L'élaboration d'un plan de sondage probabiliste pour une enquête sortie des urnes L'enquête PEOPLE2022 à Roubaix

Thomas Soubiran
Ceraps (UMR 8026 CNRS – Université de Lille)

Séminaire Vendredis Quanti

PACTE, Grenoble, 26 mai 2023

Introduction

- présentation du plan de sondage
- élaboré pour deux enquêtes sorties des urnes
- réalisées à Roubaix lors des présidentielles de 2022
- dans le cadre du projet PEOPLE2022

Introduction

Le plan de sondage des deux enquêtes est un tirage :

- à deux degrés
- par grappes
- coordonné
- équilibré
- par rejet

Introduction

- comme il serait difficile

  et par forcément pertinent

- de passer en revue tous les aspects du plan en détail
- ce qui suit vise plutôt à

  - présenter la démarche en général
  - c–à–d les problèmes et les solutions qui y ont été apportées
  - et comment la résolution de ces problèmes peut s'appuyer sur les techniques d'enquêtes
  - fondées dans la théorie des sondages
  - particulièrement dans l'utilisation de l'information auxiliaire et des ressources disponibles

Introduction

- la préparation d'une enquête nécessite en effet
- de prendre de nombreuses décisions
- qui auront des conséquences sur le déroulement de l'enquête
- ainsi que les données en résultant
- et donc les résultat du traitements des données
- en l'occurrence,
  - combien de bureaux de vote enquêtés
  - quels bureaux de votes
  - combien d'enquêteurs par bureau de vote
- les techniques d'enquête permettent d'étayer ces décisions
- en s'appuyant sur les données pour
- minimiser le biais et la variance des estimateurs

Rappels sur les plans de sondages

Rappels sur les plans de sondages

Pour commencer,

- rappel de quelques aspects
- de la théorie des plans sondages en général
- utiles pour la présentation du plan de l'enquête ensuite

Sondage

L'enquête par sondage signifie généralement

- faire des observations sur un nombre limité d'unités
- c–à–d recueillir un échantillon
- pour tirer des conclusions sur la population dont il est issu

  inférence

Toutefois,

- l'inférence à partir d'un échantillon
- ne caractérise pas un sondage en propre
- c'est p. ex. aussi le cas pour les plans d'expérience
- et pas seulement
- parce qu'en général, $n \neq \text{ALL}$
- contrairement à ce qui a pu être affirmé par le passé...

Sondage en population finie

La particularité des sondages,

- est plutôt de recueillir un échantillon
- d'une population de taille fixe
- qui se traduit par un taux de sondage

$$f = \frac{n}{N}$$

avec n la taille de l'échantillon et N la taille de la population

- on parle alors d'inférence en population finie
- par opposition à l'inférence en population infinie

comme les plans d'expérience où il n'y a aucune notion de taux de sondage

Sondage en population finie

De plus,

- on cherche généralement

  mais pas nécessairement

- à ne sélectionner les unités qu'une seule fois au plus

  sélection sans remise ou tirage sans remise (si on introduit de l'aléas dans la sélection)

Autre particularité :

- utilisation d'informations auxiliaires

  c–à–d des informations sur la population connues au préalable de la collecte

- à la fois pour organiser la collecte
- et améliorer la qualité des estimateurs

  biais et précision

Coordination d'échantillons

Cas particulier : les suivis dans le temps (panels)

- où des unités sont enquêtées plusieurs fois
- parfois dans le cadre d'une rotation
  - où on planifie la sortie progressive des unités
  - et leur remplacement
  - pour limiter la charge
  - et tempérer la contraction inéluctable de l'échantillon
- on parle alors de coordination positive
  - qui vise à maintenir des unités dans l'échantillon
  - pour un nombre prédéterminé de fois
- par opposition à la coordination négative
  - qui vise à exclure de l'échantillon des unités déjà sélectionnées
  - et qui a été appliquée ici à la sélection des lieux de vote au second tour

Plan de sondage

La théorie des sondage est une application de la théorie des probabilité

- un plan de sondage est une loi de probabilités
- un échantillon est la réalisation d'une variable aléatoire
- comme la population est finie, l'échantillonnage est généralement
- un tirage sans remise

Ce qui a de profondes conséquences sur les estimateurs

- car la théorie habituelle n'a plus cours
- le tirage sans remise induit une dépendance entre les tirages
  - les probabilités de sélectionner une unité au t^e tirage
  - dépendent en effet des $t-1$ tirages précédents

La théorie des sondage a notamment été développée pour y pallier

- ce qui a notamment pour conséquence
- que les estimateurs des paramètres d'intérêt et de leur variance est souvent différente

  comme on le verra plus loin pour les tirages stratifié et par grappes p. 46

Exemple : tirage aléatoire à probabilités égales

SASR

- tirage aléatoire à probabilités égales :

  tirage où toutes les unités ont la même probabilités d'être sélectionnées

- lorsqu'il est réalisé AVEC remise (SASR) en population infinie
- on a
  - estimateur de la moyenne :

$$\hat{\bar{y}} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{1}$$

  - estimateur de la variance de la moyenne :

$$\widehat{\mathbb{V}(\bar{y})_{SASR}} = \frac{s^2}{n} \tag{2}$$

  avec $s^2 = \sum_{i=1}^{n} (y_i - \hat{\bar{y}})^2/(n-1)$

SASSR

Lorsqu'il est réalisé SANS remise (SASSR) en population finie

- l'estimateur de la moyenne est identique à l'estimateur habituel
- par contre, l'estimateur de la variance a pour expression

$$\widehat{\mathbb{V}(\bar{y})_{SASSR}} = (1 - f)\frac{s^2}{n} \tag{3}$$

Dans ce cas,

- seule la correction de population de finie $(1 - f)$

  qu'on retrouve plus ou moins explicitement dans de nombreuses formules d'estimateurs de la variance sous le plan

- distingue la formule de la formule habituelle
- et qui suggère que la variance d'un SASSR sera plus petite que celle d'un SASR

  cf. plus loin DEFF p. 53

SASR et SASSR

Lorsque N est suffisamment grand et f petit,

- $(1 - f)$ peut être ignoré
- et on peut donc utiliser les estimateurs habituels

Mais cela vaut seulement pour le SASSR

- qui, dans les faits, est rarement utilisé
- en tout cas, pas tout seul
- parce que trop coûteux pour être mise en œuvre efficacement

C'est pourquoi on doit généralement recourir à des estimateurs propres à chaque plan

L'estimateur d'Horvitz–Thompson

Exemple d'estimateur sous le plan : l'estimateur d'Horvitz–Thompson

- l'estimateur d'Horvitz–Thompson est un estimateur linéaire non–biaisé
- p. ex. pour un total

$$t = \sum_{k \in \mathcal{U}} Y_k \tag{4}$$

avec Y_k la valeur prise par la caractéristique y dans la population

- on a

$$\hat{t}_{\text{HT}} = \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k} = \sum_{\substack{k \in \mathcal{U} \\ \pi_k > 0}} \frac{Y_k \mathbb{1}_k}{\pi_k} \tag{5}$$

avec y_k la valeur prise par la caractéristique y dans l'échantillon , $\mathbb{1}_k$ l'indicatrice de la présence de l'unité k dans l'échantillon et $\pi_k = P(k \in \mathcal{S}$ la probabilité de sélection de l'unité k pour un plan donné

- l'estimateur d'Horvitz–Thompson revient donc à diviser les valeurs de l'échantillon par leur probabilités de sélection

Note :

- $y_k \mathbb{1}_k$ est une variable aléatoire
- mais Y_k ne l'est pas...

L'estimateur d'Horvitz–Thompson

- pour la moyenne

$$\bar{y} = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k \tag{6}$$

- on a

$$\hat{\bar{y}}_{\text{HT}} = \frac{1}{N} \hat{t}_\pi = \frac{1}{N} \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k} \tag{7}$$

L'estimateur d'Horvitz–Thompson

- p. ex., pour l'ESU,
  - sondage à deux degrés
  - au 1er tour
    - tirage des lieux de vote : 15 LdV sur 30 ($\pi_h = .5$)
    - tirage des votants dans chaque LdV : 1 votant sur 5 ($\pi_{hi} = .2$)
  - comme les probabilités de sélection sont indépendantes à chaque degré,
  - il suffit de multiplier les probabilités de sélection à chaque degré
  - pour obtenir π_k

$$\pi_k = \left(\frac{15}{30}\right)\left(\frac{1}{5}\right) = .1 \text{ pour le } 1^{\text{er}} \text{ tour}$$

(en fait, pas tout à fait pour des raisons développées plus loin p. 89 mais c'est pour l'exemple)

L'estimateur d'Horvitz–Thompson

- les probabilités de sélection pour le 1^{ier} tour
- sont donc identiques pour tous les votants
  - mais ce n'est généralement pas le cas pour les plan de sondage à plusieurs degrés
  - c–à–d que des unités peuvent se retrouver sur ou sous–représentées
  - mais, en divisant par la probabilité de sélection
  - l'estimateur HT permet de redonner aux unités leur poids réel dans la population
  - aussi appelé estimateur par expansion

Probabilités de sélection

- les π_k ne doivent remplir que des conditions très générales

$$0 \leq \pi_k \leq 1 \text{ et } \sum_{k \in \mathcal{U}} \pi_k = n_{\mathcal{S}} \tag{8}$$

  - avec $\pi_k > 0$ en plus pour obtenir des estimateur non–biaisés

    c–à–d qu'il n'y a pas de défaut de couverture

  - ce qui suggère que les estimations sont, dans les faits, souvent biaisées (non–réponse)

Probabilités de sélection

- autrement dit, le tirage n'a pas à respecter de « quotas »
  - ou une quelconque similarité entre l'échantillon et la population,
  - y compris pour le tirage équilibré
- c'est une des raisons pour laquelle les statisticiens d'enquête rejettent complètement la « représentativité »
- car,
  - il n'y a aucune raison théorique de ne pas modifier les probabilités de sélection
  - il y a toutes les raisons empiriques de modifier les probabilités de sélection
- ne pas sur–représenter des unités peut en effet conduire à des estimations très biaisées ou imprécises

L'enquête selon Hajek

- pour Hajek une enquête est une stratégie composée
  - d'un plan $p(\cdot)$
  - et un estimateur $\hat{y}$

    pas de sondage omniscient, le plan est conçu uniquement pour certaines caractéristiques
- auxquels on peut ajouter
  - des ressources
  - et des variances

Inférence sous le plan

- le plan est la seule source de l'aléas
  - et c'est sur l'aléas du plan que l'estimation se fonde
  - sans postuler de distribution sous–jacente
  - les variables sont des caractéristiques ou des critères fixes
  - seules les variables indicatrices $\mathbb{1}_k$ sont aléatoires
- en conséquence de quoi les estimateurs et leur variance dépendent de la façon dont l'échantillon est sélectionné
- et sont donc propres à chaque plan

Inférence sous le plan

- la casualisation

  « randomisation »

- et donc le fondement de l'inférence

  et non pas la ressemblance

- et revêt une importance d'autant plus grande
- que, plus un échantillon sera aléatoire,

  c–à–d plus son entropie sera forte

- et meilleure sera l'inférence

Le plan de sondage de l'enquête

- le plan de sondage des deux enquêtes a été conçu dans le cadre
- du projet PEOPLE2022

  Pratiques Électorales et OPinions Lors des Élections de 2022

- associant le CERAPS et ESPOL
- et qui vise à
  - explorer la participation des citoyens français aux campagnes électorales et aux élections elles-mêmes,
  - ainsi que l'influence des médias Web et papier
  - sur leur comportement électoral.

People2022

- une partie du projet People2022 visait à réaliser
  - des enquêtes sortie des urnes
  - lors des présidentielles dans la commune de Roubaix
- l'ESU est une enquête électorale qui,
- comme son nom l'indique,
- consiste à interroger les votants à la sortie de leur bureau de vote

ESU

- l'utilisation des ESU s'est développée notamment aux États–Unis à partir des années 1960
- à fin d'estimer le résultat des élections
- les ESU ont été utilisés plus tardivement en France
  - et elles y ont connu des fortunes diverses cf. Pina (2019)
  - jusqu'à leur quasi–disparition
  - les ESU ont toutefois connu un regain d'intérêt ces dernières années
  - mais plutôt à des fins d'analyse des comportements électoraux
- les ESU comportent divers avantages et défauts

  mais comme toutes les différents types d'enquêtes électorales

ESU

- la particularité des ESU est de passer par l'entremise des bureaux de vote
- pour inclure des votants dans l'échantillon
- au moins deux façons de procéder
  - tirage stratifié
  - tirage par grappe
- recours à un tirage par grappe
- qui nécessite au préalable de sélectionner des bureaux de vote

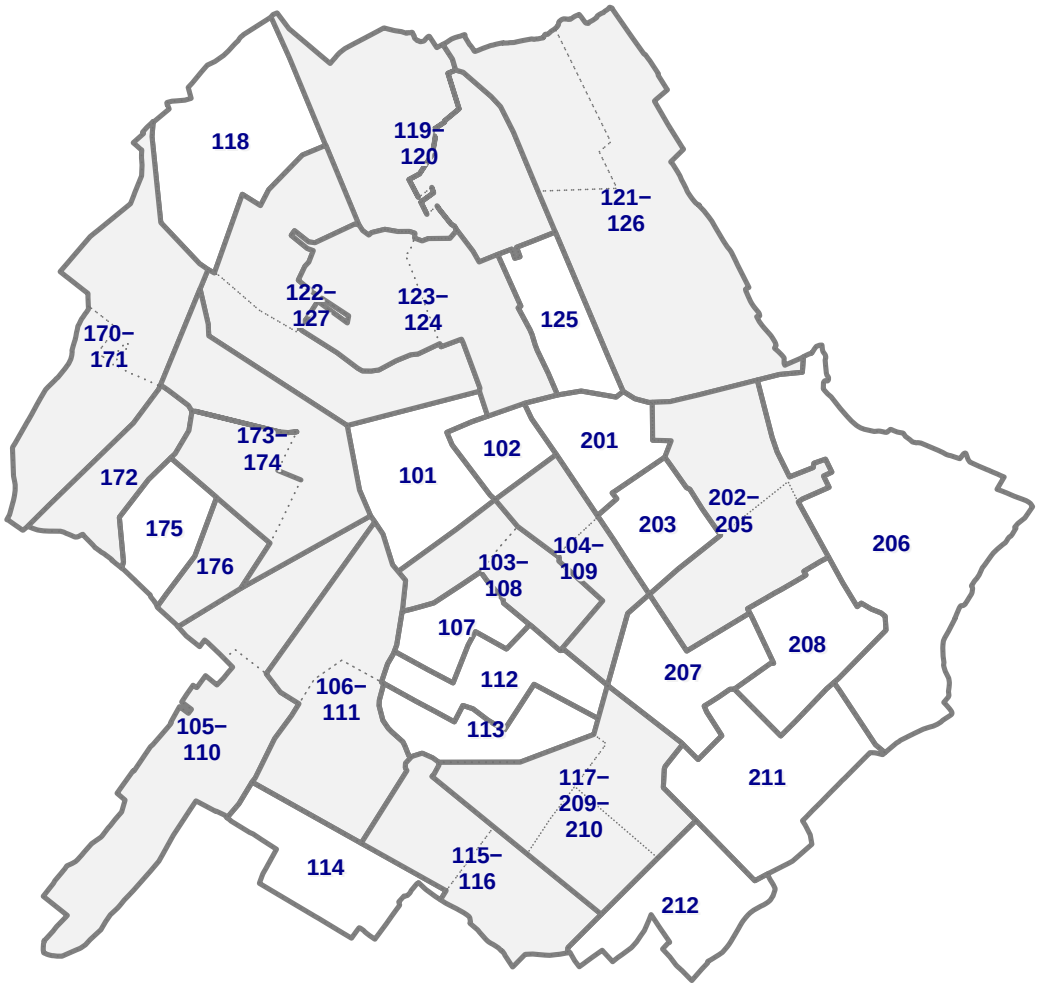
ou, pour être plus précis des lieux de vote (LdV)

Bureaux et lieux de vote

- distinction importante pour l'organisation de ESU
  - les bureaux de vote (BdV)
  - les lieux de vote (LdV)
- en effet, plusieurs bureaux peuvent correspondre au même lieu

  p. ex. le même réfectoire d'une école primaire
- ce qui complique la sélection des votants
- c'est pourquoi, les 45 bureaux ayant le même lieu de vote
- ont été fusionnés en 30 lieux de vote
- compliquant d'autant le tirage en réduisant la taille de l'univers du plan de sondage

Bureaux et lieux de vote à Roubaix



Sélection des votants

- pour présenter la sélection des LdV,
- il faut d'abord présenter la sélection des votants
- ainsi que les ressources mobilisables
- la sélection des LdV dépendant du nombre d'enquêteurs disponibles
- mais aussi des modalités de collecte dans les LdV

Le tirage des votants

- du fait de moyens limités
- l'enquête a reposé sur le volontariat étudiant
  - avec la difficulté supplémentaire que les deux tours tombaient **en plein pendant les révisions**

    les partiels commençant pour certains le lendemain du 2^{nd} tour
  - ce qui renforçait d'autant plus l'incertitude sur le nombre d'enquêteurs
  - et repoussait d'autant la conception du plan de sondage
- **le nombre d'enquêteurs** disponible étant littéralement **fondamental**
- et laissait planer un doute sur la possibilités même de concevoir un plan de sondage

  nombre minimal en–dessous duquel la sélection aurait dû être réalisée de façon *ad hoc*

Les enquêteurs

- diffusion d'un appel en janvier auprès de différentes filières de sciences sociales
- de l'Université de Lille et de la FUPL ainsi que SciencePo Lille

  Fédération universitaire et pluridisciplinaire de Lille

- on a été fixé environ début février avec
  - $\simeq$ 80 volontaires pour le 1^{ier} tour
  - $\simeq$ 40 volontaires pour le 2nd
  - au final, 101 enquêteurs en tout
- ce qui, au final, a laissé $\simeq$ un mois pour la conception du plan
- la liste des bureaux devant être transmise à l'avance

  mairie de Roubaix, préfecture

- il a donc fallu faire vite

Tirage systématique

- sélection des votants par **un tirage systématique**
- en l'absence de plus d'informations
- le tirage systématique (SY) consiste à
  - compter les votants qui se présentent
  - et à prendre tous les G$^{\text{ièmes}}$ votants
- soit plus formellement :
  - on génère un nombre h entre 1 et G
  - on sélectionne les unités avec le n° d'ordre :

    $$h, h+G, h+2G, ..., h+(n-1)G, ...$$

    c–à–d les unités dont les indices sont congruents modulo G
    —$(k-h) \equiv 0 \pmod{G}$— (arithmétique modulaire)
- ce qui revient à
  - utiliser l'ordre d'arrivée
  - pour le diviser en intervalles
  - puis à sélectionner une unité dans chaque intervalles

Tirage systématique

- intérêt du SY : **facile** à mettre en œuvre pour obtenir un tirage sans remise à probabilités égales
- mais c'est bien **le seul**
- il existe en effet une littérature conséquente pour souligner la multiplicité de ses défauts
  - tirage à faible entropie

    alors que ce qu'on veut, c'est un tirage à forte entropie
  - ses performances dépendent de l'aléas dans l'ordre des unités

    si on peut permuter les unités avant, le SY est équivalent à un SASSR
  - entre autres problèmes

    imputables à la périodicité de l'ordre des unités
  - difficultés aussi à définir un estimateur non–biaisé de la variance

    les probabilités d'inclusion de second ordre valent zéro

Mise en œuvre du tirage systématique

Motivation du SY :

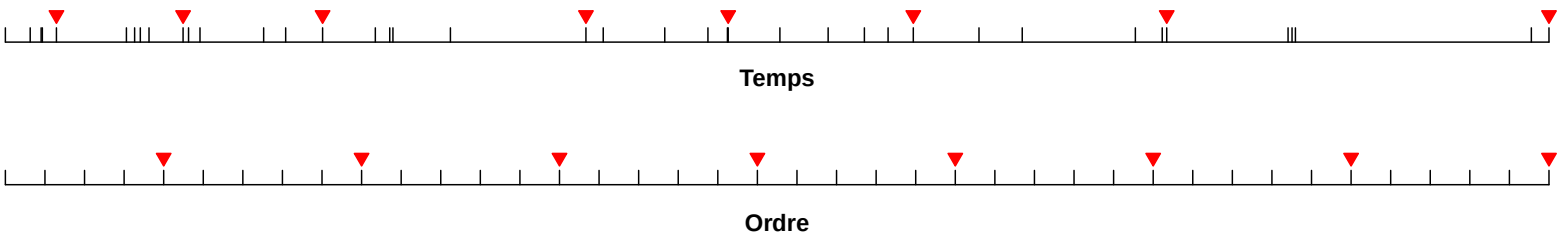
- pas de liste des votants disponible au préalable
- le tirage devait donc être réalisé
- immédiatement par les enquêteurs eux–même

Mise en œuvre du tirage systématique

- le tirage systématique implique **au moins trois enquêteurs** par LdV
  - un qui compte et sélectionne
  - un qui prend contact
  - un autre pour le cas où un autre électeur dans la pas se présenterait alors que l'autre enquêteur est occupé
- le troisième enquêteur est là pour pallier **le flot irrégulier** des électeurs

  le tirage systématique repose sur l'ordre d'arrivée et ne prend pas en compte les durées inter-événements
- les intervalles peuvent donc **se chevaucher**

  particulièrement à certains moments de la journée —fin de la matinée et le début de l'après–midi—
- et risque de **faire perdre** la cadence



Graphique : Tirage systématique

Taux de sondage

- le nombre total d'enquêteurs était ensuite déterminé au *pro rata* du nombre d'inscrits

  entre 3 et 7 enquêteurs

- un minimum de deux enquêteurs disponibles
- était aussi motivé par le choix **d'un taux de sondage** élevé
- du fait de **la faible participation** aux élections dans la commune

  38% d'abstention aux présidentielle de 2017 et 41% en 2022

- au final, un taux de $^1/_5$ a été appliqué

  - soit 20% des votants
  - c–à–d que tous les $5^{\text{ièmes}}$ votants étaient interrogés
  - taux identique pour tous les LdV

- **Note :** les électeurs souhaitant participer spontanément à l'enquête se sont vus remis un questionnaire non–numéroté

Refus de répondre

- en cas de refus
  - la personne suivante était interrogée
  - en cas de refus
  - la personne suivante était interrogée
  - ...
  - jusqu'à la $5^{\text{ième}}$ où le tirage systématique reprenait
  - si la personne répondait
  - dans le cas contraire, la personne suivante était interrogée
  - ...
- l'idée étant de prendre quelqu'un dans l'intervalle
- qui visait à ne pas trop être pénalisé par la non–réponse
- qu'on ne pouvait pas estimer *a priori*
- et qui constitue **une entorse** au plan
- **Note :** sous le plan, la non–réponse **n'existe pas**

Digression : quotas

- petite digression : la sélection des votants permet d'illustrer certains problèmes de la méthode des quotas
- à partir des listes électorales,
  - on peut déterminer les distributions marginales
  - ainsi que la distribution jointe sexe–âge par LdV
  - mais il s'agit des inscrits et non des votants
  - distorsions au moins sur l'âge
- plus généralement, dans ce cas
  - ce sont les enquêteurs qui « échantillonnent » les votants
  - mais de façon non aléatoire

    pas de casualisation
  - et ça, d'autant plus que la sélection devient alors une interaction sociale comme une autre

    avec, notamment, ce que ça implique « d'affinités électives » entre enquêteurs et enquêtés

Digression : quotas

de plus,

- les quotas font reposer une charge très lourde sur les enquêteurs
- avec notamment des difficultés cognitives

  deviner l'âge des votants

- remplir la feuille de quotas devient de plus en plus difficile au fur et mesure du déroulement de l'enquête
- or, le taux de sondage étant conséquent
- la sélection devait être opérée rapidement
- d'où l'avantage du SY dans ce cas

Digression : quotas

- souligner les défauts ne veut pas dire invalider

  rien ni personne n'est parfait en ce bas monde

- car même si la liste des défauts connus des quotas est longue
- c'est aussi une façon d'intégrer l'information auxiliaire
  - ce qui généralement bénéfique à l'estimation
  - en tout cas lorsque les informations auxiliaires sont corrélées aux caractéristiques d'intérêt
- en calant la sélection sur des distributions marginales
  - idée que l'on retrouve dans différentes méthodes
  - comme la méthode du cube qui sera présentée plus loin p. 82

    mais pour des raisons très différentes de celles généralement avancées pour les quotas

Digression : quotas

- car, à l'inverse du cas de la sélection des votants
  - dans certains cas, les méthodes reposant sur des sélections aléatoires
  - peuvent être difficiles voir impossible à mettre en œuvre
  - p. ex. lorsqu'il faut sélectionner un petit nombre d'unités dans un univers de taille réduite
- et, plus généralement,
  - des méthodes de sélection sans aléas
  - peuvent aussi être fondées théoriquement
  - inférence par le modèle
  - qui ne dispense de planifier la collecte
- l'aléas a toutefois des propriétés intéressantes
- et devrait être introduit dès que possible
- ou, pour le moins, ne pas être exclu d'office comme une impossibilité

Le tirage des lieux de vote

- maintenant qu'on a une idée
- du déroulement de la collecte lors de **la 2nde étape**
- on peut désormais envisager **la 1^{ière}**
- pour déterminer **à la fois**
  - le nombre de LdV enquêtés
  - quels sont les LdV enquêtés
  - et le nombre d'enquêteurs y étant affectés

Tirage stratifié et par grappes

- comme indiqué précédemment, un ESU
- peut correspondre à deux plans de sondage différents :
  - tirage stratifié
  - tirage par grappe

Tirage stratifié et par grappes

- dans les deux cas, la population est répartie en groupes mutuellement exclusifs

$$\bigcup_{i=1}^{M} \mathcal{U}_i = \mathcal{U} \text{ et } \mathcal{U}_i \cap \mathcal{U}_j = \emptyset, i \neq j$$

- **tirage stratifié :**

  on interroge une fraction —pas nécessairement identique— dans chaque strate h

$$\mathcal{S} = \bigcup_{h=1}^{H} \mathcal{S}_h$$

  avec $\mathcal{S}_h$ un échantillon aléatoire tiré dans la strate h avec un plan $p_h()$ et $p_h(s_h) = Pr(S_h = s_h)$. Le tirage des strates est donc indépendant.

- **tirage par grappes :**

  - on sélectionne une partie des grappes

$$\mathcal{S} = \bigcup_{i \in \mathcal{S}_I} \mathcal{U}_i$$

    avec s_I un échantillon aléatoire de grappes tiré selon un plan $p_I(s_I)$ et $\mathcal{S}_I$ un échantillon aléatoire tel que $Pr(S_I = s_I) = p_I(s_I)$ et $m = \#S_I$ le nombre de grappes sélectionnées
  - puis on interroge toutes les unités dans chaque grappe

Tirage stratifié et par grappes

- les deux tirages procèdent donc de façon très différentes
  - **tirage stratifié :** l'échantillon est l'union de H tirages indépendants
  - **tirage par grappes :** l'échantillon est un tirage de m grappes
- ce qui a d'importantes conséquences sur les propriétés des plans

Plans complexes

Dans les faits,

- tirages stratifiés et par grappes sont souvent combinés
- à d'autres méthodes tirages

Exemple : une approche courante consiste à

- d'abord réaliser un tirage stratifié
- puis un tirage par grappes à l'intérieur des strates
- tirage à deux degrés ou plus
- sans nécessairement utiliser le SASSR pour sélectionner les strates ou les grappes
- la stratification vise ici à réduire la variance d'échantillonnage dûe au tirage par grappes

Tirage à deux degrés

p. ex., pour les deux ESU,

- du fait de l'absence d'effectifs d'enquêteurs suffisant
- pour interroger tous les votants dans les grappes,
- on procède à un tirage par grappes à deux degrés :
  - on sélectionne une partie des LdV

    par un tirage équilibré avec coordination négative rejectif
  - puis on interroge une fraction des votants

    avec un tirage systématique

Tirage stratifié et par grappes

- au de–là des différences d'estimateurs
- les deux plans ne sont pas équivalents
- mais représentent deux façons différentes d'utiliser l'information auxiliaire
  - **tirage stratifié :** améliorer la précision
  - **tirage par grappes :** faciliter l'organisation de la collecte

    parfois au détriment de la précision comme on va le voir

L'effet de plan

- différents plans de sondages peuvent être comparés au moyen de l'effet de plan (*Design Effect*)
- qui consiste à diviser la variance de l'estimateur
- par la variance d'un estimateur SASSR de même taille

$$\text{DEFF} = \frac{\mathbb{V}_{p(s)}(\hat{\theta})}{\mathbb{V}_{\text{SASSR}}(\hat{\theta})} \tag{9}$$

le DEFF estime dans quelle mesure la variance d'un estimateur est sous ou sur estimé par rapport à un SAS

- en divisant la taille de l'échantillon par le DEFF

$$n_{\text{eff}} = \frac{n}{\text{DEFF}} \tag{10}$$

on obtient la taille effective de l'échantillon, c–à–d le nombre d'observations nécessaires pour obtenir le même niveau de précision qu'un SASSR

Effet de plan du SASR

- pour le SASR on a,

$$\text{DEFF}(\hat{t}_{SASR}) = \frac{\mathbb{V}(\hat{t}_{SASR})}{\mathbb{V}(\hat{t}_{SASSR})} = \frac{N^2\left(1-\frac{1}{N}\right)\frac{S^2}{n}}{N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}} = \frac{N-1}{N-n}$$

- le DEFF est donc toujours > 1 si $n \geq 2$
- ce qui confirme que le SASSR est plus précis que le SASR (cf. p. 14)
- dans ce cas, le DEFF dépend seulement de N et n
- et quand f est faible, le DEFF tend vers 1

Effet de plan du tirage stratifié

Pour un tirage stratifié

- estimateur d'un total

$$\hat{t} = N \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \tag{11}$$

- estimateur de la variance d'un total

$$\mathbb{V}(\hat{t})_{STRAT} = \sum_{h=1}^{H} \mathbb{V}(\hat{t}_h)_{SASSR} \tag{12}$$

$$= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h} \tag{13}$$

- dans le cas d'une allocation proportionnelle des strates (pour simplifier)
  - le nombre d'unités sélectionné est proportionnel à la taille de chaque strates
  - on a donc $n_h = n \times (N_h/N)$ et $n = \sum_{h=1}^{H} n_h$
  - et la variance vaut alors

$$\mathbb{V}(\hat{t})_{STRAT-P} = N^2 \left(n - \frac{1}{N}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{S_h^2}{n} \tag{14}$$

Effet de plan du tirage stratifié

- en ré–exprimant la variance du SASSR (3) pour faire apparaître explicitement les strates

$$\mathbb{V}(\hat{t})_{SASSR} = N^2 \left(n - \frac{1}{N}\right) \left[\sum_{h=1}^{H} \frac{N_h}{N} \frac{S_h^2}{n} + \sum_{h=1}^{H} \frac{N_h}{N} \frac{\bar{Y}_h - \bar{Y}}{n}\right] / n \tag{15}$$

- on obtient le $\text{DEFF}_{STRAT-P}$

$$\text{DEFF}(\hat{t}_{STRAT-P}) = \frac{\sum_{h=1}^{H} \frac{N_h}{N} \frac{S_h^2}{n}}{\frac{N_h}{N} \left[\sum_{h=1}^{H} \frac{S_h^2}{n} + \sum_{h=1}^{H} \frac{\bar{Y}_h - \bar{Y}}{n}\right]} \tag{16}$$

$$= \frac{S^2_{y(intra)}}{S_y^2} = \frac{\text{variance intra strate}}{\text{variance totale}} \tag{17}$$

Effet de plan du tirage stratifié

Pour un plan stratifié à allocation proportionnelle,

- le DEFF tendra donc à être < 1
- parce que la variance totale se réduit à la variance intra–strate

  puisqu'on collecte des informations dans toutes les strates

- et la stratification produit une variance plus faible sauf si
- les moyennes des strates sont égales
- ce qui se produit rarement dans les faits

  et d'autant moins que le nombre de strates est grand

- la stratification améliore donc d'autant plus la précision que les strates sont homogènes

  de façon générale

- et que la variance inter–strate augmente

Effet de plan du tirage stratifié

De façon, plus générale

- plus les strates sont homogènes et plus le tirage stratifié sera précis
- c'est pourquoi ne nombreux plans stratifient le tirage
- avec une variable liée à la variable d'intérêt
- la précision pouvant toutefois varier
- en fonction de la façon de déterminer la taille des strates
- on a en effet $V_{STRAT-P} < V_{opt}$
- lorsqu'on utilise l'allocation optimale de Neyman

Effet de plan du tirage par grappes

- pour le tirage par grappes, les choses sont très différentes
- et même complétement inverses

Notations pour le tirage par grappes

- N : nombre de grappes
- M_i : nombre d'unités dans la grappe i
- dans ce qui suit, on suppose

  pour simplifier

- que la taille est la même pour toues les grappes $M_i = M$

Effet de plan du tirage par grappes

- pour le DEFF du tirage par grappes, on utilise le coefficient de corrélation intra–classe

$$\rho = \frac{\mathbb{E}(y_{ij} - \bar{Y})(y_{ik} - \bar{Y}_{\mathcal{U}})}{\mathbb{E}(y_{ij} - \bar{Y}_{\mathcal{U}})} = \frac{2\sum_i \sum_{k<k}(y_{ij} - \bar{Y}_{\mathcal{U}})(y_{ik} - \bar{Y}_{\mathcal{U}})}{(M-1)(NM-1)S^2}$$

avec $S^2 = \frac{\sum_{i,j}(y_{ij} - \bar{Y}_{\mathcal{U}})^2}{NM-1}$ et $\bar{Y}_{\mathcal{U}} = 1/(NM)\sum_{i=1}^{N} y_i$

- soit le coefficient de corrélation de Pearson
- pour les $(M-1)(NM-1)$ paires y_{ij} et y_{ik}

Effet de plan du tirage par grappes

- l'ICC mesure la similarité des éléments de chaque grappe

  et n'est défini que pour des grappes de taille identique

- l'ICC peut aussi être exprimé en terme de l'ANOVA de la population

$$\rho = 1 - \frac{M}{M-1} \frac{S_{y(intra)}}{S^2}$$

avec

$$-\frac{1}{M-1} \leq \rho \leq 1$$

Effet de plan du tirage par grappes

- variance de l'estimateur de la moyenne du tirage par grappes :

$$\mathbb{V}(\hat{\bar{y}}_{GRP}) = \left(1 - \frac{1}{N}\right) \frac{\sum_i (y_i - \bar{y})^2 / (N-1)}{nM^2} \tag{18}$$

$$= \left(1 - \frac{1}{N}\right) \frac{S^2}{nM} [1 + (M-1)\rho] \tag{19}$$

avec $\bar{Y} = 1/N \sum_{i=1}^{N} y_i$ la moyenne des grappes
Note : on passe de (18) à (18) en reformulant la variance inter en terme de le variance totale et de ρ

- dans ce cas

$$\mathbb{V}(\hat{\bar{y}}_{SASSR}) = \left(1 - \frac{1}{N}\right) \frac{S^2}{nM}$$

- le DEFF vaut donc

$$\text{DEFF}_{\text{GRP}} = [1 + (M-1)\rho]$$

Effet de plan du tirage par grappes

- pour le tirage stratifié, la variabilité de l'estimateur dépend essentiellement de la variance INTRA
- pour le tirage par grappe, la variabilité de l'estimateur dépend essentiellement de la variance INTER
  - comme les grappes tendent à être homogènes
  - et hétérogènes entre elles
  - ρ est généralement > 0
  - et le tirage par grappes est généralement moins précis que le SASSR

    la perte de précision dépendant de ρ

Effets de grappes

- pour le tirage stratifié, la variance inter a moins d'importance car on recueille des informations dans toutes les strates
- ce qui, par définition, n'est pas le cas dans le tirage par grappes
- le tirage par grappes produit des situations où
  - les unités des grappes sont homogènes
  - et les grappes sont hétérogènes entre elles
- cet effet de grappe a des conséquences très concrètes sur l'organisation d'une enquête

Effets de grappes

- de façon générale, il est préférable
  - de ne retenir qu'un nombre limité d'unités enquêtées dans chaque grappe
  - et d'enquêter dans le plus grand nombre possible de grappe
  - et non pas concentrer les enquêteurs dans un nombre limité de BdV
- soit l'inverse de ce qui est spontanément fait...
- car
  - plus les unités des grappes sont homogènes et plus le gain d'information de chaque unité enquêtées se réduit
  - plus les grappes sont hétérogènes et plus on perd de l'information en interrogeant pas d'autres grappes

Remarques supplémentaires

- de plus, la variation de la taille des grappes a aussi un effet sur la précision
  - cet effet peut être atténué par un tirage proportionnel à la taille
  - ce qui est approximativement le cas pour le tirage équilibré présenté après
- dans le cas du tirage à deux degrés, si m/M est faible
  - le poids de la variance intra devient négligeable
  - mais ce n'est pas le cas ici

Effet de plan du tirage par grappes

- lorsque $\rho > 0$, les grappes

  dans le plan ou dans les données

- peuvent avoir des effets négatifs sur les tests

  comme le χ^2 ou G^2

- plus les grappes sont homogènes et plus leur variance décroît
- de ce fait, la « vraie » valeur de p sera plus grande
- que celle calculée en ignorant les grappes

  c–à–d sous l'hypothèse SASR

- et donc accepter H_0 alors qu'elle devrait être rejetée
- dit autrement, le nombre effectif d'observations est plus petit que la taille de l'échantillon
- la stratification a l'effet exactement inverse

  rejet de H_0 alors qu'elle est vraie

Tirages des BdV par grappe

- au regard de l'échantillonnage à la 2nd étape
- le tirage stratifié était inenvisageable

  pas assez d'enquêteurs

- de plus, comme une enquête était prévue pour chaque tour
- il y avait une probabilité non–nulle de réinterroger les mêmes personnes

  renforcé par le taux de sondage

- et donc un risque de refus plus élevé
- donc le tirage de LdV est différent pour le premier et le 2nd tour

Détermination du nombre de grappes retenues

- reste maintenant à déterminer le nombre de grappes retenues
- ce qui précède suggère de répartir
- en fonction de
  - ρ
  - et des ressources disponibles
- le nombre optimal de grappe peut être obtenu
- à partir de la formule en utilisant la fonction de coût C suivante

$$C = c_1 n + c_2 nm \tag{20}$$

  c_1 est donc proportionnel au nombre de LdV et c_2 est proportionnel au nombre de votants
- d'autres formules existent pour intégrer p. ex. le coût de déplacement

  qui est à peu près constant ici

Détermination du nombre de grappes retenues (notation)

- $$\bar{y} = \sum_{j=1}^{m} \frac{y_{ij}}{m} \tag{21}$$

$$\bar{\bar{y}} = \sum_{i=1}^{n} \frac{\bar{y}_i}{n} \tag{22}$$

- $$S_1^2 = \frac{\sum_{i=1}^{N} (\bar{y}_i - \bar{\bar{y}})^2}{N-1} \text{ (variance des LdV)} \tag{23}$$

$$S_2^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2}{N(M-1)} \text{ (variance des votants)} \tag{24}$$

Détermination du nombre de grappes retenues

- pour un sondage par grappes à deux degrés, $\mathbb{V}(\bar{\bar{y}})$ a pour expression

$$V(\bar{\bar{y}}) = \frac{1}{n}\left(S_1^2 - \frac{S_2^2}{M}\right) + \frac{1}{mn}S_2^2 - \frac{1}{N}S_1^2 \tag{25}$$

- minimiser V pour C fixe (ou C pour V fixe) revient à minimiser

$$\left(V(\bar{\bar{y}}) + \frac{1}{N}S_1^2\right) C = \left[\left(S_1^2 - \frac{S_2^2}{M}\right) + \frac{S_2^2}{m}\right](c_1 + c_2 m) \tag{26}$$

- d'où il vient que

$$m_{\text{opt}} = \frac{S_2}{\sqrt{S_1^2 - S_2^2/M}}\sqrt{c_1/c_2} \tag{27}$$

à la condition que $S_1^2 > S_2^2$ (division par 0)

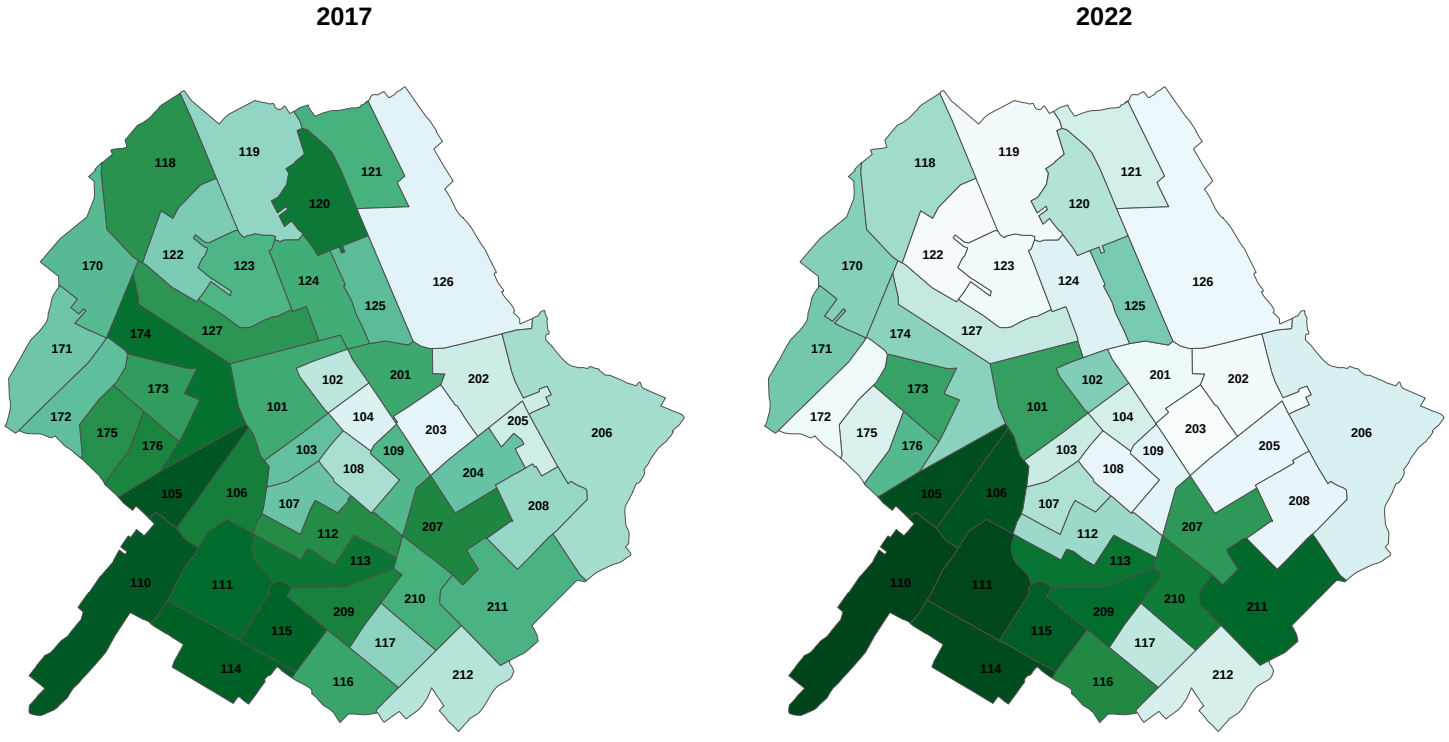
Détermination du nombre de grappes retenues

- (27) a été estimé avec le nombre de voix de J. L. Mélenchon
- au 1^{ier} tour des présidentielles de 2017
- en utilisant un budget–temps
- combiné aux tests présentés après
- et en prenant en compte qu'il fallait laisser un peu de place pour le tirage du 2^{nd} tour
- et en faisant attention à ne pas trop réduire le nombre d'enquêteurs par LdV
- m a été fixé à 15 pour le 1^{ier} tour

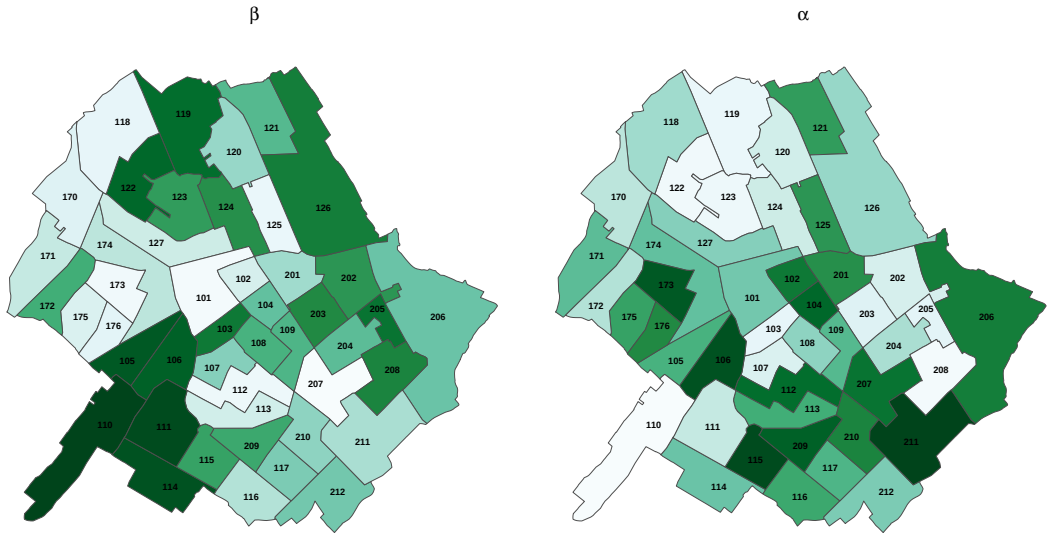
Bilan d'étape

- pour l'instant, le plan consiste dans
  - un tirage de 15 grappes

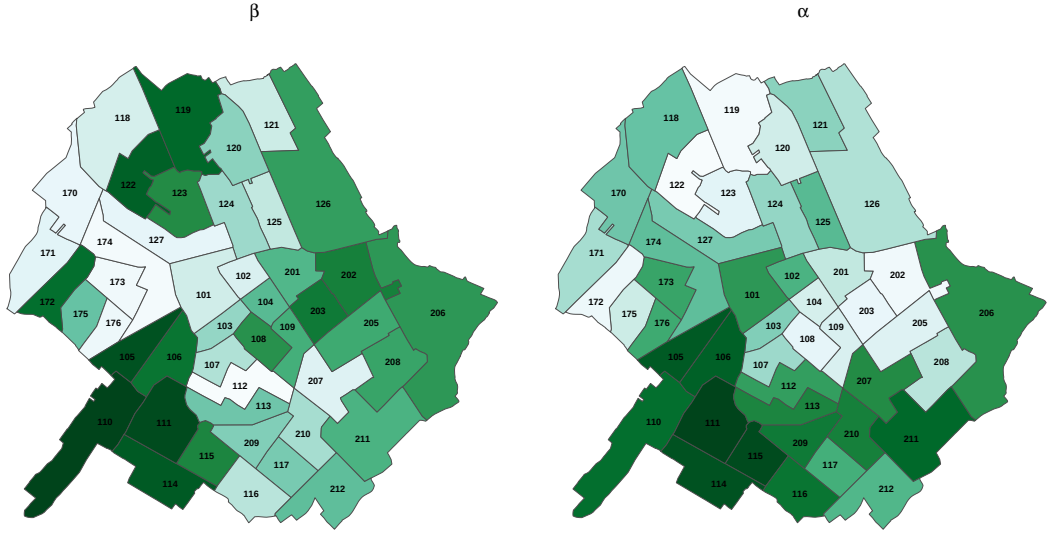
    pour le 1^{ier} tour
  - à deux degrés
- se pose maintenant la question de la sélection des grappes proprement dites
- un SASSR semble exclu
- du fait de la variabilité des résultats électoraux d'un BdV à l'autre



Graphique : Résultats de E. Macron au premier tour des présidentielles de 2017 et 2022



(a) 2017



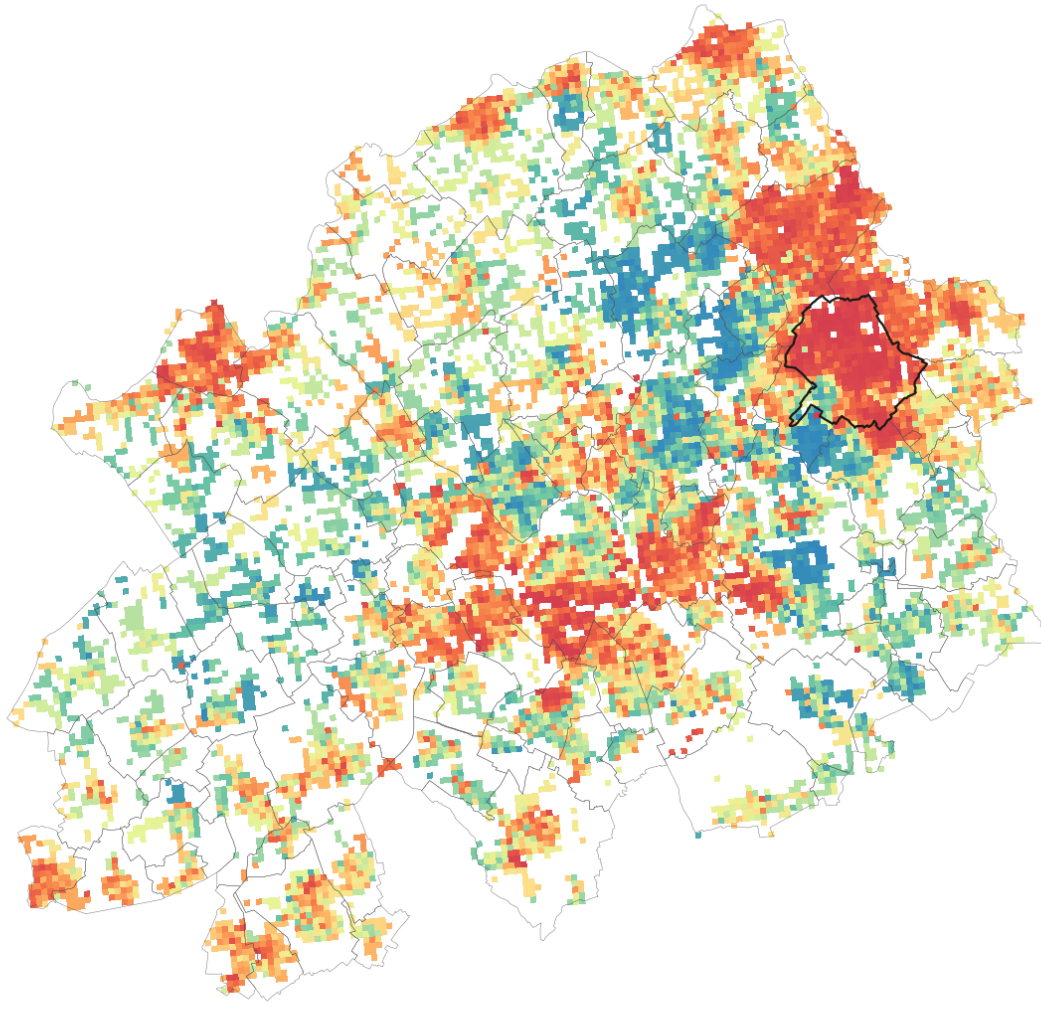
(b) 2022

Graphique : Hétérogénéité inter et intra des résultats au présidentielles

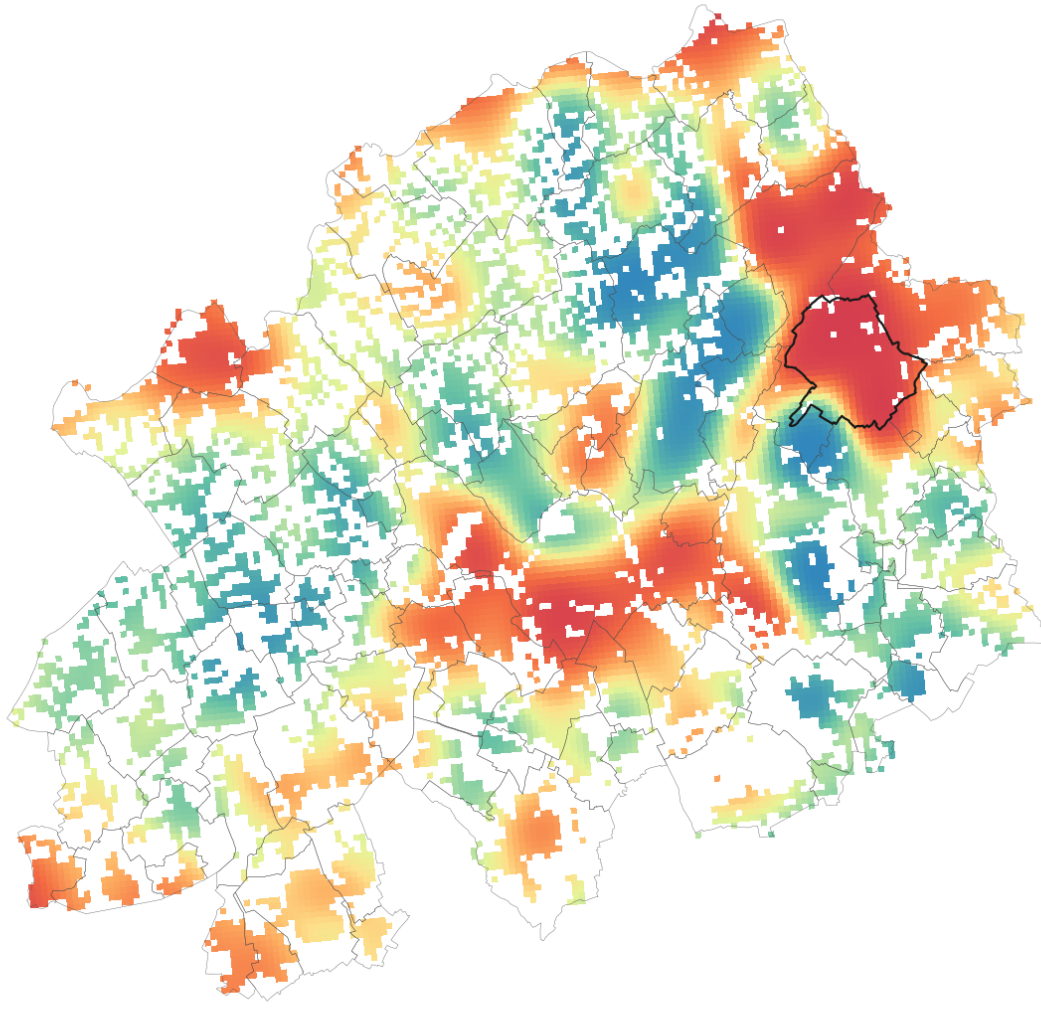
Hétérogénéité du vote et des habitants

- diversité du vote à l'échelle de la commune
- mais aussi diversité des populations
- p. ex.,
  - du point de vue des revenus
  - et du niveau de vie
- Roubaix présentant des différences particulièrement marquées

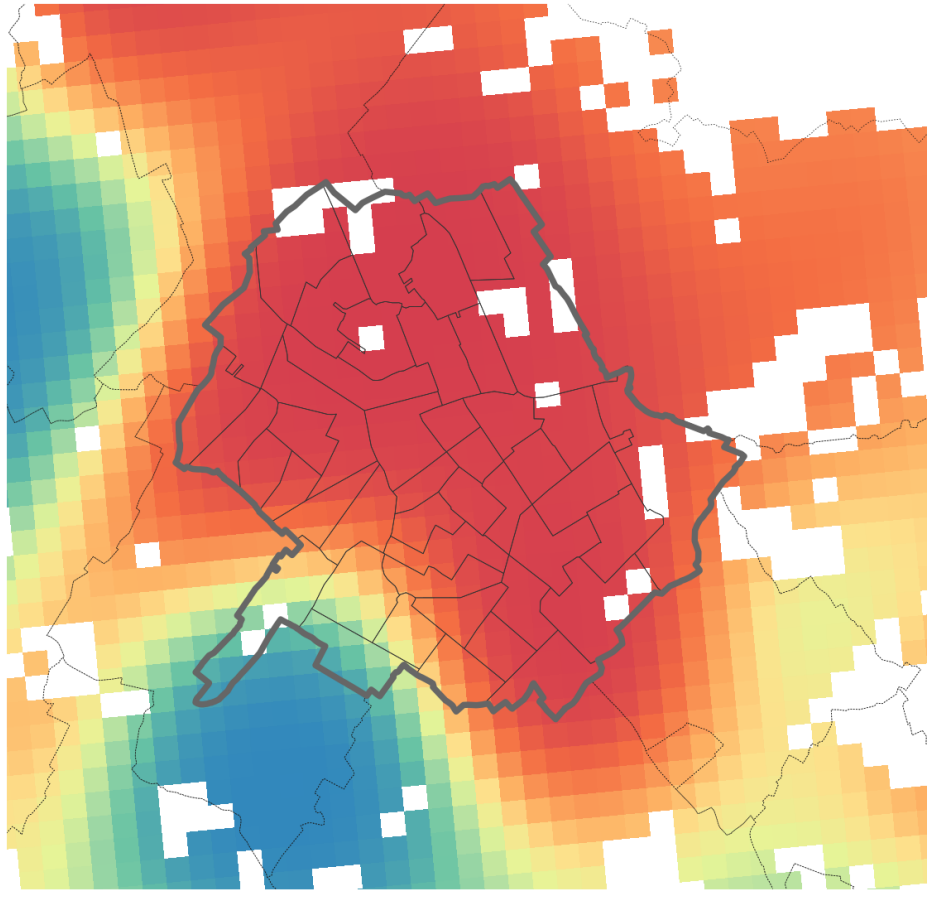
  à l'image de ce qui peut être observé à l'échelle de la métropole



Graphique : Niveau de vie moyen par carreau (MEL)
Source : Insee données carroyées



Graphique : Niveau de vie moyen lissé par carreau (MEL)
Source : Insee données carroyées



Graphique : Niveau de vie moyen lissé par carreau (Roubaix)
Source : Insee données carroyées

Couleurs primaires

Note :

- le dégradé de couleur tend **à forcer le trait**
- en opposant **les plus aisés** au reste des habitants
- parce qu'il est **difficile** de trouver un dégradé
- qui fasse correctement ressortir **les différentes nuances** de la distribution
- ce qui **homogénéise** la majorité du territoire de la commune
- qui présente toutefois des niveaux de vie **plus contrastés**
- que la carte ne le montre
- le lissage **atténue** aussi les ruptures

Hétérogénéité inter et intra

- on peut donc noter
  - l'homogénéité (variable) des résultats au BdV
  - l'hétérogénéité (elle aussi variable) entre les BdV
- mais aussi que
  - les scrutins antérieurs contiennent de l'information sur les scrutins postérieurs
  - qui suggère que l'utilisation du scrutin précédent
  - comme information auxiliaire
  - peut permettre d'aider au tirage des LdV
- toutefois, même si les scrutins contiennent de l'information sur les scrutins, ils ne les déterminent pas complètement
  - ce qui illustre un autre intérêt des tirages aléatoires
  - qui est de ne pas faire dépendre la sélection de la seule information auxiliaire
  - et de laisser la place à des facteurs inobservés

> « *the uncontrolled causes which may influence the result are always strictly innumerable* »
> — R. A. Fisher

Tirage équilibré

- différentes méthodes existent pour intégrer l'information auxiliaire

  en plus des tirages stratifié et par grappes

- comme le tirage équilibré
- un sondage est dit équilibré s'il satisfait aux équations d'équilibrage suivantes :

$$\hat{t}_{z\pi} = \sum_{k \in \mathcal{S}} \frac{\boldsymbol{z}_k}{\pi_k} = \sum_{k \in \mathcal{U}} \boldsymbol{z}_k \tag{28}$$

avec $\boldsymbol{z}_k = \{z_k, ..., z_{kP}\}$ est un vecteur de P variables auxiliaires mesurées pour l'unité k

- un tirage équilibré consiste donc
  - à sélectionner un échantillon aléatoire
  - dont les estimateurs d'Horvitz–Thompson d'un vecteurs de totaux
  - sont —approximativement— identiques dans l'échantillon et dans la population
- permet notamment de réduire la variance des estimateurs HT

  car $\mathbb{V}(\hat{t}_{z\pi}) = 0$

La méthode du cube

- le tirage équilibré a été mis en œuvre au moyen de la méthode du cube proposée par J.–C. Deville et Y. Tillé 2004
- le cube est une méthode itérative de scission
- qui consiste à scinder progressivement l'échantillon en deux
- en partant du vecteurs des probabilités d'inclusion
- à chaque étape le vecteurs des π_k est modifié aléatoirement
- de façon à ce qu'on moins un composant prenne la valeur 0 ou 1
- tout en respectant les équation d'équilibrage (28)
- l'algorithme produit donc un échantillon en plus ou moins n itérations

La méthode du cube

- l'algorithme doit son nom à la représentation géométrique des plans de sondage
- en effet, les 2^N échantillons possibles

  en incluant l'ensemble vide $\emptyset$ et le recensement

- correspondent au sommet d'un un N–cube

  hypercube $C = [0, 1]^N$

- la méthode du cube peut être définie comme
- une marche aléatoire vers un des sommets du cube satisfaisant aux équations d'équilibrage
- en arrondissant aléatoirement les π_k vers 0 ou 1

La méthode du cube

- pour arrondir dans la bonne direction
- on sélectionne un vecteur dans le noyau Q de la matrice $\check{\boldsymbol{Z}}$
- ce vecteur forme un plan dont l'intersection le cube
- qui permet « d'orienter » la marche aléatoire
- en effet, un échantillon équilibré consiste à choisir un sommet du N–cube se trouvant dans le sous–espace Q
- (28) peut être réécrite comme

$$\sum_{k\in\mathcal{U}} \frac{\boldsymbol{x}_k \mathbb{1}_{sk}}{\pi_k} = \sum_{k\in\mathcal{U}} \frac{\boldsymbol{x}_k \pi_k}{\pi_k} \tag{29}$$

$$\check{\boldsymbol{Z}}^\top \mathbb{1}_s = \check{\boldsymbol{Z}}^\top \boldsymbol{\pi} \tag{30}$$

avec $\check{\boldsymbol{Z}} = \{\frac{\boldsymbol{z}_1}{\pi_1} ... \frac{\boldsymbol{z}_k}{\pi_k} ... \frac{\boldsymbol{z}_N}{\pi_N}\}$

- ce système d'équations définit l'application affine

$$Q = \{\mathbb{1}_s \in \mathbb{R}^N | \check{\boldsymbol{Z}}^\top \mathbb{1}_s = \check{\boldsymbol{Z}}^\top \boldsymbol{\pi}\} = \boldsymbol{\pi} + \text{Ker } \check{\boldsymbol{Z}}^\top \tag{31}$$

avec Ker $\check{\boldsymbol{Z}}^\top = \{\boldsymbol{u} \in \mathbb{R}^N | \check{\boldsymbol{Z}}^\top \boldsymbol{u} = \boldsymbol{0}\}$

La méthode du cube

- l'intérêt de la méthode du cube est de permettre d'introduire plusieurs caractéristiques auxiliaires
- et des tests réalisés dans le cadre de la préparation d'une ESU lors des municipales de 2020

  qui n'a pas eu lieu pour des raisons assez évidentes

- montraient qu'un nombre limité de caractéristiques
- permettaient de réduire l'REQM des estimations

  racine carrée de l'erreur quadratique moyenne, cf. (32) p. 91

- toutefois, pour des raisons développées plus loin
- l'équilibrage a dû être limité aux voix de J. L. Mélenchon

Le 2nd tour

- comme indiqué précédemment (cf. p. 67),
- le choix a été fait de ne pas interroger les mêmes LdV au 1ier et 2nd tour
- de ce fait, 15 LdV étaient exclus de fait du tirage au 2nd tour
- cas particulier de coordination d'échantillon

Coordination des tirages

- les méthodes de coordination d'échantillons ont été développées par différents organismes de statistique publique nationaux à partir du début des années 1970
- au fil des années, ces organismes ont en effet été confrontés à la complexité grandissante de la gestion de leurs bases de sondage et à la multiplicité grandissante de leur usage
  - mises à jour fréquente pour préserver la qualité des échantillons.
  - utilisation de mêmes bases de sondages pour des enquêtes utilisant des plans différents

    y compris panels
- visaient aussi à diminuer la charge sur les ménages et les entreprises en minimisant la probabilité d'interrogation multiple

Tirage équilibré avec coordination négative

- technique proposée par Y. Tillé et A.–C. Favre 2004
- dans le cadre de l'enquête annuelle de recensement

  qui est un cas complexe de coordination négative
- et qui consiste à modifier les probabilités de sélection
- voir l'article pour plus de détails

Échantillonnage par rejet

- la méthode du cube a rencontré des difficultés pour trouver une solution
- du fait de
  - l'hétérogénéité inter et intra des LdV
  - taille de l'univers limité
  - et cela même si le nombre d'échantillons possibles demeurait conséquent
    - la sélection des LdV de 15 et 8 LdV revenait à tirer à un échantillon parmi

      $$\binom{30}{15}\binom{15}{8} = 998\ 181\ 241\ 200 \simeq 10^{12}$$

      avec $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
  - le nombre d'échantillon satisfaisant aux équations d'équilibrage étant clairement beaucoup réduit

    même s'il est impossible à estimer *a priori*
- ce qui explique aussi pourquoi l'équilibrage à été limité aux voix de J. L. Mélenchon

Échantillonnage par rejet

- pour y pallier à ces difficultés,
- ajout d'une (dernière) étape : l'échantillonnage par rejet Fuller (2009)
- qui consiste à
  - générer un grand nombre d'échantillons

    ici avec la méthode du cube avec coordination négative
  - puis sélectionner celui qui satisfait le mieux aux contraintes

Échantillonnage par rejet

- choix du tirage minimisant la racine carrée de l'erreur quadratique moyenne

$$\text{REQM}[\hat{\theta}(\mathcal{S})] = \sqrt{\mathbb{E}\left[\{\hat{\theta}(s) - \theta\}^2\right]} \tag{32}$$

$$= \sqrt{\mathbb{B}[\hat{\theta}(\mathcal{S})]^2 + \mathbb{V}[\hat{\theta}(\mathcal{S})]} \tag{33}$$

- qui permet de trouver un compromis entre
  - le biais
  - et la variance
  - des estimateurs
- d'autres métriques sont bien sûr envisageables
- la sélection a été réalisée en se calant sur les résultats de trois candidats en 2017 :
  - J. L. Mélenchon
  - E. Macron
  - M. Le Pen

Échantillonnage par rejet

- par rapport au tirage équilibré par la méthode du cube,
  - le tirage par rejet permet de contrôler l'erreur
  - mais il présente un gros désavantage
  - parce que, à la différence du cube
  - il peut modifier les probabilités d'inclusion
  - ce qui peut conduire à biaiser les estimateurs

    en fonction de la proportion d'échantillons rejetés
- des tests sur les scrutins intermédiaires

  à Lille et Roubaix
- suggèrent toutefois que le biais est faible ici

Note

- les tests sur les scrutins intermédiaires à Roubaix

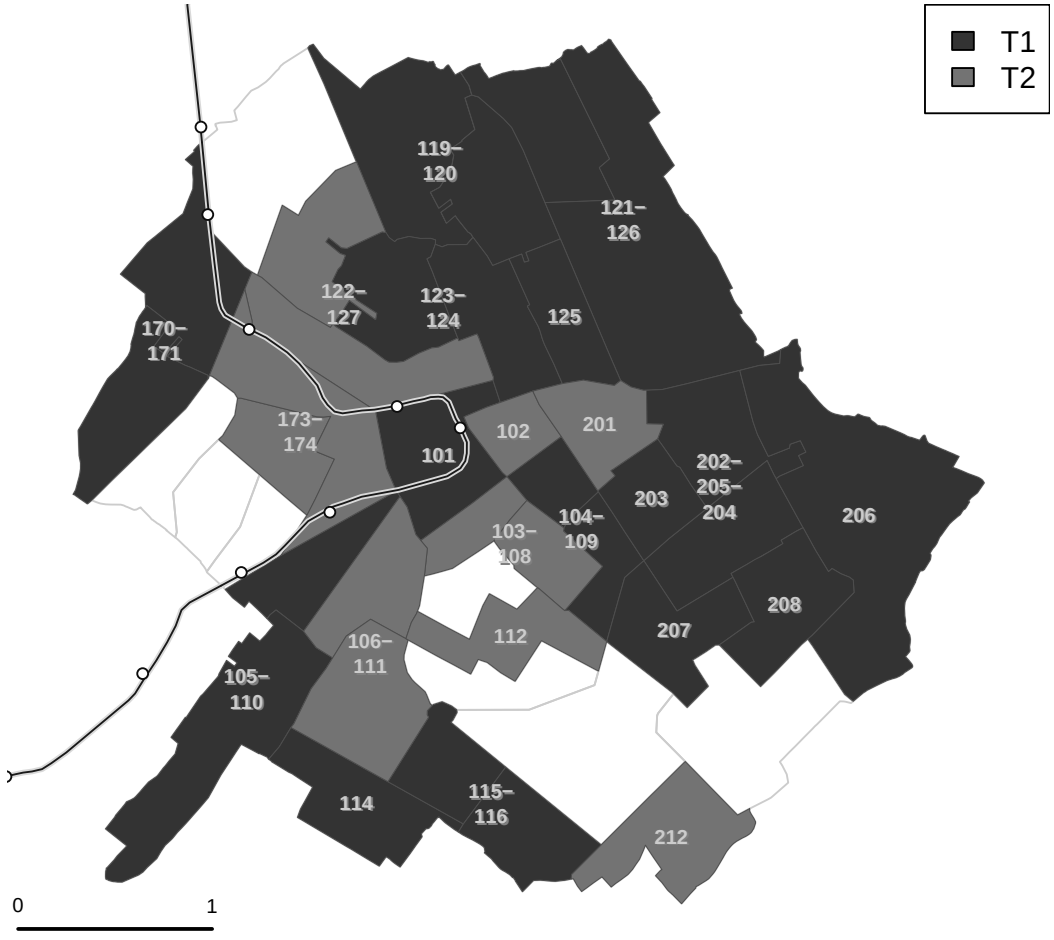
  mais aussi à Lille

- suggèrent aussi que
- même s'il paraît préférable d'équilibrer sur le scrutin de même type précédent

  présidentielles,...

- le tirage peut être réalisé à partir d'autres scrutins
- utile si le découpage des BdV à changé entre temps
- le tirage semble aussi « résilient » aux pandémies mondiales...

Résultats



Graphique : Bureaux de vote sélectionnés au premier et second tours
Source : mairie de Roubaix —fond de carte—

Fazit

- au final, l'enquête a permis de recueillir :
  - 1^{ier} tour : 15 LdV —23 BdV— et 2 372 questionnaires dans le plan

    2 795 en tout
  - 2^{nd} tour : 8 LdV —12 BdV— et 1 236 questionnaires dans le plan

    1 471 en tout
- le taux de non–réponse à l'enquête est de 3,4% pour le premier tour

1^{ier} tour

Candidat	Roubaix	Roubaix (LdV)	ESU	Δ_{LdV}	Δ_{ESU}
Jean-Luc MÉLENCHON	0.515	0.512	0.521	0.003	0.006
Emmanuel MACRON	0.194	0.198	0.200	-0.004	0.006
Marine LE PEN	0.142	0.142	0.090	0.000	-0.052
Éric ZEMMOUR	0.032	0.033	0.030	-0.001	-0.002
Yannick JADOT	0.025	0.027	0.037	-0.001	0.011
Valérie PÉCRESSE	0.019	0.020	0.021	-0.002	0.002
Fabien ROUSSEL	0.013	0.011	0.014	0.003	0.001
Nicolas DUPONT-AIGNAN	0.010	0.010	0.010	0.000	-0.000
Jean LASSALLE	0.010	0.009	0.015	0.001	0.005
Anne HIDALGO	0.010	0.010	0.018	-0.001	0.008
Nathalie ARTHAUD	0.005	0.005	0.004	0.001	-0.001
Philippe POUTOU	0.005	0.005	0.008	0.000	0.003
Blancs et nuls	0.018	0.019	0.031	-0.001	0.013

2nd tour

Candidat	Roubaix	Roubaix (LdV)	ESU
Emmanuel MACRON	.65	.671	.713
Marine LE PEN	.274	.256	.173
Blancs et nuls	.075	.073	.112

Conclusion

Conclusion

- ce qui précède visait à illustrer
- certains aspects pratiques de la théorie des sondages
- permettant de mobiliser l'information auxiliaire
- pour l'élaboration de plans de sondages
- ainsi que l'effet de différents plans

  tirages stratifié et par grappes

- sur les estimateurs

Conclusion

- l'approche par le plan permet de prendre explicitement en compte que
- les données ne sont seulement générées par des processus extérieurs à la collecte
- mais aussi par la sélection des observations
- avec la limite que le plan prend le parti inverse
- car, dans ce cadre, les variables sont fixes
- et le plan est la seule source de l'aléas
- de plus, l'approche par le plan a plutôt été conçue pour estimer des statistiques descriptives

  totaux, moyennes, médianes,...

- que, p. ex., des modèles de régression

Conclusion

- ce cadre un peu étroit peut toutefois être étendu
- pour l'estimation
- en combinant le plan avec une approche par le modèle
- p. ex. avec les modèles de superpopulation
- qui permet d'ajouter d'autres sources d'aléas
- et donc d'estimer des modèles

Conclusion

- l'approche par le plan présente d'autres limites
- car c'est un cadre général
- mais largement développé pour les enquêtes de la statistique publique
- p. ex., la méthode du cube a d'abord servi à réaliser des tirages
- dans des fichiers de millions d'adresses
- donc il faut parfois faire preuve d'un peu d'imagination…

Bibliographie I

Cochran, William G. (1977), *Sampling Techniques, 3rd Edition.* New York, John Wiley & Sons.

Deville, Jean–Claude et Yves Tillé (2004), « Efficient balanced sampling : The cube method », *Biometrika*, nº 4, vol. 91, p. 893-912.

Fuller, Wayne (2009), « Some design properties of a rejective sampling procedure », *Biometrika*, nº 4, vol. 96, p. 933-944.

Lohr, Sharon L. (2019), *Sampling Design and Analysis,* New York, Chapman et Hall/CRC.

Pina, Christine (2019), « Que sont les SSU devenus ? Les sondages 'sortie des urnes' en France et aux États-Unis », *Genèses*, vol. 1, p. 117-133.

Tillé, Yves (2001), *Théorie des sondages. Échantillonnage et estimation en populations finies.Cours et exercices avec solution,* Paris, Dunod.

Tillé, Yves et Anne–Catherine Favre (2004), « Coordination, combination and extension of balanced samples », *Biometrika*, nº 4, vol. 91, p. 913-927.

Merci pour votre attention

Des questions ?