Plan de sondage informatif : comment le définir et en tenir compte ?

Séminaire en ligne sur les sondages du groupe Enquêtes de la Société Française de Statistique organisé en collaboration avec l'Université de Neuchâtel

20 novembre 2025

D.Bonnéry

1

La sélection est informative lorsqu'il convient d'en tenir compte

Pfeffermann, Danny, Krieger, AM Abba M AM, & Rinott, Yosef. 1998.
Parametric distributions of complex survey data under informative probability sampling.
Statistica Sinica, **8**(4), 1087–1114.

Les données d'enquête sont le produit de deux processus aléatoires.

- Le processus dont est issu la population d'intérêt et ses caractéristiques
- Le processus de sélection

La sélection est informative quand il convient d'en tenir compte lors de l'inférence sur le modèle de population à partir des données d'enquête.

Lorsque les observations sont le produit

- d'un processus d'intérêt
- d'un processus de nuisance.

le processus de nuisance est informatif lorsqu'il convient d'en tenir compte lors de l'inférence.
Processus de nuisance:

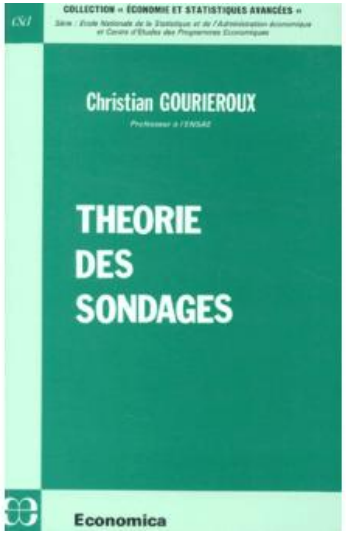
- erreur de mesure
- caviardage
- censure
- non réponse
- nombre de tirages par unité
- taille de la population
- taille de l'échantillon
- étiquetage

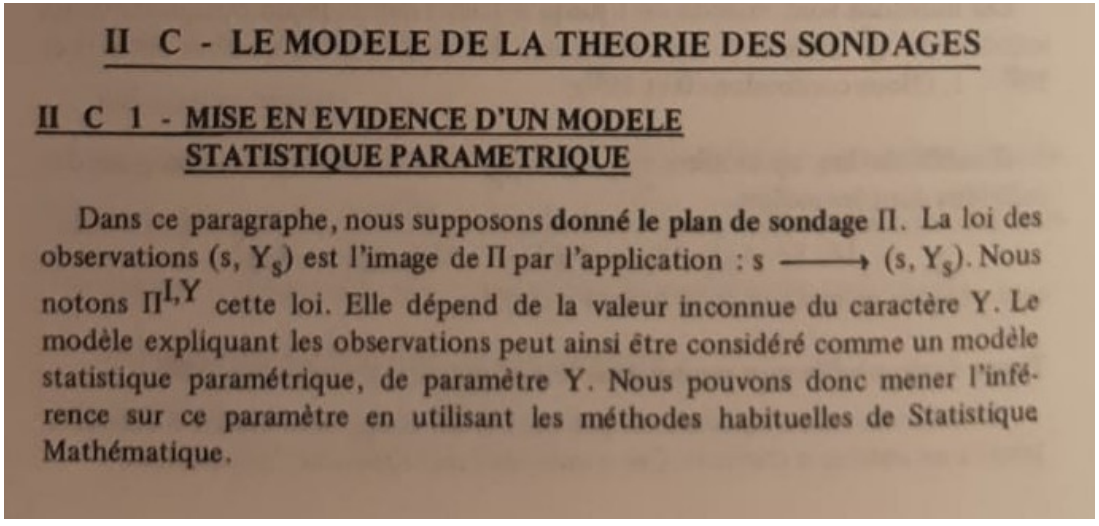
- Modèle de la variable d'intérêt: $V \sim P$;
- Tout le reste est nuisance: $\bar{V} \mid V \sim P'$

$$\text{Observation} = X = \text{function}(V, \bar{V}).$$

2

Modèle générique pour les sondages et sélection informative dans la littérature scientifique





II C - LE MODELE DE LA THEORIE DES SONDAGES

II C 1 - MISE EN EVIDENCE D'UN MODELE STATISTIQUE PARAMETRIQUE

Dans ce paragraphe, nous supposons **donné le plan de sondage** Π. La loi des observations (s, Y_s) est l'image de Π par l'application : $s \longrightarrow (s, Y_s)$. Nous notons $\Pi^{I,Y}$ cette loi. Elle dépend de la valeur inconnue du caractère Y. Le modèle expliquant les observations peut ainsi être considéré comme un modèle statistique paramétrique, de paramètre Y. Nous pouvons donc mener l'inférence sur ce paramètre en utilisant les méthodes habituelles de Statistique Mathématique.

Gourieroux, C. 1981.
Théorie des sondages.

La distribution des observations sous le modèle fixe de population pour l'inférence basée sur le plan appartient à un **modèle parametrique**, les caractéristiques de la population étant le paramètre.

Pour une population de taille N, un plan de sondage P^S, le modèle

- de population est: $P^Y \in \{\text{dirac}_\theta \mid \theta \in \Theta = \mathbb{R}^N\}$
- des observations est: $P^{S,Y[S]} \in \{\left(P^S\right)^{\mathrm{x}\to\theta[\mathrm{x}]} \mid \theta \in \Theta = \mathbb{R}^N\}$


Godambe, V. P. 1966.
A New Approach to Sampling from Finite Populations. I Sufficiency and Linear Estimation.
Journal of the Royal Statistical Society, **28**(2), 310–319.


Ericson, W . A . 1969.
Subjective Bayesian Models in Sampling Finite Populations.
Journal of the Royal Statistical Society, **31**(2), 195–233.

Les auteurs expliquent que dans le cas du modèle de population fixe, la vraisemblance est "uninformative".

Scott, Alastair J. 1975.
Some comments on the problem of randomisation in surveys.
In: Int. Assoc. of Survey Statisticians, 2nd Meeting, Warsaw, September 1975.

On the problem of randomization in survey sampling.
In: Sankhya, 39, 1977.

“ Scott (1977) ”

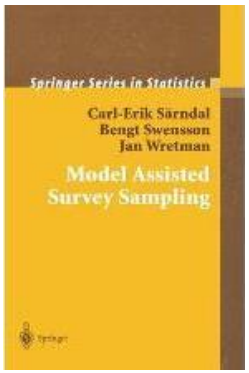
[If the] posterior distribution depends only on the sample actually drawn and not on the sampling desing used to draw it, provided that [the sample of the population elements] and [the study variable on the population] are independent random variables. (If this condition of independence is satisfied the design is said to be non-informative.)

1. Premier article (parmi ceux que j'ai lu) qui proposent une définition d'un plan de sondage non informatif.
2. Autre aspect important: la définition n'est pas figée à un cadre particulier, mais s'applique à des cas différents, ou les observations peuvent contenir plus ou moins d'information (variables auxiliaires sur l'échantillon, la population...).

Rubin, DB Donald B. 1976.
Inference and missing Data.
Biometrika, **63**(3), 591–592.

1. Ne définit pas ce qu'est un processus informatif en général, mais, dans un cas particulier
2. Définit ce qu'ignorer un processus de nuisance particulier (la non réponse) signifie
3. Énonce des conditions pour que la vraisemblance lorsque le processus de nuisance est ignoré est la même que la vriasemblance lorsqu'il ne l'est pas.
4. La vraisemblance est la vraisemblance conditionelle au processus de nuisance.

Le terme ignorable n'apparait pas, mais le terme "at random" oui.



Cassel, CM, Sarndal, CE, & Wretman, JH. 1977.
Foundations of inference in survey sampling.

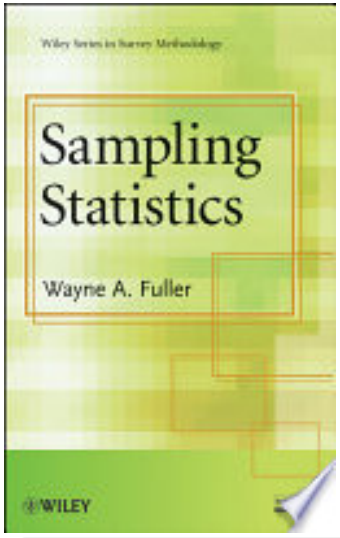
Définition tautologique limitée au modèle de population fixe: le plan est non informatif lorsque les probabilités d'inclusion et les variables d'intérêt sont indépendantes, ce qui est toujours le cas dans un modèle de population fixe.

Little, Roderick J. A. RJA. 1982.
Models for Nonresponse in Sample Surveys.
Journal of the American Statistical Association, **77**(378), 237–250.

Dans la lignée de Rubin, 1976, donne la définition d'un plan de sondage ignorable.

Sugden, RA, & Smith, T. M. F. 1984.
Ignorable and informative designs in survey sampling inference.
Biometrika, **71**(3), 581–592.

- Premier article (à ma connaissance) à opposer to plan ignorables et informatifs dans un article.



Fuller, Wayne A. 2011.
Sampling statistics.
Vol. 560.
Hoboken, NJ: Wiley.

> “If $E[x_i \pi_i e_i] \neq 0$, it is sometimes said that the design is “informative” for the model.“
>
> Fuller (2011)[p. 355 sec. 6.3.2]

Les objets x_i, π_i et e_i sont la variable d'intérêt, la probabilité d'inclusion, le résidu de la régression linéaire de x_i sur la variable de sondage.

Une définition contradictoire

Le sondage est informatif $\Leftrightarrow$ *dependance entre* π_k *et* Y_k *après conditionement sur les covariables du modèle* X_k.

Ce n'est

- ni suffisant,
- ni nécessaire.

Plan informatif: $\pi \perp Y$

Sondage par grappe:

- $(Y_k)_{k \in U} \sim$ iid $\mathscr{B}(1/2)$,
- $J = Y + \varepsilon$,
- Grappe U_0, U_1 obtenues après tri selon J,
- Probabilité de sélection d'un cluster : $\frac{1}{2}$,
- $\pi_k = \frac{1}{2}$,
- $(Y_k)_{k \in S}$ ne sont pas iid $\mathscr{B}(1/2)$.

Plan non-informatif: π et Y sont dépendants

- $(Y_k)_{k\in U} \sim$ iid $\mathscr{B}(1/2)$,
- $\pi_k = \begin{cases} 1/2 & \text{if } Y_k = 0 \\ \left| \begin{array}{l} 1/4 \text{ w.p. } 1/2 \\ 3/4 \text{ w.p. } 1/2 \end{array} \right. & \text{if } Y_k = 1, \end{cases}$
- Plan: $\mathrm{Poisson}(\pi)$,
- $(Y_k)_{k\in S}$ sont iid $\mathscr{B}(1/2)$.

Pfeffermann, Danny, Krieger, AM Abba M AM, & Rinott, Yosef. 1998.
Parametric distributions of complex survey data under informative probability sampling.
Statistica Sinica, **8**(4), 1087–1114.

D. Pfeffermann introduit les notions de distribution de population et d'échantillon, ces distributions pouvant différer en présence de sélection informative.

- La distribution d'échantillon est une loi dominée par la distribution de population
- La dérivée ρ de Radon Nikodyn est la probabilité d'être sélectionnée ou l'espérance des probabilités de sondages conditionellement à la caractéristique d'intérêt.
- de la connaissance ou modélisation de ρ peut être dérivée une vraisemblance approchée.

$$\rho(y) = [E[I_k \mid Y_k = y]],$$

Approximativement : $(Y_k)_{k \in S} \overset{\text{i.i.d.}}{\sim} \rho_Y$

- Intérêt pratique pour des modèles de superpopulation avec résidus par unités indépendants.
- Estimateurs plus efficaces que les estimateurs de pseudo vraisemblance.


Daniel Bonnéry, Breidt, F. J., and Coquet, F. (2012). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli Journal, 2012*.


Daniel Bonnéry, F. Jay Breidt, and François Coquet. *Kernel Estimation for a Superpopulation Probability Density Function under Informative Selection* *Metron*, 75(3), 2017.

http://rdcu.be/wtPL.


Daniel Bonnéry, F. Jay Breidt, and François Coquet. Asymptotics for the maximum sample likelihood estimator under informative selection from a finite population. *Bernoulli*, 11(2):655–679.

2.1. Notation conventions and definitions

- Généralisation des définitions de processus ignorables.

Daniel Bonnery and Joseph Sedransk, 2020.
On the definition of informative vs. ignorable nuisance process.
arXiv, **math.ST**1906.02733.

> “ it is inevitable, [...] that a paper dealing with [the] problem [of what conclusions regarding sufficient statistics may be drawn from the existence of uniformly most powerful tests, or vice versa] should bear some mark of the theory of functions, in spite of its concern with statistical questions ”
>
> Neyman & Pearson (1936)

2.1.1. Théorie des ensembles

- Soit $h : \mathscr{E} \to$, $\sim_h$ est la relation d'équivalence sur $\mathscr{E}$: $x \sim_h x' \leftrightarrow h(x) = h(x')$. Les classes d'équivalence sur $\mathscr{E}$ pour $\sim_h$ sont notées $\mathscr{E}/h$, la classe de x pour $\sim_h$ est $\mathrm{class}_h(x) = h^{-1}(\{h(x)\})$.
- $h : \mathscr{E} \to$ dépend déterministiquement $h' : \mathscr{E} \to$ seulement (on note $h \prec h'$) si et ssi $\exists g :\to$ tel que $h = g \circ h'$.
- le complément de $h : \mathscr{E} \to$ est toute fonction $\bar{h} : \mathscr{E} \to$ telle que $(\mathrm{class}_h, \mathrm{class}_{\bar{h}}) : \mathscr{E} \to (\mathscr{E}/h) \times (\mathscr{E}/\bar{h}), x \mapsto (\mathrm{class}_h(x), \mathrm{class}_{\bar{h}}(x))$ est injective. Si une fonction $\bar{h}$ satisfait cette propriété, on écrit $(h, \bar{h}) \trianglerighteq \mathscr{E}$. Pour un complément $\bar{h}$ de h, on définit la fonction $\sqcap_{h,\bar{h}} : \{h(x), \bar{h}(x) : x \in \mathscr{E}\} \to \mathscr{E}$ comme l'inverse de $(\mathrm{class}_h, \mathrm{class}_{\bar{h}})\Big|^{\mathrm{image}((\mathrm{class}_h, \mathrm{class}_{\bar{h}}))}$.

Si $\mathrm{image}(h, h') = \mathrm{image}(h) \times \mathrm{image}(h')$ on dit que h et h' sont variation-independants ou séparés et on note $h \perp h'$ et $(h, h') \sqcap \mathscr{E}$.

2.1.2. Modèle statistique général

- $\{(\Omega, \mathfrak{S}_\Omega, P)\}_{P \in \mathscr{P}}$, un modèle statistique
- $X : (\Omega, \mathfrak{S}_\Omega) \to (\mathscr{X}, \mathfrak{S}_{\mathscr{X}})$, la statistique observée
- $V : (\Omega, \mathfrak{S}_\Omega) \to (\mathscr{V}, \mathfrak{S}_{\mathscr{V}})$, latente, ou non. Non latente: $V \prec X$. Le **processus d'intérêt**
- La cible de l'inférence est d'estimer $\boldsymbol{\theta}(P^V)$ ou de prédire $\boldsymbol{\theta}(V)$. en résumé: $\boldsymbol{\theta} : (\{P^V : P \in \mathscr{P}\} \times V(\Omega)) \to$.
- Soit $\bar{V}$ un complément (séparé ou pas) de V, latent ou non. La **nuisance**
- Soit $\mathscr{P}$ l'ensemble des distributions possibles de $\bar{V}$ une fois le processus ignoré.

Puisque $(V, \bar{V}) \trianglerighteq \Omega$, $\exists \mathbf{x} : (V(\Omega) \times \bar{V}(\Omega))$ est tel que $X = \mathbf{x}(V, \bar{V})$, $\mathbf{x}$ est donné par la relation: $\mathbf{x} = X[\sqcap_{V,\bar{V}}]$.

2.2. Ignorer le processus de nuisance

2.2.1. Transformation du modèle

Pour $v \in V(\Omega)$, on définit l'ensemble mesurable

$$\Phi_{v,V,\bar{V}} = \bar{V}^{-1}\left(\bar{V}\left(V^{-1}\left(\{v\}\right)\right)\right).$$

Si $(V, \bar{V}) \sqsupseteq \Omega$, alors $\forall \bar{v} \in \bar{V}(\Omega)$, $\Phi_{\bar{v},\bar{V},V} = \Omega$, et
$: \omega \mapsto \left(\sqcap_{\bar{V},V}\right)(\bar{v}, V(\omega))$ est mesurable : $(\Omega, \mathfrak{S}_\Omega) \to (\Omega, \mathfrak{S}_\Omega)$.
Si $\forall P \in \mathscr{P}$, $\forall \bar{v} \in \bar{V}(\Omega)$, $P\left(\Phi_{\bar{v},\bar{V},V}\right) > 0$, on définit:

$$\mathscr{P}^\star = \left\{ \left(\left\langle \bar{v} \mapsto P^{(V,\bar{v})|\Phi_{\bar{v},\bar{V},V}} \middle| P' \right\rangle_{(V,\bar{V})(\Omega),\bar{V}(\Omega)} \right)^{\sqcap_{V,\bar{V}}} \middle| (P, P') \in \mathscr{P} \times \mathscr{P}' \right\}.$$

Le modèle obtenu après avoir ignoré $\bar{V}$ est le modèle
$\left(\mathrm{codomain}(X), \mathfrak{S}_{\mathrm{codomain}(X)}, \{(P^\star)^X \mid P^\star \in \mathscr{P}^\star\}\right)$.

2.2.2. Transformation de la cible de l'inférence

Etant donné une cible $\boldsymbol{\theta}$, quelle est la cible dans le modèle correspondant $\boldsymbol{\theta}^\star$ dans le modèle $\mathscr{P}^\star$? Dans certains cas, il y a une réponse naturelle

1. S'il existe $Y : (\Omega, \mathfrak{S}_\Omega) \to$ tel que $\boldsymbol{\theta} : (\Omega \times \mathscr{P}) \to, (\omega, P) \mapsto Y(\omega)$, un choix naturel est $\boldsymbol{\theta}^\star : \mathscr{P}^\star \times \Omega \to, (P^\star, \omega) \mapsto (P^\star, \omega) \mapsto Y(\omega)$.
2. S'il existe $Y : (\Omega, \mathfrak{S}_\Omega) \to$ tel que $\boldsymbol{\theta} \prec (P, \omega \mapsto P^Y)$: $\boldsymbol{\theta} = \boldsymbol{\theta'} \circ (P \mapsto P^Y)$ et si $\{(P^\star)^Y \mid P^\star \in \mathscr{P}^\star\} \subset \mathrm{codomain}(\boldsymbol{\theta'})$, alors on définit: $\boldsymbol{\theta}^\star : P^\star \mapsto \boldsymbol{\theta'}((P^\star)^Y)$.
3. Quand $\mathscr{P}^\star \subset \mathscr{P}$, un choix naturel est $\boldsymbol{\theta}^\star = \boldsymbol{\theta}\Big|_{\mathscr{P}^\star \times \Omega}$.

2.3. Transformation de l'ensemble des distributions a priori

Après être passé de $\mathscr{P}$ à $\mathscr{P}^\star$ après avoir ignoré un processus, il est difficile de définir naturellement un nouveau jeu de probabilités a priori $\mathscr{Q}^\star$ parce que Q et $Q^\star$ ne sont pas des distributions définies sur le même ensemble. Quand chaque distribution est le produit d'un a priori sur le processus d'intérêt et d'une distribution sur le processus de nuisance, un modèle naturel peut être proposé:

1. On définit l'ensemble $\tilde{\mathscr{Q}}$ de distributions à priori (pas forcément propre) sur $\mathscr{P}'$.
2. Define $\mathscr{Q}^\star =$

$$\left\{ \left(Q^{P \mapsto \left(\bar{v} \mapsto P^{(V,\bar{v})|\Phi_{\bar{v},\bar{V},V}} \right)} \otimes Q' \right)^{(P'',P') \mapsto \left(\left(\langle P''|P' \rangle_{(V,\bar{V})(\Omega),\bar{V}(\Omega)} \right)^{\sqcap_{V,\bar{V}}} \right)} : \right.$$
$$\left. Q \in \mathscr{Q}, Q' \in \mathscr{Q}' \right\}.$$

2.4. Inférences équivalentes

- Inférence basée sur la vraisemblance: L'inférence basée sur $\boldsymbol{\theta} : \mathscr{P} \to$ est équivalente à l'inférence basée sur $\boldsymbol{\theta}^\star : \mathscr{P}^\star \to$ lorsque $X = x$ if $\mathrm{codomain}(\boldsymbol{\theta}) = \mathrm{codomain}(\boldsymbol{\theta}^\star)$ et $\Big(\theta \mapsto \mathscr{L}^\star_{\boldsymbol{\theta}^\star;X}(\theta; x)\Big) = \Big(\theta \mapsto \mathscr{L}_{\boldsymbol{\theta},X}(\theta; x)\Big)$.
- Estimation fréquentiste:
  L'inférence basée sur l'estimation de $\boldsymbol{\theta} : \mathscr{P} \to$ par $\hat{\boldsymbol{\theta}} : \Omega \to \mathrm{codomain}(\boldsymbol{\theta})$ est équivalente à l'estimation de $\boldsymbol{\theta}^\star : \mathscr{P} \to$ par $\hat{\boldsymbol{\theta}}$ si $\mathrm{codomain}\boldsymbol{\theta} = \mathrm{codomain}\boldsymbol{\theta}^\star$ et $\forall \theta \in \mathrm{codomain}(\boldsymbol{\theta})$, $\{P^{\hat{\boldsymbol{\theta}}} \mid P \in \boldsymbol{\theta}^{-1}(\theta)\} = \{(P^\star)^{\hat{\boldsymbol{\theta}}} \mid (P^\star) \in (\boldsymbol{\theta}^\star)^{-1}(\theta)\}$.

- Inférence Bayesienne: L'inférence Bayesienne sur $\boldsymbol{\theta} : (\mathscr{P} \times \Omega) \rightarrow$ avec l'ensemble de distributions à priori $\mathscr{Q}$ est equivalent à l'inférence basée sur la vraisemblance sur $\theta^\star : (\mathscr{P}^\star \times \Omega) \rightarrow$ avec le jeu de distributions a priori $\mathscr{Q}$ quand $X = x$ si $\{(Q^\star)^{\boldsymbol{\theta}^\star | X=x} \mid Q^\star \in \mathscr{Q}^\star\} = \{Q^{\boldsymbol{\theta}|X=x} \mid Q \in \mathscr{Q}\}$

2.5. Processus ignorable vs. processus informatif : définition et caractérisation

2.5.1. Definition

Le processus $\bar{V}$ sera dit ignorable pour un type particulier d'inférence si pour $\mathscr{P}' = \left\{\text{Dirac}_{\bar{v}} : \bar{v} \in \bar{V}(\Omega)\right\}$,l'inférence basée sur $(\mathscr{P}, \boldsymbol{\theta}, \mathscr{L}, \mathscr{Q}, ...)$ est équivalente à l'inférence basée sur $(\mathscr{P}^\star, \boldsymbol{\theta}^\star, \mathscr{L}^\star, \mathscr{Q}^\star, ...)$. Dans le cas ou les deux ne sont pas équivalents, le processus $\bar{V}$ est dit informatif.

De façon regrettable, "V est ignorable" dans ce sens ne signifie pas que "V is ignorable" dans le sens habitiuel de l'ignorabilité en statistiques. Une statistique ignorable est toute statistique indépendante d'au moins une variable exhaustive (Schervish 1995, p. 142, Exercise 33).

Exemple:
$Y \sim \mathrm{Normal}(0, \sigma^2)^{\otimes N}$, et T est la sélection associée avec $R = (\arg\min(Y_k \mid k \in \{1, \ldots, N\}), \arg\max(Y_k \mid k \in \{1, \ldots, N\}))$.
On observe T et $T[Y]$ Ici: $T[Y] = (\min(Y), \max(Y))$ et $T[Y]$ est independant de T. Etant donné $R = (1, 5)$, T est la fonction $t : y \mapsto (y_1, y_5)$, et $P^{t(Y)} = P^{(Y_1, Y_5)} \neq P^{T[Y]|T=t} = P^{\min(Y), \max(Y)}$.
Donc la sélection est non ignorable et informative, bien que $P^T_{\theta, \bar{\theta}}$ soit indépendant de θ, et inconditionnellement à Y, T est le sondage aléatoire simple de 2 parmi N. donc la variable T est libre, et ignorable puis que indépendante de la statistique exhaustive $T[Y]$.

2.6. Au hasard (at random)

La condition "At random" est la condition ou $\mathscr{P} = \{P^V \otimes P^{\bar{V}} : P \in \mathscr{P}\}$.

3

Conditionnellement au modèle, quelle est la meilleure inférence ?

Pour résumer:

- Rubin,1976, donne une définition rigoureuse et généralisable d'un processus de non réponse ignorable.
- On se retrouve à se poser la question de la meilleure inférence en présence d'un processus de nuisance.
- Quelles sont les meilleures prédictions pour un modèle donné ?

Premier écueil: ne pas spécifier le modèle ! Une inférence est optimale pour un modèle donné.

Basu, Debabrata. 1977.
On the Elimination of Nuisance Parameters.
Journal of the American Statistical Association, **72**(358), 355—-366.

Pour l'inférence classique, il n'y a pas de définition totalement satisfaisante d'une statistique exhaustive en présence d'un paramêtre de nuisance.

Marginaliser

$$\mathscr{L}_{S,Y[S]}(\mathbf{x},\mathbf{y};\theta,\xi) = f_{Y[\mathbf{x}];\theta}(\mathbf{y}) \times f_{S|Y[\mathbf{x}];\theta,\bar{\theta}}(\mathbf{x} \mid \mathbf{y})$$

ou
conditionner

$$\mathscr{L}_{Y[S]|S}(\mathbf{y} \mid \mathbf{x};\theta,\bar{\theta}) = f_{Y[\mathbf{x}];\theta}(\mathbf{y}) \times \frac{f_{S|Y[\mathbf{x}];\theta,\bar{\theta}}(\mathbf{x} \mid \mathbf{y})}{f_{S;\theta,\bar{\theta}}(\mathbf{x})}$$

?

- Conflit Fisher - Barnard

Pas de résultat d'optimalité.

4

Quel est le meilleur modèle ?

Dans un institut de sondage, en l'absence de non réponse, l'estimateur d'Horvitz Thompson est:

$$\sum_{k \in U} \frac{Y_k}{\pi_k} \mathbb{1}_S(k).$$

et l'estimateur de la variance est:

$$\sum_{k \in U} Y_k Y_\ell (\pi_{k,l})^{-1} (\pi_{k,\ell} - \pi_k \pi_\ell) \mathbb{1}_S(k) \mathbb{1}_S(\ell).$$

avec $\pi_{k,\ell} = E[\mathbb{1}_S(k)\mathbb{1}_S(\ell)]$
L'estimation de la précision de notre mesure est basée sur ce que d'autres échantillons auraient pu nous donner:

$$V[\hat{t}_y(S)] = \sum_{s \in S(\Omega)} (\hat{t}_y(s) - t_y)^2 P(S = s).$$

Un directeur d'institut a 3 sondeurs préférés, plus ou moins matinaux.

- Le premier interroge les gens de son village
- Le deuxième fait du sondage aléatoire simple
- Le troisième un plan stratifié avec allocation constante.

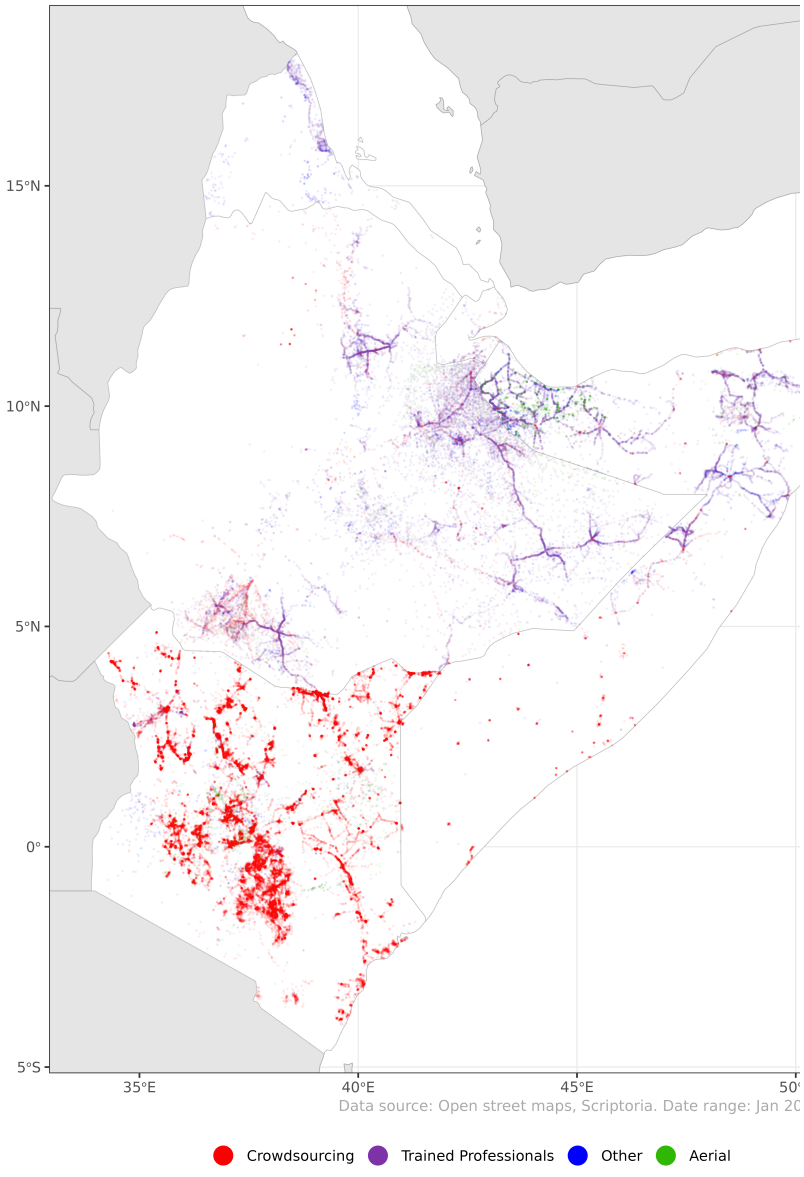
Il décide que le premier à lui apporter un café sera en charge du prochain plan de sondage.
Quelles sont les probabilités d'inclusion et les probabilités de double inclusion ?

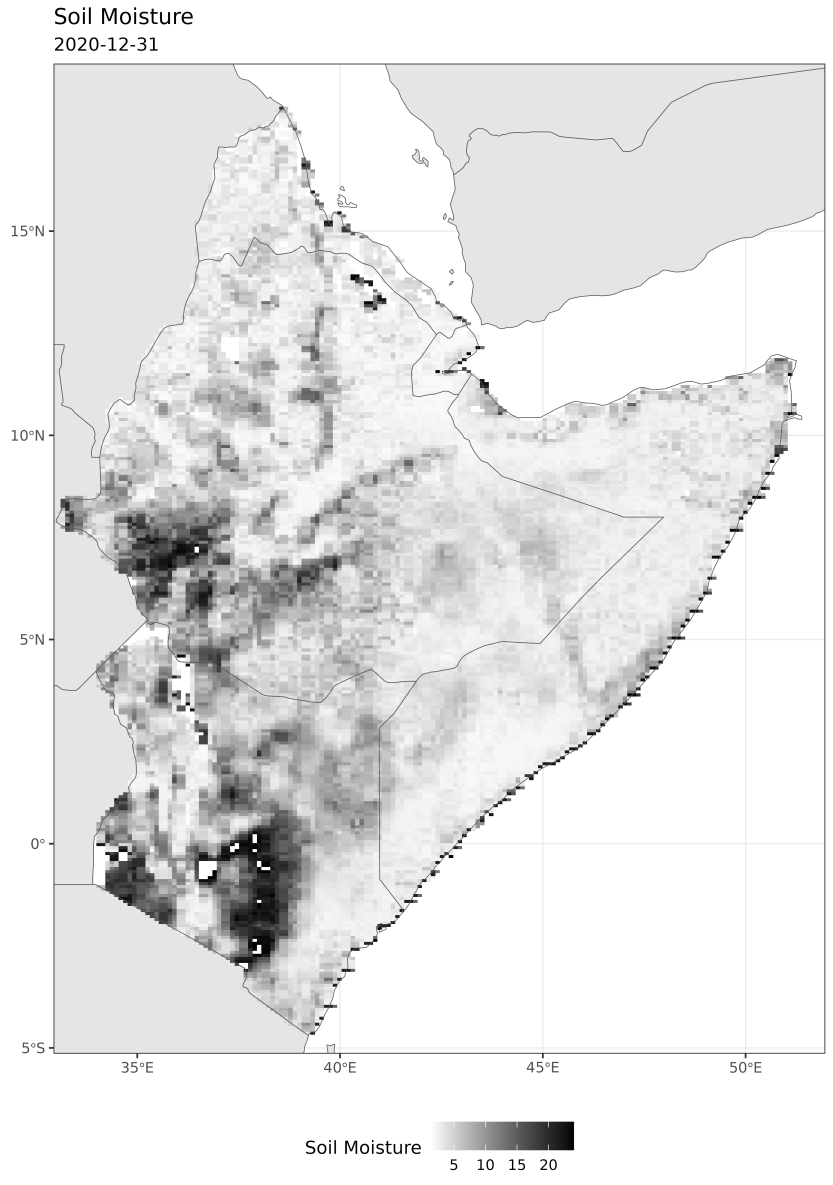
Eléphants de Basu : Pour un modèle fixe de population,

- le refus de sortir du modèle fixe de population empèche de formuler le lien entre information auxiliaire et variable d'intérêt.

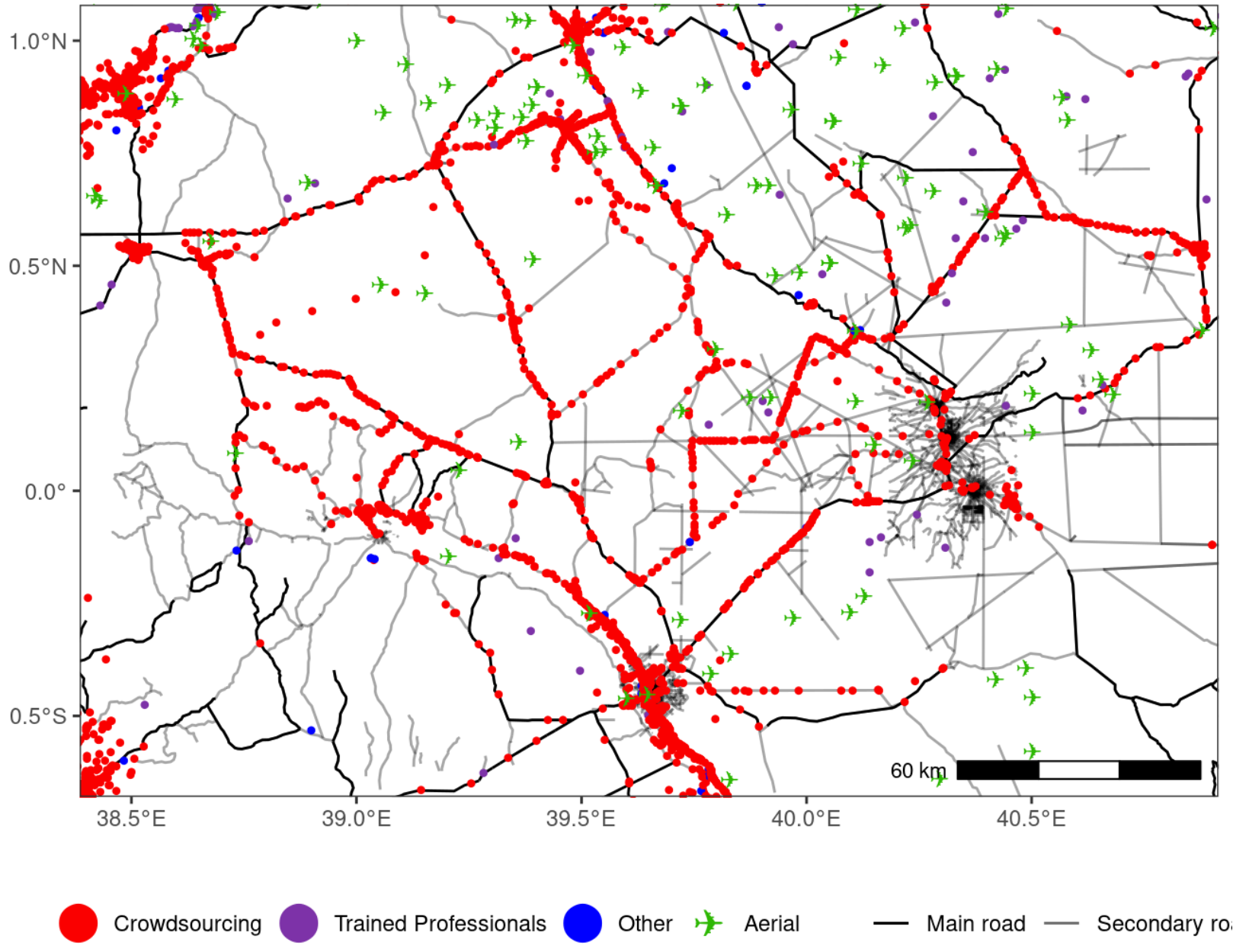
Le choix optimal (déjà appliqué):
Hors enquêtes de type enquête nationales, ou l'estimateur est suffisament concentré autour du paramètre d'intérêt (petits domaines, non réponse, données participatives) :

- dépasser le dogme de l'inférence basée sur le plan
- modéliser la sélection conditionnellement à la variable d'intérêt.
- essayer de proposer une inférence raisonnable


15°N
10°N
5°N
0°
5°S
35°E
40°E
45°E
Data source: Open street maps, Scriptoria. Date range: Jan 2
Crowdsourcing
Trained Professionals
Other
Aerial


Soil Moisture
2020-12-31
15°N
10°N
5°N
0°
5°S
35°E
40°E
45°E
50°E
Soil Moisture
5
10
15
20

Dans la littérature on trouve l'utilisation des images à haute résolution pour expliquer pourquoi un criquet était à un endroit donné et pas 10 mêtres plus loin.


1.0°N
0.5°N
0.0°
0.5°S
38.5°E
39.0°E
39.5°E
40.0°E
40.5°E
60 km
Crowdsourcing
Trained Professionals
Other
Aerial
Main road
Secondary ro

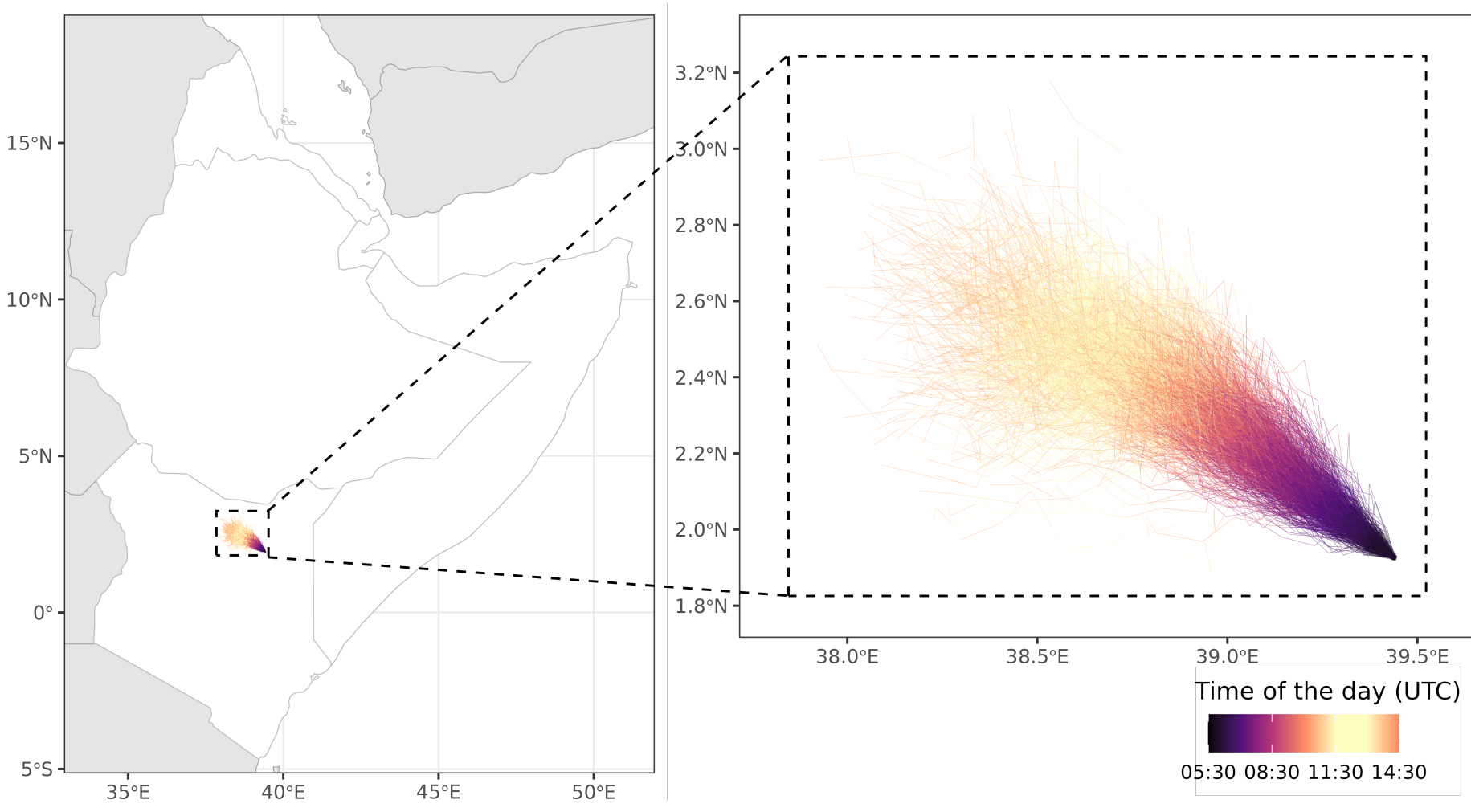
- Données participatives
- Erreur de mesure (spatiale, temporelle, sur la nature des observations)
- Campagnes de mesure (stratification temporelle)
- Non uniformité spatiale
- Pas d'information sur la planification des campagnes

Nous sommes en présence de deux processus:

- Le processus de nuisance c'est la rédaction d'un rapport.
- Le processus d'intérêt c'est la présence de criquets.

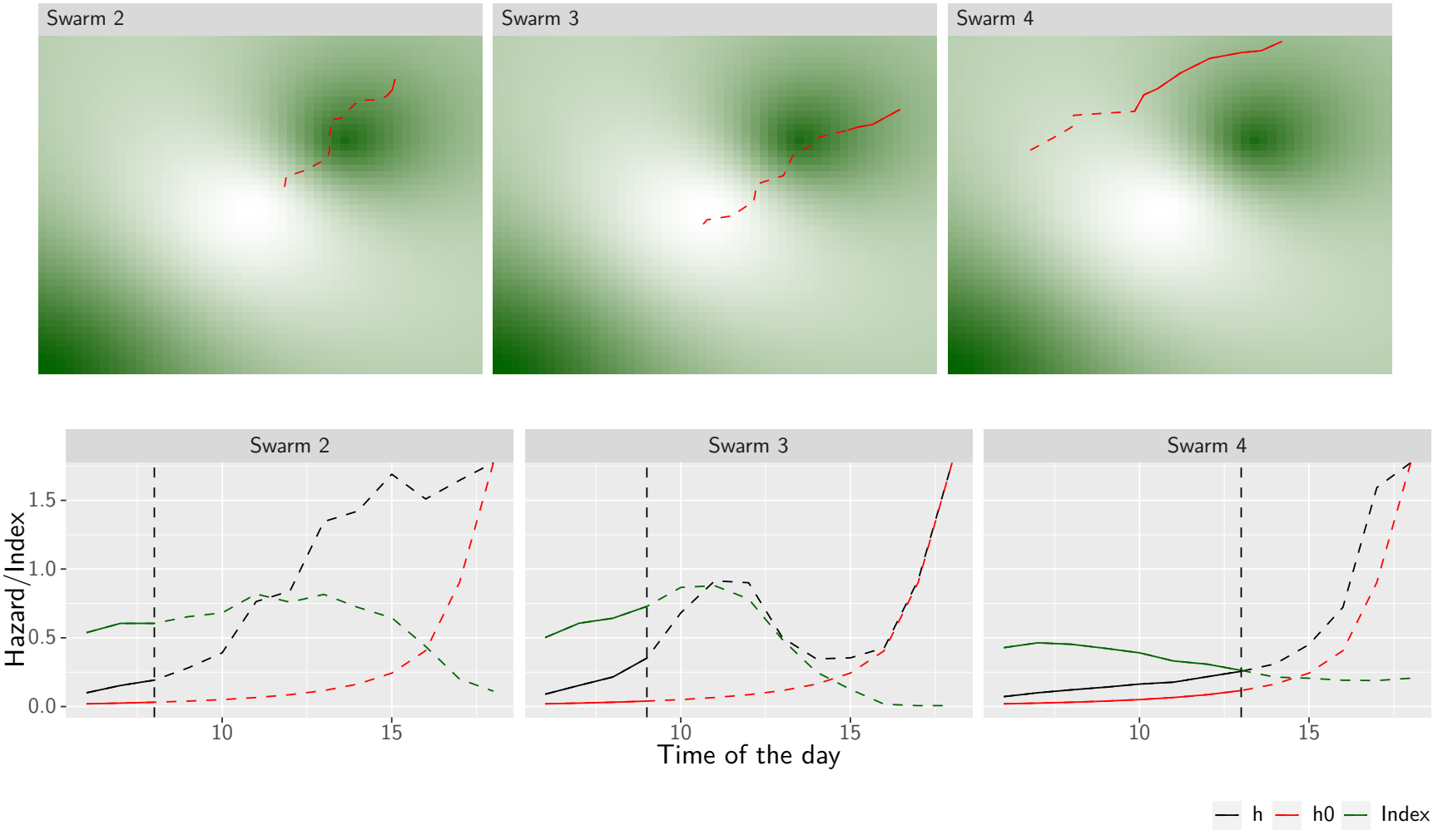
Les deux sont dépendants des variables environnementales.

Objectif: prédiction de court terme sur la direction des essaims.


15°N
10°N
5°N
0°
5°S
35°E
40°E
45°E
50°E
3.2°N
3.0°N
2.8°N
2.6°N
2.4°N
2.2°N
2.0°N
1.8°N
38.0°E
38.5°E
39.0°E
39.5°E
Time of the day (UTC)
05:30 08:30 11:30 14:30

- Modèle de survie pour la distribution des heures d'atterissage
- Modélise l'intensité de surveillance comme une fonction de la présence d'une route et de la densité de population et du pays pour chaque type de surveillance

Figure 4.1: Single swarm potential and realised trajectories



Principe de l'inférence (Sequential Monte Carlo ABC):

- Tirer les paramètres dans la loi a priori (initialisation)
- puis récursivement :
  - Simuler les observations et les vols de criquets
  - Calculer une distance entre observations et observations simulées
  - Conserver 10% des paramètres donnant les plus petites distances
  - Retirer les paramètres dans le voisinage des paramètres sélectionnés

Sont prédits à la fois:

- Le nombre de criquets
- Le nombre de signalements

Est obtenu la loi jointe a posteriori des paramètres de nuisance et d'intérêt.

5

Sélection optimale

Soit P' un ensemble de lois pour le processus de nuisance (plan à taille fixe n) par exemple.
Un critère d'optimalité possible serait :

$$\arg\min_{P' \in \mathscr{P}'} \int_{P \in \mathscr{P}} \mathrm{E}\left[\left(\hat{\theta}(X) - \theta(V,P)\right)^2\right] \mathrm{d}Q(P)$$

avec $\hat{\theta}(X) = E[\theta(V,P) \mid X]$.
Problèmes :

- ce critère n'intègre pas le cout de calcul de $\hat{\theta}(X)$, qui n'est pas calculable quelque soit P'
- l'optimum est incalculable.

Exemple :

1. Données méta génomiques
2. Echantillonnage adaptatif

6

Conclusion

Merci de votre attention