Estimation de la précision en présence de données imputées par un modèle

David Levy

INSEE, direction régionale de Rhône-Alpes

Estimation de la précision en présence de données imputées par un modèle

- Présentation de l'enquête
- Plan de sondage
- Procédure d'imputation
- Estimation de la variance
- Résultats

Présentation de l’enquête

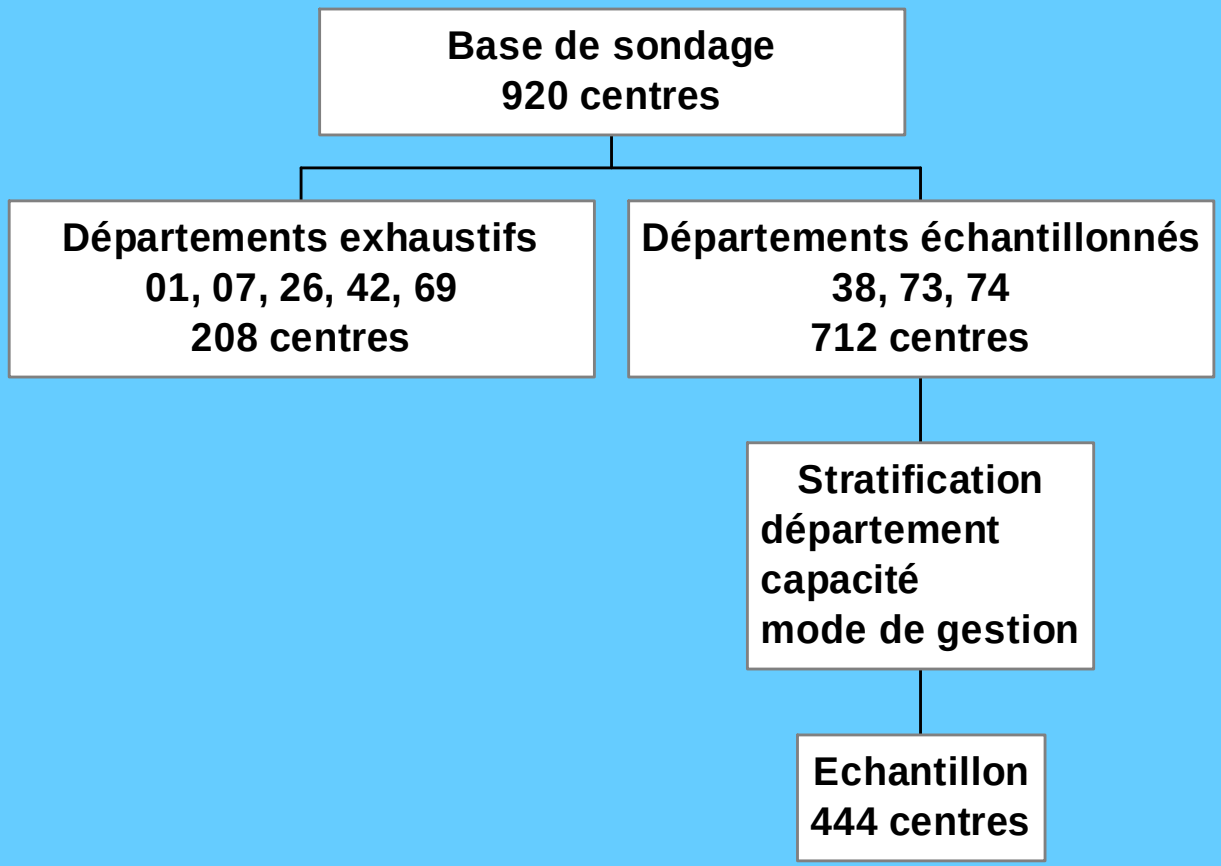
- Enquête auprès des hébergements collectifs

  Villages de vacances, auberges de jeunesse, centre de vacances

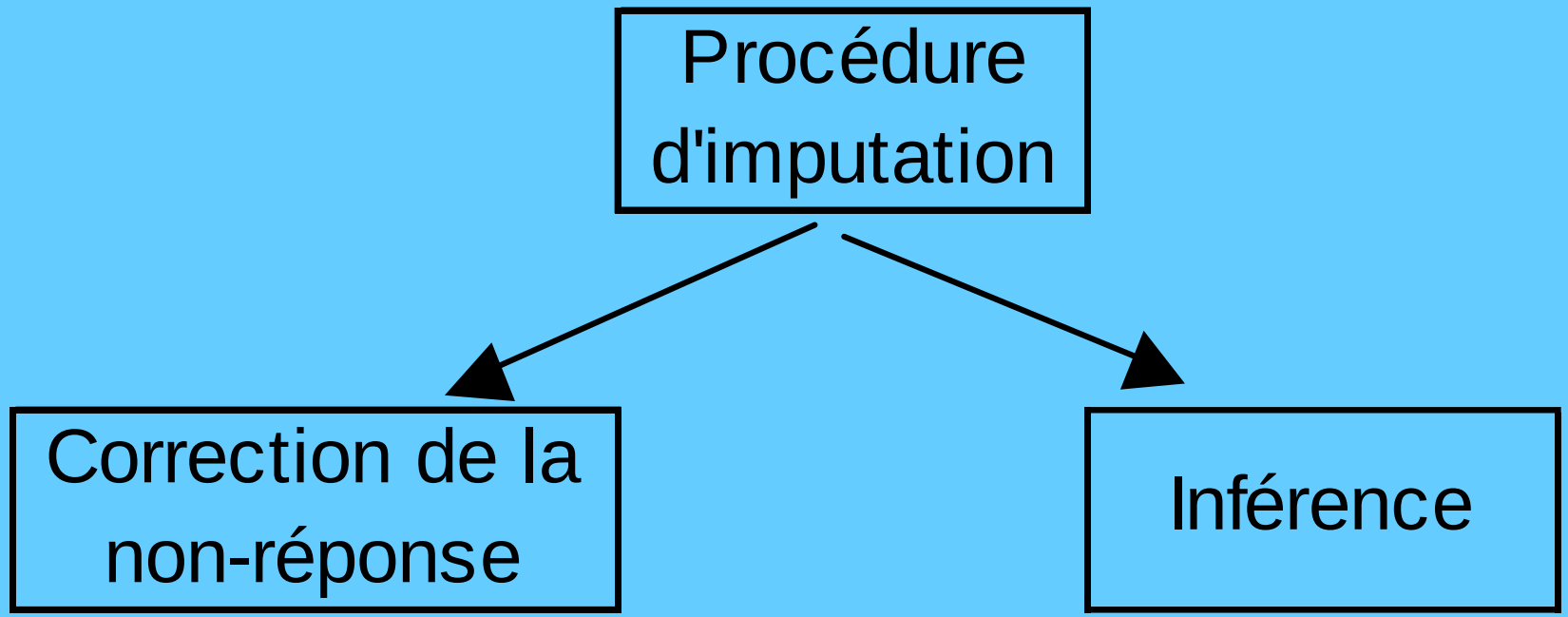
- Enquête mensuelle par voie postale de décembre à septembre

- Objectif : fournir des estimations départementales d’indicateurs d’activité (taux d’occupation, nombre de nuitées…)

Plan de sondage



Procédure d’imputation



Modèle d’imputation

Etude des nuages de points entre les variables auxiliaires (équipements offerts) et les variables à estimer (équipements occupés)

⇒ Relation linéaire

⇒ Relations différentes dans chaque strate

Procédure d'imputation

Modèle de type $y_i = a + bx_i + e_i$

$\text{avec} : \mathrm{E}(e_i) = 0$

$\mathrm{V}(e_i) = \sigma^2$

On s'intéresse à un total $Y = \sum_N Y_i$

L'estimateur s'écrit $\hat{Y} = \sum_N y_i = \sum_r y_i + \sum_{N-r} \hat{y}_i$

Soit, $\hat{Y} = \sum_r y_i + \sum_{N-r} \hat{a} + \hat{b}x_i$

En développant,

$$\hat{Y} = r\bar{y}_r + \left(\sum_N x_i - \sum_r x_i \right) \hat{b} + (N - r)\hat{a}$$

$$= r\bar{y}_r + (X - r\bar{x}_r)\hat{b} + (N - r)(\bar{y}_r - \hat{b}\bar{x}_r)$$

Finalement, $\boxed{\hat{Y} = N\left[\hat{b}\left(\bar{X} - \bar{x}_r\right) + \bar{y}_r\right]}$

On retrouve **l'estimateur par la régression**

Dans le cadre d'un plan stratifié, on utilise l'estimateur « séparé »

$$\boxed{\hat{Y} = N \sum_h W_h \left(\bar{y}_h + \hat{b}_h \left(\bar{X}_h - \bar{x}_h \right) \right)}$$

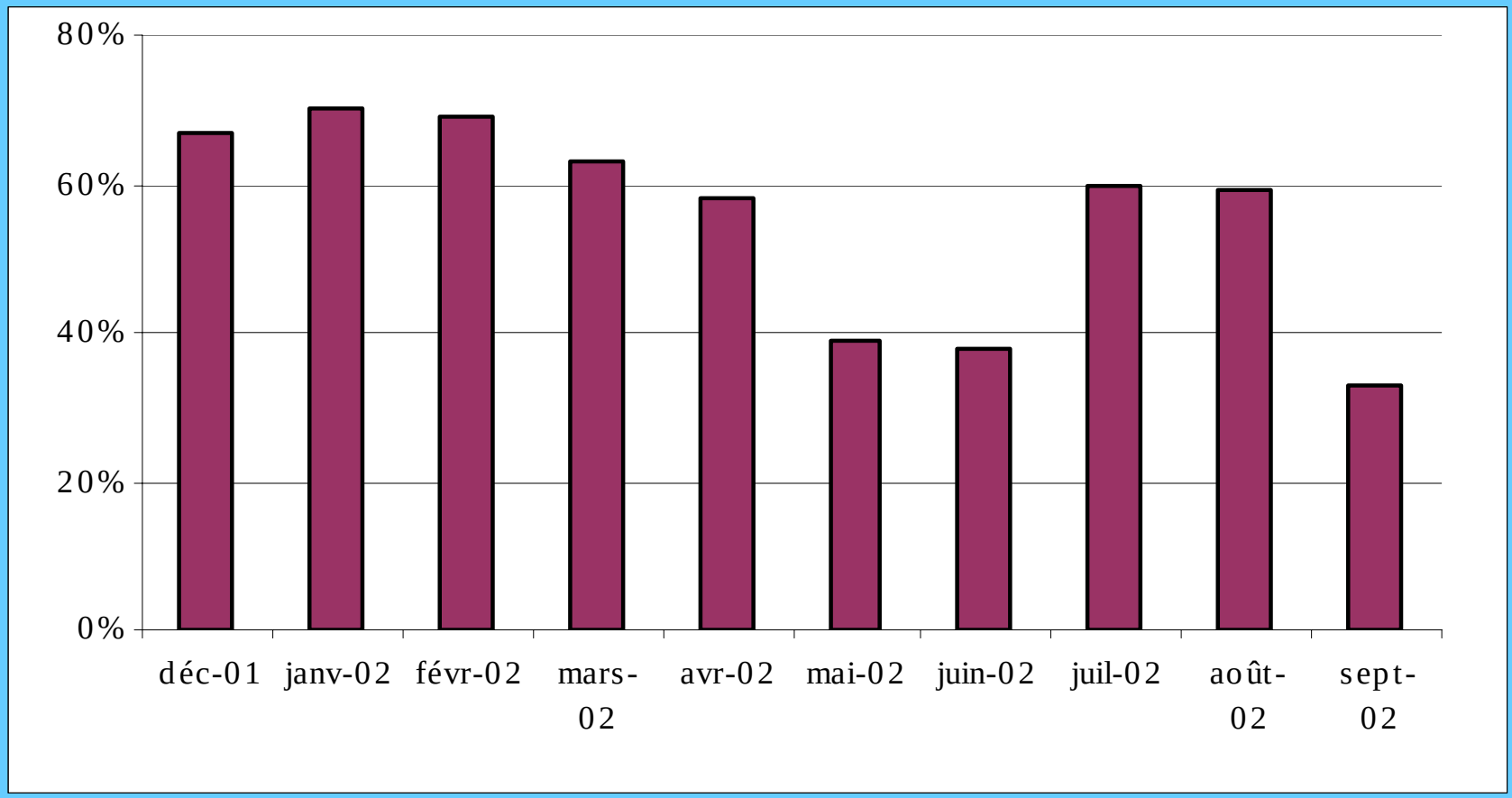
Estimation de la variance

- Formule ordinaire de la variance calculée sur les données après imputation n'est pas suffisante

- L'imputation a pour effet d'accroître la variance des estimateurs

- Compte tenu des taux de réponse, il est nécessaire de prendre en compte le modèle d'imputation dans l'expression de la variance totale

Taux de réponse mensuels



Estimation de la variance - partie échantillonnée

Données après imputation par régression

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_k = x_k \hat{B} & \text{si } k \in s \end{cases}$$

On cherche à estimer un total $\hat{Y}_{\bullet s} = \sum_s w_k y_{\bullet k}$

La variance totale s'exprime :

$$V_{totale} = E_{imp} E_p E_q \left[\left(\hat{Y}_{\bullet s} - Y_U \right)^2 \right]$$

En écrivant $\left(\hat{Y}_{\bullet s} - Y_U\right)^2 = \left[\left(\hat{Y}_s - Y_U\right) + \left(\hat{Y}_{\bullet s} - \hat{Y}_s\right)\right]^2$

$$V_{totale} = E_{imp} E_p E_q \left(\hat{Y}_s - Y_U\right)^2 + E_{imp} E_p E_q \left[\left(\hat{Y}_{\bullet s} - \hat{Y}_s\right)^2 + 2\left(\hat{Y}_s - Y_U\right)\left(\hat{Y}_{\bullet s} - \hat{Y}_s\right)\right]$$

La variance totale se décompose en deux termes :

$$\boxed{V_{tot} = V_s + V_{imp}}$$

Variance due à l'échantillonnage + Variance due à l'imputation

Estimation de la variance - Partie exhaustive

Soit N la taille de la population et r le nombre de répondants. On suppose de plus

$$r \rightarrow \beta(N, p)$$

Variance conditionnée à r :

$$V(\hat{Y}_{reg}) = V_r\left[E\left(\hat{Y}_{reg} / r\right)\right] + E_r\left[V\left(\hat{Y}_{reg} / r\right)\right]$$

En supposant l'estimateur par la régression sans biais, soit $E\left(\hat{Y}_{reg} / r\right) = Y$

On a : $V(\hat{Y}_{reg}) \approx E_r\left[V\left(\hat{Y}_{reg} / r\right)\right]$

La variance de l'estimateur par la régression peut également s'écrire sous la forme :

$$V(\hat{Y}_{reg}) = \frac{1-f}{n} S_u^2$$

Ainsi : $$V(\hat{Y}_{reg}) = E_r\left[\left(1-\frac{r}{n}\right)\frac{S_u^2}{r}\right]$$

$$V(\hat{Y}_{reg}) = \left[E_r\left(\frac{1}{r}\right) - \frac{1}{N}\right] S_u^2$$

Le calcul de $E\left(\frac{1}{r}\right)$ se fait par linéarisation

sachant que $r \to \beta(N, p)$

$$E(r) = Np$$

$$V(r) = (1-p)Np$$

On trouve : $E\left(\frac{1}{r}\right) \approx \frac{1}{Np}\left(1+\frac{1-p}{Np}\right)$

Au final, $V(\hat{Y}_{reg}) = \left(\frac{1}{Np}\left(1+\frac{1-p}{Np}\right)-\frac{1}{N}\right)S_u^2$

Avec p estimé par $\hat{p} = \frac{r}{N}$

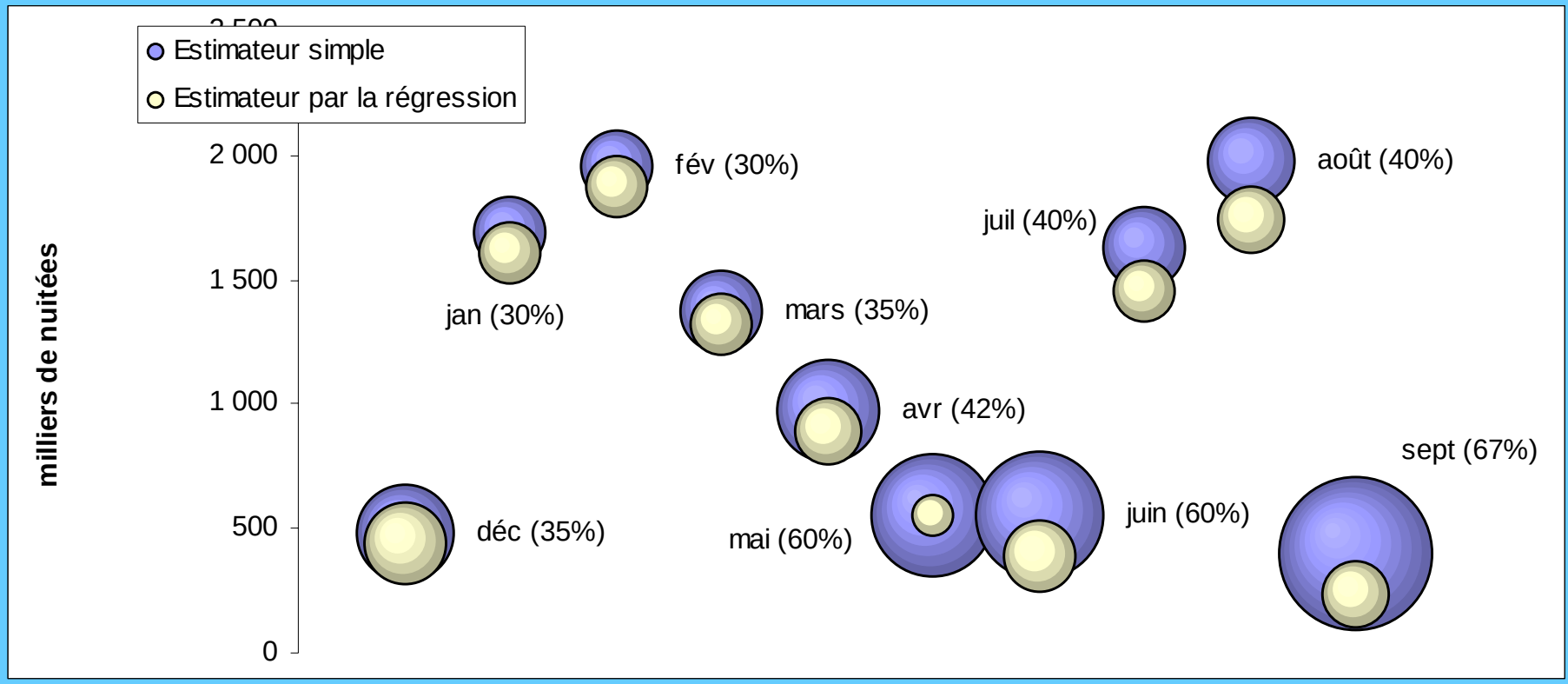
Résultats

- On estime le total des nuitées
- Modèle d'imputation : nuitées = a + b.CAPACITE + u
- Estimations calculées de décembre 01 à septembre 02

  - Comparaison de l'estimateur par la régression avec l'estimateur simple

  - Part de la variance due au modèle dans la variance totale

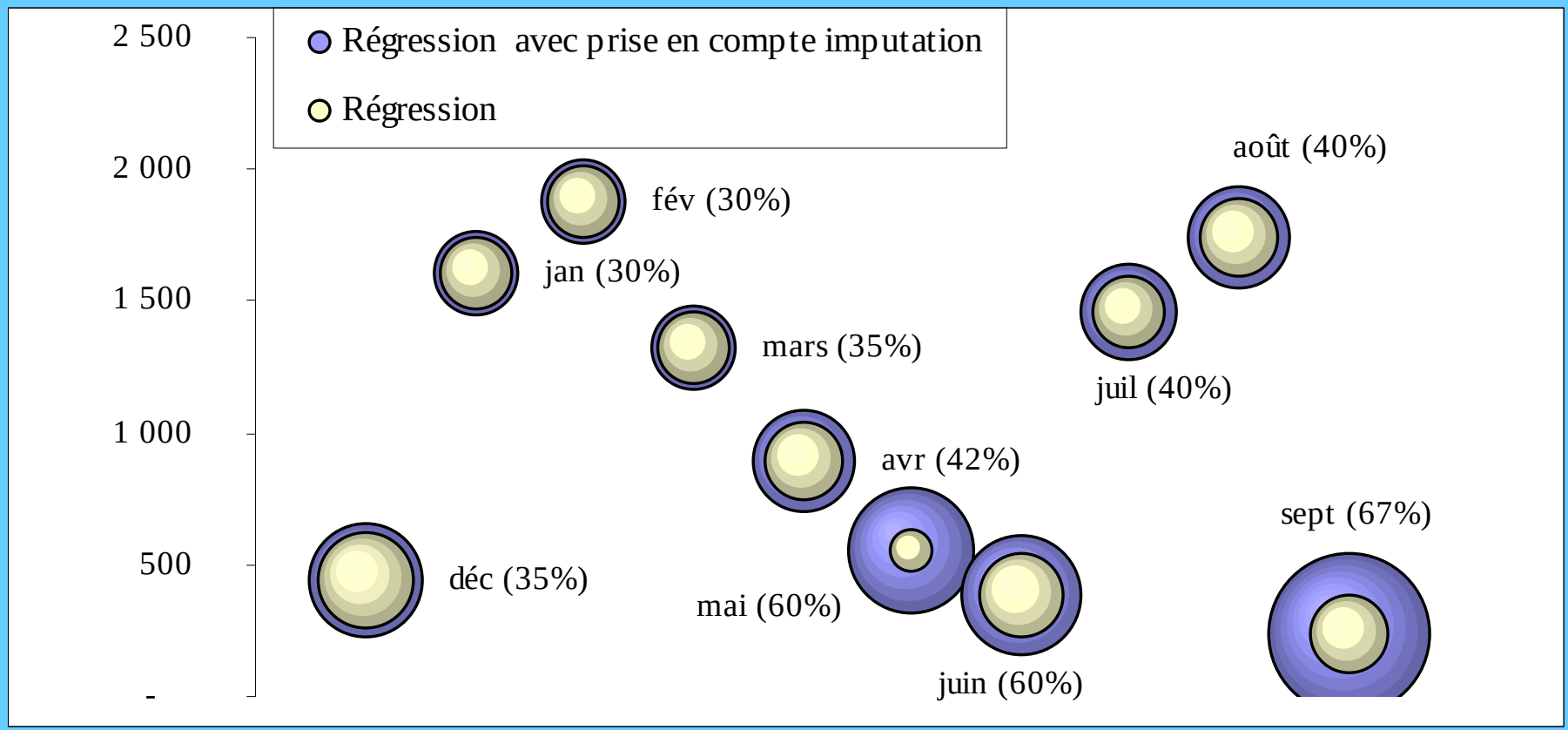
Estimateur par la régression / estimateur simple



Le centre d'un cercle représente l'estimation d'un mois, le diamètre l'importance de l'intervalle de confiance

Les % indiquent le taux d'imputation

Estimateur par la régression avec prise en compte du modèle d'imputation



Le centre d'un cercle représente l'estimation d'un mois, le diamètre l'importance de l'intervalle de confiance

Les % indiquent le taux d'imputation

Part de la variance d'imputation dans la variance totale et taux d'imputation