter des réponses à certaines questions mais aussi à formuler ou soulever des questions pour une quête des réponses.

Bibliographie

1. Aprodec (2011) - Allocation prononcée en date du 20 octobre devant le Parlement Européen, 19 pages (accéder en ligne sur le lien suivant http://static.blog4ever.com/2011/02/467504/artfichier_467504_292763_201110241213387.pdf, 11 novembre 2011).

2. Journal officiel de la RDC (2006) – Loi no 06/006 portant organisation des élections présidentielle, législatives, provinciales, urbaines, municipales et locales, 31 pages (accéder en ligne sur le lien suivant : http://www.leganet.cd/Legislation/JO/2006/JO.10.03.2006.pdf).

3. Journal officiel de la RDC (2011) – Loi no 11/014 du 17 août 2011 portant répartition des sièges par circonscription électorale pour les élections legislatives et provinciales, 18 pages (accéder en ligne sur le lien suivant : http://www.leganet.cd/Legislation/Droit%20Public/Divers/Loi.11.014.17.08.2011.htm).