

Fraude ou Transferts ? Une piste à approfondir

Alain Gély, juillet 2006.

Ce travail ayant été avancé pour l'essentiel le 14 juillet, je crains bien que Thomas ne le marque du sceau de l'infamie "souverainiste" ; tant pis, je prends le risque :<))

Pour exploiter un fichier de données :

* Tout d'abord se renseigner avec précision sur la manière dont il a été constitué : en particulier, des « biais d'observations » des effets de génération ou de grappes ont-ils pu affecter la qualité des données OU LEUR ORDONNANCEMENT ?

* Il faut ensuite « l'apurer » et le redresser : corriger les erreurs qu'il contient (1) et compléter les éventuels "blancs". Le fichier de l'huissier (dans la version qui se trouve sur le site de Michel Husson) en contient au moins 2 grosses qui affectent le lot L2 (suffrages de G.Azam) et B2-3 (124 voix attribués). C'est une première précaution nécessaire que tout aspirant statisticien doit prendre. J'ai remplacé ici le 28 de G.Azam par 128 ; les "124 multiples" ne semblent pas affecter les résultats, mais il faudra corriger cette erreur.

* Ensuite appliquer des méthodes d'analyse des données qui prennent en compte TOUTE L'INFORMATION DISPONIBLE. Il faut notamment éviter de n'en sélectionner qu'une partie pour en tirer des conclusions générales qui risquent fort d'être gravement erronées (2).

Une proposition

Je ne dispose pas actuellement d'outils d'analyse stat très performants mais on peut faire pas mal de choses avec Excel (ou la version OpenOffice que je joins au présent courriel). Après avoir comparé, par précaution et par intérêt, le nombre moyen de candidats retenus de chaque lot avec la moyenne générale (21,2) et avoir constaté que cela n'isolait que deux lots relativement marginaux, je propose donc – et j'ai commencé à mettre en oeuvre – la méthode un peu lourde mais relativement simple suivante :

calculer pour chaque candidat et pour chaque lot l'écart qui sépare le % qu'il a obtenu dans ce lot du % qu'il a obtenu au total (3) en déduire, pour chaque lot et pour divers regroupements de lots (par date etc), l'écart absolu moyen EAM qui le sépare de l'ensemble (somme « en colonne » des valeurs absolues des écarts sur les %, divisée ici par 62 (4))

le principal (à appliquer avec quelque autre indicateur de distance que ce soit) : ranger les lots dans l'ordre (ici : décroissant) des EAM ainsi obtenus. Les lots « les plus éloignés du total général » apparaissent donc en tête

à partir de là, il faut introduire des hypothèses autres que la seule, ici admise jusqu'alors, selon laquelle l'EAM serait un outil pertinent. Notamment : fixer un seuil minimum de prise en compte des lots, car les « petits lots » ont souvent une EAM élevée du fait même de leur taille, qui affecte d'ailleurs importants les "petits pourcentages" (5) ; admettons 24, sans enthousiasme ; dans ce cas, les lots qui viennent en tête sont dans l'ordre B4 P4 et VPRO2 qui se **détachent nettement** avec des EAM de l'ordre de 15% puis T3 PM1 VDIRECT2 B3 R2 J1 VPRO3 VPRO5 B5 L4 et VPRO4 (il y a ici une **seconde discontinuité** « significative (6) » dans la succession des EAM, après celle qui suivait les trois premiers lots) le tableau joint cache les "petits lots" mais ils ne sont que cachés et non détruits : on peut les faire réapparaître sans difficulté

Les lots « de plus de 24 bulletins les plus éloignés du total général » ne sont pas 10 (selon une contribution précédente) ni 11 (source MH, dont je ne vois pas la justification ,mais peut-être celle-ci m'a-t-elle échappé) mais 3 ou 14 avec ces hypothèses. Peu importe, d'ailleurs, leur nombre exact : le fait important est qu'on retrouve, certes, plusieurs cas déjà signalés mais pas tous et, surtout, on y trouve beaucoup d'autres, y compris parmi ceux qui avaient été dépouillés avant le 17. **Il n'y a donc absolument pas 10 ou 11 lots "atypiques" mais - puisque nous faisons dans le néologisme avec les concepts récemment créés d'aberration et d'anomalie statistiques - des "degrés d'atypicité" de chaque lot.** Il faut donc approfondir cette piste : je ne dis pas qu'elle soit la seule mais il me semble que tout statisticien compétent et non aveuglé par ses a priori aurait dû pouvoir la découvrir (avec l'EAM ou avec d'autres "indicateurs de distance").

Il est bien sûr possible de continuer, en regroupant des lots OU EN PROCEDANT POUR LES CANDIDATS COMME ON L'A FAIT POUR LES LOTS (classer les lots dans l'ordre décroissant des écarts au % général obtenu par le candidat) et analyser.. Ce que je n'ai pas encore eu le temps de faire.

Ce qui saute tout de suite aux yeux, quand même, c'est que la lettre B se trouve à plusieurs reprises dans ces lots. Si on regroupe B 3 4 et 5 avec le lot B2, malaisément formalisable mais sans doute "agrégeable" avec les suivants, le lot ainsi obtenu en regroupant ces "lots atypiques" « rentre dans le rang » (EAM de 3,2%). Les « bizarreries » de B4, notamment, sont alors effacées. L'hypothèse la plus plausible est alors le transfert entre lots de paquets de bulletins « non constitués aléatoirement ». Ce serait une caractéristique du dépouillement certes regrettable car ne permettant pas de retenir l'hypothèse - toujours forte - de constitution aléatoire des populations étudiées mais totalement distincte de "l'hypothèse fraude". Cette "hypothèse regroupement/transfert" me semble non seulement préférable pour l'avenir d'ATTAC (un scrutin un peu "merdique", voire très, peut plus facilement être corrigé qu'une éventuelle fraude) mais de loin plus vraisemblable que la substitution nocturne de paquets de bulletins - qu'il faudrait de toute façon étayer de manière sérieuse.

Remarque : à ma connaissance, un seul jugement s'est prononcé "en probabilité" pour une conclusion "presque sûre" (probabilité "extrêmement voisine" de 1, risque d'accepter une hypothèse fautive inférieure à une proba très faible, donnée a priori) ; il s'agissait d'une commune de la banlieue parisienne où plusieurs paquets de 100 s'écartaient notablement de la répartition moyenne (environ 50/50 pour chacune des deux listes) ; ces paquets étaient constitués, si j'ai bonne mémoire, de 95 à 97 bulletins pour la liste A et 3 à 5 pour la liste B. Il avait été lors possible d'établir que ces "lots" étaient tous fort éloignés du "lot normal" le plus proche et de faire un calcul de proba relativement simple montrant que la proba d'un tel événement était infime. Un tel calcul n'a pas été produit dans le cas qui nous occupe. Pouvait-il l'être ? Je ne crois pas. Il aurait de toute façon été presque sûrement fallacieux pour des raisons que Michel Fenayon a indiquées dans la seconde partie de son message (le première ne me semble pas probante).

Autre remarque : l'hypothèse d'un « effet de génération » n'est ni confirmée ni démentie par cette méthode, à ce stade de l'analyse. Je persiste à penser qu'il a joué. J'ai ouvert une piste en classant les EAM dans l'ordre chronologique. Le graphique qui s'en déduit (onglet du fichier joint) est à "nettoyer" des petits lots... Je ne suis pas sûr que cette piste-là puisse aboutir. A voir.

Par ailleurs, les EAM calculés sur les regroupements quotidiens ne mettent réellement en évidence que le 17 juin (EAM de 7%, les autres se situant aux alentours de 3%, il n'apparaît pas ici de "jour aberrant" autre que le 17, mais le dépouillement du 17 n'est pas contesté à ma connaissance)

Je ne promets pas de progresser beaucoup dans les 15 jours - de vacances - qui viennent mais l'utilisation de l'EAM, parmi d'autres possibilités, me semble prometteuse. Elle nous dirige vers des conclusions très différentes de certaines affirmations péremptoires. En tout cas **j'insiste sur la nécessité d'utiliser toutes les informations disponibles (qualitatives et quantitatives) quand on prétend analyser des données.**

Je ne peux évidemment conclure à ce stade. Une ou deux choses (avec le sourire, quand même) :

- il vaut mieux ne pas communiquer à l'extérieur les premiers travaux diffusés (un peu précipitamment : sans doute l'euphorie de la pseudo-découverte) par certains : moins par la crainte que cela n'influence ce qui se fait que parce que cela risquerait de discréditer ATTAC et son conseil scientifique, ou du moins certains de ses membres, si les premières affirmations péremptoires - voire diffamatoires - s'avèrent effectivement imprudentes ; gardons quand même ces documents : cela pourra éventuellement servir à des actions d'éducation populaire avec la "découverte" du rayon N et quelques autres errements commis et validés par des "scientifiques" (en toute "bonne foi", éventuellement :<))

- je ne peux pour ma part conclure, à ce stade, ni à l'existence d'une ou de plusieurs fraudes ni à leur impossibilité, ni évaluer sérieusement leur probabilité (qui n'est ni nulle ni "infiniment proche de 1"; **je peux encore moins conclure à une fraude "de type souverainiste"** (conclusion de Thomas, qui n'a pas produit les équations qui le conduisaient à cette intéressante conclusion) **ni à une fraude "de type trotskyste"** (la lecture de "Leur morale et la nôtre" de LDBronstein m'a édifié sur la notion de "fracture éthique" mais je ne considère pas pour autant comme probable l'irruption nocturne d'un commando gauchiste procédant à des substitutions de bulletins dans des conditions nécessairement très acrobatiques et en commettant des erreurs absolument grossières) ; **reste l'hypothèse d'une action de la CIA, de plombiers soudoyés par une agence de compensation luxembourgeoise ou toute autre personne "morale"** qui verrait d'un bon œil l'échec d'ATTAC, mais j'avoue que je manque d'éléments tangibles et que je n'ai pas le talent ni le temps de Michel Husson pour étayer ce genre d'hypothèses - qui pourtant me satisferaient idéologiquement- et l'accompagner de graphiques et d'équations économétriques :<))

Je joins le fichier dans son état actuel (en devenir, donc...) pour celles et ceux qui auraient le temps et le goût d'y travailler, d'y rechercher des erreurs... toujours possibles et de progresser dans l'analyse de ces résultats ! Pour ceux qui n'ont pas OpenOffice et qui brûleraient de se lancer dans ce travail, OpenOffice est gratuit...

(1) Corriger la non-réponse (partielle ou totale) ou plus généralement les « blancs » est un aspect important de cette étape, mais elle est a priori hors sujet ici !

(2) Ou alors, si on veut procéder par sondage (ce qui n'est sans doute pas opportun ici), il faut des procédures rigoureuses.

(3) On peut évidemment comparer chaque lot à d'autres ensembles : les lots dépouillés le même jour, les lots commençant par la même lettre etc Je n'ai fait qu'esquisser ce travail assez simple mais relativement lourd

(4) Peu connu, l'écart absolu moyen EAM a pourtant des propriétés intéressantes : a) il est un indicateur de distance entre deux distributions plus compréhensible par le « profane » que ses concurrents, notamment le khi-deux b) il est plus simple que le coefficient de corrélation (qui a en outre au moins un inconvénient : il risque d'aggraver la confusion fréquente entre corrélation et causalité) c) il ne fait pas intervenir de « dénominateur déstabilisateur » (ceux-ci en revanche « polluent » les pourcentages sur les petits lots, plus affectés par le hasard) d) il a des propriétés intéressantes : c'est par rapport à la médiane que l'EAM d'une distribution est minimum ; or la médiane est indicateur de tendance centrale plus « citoyen » que la moyenne, par exemple ; je préciserai cette affirmation si nécessaire, mais ce n'est pas le lieu :<))

(5) Aucun seuil ne s'impose. 50 me semble excessif en ce qu'il risque d'écarter des informations pertinentes J'aurais bien pris 31 (la moitié de 62 ; nombre par ailleurs proche de 30, que certains considèrent comme le plus petit des grands nombres... on peut évidemment prendre un autre seuil) mais va pour 24 qui a été retenu dans une contribution précédente

(6) En définissant comme significatif un écart entre deux EAM successifs, caractérisé par une différence de 0,5 entre ces % ou une "distance relative" de 5% ; ce choix n'est pas d'enature à affecter les conclusions, puisque l'ordre des lots ne changera pas, mais on peut évidemment en retenir d'autres si l'on désire s'en assurer ;

(7) Regrouper les bulletins où la case « Harribey » était cochée, si c'est bien de cela qu'il s'agit, était une initiative qui pouvait se défendre (léger gain dans le temps de dépouillement, fort long) ; cette initiative s'avère malencontreuse si, comme il est vraisemblable, elle a égaré certains sur le « chemin de l'explication par la fraude » ; pour ma part, je n'accablerai pas celles et ceux qui, participant au dépouillement, ont cherché à le faciliter. Qu'on puisse traiter ces militants d'ânes ou de bureaucrates est carrément injuste, pour ne pas employer d'autre mot...