Reports de votes entre les deux tours d'une élection présidentielle

Estimation statistique et sociologie électorale

Yannick FENDRICH, Emmanuel L'HOUR, Guillaume RATEAU

avec l'encadrement de Nicolas CHOPIN (Ensae-Crest) et Robin RYDER (Ceremade-Crest)

mai 2014

Groupe de travail ENSAE 3A

1 Introduction

1.1 Élections par scrutins à deux tours

L'élection du Président de la République française constitue un exemple important de mode de scrutin à deux tours. Dans le cadre de ce type de mode de scrutin, l'intégralité des candidats [1] s'affrontent et sollicitent le vote des électeurs au cours d'un premier tour. Si aucun des candidats n'a recueilli une majorité absolue des suffrages exprimés à l'issue du vote, un second tour de scrutin est alors organisé. Les règles de qualification pour le second tour diffèrent sensiblement d'un pays et d'une élection à l'autre. Elles peuvent néanmoins être pour la plupart regroupées dans deux grandes catégories. Une première possibilité consiste à ce que le second tour soit restreint à deux candidats et que les deux candidats à y être qualifiés soient alors ceux qui ont obtenu les scores les plus élevés au premier tour de scrutin. Alternativement, les candidats sont parfois qualifiés au second tour dès lors que leur score du premier tour excède un seuil donné. Dans ce cas, le second tour peut opposer plus de deux candidats. Lorsqu'un second tour de scrutin est organisé, le candidat qui y a recueilli le plus de suffrages est déclaré élu s'il s'agit d'un scrutin uninominal et il bénéficie, si elle est prévue, d'une prime majoritaire dans le cas d'un scrutin de liste à la proportionnelle [2].

L'issue des deux tours d'une telle élection est bien évidemment liée. Les résultats du premier permettent souvent d'avoir une idée de l'issue du second tour. Il est en effet plausible que les électeurs ayant voté au premier tour pour l'un des candidats qualifiés au second tour confortent leur choix.

En revanche, le vote des électeurs ayant accordé au premier tour leurs suffrages à des candidats éliminés du second tour est entaché de plus d'incertitude. Le choix de ces électeurs n'est certes pas totalement aléatoire car il est notamment susceptible de dépendre de la proximité idéologique et partisane entre le candidat pour lequel ils ont voté au premier tour et chacun des candidats qualifiés. Ce choix est également susceptible d'être orienté par les consignes de vote que communiquent parfois à leurs électeurs les candidats éliminés à l'issue du premier tour. Cependant, connaître avec précision le comportement électoral des électeurs des candidats malheureux du premier tour demeure difficile pour plusieurs raisons. D'abord, le degré d'information politique des électeurs est hétérogène et certains d'entre eux sont susceptibles de ne pas identifier correctement les proximités politiques existant. Ensuite, l'espace des options politiques défendues par les candidats peut avoir plusieurs dimensions, et les électeurs n'accordent pas nécessairement le même ordre de priorité à ces options que leur candidat. Ainsi, certains électeurs accordent beaucoup d'importance à des thématiques que leur candidat juge plus secondaires. Ils peuvent donc être amenés à ne pas choisir au second tour le candidat qui semble le plus proche de leur choix du premier tour.

A titre d'illustration assez simple, considérons l'exemple d'une élection présidentielle. Trois candidats s'affrontent au premier tour : un candidat de gauche, un candidat centriste et un candidat de droite. Le candidat de gauche défend des options interventionnistes en matière d'économie et progressistes d'un point du vue culturel, le candidat centriste défend des positions libérales en économie et progressistes au plan culturel et le candidat de droite a un programme libéral en économie et conservateur du point de vue sociétal. Le candidat centriste est éliminé à l'issue du premier tour et le second tour oppose alors les deux autres candidats. Dans ce cas, un électeur du candidat centriste peut décider de voter au second tour pour le candidat de gauche s'il est plus sensible aux questions d'ordre culturel qu'aux enjeux économiques, et ce quand bien même la plupart des autres électeurs jugerait que l'économie constitue le thème prioritaire.

Enfin, le positionnement de certains candidats au premier tour ne permet tout simplement pas de les situer par rapport aux candidats du second tour. La difficulté à placer ces candidats sur le marché électoral tient

1. Les candidats peuvent être des personnes dans le cas de scrutins uninominaux mais aussi des listes de personnes dans le cas de scrutins dits « de liste ».

2. Un scrutin de liste « à la proportionnelle » consiste à attribuer tout ou partie de la composition d'une assemblée de manière proportionnelle aux scores réalisés par les listes. La représentation au sein d'une assemblée est néanmoins parfois conditionnée à l'obtention d'un score minimal.

parfois aux thèmes qu'ils ont développés au cours de leur campagne et qui les rendent atypiques. Le vote de second tour des électeurs de ces candidats est alors *a priori* très incertain.

1.2 Estimer les reports de vote entre les deux tours

En science politique, les choix électoraux au second tour de scrutin des électeurs ayant voté au premier tour pour un candidat éliminé sont appelés les ń reports de vote ż. Ils sont généralement appréhendés de manière agrégée et sous forme de proportions : telle fraction de l'électorat de premier tour de tel candidat se reporte sur un certain candidat au second tour. La connaissance de ces reports de votes comporte principalement deux intérêts. D'une part, les reports de votes constituent un élément de science politique à part entière. Ils renseignent au sujet du positionnement respectif des électorats, de l'ordre de priorité qu'ils attribuent aux différentes thématiques du débat public, de leur fidélité partisane et de leur degré d'information politique. A ce titre, ils sont susceptibles d'intéresser les formations politiques et les candidats eux-mêmes dans l'élaboration *ex ante* de leurs stratégies électorales et dans leur évaluation *ex post*. D'autre part, connaître les reports de votes permet de construire des prédictions des résultats du second tour à l'issue du premier tour. De telles prédictions peuvent par exemple s'avérer utiles lors des élections législatives françaises pour produire des projections de répartition des sièges à partir des résultats du seul premier tour [3].

Les estimations de reports de votes publiées, notamment dans la presse, résultent le plus souvent de sondages effectués directement auprès des électeurs. Les estimations produites par inférence à partir des résultats de scrutin par bureaux de vote restent rares [4]. Il faut en effet avoir à l'esprit que le scrutin à deux tours n'est pas le plus répandu pour les élections d'importance nationale, ce qui limite l'intérêt de la recherche dans ce champ, à l'étranger notamment. Ainsi, ni le Président des États-Unis, ni les parlements de la plupart des pays européens ne sont élus au moyen d'un scrutin à deux tours.

Cette étude a pour objectif d'estimer les reports de votes entre les deux tours de l'élection présidentielle française de 2007.

L'élection présidentielle de 2007 concerne, en métropole, environ 42 086 000 électeurs inscrits, répartis dans 64 020 bureaux de vote. Elle se déroule au terme du second mandat de Jacques Chirac [5] qui n'est pas candidat à sa succession. Le premier tour oppose douze candidats. Ces candidats n'ont cependant pas le même poids politique *a priori*. Une distinction est ainsi usuellement opérée entre les « grands candidats »et les « petits candidats ». Les « grands candidats »sont soutenus par de grandes formations politiques installées dans le paysage politique depuis longtemps, tandis que ce n'est pas le cas des « petits candidats ». Cette distinction se retrouve d'ailleurs de fait dans les résultats du premier tour, puisque quatre candidats rassemblent à eux seuls plus de 85% des suffrages exprimés.

- Nicolas Sarkozy, alors ministre de l'Intérieur, est le candidat de l'UMP (Union pour un Mouvement Populaire). Il s'agit du principal parti de la droite parlementaire qui a été créé en 2002 par des membres du RPR (Rassemblement pour la République), de DL (Démocratie Libérale) et de l'UDF (Union pour la Démocratie Française).

- La candidate du Parti Socialiste est Ségolène Royal. Son investiture par le Parti Socialiste fin 2006 constitue une surprise. La plupart des commentateurs de la vie politique ne commencent ainsi à évoquer son nom dans la compétition que quelques mois avant l'élection [6]. Cette ancienne ministre s'est

3. L'exploitation des reports de votes dans ce contexte repose toutefois implicitement sur l'hypothèse qu'ils sont relativement stables dans le temps.

4. Il existe néanmoins quelques études à ce sujet. L'une d'elles a notamment été conduite par Alain Bernard du département d'économie de l'école polytechnique [3]

5. Le second mandat de Jacques Chirac a également constitué le premier quinquennat de la $V^{\text{ème}}$ République, exception faite du quinquennat involontaire de Georges Pompidou qui est décédé en cours de mandat.

6. Une anecdote illustre bien le caractère inattendu de cette candidature. Ainsi, Alain Duhamel, journaliste politique français,

notamment fait remarquer en remportant largement l'élection régionale de Poitou-Charentes en 2004.

- François Bayrou est le candidat de l'UDF (Union pour la Démocratie Française). Déjà candidat en 2002, il a refusé de rallier l'UMP lors de sa création. De plus en plus critique vis-à-vis du gouvernement au cours du second mandat de Jacques Chirac, il s'est progressivement éloigné de la majorité de droite.

- Enfin, Jean-Marie Le Pen est le candidat du parti d'extrême-droite qu'il dirige depuis 1972, le Front National. Candidat pour la cinquième fois, il est parvenu à se qualifier au second tour en 2002, en battant notamment le Premier Ministre sortant, Lionel Jospin.

- Les autres candidats du premier tour sont, dans l'ordre alphabétique : Olivier Besancenot, José Bové, Marie-George Buffet, Arlette Laguiller, Frédéric Nihous, Gérard Schivardi, Philippe de Villiers et Dominique Voynet.

A l'issue du premier tour, Nicolas Sarkozy et Ségolène Royal se qualifient pour le second tour avec 31,2% des suffrages exprimés pour le premier et 25,9% pour la seconde. Suivent François Bayrou avec 18,6% puis Jean-Marie Le Pen avec 10,4% des voix. Au second tour, Nicolas Sarkozy est élu Président de la République avec 53,1% des suffrages exprimés.

Au-delà des aspects purement statistiques, l'estimation des reports de votes entre les deux tours de la présidentielle de 2007 est susceptible d'éclairer le dénouement de cette élection à (au moins) deux égards.

D'une part, le report des électeurs du candidat centriste François Bayrou est un paramètre clé du second tour. Ce dernier vient historiquement du centre-droit chrétien-démocrate mais il a mené une campagne très agressive contre Nicolas Sarkozy au premier tour. Il prétend ne s'ancrer ni à gauche, ni à droite et ainsi construire une force politique centriste autonome à long terme. Il tente de séduire une partie de l'électorat de centre-gauche en prétendant, sondages à l'appui, qu'il serait mieux à même que Ségolène Royal de battre Nicolas Sarkozy au second tour. Par ailleurs, il adopte des positions plus progressistes en matière culturelle en se déclarant par exemple favorable à l'extension du mariage aux couples homosexuels. Au final, il n'est donc pas évident qu'à l'issue du premier tour, ses électeurs vont se reporter sur le candidat de droite, Nicolas Sarkozy. Les deux candidats finalistes ont pleinement conscience que le report des électeurs de François Bayrou constitue un enjeu décisif pour le second tour. Ainsi, Ségolène Royal qui est distancée de plus de cinq points à l'issue du premier tour mène entre les deux tours une campagne très active en direction des électeurs de François Bayrou. Elle propose même à ce dernier d'organiser un débat pour éclairer le choix des électeurs, ce qu'il accepte. Il est intéressant d'évaluer rétrospectivement l'efficacité de cette stratégie.

D'autre part, le report des électeurs du candidat d'extrême droite Jean-Marie Le Pen constitue aussi, quoique dans une moindre mesure, un élément important du résultat final de l'élection. Le Front National, le parti qu'il dirige, est le plus souvent décrit comme un parti d'extrême-droite. A ce titre, il semble logique d'anticiper que ses électeurs vont se reporter sur Nicolas Sarkozy. Pour autant, il n'est pas évident que ce report sur le candidat de droite s'opère de manière nette. En effet, Nicolas Sarkozy a explicitement adopté un discours visant à attirer à lui des électeurs proches du Front National dés le premier tour. Il a notamment beaucoup mis en avant la question identitaire à cette fin. Cette stratégie semble avoir été payante au premier tour. Il rassemble en effet sous son nom plus de suffrages que tous les candidats de la droite modérée lors de la précédente élection, en 2002 [7]. A l'inverse, Jean-Marie Le Pen accuse pour sa part un retrait de plus de six points. A l'issue du premier tour, il est donc possible que Nicolas Sarkozy ait déjà rallié la plupart des électeurs du Front National susceptibles de le soutenir. Sa capacité à obtenir au second tour le vote des électeurs restés fidèles à Jean-Marie Le Pen est ainsi inconnue. Par ailleurs, la nature du vote frontiste fait débat parmi les observateurs de la vie politique. Dans quelle mesure ce vote traduit une adhésion au projet de droite radicale

ne la mentionne même pas dans son essai de 2006 qui concerne l'élection présidentielle en préparation [4].

7. Il s'agit de Jacques Chirac, François Bayrou, Alain Madelin et Christine Boutin.

porté par Jean-Marie Le Pen ou une forme de protestation envers les institutions constitue une question de la science politique à part entière. Dans le second cas, le report de ces électeurs apparaît plus incertain. Il n'est en effet pas exclu qu'ils votent blanc, s'abstiennent ou même qu'ils votent pour la candidate de gauche.

Ce document présente la construction de méthodes statistiques qui visent à estimer les reports de votes s'étant opérés entre les deux tours de la présidentielle de 2007. Ces méthodes procèdent par inférence à partir des résultats des deux tours du scrutin au niveau de chaque bureau de vote.

Pour ce faire, nous commençons par définir les différents cadres de modélisation statistique considérés. Ces modèles sont paramétriques et l'inférence de leurs paramètres constitue des problèmes statistiques complexes, pour lesquels nous devons élaborer des stratégies de résolution spécifiques. L'ensemble de ces considérations théoriques sont regroupées dans la partie 2. Explicitées, ces stratégies doivent être mises en œuvre de manière efficace étant donné la taille des problèmes considérés. En outre, pour la crédibilité des résultats, leur implémentation doit être testée ainsi que leur validité et leur robustesse. Ces aspects numériques et pratiques forment la partie 3. Enfin, ce travail peut être appliqué au cas de l'élection présidentielle de 2007. Pour ce faire, après une mise en forme des données, les modélisations statistiques sont déclinées sous l'angle de la sociologie politique, et les résultats obtenus sont précisés et analysés. Ces applications constituent la dernière partie 4.

2 Modélisation statistique et stratégies d'inférence

En démocratie, le vote de chaque électeur est confidentiel. Suivant ce principe, à chaque tour d'une élection, seul le résultat agrégé des votes est connu, et le niveau le plus fin d'agrégation est donné par le bureau de vote. Dans ces conditions, les reports d'un vote à un autre sont inconnus, et leur estimation à partir des résultats agrégés nécessite de définir un modèle. Comme chacun peut le comprendre, la décision individuelle de chaque électeur n'obéit à aucune fonction déterministe, de sorte que la modélisation est nécessairement statistique. Cette modélisation choisie, l'inférence des reports de vote prend la forme d'un problème mathématique complexe dont la résolution mobilise l'élaboration de stratégies spécifiques.

Dans cette partie, nous précisons successivement ces deux aspects en présentant et motivant les modélisations statistiques considérées et les techniques de résolution employées.

2.1 Modèles considérés des reports de vote

La modélisation des reports de vote s'appuie sur une conceptualisation d'une élection française à deux tours que nous détaillons. Une telle élection mobilise un corps électoral composé de N électeurs répartis dans K bureaux de vote, à raison de N^k électeurs dans le bureau k. Les électeurs du corps électoral se sont ou ont été inscrits avant le premier tour de l'élection et rattachés à un bureau de vote donné. En principe, les nombres N^k d'électeurs dans chaque bureau demeurent donc identiques lors des deux tours.

Pour chaque tour et chaque électeur, le vote peut être vu comme la sélection d'une unique modalité dans un ensemble fini de choix commun à tous les électeurs. Pour l'élection présidentielle de 2007, il y avait ainsi $I = 14$ choix possibles au premier tour : s'abstenir, voter blanc ou pour un des douze candidats ; et $J = 4$ choix possibles au second : s'abstenir, voter blanc ou pour les candidats Royal ou Sarkozy.

Nous introduisons deux concepts essentiels pour la suite. Premièrement, étant donné l'hypothèse de stabilité du corps électoral, il existe pour chaque bureau k, un **tableau de contingence** constitué des nombres $N^k_{i,j}$ d'électeurs ayant opté pour le choix i au premier tour et j au second. De ces tableaux, seules les marges suivant les lignes $N^k_{i,*}$ et suivant les colonnes $N^k_{*,j}$ sont connues et correspondent respectivement

aux résultats du premier et du second tour. Par construction, les identités suivantes sont vérifiées $\forall k = 1 \dots K$

$$\sum_{i=1}^{I} N_{i,*}^{k} = \sum_{j=1}^{J} N_{*,j}^{k} = N^{k} \tag{1}$$

Deuxièmement, les conditions de vote (isoloir, bulletin sous enveloppe, dépouillement à la fin, ...) font que lors du scrutin, les électeurs sont globalement ignorants des choix des autres électeurs, à l'exception de l'abstention qui fait l'objet d'une estimation communiquée et révisée régulièrement en cours de journée. Il apparaît toutefois que cette communication modifie peu la participation des électeurs, ce qui laisse supposer qu'il en est de même pour leur comportement. Par conséquent, il semble relativement raisonnable de considérer que les décisions de vote sont prises simultanément et qu'elles sont donc indépendantes entre électeurs.

Nous nous concentrons alors sur le report de vote d'un électeur ayant choisi la modalité i au premier tour. Ce report est fonction de caractéristiques qui lui sont propres, de son environnement et des éventuelles croyances qu'il a du choix des autres électeurs. Ces variables étant pour la plupart inobservables, nous modélisons le choix de la modalité j au second tour de manière probabiliste, et introduisons des **probabilités de report** $P(j \mid i)$. Ces probabilités diffèrent *a priori* entre électeurs d'un même bureau et entre bureaux. Ce faisant, pour des raisons élémentaires d'identification, il est nécessaire de donner un modèle pour ces probabilités. Dans ce travail, nous considérons trois modèles, que nous précisons et discutons.

2.1.1 Probabilités de report constantes par ensemble de bureaux

Une première modélisation consiste à considérer une partition de l'ensemble des bureaux

$$\{1, \dots, K\} = \bigsqcup_{z=1}^{Z} \mathbb{K}_z \tag{2}$$

et à supposer que pour chaque ensemble de bureaux $\mathbb{K}_z$, les probabilités de report sont identiques pour tous les électeurs rattachés aux bureaux composant cet ensemble. On note ces probabilités $p_{j|i}^{(z)}$.

Ce faisant, pour tous les bureaux k appartenant à l'ensemble $\mathbb{K}_z$, étant donné l'identité des probabilités de report entre électeurs et l'indépendance de leur vote, les tableaux de contingence $(N_{i,j}^{k})$ sont indépendants entre bureaux et suivent des lois multinomiales

$$\forall i = 1 \dots I \qquad (N_{i,1}^{k}, ..., N_{i,J}^{k}) \sim \mathcal{M}(N_{i,*}^{k}, p_{1|i}^{(z)}, ..., p_{J|i}^{(z)}) \tag{3}$$

que nous conditionnons à partir de la connaissance des marges $N_{*,j}^{k}$

$$\forall j = 1 \dots J \qquad \sum_{i=1}^{I} N_{i,j}^{k} = N_{*,j}^{k} \tag{4}$$

Cette modélisation étant posée, la détermination des reports de vote se ramène à l'identification des probabilités de report $p_{j|i}^{(z)}$ à partir de l'observation des marges $N_{i,*}^{k}$ et $N_{*,j}^{k}$. Étant donné l'indépendance des votes, cette identification peut être effectuée séparément pour chaque ensemble de bureaux $\mathbb{K}_z$.

Considérons un ensemble de bureaux $\mathbb{K}_z$ de cardinal K_z. Pour tout bureau k, les marges $N_{i,*}^{k}$ sont par construction vérifiées. De même, les marges $N_{*,j}^{k}$ sont liées par l'équation (1) et les probabilités de report somment chacune à 1. Il y a donc $I \times (J-1)$ paramètres indépendants à identifier à partir de $K_z \times (J-1)$ observations indépendantes. Par conséquent, les ensembles $\mathbb{K}_z$ doivent être suffisamment grands pour que l'identification puisse se faire. En particulier, il est impossible d'identifier ces probabilités par bureaux.

Plusieurs partitions d'ensembles de bureaux sont considérées dans cette étude et sont construites *a priori* sur la base de critères géographiques identifiés comme pertinents par la littérature sur la sociologie du vote.

Ces critères et ces constructions sont détaillées en partie 4.2. Par ailleurs, le cas d'un unique ensemble de bureaux est également étudié, ce qui revient à supposer des probabilités de report $p_{j|i}$ identiques pour tous les électeurs et à les identifier de manière nationale.

Cette première modélisation s'appuie sur le postulat d'identité des probabilités de report à la fois au sein des bureaux de vote et dans un grand nombre de bureaux. Cette hypothèse est forte et sa crédibilité, limitée. Elle conduit néanmoins à des jeux de paramètres simples à comprendre.

2.1.2 Probabilités de report logistiques

Une seconde modélisation est plus générale, et se propose de permettre aux probabilités de report de varier entre bureaux en fonction de variables socio-démographiques. Autrement dit, les probabilités de report restent identiques entre électeurs d'un même bureau k et sont notées $p^k_{j|i}$. Les tableaux de contingence $(N^k_{i,j})$ sont donc indépendants et suivent $\forall k = 1 \dots K$ des lois multinomiales

$$\forall i = 1 \dots I \qquad (N^k_{i,1}, ..., N^k_{i,J}) \sim \mathcal{M}(N^k_{i,*}, p^k_{1|i}, ..., p^k_{J|i}) \tag{5}$$

$$\text{conditionnées par} \quad \forall j = 1 \dots J \qquad \sum_{i=1}^{I} N^k_{i,j} = N^k_{*,j} \tag{6}$$

Par ailleurs, les probabilités sont fonction de variables socio-démographiques définies au niveau des bureaux et notées vectoriellement $\boldsymbol{X}^k$. Comme ces probabilités sont positives et doivent sommer à 1, nous choisissons de représenter leur lien fonctionnel avec les variables socio-démographiques par une fonction logistique de paramètres $(\boldsymbol{\beta}_{i,j})$:

$$\forall k = 1 \dots K, i = 1 \dots I, \qquad p^k_{j|i} = \frac{\exp \boldsymbol{\beta}'_{i,j} \boldsymbol{X}^k}{\sum_{\ell=1}^{J} \exp \boldsymbol{\beta}'_{i,\ell} \boldsymbol{X}^k} \tag{7}$$

Pour identifier ces paramètres, une condition nécessaire et classique consiste à supposer que les variables $\boldsymbol{X}_k$ sont linéairement indépendantes et à choisir une modalité de référence en posant par exemple

$$\forall i = 1 \dots I \qquad \boldsymbol{\beta}_{i,1} = \mathbf{0} \tag{8}$$

De cette façon, l'interprétation des paramètres $\boldsymbol{\beta}_{i,j}$ se fait par comparaison, puisqu'ils représentent la sensibilité des risques relatifs de vote pour la modalité j par rapport à la modalité de référence aux différentes variables socio-démographiques $\boldsymbol{X}^k$. Une manière plus commode consiste à étudier leur impact sur les probabilités $p^k_{j|i}$, ce qui peut se faire par le calcul de la moyenne de ces probabilités et leurs effets marginaux moyens :

$$\overline{p_{j|i}} = \frac{1}{N^*_{i,*}} \sum_{k=1}^{K} p^k_{j|i} N^k_{i,*} \qquad \overline{\frac{\partial p_{j|i}}{\partial \boldsymbol{X}}} = \frac{1}{N^*_{i,*}} \sum_{k=1}^{K} \frac{\partial p^k_{j|i}}{\partial \boldsymbol{X}^k} N^k_{i,*} \qquad \text{avec} \quad N^*_{i,*} = \sum_{k=1}^{K} N^k_{i,*} \tag{9}$$

Cette modélisation est ainsi plus générale mais plus complexe à interpréter. Le lien fonctionnel (7) présente, en outre, un caractère mécanique et déterministe peu crédible. Ce défaut pourrait être atténué en considérant des effets aléatoires propres à chaque bureau, par exemple sous la forme suivante

$$\forall k = 1 \dots K, i = 1 \dots I, \qquad p^k_{j|i} = \frac{\exp[\boldsymbol{\beta}'_{i,j} \boldsymbol{X}^k + u^k_{i,j}]}{\sum_{\ell=1}^{J} \exp[\boldsymbol{\beta}'_{i,\ell} \boldsymbol{X}^k + u^k_{i,j}]} \tag{10}$$

où les variables $u^k_{i,j}$ sont indépendantes et identiquement distribuées suivant une loi normale centrée de variance $\sigma^2_{i,j}$ inconnue avec pour l'identification, $u^k_{i,1} = \sigma^2_{i,1} = 0$. Une telle extension n'a pas été considérée dans ce travail faute de temps.

2.1.3 Probabilités de report constantes par population

Une dernière modélisation se propose de lever l'hypothèse d'identité des probabilités de report entre électeurs d'un même bureau. Pour ce faire, l'idée consiste à considérer des partitions du corps électoral par groupes d'électeurs que nous appelons « population »,

$$\{1,\dots,N\} = \bigsqcup_{\pi=1}^{\Pi} \mathbb{N}^{\pi} \tag{11}$$

et à supposer que les comportements sont identiques pour tous les électeurs appartenant à un même groupe. Ces populations sont formées *e.g.* par des tranches d'âge, des catégories socio-professionnelles ou encore des niveaux de formation initiale.

Pour lier ces partitions aux résultats électoraux, nous considérons leur intersection avec la partition formée par la définition des bureaux de vote

$$\forall \pi = 1 \dots \Pi \qquad \mathbb{N}^{\pi} = \bigsqcup_{k=1}^{K} \mathbb{N}^{k,\pi} \tag{12}$$

où les cardinaux $N^{k,\pi}$ des ensembles $\mathbb{N}^{k,\pi}$ doivent être connus.

A ce niveau, nous étendons les deux concepts introduits en début de partie, en définissant d'une part, du fait de la stabilité du corps électoral, des tableaux de contingence $N^k_{i,j,\pi}$ donnant le nombre d'électeurs du bureau k appartenant à la population π ayant pris les choix i et j au premier et au second tours, respectivement, et pour lesquels seules les marges $N^k_{i,*,*} = N^k_{i,*}$, $N^k_{*,j,*} = N^k_{*,j}$ et $N^k_{*,*,\pi} = N^k_{\pi}$ sont connues.

D'autre part, nous introduisons des probabilités de vote que nous supposons identiques pour tous les électeurs appartenant à une même population. Nous ne considérons pas directement les probabilités de report par population, car la répartition des résultats du premier tour $N^k_{i,*}$ entre les Π populations d'électeurs est inconnue et ne peut raisonnablement être supposée homogène. Nous considérons, au lieu, les probabilités jointes $p_{i,j|\pi}$ de choisir i au premier tour, j au second tour, conditionnellement au fait d'appartenir à la population π.

Dans ces conditions, étant donné l'indépendance des votes des électeurs, les tableaux de contingence $(N^k_{i,j,\pi})$ sont indépendants et $\forall k = 1 \dots K$ suivent des lois multinomiales

$$\forall \pi = 1 \dots \Pi \qquad (N^k_{1,1,\pi}, ..., N^k_{1,J,\pi}, ..., N^k_{I,J,\pi}) \sim \mathcal{M}(N^k_{\pi}, p_{1,1|\pi}, ..., p_{1,J|\pi}, ..., p_{I,J|\pi}) \tag{13}$$

que nous conditionnons par la connaissance des marges suivant les deux dimensions

$$\forall i = 1 \dots I \quad \sum_{j=1}^{J} \sum_{\pi=1}^{\Pi} N^k_{i,j,\pi} = N^k_{i,*} \qquad\qquad \forall j = 1 \dots J \quad \sum_{i=1}^{I} \sum_{\pi=1}^{\Pi} N^k_{i,j,\pi} = N^k_{*,j} \tag{14}$$

Dans les données, les nombres N^k_{π} ont fait l'objet d'estimation et concernent l'ensemble de la population vivant dans la zone géographique rattaché à chaque bureau. Ce faisant, la somme de ces nombres égale rarement la valeur du nombre d'inscrits N^k dans chaque bureau. En conséquence, le modèle (13) vérifiant exactement la valeur de ces nombres est probablement trop contraint. Pour l'atténuer, nous considérons, à la place, les proportions ρ^k_{π} de la population π dans le corps électoral du bureau k, et écrivons la probabilité $p^k_{i,j,\pi}$ qu'un électeur du bureau k appartiennent à la population π et fassent les choix i et j aux deux tours comme suit

$$p^k_{i,j,\pi} = p_{i,j|\pi} \times \rho^k_{\pi} \tag{15}$$

De cette façon, les tableaux de contingence $(N^k_{i,j,\pi})$ suivent $\forall k = 1 \dots K$ les lois multinomiales

$$(N^k_{1,1,1}, ..., N^k_{1,1,\Pi}, ..., N^k_{1,J,\Pi}, ..., N^k_{I,J,\Pi}) \sim \mathcal{M}(N^k, p^k_{1,1,1}, ..., p^k_{1,1,\Pi}, ..., p^k_{1,J,\Pi}, ..., p^k_{I,J,\Pi}) \tag{16}$$

conditionnées par la connaissance des marges (14).

Cette modélisation par « population »permet de considérer des comportements électoraux hétérogènes au sein d'un même bureau de vote. Toutefois, la prise en compte de cette hétérogénéité nécessite des données précises sur la décomposition par bureau des effectifs du corps électoral suivant certaines catégories socio-démographiques ou d'effectuer des approximations ne corrigeant que partiellement le problème.

2.1.4 Difficulté de l'inférence

Pour chacune des trois modélisations, l'étude des reports de vote passe par l'estimation de leurs paramètres : probabilités de report pour la première, coefficients $\boldsymbol{\beta}_{i,j}$ pour la seconde et probabilités de vote par population $p_{i,j|\pi}$ pour la dernière. Pour ce faire, le principe consiste à chercher le jeu de paramètres rendant maximale, dans le cadre du modèle, la vraisemblance des résultats électoraux des deux tours agrégés par bureaux. Si l'idée est simple, sa mise en œuvre pratique s'avère particulièrement délicate.

En effet, pour les trois modèles, l'expression de la loi jointe des marges $(N^k_{i,*}, N^k_{*,j})$ prend la forme de produits de convolution de lois qui étant donné le nombre de bureaux de vote et d'électeurs par bureaux, ne peuvent être calculés en un temps raisonnable, y compris en inversant les fonctions caractéristiques associées par transformées de Fourier rapides. La prise en compte des tableaux de contingence $(N^k_{i,j})$ ou $(N^k_{i,j,\pi})$ conduit à des problèmes d'optimisation contraints présentant une fonction objectif simple à exprimer et à calculer. Cette simplification apparente se fait toutefois au prix d'un élargissement considérable de l'espace de recherche par l'ajout d'une multitude de multiplicateurs de Lagrange et d'un nombre encore plus grand de paramètres de nuisance pour décrire chacun des tableaux de contingence. Ces derniers paramètres étant des entiers positifs, les programmes prennent la forme de problèmes d'optimisation contraints à valeurs entières de très grande taille, généralement réputés parmi les plus complexes à résoudre.

Dans ces conditions, les problèmes posés par l'estimation des paramètres nécessitent la mise en œuvre de stratégies de résolution spécifiques.

2.2 Revue de la littérature afférente à ces modèles

A notre connaissance et à l'exception d'un travail [17] accompli par des étudiants en 2013, ces problèmes statistiques n'apparaissent pas directement dans la littérature. Par plusieurs aspects, ils se rapprochent toutefois de problèmes ayant été largement étudiés et pour lesquels les résultats et les méthodes sont riches d'enseignements que nous résumons et discutons en rapport avec notre propos.

Un premier pan de la littérature concerne le calage sur marges dont le problème consiste à estimer au mieux les cellules d'une table de contingence dont on connaît la valeur des marges et pour laquelle on dispose d'une première estimation. A ce sujet, nous citons notamment l'algorithme de Deming et Stephan [16], la preuve de sa convergence faite par [42] en interprétant l'algorithme comme une minimisation de la divergence de Kullback-Leibler, et l'estimateur asymptotiquement efficace établi par [23]. La difficulté réside dans la prise en compte *simultanée* des équations de marge et sur le traitement des zéros pouvant apparaître dans la table.

Un second pan concerne les tests d'indépendance à distance finie de deux variables discrètes suivant la méthode du χ^2 de Pearson. Pour calculer la loi de la statistique de test [37], tous les tableaux de contingence ayant les mêmes marges que le tableau d'origine doivent être énumérés ou suivant l'approximation de Monte-Carlo, tirés efficacement. La distribution des tableaux de contingence $(N_{i,j})$ conditionnellement à la connaissance des marges et sous l'hypothèse d'indépendance des variables s'exprime simplement. Sur la base de cette expression, Boyett [24] puis Patefield [51] ont proposé des algorithmes d'efficacité croissante de tirage de tels tableaux.

Dans ces deux premiers cas et contrairement à nos problèmes, les approches se placent dans le cas d'une distribution particulière des tableaux de contingence ou supposent la connaissance approchée de la distribution.

Un troisième pan correspond à l'inférence écologique pour laquelle il s'agit d'estimer, à partir de données $X^1 \dots X^K$ agrégées sur I populations et de la connaissance de leur taille $N_1^k \dots N_I^k$, les contributions moyennes $\bar{p}_i$ de chacune d'elles à la variable d'intérêt X. Les contributions sont susceptibles de varier en fonction de k et leurs moyennes sont définies comme suit

$$X^k = \left(\sum_{i=1}^{I} p_i^k N_i^k \right) / \left(\sum_{i=1}^{I} N_i^k \right) \qquad \bar{p}_i = \left(\sum_{k=1}^{K} N_i^k p_i^k \right) / \left(\sum_{k=1}^{K} N_i^k \right) \tag{17}$$

Ces inférences sont fréquemment utilisées en sciences politiques. Dans notre cas, les populations sont les électeurs ayant opté pour une modalité i au premier tour ; les paramètres d'intérêt, les votes pour les options du second tour ; et $\bar{p}_i$, les probabilités de report.

L'écueil principal de ce type d'inférence est l'*illusion écologique* [12] et correspond à l'erreur importante qu'on peut commettre en estimant des comportements individuels à partir de données agrégées et notamment, en supposant à tort une identité de ces comportements par population.

Plusieurs méthodes ont néanmoins été proposées. La première de Duncan-Davis [10] utilise le fait que les contributions sont bornées et encadre leurs valeurs, ce qui conduit à des intervalles souvent peu informatifs. La deuxième formalisée par Goodman [30] effectue la régression aux moindres carrés de la première équation de (17) en y substituant les p_i^k par $\bar{p}_i$. Cette régression est valable à condition que la proportion de population soit décorrélée des contributions. Elle conduit toutefois à des contributions pouvant sortir de l'intervalle auquel elles doivent appartenir.

King [32] [33] propose une approche permettant une synthèse des précédentes en utilisant un modèle bayésien hiérarchique simulé par chaînes de Markov Monte Carlo. Le premier niveau distribue les paramètres d'intérêt X_j^k suivant une loi multinomiale dont les probabilités s'expriment comme dans (17) sous forme d'une somme de contributions pondérées par les proportions des populations. Le second contraint ces contributions à l'intervalle $[0, 1]$ en les distribuant suivant des lois de Dirichlet dont les paramètres sont des fonctions logistiques de covariables. Enfin, un dernier niveau donne sans raison un *a priori* exponentiel pour chacun des paramètres de ces fonctions logistiques. L'auteur justifie son modèle en expliquant l'avoir testé avec succès sur un très grand nombre de cas et de données simulées. La pertinence de ce modèle vis à vis du phénomène d'illusion écologique fait toutefois débat [12] [18].

Judge et al [28] insistent sur le fait que l'inférence écologique est un problème inverse généralement mal posé. Ce faisant, les auteurs proposent d'estimer les contributions en minimisant une divergence de Read-Cressie [46] contrainte par l'équation (17) et les autres informations dont on dispose. Cette estimation est semi-paramétrique, généralise les méthodes de minimisation de Kullback-Leibler, de maximum d'entropie et de vraisemblance empirique, et s'appuie sur des équations de moment.

Face au problème d'illusion écologique, les méthodes s'orientent donc soit vers des modèles complexes étudiés suivant des techniques numériques bayésiennes, soit vers des inférences semi-paramétriques ne retenant que quelques équations de moments. Nos problèmes sont proches de ceux de l'inférence écologique et les diverses hypothèses d'identité de comportement laissent planer un risque d'illusion écologique. Toutefois, ils s'en distinguent dans la mesure où les variables d'intérêt et les effectifs de population, ie. les résultats associés au premier et au second tour, sont pris en compte quelque soit la réalisation et non en espérance.

Signalons enfin que le modèle en probabilité de report constant par ensembles de bureaux a fait l'objet du mémoire [17]. Cette modélisation a semblé naturelle aux auteurs et n'a pas été discutée. Devant les difficultés de résolution, la stratégie a consisté à se placer dans un cadre bayésien avec un *a priori* non informatif, et à

utiliser l'algorithme de Gibbs pour simuler les probabilités de report et les tableaux de contingence. L'espace de ces tableaux à marges contraintes a été exploré suivant une marche aléatoire contrôlée par l'algorithme de Metropolis-Hastings et consistant à chaque étape à modifier *a minima* chacun des tableaux. Dans le calcul du ratio d'acceptation, les auteurs négligent le caractère contraint de la marche aléatoire, ce qui conduit à une déformation de la distribution des tableaux. Au vu des résultats obtenus, cette exploration s'est révélée numériquement très peu efficace. En particulier, ont seulement été présentées des estimations par département, par ailleurs peu crédibles au regard de la littérature de sociologie du vote en France.

2.3 Stratégies de résolution adoptées

Étant donné les éléments bibliographiques, nous cherchons à estimer les paramètres des modèles de reports de vote par des méthodes numériques bayésiennes. Ce faisant, nous reprenons la démarche adoptée par [17] et la développons suivant le double objectif d'en améliorer l'efficacité et de l'étendre à l'inférence des trois types de modélisations considérés.

Pour ce faire, nous considérons des lois *a priori* de faible poids pour les différents jeux de paramètres, et de préférence conjuguée à la vraisemblance afin de simplifier certains calculs. Nous supposons ainsi que les probabilités de report par ensembles de bureaux et que les probabilités de vote par population suivent *a priori* des lois de Dirichlet indépendantes

$$\forall z = 1 \dots Z, i = 1 \dots I \qquad (p^{(z)}_{1|i}, ..., p^{(z)}_{J|i}) \sim \mathcal{D}(\alpha^{(z)}_{i,1}, ..., \alpha^{(z)}_{i,J}) \tag{18}$$

$$\forall \pi = 1 \dots \Pi \qquad (p_{1,1|\pi}, ..., p_{1,J|\pi}, ..., p_{I,J|\pi}) \sim \mathcal{D}(\alpha_{1,1,\pi}, ..., \alpha_{1,J,\pi}, ..., \alpha_{I,J,\pi}) \tag{19}$$

dont les paramètres $(\alpha^{(z)}_{i,j})$ et $(\alpha_{i,j,\pi})$ sont petits, non nuls et non informatifs. Par exemple

$$\forall i = 1 \dots I, j = 1 \dots J, z = 1 \dots Z, \pi = 1 \dots \Pi \qquad \alpha^{(z)}_{j|i} = \alpha_{i,j,\pi} = 1 \tag{20}$$

Pour le modèle à probabilité de report logistique, il n'apparaît pas de lois conjuguées simples. En revanche, étant définies sur tout un espace vectoriel réel, il semble assez naturel de considérer des lois *a priori* normales centrées et que nous supposons indépendantes et identiquement distribuées pour simplifier

$$\forall i = 1 \dots I, j = 2 \dots J \qquad \boldsymbol{\beta}_{i,j} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{21}$$

où $\boldsymbol{I}$ désigne la matrice identité et où σ est un réel strictement positif et suffisamment grand mais pas trop pour des raisons de stabilité numérique (*e.g.* σ^2=10).

Par ailleurs, la loi jointe des marges $(N^k_{i,*}, N^k_{*,j})$ ou $(N^k_{i,*,*}, N^k_{*,j,*}, N^k_{*,*,\pi})$ étant particulièrement complexe à calculer et à simuler directement, nous considérons les tableaux de contingence $(N^k_{i,j})$ et $(N^k_{i,j,\pi})$, qui représentent des variables latentes et des paramètres de nuisance. Étant donné la nature distincte et duale de ces tableaux et des paramètres, nous inscrivons assez naturellement les simulations dans le cadre d'un algorithme de Gibbs à deux niveaux. Autrement dit, nous construisons une suite de paramètres et de tableaux de contingence en alternant, par récurrence et selon les modèles, les tirages suivants

$$\begin{cases} \forall k = 1 \dots K & ({}^nN^k_{i,j}) \sim (N^k_{i,j}) \mid ({}^{n-1}p^{(z)}_{j|i}), (N^k_{i,*}), (N^k_{*,j}) \\ \forall i = 1 \dots I & ({}^np^{(z)}_{1|i}, ..., {}^np^{(z)}_{J|i}) \sim (p^{(z)}_{1|i}, ..., p^{(z)}_{J|i}) \mid ({}^nN^k_{i,j}) \end{cases} \tag{22}$$

$$\begin{cases} \forall k = 1 \dots K & ({}^nN^k_{i,j}) \sim (N^k_{i,j}) \mid ({}^{n-1}\boldsymbol{\beta}_{i,j}), (N^k_{i,*}), (N^k_{*,j}) \\ \forall i = 1 \dots I & ({}^n\boldsymbol{\beta}_{i,2}, ..., {}^n\boldsymbol{\beta}_{i,J}) \sim (\boldsymbol{\beta}_{i,2}, ..., \boldsymbol{\beta}_{i,J}) \mid ({}^nN^k_{i,j}) \end{cases} \tag{23}$$

$$\begin{cases} \forall k = 1 \dots K & ({}^nN^k_{i,j,\pi}) \sim (N^k_{i,j,\pi}) \mid ({}^{n-1}p_{i,j|\pi}), (N^k_{i,*}), (N^k_{*,j}), (N^k_{\pi}) \\ \forall \pi = 1 \dots \Pi & ({}^np_{1,1|\pi}, ..., {}^np_{I,J|\pi}) \sim (p_{1,1|\pi}, ..., p_{I,J|\pi}) \mid ({}^nN^k_{i,j,\pi}) \end{cases} \tag{24}$$

où n indexe les suites de paramètres et de tableaux de contingence. Nous détaillons et justifions dans les paragraphes suivants les principes des procédures mises en œuvre pour simuler efficacement tous ces tirages.

2.3.1 Simulation des probabilités

Étant donné les lois des tableaux de contingence (3), (13) , (16) et les lois *a priori* (18)-(19), les distributions *a posteriori* des probabilités de report et de vote par population s'écrivent simplement

$$f(p^{(1)}_{1|1},...,p^{(Z)}_{J|I} \mid (N^k_{i,j})) \propto \prod_{z=1}^{Z}\prod_{k\in\mathbb{K}_z}\prod_{i=1}^{I}\prod_{j=1}^{J}(p^{(z)}_{j|i})^{N^k_{i,j}} \prod_{z=1}^{Z}\prod_{i=1}^{I}\prod_{j=1}^{J}(p^{(z)}_{j|i})^{\alpha^{(z)}_{i,j}-1} \propto \prod_{z=1}^{Z}\prod_{i=1}^{I}\prod_{j=1}^{J}(p^{(z)}_{j|i})^{\tilde{\alpha}^{(z)}_{i,j}-1} \tag{25}$$

$$f(p_{1,1|1},...,p_{I,J|\Pi} \mid (N^k_{i,j,\pi})) \propto \prod_{\pi=1}^{\Pi}\prod_{k=1}^{K}\prod_{i=1}^{I}\prod_{j=1}^{J}(p_{i,j|\pi})^{N^k_{i,j,\pi}} \prod_{\pi=1}^{\Pi}\prod_{i=1}^{I}\prod_{j=1}^{J}(p_{i,j|\pi})^{\alpha_{i,j,\pi}-1} \propto \prod_{\pi=1}^{\Pi}\prod_{i=1}^{I}\prod_{j=1}^{J}(p_{i,j|\pi})^{\tilde{\alpha}_{i,j,\pi}-1} \tag{26}$$

avec $\forall i = 1\dots I, j = 1\dots J, z = 1\dots Z, \pi = 1\dots\Pi$

$$\tilde{\alpha}^{(z)}_{i,j} = \alpha^{(z)}_{i,j} + \sum_{k\in\mathbb{K}_z} N^k_{i,j} \qquad \tilde{\alpha}_{i,j,\pi} = \alpha_{i,j,\pi} + \sum_{k=1}^{K} N^k_{i,j,\pi} \tag{27}$$

En d'autres termes, ces probabilités suivent *a posteriori* des lois de Dirichlet indépendantes

$$\forall z = 1\dots Z, i = 1\dots I \qquad (p^{(z)}_{1|i},...,p^{(z)}_{J|i}) \mid (N^k_{i,j}) \sim \mathcal{D}(\tilde{\alpha}^{(z)}_{i,1},...,\tilde{\alpha}^{(z)}_{i,J}) \tag{28}$$

$$\forall \pi = 1\dots\Pi \qquad (p_{1,1|\pi},...,p_{I,J|\pi}) \mid (N^k_{i,j,\pi}) \sim \mathcal{D}(\tilde{\alpha}_{1,1,\pi},...,\tilde{\alpha}_{I,J,\pi}) \tag{29}$$

Ce faisant, les deuxièmes étapes des récurrences (22) et (24) s'effectuent sans difficulté.

2.3.2 Simulation des tableaux de contingence

La simulation des tableaux de contingence en première étape des récurrences (22)-(24) est plus complexe à réaliser. Les contraintes de marges imposées suivant différentes dimensions des tableaux interdisent de tirer indépendamment chacune des cellules. Étant donné en outre le caractère entier et positif de ces coefficients, ces tableaux correspondent à des points évoluant dans des parties finies de réseaux de grandes dimensions et à la géométrie complexe délimitée par des polyèdres.

Le mémoire [17] avait proposé d'explorer cet espace suivant une marche aléatoire contrôlée selon la méthode de Metropolis-Hastings, consistant, à chaque itération, à modifier d'une unité quatre cellules de chaque tableau de sorte à respecter les contraintes de marge. Étant donné les dimensions de l'espace à parcourir, cette stratégie s'est révélée peu efficace. En outre, les auteurs ont considéré à tort que cette marche était parfaitement symétrique et ont par là-même ignoré le caractère fini du réseau dans lequel évoluent ces tableaux. Dans ce travail, nous reprenons l'idée d'utiliser la méthode de Metropolis-Hastings, mais cherchons à reproduire fidèlement la distribution des tableaux de contingence et à améliorer radicalement la procédure d'exploration de l'espace de ces tableaux.

Le parcours d'un tel espace rejoint le problème de tirage de tableaux de contingence à marges constantes devant être effectué pour les tests d'indépendance à distance finie. Pour ce faire, l'algorithme de Patefield [51] a la réputation dans la littérature d'être particulièrement performant. Toutefois, étant donné le nombre élevé d'électeurs par bureaux, les espaces des tableaux sont particulièrement grands, de sorte que le parcours de ces espaces suivant cet algorithme demeure en l'état trop peu efficace. Sur des essais de cette méthode réalisés à partir de données simulées, on observe un taux d'acceptation de Metropolis-Hastings extrêmement faible. Il convient donc d'améliorer la procédure en cherchant à construire des tableaux candidats respectant les marges et ayant une distribution relativement proche de la distribution cible. La construction de ces candidats s'effectue en tenant compte des probabilités inégales de report et en mimant le fonctionnement de l'algorithme de Patefield.

Principe de l'algorithme de Patefield

Avant de détailler cette construction, reprenons en détail les principes de la méthode de Patefield. Comme indiqué au §2.2, la loi de probabilité des tableaux de contingence $(N_{i,j})$ conditionnellement à la connaissance des marges et sous l'hypothèse de leur indépendance prend la forme simple [25] suivante

$$P(N_{i,j} = n_{i,j} \mid N_{i,*}, N_{*,j}) = \frac{\prod_{i=1}^{I} N_{i,*}! \prod_{j=1}^{J} N_{*,j}!}{N! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{i,j}!} \qquad \text{avec} \quad N = \sum_{i=1}^{I} N_{i,*} = \sum_{j=1}^{J} N_{*,j} \tag{30}$$

En marginalisant cette expression, on trouve que

$$P(N_{1,1} = x \mid (N_{i,*}), (N_{*,j})) = C^{x}_{N_{1,*}} C^{N_{*,1}-x}_{N-N_{1,*}} / C^{N_{*,1}}_{N} \tag{31}$$

Ainsi, on peut générer le tableau de contingence en le remplissant ligne par ligne, de gauche à droite et de haut en bas, car l'expression de la probabilité de chaque cellule sachant les cellules déjà remplies et les marges est connue. En effet, pour une cellule (i, j) donnée, on note (cf. figure 1) l'ensemble des coefficients déjà tirés $\mathcal{N}_{i,j} = \{N_{1,1}...N_{i-1,J}, N_{i,1}, ..., N_{i,j-1}\}$, les restes de marges à satisfaire $N'_{i,*} = N_{i,*} - N_{i,1} - ... - N_{i,j-1}$ et $N'_{*,j} = N_{*,j} - N_{1,j} - ... - N_{i-1,j}$ et le nombre à répartir depuis la cellule $N' = N - N_{1,*} - ... - N_{i-1,*} - (N_{i,1} - ... - N_{i,j-1}) - ... - (... - N_{I,j-1})$. Avec ces notations, on déduit de (31) que

$$P(N_{i,j} = x \mid \mathcal{N}_{i,j}, (N_{k,*}), (N_{*,\ell})) = C^{x}_{N'_{i,*}} C^{N'_{*,j}-x}_{N'-N'_{i,*}} / C^{N'_{*,j}}_{N'} \tag{32}$$

Autrement dit, conditionnellement aux marges et aux coefficients déjà traités, $N_{i,j}$ suit une loi hypergéométrique. Dans ce résultat, l'identité (30) joue un rôle essentiel. Toutefois, on peut réinterpréter ce résultat en remarquant que

$$P(N_{i,j} = x \mid \mathcal{N}_{i,j}, (N_{k,*}), (N_{*,\ell})) = P(N_{i,j} = x \mid N'_{i,*}, N'_{*,j}, N') \tag{33}$$

et que ce faisant, l'algorithme se ramène le tirage de chaque cellule au cas d'un tableau de contingence 2×2 dont les marges sont imposées (cf. figure 1).

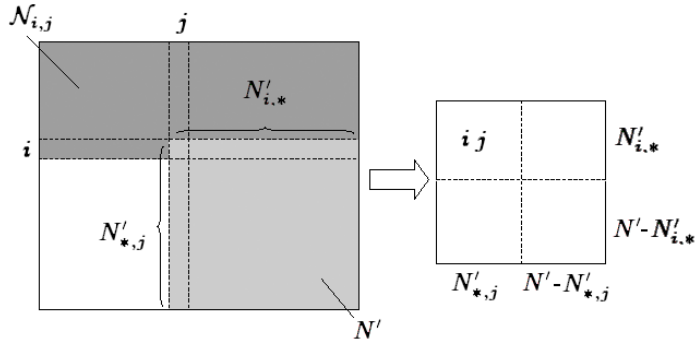

FIGURE 1 – Principe de l'algorithme de Patefield

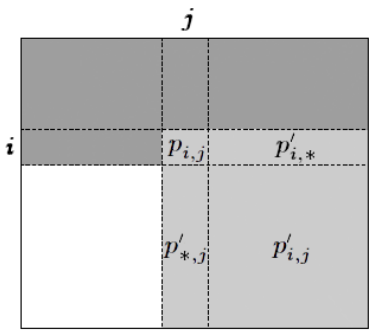

FIGURE 2 – Probabilité réduite

Construction des tableaux candidats à deux dimensions

Pour construire les tableaux de contingence candidats, nous procédons par analogie et faisons comme si l'égalité (33) était vérifiée alors que les votes du premier et du second tour ne sont pas indépendants. Ce faisant, nous ramenons le tirage de chaque cellule au cas d'un tableau de contingence 2×2. Pour ce faire, il est préférable de considérer des probabilités jointes de vote $p^{k}_{i,j}$ et d'écrire que sans conditionnement par les marges, les coefficients du tableau de contingence k suivent des lois multinomiales

$$(N^{k}_{1,1}, ..., N^{k}_{I,J}) \sim \mathcal{N}(N^{k}, p^{k}_{1,1}, ..., p^{k}_{I,J}) \tag{34}$$

Le passage des probabilités de report à ces probabilités jointes se fait selon l'approximation suivante

$$p^k_{i,j} = p_{j|i} \frac{N^k_{i,*}}{N^k} \tag{35}$$

Dans ces conditions, pour un bureau k et une cellule (i,j) donnés, la condensation sur un tableau 2×2 s'écrit, en définissant $N'^k_{i,*}$, $N'^k_{*,j}$, N'^k comme précédemment et en posant (cf. figure 2) $p'^k_{i,*} = p^k_{i,j+1} + ... + p^k_{i,J}$, $p'^k_{*,j} = p^k_{i+1,j} + ... + p^k_{I,j}$, $p'^k = p^k_{i,j} + ... + p^k_{I,J}$ et $p'^k_{i,j} = p'^k - p'^k_{i,*} - p'^k_{*,j} + p'^k_{i,j}$,

$$(N^k_{i,j}, N'^k_{i,*} - N^k_{i,j}, N'^k_{*,j} - N^k_{i,j}, N'^k - N'^k_{i,*} - N'^k_{*,j} + N^k_{i,j}) \sim \mathcal{M}\left(N'^k, \frac{p^k_{i,j}}{p'^k}, \frac{p'^k_{i,*}}{p'^k}, \frac{p'^k_{*,j}}{p'^k}, \frac{p'^k_{i,j}}{p'^k}\right) \tag{36}$$

On en déduit la loi de $N^k_{i,j}$ conditionnellement aux marges de ce tableau réduit

$$P(N^k_{i,j} = x \mid N'^k_{i,*}, N'^k_{*,j}, N'^k) \propto C^x_{N'^k_{i,*}} C^{N'^k_{*,j}-x}_{N'^k - N'^k_{i,*}} \omega^x \quad \text{avec} \quad \omega = \frac{p^k_{i,j}\, p'^k_{i,j}}{p'^k_{i,*} p'^k_{*,j}} \tag{37}$$

Autrement dit, $N^k_{i,j}$ conditionnellement aux marges restantes agrégées, suit une loi hypergéométrique décentrée de Fisher, pour laquelle ont été proposés dans la littérature, des algorithmes de simulation et de calcul de la loi relativement efficaces que nous détaillons en partie 3.1.1. Notons que dans le cas où les probabilités de vote au premier et au second tour sont indépendantes, nous retrouvons $\omega = 1$ et la loi hypergéométrique (32).

La qualité du candidat obtenu dépend sensiblement de l'erreur commise par l'approximation (33), qui revient à négliger l'action précise des contraintes de marges opérant sur les colonnes et les lignes situées à droite et en bas de la cellule considérée. Ce faisant, contrairement à l'algorithme original de Patefield, la distribution des tableaux de contingence candidats dépend de l'ordre des lignes et des colonnes suivi par l'algorithme, et l'approximation est d'autant meilleure que les cellules les plus contraintes par les marges sont tirées en premier. Suivant cette analyse, nous optimisons la simulation en nous inscrivant dans le cadre d'une stratégie hybride au sein des itérations de Gibbs, consistant à permuter aléatoirement, à chaque tirage d'un tableau de contingence, l'ordre selon lequel les lignes et les colonnes sont traitées.

Nous obtenons ainsi un algorithme de simulation efficace pour les premières étapes des récurrences (22) et (23) où les $(p^k_{j|i})$ sont données par (7). Nous observons, en effet, dans l'ensemble des calculs effectués que cette procédure de tirage conduit à des taux d'acceptation de Metropolis-Hastings supérieurs, en moyenne sur les bureaux et les itérations, à environ 30%.

Construction des tableaux candidats à trois dimensions

Pour les modèles en population, le tirage des tableaux de contingence à trois dimensions en première étape de la récurrence (24) peut s'effectuer avec les mêmes idées et les mêmes procédures, en écrivant $\forall k = 1 \dots K$

$$P[(N^k_{i,j,\pi}) \mid (N^k_{i,*}), (N^k_{*,j}), (N^k_\pi)] = P[(N^k_{i,j}) \mid (N^k_{i,*}), (N^k_{*,j}), (N^k_\pi)] \times P[(N^k_{i,j,\pi}) \mid (N^k_{i,j}), (N^k_\pi)] \tag{38}$$

$$\text{où} \quad \forall i = 1 \dots I, j = 1 \dots J \qquad N^k_{i,j} = \sum_{\pi=1}^{\Pi} N^k_{i,j,\pi} \tag{39}$$

En outre, similairement à (35), nous faisons l'approximation (15) avec $\rho^k_\pi = N^k_\pi / N^k$

$$p^k_{i,j,\pi} = p_{i,j|\pi} \frac{N^k_\pi}{N^k} \tag{40}$$

Ce faisant, si on tient compte des trois contraintes de marge, le tableau de contingence candidat peut être obtenu par l'application répétée de la construction décrite ci-dessus : une première fois avec les contraintes

de marge $(N^k_{i,*})$, $(N^k_{*,j})$ et les probabilités $p^k_{i,j} = p^k_{i,j,1} + ... + p^k_{i,j,\Pi}$, une seconde avec les marges $(N^k_{i,j})$, (N^k_{π}) et les probabilités $p^k_{i,j,\pi}$. Notons que ce faisant, les marges liées aux résultats électoraux sont prises en compte avant celles liées aux effectifs des populations et que cet ordre est fixé.

Si on considère la version atténuée (16), le tableau candidat peut similairement être obtenu en effectuant la même première étape et en tirant les cellules $(N^k_{i,j,\pi})$ suivant les lois multinomiales

$$\forall i = 1 \dots I, j = 1 \dots J \qquad (N^k_{i,j,1}, ..., N^k_{i,j,\Pi}) \sim \mathcal{M}\left(N^k_{i,j}, \frac{p^k_{i,j,1}}{p^k_{i,j}}, ..., \frac{p^k_{i,j,\Pi}}{p^k_{i,j}}\right) \tag{41}$$

Ces procédures sont détaillées en annexe au §A.3. Sur l'ensemble des calculs effectués, on observe suivant cette seconde stratégie, des taux d'acceptation de Metropolis-Hastings moyens légèrement inférieurs à 20%.

2.3.3 Simulation des paramètres $\boldsymbol{\beta}_{i,j}$

Parmi les trois récurrences considérées, reste à détailler la deuxième étape de (23) correspondant aux tirages des paramètres $(\boldsymbol{\beta}_{i,j})$. Leur simulation est également complexe, comme le montre l'expression[8] de leur densité *a posteriori*

$$\forall i = 1 \dots I \qquad f(\boldsymbol{\beta}_{i,2}, ..., \boldsymbol{\beta}_{i,J} \mid (N^k_{i,j})) \propto \prod_{j=2}^{J} \exp\left(-\frac{\|\boldsymbol{\beta}_{i,j}\|^2}{2\sigma^2}\right) \prod_{k=1}^{K} \prod_{j=1}^{J} (p^k_{j|i})^{N^k_{i,j}} \tag{42}$$

$$\text{avec} \qquad p^k_{j|i} = \frac{\lambda^k_{i,j}}{\lambda^k_{i,*}} \qquad \lambda^k_{i,j} = \exp\left(\boldsymbol{\beta}'_{i,j} \boldsymbol{X}^k\right) \qquad \lambda^k_{i,*} = \sum_{j=1}^{J} \lambda^k_{i,j} \tag{43}$$

Selon cette expression, le tirage des $\boldsymbol{\beta}_{i,j}$ revient à simuler dans la loi *a posteriori* d'une régression logistique multinomiale. Dans la littérature, ce problème est réputé difficile et a fait l'objet de plusieurs travaux. Nous pouvons notamment citer à ce sujet les approches de Holmes & Held [31] (2006), Frühwirth-Schnatter & Frühwirth [38] (2010) et Polson & Scott [27] (2013).

La plupart de ces méthodes ne traite pas directement le cas d'une régression logit multinomiale mais se ramène au cas binomial en figeant tous les paramètres $\boldsymbol{\beta}_{i,j}$ sauf un, selon l'algorithme de Gibbs. En revanche, elles ont toutes en commun de suivre l'approche proposée par Albert & Chib [40] (1993) pour les régressions probits, consistant à représenter la vraisemblance de chaque observation $N^k_{i,j}$ comme la marginale d'un mélange de lois normales. En d'autres termes, l'idée est d'introduire des variables latentes telles qu'en marginalisant, on retrouve les distributions (42) et que conditionnellement aux autres, une de ces variables soit distribuée suivant une loi normale. Ce faisant, on est ramené au cas trivial de la détermination de la distribution *a posteriori* d'un *a priori* normal par une vraisemblance normale, et à la simulation des paramètres $\boldsymbol{\beta}_{i,j}$ suivant les lois normales résultantes.

La méthode d'**Holmes & Held** [31] s'appuie sur un résultat d'Andrews-Mallows [11] et représente la distribution logistique comme la marginale d'un mélange de lois normales par des variables aléatoires indépendantes suivant des lois de Kolmogorov-Smirnov. Ainsi, leur méthode considère essentiellement le cas binaire $J = 2$ et $K = 1$, auquel on peut se ramener en appliquant l'algorithme de Gibbs aux paramètres $\boldsymbol{\beta}_{i,2}$ ... $\boldsymbol{\beta}_{i,J}$ et en considérant les nombres de vote $N^k_{i,j}$ comme la somme de $N^k_{i,*}$ variables binaires. Ce faisant, on introduit pour des simulations au niveau national, environ 90 millions de variables latentes, ce qui conduit à des calculs aux coûts prohibitifs.

La méthode de **Polson & Scott** [27] considère le cas des régressions binomiales logistiques, auquel on peut se ramener avec l'algorithme de Gibbs. Dans ce cas, les auteurs montrent que le terme associé à la

8. compte tenu de la condition d'identification (8)

vraisemblance dans la distribution *a posteriori* des paramètres peut être représenté comme la distribution marginale d'un mélange de lois normales par une variable suivant une loi Pólya-gamma, ce qui s'écrit avec $J = 2$

$$\frac{(\lambda_{i,2}^k)^{N_{i,j}^k}}{(\lambda_{i,*}^k)^{N_{i,*}^k}} \propto \exp\left(\frac{1}{2}\boldsymbol{\beta}_{i,2}'\boldsymbol{X}^k(N_{i,2}^k - N_{i,1}^k)\right)\int_0^\infty \exp\left(-\omega_{i,2}^k(\boldsymbol{\beta}_{i,2}'\boldsymbol{X}^k)^2\right)p_\gamma(\omega_{i,2}^k)\,d\omega_{i,2}^k \tag{44}$$

où p_γ désigne la densité de la loi Pólya-Gamma $PG(N_{i,*}^k, 0)$.

A ce niveau, comme on a également $\omega_{i,2}^k \mid \boldsymbol{\beta}_{i,2}, (N_{i,j}^k) \sim PG(N_{i,*}^k, \boldsymbol{\beta}_{i,2}'\boldsymbol{X}^k)$, toute la difficulté se concentre sur la simulation efficace des lois de Pólya-Gamma. Pour de très petites valeurs de $N_{i,*}^k$, les auteurs utilisent un algorithme d'acceptation / rejet basé sur la technique des séries alternées [29]. Pour des valeurs supérieures à quelques unités seulement, la simulation passe par des approximations obtenues suivant la méthode du point col. Enfin pour des valeurs dépassant quelques centaines, la loi est approchée par une gaussienne.

La méthode de **Frühwirth-Schnatter & Frühwirth** [38] traite directement du cas de la régression multinomiale logistique. Pour ce faire, les auteurs décomposent les nombres $N_{i,j}^k$ sous la forme de $N_{i,*}^k$ variables binaires $z_{i,j,\ell}^k$, qu'on peut voir comme le vecteur de choix de chaque électeur ℓ du bureau k ayant opté pour la modalité i au premier tour. Ces choix sont donnés par la maximisation de variables latentes d'utilité $u_{i,j,\ell}^k$, qui indépendantes et identiquement distribuées suivant une loi Gumbel de paramètres $(\log\lambda_{i,j}^k, 1)$, donnent une représentation des probabilités logistiques $p_{j|i}^k$:

$$z_{i,j,\ell}^k = 1_{(u_{i,j,\ell}^k > u_{i,(j),\ell}^k)} \qquad u_{i,j,\ell}^k \overset{\text{iid}}{\sim} \mathcal{G}(\log\lambda_{i,j}^k, 1) \qquad u_{i,(j),\ell}^k = \max_{m\neq j} u_{i,m,\ell}^k \qquad P(z_{i,j,\ell}^k = 1) = p_{j|i}^k \tag{45}$$

Pour éviter l'introduction d'un nombre inconsidéré de variables latentes, les auteurs proposent de transformer ces utilités en des variables $v_{i,j,\ell}^k$ suivant des lois exponentielles et de les agréger en des variables $u_{i,j,*}^k$, indépendantes et identiquement distribuées, à une translation près, suivant une loi logistique de type III.

$$v_{i,j,\ell}^k = \exp(-u_{i,j,\ell}^k) \overset{\text{iid}}{\sim} \mathcal{E}(\lambda_{i,j}^k) \qquad u_{i,j,*}^k = -\log\sum_{\ell=1}^{N_{i,*}^k}\left[v_{i,j,\ell}^k - v_{i,1,\ell}^k\right] = \log\lambda_{i,j}^k + r_{i,j}^k \qquad r_{i,j}^k \overset{\text{iid}}{\sim} \mathrm{Lo}_3(N_{i,*}^k) \tag{46}$$

Cette loi étant centrée et fonction d'un unique paramètre discret, les auteurs en donnent une représentation paramétrée sous forme d'une mélange fini de lois normales centrées :

$$\mathrm{Lo}_3(N_{i,*}^k) \sim \sum_{q=1}^{Q} w_q\,\mathcal{N}(0, \sigma_q^2) \tag{47}$$

Le nombre de normales P, les poids w_q et les variances σ_q^2 sont tabulés jusqu'à $N_{i,*}^k \leqslant 60$. Entre 60 et 500, Q vaut 2, les variances σ_q^2 sont données par des fractions rationnelles fonctions de $N_{i,*}^k$, et les poids en sont déduits. Au delà, le résidu est directement approché par une unique loi normale.

Entre ces trois méthodes, la méthode d'Holmes & Held conduit à un calcul prohibitif. Les méthodes de Polson & Scott et de Frühwirth-Schnatter & Frühwirth semblent toutes deux pertinentes. Toutefois, la première n'est pas directement adaptée à notre problème et exige de considérer des itérations de Gibbs supplémentaires. En outre, elle semble conçue pour de petites valeurs $N_{i,*}^k$, ce qui n'est généralement pas notre cas. En conséquence, nous choisissons de mettre en œuvre celle de Frühwirth-Schnatter & Frühwirth. Pour la complétude du document, les détails de cette implémentation sont précisés en annexe au §A.4.1.

Outre cette méthode, comme les paramètres $\boldsymbol{\beta}_{i,j}$ évoluent sans contrainte dans un espace vectoriel, donc dans un espace géométriquement simple, nous pouvons effectuer leur simulation à partir d'une marche aléatoire contrôlée selon la méthode de Metropolis-Hastings. Comme l'espace est toutefois de grande dimension

et que le gradient suivant $\boldsymbol{\beta}_{i,j}$ de (42) s'exprime simplement, nous guidons, pour l'efficacité, ces marches aléatoires suivant la méthode **HMC** [39].

Cette technique consiste à définir une surface au-dessus de l'espace des paramètres à partir de l'opposé du logarithme de la distribution (42) et à y faire rouler une petite bille. Dans cette configuration, on assimile les jeux de paramètres $(\boldsymbol{\beta}_{i,j})$ à la position de la bille. L'algorithme consiste alors à lancer la bille, à partir du point courant, dans une direction aléatoire et avec une vitesse unitaire, et à regarder sa position au bout d'un temps fixe T, cette position donnant le jeu de paramètres candidat $(\boldsymbol{\beta}'_{i,j})$. Ce faisant, la bille soumise à la gravité est attirée par les déprécions de la surface, correspondant à des maxima locaux de la distribution (42). Du point de vue statistique, on ne considère plus uniquement la distribution des paramètres $(\boldsymbol{\beta}_{i,j})$, mais celle du couple (position, vitesse) = $(\boldsymbol{\beta}_{i,j}, \boldsymbol{v}_{i,j})$. Ce faisant, comme la direction initiale est tirée suivant une loi normale et que la bille est assimilée à un point de masse unitaire, la densité de ce couple s'écrit

$$f(\boldsymbol{\beta}_{i,2}, ..., \boldsymbol{\beta}_{i,J}, \boldsymbol{v}_{i,2}, ...\boldsymbol{v}_{i,J} \mid (N^k_{i,j})) \propto f(\boldsymbol{\beta}_{i,2}, ..., \boldsymbol{\beta}_{i,J} \mid (N^k_{i,j})) \times \exp\left(-\sum_{j=2}^{J} \frac{\|\boldsymbol{v}_{i,j}\|^2}{2}\right) \tag{48}$$

Autrement dit, cette densité correspond à l'exponentielle de l'opposé de l'énergie totale. Comme cette énergie est conservée le long de chaque trajectoire, la marche aléatoire ainsi définie est symétrique, et le ratio d'acceptation de Metropolis-Hastings, théoriquement égal à 1. En pratique, ce ratio varie car la trajectoire doit être calculée numériquement. Pour ce faire, la méthode utilise le schéma classique d'Euler saute-mouton, qui permet une assez bonne conservation de l'énergie totale. Les détails de cette implémentation sont également donnés en annexe au §A.4.2.

Notons que cette méthode nécessite de définir le temps de lancer T et le pas du schéma d'intégration ε. Le choix optimal de ces paramètres pour notre problème, ainsi qu'une comparaison des performances entre cette méthode et celle de Frühwirth-Schnatter & Frühwirth font l'objet d'une étude présentée §B.2 en annexe.

3 Mise en œuvre et validation de la méthode d'inférence

Entre la définition des stratégies de résolution et leur application sur les données, il existe un travail important d'élaboration et de validation sur le plan pratique. Le travail d'élaboration recouvre de nombreuses décisions essentielles pour la faisabilité des calculs et la pertinence des résultats. Ces décisions vont du choix des outils numériques, des algorithmes de simulation à celui des techniques de détection de la convergence. Les tests de validation sont tout aussi importants pour le crédit du travail présenté, et comprennent la vérification statistique des sorties des algorithmes de simulation, l'étude du comportement des stratégies de résolution sur des données simulées, celle de leur robustesse et de leur performance. Dans cette partie, nous précisons et discutons ces deux aspects.

3.1 Élaboration des stratégies de résolution

3.1.1 Choix et validation des méthodes numériques

Les problèmes considérés sont de très grande taille et exigent un volume important de calculs. Pour l'élection et le champ étudiés, on compte environ 64000 bureaux de vote et on considère jusqu'à 7 populations différentes. Dans ces conditions, il y a 64000 tableaux de contingence inconnus, de dimension au plus $14 \times 4 \times 7$, ce qui conduit, pour les plus grands problèmes, à devoir estimer une distribution *a posteriori* d'une dimension supérieure à 20 millions. Utilisant l'algorithme de Gibbs, les chaînes présentent des longueurs d'auto-corrélation importantes, de sorte que plusieurs centaines de milliers d'itérations de cet algorithme doivent être calculées pour obtenir des estimations suffisamment précises. Devant ces ordres de grandeur, l'efficacité des calculs est au cœur de nos priorités.

Si les travaux mathématiques de la partie 2.3 permettent d'accroître cette efficacité, le choix du langage de développement constitue également une source importante de gain. Les logiciels intégrés, tels `Matlab` ou `R`, donnent accès à de nombreuses fonctions statistiques pré-codées, mais présentent une gestion de la mémoire et un mécanisme d'interprétation du code coûteux pour de grands calculs. Nous avons donc préféré écrire le code dans un langage de plus bas niveau et avons choisi de le faire en `C++`. Ce langage a la réputation d'être particulièrement optimisé pour les opérations mathématiques et de permettre les calculs les plus rapides. En outre, son orientation objet permet de structurer et de mutualiser le code autour des concepts théoriques introduits dans la modélisation : modèles, tableaux de contingence, matrices ...

Ce faisant, nous avons dû mettre en place un module statistique comprenant un générateur aléatoire et des simulateurs de lois usuelles. Pour ce faire, nous avons choisi le Mersenne Twister MT19937 [47] pour le générateur et les algorithmes [48] pour la loi exponentielle, [2] pour la loi normale, [49] et [50] pour la loi gamma, et [45] pour la loi binomiale. Leur génération se fonde toute sur des méthodes d'inversion ou d'acceptation-rejet. En outre, par souci d'efficacité et de sécurité, les codes de ces algorithmes s'appuient sur des sources libres développées par le *R core team* pour le logiciel `R`.

Les codes de simulation des lois vectorielles normale, multinomiale ou de Dirichlet sont classiquement construits à partir d'appels répétés à ces algorithmes. Par exemple, la simulation selon une loi de Dirichlet utilise le fait qu'un vecteur de m variables aléatoires indépendantes de loi gamma $\Gamma(\alpha_1, 1)$, ..., $\Gamma(\alpha_m, 1)$, normalisé par sa norme L^1, suit une loi de Dirichlet de paramètres $(\alpha_1, ..., \alpha_m)$.

La simulation suivant une loi hypergéométrique décentrée de Fisher (37) est moins classique et se base sur les travaux [6] et le code en source libre [5] d'Agner Fog. Pour de petites valeurs des paramètres N'^k, $N'^k_{i,*}$, $N'^k_{*,j}$ et une pondération ω relativement proche de 1, l'approche calcule l'ensemble de la loi et procède suivant la méthode d'inversion. Pour les autres valeurs, comme la loi hypergéométrique décentrée de Fisher est unimodale et que les queues de distribution sont sous-quadratiques, l'auteur utilise la méthode du ratio d'uniformes proposée par Kinderman-Monahan [26] [15] consistant à envelopper la densité cible à partir d'une densité qu'on peut obtenir en tirant indépendamment deux lois uniformes, et à appliquer la méthode d'acceptation-rejet. Cette méthode nécessite très peu de calculs et conduit à des taux d'acceptation élevés.

Toutes les implémentations de ces algorithmes ont été testées statistiquement : le générateur avec la batterie de tests Diehard [19], les simulateurs de lois en comparant la distribution empirique obtenue et celle théorique par les tests de Kolmogorov-Smirnov pour les lois continues et du χ^2 pour les lois discrètes. Ces tests ont également été effectués sur l'algorithme décrit au §2.3.2, avec et sans contrôle par l'algorithme de Metropolis-Hastings, dans le cas de tableaux de contingence de petites dimensions et concernant un petit nombre d'électeurs (cf. §B.1 en annexe).

3.1.2 Choix des méthodes de contrôle de la convergence

Pour pouvoir exploiter les chaînes produites par le code et appliquer les résultats d'ergodicité, il nous faut classiquement éliminer la phase de *burn in* correspondant à l'amorce des chaînes qui présente un comportement transitoire et éloigné de l'état stationnaire. Sur la partie restante, il nous faut vérifier que les chaînes ont atteint un état suffisamment proche de celui stationnaire, qu'elles comportent suffisamment de points indépendants et que les modes les plus probables ont été visités. Pour ce faire, nous suivons les recommandations de [20], et mettons en œuvre quatre procédures de contrôle, à la fois visuelles et basées sur des tests statistiques. Pour ces derniers, nous utilisons l'implémentation faite dans la libraire `coda` de `R`.

Diagnostics graphiques
Une première procédure de contrôle consiste à représenter les sorties des chaînes simulées, comme dans les graphiques 3 et 4, afin de détecter d'éventuels comportements anormaux ou non stationnaires.

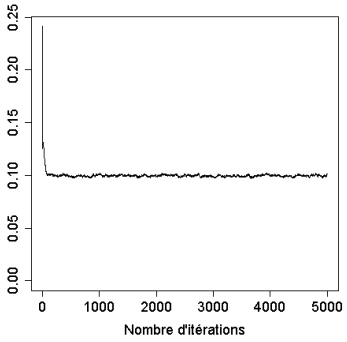

FIGURE 3 – Exemple d'une chaîne produite par le code

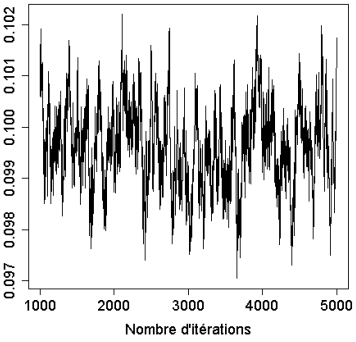

FIGURE 4 – Phase stationnaire de la chaîne du graphique 3

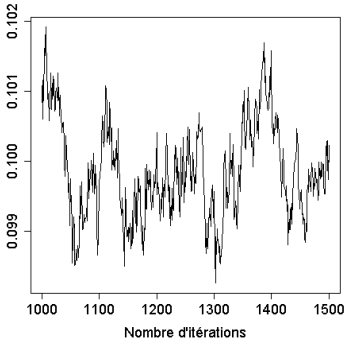

FIGURE 5 – Zoom sur l'état stationnaire du graphique 4

Test de la convergence vers une loi stationnaire
Plus rigoureusement, nous mettons en œuvre des tests non-paramétriques de la stationnarité d'une chaîne (ou d'un morceau de la chaîne). Plutôt qu'un test de Kolmogorov-Smirnoff, qui pose des problèmes d'interprétation, nous privilégions, comme le suggère [21], la méthode de Heidelberger & Welch (1981) [35]. Ce test se fonde sur la statistique de Cramer-von Mises qui approche la distance L_2 entre deux distributions. Si F_n est la distribution empirique, dont on cherche à tester l'adéquation avec une distribution théorique F , cette statistique s'écrit :

$$C = \int [F(x) - F_n(x)]^2 dF(x) \tag{49}$$

Dans l'implémentation du test, la chaîne $(x_1, ..., x_{N+M})$ est partitionnée suivant deux parties, qu'on peut noter $y = (y_1, ..., y_N) = (x_1, ..., x_N)$ et $z = (z_1, ..., z_M) = (x_{N+1}, ..., x_M)$. Puis les deux distributions (empiriques) de y et z sont comparées, selon une statistique construite en normalisant la variance des rangs :

$$T = \frac{U}{NM(N+M)} - \frac{4MN-1}{6(M+N)} \qquad \text{avec} \quad U = N\sum_{n=1}^{N}(r_n - n)^2 + M\sum_{m=1}^{M}(s_m - m)^2 \tag{50}$$

où $(r_1, ..., r_N)$ et $(s_1, ..., s_M)$ désignent les rangs des éléments des vecteurs y et z, respectivement. Si la valeur de la statistique de test T est plus grande que des valeurs tabulées, l'hypothèse d'équi-distribution des deux sous-échantilllons y et z est rejetée.

En outre, pour tenir compte de la présence en début des chaînes d'une phase transitoire, l'implémentation reproduit le test sur les chaînes amputées d'une amorce de 10%, 20%, ... jusqu'à 50% de chaque réalisation. A l'issue, la p-valeur du test et la longueur considérée de la phase de *burn in* sont indiquées.

Mesure de la taille effective et *thinning*
Dans le régime stationnaire, le caractère auto-régressif d'une chaîne simulée (voir graphique 5) induit l'existence d'une différence entre le nombre de points de la chaîne et le nombre de périodes indépendantes qu'elle comporte. Autrement dit, ce nombre, appelé « taille effective » de la chaîne correspond au nombre maximal de points indépendants qu'on peut extraire de la chaîne, et la sélection de ces points correspond à l'opération de *thinning*. Dans l'application des résultats d'ergodicité, seule cette taille importe. Aussi, afin de garantir la précision de nos résultats, est-il essentiel de l'estimer, ce qui peut se faire *e.g.* à l'aide d'auto-corrélogrammes.

Étude de la convergence des moyennes
Avec les procédures précédentes, rien n'indique que la chaîne ne reste pas « bloquée » dans une région particulière de l'espace des paramètres, éloignée de la (ou des) masse(s) principale(s) de la distribution cible.

En effet, dans les cas d'une forte attraction d'un mode local, la chaîne se comporte comme si elle était simulée selon la restriction de la « vraie » densité au voisinage de ce mode et conduit ainsi à un diagnostic de convergence positif. Lancer plusieurs chaînes avec des valeurs initiales dispersées permet de se prémunir contre ce genre de mauvaises conclusions. À partir de H chaînes simulées en parallèle $((x_t^{(1)})_{t=1,...T}, ..., (x_t^{(H)})_{t=1,...T})$, le critère d'arrêt de Gelman & Rubin [13] analyse leur variance, en la décomposant classiquement en la somme d'une variance « intra-chaînes » et d'une variance « inter-chaînes ». Dans le cas de convergence, la variance « inter-chaînes » traduit les différences d'initialisation, et devient négligeable à mesure que les chaînes convergent ensemble. Toutefois, la méthode est limitée, car ne permettant pas de détecter la présence de modes qui n'auraient pas été visités au cours des simulations.

3.2 Validation des stratégies de résolution

Pour la validation, nous menons trois types de test répondant successivement aux questions suivantes : Quelle est la fiabilité des estimations ? Quelle est la robustesse des inférences faites dans le cadre des modèles considérés ? Comment se comportent comparativement des méthodes alternatives ? Pour y répondre, nous considérons des jeux de données simulées tirés aléatoirement et pour lesquels les tableaux de contingence, les coefficients et les probabilités de vote exactes sont connus. En outre, dans le dépouillement des résultats, la stationnarité des chaînes et leur convergence des moyennes sont systématiquement vérifiées.

3.2.1 Comportement sur données simulées

Il s'agit premièrement de vérifier que pour les trois modèles, les procédures d'estimation appliquées à un jeu de données simulées permettent de retrouver les probabilités de vote ayant servi à constituer ce jeu.

Modèles à probabilités de report constantes
Pour ce premier modèle, le jeu de données est constitué de manière à ressembler au cas de l'élection présidentielle de 2007. On simule 50 000 bureaux, la distribution du nombre d'électeurs par bureaux est reproduite et les probabilités de report sont choisies de manière à ressembler aux résultats attendus (cf §B.3 en annexe). Sur ces simulations, l'état stationnaire est atteint assez rapidement et le *burn in* maximal est de l'ordre du millier de points. Globalement, les estimations donnent des probabilités de report légèrement biaisées. Pour environ la moitié des chaînes seulement, les intervalles de plus haute crédibilité à 95% contiennent la valeur ayant servi à la simulation. Les deux cas de figure sont illustrés sur les graphiques 6 et 7. Le détail de ces estimations est donné au §B.3 en annexe. Les biais sont toutefois à nuancer, puisqu'ils sont, dans le pire de nos cas de simulations, inférieurs à 2 points en écart absolu. En cherchant, leur origine semble être liée au calcul des probabilités des tableaux candidats et pourrait être due à l'accumulation d'approximations numériques.

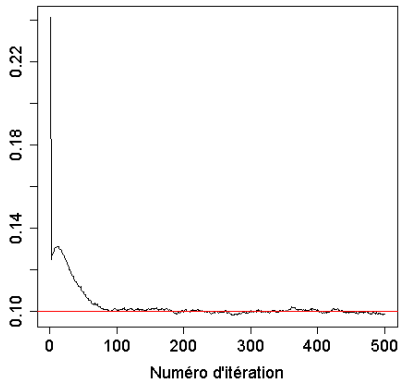

FIGURE 6 – **Estimation correcte** d'une probabilité de report (valeur exacte en horizontale)

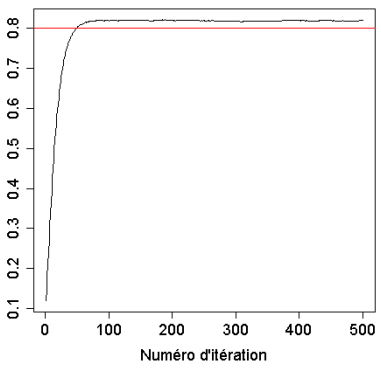

FIGURE 7 – **Estimation biaisée** d'une probabilité de report (valeur exacte en horizontale)

Modèles à probabilités de report logistiques et à probabilités de vote par population
Pour les modèles à probabilités de report logistiques et à probabilités de vote constantes par population, les probabilités et les variables socio-démographiques sont choisies aléatoirement. En outre, afin d'alléger les calculs, seuls 10 000 bureaux de vote sont considérés et on limite à deux le nombre d'explicatives pour le modèle logistique et le nombre de populations pour le modèle en population. Les estimations sur ces modèles présentent des variances plus élevées (voir graphiques 8 et 9, et §B.3 pour le détail des estimations) et masquent les éventuels biais observés dans le modèle précédent. Ce faisant, toutes les valeurs des coefficients $\boldsymbol{\beta}_{i,j}$ et les probabilités de vote $p_{i,j|\pi}$ sont toutes dans les intervalles de plus haute crédibilité à 95%. A noter que ces deux modèles nécessitent beaucoup plus d'itérations pour atteindre un état quasi stationnaire et pour obtenir la même taille effective que dans le modèle à probabilités constantes. Notamment, le modèle en population conduit à des longueurs d'auto-corrélation particulièrement élevées, ce qui montre le caractère faiblement mélangeant. Ces observations sont vraisemblablement à rapprocher de la stratégie de simulation des tableaux de contingence qui fixent un ordre précis de traitement des marges.

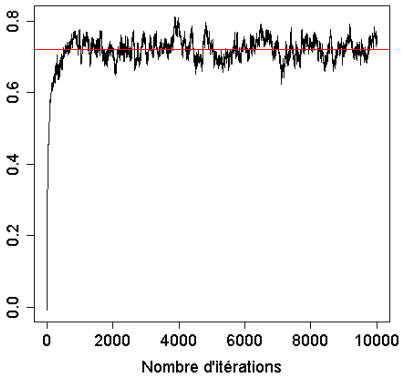

FIGURE 8 – Estimation d'un coefficient logistique (valeur exacte en horizontale) : les chaînes sont **plus dispersées et moins mélangeantes**

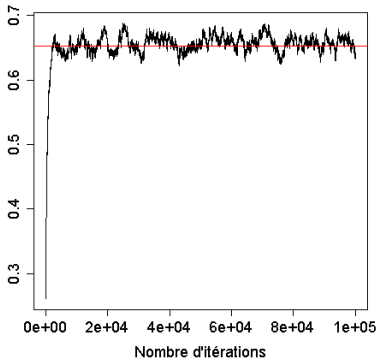

FIGURE 9 – Estimation d'une probabilité de vote (valeur exacte en horizontale) : les chaînes sont **peu mélangeantes** et mettent plus de temps à converger

3.2.2 Étude de la robustesse

Les modèles considérés font l'hypothèse d'un comportement identique pour les électeurs d'un même bureau de vote, voire pour plusieurs ou tous les bureaux. En outre, les inférences s'appuient sur des données agrégées correspondant aux résultats électoraux du premier et du second tour. Dans ces conditions, se pose la question de la présence d'un **biais écologique**. Pour répondre à cette question, nous considérons deux populations d'électeurs aux comportements électoraux distincts pour simuler les jeux de données. Pour simplifier, nous supposons en outre que chaque bureau de vote comporte un seul type de population. Chacune de ces populations présente un jeu de probabilités de report ou de coefficients logistiques qui lui est propre et qui est choisi de manière aléatoire. Les simulations des résultats électoraux obtenus sont agrégées dans un unique fichier que nous soumettons aux procédures d'estimation du modèle à probabilités de report constantes et du modèle à probabilités de report logistiques.

Ce faisant, nous observons globalement que les estimations des probabilités de report et des coefficients logistiques se situent entre les deux valeurs ayant servi à la simulation des données (cf figures 10 et 11, et §B.3 pour le détail des estimations), ce qui suggère une certaine robustesse des inférences effectuées dans le cadre des modélisations à probabilités de report constantes et logistiques.

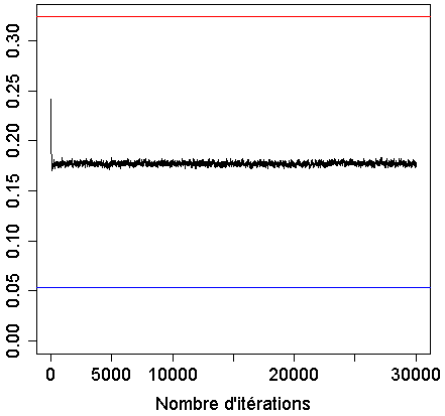

FIGURE 10 – Estimation d'une probabilité de report constante avec deux populations (valeurs exactes en horizontale)

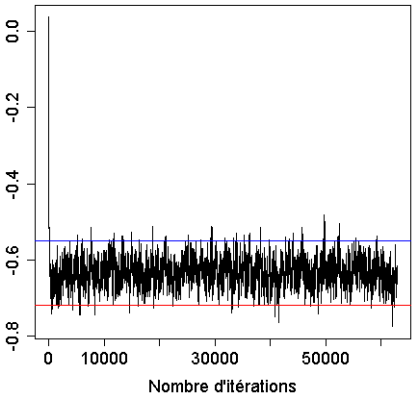

FIGURE 11 – Estimation d'un coefficient logistique avec deux populations (valeurs exactes en horizontale)

3.2.3 Étude de la performance

Dans la littérature, nous avons vu, au paragraphe 2.2, que les problèmes écologiques, semblables à ceux que nous traitons, sont traités soit par des méthodes numériques bayésiennes, soit par des approches semi-paramétriques s'appuyant sur des équations de moments. Dans ce contexte, comme nous avons opté pour la famille bayésienne, nous pouvons nous interroger sur la pertinence de ce choix et comparer les résultats obtenus par une méthode de la famille alternative. Pour ce faire, nous considérons l'une de ces approches basée sur les équations de moment d'ordre 1 et pour laquelle la convergence asymptotique des estimateurs est théoriquement établie. Les détails de cette méthode sont précisés au §B.4 de l'annexe. Son application sur les données simulées pour le modèle à probabilités constantes permet de retrouver les valeurs de simulation avec une précision inférieure à $2\ 10^{-4}$. Néanmoins, nous supposons que cette précision remarquable se fait au prix d'une moindre robustesse des estimations.

4 Résultats des estimations sur données réelles

4.1 Appariement de deux sources de données

Les données exploitées dans le cadre de cette analyse statistique proviennent principalement de deux sources. D'une part, les **données électorales** mobilisées ont été collectées sur la plate-forme gouvernementale de mise à disposition de données publiques [1] par un des encadrants du projet R. Ryder. [9] Ces données produites par le ministère de l'Intérieur sont constituées des résultats détaillés des deux tours de l'élection présidentielle de 2007 au niveau de chaque bureau de vote du territoire [10]. Pour chaque bureau de vote et pour chaque tour de scrutin, y sont reportés un identifiant de la commune, un identifiant du bureau, le nombre d'électeurs inscrits, le nombre de votants, le nombre de votes blancs et nuls, le nombre de suffrages exprimés ainsi que le nombre de suffrages recueillis par chacun des candidats en compétition.

D'autre part, des **données socio-démographiques** ont été intégrées aux données électorales. Ces données complémentaires sont issues d'un jeu de données intégralement construit par une équipe d'universitaires à l'occasion d'un projet de recherche. Ce projet intitulé « Cartelec »[43][41] a pour objectif de développer un outil performant d'analyse de la sociologie et de la géographie électorales au niveau le plus fin possible, à savoir le bureau de vote. Ce jeu de données comprend un identifiant du bureau et de la commune ainsi que la

9. Pour une raison inconnue, ces données ne sont plus à libre disposition sur le site actuellement.
10. Pour rappel, le champ de l'étude est restreint aux départements de France métropolitaine.

distribution pour la population des bureaux de plusieurs variables qualitatives ou ayant été discrétisées : sexe, âge, niveau de diplôme, catégorie socioprofessionnelle, statut d'occupation du logement (propriétaire, locataire du parc privé...), situation vis-à-vis du marché du travail et ancienneté résidentielle [11]. Pour construire ces données, l'équipe de Cartelec a mené au niveau des Iris [12] un travail minutieux de rapprochement des résultats des recensements de la population conduits par l'Insee avec la géographie précise des bureaux de vote.

Pour faciliter leur exploitation, ces deux sources de données ont été **appariées** en une seule. Ce travail a néanmoins posé **quelques difficultés** car les données socio-démographiques issues de Cartelec sont incomplètes. Elles ne comprennent en effet pas tous les bureaux de vote. Lorsque les données de Cartelec étaient disponibles pour tous les bureaux de vote d'une commune donnée, les observations des deux sources ont été appariées par bureau. En revanche, lorsque les données socio-démographiques n'étaient renseignées qu'au niveau de la commune pour les communes comportant plusieurs bureaux, les observations ont été appariées par commune. Ce choix a impliqué d'agréger les données électorales associées au niveau communal. Par ailleurs, les quelques communes dont aucune donnée socio-démographique n'était disponible ont été exclues du champ des études faisant intervenir des éléments de nature sociologique ou démographique. La déformation de la distribution du nombre d'inscrits par bureaux de vote par ces opérations est étudiée au §C.2

4.2 Modélisations mises en œuvre

Une démarche « naïve » pour l'estimation des reports de votes repose sur l'hypothèse que les comportements de report sont homogènes entre les bureaux de vote (cf §2.1). Si cette hypothèse est nécessaire à la mise au point d'une stratégie d'estimation, sa crédibilité sur le territoire national est sujette à caution. La validité de cette hypothèse est en effet *a priori* contestable en raison de deux enseignements bien établis et documentés en science politique. En premier lieu, nombre d'études de la littérature en science politique attestent que les comportements électoraux sont corrélés à des caractéristiques individuelles telles que le niveau de diplôme ou le niveau de revenu. Il est donc raisonnable d'envisager que ces caractéristiques aient également une influence sur les comportements de report eux-mêmes. Si ce lien entre reports de votes et composition sociale est avéré, les reports de votes ne sont pas identiques selon les quartiers ou les villages puisque la population n'est pas répartie de manière totalement aléatoire sur le territoire national. Par ailleurs, la science politique a depuis longtemps [13] mis en évidence qu'une permanence géographique existait dans les comportements électoraux. Par exemple, le sud-ouest de la France penche majoritairement à gauche tandis que l'Alsace est ancrée à droite depuis plusieurs décennies. Ces permanences observées au niveau local tiennent à des particularités culturelles et historiques. Certains auteurs en avancent même des explications d'ordre anthropologique *e.g.* [22]. Il s'ensuit que les comportements de reports de votes sont eux aussi susceptibles de varier selon les zones géographiques, à l'instar de ce qui est ainsi observé pour les préférences partisanes.

Ce faisant, **deux démarches visant à assouplir l'hypothèse d'homogénéité des reports de votes** ont été adoptées. D'une part, une **stratification électorale** du territoire français a été construite. Cette stratification vise notamment à établir les contours de zones géographiques de culture politique homogène au sein desquelles l'hypothèse de stabilité des reports de votes serait plus plausible. Il aurait été possible de simplement estimer les reports de votes en segmentant le territoire français par régions administratives. Néanmoins, cette démarche n'a pas été retenue car les limites des régions administratives ont été fixées sur la base de critères indépendants de la géographie électorale. En outre, elles sont trop nombreuses et certaines d'entre elles comportent un nombre limité de bureaux de votes susceptible de fragiliser la qualité des estimations.

D'autre part, la construction de certains estimateurs mobilisés par la suite fait intervenir les données socio-démographiques recueillies grâce au projet Cartelec dans un modèle de report de votes à **probabilités logistiques** (cf. §2.1.2) ou par **populations** (cf. §2.1.3).

11. depuis combien de temps les personnes habitent la zone couverte par le bureau de vote
12. « Ilots regroupés pour l'information statistique » : il s'agit d'unités territoriales infra-communales délimitées par l'Insee.
13. André Siegfried a été un précurseur en la matière, notamment avec son ouvrage de 1913 [9].

4.2.1 Construction de zones géographiques de culture politique homogène

La partition des bureaux de vote a été opérée suivant deux directions complémentaires. La première est géographique et consiste à former des zones connexes du territoire ayant des comportements électoraux homogènes. La seconde joue sur la taille de la commune et sa distance à un grand centre urbain.

Classification des départements

La constructions des zones de culture politique homogène tient compte de deux objectifs complémentaires. Le premier est logiquement que les électeurs aient des comportements électoraux observés (participation électorale et préférences partisanes) relativement proches au sein de chaque zone mais relativement hétérogènes entre les zones. Le second objectif est que les zones ainsi constituées soient connexes, c'est à dire que leur limite extérieure puisse être représentée par une seule courbe continue. Ce second objectif est destiné à éviter la formation de zones chaotiques n'ayant aucune pertinence culturelle et électorale.

D'un point de vue technique, les zones sont constituées à l'aide d'une classification ascendante hiérarchique. Afin que cette construction ne soit pas trop lourde, nous ne considérons pas directement les bureaux de vote mais les départements. Cette approximation, tout en présentant un grain de finesse encore suffisant, facilite considérablement la description géographique du voisinage des éléments à classifier, et est d'autant plus valable que les zones constituées sont vastes.

La distance utilisée pour la classification agrège à la fois la distance entre départements en termes de profil électoral et leur proximité géographique. D'une part, la distance « électorale » est calculée en prenant la distance de Mahalanobis [14] entre les profils électoraux des départements. Ces profils sont définis au niveau départemental et sont constitués des scores agrégés de chacun des candidats du premier tour de l'élection présidentielle de 2007, de la proportion de votes blancs et nuls [15] et du taux de participation. Nous n'avons pas tenu compte des résultats du second tour, car à scores du premier tour donnés, ils sont évidemment corrélées aux reports de votes. La présence d'une telle endogénéité dans le découpage exclurait par la suite toute possibilité de donner une dimension explicative aux différences locales de culture politique, ce qui n'est pas souhaitable. D'autre part, la proximité géographique est définie, pour chaque paire de départements, par le nombre minimal de limites départementales à franchir pour aller d'un département à l'autre. Enfin, la distance de classification est obtenue par une somme pondérée de ces deux distances. La pondération est calibrée par tâtonnements de manière à minimiser le poids relatif de la composante géographique tout en permettant de former des zones connexes. Le nombre de zones à construire est fixé en examinant le gain marginal à l'ajout ou au retrait d'une zone. Cette procédure nous conduit à distinguer neuf grandes régions de culture politique homogène, à la fois connexes et vastes (cf. carte 12 et §C.1 en annexe).

Segmentation pour prendre en compte la localisation résidentielle

Au-delà des caractéristiques individuelles des électeurs et des permanences géographiques, la localisation résidentielle au sein d'une région donnée est également parfois mise en avant comme facteur corrélé aux choix électoraux. De cette façon, les comportements électoraux ne sont pas identiques en milieu urbain et rural. Ainsi, suivant l'idée que si les choix électoraux diffèrent, les reports de votes peuvent aussi varier, nous complétons le découpage du territoire français en neuf zones en distinguant les communes de plus 100 000 habitants des autres.

Toutefois, la nature de ce lien est plus ambiguë et peut refléter la combinaison de plusieurs mécanismes.

14. Cette distance est fondée sur les corrélations existant entre les valeurs prises par les variables associées aux individus, ici les départements.

15. Le dénominateur de cette proportion est constitué du nombre d'électeurs inscrits. Pour rappel, les scores des candidats sont définis par le rapport du nombre de suffrages qu'ils ont recueillis sur le nombre total de suffrages exprimés, c'est à dire à l'exclusion des votes blancs et nuls.

En premier lieu, de par les dynamiques de ségrégation sociale et spatiale qui peuvent exister, le lieu de résidence des électeurs est parfois lié à leurs caractéristiques individuelles, en particulier à leur niveau de revenu. En effet, la présence intensive de ménages aisés en un lieu donné exerce une pression à la hausse sur les prix (à l'achat aussi bien qu'à la location) du marché immobilier local qui tend à évincer les ménages plus modestes. De fait, plusieurs grandes villes françaises, en particulier Paris, sont le théâtre d'un processus dit de « *gentrification* ». Or, l'incidence électorale de telles dynamiques est bien établie [16]. Ces mécanismes de ségrégation spatiale demeurent cependant difficiles à capter à l'aide de caractéristiques territoriales telles que la taille des communes ou leur distance au centre urbain le plus proche. Ils ne seront donc traités que très partiellement sous l'angle territorial dans ce qui suit. Par ailleurs, le choix du lieu de résidence est guidé par un certain nombre d'arbitrages qui sont susceptibles d'être motivés par les préférences politiques. Ainsi, des électeurs qui accordent plus d'importance à leur confort matériel individuel (taille et qualité du domicile, jouissance d'une voiture, ...) ou à l'opportunité de constituer un patrimoine immobilier qu'au fait de bénéficier à leur proximité immédiate de réseaux de transports en commun ou d'une forte densité d'aménités urbaines sont susceptibles d'être plus sensibles à des discours de droite (réduction de la pression fiscale, soutien à l'accession à la propriété, ...). Or, compte tenu de la mobilité des populations plus soutenue au sein des régions qu'entre les régions et étant donnée la disponibilité de foncier plus abondante en dehors des villes, ce type d'électeurs peut tendre à s'installer en milieu péri-urbain voire rural. Enfin, une autre piste d'explication du lien entre le degré d'urbanisation de l'environnement immédiat des électeurs et leurs comportements électoraux n'est pas à exclure bien qu'elle soit un peu plus audacieuse. Il est ainsi possible que leur vie quotidienne et les rencontres qu'ils expérimentent façonnent les convictions des électeurs. Dans ce cas, les comportements électoraux différenciés observés entre les villes et les campagnes peuvent aussi résulter de réalités et de préoccupations qui diffèrent entre ces milieux.

Dans ces conditions, nous estimons, dans le cadre du modèle du §2.1, des probabilités de report sur l'ensemble du territoire métropolitain, sur chacune des neuf zones du découpage et puis sur chacune des intersections entre les zones du découpage territorial et les segments de communes.

4.2.2 Variables exogènes pour l'estimation de probabilité de report logistiques

Une modélisation alternative à cette partition est donnée par le modèle à probabilités de report logistiques (cf. §2.1.2). Dans le cadre de ce modèle, nous expliquons les probabilités de report de chaque bureau par plusieurs caractéristiques socio-démographiques attachées aux bureaux. Suivant la littérature sociologique et tenant compte des variables mobilisables, nous retenons, comme explicatives, les proportions, parmi les habitants de la zone couverte par le bureau de vote, de mineurs, de retraités, d'étrangers et de propriétaires occupant leur logement. Outre ces proportions, nous considérons également le taux de chômage au niveau départemental, de manière à tenir compte de l'impact géographiquement diffus de cette variable. Enfin, sont insérées en explicatives la ventilation des niveaux de formation initiale de la population résidant sur la zone couverte par le bureau de vote ainsi que l'appartenance à une grande région de culture politique homogène définie au §4.2.1.

4.2.3 Variables exogènes pour la définition des « populations » d'électeurs

Une dernière modélisation est obtenue dans le cadre du modèle du §2.1.3 en définissant des « populations » d'électeurs. Ces partitions du corps électoral sont définies à partir de distributions de variables socio-économiques fournies par le projet Cartelec sur l'ensemble des habitants de la zone couverte par chaque bureau de vote. La modélisation est approximative et assimile ces distributions concernant les habitants à celles concernant les électeurs. Ce faisant, nous considérons différentes partitions du corps électoral : suivant

16. L'incidence électorale de ce type de processus en milieu urbain est d'ailleurs plus ambiguë que d'aucuns pourraient le croire. Par exemple, la hausse de revenu moyen des parisiens s'est accompagnée d'une progression électorale des formations politiques de gauche, en particulier dans l'est de la capitale.

des tranches d'âge [17], suivant le niveau de formation atteint [18], suivant les catégories socio-professionnelles [19] et suivant le fait que les électeurs sont propriétaires ou locataires.

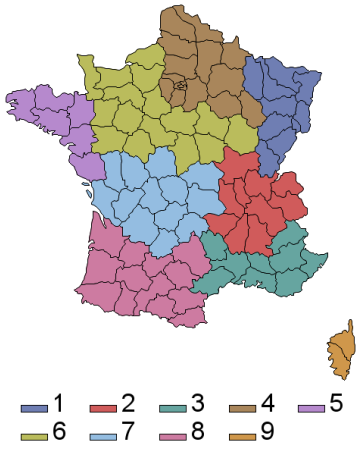

FIGURE 12 – Zones de culture politique homogène

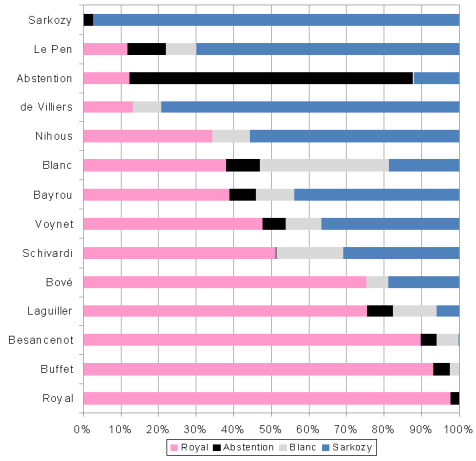

FIGURE 13 – Reports de vote au niveau de la métropole

4.3 Résultats

4.3.1 Estimation au niveau de la métropole

Une estimation a d'abord été produite au niveau de la métropole (cf. figure 13 et §C.3.1 en annexe). D'après cette estimation, les électeurs ayant voté au premier tour pour l'un des deux candidats qualifiés au second tour confirment comme attendu presque unanimement leur choix. Ainsi, les électeurs de Ségolène Royal sont 98% à rééditer leur vote tandis que ceux de Nicolas Sarkozy sont 97% dans ce cas. Leurs quelques électeurs résiduels se sont pour l'essentiel abstenus au second tour.

Les **électeurs de François Bayrou** sont quant à eux très partagés. Ils accordent néanmoins un léger avantage à Nicolas Sarkozy : 44% contre 39% pour Ségolène Royal. Les électeurs de François Bayrou sont en outre 7% à s'abstenir au second tour et 10% à émettre un vote blanc ou nul. Le report des électeurs de François Bayrou atteste que l'électorat de ce dernier est bien composite. En dépit de son histoire chrétienne-démocrate, le candidat centriste a recueilli les suffrages de citoyens de sensibilité de gauche. Pour autant, l'électorat de François Bayrou demeure hétérogène et Ségolène Royal ne parvient pas à en attirer une majorité à elle. A cet égard, la démarche de séduction électorale que cette dernière a entreprise entre les deux tours en direction des électeurs centristes se solde par un échec relatif. Compte tenu de la faiblesse des scores cumulés des candidats de gauche au premier tour [20], la candidate socialiste aurait en effet eu besoin de reports de votes bien plus favorables de la part des électeurs de François Bayrou pour l'emporter au second tour face à Nicolas Sarkozy. Le report des électeurs de François Bayrou constitue donc une clé décisive pour l'issue du scrutin.

Pour ce qui concerne à présent les **électeurs de Jean-Marie Le Pen**, ils se reportent massivement sur Nicolas Sarkozy, pour plus de deux tiers d'entre eux : 70%. Bien qu'il ait vraisemblablement déjà rallié à lui dés le premier tour une fraction notable d'électeurs proches de l'extrême-droite, Nicolas Sarkozy bénéficie

17. de 18 à 24 ans, de 25 à 39, de 40 à 54, de 55 à 64, de 65 à 79, et plus de 80 ans

18. sans diplôme, titulaire d'un CEP, d'un BEPC, d'un CAP ou d'un BEP, bachelier, ayant suivi des études supérieures jusqu'à bac + 2, ou au delà

19. agriculteurs, artisans et commerçants, cadres et professions libérales, professions intermédiaires, employés, ouvriers

20. Le score cumulé de tous les candidats de gauche au premier tour (en y incluant Ségolène Royal, Dominique Voynet, José Bové, Marie-George Buffet, Olivier Besancenot, Gérard Schivardi et Arlette Laguiller) s'élève en effet à près de 36%. A titre de comparaison, le score de la gauche était de l'ordre de 43% au premier tour de 2002.

d'un report de votes très favorable de la part des électeurs de Jean-Marie Le Pen. Ce report est un véritable succès pour Nicolas Sarkozy à deux égards. D'une part, des divergences sérieuses existent entre les options défendues par les deux candidats, notamment en ce qui concerne la construction européenne. D'autre part, Jean-Marie Le Pen a explicitement appelé ses électeurs à s'abstenir et à ne pas voter pour Nicolas Sarkozy au second tour [21]. Il apparaît ainsi que l'électorat de Jean-Marie Le Pen adopte un comportement électoral qui est politiquement structuré par une proximité partisane avec la droite. Nicolas Sarkozy étant à l'époque l'un des ministres les plus importants, ce résultat plaide pour appréhender le vote en faveur du Front National comme un vote d'adhésion et non comme un vote de protestation même s'il convient de rester prudent en l'absence d'analyses complémentaires. Outre Nicolas Sarkozy, les électeurs de Jean-Marie Le Pen sont 12% à voter pour Ségolène Royal, 8% à émettre un vote blanc ou nul et seulement 10% à se conformer à la consigne de leur candidat, à savoir ne pas voter. Bien que ce ne soit qu'un facteur secondaire du dénouement de l'élection compte tenu de la contre-performance relative de Jean-Marie Le Pen au premier tour, la forte attraction électorale exercée par Nicolas Sarkozy sur les électeurs de Jean-Marie Le Pen amplifie la victoire de ce dernier [22].

Le report des voix des **autres candidats** du premier tour est d'intérêt relativement marginal vu leur faible score. Dans la plupart des cas, les voix de ces candidats se reportent en fonction des proximités idéologiques qui existent avec les finalistes du second tour. Ainsi, les électeurs de « petits candidats »de gauche se reportent sur Ségolène Royal. Tel est de fait le cas de plus de 75% des électeurs d'Arlette Laguiller, d'Olivier Besancenot, de José Bové et de Marie-George Buffet [23]. *A contrario*, 79% des électeurs de Philippe de Villiers votent pour Nicolas Sarkozy au second tour. Les deux seuls candidats dont les reports de votes estimés ici peuvent paraître surprenants sont Dominique Voynet et Frédéric Nihous. Dominique Voynet est la candidate du parti écologiste « Les Verts »qui a déjà conclu plusieurs accords électoraux avec le Parti Socialiste par le passé. Pourtant, ses électeurs ne sont que 48% à voter pour Ségolène Royal au second tour. Ils sont en outre 36% à choisir Nicolas Sarkozy, ce qui est loin d'être négligeable. Quant à Frédéric Nihous, il est le candidat d'une formation politique dénommée « Chasse, pêche, nature et traditions »(CPNT). Ce nom laisse penser qu'il s'agit d'un parti de droite car il fait référence aux traditions qui constituent habituellement un marqueur idéologique de la droite. La défense des traditions reflète en effet un positionnement relativement conservateur [24]. Ce préjugé concernant le positionnement politique de CPNT apparaît de plus conforté par l'accord politique scellé en 2010 entre le parti de Frédéric Nihous et l'UMP[36]. Pour autant, le report des électeurs de Frédéric Nihous sur Nicolas Sarkozy n'est pas massif. Ils votent certes à 55% pour Nicolas Sarkozy mais Ségolène Royal rallie tout de même à elle 35% de ces électeurs. A titre plus anecdotique, il est à noter que les électeurs de CPNT ont un sens civique très développé : quasiment aucun d'entre eux ne s'abstient au second tour.

Enfin, plus des trois quarts des **abstentionnistes** du premier tour adoptent la même attitude au second tour. Pour le reste, les autres abstentionnistes accordent un très léger avantage à Ségolène Royal sur Nicolas Sarkozy : 13% contre 12%. La proportion d'abstentionnistes qui se mobilisent au second tour pour déposer un bulletin blanc ou nul dans l'urne est quant à elle marginale.

21. Il a communiqué cette consigne de (non-)vote à ses électeurs à l'occasion de son traditionnel discours du 1er mai en l'honneur de Jeanne D'arc.

22. Il s'agit d'un facteur secondaire du dénouement à résultats du premier tour donnés. En revanche, la capacité de Nicolas Sarkozy à rassembler derrière lui toutes les traditions de la droite française (droite orléaniste, droite bonapartiste et droite catholique) est souvent mise en avant pour expliquer son succès et notamment son score du premier tour. Tel est notamment le cas dans [44]

23. En revanche, le report des électeurs de Gérard Schivardi est plus diversifié : 51% vers Ségolène Royal et 30% vers Nicolas Sarkozy.

24. Dans [34], Frédéric Nihous prétend officiellement n'appartenir à aucun camp lorsqu'il est interrogé à ce sujet par des journalistes au cours de la campagne. Il est toutefois permis de douter de la sincérité de telles déclarations qui sont courantes, même de la part de mouvements dont le positionnement est pourtant clairement établi. Ainsi, le Front National affirme n'être ni de droite ni de gauche alors que la plupart des observateurs de la vie politique s'accordent à le placer à l'extrême-droite de l'échiquier politique.

Les estimations commentées ici diffèrent quelque peu des enquêtes d'opinion menées par les instituts de sondage auprès des électeurs, notamment l'enquête post-présidentielle 2007 du Cevipof[25]. Contrairement aux résultats présentés précédemment, le sondage du Cevipof conclut que les électeurs de François Bayrou se sont davantage reportés sur Ségolène Royal, bien que les ordres de grandeur soient compatibles : 49% pour Ségolène Royal contre 38% pour Nicolas Sarkozy. En revanche, les enseignements du sondage convergent plutôt avec les estimations produites par l'algorithme pour ce qui est du report des électeurs de Jean-Marie Le Pen, à l'exception de l'abstention et des votes blancs et nuls qui sont moindres d'après le sondage. Par ailleurs, le sondage affiche plus de déperditions des « petits candidats »vers les finalistes de leur camp, notamment pour José Bové et Arlette Laguiller. Au contraire, les reports des électeurs de respectivement Dominique Voynet et Frédéric Nihous vers respectivement Ségolène Royal et Nicolas Sarkozy seraient à en croire l'enquête du Cevipof relativement nets : respectivement 77% et 70%.

4.3.2 Estimations par grandes régions politiques

Les estimations par grandes régions de culture politique homogène font apparaître des écarts inter régionaux dans les reports de votes qui restent modestes. Elles permettent néanmoins d'affiner légèrement les enseignements extraits de l'estimation nationale, en particulier au sujet des électorats de François Bayrou et de Jean-Marie Le Pen (cf. figures 14 - 15 et §C.3.2 en annexe).

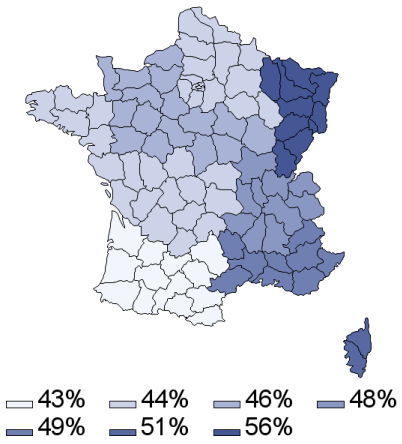

FIGURE 14 – Probabilité de report de F. Bayrou vers N. Sarkozy par régions politiques

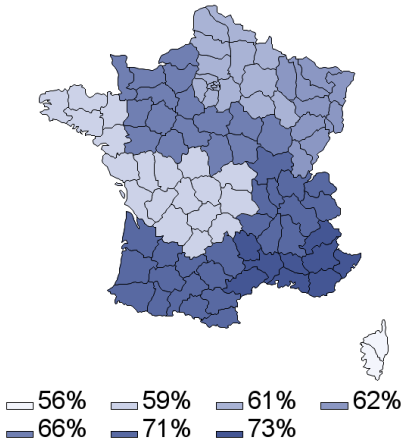

FIGURE 15 – Probabilité de report de J.M Le Pen vers N. Sarkozy par régions politiques

Les reports des électeurs de François Bayrou présentent peu de variations selon les zones à l'exception du nord-est de la France. La proportion de ces électeurs qui se reportent sur Nicolas Sarkozy varie de 43% à 51% suivant ces huit zones. Il est en revanche un peu plus élevé dans le nord-est : 56%. De même, le report des électeurs centristes sur Ségolène Royal s'avère assez stable selon les zones. Le nord-est et la Corse mis à part, il s'échelonne en effet de 35% à 43%. Il s'établit par ailleurs à 29% en Corse et à 30% dans le nord-est. De manière générale, un contraste se dégage entre l'est et l'ouest de la France en ce qui concerne l'avantage relatif accordé par les électeurs du candidat centriste à Nicolas Sarkozy. La différence entre les reports centristes en direction des finalistes est en effet moindre à l'ouest où elle n'excède jamais 10 points, au contraire de ce qui est constaté dans l'est.

Pour ce qui concerne les électeurs de Jean-Marie Le Pen, ils semblent se reporter sur Nicolas Sarkozy dans des proportions qui changent un peu d'une zone à l'autre. Le report des électeurs du leader frontiste sur

25. Les reports de votes estimés à partir de cette enquête sont notamment reproduits dans la publication qui suit [14].

Nicolas Sarkozy est de l'ordre de 70% dans le sud de la France tandis qu'il s'établit à un niveau plus proche de 60% au nord. De même, les reports de Jean-Marie Le Pen sur Ségolène Royal présentent également quelques disparités géographiques. La zone la moins favorable à Ségolène Royal de ce point de vue est la zone sud-est où elle ne recueille les suffrages que de 10% des électeurs de Jean-Marie Le Pen. A l'opposé, elle en recueille jusqu'à 26% dans la zone nord-ouest.

Globalement, la faible ampleur de ces variations géographiques suggère que l'hypothèse d'homogénéité des comportements de report exploitée pour l'estimation nationale est somme toute raisonnable.

La distinction entre villes de plus de 100 000 habitants et autres communes permet d'identifier les différences de comportements de report entre les électeurs de milieu urbain et les autres (cf. §C.3.3 en annexe). Là aussi les variations restent limitées. Les reports des électeurs de François Bayrou sur Nicolas Sarkozy sont dans la plupart des grandes régions moindres en milieu urbain qu'ailleurs. Ce déficit en milieu urbain bénéficie dans le nord-est ainsi que dans les parties sud et ouest du bassin parisien à Ségolène Royal. En revanche, il se traduit par davantage d'abstention dans les grandes villes franciliennes, du nord et du sud-est. Il est plus difficile de tirer des conclusions des disparités de report des électeurs frontistes entre grandes villes et communes plus rurales. Les électeurs urbains de Jean-Marie Le Pen se reporter légèrement plus que les autres sur Ségolène Royal dans toute la partie centrale de l'ouest de la France, autour des Deux-Sèvres, terre d'élection de la candidate socialiste. Ils se reportent au contraire plus souvent sur Nicolas Sarkozy dans le nord-est ainsi que dans le sud-ouest. Enfin, l'électorat urbain de Jean-Marie Le Pen apparaît un peu plus abstentionniste dans le sud-est et au nord de la France.

4.3.3 Estimation par proportions

Afin de tenir compte de l'éventuelle incidence des caractéristiques socio-économiques sur les comportements de report, une technique d'estimation alternative a été mise au point. Cette technique mobilise notamment les distributions de variables socio-économiques fournies par Cartelec. Elle permet d'estimer des reports de votes pour plusieurs sous-populations données telles que les jeunes de 18 à 24 ans ou les ouvriers. Les estimations qui en sont issues sont néanmoins à appréhender avec précaution car elles sont susceptibles d'être entachées d'un biais écologique. Plus précisément, elles conduisent par exemple à conclure que les jeunes électeurs adoptent davantage un comportement de report donné alors qu'en réalité, elles peuvent traduire un comportement différent de l'ensemble des électeurs des bureaux de vote comportant beaucoup de jeunes électeurs. Il est néanmoins par la suite supposé que ce biais écologique est négligeable.

D'un point de vue qualitatif, plusieurs enseignements émergent de la lecture des estimations produites par cette technique alternative (cf. §C.3.4 en annexe).

Tout d'abord, en termes d'âge, les jeunes électeurs de François Bayrou tendent à se reporter davantage sur Ségolène Royal tandis que ses électeurs plus âgés sont plus attirés au second tour par Nicolas Sarkozy. Pour ce qui est des électeurs de Jean-Marie Le Pen, ceux qui sont relativement âgés semblent se reporter assez massivement sur Nicolas Sarkozy alors que les plus jeunes se dispersent davantage, notamment entre l'abstention et le vote en faveur de Ségolène Royal.

Ensuite, du point de vue des catégories socioprofessionnelles, les agriculteurs ayant voté pour François Bayrou au premier tour se reportent plus nettement qu'en moyenne sur Nicolas Sarkozy, de même que les ouvriers. En revanche, le report des voix de François Bayrou est plus équilibré entre les deux finalistes pour les cadres et les professions intermédiaires. Paradoxalement, les ouvriers et les agriculteurs qui se reportent plus souvent sur Nicolas Sarkozy lorsqu'ils ont voté pour François Bayrou se reportent en parallèle en légèrement plus forte proportion que la moyenne sur Ségolène Royal quand ils ont choisi Jean-Marie Le Pen au premier tour. Quant aux artisans et commerçants, ils se reportent massivement sur Nicolas Sarkozy, qu'ils aient voté au premier tour pour François Bayrou ou pour Jean-Marie Le Pen.

Par ailleurs, aucune conclusion sérieuse ne peut être extraite des estimations par niveaux de formation initiale. Lorsqu'elles ne souffrent pas d'un manque de précision important, ces estimations indiquent des reports intégraux sur l'un des finalistes, ce qui n'est pas très crédible. La seule tendance qui s'en dégage est que, parmi les électeurs de François Bayrou, les moins diplômés tendent à se reporter sur Nicolas Sarkozy tandis que les plus qualifiés accordent plus souvent leurs suffrages à Ségolène Royal.

Enfin, des variations de reports sont également identifiées selon le statut d'occupation des électeurs de leur logement. Ainsi, les électeurs de François Bayrou choisissent plus souvent Nicolas Sarkozy au second tour lorsqu'ils sont propriétaires et plus souvent Ségolène Royal quand ils sont locataires,, que ce soit en HLM ou dans le parc privé. Pour ce qui est des électeurs de Jean-Marie Le Pen, ils accordent plus massivement leurs suffrages à Nicolas Sarkozy lorsqu'ils sont locataires du parc privé. Peut-être faut-t-il y voir un effet des propositions du candidat de l'UMP en matière de soutien à l'accession à la propriété[26].

4.3.4 Estimation d'un modèle de probabilités de report logistiques

L'estimation d'un modèle de probabilité de report de vote logistique permet d'évaluer l'incidence marginale de certaines variables ou modalités sur le report de voix d'un candidat sur un autre à autres variables du modèle donné, c'est à dire « toutes choses égales par ailleurs ». Les principales conclusions qui peuvent être extraites de l'estimation de ce modèle concernent les reports des électeurs de François Bayrou et de Jean-Marie Le Pen sur chacun des deux finalistes de l'élection. Les résultats de l'estimation sont appréhendés en prenant comme référence le report vers le vainqueur de l'élection, Nicolas Sarkozy.

Concernant tout d'abord le report des électeurs de François Bayrou, il présente (cf §C.3.5 en annexe) de faibles variations en fonction des variables explicatives mobilisées dans le modèle. L'effet marginal le plus ample concerne les niveaux de diplôme. Ainsi, une augmentation d'un point de la part des diplômés de niveau bac +2 parmi les électeurs d'un bureau de vote donné est estimée conduire en moyenne à un accroissement de +0,43 point de la proportion des électeurs centristes qui y votent pour Ségolène Royal au second tour. Le niveau de diplôme pris pour référence étant supérieur à bac +2, ce résultat implique que les diplômés de l'enseignement supérieur court qui votent pour François Bayrou sont à autres caractéristiques comparables plus favorables à Ségolène Royal que les diplômés du supérieur long. De manière similaire, le modèle permet d'affirmer que les électeurs de François Bayrou disposant d'un niveau de formation initiale modeste sont eux aussi relativement plus favorables à la candidate socialiste que les électeurs centristes ayant au moins une licence. Par ailleurs, le report de Bayrou vers Royal augmente avec la part d'étrangers sur la zone couverte par les bureaux de vote et il diminue avec la part des retraités inscrits dans les bureaux. D'un point de vue géographique, le report des électeurs de François Bayrou est plus favorable à Ségolène Royal dans toutes les zones à l'exception de la Corse par rapport au nord-est. Les zones qui se distinguent le plus du nord-est sont le sud-est et toutes celles de la façade atlantique.

Pour ce qui est à présent du report des électeurs frontistes, certaines variables ont contrairement au cas de François Bayrou des effets marginaux tout à fait notables. Ainsi, un accroissement d'un point de la part des étrangers en un lieu donné y réduit le report de votes dont bénéficie Ségolène Royal de la part des électeurs de Jean-Marie Le Pen de plus de deux points par rapport à Nicolas Sarkozy. De même, chaque point de pourcentage supplémentaire que représente la part des diplômés du supérieur court s'accompagne d'une contraction de près de deux points du report des électeurs frontistes vers Ségolène Royal par rapport à Nicolas Sarkozy. A l'inverse, un point de taux de chômage en plus à un endroit donné tend à y accroître le report frontiste en faveur de Ségolène Royal. Sous l'angle de la géographie, le nord-est constitue la zone au sein de laquelle les électeurs de Jean-Marie Le Pen accordent le soutien relativement le plus important à Ségolène Royal à autres caractéristiques semblables. Les trois zones qui se distinguent le plus nettement du nord-est à cet égard sont le centre-est, le sud-est ainsi que la Corse.

26. Nicolas Sarkozy propose notamment de rendre déductibles de l'assiette de l'impôt sur le revenu les intérêts des emprunts immobiliers destinés à financer l'acquisition d'une résidence principale.

4.3.5 Estimations fréquentistes

L'estimation des reports au niveau national par une méthode fréquentiste (cf §C.3.6) produit des résultats très proches en ce qui concerne les reports entre des candidats précis. En revanche, les estimations concernant d'une manière ou d'une autre l'abstention et dans une moindre mesure les votes blancs et nuls ressortent profondément déstabilisées de la démarche fréquentiste. En l'occurrence, les estimations fréquentistes de reports de l'abstention et des votes blancs et nuls au premier tour ne sont pas du tout crédibles. Ainsi, lorsque les électeurs concernés auraient finalement choisi d'exprimer une préférence entre les deux candidats du second tour, ils auraient d'après ces estimations unanimement accordé leurs suffrages à Ségolène Royal.

Pour le reste, les estimations fréquentistes convergent globalement avec les estimateurs bayésiens. Elles concluent juste à un écart légèrement plus serré entre les reports des électeurs de François Bayrou et de Frédéric Nihous en faveur des deux finalistes. Le seul candidat pour lequel les estimations fréquentistes diffèrent sensiblement des résultats bayésiens est Jean-Marie Le Pen. La méthode fréquentiste conclut à cet égard à un report nettement plus favorable à Nicolas Sarkozy. A l'inverse, presque aucun (1%) électeur frontiste n'aurait à en croire ces chiffres voté pour Ségolène Royal au second tour.

5 Conclusion

Dans le cadre de ce travail, nous proposons des modèles de report de vote flexibles, permettant l'analyse de manière globale, locale et suivant une ou plusieurs dimensions socio-démographiques. L'inférence suivant ces modèles soulève des problèmes complexes, résolus en adoptant une optique bayésienne et en mobilisant de nombreuses techniques algorithmiques. Le comportement de ces inférences sur des données simulées montre qu'elles conduisent à des résultats relativement satisfaisants et robustes, même si quelques corrections d'ordre numérique ou informatique restent à faire. En outre, elles permettent d'aboutir à des estimations crédibles de reports de votes entre les deux tours de l'élection présidentielle de 2007, ces estimations étant à la fois compatibles avec les enseignements extraits de la littérature en science politique et convergentes avec les reports de vote reconstitués au moyen d'enquêtes d'opinion par les instituts de sondage.

L'application de cette méthodologie aux données de l'élection présidentielle de 2007 permet plus particulièrement de formuler trois conclusions sur le scrutin. Premièrement, les électeurs ayant porté leurs suffrages sur François Bayrou au premier tour ont accordé un léger avantage à Nicolas Sarkozy au second tour. Il s'agit là d'un élément décisif du dénouement de l'élection. Par ailleurs, les électeurs de Jean-Marie Le Pen se sont pour leur part très majoritairement reportés sur Nicolas Sarkozy. Ce résultat suggère que le vote en faveur du Front National traduit bien une adhésion au projet de droite radicale porté par cette formation politique, et non uniquement une défiance à l'égard des institutions. Enfin, la plupart des électeurs qui ont opté au premier tour pour un « petit candidat » vote au second tour pour le finaliste appartenant au même camp que le candidat en question. Ainsi, le clivage entre la gauche et la droite demeure une grille de lecture pertinente de la vie politique française.

La méthode d'estimation construite à l'occasion de cette étude est évidemment susceptible d'être exploitée pour évaluer les reports de votes intervenus entre les deux tours d'autres scrutins tels que les élections régionales ou municipales dans les grandes villes. Elle pourrait aussi plus généralement être mobilisée pour appréhender la mobilité électorale se produisant entre élections successives. La réplication de la méthode à ce cadre permettrait notamment d'explorer deux pistes de recherche. La première est générale. Il pourrait être opportun de tester l'assertion selon laquelle les électeurs du camp défait lors d'une élection présidentielle s'abstiennent davantage que ceux du camp vainqueur aux élections législatives qui suivent directement la présidentielle en question. La seconde piste est liée à l'actualité récente. Estimer les reports de votes entre le premier tour de l'élection présidentielle de 2012 et les dernières élections européennes permettrait de mieux décrypter la performance électorale réalisée par le Front National le 25 mai dernier.

A Annexe : algorithmes de simulation

Pour la complétude du document, nous détaillons l'ensemble des algorithmes de simulation utilisés. Ces algorithmes sont spécifiques à chaque modélisation considérée et peuvent admettre pour chacune quelques variantes que nous précisons. Auparavant, nous écrivons sous forme de pseudo-code, l'algorithme défini au §2.3.2 construisant les tableaux de contingence candidats à deux dimensions.

A.1 Construction des tableaux candidats à deux dimensions

L'algorithme prend en entrée les probabilités jointes $p^k_{i,j}$, les marges $(N^k_{i,*})$ suivant les lignes et $(N^k_{j,*})$ suivant les colonnes et le tableau de contingence courant $N^k_{i,j}$. En sortie, il fournit le tableau candidat $(N'^k_{i,j})$, sa probabilité $P' = P((N'^k_{i,j}) \mid (p^k_{i,j}))$ et la probabilité $P = P((N^k_{i,j}) \mid (p^k_{i,j}))$ de tirage du tableau courant.

Algorithme 1 Construction d'un tableau de contingence candidat

tirage d'une permutation des lignes $(\sigma^I_1, ..., \sigma^I_I) \in \mathfrak{S}_I$ et des colonnes $(\sigma^J_1, ..., \sigma^J_J) \in \mathfrak{S}_J$
$\forall i = 1 \dots I$, $p'_{i,*} \leftarrow p^k_{i,1} + ... + p^k_{i,J}$, $N'_{i,*} \leftarrow N^k_{i,*}$, $N_{i,*} \leftarrow N^k_{i,*}$
$\forall j = 1 \dots J$, $p'_{*,j} \leftarrow p^k_{1,j} + ... + p^k_{I,j}$, $N'_{*,j} \leftarrow N^k_{*,j}$, $N_{*,j} \leftarrow N^k_{i,*}$
$N'_0 \leftarrow N^k$, $p'_0 \leftarrow 1$, $P' \leftarrow 1$, $P \leftarrow 1$
↻ $i = 1 \dots I - 1$
    $p'_0 \leftarrow p'_0 - p'_{\sigma^I_i,*}$, $N' \leftarrow N'_0$, $N \leftarrow N'_0$
    $p' \leftarrow p'_0$ et $N'_0 \leftarrow N'_0 - N'_{\sigma^I_i,*}$
    ↻ $j = 1 \dots J - 1$
        $p'_{\sigma^I_i,*} \leftarrow p'_{\sigma^I_i,*} - p^k_{i,j}$, $p'_{*,\sigma^J_j} \leftarrow p'_{*,\sigma^J_j} - p^k_{i,j}$
        $p' \leftarrow p' - p^k_{*,\sigma^J_j}$, $\omega \leftarrow (p^k_{\sigma^I_i,\sigma^J_j} \times p')/(p'_{\sigma^I_i,*} \times p'_{*,\sigma^J_j})$
        tirage de $N'^k_{\sigma^I_i,\sigma^J_j} \sim \mathrm{HNF}(N'_{\sigma^I_i,*}, N'_{*,\sigma^J_j}, N', \omega)$
        $P' \leftarrow P' \times P(N'^k_{\sigma^I_i,\sigma^J_j} \mid N'_{\sigma^I_i,*}, N'_{*,\sigma^J_j}, N', \omega)$
        $N' \leftarrow N' - N'^k_{\sigma^I_i,\sigma^J_j}$, $N'_{\sigma^I_i,*} \leftarrow N'_{\sigma^I_i,*} - N'^k_{\sigma^I_i,\sigma^J_j}$, $N'_{*,\sigma^J_j} \leftarrow N'_{*,\sigma^J_j} - N'^k_{\sigma^I_i,\sigma^J_j}$
        $P \leftarrow P \times P(N^k_{\sigma^I_i,\sigma^J_j} \mid N_{\sigma^I_i,*}, N_{*,\sigma^J_j}, N, \omega)$
        $N \leftarrow N - N^k_{\sigma^I_i,\sigma^J_j}$, $N_{\sigma^I_i,*} \leftarrow N_{\sigma^I_i,*} - N^k_{\sigma^I_i,\sigma^J_j}$, $N_{*,\sigma^J_j} \leftarrow N_{*,\sigma^J_j} - N^k_{\sigma^I_i,\sigma^J_j}$
    ↺
    $N'^k_{\sigma^I_i,\sigma^J_j} \leftarrow N'_{\sigma^I_i,*}$
↺
↻ $j = 1 \dots J - 1$
    $N'^k_{\sigma^I_I,\sigma^J_j} \leftarrow N'_{*,\sigma^J_j}$, $N'_{\sigma^I_I,*} \leftarrow N'_{\sigma^I_I,*} - N'_{*,\sigma^J_j}$
↺
$N'^k_{\sigma^I_I,\sigma^J_J} \leftarrow N'_{\sigma^I_I,*}$

où HNF désigne la loi de Hypergéométrique décentrée de Fisher (37) et où $P(\cdot \mid \cdot,\cdot,\cdot,\cdot)$ désigne la loi de probabilité associée.

A.2 Modèle à probabilités de report constantes par ensemble de bureaux

Étant donné l'indépendance des variables par rapport à l'ensemble de bureaux, les simulations sont réalisées en considérant un par un chaque ensemble de bureaux. Ainsi, pour l'ensemble z, l'algorithme de simulation s'écrit comme suit.

Algorithme 2 Modèle à probabilités de report constantes par ensemble de bureaux

$\forall i = 1 \dots I$ tirage de $({}^0p^{(z)}_{1|i}, ..., {}^0p^{(z)}_{J|i}) \sim \Gamma(\alpha^{(z)}_{i,1}, ..., \alpha^{(z)}_{i,J})$
$\circlearrowright\ k = 1 \dots K_z$
  $\forall i = 1 \dots I, j = 1 \dots J,\ {}^0p^k_{i,j} = \frac{N^k_{i,*}}{N^k}\, {}^0p^{(z)}_{j|i}$
  tirage de $({}^0N^k_{i,j})$ suivant l'algorithme 1
$\circlearrowleft$
$\circlearrowright\ n = 1 \dots N$ ▷ itérations de Gibbs
  $\forall i = 1 \dots I, j = 1 \dots J,\ {}^n\tilde{\alpha}^{(z)}_{i,j} \leftarrow \alpha^{(z)}_{i,j} + {}^{n-1}N^1_{i,j} + ... + {}^{n-1}N^{K_z}_{i,j}$
  $\forall i = 1 \dots I$, tirage de $({}^np^{(z)}_{1|i}, ..., {}^np^{(z)}_{J|i}) \sim \Gamma({}^n\tilde{\alpha}^{(z)}_{i,1}, ..., {}^n\tilde{\alpha}^{(z)}_{i,J})$
  $\circlearrowright\ k = 1 \dots K_z$
    calcul des probabilités $({}^np^k_{i,j})$ à partir de $({}^np^{(z)}_{j|i})$ suivant la formule (35)
    tirage du candidat $({}^nN'^k_{i,j})$ suivant l'algorithme 1
    ${}^n\rho^k = \min\left[\frac{P((^{n-1}N^k_{i,j})|({}^np^{(z)}_{j|i}))}{P(({}^nN'^k_{i,j})|({}^np^{(z)}_{j|i}))} \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{^{n-1}N^k_{i,j}!}{{}^nN'^k_{i,j}!} \frac{({}^np^{(z)}_{j|i})^{{}^nN'^k_{i,j}}}{({}^np^{(z)}_{j|i})^{^{n-1}N^k_{i,j}}}, 1\right]$ ▷ ratio de Metropolis-Hastings
    $({}^nN^k_{i,j}) \leftarrow \begin{cases} ({}^nN'^k_{i,j}) & \text{avec probabilité } {}^n\rho^k \\ ({}^{n-1}N^k_{i,j}) & \text{sinon} \end{cases}$
  $\circlearrowleft$
$\circlearrowleft$

où $P((N^k_{i,j}) \mid (p^{(z)}_{j|i}))$ est donnée par le produit des probabilités des $N^k_{i,j}$ dans les lois hypergéométriques décentrées de Fisher associées à l'algorithme du §2.3.2.

A.3 Modèle à probabilités de vote constantes par population

La simulation pour le modèle à probabilités de vote constantes par population s'effectue de la manière, à l'exception de la procédure donnée à la fin du paragraphe 2.3.2 de tirage des tableaux de contingence à trois dimensions $(N^k_{i,j,\pi})$. L'enchaînement correspondant des calculs sont détaillés dans les algorithmes 3 et 4

A.4 Modèle à probabilités de report logistiques

A.4.1 Méthode de Frühwirth-Schnatter & Frühwirth

Nous détaillons la méthode de Frühwirth-Schnatter & Frühwirth. Pour ce faire, nous rappelons le modèle.

$$\forall i = 1 \dots I \qquad f(\boldsymbol{\beta}_{i,2}, ..., \boldsymbol{\beta}_{i,J} \mid (N^k_{i,j})) \propto \prod_{j=2}^{J} \exp\left(-\frac{\|\boldsymbol{\beta}_{i,j}\|^2}{2\sigma^2}\right) \prod_{k=1}^{K}\prod_{j=1}^{J} (p^k_{j|i})^{N^k_{i,j}} \tag{51}$$

$$\text{avec} \qquad p^k_{j|i} = \frac{\lambda^k_{i,j}}{\lambda^k_{i,*}} \qquad \lambda^k_{i,j} = \exp\left(\boldsymbol{\beta}'_{i,j}\boldsymbol{X}^k\right) \qquad \lambda^k_{i,*} = \sum_{j=1}^{J} \lambda^k_{i,j} \tag{52}$$

On a la **représentation** suivante

$$N^k_{i,j} = \sum_{\ell=1}^{N^k_{i,*}} z^k_{i,j,\ell} \qquad z^k_{i,j,\ell} = 1_{(u^k_{i,j,\ell} > u^k_{i,(j),\ell})} \qquad u^k_{i,j,\ell} \overset{\text{iid}}{\sim} \mathcal{G}(\log \lambda^k_{i,j}, 1) \qquad u^k_{i,(j),\ell} = \max_{m \neq j} u^k_{i,m,\ell} \tag{53}$$

Algorithme 3 Modèle à probabilités de vote constantes par population avec marges exactes

$\forall \pi = 1 \dots \Pi$, tirage de $({}^0p_{1,1|\pi}, ..., {}^0p_{I,J|\pi}) \sim \Gamma(\alpha_{1,1,\pi}, ..., \alpha_{I,J,\pi})$
$\circlearrowright\ k = 1 \dots K$
  $\forall i = 1 \dots I, j = 1 \dots J, \pi = 1 \dots \Pi$, ${}^0p^k_{i,j,\pi} = \frac{N^k_\pi}{N^k} {}^0p_{i,j|\pi}$ et ${}^0p^k_{i,j} = {}^0p^k_{i,j,1} + ... + {}^0p^k_{i,j,\Pi}$
  tirage de $({}^0N^k_{i,j})$ suivant l'algorithme 1
  tirage de $({}^0N^k_{i,j,\pi})$ suivant l'algorithme 1 avec les marges $({}^0N^k_{i,j})$
$\circlearrowleft$
$\circlearrowright\ n = 1 \dots N$ ▷ itérations de Gibbs
  $\forall i = 1 \dots I, j = 1 \dots J, \pi = 1 \dots \Pi$, ${}^n\tilde{\alpha}_{i,j,\pi} \leftarrow \alpha_{i,j,\pi} + {}^{n-1}N^1_{i,j,\pi} + ... + {}^{n-1}N^K_{i,j,\pi}$
  $\forall \pi = 1 \dots \Pi$, tirage de $({}^np_{1,1|\pi}, ..., {}^np_{I,J|\pi}) \sim \Gamma({}^n\tilde{\alpha}_{1,1,\pi}, ..., {}^n\tilde{\alpha}_{I,J,\pi})$
  $\circlearrowright\ k = 1 \dots K$
    $\forall i = 1 \dots I, j = 1 \dots J, \pi = 1 \dots \Pi$, ${}^np^k_{i,j,\pi} = \frac{N^k_\pi}{N^k} {}^np_{i,j|\pi}$ et ${}^np^k_{i,j} = {}^np^k_{i,j,1} + ... + {}^np^k_{i,j,\Pi}$
    tirage du candidat $({}^nN'^k_{i,j})$ suivant l'algorithme 1
    tirage du candidat $({}^nN'^k_{i,j,\pi})$ suivant l'algorithme 1 avec les marges $({}^nN'^k_{i,j})$
    ${}^n\rho^k = \min\left[\frac{P[({}^{n-1}N^k_{i,j,\pi})|({}^{n-1}N^k_{i,j}),({}^np_{i,j|\pi})]P[({}^{n-1}N^k_{i,j})|({}^np_{i,j|\pi})]}{P[({}^nN'^k_{i,j,\pi})|({}^nN'^k_{i,j}),({}^np_{i,j|\pi})]P[({}^nN'^k_{i,j})|({}^np_{i,j|\pi})]} \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{\pi=1}^{\Pi} \frac{{}^{n-1}N^k_{i,j,\pi}!}{{}^nN'^k_{i,j,\pi}!} \frac{({}^np_{i,j|\pi})^{{}^nN'^k_{i,j,\pi}}}{({}^np_{i,j|\pi})^{{}^{n-1}N^k_{i,j,\pi}}}, 1\right]$
    $({}^nN^k_{i,j,\pi}) \leftarrow \begin{cases} ({}^nN'^k_{i,j,\pi}) & \text{avec probabilité } {}^n\rho^k \\ ({}^{n-1}N^k_{i,j,\pi}) & \text{sinon} \end{cases}$
  $\circlearrowleft$
$\circlearrowleft$

Pour simplifier les écritures, on pose

$$v^k_{i,j,\ell} = e^{-u_{ijk\ell}} \overset{\text{iid}}{\sim} \mathcal{E}(\lambda^k_{i,j}) \qquad v^k_{i,(j),\ell} = \min_{m \neq j} v^k_{i,m,\ell} \overset{\text{iid}}{\sim} \mathcal{E}(\lambda^k_{i,(j)}) \qquad \lambda^k_{i,(j)} = \sum_{m \neq j} \lambda^k_{i,m} \tag{54}$$

Justification :

$$P(z^k_{i,j,\ell} = 1) = P(v^k_{i,j,\ell} < v^k_{i,(j),\ell}) = 1 - \int_0^\infty e^{-\lambda^k_{i,j}x} \lambda^k_{i,(j)} e^{-\lambda^k_{i,(j)}x} dx = 1 - \frac{\lambda^k_{i,(j)}}{\lambda^k_{i,*}} = p^k_{j|i} \quad \square \tag{55}$$

La procédure d'**agrégation** s'écrit

$$u^k_{i,j,*} = -\log \sum_{\ell=1}^{N^k_{i,*}} v^k_{i,j,\ell} + \log \sum_{\ell=1}^{N^k_{i,*}} v^k_{i,1,\ell} = \log \lambda^k_{i,j} + r^k_{i,j} \qquad r^k_{i,j} \overset{\text{iid}}{\sim} \text{Lo}_3(N^k_{i,*}) \tag{56}$$

Preuve : $r^k_{i,j} = \log \frac{X}{Y}$ avec $X \sim \Gamma(N^k_{i,*})$ et $Y \sim \Gamma(N^k_{i,*})$ indépendantes. Pour h mesurable bornée

$$\begin{aligned}
E\left[h\left(\log \frac{X}{Y}\right)\right] &= \int\int h\left(\log \frac{x}{y}\right) \frac{x^{N^k_i - 1} y^{N^k_i - 1}}{(N^k_i - 1)!^2} e^{-x} e^{-y} dx dy & (57)\\
&= \int\int h(\log z) \frac{z^{N^k_{i,*}-1} y^{2N^k_{i,*}-1}}{(N^k_{i,*} - 1)!^2} e^{-y(z+1)} dz dy & (58)\\
&= \int h(\log z) \frac{z^{N^k_{i,*}-1}}{(1+z)^{2N^k_{i,*}}} \frac{(2N^k_{i,*} - 1)!}{(N^k_{i,*} - 1)!^2} dz & (59)\\
&= \int h(z) \frac{(e^z)^{N^k_{i,*}}}{(1+e^z)^{2N^k_{i,*}}} \frac{1}{B(N^k_{i,*}, N^k_{i,*})} dz & (60)
\end{aligned}$$

où $B(\cdot,\cdot)$ désigne la fonction bêta. $\square$

Algorithme 4 Modèle à probabilités de vote constantes par population avec marges approchées

$\forall \pi = 1 \dots \Pi$, tirage de $({}^0p_{1,1|\pi}, ..., {}^0p_{I,J|\pi}) \sim \Gamma(\alpha_{1,1,\pi}, ..., \alpha_{I,J,\pi})$
$\circlearrowright\ k = 1 \dots K$
  $\forall i = 1 \dots I, j = 1 \dots J, \pi = 1 \dots \Pi$, ${}^0p^k_{i,j,\pi} = \frac{N^k_\pi}{N^k} {}^0p_{i,j|\pi}$, ${}^0p^k_{i,j} = {}^0p^k_{i,j,1} + ... + {}^0p^k_{i,j,\Pi}$ et ${}^0p^k_{\pi|i,j} = \frac{{}^0p^k_{i,j,\pi}}{{}^0p^k_{i,j}}$
  tirage de $({}^0N^k_{i,j})$ suivant l'algorithme 1
  $\forall i = 1 \dots I, j = 1 \dots J$, tirage de $({}^0N^k_{i,j,1}, ..., {}^0N^k_{i,j,\Pi}) \sim \mathcal{M}({}^0N^k_{i,j}, {}^0p^k_{1|i,j}, ..., {}^0p^k_{\Pi|i,j})$
$\circlearrowleft$
$\circlearrowright\ n = 1 \dots N$ ▷ itérations de Gibbs
  $\forall i = 1 \dots I, j = 1 \dots J, \pi = 1 \dots \Pi$, ${}^n\tilde{\alpha}_{i,j,\pi} \leftarrow \alpha_{i,j,\pi} + {}^{n-1}N^1_{i,j,\pi} + ... + {}^{n-1}N^K_{i,j,\pi}$
  $\forall \pi = 1 \dots \Pi$, tirage de $({}^np_{1,1|\pi}, ..., {}^np_{I,J|\pi}) \sim \Gamma({}^n\tilde{\alpha}_{1,1,\pi}, ..., {}^n\tilde{\alpha}_{I,J,\pi})$
  $\circlearrowright\ k = 1 \dots K$
    $\forall i = 1 \dots I, j = 1 \dots J, \pi = 1 \dots \Pi$, ${}^np^k_{i,j,\pi} = \frac{N^k_\pi}{N^k} {}^np_{i,j|\pi}$, ${}^np^k_{i,j} = {}^np^k_{i,j,1} + ... + {}^np^k_{i,j,\Pi}$ et ${}^np^k_{\pi|i,j} = \frac{{}^np^k_{i,j,\pi}}{{}^np^k_{i,j}}$
    tirage du candidat $({}^nN^k_{i,j})$ suivant l'algorithme 1
    $\forall i = 1 \dots I, j = 1 \dots J$, tirage de $({}^nN'^k_{i,j,1}, ..., {}^nN'^k_{i,j,\Pi}) \sim \mathcal{M}({}^nN'^k_{i,j}, {}^np^k_{1|i,j}, ..., {}^np^k_{\Pi|i,j})$
    $${}^n\rho^k = \min\left[\frac{P[({}^{n-1}N^k_{i,j})|({}^np_{i,j|\pi})]}{P[({}^nN'^k_{i,j})|({}^np_{i,j|\pi})]} \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{{}^{n-1}N^k_{i,j}!}{{}^nN'^k_{i,j}!} \frac{({}^np^k_{i,j})^{{}^nN'^k_{i,j}}}{({}^np^k_{i,j})^{{}^{n-1}N^k_{i,j}}}, 1\right]$$
    $$({}^nN^k_{i,j,\pi}) \leftarrow \begin{cases} ({}^nN'^k_{i,j,\pi}) & \text{avec probabilité } {}^n\rho^k \\ ({}^{n-1}N^k_{i,j,\pi}) & \text{sinon} \end{cases}$$
  $\circlearrowleft$
$\circlearrowleft$

La fonction caractéristique de la logistique de type III (N^k_i) vaut, en notant Γ et ψ les fonctions gamma et digamma

$$\frac{\Gamma(N^k_{i,*} - \boldsymbol{i}t)\Gamma(N^k_{i,*} + \boldsymbol{i}t)}{\Gamma(N^k_{i,*})^2} = 1 - t^2\psi'(N^k_{i,*}) + o(t^2) \tag{61}$$

car $\psi = \frac{\Gamma'}{\Gamma}$ et $\psi' = \frac{\Gamma''}{\Gamma} - \frac{\Gamma'^2}{\Gamma^2}$. D'où $Er^k_{i,j} = 0$ et $Vr^k_{i,j} = 2\psi'(N^k_{i,*})$.

La **simulation** de cette variable agrégée $u^k_{i,j,*}$ conditionnellement au tableau $(N^k_{i,j})$ se fait simplement en écrivant sa fonction caractéristique.

$$E\left[e^{\boldsymbol{i}\sum_j t_j \sum_\ell v^k_{i,j,\ell}} \,\middle|\, z^k_{i,j,\ell}\right] = \prod_{\ell=1}^{N^k_{i,*}} E\left[e^{\boldsymbol{i}\sum_j t_j v^k_{i,j,\ell}} \,\middle|\, 1_{(v^k_{i,j,\ell} < v^k_{i,(j),\ell})}\right] \tag{62}$$

$$\text{Or,} \quad E\left[e^{\boldsymbol{i}\sum_m t_m v^k_{i,m,\ell}} \,\middle|\, v^k_{i,j,\ell} < v^k_{i,(j),\ell}\right] = \left(1 - \boldsymbol{i}\frac{t_*}{\lambda^k_{i,*}}\right)^{-1} \prod_{m \neq j} \left(1 - \boldsymbol{i}\frac{t_m}{\lambda^k_{i,m}}\right)^{-1} \quad \text{avec} \quad t_* = \sum_{j=1}^{J} t_j \tag{63}$$

$$\text{D'où} \quad E\left[e^{\boldsymbol{i}\sum_j t_j \sum_\ell v^k_{i,j,\ell}} \,\middle|\, N^k_{i,j}\right] = \left(1 - \boldsymbol{i}\frac{t_*}{\lambda^k_{i,*}}\right)^{-N^k_{i,*}} \prod_{j=1}^{J} \left(1 - \boldsymbol{i}\frac{t_j}{\lambda^k_{i,j}}\right)^{-N^k_{i,*}+N^k_{i,j}} \tag{64}$$

$$\text{Ainsi} \quad \sum_{\ell=1}^{N^k_{i,*}} v^k_{i,j,\ell} \mid (N^k_{i,j}) = U^k_i + V^k_{i,j} \quad \text{avec} \quad U^k_i \overset{\text{iid}}{\sim} \Gamma(N^k_{i,*}, \lambda^k_{i,*}) \quad \text{et} \quad V^k_{i,j} \overset{\text{iid}}{\sim} \Gamma(N^k_{i,*} - N^k_{i,j}, \lambda^k_{i,j}) \tag{65}$$

$$\text{et} \quad u^k_{i,j,*} \mid (N^k_{i,j}) = -\log \sum_{\ell=1}^{N^k_{i,*}} v^k_{i,j,\ell} + \log \sum_{\ell=1}^{N^k_{i,*}} v^k_{i,1,\ell} \mid (N^k_{i,j}) = -\log \frac{U^k_i + V^k_{i,j}}{U^k_i + V^k_{i,1}} \tag{66}$$

où on convient que $\Gamma(0, \cdot) = \delta_0$.

La loi logistique de type III est représenté par un **mélange fini de normales**. Autrement dit, pour i et k donné

$$r_{i,j}^k \overset{\text{iid}}{\sim} \sum_{q=1}^{Q} w_q \mathcal{N}(0, \sigma_q^2) \tag{67}$$

où les poids w_q, les variances σ_q^2 et le nombre Q sont fonction de $N_{i,*}^k$.

Les paramètres de mélange sont extraits du code [7]. Jusqu'à $N_{i,*}^k \leqslant 60$, ils sont tabulés. Entre 60 et 500, $Q = 2$, les deux variances normalisées par $2\psi'(N_{i,*}^k)$ sont données des fractions rationnelles de la forme

$$\frac{a(N_i^k)^2 + bN_i^k + 1}{cN_i^k + d} \tag{68}$$

et les poids sont déduits du système d'équations $q_1 + q_2 = 1$ et $q_1\sigma_1^2 + q_2\sigma_2^2 = 2\psi'(N_{i,*}^k)$. Au delà de 500, une seule normale $\mathcal{N}(0, 2\psi'(N_{i,*}^k))$ est considérée.

Pour chaque triplet (i, j, k) et à chaque itération de l'algorithme de Gibbs, correspond une seule valeur de q qu'on note $q_{i,j}^k$. Cet entier conditionnellement aux autres variables suit la loi discrète

$$P(q_{i,j}^k = q \mid u_{i,j,*}^k, \boldsymbol{\beta}_{i,j}) \propto \frac{w_q}{\sigma_q} \exp - \frac{(u_{i,j,*}^k - \boldsymbol{\beta}'_{j|i}\boldsymbol{X}^k)^2}{2\sigma_q^2} \tag{69}$$

et sa valeur est tirée suivant la méthode d'inversion.

Conditionnellement à cet entier, la variable latente agrégée $u_{i,j,*}^k \mid q_{i,j}^k \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\beta}'_{i,j}\boldsymbol{X}^k, \sigma^2_{q_{i,j}^k})$.
Comme *a priori* $\boldsymbol{\beta}_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, $\boldsymbol{\beta}_{i,j} \mid (u_{i,j,*}^k), (q_{i,j}^k) \sim \mathcal{N}(\boldsymbol{b}_{i,j}, \boldsymbol{B}_{i,j})$ avec

$$\boldsymbol{B}_{i,j} = \left(\frac{1}{\sigma^2}\boldsymbol{I} + \sum_{k=1}^{K} \frac{\boldsymbol{X}^k\boldsymbol{X}'^k}{\sigma^2_{q_{i,j}^k}} \right)^{-1} \qquad \boldsymbol{b}_{i,j} = \boldsymbol{B}_{i,j} \sum_{k=1}^{K} \frac{u_{i,j,*}^k \boldsymbol{X}^k}{\sigma^2_{q_{i,j}^k}} \tag{70}$$

Les **coefficients** $\boldsymbol{\beta}_{i,j}$ conditionnellement aux autres variables sont donc tirés suivant cette loi normale.

L'ensemble de ces éléments se combinent dans l'algorithme 5.

A.4.2 Méthode HMC

Pour la méthode HMC, on définit les potentiels U_i à partir de l'opposé du logarithme de la densité (51) et on calcule leur gradients $\boldsymbol{\nabla} U_i$

$$U_i(\boldsymbol{\beta}_{i,j}) = \sum_{j=1}^{J} \frac{\|\boldsymbol{\beta}_{i,j}\|^2}{2\sigma^2} - \sum_{j=1}^{J}\sum_{k=1}^{K} N_{i,j}^k \log p_{j|i}^k \tag{71}$$

$$\boldsymbol{\nabla} U_i(\boldsymbol{\beta}_{i,j}) = \frac{1}{\sigma^2}\boldsymbol{\beta}_{i,j} + \sum_{k=1}^{K} \boldsymbol{X}^k \left(p_{j|i}^k N_{i,*}^k - N_{i,j}^k \right) \tag{72}$$

$$\text{avec} \qquad p_{j|i}^k = \frac{\lambda_{i,j}^k}{\lambda_{i,*}^k} \qquad \lambda_{i,j}^k = \exp\left(\boldsymbol{\beta}'_{i,j}\boldsymbol{X}^k\right) \qquad \lambda_{i,*}^k = \sum_{j=1}^{J} \lambda_{i,j}^k \tag{73}$$

Algorithme 5 Modèle à probabilités de report logistique suivant la méthode de Frühwirth-Schnatter

$\forall i = 1 \dots I, j = 2 \dots J, k = 1 \dots K$, tirage ${}^0q_{i,j}^k$ par inversion de la loi $P({}^0q_{i,j}^k = q) \propto \frac{w_q}{\sigma_q}$
$\forall i = 1 \dots I, j = 2 \dots J$, calcul des matrices ${}^0\boldsymbol{B}_{i,j}$ suivant la formule (70)
$\forall i = 1 \dots I, j = 2 \dots J$, tirage de ${}^0\boldsymbol{\beta}_{i,j} \sim \mathcal{N}(\boldsymbol{0}, {}^0\boldsymbol{B}_{i,j})$ ▷ initialisation des coefficients
↻ $k = 1 \dots K$
    $\forall i = 1 \dots I, j = 1 \dots J$, ${}^0p_{j|i}^k \leftarrow \frac{{}^0\lambda_{i,j}^k}{{}^0\lambda_{i,*}^k}$ et ${}^0p_{i,j}^k \leftarrow \frac{N_{i,*}^k}{N^k} {}^0p_{j|i}^k$
    tirage de $({}^0N_{i,j}^k)$ suivant l'algorithme 1
    $\forall i = 1 \dots I, j = 1 \dots J$, tirage de ${}^0U_i^k \sim \Gamma(N_{i,*}^k, {}^0\lambda_{i,*}^k)$ et ${}^0V_{i,j}^k \sim \Gamma(N_{i,*}^k - {}^0N_{i,j}^k, {}^0\lambda_{i,j}^k)$
    $\forall i = 1 \dots I, j = 2 \dots J$, ${}^0u_{i,j,*}^k \leftarrow -\log \frac{{}^0U_i^k + {}^0V_{i,j}^k}{{}^0U_i^k + {}^0V_{i,1}^k}$
↺
↻ $n = 1 \dots N$ ▷ itérations de Gibbs
    $\forall i = 1 \dots I, j = 2 \dots J, k = 1 \dots K$, tirage de ${}^nq_{i,j}^k$ par inversion de la loi (69)
    $\forall i = 1 \dots I, j = 2 \dots J$, calcul des matrices ${}^n\boldsymbol{B}_{i,j}$ et des vecteurs ${}^n\boldsymbol{b}_{i,j}$ suivant les formules (70)
    $\forall i = 1 \dots I, j = 2 \dots J$, tirage de ${}^n\boldsymbol{\beta}_{i,j} \sim \mathcal{N}({}^n\boldsymbol{b}_{i,j}, {}^n\boldsymbol{B}_{i,j})$
    ↻ $k = 1 \dots K$
        $\forall i = 1 \dots I, j = 1 \dots J$ ${}^np_{j|i}^k \leftarrow \frac{{}^n\lambda_{i,j}^k}{{}^0\lambda_{i,*}^k}$ et ${}^np_{i,j}^k \leftarrow \frac{N_{i,*}^k}{N^k} {}^np_{j|i}^k$
        tirage du candidat $({}^nN_{i,j}'^k)$ suivant l'algorithme 1
        ${}^n\rho^k = \min\left[\frac{P(({}^{n-1}N_{i,j}^k)|({}^np_{j|i}^k))}{P(({}^nN_{i,j}'^k)|({}^np_{j|i}^k))} \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{{}^{n-1}N_{i,j}^k!}{{}^nN_{i,j}'^k!} \frac{({}^np_{j|i}^k)^{{}^nN_{i,j}'^k}}{({}^np_{j|i}^k)^{{}^{n-1}N_{i,j}^k}}, 1\right]$ ▷ ratio de Metropolis-Hastings
        $({}^nN_{i,j}^k) \leftarrow \begin{cases} ({}^nN_{i,j}'^k) & \text{avec probabilité } {}^n\rho^k \\ ({}^{n-1}N_{i,j}^k) & \text{sinon} \end{cases}$
        $\forall i = 1 \dots I, j = 1 \dots J$, tirage de ${}^nU_i^k \sim \Gamma(N_{i,*}^k, {}^n\lambda_{i,*}^k)$ et ${}^nV_{i,j}^k \sim \Gamma(N_{i,*}^k - {}^nN_{i,j}^k, {}^n\lambda_{i,j}^k)$
        $\forall i = 1 \dots I, j = 2 \dots J$, ${}^nu_{i,j,*}^k \leftarrow -\log \frac{{}^nU_i^k + {}^nV_{i,j}^k}{{}^nU_i^k + {}^nV_{i,1}^k}$
    ↺
↺

On note $\boldsymbol{v}_{i,j}$ les vitesses associées aux paramètres de position $\boldsymbol{\beta}_{i,j}$, et on définit, en considérant des masses unitaires, les énergies totales suivantes

$$H_i(\boldsymbol{\beta}_{i,j}, \boldsymbol{v}_{i,j}) = U_i(\boldsymbol{\beta}_{i,j}) + \sum_{j=1}^{J} \frac{\|\boldsymbol{v}_{i,j}\|^2}{2} \tag{74}$$

Dans la méthode, la simulation de la trajectoire est effectuée suivant un schéma d'Euler « saute-mouton » avec un pas de temps ε pour un temps de lancer T. Dans ces conditions, on considère $L = \lceil \frac{T}{\varepsilon} \rceil$ itérations. Ce faisant, on obtient l'algorithme 6.

Algorithme 6 Modèle à probabilités de report logistique suivant la méthode HMC

$\forall i = 1 \dots I, j = 2 \dots J$, ${}^0\boldsymbol{\beta}_{i,j} \leftarrow \mathbf{0}$
$\circlearrowright\ k = 1 \dots K$
  $\forall i = 1 \dots I, j = 1 \dots J$, ${}^0p^k_{j|i} \leftarrow \frac{{}^0\lambda^k_{i,j}}{{}^0\lambda^k_{i,*}}$ et ${}^0p^k_{i,j} \leftarrow \frac{N^k_{i,*}}{N^k}\,{}^0p^k_{j|i}$
  tirage de $({}^0N^k_{i,j})$ suivant l'algorithme 1
$\circlearrowleft$
$\circlearrowright\ n = 1 \dots N$ ▷ itérations de Gibbs
  $\forall i = 1 \dots I, j = 2 \dots J$, ${}^n\boldsymbol{\beta}^0_{i,j} = {}^{n-1}\boldsymbol{\beta}_{i,j}$ et tirage de ${}^n\boldsymbol{v}^0_{i,j} \sim \mathcal{N}\left(-\frac{\varepsilon}{2}\boldsymbol{\nabla}U_i({}^n\boldsymbol{\beta}^0_{i,j}), \boldsymbol{I}\right)$
  $\circlearrowright\ \ell = 1 \dots L-1$ ▷ Euler saute-mouton
    $\forall i = 1 \dots I, j = 2 \dots J$, ${}^n\boldsymbol{\beta}^\ell_{i,j} = {}^n\boldsymbol{\beta}^{\ell-1}_{i,j} + \varepsilon\,{}^n\boldsymbol{v}^{\ell-1}_{i,j}$ et ${}^n\boldsymbol{v}^\ell_{i,j} = {}^n\boldsymbol{v}^{\ell-1}_{i,j} - \varepsilon\boldsymbol{\nabla}U_i({}^n\boldsymbol{\beta}^\ell_{i,j})$
  $\circlearrowleft$
  $\forall i = 1 \dots I, j = 2 \dots J$, ${}^n\boldsymbol{\beta}^L_{i,j}l = {}^n\boldsymbol{\beta}^{Ll-1}_{i,j} + \varepsilon\,{}^n\boldsymbol{v}^{Ll-1}_{i,j}$ et ${}^n\boldsymbol{v}^L_{i,j}l = {}^n\boldsymbol{v}^{L-1}_{i,j} - \frac{\varepsilon}{2}\boldsymbol{\nabla}U_i({}^n\boldsymbol{\beta}^L_{i,j})$
  $\forall i = 1 \dots I$, ${}^n\rho_i = \min\left[\exp\left(H_i({}^n\boldsymbol{\beta}^0_{i,j}, {}^n\boldsymbol{v}^0_{i,j}) - H_i({}^n\boldsymbol{\beta}^L_{i,j}, {}^n\boldsymbol{v}^L_{i,j})\right), 1\right]$ ▷ Ratio de Metropolis-Hastings
  $\forall i = 1 \dots I$, $({}^n\boldsymbol{\beta}_{i,j}) \leftarrow \begin{cases} ({}^n\boldsymbol{\beta}^L_{i,j}) & \text{avec probabilité } {}^n\rho_i \\ ({}^{n-1}\boldsymbol{\beta}_{i,j}) & \text{sinon} \end{cases}$
  $\circlearrowright\ k = 1 \dots K$
    $\forall i = 1 \dots I, j = 1 \dots J$ ${}^np^k_{j|i} \leftarrow \frac{{}^n\lambda^k_{i,j}}{{}^0\lambda^k_{i,*}}$ et ${}^np^k_{i,j} \leftarrow \frac{N^k_{i,*}}{N^k}\,{}^np^k_{j|i}$
    tirage du candidat $({}^nN'^k_{i,j})$ suivant l'algorithme 1
    ${}^n\rho^k = \min\left[\frac{P(({}^{n-1}N^k_{i,j})|({}^np^k_{j|i}))}{P(({}^nN'^k_{i,j})|({}^np^k_{j|i}))}\prod_{i=1}^{I}\prod_{j=1}^{J}\frac{{}^{n-1}N^k_{i,j}!}{{}^nN'^k_{i,j}!}\frac{({}^np^k_{j|i})^{{}^nN'^k_{i,j}}}{({}^np^k_{j|i})^{{}^{n-1}N^k_{i,j}}}, 1\right]$ ▷ ratio de Metropolis-Hastings
    $({}^nN^k_{i,j}) \leftarrow \begin{cases} ({}^nN'^k_{i,j}) & \text{avec probabilité } {}^n\rho^k \\ ({}^{n-1}N^k_{i,j}) & \text{sinon} \end{cases}$
  $\circlearrowleft$
$\circlearrowleft$

B Annexe : validation des stratégies de résolution

B.1 Validation de la construction des tableaux candidats

Dans ce paragraphe, nous donnons un exemple de test effectué pour valider les algorithmes de simulation et leur implémentation. Cet exemple concerne ici la validation de l'algorithme 1.

Dans ce test, nous considérons un bureau avec $N^k = 15$ électeurs et $I = J = 3$ possibilités de choix au premier et second tour. Les résultats agrégés des deux tours valent $N^k_{i,*} = \{8, 5, 2\}$ et $N^k_{*,j} = \{9, 3, 3\}$. De cette façon, on dénombre seulement 64 combinaisons possibles de tableaux de contingence. En outre, nous considérons le jeu de probabilité de report suivant

	$j=1$	$j=2$	$j=3$
$i=1$	19,27%	20,51%	60,22%
$i=2$	40,96%	1,09%	57,95 %
$i=3$	12,03%	21,06%	66,91 %

TABLE 1 – Probabilités de report considérées

Dans un premier temps, nous vérifions que sans le contrôle de Metropolis-Hastings, l'algorithme 1 tire bien des tableaux de contingence suivant la distribution $P[(N^k_{i,j}) \mid (p^k_{i,j})]$ donnée par le produit des probabilités de la loi hypergéométrique décentrée de Fisher. Nous figeons les permutations σ^I et σ^J en prenant l'identité

pour chacune. La comparaison des distributions théorique et empirique sur un échantillon de taille 100000 est représentée sur la figure 16. Dans un second temps, nous activons le contrôle de Metropolis-Hastings et libérons le choix des permutations. La comparaison entre la distribution théorique cible et celle obtenu sur un échantillon de même taille est indiquée sur la figure 17. Lors de ce calcul, le ratio d'acceptation moyen est de $17,9\%$.

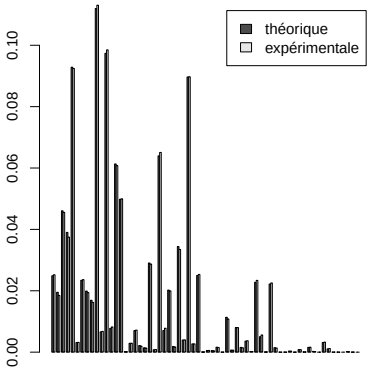

FIGURE 16 – Distributions empirique et théorique des tableaux sans contrôle de Metropolis-Hastings

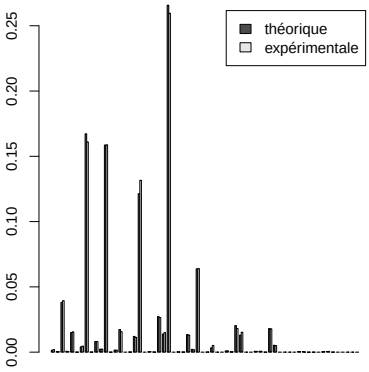

FIGURE 17 – Distributions empirique et théorique des tableaux avec contrôle de Metropolis-Hastings

Visuellement l'adéquation semble bonne. Nous précisons cette dernière avec un test du χ^2 résumé dans le tableau suivant.

Distribution	Statistique de test	p-valeur	Conclusion
Sans contrôle	0.0007	1	accepté
Avec contrôle	0.0039	1	accepté

TABLE 2 – Test du χ^2 pour les deux simulations

Ce test montre ainsi sur un cas simple la pertinence de l'algorithme proposé et de son implémentation.

B.2 Choix des paramètres de la méthode HMC

La méthode HMC a l'inconvénient de dépendre de deux paramètres qui sont le temps de lancer T et le pas du schéma ε. Ces paramètres règlent la qualité de simulation et la longueur des trajectoires. Si la trajectoire est mal simulée, l'énergie totale n'est pas bien conservée et le taux d'acceptation des candidats est altéré. Si la longueur de la chaîne est trop courte, le taux d'acceptation est grand mais la longueur d'auto-corrélation de la chaîne produite est grande. Dans ces conditions, nous utilisons comme **critère de performance**, le **rapport de la longueur efficace de l'échantillon par le temps de calcul**.

Pour mener assez rapidement ces comparaisons, nous ne considérons pas les vraies données mais utilisons des données simulées comprenant 10000 bureaux de vote et seulement deux variables explicatives. Les résultats de ces simulations sont résumés dans le tableau 3, où les tailles efficaces ont été calculées à partir de 4 simulations distinctes, la vitesse est calculée en mesurant en seconde le temps mis par 100 itérations et $L = \frac{T}{\varepsilon}$ représente le nombre d'itérations nécessaire à la simulation des trajectoires. Globalement, les résultats indiquent qu'un temps de lancer d'environ 30 est optimal et qu'on peut utiliser un pas ϵ relativement grossier.
La comparaison avec la méthode de Frühwith-Schnatter & Frühwirth est effectuée sur le même jeu de données et la performance du calcul est également indiqué dans le tableau 3. A ce niveau, il nous est toutefois difficile de savoir si ces résultats peuvent être extrapolés à d'autres cas de calcul. C'est pourquoi, nous choisissons finalement, pour mener à bien les inférences du modèle à probabilité de report logistique, la méthode de Frühwith-Schnatter & Frühwirth qui présente des performances moyennes par rapport à un réglage en aveugle des paramètres de la méthode HMC.

Méthode	T	L	Taux d'accept.	Vitesse	Taille efficace	Performance
HMC	10	2	88,4%	44	38,9	0,88
HMC	10	5	88,4%	60	38,3	0,64
HMC	10	8	88,3%	60	36,6	0,61
HMC	10	10	88,3%	80	36,0	0,45
HMC	20	4	88,3%	50	89,9	1,80
HMC	20	10	88,3%	74	98,4	1,33
HMC	20	15	88,3%	86	96,2	1,12
HMC	20	20	88,3%	102	93,4	0,92
HMC	30	6	89,3%	60	143,9	2,40
HMC	30	14	89,1%	90	139,8	1,55
HMC	30	22	89,1%	112	145,5	1,30
HMC	30	30	89,1%	140	142,5	1,02
HMC	50	10	88,5%	75	178,1	2,37
HMC	50	20	88,7%	118	177,3	1,50
HMC	50	30	88,8%	150	174,7	1,16
Frühwirth			88,5%	63	62,1	0,99

TABLE 3 – Performances des paramètres de la méthode HMC et de la méthode de Frühwirth-Schnatter

B.3 Tableaux de simulation

Chaîne	Vraie val.	IC à 95 %	Chaîne	Vraie val.	IC à 95 %
Abstention→Abstention	80 %	[82 % ; 82 %]	Le Pen→Abstention	15 %	[15 % ; 15 %]
Abstention→Blanc	0 %	[0 % ; 0 %]	Le Pen→Blanc	0 %	[0 % ; 0 %]
Abstention→Royal	10 %	[9 % ; 9 %]	Le Pen→Royal	10 %	[9 % ; 9 %]
Abstention→Sarkozy	10 %	[9 % ; 9 %]	Le Pen→Sarkozy	75 %	[76 % ; 76 %]
Blanc→Abstention	0 %	[0 % ; 0 %]	Nihous→Abstention	10 %	[10 % ; 10 %]
Blanc→Blanc	50 %	[50 % ; 50 %]	Nihous→Blanc	0 %	[0 % ; 0 %]
Blanc→Royal	25 %	[25 % ; 25 %]	Nihous→Royal	20 %	[19 % ; 19 %]
Blanc→Sarkozy	25 %	[25 % ; 25 %]	Nihous→Sarkozy	70 %	[71 % ; 71 %]
Bayrou→Abstention	10 %	[10 % ; 10 %]	Royal→Abstention	5 %	[4 % ; 5 %]
Bayrou→Blanc	0 %	[0 % ; 0 %]	Royal→Blanc	0 %	[0 % ; 0 %]
Bayrou→Royal	55 %	[55 % ; 55 %]	Royal→Royal	95 %	[95 % ; 96 %]
Bayrou→Sarkozy	35 %	[35 % ; 35 %]	Royal→Sarkozy	0 %	[0 % ; 0 %]
Besancenot→Abstention	5 %	[5 % ; 5 %]	Sarkozy→Abstention	5 %	[4 % ; 4 %]
Besancenot→Blanc	0 %	[0 % ; 0 %]	Sarkozy→Blanc	0 %	[0 % ; 0 %]
Besancenot→Royal	90 %	[91 % ; 91 %]	Sarkozy→Royal	0 %	[0 % ; 0 %]
Besancenot→Sarkozy	5 %	[4 % ; 4 %]	Sarkozy→Sarkozy	95 %	[96 % ; 96 %]
Bové→Abstention	10 %	[10 % ; 10 %]	Schivardi→Abstention	20 %	[20 % ; 20 %]
Bové→Blanc	0 %	[0 % ; 0 %]	Schivardi→Blanc	0 %	[0 % ; 0 %]
Bové→Royal	70 %	[70 % ; 71 %]	Schivardi→Royal	70 %	[70 % ; 71 %]
Bové→Sarkozy	20 %	[19 % ; 20 %]	Schivardi→Sarkozy	10 %	[9 % ; 9 %]
Buffet→Abstention	5 %	[4 % ; 5 %]	de Villiers→Abstention	5 %	[5 % ; 5 %]
Buffet→Blanc	0 %	[0 % ; 0 %]	de Villiers→Blanc	0 %	[0 % ; 0 %]
Buffet→Royal	95 %	[95 % ; 96 %]	de Villiers→Royal	15 %	[14 % ; 14 %]
Buffet→Sarkozy	0 %	[0 % ; 0 %]	de Villiers→Sarkozy	80 %	[81 % ; 81 %]
Laguiller→Abstention	15 %	[15 % ; 15 %]	Voynet→Abstention	5 %	[5 % ; 5 %]
Laguiller→Blanc	0 %	[0 % ; 0 %]	Voynet→Blanc	0 %	[0 % ; 0 %]
Laguiller→Royal	65 %	[65 % ; 66 %]	Voynet→Royal	70 %	[70 % ; 71 %]
Laguiller→Sarkozy	20 %	[19 % ; 20 %]	Voynet→Sarkozy	25 %	[24 % ; 25 %]

TABLE 4 – Estimation sur des données simulées issues d'un modèle à probabilités constantes

No chaîne	Vraie val.	IC à 95 %	No chaîne	Vraie val.	IC à 95 %
1	-68 %	[-73 % ; -64 %]	43	-6 %	[-11 % ; 0 %]
2	100 %	[97 % ; 106 %]	44	-25 %	[-33 % ; -20 %]
3	-98 %	[-105 % ; -92 %]	45	-72 %	[-80 % ; -67 %]
4	27 %	[21 % ; 33 %]	46	-9 %	[-17 % ; -3 %]
5	-5 %	[-10 % ; 0 %]	47	-95 %	[-104 % ; -89 %]
6	-41 %	[-49 % ; -36 %]	48	12 %	[4 % ; 18 %]
7	-20 %	[-25 % ; -15 %]	49	-4 %	[-11 % ; 2 %]
8	72 %	[67 % ; 78 %]	50	58 %	[53 % ; 66 %]
9	-65 %	[-71 % ; -59 %]	51	-15 %	[-22 % ; -8 %]
10	70 %	[63 % ; 75 %]	52	21 %	[15 % ; 28 %]
11	-34 %	[-40 % ; -29 %]	53	-60 %	[-70 % ; -49 %]
12	30 %	[24 % ; 36 %]	54	-45 %	[-53 % ; -35 %]
13	75 %	[70 % ; 81 %]	55	-10 %	[-19 % ; 0 %]
14	-27 %	[-33 % ; -22 %]	56	-71 %	[-85 % ; -64 %]
15	16 %	[10 % ; 23 %]	57	-43 %	[-53 % ; -35 %]
16	-1 %	[-8 % ; 4 %]	58	62 %	[55 % ; 71 %]
17	79 %	[75 % ; 86 %]	59	34 %	[27 % ; 42 %]
18	-18 %	[-24 % ; -13 %]	60	6 %	[-1 % ; 14 %]
19	-26 %	[-32 % ; -19 %]	61	-7 %	[-16 % ; 0 %]
20	53 %	[47 % ; 58 %]	62	30 %	[20 % ; 38 %]
21	-23 %	[-29 % ; -16 %]	63	-8 %	[-15 % ; 0 %]
22	12 %	[3 % ; 18 %]	64	62 %	[54 % ; 69 %]
23	28 %	[24 % ; 35 %]	65	43 %	[36 % ; 51 %]
24	46 %	[41 % ; 52 %]	66	0 %	[-9 % ; 9 %]
25	54 %	[49 % ; 62 %]	67	5 %	[-4 % ; 18 %]
26	38 %	[32 % ; 44 %]	68	-26 %	[-39 % ; -19 %]
27	36 %	[30 % ; 44 %]	69	81 %	[76 % ; 92 %]
28	-25 %	[-34 % ; -19 %]	70	-75 %	[-88 % ; -68 %]
29	35 %	[30 % ; 43 %]	71	96 %	[91 % ; 106 %]
30	48 %	[43 % ; 56 %]	72	-23 %	[-30 % ; -17 %]
31	82 %	[76 % ; 92 %]	73	29 %	[17 % ; 40 %]
32	-96 %	[-107 % ; -88 %]	74	6 %	[-6 % ; 17 %]
33	54 %	[50 % ; 63 %]	75	88 %	[82 % ; 98 %]
34	44 %	[38 % ; 50 %]	76	88 %	[80 % ; 99 %]
35	13 %	[5 % ; 20 %]	77	-50 %	[-66 % ; -41 %]
36	2 %	[-5 % ; 9 %]	78	32 %	[20 % ; 43 %]
37	-79 %	[-88 % ; -75 %]	79	-49 %	[-58 % ; -42 %]
38	92 %	[89 % ; 100 %]	80	61 %	[54 % ; 72 %]
39	-17 %	[-24 % ; -11 %]	81	-52 %	[-64 % ; -47 %]
40	62 %	[56 % ; 69 %]	82	45 %	[36 % ; 54 %]
41	-37 %	[-46 % ; -30 %]	83	-98 %	[-110 % ; -91 %]
42	-79 %	[-92 % ; -71 %]	84	57 %	[50 % ; 68 %]

TABLE 5 – Estimation des $\boldsymbol{\beta}_{i,j}$ sur des données simulées issues d'un modèle à probabilités logistiques

Chaîne	Vraie val.	IC à 95 %	Chaîne	Vraie val.	IC à 95 %	Chaîne	Vraie val.	IC à 95 %
1	37 %	[36 % ; 40 %]	38	93 %	[92 % ; 94 %]	76	11 %	[10 % ; 13 %]
2	22 %	[21 % ; 23 %]	39	2 %	[2 % ; 3 %]	77	17 %	[16 % ; 20 %]
3	39 %	[37 % ; 41 %]	40	0 %	[0 % ; 1 %]	78	74 %	[70 % ; 73 %]
4	2 %	[0 % ; 3 %]	41	19 %	[16 % ; 24 %]	79	2 %	[3 % ; 7 %]
5	21 %	[18 % ; 24 %]	42	71 %	[69 % ; 74 %]	80	7 %	[4 % ; 8 %]
6	13 %	[11 % ; 15 %]	43	6 %	[2 % ; 10 %]	81	27 %	[25 % ; 28 %]
7	65 %	[63 % ; 68 %]	44	4 %	[0 % ; 7 %]	82	1 %	[1 % ; 3 %]
8	1 %	[0 % ; 3 %]	45	42 %	[40 % ; 44 %]	83	60 %	[57 % ; 60 %]
9	0 %	[0 % ; 3 %]	46	0 %	[0 % ; 3 %]	84	13 %	[12 % ; 14 %]
10	3 %	[1 % ; 6 %]	47	4 %	[1 % ; 4 %]	85	1 %	[0 % ; 2 %]
11	65 %	[63 % ; 68 %]	48	54 %	[53 % ; 55 %]	86	13 %	[10 % ; 13 %]
12	32 %	[28 % ; 32 %]	49	21 %	[20 % ; 25 %]	87	79 %	[79 % ; 83 %]
13	19 %	[6 % ; 25 %]	50	61 %	[59 % ; 64 %]	88	7 %	[5 % ; 8 %]
14	52 %	[40 % ; 59 %]	51	2 %	[0 % ; 4 %]	89	4 %	[3 % ; 9 %]
15	3 %	[0 % ; 16 %]	52	16 %	[11 % ; 17 %]	90	2 %	[1 % ; 6 %]
16	26 %	[20 % ; 38 %]	53	13 %	[11 % ; 20 %]	91	22 %	[16 % ; 22 %]
17	7 %	[4 % ; 20 %]	54	8 %	[1 % ; 7 %]	92	73 %	[69 % ; 75 %]
18	9 %	[0 % ; 10 %]	55	67 %	[64 % ; 75 %]	93	18 %	[11 % ; 31 %]
19	70 %	[62 % ; 80 %]	56	13 %	[9 % ; 16 %]	94	8 %	[0 % ; 16 %]
20	14 %	[8 % ; 18 %]	57	63 %	[63 % ; 67 %]	95	72 %	[60 % ; 82 %]
21	0 %	[0 % ; 3 %]	58	13 %	[10 % ; 14 %]	96	1 %	[0 % ; 7 %]
22	6 %	[4 % ; 8 %]	59	18 %	[16 % ; 22 %]	97	1 %	[0 % ; 3 %]
23	83 %	[80 % ; 84 %]	60	6 %	[2 % ; 6 %]	98	85 %	[86 % ; 89 %]
24	11 %	[8 % ; 12 %]	61	29 %	[21 % ; 36 %]	99	13 %	[9 % ; 13 %]
25	58 %	[51 % ; 60 %]	62	14 %	[7 % ; 18 %]	100	1 %	[0 % ; 3 %]
26	4 %	[1 % ; 4 %]	63	56 %	[48 % ; 63 %]	101	5 %	[3 % ; 10 %]
27	3 %	[1 % ; 11 %]	64	0 %	[0 % ; 9 %]	102	10 %	[4 % ; 10 %]
28	34 %	[31 % ; 38 %]	65	12 %	[8 % ; 22 %]	103	22 %	[21 % ; 30 %]
29	55 %	[51 % ; 56 %]	66	5 %	[1 % ; 12 %]	104	63 %	[59 % ; 65 %]
30	2 %	[2 % ; 6 %]	67	43 %	[28 % ; 47 %]	105	20 %	[10 % ; 20 %]
31	36 %	[31 % ; 37 %]	68	40 %	[33 % ; 49 %]	106	11 %	[6 % ; 14 %]
32	7 %	[7 % ; 10 %]	69	51 %	[49 % ; 54 %]	107	62 %	[62 % ; 72 %]
33	26 %	[19 % ; 28 %]	70	4 %	[3 % ; 7 %]	108	7 %	[6 % ; 13 %]
34	11 %	[5 % ; 14 %]	71	9 %	[5 % ; 9 %]	109	29 %	[26 % ; 33 %]
35	17 %	[18 % ; 25 %]	72	36 %	[33 % ; 38 %]	110	1 %	[0 % ; 3 %]
36	46 %	[41 % ; 49 %]	73	71 %	[66 % ; 70 %]	111	65 %	[60 % ; 67 %]
37	5 %	[3 % ; 5 %]	74	7 %	[7 % ; 10 %]	112	5 %	[3 % ; 9 %]
			75	11 %	[9 % ; 14 %]			

TABLE 6 – Estimation sur des données simulées issues d'un modèle à probabilités constantes par population

B.4 Méthode d'estimation alternative

La méthode considérée est semi-paramétrique et part des équations de moment d'ordre 1. Ce faisant, nous écrivons ces moments en déduisant respectivement des espérances des lois des modèles (3), (5), (13) et (16) considérés, les équations suivantes

$$r^k_{*,j}(p^{(z)}_{i,j}) = \sum_{i=1}^{I} p^{(z)}_{j|i} N^k_{i,*} - N^k_{*,j} = 0 \qquad r^k_{*,j}(\boldsymbol{\beta}_{i,j}) = \sum_{i=1}^{I} p^k_{j|i} N^k_{i,*} - N^k_{*,j} = 0 \tag{75}$$

$$r^k_{i,*}(p_{i,j|\pi}) = \sum_{j=1}^{J}\sum_{\pi=1}^{\Pi} p_{i,j|\pi} N^k_{\pi} - N^k_{i,*} = 0 \qquad r^k_{*,j}(p_{i,j|\pi}) = \sum_{i=1}^{I}\sum_{\pi=1}^{\Pi} p_{i,j|\pi} N^k_{\pi} - N^k_{*,j} = 0 \tag{76}$$

$$r^k_{i,*}(p_{i,j|\pi}) = \sum_{j=1}^{J}\sum_{\pi=1}^{\Pi} p_{i,j|\pi} \rho^k_{\pi} N^k - N^k_{i,*} = 0 \qquad r^k_{*,j}(p_{i,j|\pi}) = \sum_{i=1}^{I}\sum_{\pi=1}^{\Pi} p_{i,j|\pi} \rho^k_{\pi} N^k - N^k_{*,j} = 0 \tag{77}$$

Pour simplifier les écritures, nous notons vectoriellement $\boldsymbol{r}_I = (r^1_{1,*}, ..., r^K_{I,*})$ et $\boldsymbol{r}_J = (r^1_{*,1}, ..., r^K_{*,J})$. Ce faisant, nous suivons la méthode du quasi-maximum de vraisemblance [8] en supposant que les résidus empiriques $\boldsymbol{r}_I$ et $\boldsymbol{r}_J$ définis à partir des équations (75)-(77) sont indépendants et équi-distribués suivant une loi normale centrée. Considérant alors les équations de vraisemblance, l'approche rejoint celle des moindres carrés non linéaires et conduit aux problèmes de minimisation suivant

$$\min_{(p^{(z)}_{i,j})} \left[\boldsymbol{r}'_J(p^{(z)}_{i,j})\, \Omega_J\, \boldsymbol{r}_J(p^{(z)}_{i,j})\right] \qquad \min_{(\boldsymbol{\beta}_{i,j})} \left[\boldsymbol{r}'_J(\boldsymbol{\beta}_{i,j})\, \Omega_J\, \boldsymbol{r}_J(\boldsymbol{\beta}_{i,j})\right] \tag{78}$$

$$\min_{(p_{i,j|\pi})} \left[\boldsymbol{r}'_I(p_{i,j|\pi})\, \Omega_I\, \boldsymbol{r}_I(p_{i,j|\pi}) + \boldsymbol{r}'_J(p_{i,j|\pi})\, \Omega_J\, \boldsymbol{r}_J(p_{i,j|\pi})\right] \tag{79}$$

Dans une première étape, les problèmes sont résolus en prenant l'identité pour les métriques Ω_I et Ω_J. Dans une seconde, pour améliorer la précision, les problèmes sont à nouveau résolus en prenant respectivement pour ces métriques l'inverse des matrices de variance-covariance des résidus estimés $\boldsymbol{r}_I$ et $\boldsymbol{r}_J$ pour le jeu de paramètres trouvés à la première étape. Par ailleurs, pour tenir compte des contraintes de sommation à 1 des probabilités $(p^{(z)}_{j|i})$ et $(p_{i,j|\pi}$, nous adoptons la paramétrisation suivante :

$$p^{(z)}_{j|i} = \frac{\xi^{(z)}_{i,j}}{\xi^{(z)}_{i,*}} \qquad \xi^{(z)}_{i,*} = \sum_{j=1}^{J} \xi^{(z)}_{i,j} \qquad p_{i,j|\pi} = \frac{\xi_{i,j,\pi}}{\xi_{i,j,*}} \qquad \xi_{i,j,*} = \sum_{\pi=1}^{\Pi} \xi_{i,j,\pi} \tag{80}$$

où $\xi_{i,j}$ et $\xi_{i,j,\pi}$ sont des réels positifs, et où pour l'identification, on pose $\xi_{i,1} = \xi_{i,j,1} = 1\ \forall i = 1 \ldots I, j = 1 \ldots J$. Ce faisant, la théorie [8] indique qu'asymptotiquement, les solutions des programmes (78)-(79) convergent presque sûrement et normalement vers la valeur des paramètres obtenus en maximisant la vraisemblance associée à la loi exacte des marges $N_{*,j}$ et, pour les modèles en population, $N_{i,*}$.

La résolution des problèmes (78)-(79) pose toutefois une difficulté numérique car certaines probabilités tendent vers 1 et conduisent à des hessiennes de la fonction objectif extrêmement mal conditionnées. Pour contourner ce problème, nous utilisons la méthode `BFGS` qui est un algorithme de descente de gradient approchant au fur et à mesure des itérations la hessienne à partir des gradients de la fonction objectif, et évitant ainsi les directions quasi singulières. Pour tenir compte de la positivité des $\xi_{i,j}$ et $\xi_{i,j,\pi}$, nous utilisons sa variante `L-BFGS-B` qui permet de contraindre les variables dans des intervalles donnés. Enfin, la précision de l'estimation est obtenue en calculant les bornes d'un intervalle symétrique à 95% de la distribution bootstrap d'Efron construite à partir de $B = 999$ tirages.

C Annexe : détails des résultats

C.1 Détail des zones de culture politique homogène

Les neuf zones ainsi formées peuvent être décrites de la manière suivante. Une première zone regroupe les trois régions du nord-est de la France : l'Alsace, la Lorraine et la Franche-Comté. Une seconde zone correspond exactement à la région Rhône-Alpes à laquelle s'ajoute la Haute-Loire ainsi que la Saône-et-Loire. Au sud-est de la France, une troisième zone agrège l'intégralité de Provence-Alpes-Côte-D'azur ainsi que la partie est de Languedoc-Roussillon : le Gard, l'Hérault et la Lozère. Une quatrième zone rassemble l'Île-de-France, la Picardie, le Nord-Pas-de-Calais et Champagne-Ardenne. A l'extrême nord-ouest de la France, une cinquième zone regroupe la Bretagne et les départements de la Vendée et de Loire-Atlantique. Une vaste sixième zone correspond à toute la périphérie de la région parisienne qui n'est pas incluse dans la quatrième zone. Plus précisément, cette zone agglomère la Normandie, les Pays-de-la-Loire à l'exclusion de Loire-Atlantique et de la Vendée, la région Centre à l'exception de l'Indre et la Bourgogne, sauf la Saône-et-Loire. Une septième zone regroupe Poitou-Charentes, le Limousin et l'Auvergne exception faite de la Haute-Loire ainsi que les départements du Lot, de la Dordogne et de l'Indre. Au sud-ouest, une huitième zone est constituée de l'Aquitaine sans la Dordogne, de Midi-Pyrénées à l'exclusion du Lot et de la partie ouest de Languedoc-Roussillon, c'est-à-dire l'Aude et les Pyrénées-Orientales. Enfin, la neuvième et dernière zone est réduite au territoire corse.

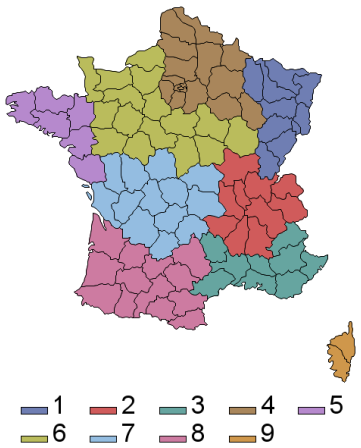

FIGURE 18 – Zones politiquement homogènes

Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7	Zone 8	Zone 9
67	1	13	60	35	27	19	81	2A
68	69	84	95	56	61	24	82	2B
88	73	30	93	44	53	16	9	
90	74	34	94	22	72	87	66	
25	7	6	77	29	45	23	64	
70	43	83	91	85	89	36	65	
57	26	4	59		14	15	31	
39	38	5	80		50	46	32	
54	42	48	78		18	17	33	
55	71		92		58	79	47	
			10		28	86	12	
			52		41	3	40	
			2		49	63	11	
			8		37			
			51		21			
			62		76			
			75					

TABLE 7 – Départements composant les zones de culture politique homogène

C.2 Distribution du nombre d'électeurs par bureaux de vote

Pour avoir une idée de la déformation opérée en regroupant les bureaux dont les données socio-démographiques sont décrites par commune seulement et en éliminant ceux pour lesquels ces données ne sont pas disponibles. Nous traçons la distribution du nombre d'électeur par bureaux et pseudo-bureaux agrégés (cf. figure 19 et 20). La distribution avant les opérations de fusion et d'élimination est quasi identique à celle du fichier original issu du ministère de l'Intérieur. Seuls trois bureaux sont écartés car l'ensemble des votes exprimés au premier ou second tour y sont blancs, ce qui suggère une invalidation locale du suffrage. En revanche, celle après les opérations est essentiellement déformée par un épaississement et un allongement de la queue de distribution à droite. Les (pseudo) bureaux sont au nombre de 46063 et regroupent environ 41 306 000 électeurs, ce qui représente une perte d'environ 780 000 électeurs.

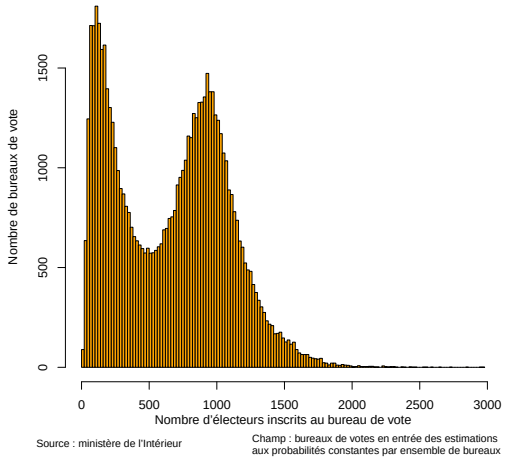

FIGURE 19 – Distribution pour les estimations sans variables socio-démographiques

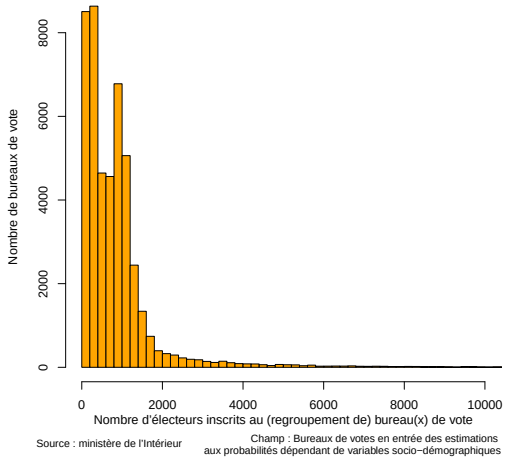

FIGURE 20 – Distribution pour les estimations avec variables socio-démographiques

C.3 Tableaux de probabilité de report de votes

C.3.1 Au niveau de la métropole

	S. Royal	N. Sarkozy	Abstention	Blancs
F. Bayrou	39 ± 0.1%	43.8 ± 0.1%	6.9 ± 0.1%	10.1 ± 0%
O. Besancenot	89.8 ± 0.5%	0.1 ± 0.3%	4.1 ± 0.5%	5.8 ± 0.3%
J. Bové	75.3 ± 0.8%	18.7 ± 0.7%	0 ± 0.1%	5.8 ± 0.6%
M.G. Buffet	93.1 ± 0.4%	0 ± 0%	4.4 ± 0.4%	2.3 ± 0.3%
A. Laguiller	75.6 ± 1%	5.9 ± 0.7%	6.7 ± 0.9%	11.7 ± 0.8%
J.M. Le Pen	11.7 ± 0.1%	69.8 ± 0.1%	10.2 ± 0.2%	8 ± 0.1%
F. Nihous	34.4 ± 0.6%	55.5 ± 0.7%	0 ± 0%	9.9 ± 0.5%
S. Royal	97.7 ± 0%	0 ± 0%	2.2 ± 0%	0 ± 0%
N. Sarkozy	0 ± 0%	97.4 ± 0%	2.5 ± 0%	0 ± 0%
G. Schivardi	51.2 ± 1.5%	30.8 ± 1.6%	0 ± 0.1%	17.8 ± 1.6%
P. de Villiers	13.2 ± 0.5%	79.1 ± 0.5%	0 ± 0.1%	7.5 ± 0.3%
D. Voynet	47.7 ± 0.9%	36.6 ± 0.9%	6.2 ± 0.8%	9.3 ± 0.7%
Abstention	12.3 ± 0.1%	12 ± 0.1%	75.3 ± 0.1%	0.2 ± 0.1%
Blancs	38 ± 1%	18.5 ± 0.8%	8.8 ± 0.9%	34.4 ± 0.8%

TABLE 8 – Valeur moyenne et intervalles de crédibilité à 95% des reports de votes au niveau de la métropole

C.3.2 Par zones politiques

	F. Bayrou		J.M. Le Pen	
	S. Royal	N. Sarkozy	S. Royal	N. Sarkozy
1 Nord-Est	29.8 ± 0.6%	56.2 ± 0.6%	17.1 ± 0.7%	62.4 ± 0.7%
2 Rhône-Alpes	36.1 ± 0.5%	48 ± 0.5%	11.1 ± 0.6%	71.1 ± 0.7%
3 Sud-Est	36.1 ± 0.6%	49.2 ± 0.6%	10.1 ± 0.5%	72.9 ± 0.7%
4 ÎdF-Nord	34.8 ± 0.4%	44.1 ± 0.3%	20.3 ± 0.5%	60.5 ± 0.3%
5 Bretagne-Vendée	42.7 ± 0.5%	43.5 ± 0.6%	25.7 ± 1.3%	59.3 ± 1%
6 Normandie-Bourgogne	39.6 ± 0.3%	46.1 ± 0.5%	15.7 ± 0.6%	66 ± 0.5%
7 Centre Ouest	40.1 ± 0.7%	44 ± 0.7%	21 ± 1.1%	59 ± 1.1%
8 Sud-Ouest	43.1 ± 0.3%	42.8 ± 0.4%	14.7 ± 0.7%	71 ± 0.7%
9 Corse	28.9 ± 3%	51.2 ± 3.9%	20.1 ± 2.8%	56.2 ± 3.2%

TABLE 9 – Valeur moyenne et intervalles de crédibilité à 95% des reports de votes de F. Bayrou et de J.M. Le Pen en faveur de S. Royal et N. Sarkozy par zones de culture politique

C.3.3 Par zones politiques et par taille de communes

	S. Royal		N. Sarkozy	
	Petite commune	Grande commune	Petite commune	Grande commune
1 Nord-Est	27.5 ± 0.6%	40.8 ± 1.5%	58.6 ± 0.6%	46.6 ± 1.7%
2 Rhône-Alpes	34.7 ± 0.5%	34.5 ± 1.3%	50.7 ± 0.5%	41.6 ± 1.3%
3 Sud-Est	35.6 ± 0.8%	34.7 ± 1.3%	51.5 ± 0.7%	45.2 ± 1%
4 ÎdF-Nord	34.4 ± 0.2%	34.2 ± 0.5%	46.6 ± 0.3%	39.4 ± 0.5%
5 Bretagne-Vendée	42.1 ± 0.5%	40.7 ± 1.7%	43.8 ± 0.5%	42.3 ± 1.5%
6 Normandie-Bourgogne	38.7 ± 0.4%	42.5 ± 1.5%	47.1 ± 0.4%	41.5 ± 1.1%
7 Centre Ouest	40.2 ± 0.6%	38 ± 3.5%	44.1 ± 0.9%	48.3 ± 3.5%
8 Sud-Ouest	43.1 ± 0.5%	37.7 ± 1.4%	42.2 ± 0.4%	38 ± 1.2%
9 Corse	28.9 ± 2.9%		51.2 ± 3.8%	

TABLE 10 – Valeur moyenne et intervalles de crédibilité à 95% des reports de votes de F. Bayrou en faveur de S. Royal et N. Sarkozy par zones de culture politique et taille de communes

	S. Royal		N. Sarkozy	
	Petite commune	Grande commune	Petite commune	Grande commune
1 Nord-Est	19.5 ± 0.6%	10.1 ± 2.9%	58 ± 0.7%	70.4 ± 2.9%
2 Rhône-Alpes	11.6 ± 0.8%	6.8 ± 3.5%	68.8 ± 0.7%	67.2 ± 3.2%
3 Sud-Est	10.6 ± 0.6%	8.7 ± 1.4%	71.4 ± 0.7%	72.1 ± 1.4%
4 ÎdF-Nord	21.9 ± 0.4%	13.1 ± 1.7%	58.9 ± 0.4%	49.8 ± 1.7%
5 Bretagne-Vendée	25.5 ± 1%	25 ± 5.4%	57.2 ± 1.1%	43.8 ± 4.9%
6 Normandie-Bourgogne	16.1 ± 0.7%	22.4 ± 4.3%	64.5 ± 0.6%	59.8 ± 2.7%
7 Centre Ouest	20.2 ± 1%	39.4 ± 7.4%	59.4 ± 1.1%	39.7 ± 7%
8 Sud-Ouest	14.6 ± 0.7%	10.4 ± 3.1%	68.8 ± 0.8%	74.7 ± 2.4%
9 Corse	20.1 ± 2.8%		56.2 ± 3.1%	

TABLE 11 – Valeur moyenne et intervalles de crédibilité à 95% des reports de votes de J.M. Le Pen en faveur de S. Royal et N. Sarkozy par zones de culture politique et taille de communes

C.3.4 Par « populations »

	F. Bayrou		J.M. Le Pen	
	S. Royal	N. Sarkozy	S. Royal	N. Sarkozy
18 à 24 ans	62.1% – 74.4%	17.1% – 29.1%	1% – 77.4%	0% – 63.7%
25 à 39 ans	30.1% – 34.2%	43.3% – 47%	39.1% – 64.4%	0.2% – 9.8%
40 à 54 ans	42.3% – 45.3%	40.2% – 44%	4.4% – 6.6%	81% – 82.9%
55 à 64 ans	30.1% – 34%	51.8% – 55.4%	0% – 6.9%	84.4% – 99.1%
65 à 79 ans	0.2% – 25.8%	62% – 95.7%	11.4% – 15.7%	50.2% – 53.7%
$\geqslant$ 80 ans	27.8% – 32.1%	43.9% – 50.3%	0.8% – 77.1%	0.8% – 70.5%

TABLE 12 – Intervalles de crédibilité à 95% des reports de votes par tranches d'âge de F. Bayrou et de J.M. Le Pen en faveur de S. Royal et N. Sarkozy

	F. Bayrou		J.M. Le Pen	
	S. Royal	N. Sarkozy	S. Royal	N. Sarkozy
agriculteurs	32.1% – 35%	55% – 57.8%	16.7% – 30%	45.8% – 57.1%
commerçants	0% – 0.2%	84.3% – 88.5%	0% – 0.1%	98.8% – 99.9%
cadres - prof. lib.	34.3% – 36%	29.9% – 31.4%	0% – 72.2%	0% – 71.7%
prof. int.	41.5% – 43.3%	41.5% – 43.5%	0% – 0.1%	87.1% – 93.7%
employés	84.2% – 90.4%	0% – 5.8%	0% – 0.2%	96.1% – 99.6%
ouvriers	23.3% – 26.8%	66.8% – 70.7%	15.2% – 16.2%	54.1% – 56%

TABLE 13 – Intervalles de crédibilité à 95% des reports de votes par catégories socio-professionnelles de F. Bayrou et de J.M. Le Pen en faveur de S. Royal et N. Sarkozy

	F. Bayrou		J.M. Le Pen	
	S. Royal	N. Sarkozy	S. Royal	N. Sarkozy
sans diplôme	1% – 72.3%	1.1% – 77.1%	16.6% – 18.9%	56.6% – 60.1%
CEP	20.8% – 28.9%	63.1% – 68%	38.1% – 44.5%	0% – 2.5%
BEPC	0.9% – 74.6%	0.6% – 64.3%	0% – 0.4%	98.5% – 99.8%
CAP-BEP	48.5% – 53.7%	38.5% – 41.1%	0% – 0.1%	86.5% – 90.3%
BAC	15.1% – 31.2%	0.4% – 26.6%	0% – 1.7%	94.3% – 99.5%
BAC +2	39.8% – 41.6%	44.1% – 46.2%	0% – 64.5%	1.8% – 81.6%
> BAC +2	37.6% – 38.8%	27.1% – 29.5%	0% – 60.7%	0% – 80%

TABLE 14 – Intervalles de crédibilité à 95% des reports de votes par niveau de formation de F. Bayrou et de J.M. Le Pen en faveur de S. Royal et N. Sarkozy

	F. Bayrou		J.M. Le Pen	
	S. Royal	N. Sarkozy	S. Royal	N. Sarkozy
propriétaire	37.8% – 38.6%	45.8% – 46.7%	9% – 9.6%	68.4% – 69.8%
locataire non HML	37.2% – 38.6%	29.7% – 31.1%	11.6% – 19.2%	73.4% – 82.7%
locataire HML	47.5% – 49.6%	35.6% – 40.3%	14% – 16.8%	59.5% – 61.7%
logement gratuit	0% – 2.4%	12.6% – 84%	0% – 0.2%	53% – 72.5%

TABLE 15 – Intervalles de crédibilité à 95% des reports de votes par type de propriété de F. Bayrou et de J.M. Le Pen en faveur de S. Royal et N. Sarkozy

C.3.5 Par le modèle logistique »

	F. Bayrou		J.M. Le Pen	
	Probabilité de report moyenne	Effet marginal moyen	Probabilité de report moyenne	Effet marginal moyen
constante	46.8% – 47.2%	-12.4% – -7.4%	11.4% – 12%	12.8% – 28.9%
prop. mineurs	46.8% – 47.2%	-23.4% – -12.2%	11.4% – 12%	22.4% – 48.1%
prop. retraités	46.8% – 47.2%	-39.5% – -32.3%	11.4% – 12%	30.4% – 54.6%
prop. étrangers	46.8% – 47.2%	34.2% – 46.6%	11.4% – 12%	-221.8% – -199.9%
prop. propriétaires	46.8% – 47.2%	-5.2% – -1.8%	11.4% – 12%	-6.4% – -1.6%
taux de chômage	46.8% – 47.2%	-0.9% – 18.8%	11.4% – 12%	89.7% – 132.2%
prop. sans diplôme	46.8% – 47.2%	-8.2% – -0.6%	11.4% – 12%	-71.3% – -46%
prop. CEP	46.8% – 47.2%	8.2% – 19.9%	11.4% – 12%	-73.6% – -42.7%
prop. BEPC	46.8% – 47.2%	22.3% – 38%	11.4% – 12%	-98.9% – -65.4%
prop. CAP-BEP	46.8% – 47.2%	27.6% – 35.1%	11.4% – 12%	-69.8% – -51.2%
prop. BAC	46.8% – 47.2%	-41.7% – -29.9%	11.4% – 12%	-33.9% – -6%
prop. BAC + 2	46.8% – 47.2%	36.1% – 49.8%	11.4% – 12%	-192.6% – -140.6%
en zone 2	46.8% – 47.2%	6.2% – 8.4%	11.4% – 12%	-421.7% – -179.2%
en zone 3	46.8% – 47.2%	10.6% – 12.8%	11.4% – 12%	-269.4% – -83.7%
en zone 4	46.8% – 47.2%	6.2% – 7.8%	11.4% – 12%	-6.1% – -3.8%
en zone 5	46.8% – 47.2%	10.8% – 12.3%	11.4% – 12%	-5.9% – -2.9%
en zone 6	46.8% – 47.2%	12.8% – 14.1%	11.4% – 12%	-16.1% – -12.2%
en zone 7	46.8% – 47.2%	12.9% – 15%	11.4% – 12%	-11.8% – -7.1%
en zone 8	46.8% – 47.2%	14.4% – 15.8%	11.4% – 12%	-23.2% – -19.2%
en zone 9	46.8% – 47.2%	-14.5% – -1.5%	11.4% – 12%	-233.6% – -119.3%
péri-urbain	46.8% – 47.2%	-3.1% – -1.9%	11.4% – 12%	-2.1% – -0.1%
rural	46.8% – 47.2%	-3.9% – -2.2%	11.4% – 12%	-3.7% – -1.6%
autre commune	46.8% – 47.2%	-2.2% – -1%	11.4% – 12%	-4.8% – -3.1%

TABLE 16 – Intervalles de crédibilité à 95% des reports de votes et des effets marginaux moyens suivant le modèle logistique de F. Bayrou et de J.M. Le Pen en faveur de S. Royal et N. Sarkozy

C.3.6 Par une méthode fréquentiste »

	S. Royal	N. Sarkozy	Abstention	Blancs
F. Bayrou	40.7 ± 0.1%	42.6 ± 0.1%	6.3 ± 0%	10.1 ± 0%
O. Besancenot	95.9 ± 0%	0 ± 0%	0.4 ± 0%	3.5 ± 0%
J. Bové	82.6 ± 0%	15.5 ± 0%	1.7 ± 0%	0 ± 0%
M.G. Buffet	92.4 ± 0%	0 ± 0%	4.4 ± 0%	3.1 ± 0%
A. Laguiller	89.7 ± 0%	0 ± 0%	4.7 ± 0%	5.5 ± 0%
J.M. Le Pen	1.4 ± 0%	85.5 ± 0.1%	4.4 ± 0%	8.6 ± 0.1%
F. Nihous	40.4 ± 0%	52.2 ± 0%	0 ± 0%	7.2 ± 0%
S. Royal	98.1 ± 0%	0 ± 0%	1.8 ± 0%	0 ± 0%
N. Sarkozy	0 ± 0%	99.8 ± 0%	0 ± 0%	0.1 ± 0%
G. Schivardi	31.5 ± 0%	49.8 ± 0%	0 ± 0%	18.6 ± 0%
P. de Villiers	0 ± 0%	88.9 ± 0%	3.3 ± 0%	7.6 ± 0%
D. Voynet	51.5 ± 0%	39.3 ± 0%	0.1 ± 0%	8.8 ± 0%
Abstention	13.3 ± 0.1%	0 ± 0%	86.6 ± 0.1%	0 ± 0%
Blancs	50.4 ± 0%	0 ± 0%	3.6 ± 0%	45.8 ± 0%

TABLE 17 – Valeur moyenne et intervalles de confiance à 95% des reports de votes au niveau de la métropole suivant la méthode fréquentiste du §B.4

Bibliographie

[1] www.data.gouv.fr.

[2] Inverse normal computation, algorithm as 241. *Applied Statistics*, Vol 37, 1988.

[3] Bernard A. Comment vont s'effectuer les reports de voix au second tour des présidentielles ? les enseignements tirés d'une analyse statistique des présidentielles de 2007 et d'élections intermédiaires. *Ecole Polytechnique. Cahier*, 2012-13, avril 2012.

[4] Duhamel A. *Les Prétendants 2007*. Plon, 2006.

[5] Fog A. www.agner.org/random.

[6] Fog A. Sampling methods for wallenius' and fisher's noncentral hypergeometric distributions. *Communications in Statistics, Simulation and Computation*, Vol 37 No 2 :241–257, 2008.

[7] Fussl A. *binomlogit : Efficient MCMC for Binomial Logit Models*. cran.r-project.org/web/packages/binomlogit/, 2014.

[8] Gourieroux C Monfort A Trognon A. Pseudo maximum likelihood methods : Theory. *Econometrica*, Vol 52 No 3 :681–700, May 1984.

[9] Siegfried A. *Tableau politique de la France de l'Ouest sous la Troisième République*. 1913.

[10] Duncan O Davis B. An alternative to ecological correlation. *American Sociological Review*, No 18 :665–666, 1953.

[11] Andrews D Mallows C. Bayesian analysis of binary and polychotomous response data. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol 36 No 1 :99–102, 1974.

[12] Freedman D. Ecological inference and the ecological fallacy. *Prepared for the International Encyclopedia of the Social & Behavioral Sciences. Technical Report*, No 549, october 1999.

[13] Gelman A Rubin D. Inference from iterative simulation using multiple sequences. *Statistical Science*, Vol 7 No 4, 1992.

[14] Dupoirier E. L'électorat présidentiel de ségolène royal, premiers éléments d'analyse. *Revue française de science politique Presses de Sciences Po*, Vol 57, 2007.

[15] Stadlober E. The ratio of uniforms approach for generating discrete random variates. *Journal of Computational and Applied Mathematics*, Vol 31 No 1 :181–189, 1990.

[16] Deming E Stepan F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, Vol 11 No 4 :427–444, 1940.

[17] Baret M Caisson M Meinzel P Thiercelin G. *Analyse du report de votes aux élections présidentielles françaises*. Tuteurs : Chopin N, Ryder R, Projet de Statistique Appliquée ENSAE, 2013.

[18] King G. Letters to the editor. *Journal of the American Statistical Association*, Vol 94 No 445, March 1999.

[19] Marsaglia G. www.stat.fsu.edu/pub/diehard.

[20] Robert C Casella G. *Monte Carlo Statistical Methods*. Springer, 2004.

[21] Robert C Casella G. *Méthodes de Monte Carlo avec R*. Springer, 2011.

[22] Todd E Le Bras H. *Le mystère français*. Seuil et La République des idées, 2013.

[23] Bickel P Ritov Y Wellner J. Efficient estimation of linear functionals of a probability measure p with known marginal distributions. *The Annals of Statistics*, Vol 19 No 3 :1316–1346, 1991.

[24] Boyett J. Algorithm as 144 : Random r×c tables with given row and column totals. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol 28 No 3 :329–332, 1979.

[25] Halton J. A rigorous derivation of the ekact contingency formula. *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol 65 No 2 :527–530, 1969.

[26] Kinderman A Monahan J. Computer generation of random variables using the ratio of uniform deviates. *ACM transactions on mathematical software*, Vol 3 :257–260, 1977.

[27] Polson N Scott J Windle J. Bayesian inference for logistic models using polya-gamma latent variables. *Journal of American Statistical Association*, Vol 108 :1339–1349, 2013.

[28] Tam Cho W Judge G Miller D. An information theoretic approach to ecological estimation and inference. In King G Rosen O Tanner M, editor, *Ecological Inference : New Methodological Strategies*, pages 162–187. Cambridge University Press, New York, 2004.

[29] Devroye L. *Non-uniform random variate generation*. Springer, 1986.

[30] Goodman L. Some alternatives to ecological correlation. *American Journal of Sociology*, No 64 :610–624, 1959.

[31] Holmes C Held L. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, Vol 1 No 1 :145–168, 2006.

[32] King G Rosen O Tanner M. *Ecological Inference : New Methodological Strategies*. Cambridge University Press, New York, 2004.

[33] Rosen O Jiang W King G Tanner M. Bayesian and frequentist inference for ecological inference : The r×c case. *Statistica Neerlandica*, No 55 :134–156, 2001.

[34] Le Nouvel Observateur. tempsreel.nouvelobs.com/elections-2007/20070327.OBS9160/frederic-nihous-se-dit-ni-a-droite-ni-a-gauche.html.

[35] Heidelberger P Welch P. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, Vol 24 No 4 :233–245, April 1981.

[36] Le Point. *Frédéric Nihous (CPNT) confirmé sur une liste UMP en Pyrénées-Atlantiques*. www.lepoint.fr/actualites-politique/2010-02-01/frederic-nihous-cpnt-confirme-sur-une-liste-ump-en-pyrenees/917/0/419367, 1er février 2010.

[37] Fisher R. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1934.

[38] Frühwirth-Schnatter S Frühwirth R. Data augmentation and mcmc for binary and multinomial logit models. In Kneib T Tutz G, editor, *Statistical Modelling and Regression Structures - Festschrift in Honour of Ludwig Fahrmeir*, pages 111–132. Physica-Verlag, Heidelberg, 2010.

[39] Neal R. Mcmc using hamiltonian dynamics. In Brooks S Gelman A Jones G Meng X, editor, *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman & Hall CRC Press, 2011.

[40] Albert J Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, Vol 88 No 422 :669–679, June 1993.

[41] Colange C Beauguitte L Freire-Diaz S. *Base de données socio-électorales Cartelec (2007-2010)*. www.cartelec.net, 2013.

[42] Ireland CT Kullback S. Contingency tables with given marginals. *Biometrika*, Vol 55 No 1 :179–188, March 1968.

[43] Jadot A Bussi M Colange C Freire-Diaz S. Un outil d'analyse électorale en cours de création. cartelec, un sig au niveau des bureaux de vote français. *CFC*, Vol 205, 2010.

[44] Strudel S. L'électorat de nicolas sarkozy : « rupture tranquille » ou syncrétisme tourmenté ? *Revue française de science politique Presses de Sciences Po*, Vol 57, 2007.

[45] Kachitvichyanukul Schmeiser. Binomial random variate generation. *Communications of the ACM*, Vol 31 :216–222, 1988.

[46] Cressie N Read T. Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B*, Vol 46 :400–464, 1984.

[47] Matsumoto M Nishimura T. Mersenne twister : A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, Vol 8 No 1 :3–30, Junuary 1998.

[48] Ahrens J Dieter U. Computer methods for sampling from the exponential and normal distributions. *Communications of the ACM*, Vol 15 :873–882, 1972.

[49] Ahrens J Dieter U. Computer methods for sampling from gamma, beta, poisson and binomial distributions. *Computing*, Vol 12 :223–246, 1974.

[50] Ahrens J Dieter U. Generating gamma variates by a modified rejection technique. *Communications of the ACM*, Vol 25 :47–54, 1982.

[51] Patefield W. Algorithm as 159 : An efficient method of generating random r $\times$ c tables with given row and column totals. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol 30 No 1 :91–97, 1981.