Les sondages Internet : nouveaux défis et défis revisités

Claire Durand,
Professeur titulaire,
Département de sociologie,
Université de Montréal

Dans le cadre du Symposium de Statistiques Canada “Produire des estimations fiables à partir de bases imparfaites”
Ottawa, 15-18 octobre 2013

Plan de la présentation

- Pourquoi parler de sondages Internet?
- Qu'est-ce qu'un sondage Internet?
- Défis méthodologiques classiques liés aux sondages
  - Couverture
  - Base de sondage/ recrutement
  - Échantillon
  - Taux de réponse
  - Correctifs possibles: pondération et redressements
- Le questionnaire: nouveaux défis
- État des lieux et perspectives d'avenir

Pourquoi parler de sondages Internet?

- Prolifération de sondages électoraux ou sociopolitiques utilisant des panels Internet de volontaires.
- Les recherches académiques utilisent de plus en plus les sondages internet.
- Des sondages “internationaux” présentent une “image” de plusieurs pays d’Afrique, d’Asie basée sur des sondages Internet.
- L’internet est utilisé également comme mode d’administration principal ou complémentaire pour les enquêtes des agences statistiques:
  - Voir le recensement canadien de 2011 et l’Enquête nationale auprès des ménages.

Pourquoi utiliser Internet pour faire des sondages?

- L'accès – incluant sans fil – augmente rapidement, y compris dans les pays moins développés.
  - Pénétration en 2012 selon Internet World Stats:
    - Canada: 83%; États-Unis: 78%;Groenland: 90%
    - Bermudes: 88%; Argentine: 66%; Mexique: 36%
    - Islande: 97%;France 80%;United Kingdom, 84%
    - Corée du sud, 83%, Japon, 79%, Vietnam 34%
    - Maroc, 51%; Algérie, 14%; Togo 5%
- Conduire des sondages internet est peu coûteux.
- Les sondages internet ont
  - les avantages des sondages auto-administrés (moins de conformisme) et
  - certains des avantages des sondages en face à face (présentation de matériel visuel, audio, etc.)

Qu’est-ce qu’un sondage Internet?

- Bien sûr, une enquête via site web mais,
  - Quelle base de sondage, quel mode de recrutement?

Couper (2000) présente 8 modes de recrutement différents:

- Non probabilistes:
  - 1. Vox pop sur les sites des medias
  - 2. Recrutement de volontaires pour 1 enquête ($)
  - 3. Recrutement de volontaires pour « panel » ($)
- Probabilistes:
  - 4. Sondages de sortie de site
  - 5. Listes courriel complètes (organisations, associations)
  - 6. Sondages multi-modes avec possibilité d’utiliser Internet
  - 7. Panels d’utilisateurs d’Internet recrutés par un autre mode
  - 8. Échantillons probabilistes de l’ensemble de la population où l’on fournit l’accès internet lorsque le ménage ne l’a pas.

Qu’est-ce qu’un sondage Internet?

En ce moment

- Seul le mode 8, où on recrute par un autre moyen et où on fournit l’accès Web aux ménages non équipés, permet de constituer une base de sondage probabiliste de l’ensemble de la population.
- Ce mode est très cher et peu disponible.
- La quasi-totalité des sondages Internet publiés sur des questions d’actualité utilise un échantillon non probabiliste d’internautes inscrits à un panel (mode 3) ... **même s’ils pourraient utiliser le mode 7**, plus approprié.

Quels problèmes pour représenter l'ensemble de la population?

Les problèmes de couverture/ d'exclusion

- Une partie de la population, plus ou moins importante selon les pays, n'a pas accès à Internet. Elle a des caractéristiques distinctes - résidence, scolarité, âge, **style de vie**.
- À partir de 80% de la population couverte, l'impact de l'exclusion devient moins important mais néanmoins existant.
- Avec la méthode 8, en fournissant l'accès à Internet, on réduit le problème d'exclusion mais on ne l'élimine pas (il faut pouvoir apprendre à se servir de l'outil).

La couverture / l'exclusion

Bigot, Croutte et Recours (2010)

- Différences entre internautes et non- internautes en France:
  - Non redressé: différences significatives pour 71 variables sur 191 (34%) portant sur
    - L'ensemble des caractéristiques socio-démographiques à l'exception du genre,
    - L'équipement des ménages
    - Les opinions en matière de mœurs
  - Après redressement: différences sur **12%** des variables dont
    - L'équipement des ménages
    - Les opinions en matière de mœurs
    - Le logement
    - Les pratiques culturelles

La constitution de la base de sondage

- On est passé d'une base de ménage à une base individuelle (y inclus dans le téléphonique avec les cellulaires).
- On aurait pu penser…
  - Que les firmes tenteraient de constituer une "base" non biaisée des adresses courriel de la population.
  - Que les sondeurs tenteraient de constituer une base de sondage à partir d'invitations téléphoniques. Il resterait alors surtout des biais relatifs à la couverture et à la non-réponse.
- Ce n'est pas ce qui s'est passé.
  - Les modes de recrutement pour les panels varient selon les firmes et ne rejoignent pas tous les internautes.
  - Pour beaucoup de panels, les répondants peuvent s'inscrire en allant sur le site du sondeur.

La constitution de la base de sondage

Le recrutement

- Fait de manière variable selon les firmes (Baker et al., AAPOR task force, 2013)
  - Sollicitation téléphonique ou postale.
  - River sampling.
  - Sollicitation sur les medias sociaux.
  - Sollicitation sur divers sites web:
    - La variété des sites web utilisés est très importante.
- Certains modes de sollicitation sont probabilistes (recrutement par sondage téléphonique)
  - Certains sondeurs utilisent seulement ce mode.
  - Certains combinent ce mode au recrutement via site web et au recrutement sur le site de la firme. Il est possible de restreindre la base de sondage aux personnes recrutées via sondage téléphonique, mais cela semble peu utilisé.

La constitution de la base de sondage

La confidentialité

- Quelles informations doit-on recueillir sur les membres du panel?
  - Sans doute pour prévenir les inscriptions multiples, pour mieux ajuster et pour pouvoir rejoindre des clientèles spécifiques, certaines firmes recueillent des informations très détaillées – date de naissance, adresse, numéro de téléphone, consommation, valeurs, etc.. . Biais supplémentaire?
- Les répondants ont-ils autant confiance dans la confidentialité de leurs réponses dans un sondage Internet? Comment les rassurer? (Lozar Manfreda et coll., 2008)

La constitution de l'échantillon de départ

- On aurait pu penser...
  - Que les firmes procéderaient comme pour les sondages probabilistes, soit:
    - Sélectionner un échantillon probabiliste "fini" stratifié dans la base "imparfaite".
    - Prendre tous les moyens pour contacter les membres de l'échantillon et les convaincre de répondre.
- La manière de faire consiste plutôt
  - À lancer des invitations en très grand nombre à un échantillon de membres du panel – et même à d'autres;
  - À utiliser des quotas et à clore la collecte des données au fur et à mesure que les quotas sont remplis.

Le taux de réponse/ de participation

- Lozar Manfreda et coll., (2008) font une méta-analyse de 45 expériences comparant le mode Internet à d'autres modes.
- Ils montrent que les taux de réponse aux sondages Internet sont généralement plus bas de 11% en moyenne (6%-15%) que ceux des autres modes.
- Pour les **sondages de type panel** de volontaires, on parle de taux aussi bas que moins de 1%, ce qui empire probablement le problème de représentation et amène les firmes à ne pas renouveler leurs panels.
- On pourrait atteindre des taux de réponse similaires aux sondages auto-administrés "classiques" en ayant recours à une gestion très serrée, des rappels, etc. (Dillman, 2000)

L’échantillon et le taux de réponse

Impact de la manière de procéder

- Multiplication du nombre de requêtes reçues
  - Augmentation du fardeau pour les répondants, et encore plus pour certains répondants plus “rares”
    - Qui finissent par ne plus répondre,
    - Ce qui provoque une nouvelle hausse du nombre de requêtes,
    - Ce qui entraîne le taux de réponse à la baisse.
  - Au final: des échantillons de répondants professionnels qui ont du temps et sont attirés par les possibilités de rémunération ou par le sujet de l’enquête. On estime que 3% des internautes complètent plus de 80% des sondages Internet (Rivers, Yougov).
- **Les problèmes se combinent pour biaiser les échantillons.**

Blasius et Brandt (2010)

- Comparent un échantillon de panel représentatif des 18-49 ans à des échantillons “étalon” en face à face (GSS et Micro-recensement allemands).
  - Ils constatent qu’il est impossible d’avoir suffisamment de répondants de 50 ans et plus.
  - Ils réussissent à faire un échantillon des 18 à 49 ans représentant cette population de façon proportionnelle selon l’âge, le sexe et le niveau d’éducation.
- Les comparaisons (y compris après pondération) avec les 2 autres échantillons montrent, dans le sondage Internet:
  - Plus de célibataires ou divorcés, de personnes sans enfants.
  - Moins de personnes fréquentant l’église.
  - Plus de valorisation du laisser faire, de la richesse, de la réalisation de soi.

Stephenson et Crête, 2011

- Comparent deux sondages faits par Léger Marketing en 2007 avec le même questionnaire, un par panel web, l'autre par téléphone.
- 36 des 52 variables ont des distributions significativement différentes même après pondération.
- Dans le panel web, moins de pratiquants religieux (idem de Blasius et Brandt, 2010), des répondants plus éduqués, plus susceptibles de penser que l'on a été trop loin pour accommoder les minorités culturelles au Québec (82% vs 76%).

Malhotra & Krosnick, 2007

- Comparent l'American National Election Study de 2000 et de 2004 à des sondages Internet de type panel de volontaires (U.S.A.).
- Même après pondération, les sondages internet comportent:
  - Moins de Noirs, moins de personnes peu scolarisées et près de deux fois plus de personnes ayant une scolarité moyenne.
  - En 2004, plus de personnes ayant voté, plus de partisans de Bush (que de Kerry), plus de partisans de la guerre en Irak, plus de personnes intéressées par la politique.
  - En 2000, à peu près les mêmes différences qu'en 2004. À peu près deux fois plus de "strong Republicans".
  - Les relations entre les variables prédisant le vote et le vote lui-même étaient significativement différentes dans presque tous les cas.

Pasek et Krosnick (2010)

- Comparent un sondage téléphonique de type RDD et un panel Internet de volontaires sur l'intention de collaborer au recensement de 2010 et la collaboration effective. Les sondages Internet de type panel...
  - Ont une moins bonne distribution démographique.
  - Diffèrent en moyenne de 13 points et jusqu'à 30 points dans les proportions des réponses modales.
  - Présentent des différences significatives et non insignifiantes dans la prédiction de la participation au recensement de même que dans l'évolution dans le temps des opinions et dans les relations entre les variables.

Durand, 2012, 2013

- Analyse des échantillons de deux firmes:
  - Sous-représentation des moins de 35 ans.
  - Profil des minorités linguistiques non plausibles (ex: forte proportion de non francophones vivant hors Montréal).
  - Profil politique de certains groupes invraisemblable:
    - Trop forte proportion d'électeurs PQ chez les 18-24 ans.
  - Échantillon difficile à redresser (relation vote précédent - vote actuel improbable).
- Qualité des estimations des sondages internet électoraux:
  - Canada 2011: sous-estimation des Conservateurs.
  - Alberta 2012: sous-estimation du Wild Rose Party.
  - Québec 2012: surestimation de la CAQ.
  - BC 2013: surestimation du NPD.

Peut-on corriger la situation?

Plusieurs méthodes (AAPOR 2013)

- Pondération par score de propension
- Pondération utilisant des variables de style de vie (ex: CROP 3SC)
- Sample matching avec utilisation de variables externes supplémentaires.

Loosveldt et Sonck (2008)

- Comparent un panel Internet de volontaires et un sondage face à face en Belgique (Flandre).
- Corrigent par un score de propension basé sur l'accès à Internet (comme Bigot et coll.).
  - 18% des personnes ayant moins de 10 ans de scolarité ou ayant 60 ans et plus ont accès à Internet, comparé à plus de 80% des personnes en emploi, ayant une scolarité universitaire, ayant 30 ans ou moins.
- La pondération par score de propension
  - Permet d'ajuster pour les différences dans la proportion d'urbains et de personnes en emploi.
  - Mais les différences demeurent significatives pour: la satisfaction face à l'emploi (Internet -), l'intérêt pour la politique (+), les attitudes face aux immigrants (plus négatives chez les répondants au panel Internet).

Tourangeau et coll. 2013 In Baker (2013)

- Font la synthèse de 8 études qui ont tenté de réduire les biais des panels de volontaires en utilisant différents ajustements et pondérations.
- Ils concluent que…
  - Les ajustements ne corrigent qu'une partie des biais, au plus 60%.
  - Les ajustements augmentent parfois le biais des estimations non ajustées, jusqu'à un facteur de plus de 2.
  - Les biais peuvent être importants, après ajustement, les estimations se modifiant d'autant que 20 points.
  - Il y a d'importantes différences selon les variables, les ajustements ayant pour effet tantôt d'éliminer le biais, tantôt de l'augmenter énormément.

Pour ce qui est des défis classiques,…

- Différences fréquentes entre les sondages Internet de type panel de volontaires et les méthodes probabilistes en face à face ou au téléphone.
  - Des différences non constantes, non systématiques, plus importantes que les différences entre utilisateurs et non utilisateurs d'Internet sélectionnés au hasard.
  - Une différence constante: moins de pratique religieuse, attitude plus négative face à la diversité.
- Lorsque des quotas sont appliqués avec succès, on constate des problèmes similaires à ceux d'autres échantillons par quotas: le respect des quotas n'assure pas une représentation sociopolitique adéquate.
- Les ajustements réduisent parfois mais n'éliminent pas les biais dus à la couverture, à la sélection et à la non- réponse inhérents aux panels de volontaires.

Les questions et le questionnaire: de nouveaux défis

- Le questionnaire est un instrument de mesure **et** une interaction sociale.
  - Dans une entrevue, l'interviewer peut "aider" le répondant au besoin, expliquer.
  - Dans un questionnaire en format papier, le répondant voit l'instrument au complet, les questions qui viennent avant et après, leur nombre, leur forme.
- Dans un sondage internet, on présente des "pages", une par une. Le répondant perd de vue ce qui précède et n'a pas de vision de ce qui s'en vient. Il faut adapter le questionnaire en conséquence.

Des défis qui valent pour tous les sondages internet

- Les questions “disparaissent” aussitôt répondues, ce qui rend la tâche du répondant plus difficile. Ça demande de...
  - Donner des indications sur la progression du questionnaire.
  - Grouper les variables en format tableau lorsque le choix de réponse est le même pour une série d’énoncés
    - voir problème de satisficing.
  - Insérer des phrases de transition lorsqu’on fait sauter des questions pour permettre au répondant de mieux se situer.
  - Permettre la non réponse, sauf lorsque la réponse est essentielle et la question non sensible, pour éviter les réponses “au hasard”.
  - Permettre le retour en arrière.
- Certaines de ces recommandations sont implantées mais pas toujours.

Des défis qui valent pour tous les sondages internet

- Certains sondages empêchent la non réponse sur des questions d'opinion.
  - Problème de questionnaires incomplets et de fiabilité des réponses.
- Certains questionnaires sont beaucoup trop longs.
  - Problème de réponses trop rapides.
    - La plupart des firmes rejettent les questionnaires répondus trop rapidement.
- Certaines questions demandent une réflexion en profondeur ou le recours à des ressources externes.
  - Problème de fiabilité des réponses.

État des lieux et perspectives d'avenir

État des lieux et perspectives d'avenir

La couverture

- État des lieux :
  - Internet devient de plus en plus accessible.
  - Il existe des moyens coûteux de rejoindre presque toute la population en installant internet à ceux qui ne l'ont pas. Pas la solution idéale.
- L'avenir:
  - Avancées technologiques dans l'accès à Internet partout **et** à moindre frais.
  - L'adresse internet appelée à devenir une adresse unique à chacun, comme l'adresse postale et le n.a.s.?

État des lieux et perspectives d'avenir

La base de sondage

- État des lieux :
  - Développement de diverses méthodes de recrutement non probabilistes, plus diversifiées.
  - Développement de méthodes probabilistes (recrutement téléphonique, face à face et postal).
- L'avenir:
  - Constitution d'une liste d'adresses de toute la population, type du processus RDD pour le téléphone.
    - lié à la couverture et aux avancées technologiques.
  - Développer les bases existantes pour inclure une plus grande partie de la population via recrutement téléphonique.
  - Utiliser le plein potentiel des bases permettant des panels et utiliser les informations disponibles dans ces bases.

État des lieux et perspectives d'avenir

La constitution de l’échantillon de départ

- État des lieux:
  - On semble être retourné à l’époque des échantillons ouverts remplis par quotas, pour la plupart sinon tous les panels internet de volontaires. Cela a un impact sur la collaboration.
- L’avenir:
  - Revenir à des méthodes fiables connues: échantillons fermés et processus de maximisation de la collaboration.

État des lieux et perspectives d'avenir

La collaboration, le taux de réponse

- État des lieux:
  - Le taux de réponse n’est pas considéré important avec des échantillons de type quotas.
    - Incitatifs monétaires.
- L’avenir:
  - Faire diminuer le nombre de requêtes aux membres des panels, entre autres,...
    - Améliorer les méthodes pour maximiser le taux de réponse: Rappels avec des messages diversifiés Dillman (2000).

État des lieux et perspectives d'avenir

La pondération, les ajustements

- État des lieux :
  - Beaucoup de recherches basées sur le pairage, la probabilité d'être inclus (score de propension), le recours à des variables de style de vie.
  - Résultats décevants en général.
- L'avenir:
  - **Poser la question**: Existe-t-il un, des moyens de s'assurer qu'un échantillon de volontaires est représentatif de la population dans toutes les circonstances? Ou dans certaines circonstances? Ou d'en arriver à des biais prévisibles?
  - Se concentrer sur l'ajustement d'échantillons probabilistes d'utilisateurs?
  - Se concentrer sur la production d'échantillons probabilistes?

État des lieux et perspectives d'avenir

Les questions, le questionnaire

- État des lieux :
  - Beaucoup de recherches faites
    - Sur les questions
    - Sur les listes (format tableau)
    - Sur le satisficing
    - Sur la non réponse (permise ou pas)
    - Sur les indications au répondant (sur la progression, entre autres).
- L'avenir :
  - Utiliser le plein potentiel offert par le mode internet, entre autres utiliser les questions ouvertes, devenues beaucoup plus faciles à traiter.
  - Trouver les moyens techniques pour permettre au répondant de répondre comme si c'était en format papier (avec une vue d'ensemble).
  - Établir des balises pour les filtres, la longueur maximale d'une liste et d'un questionnaire.

En conclusion

- Les sondages Internet sont l'avenir mais... on est dans le présent.
- Potentiel très intéressant à condition de résoudre les problèmes de base:
  - En l'absence de base de sondage fiable, les coûts pour faire un sondage avec échantillon probabiliste sont prohibitifs.
  - La manière de procéder entraîne des taux de réponse inacceptables.
  - Les tentatives d'ajuster les échantillons après coup n'ont pas donné les résultats espérés.
  - Le questionnaire n'est pas encore un instrument global et convivial.
- Il faut centrer la recherche sur la constitution de bases de sondage.

Références

- Baker et coll. (2013). Report of the AAPOR task force on non-probability sampling. http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=6055, 128 p.
- Bigot, R, P. Croutte et F. Recours (2010). *Enquêtes en ligne : peut-on extrapoler les comportements et les opinions des internautes à la population générale?* Centre de recherche pour l'étude et l'observation des conditions de vie (CREDOC), texte manuscrit, 63 p.
- Blasius et Brandt (2010) Representativeness in Online Surveys through stratified samples, *BMS, 107,* p. 5-21.
- Dillman, D. (2000). *Mail and Internet Surveys, the Tailored Design Method*, New York: Wiley and Sons, 463 p.
- Durand, C. (2013). Why do Polls go Wrong Sometimes? The Canadian Case. Présenté à UBC, 23 septembre 2013.
- Durand, C. (2013). How do Internet Polls Fare in Predicting Election Results? Présenté au World Social Science Forum, Montréal, 15 octobre 2013.
- Loosveldt et Sonck (2008) An evaluation of the weighting procedures for an online access panel survey, *Survey Research Methods, 2 (2),* 93-105.
- Lozar Manfreda, K., Bosnjak, M., Berzelak, J. Haas, I. et V. Vehovar (2008). Web surveys versus other survey modes. A meta-analysis comparing response rates. International Journal of Market Research, *50 (1),* 79-104.
- Malhotra et Krosnick (2007)Malhotra, N et J. A. Krosnick (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples *Political Analysis, 15(3)*, 286-323.
- Pasek et Krosnik (2010) *Measuring Intention to Participate and Participation in the 2010 Census and their Correlates and Trends: Comparison of RDD Telephone and Non-Probability Internet Survey Data*, Statistical Research Division, US Census Bureau, Washington, Study Series Survey Methodology #2010-15, 71 pages
- Stephenson et Crête (2011) Studying Political Behavior: A Comparison of Internet and Telephone Surveys, *International Journal of Public Opinion Research, 23 (1)*, 24-49.

Pour obtenir une copie, d'autres présentations

http://www.mapageweb.umontreal.ca/durandc