Estimateurs non paramétriques de la fonction de répartition d'une variable censurée à droite sur petits domaines : approche basée sur un modèle

Eve Leconte [1] & Sandrine Casanova [2]

[1] *TSE-R, Université Toulouse 1 Capitole,*
21, allée de Brienne, 31042 TOULOUSE, France, eve.leconte@tse-fr.eu
[2] *TSE-R, Université Toulouse 1 Capitole,*
21, allée de Brienne, 31042 TOULOUSE, France, sandrine.casanova@tse-fr.eu

Résumé

Nous nous intéressons au cas de l'estimation de la fonction de répartition (fdr) sur petits domaines lorsque la variable d'intérêt est censurée à droite, cas n'ayant, à notre connaissance, jamais été étudié dans le cadre des sondages. Si les domaines sont de taille suffisante, l'estimation peut être basée uniquement sur les données relatives au domaine et les estimateurs produits sont de précision acceptable. Cependant, dans beaucoup d'applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. De l'information doit alors être "empruntée" aux autres domaines pour améliorer la précision. Nous nous plaçons dans une approche basée sur un modèle non paramétrique et disposons d'une information auxiliaire fournie par une covariable, connue sur toute la population. Notre estimateur est une adaptation au cas censuré de la technique de Casanova (2012) : la variable d'intérêt d'un individu hors échantillon est prédite à l'aide d'un quantile conditionnel dont l'ordre doit être préalablement estimé. Les estimations de ces deux étapes utilisent les échantillons de tous les domaines afin d'emprunter de la force aux voisins, en se basant sur l'estimateur de Kaplan-Meier généralisé lissé (Leconte *et al.*, 2002) de la fdr conditionnelle. Des simulations comparent les performances du nouvel estimateur avec l'estimateur *model-based* de Casanova et Leconte (2015) et l'estimateur de Kaplan-Meier, qui n'utilisent que l'information relative au domaine. Des données sur le temps d'accès au premier emploi de jeunes filles diplômées de la région Occitanie illustrent notre méthode, les domaines étant le niveau et le type de la formation suivie.

Mots-clés. Fonction de répartition, information auxiliaire, données censurées, petits domaines, quantiles conditionnels non paramétriques, estimateur de Kaplan-Meier généralisé, estimation model-based.

Abstract. In survey analysis, the estimation of the cumulative distribution function (cdf) is of great interest in order to derive mean or median estimators for the population or for sub-populations (domains). We consider the case where the response variable is right censored. In this framework, a nonparametric model-based estimator of the cdf in small domains is proposed. The estimator is the extension to the censored case of the

cdf estimator of Casanova (2012) : it uses auxiliary information brought by a continuous covariate and is based on nonparametric quantile regression. The estimator has been compared by simulations with the cdf model-based estimator of Casanova and Leconte (2015) and the Kaplan-Meier estimator, which only use information of the domain. Access times to the first job for females graduates in the Occitania region are used to illustrate the new methodology.

Keywords. Cumulative distribution function, auxiliary information, censored data, generalized Kaplan-Meier estimator, nonparametric conditional quantiles, small area estimation.

1 Introduction

Un axe de recherche important en sondages est l'estimation sur petits domaines, qui correspond à l'estimation des quantités d'intérêt sur des sous-populations de petite taille. Si un domaine est de taille suffisante, l'estimation des paramètres d'intérêt peut se restreindre aux données relatives aux individus du domaine et les estimateurs produits sont de précision acceptable. Cependant, dans la plupart des applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. L'estimation se fonde alors sur une information auxiliaire fournie par une covariable et de l'information est "empruntée" aux autres domaines.

Le modèle classique qui permet de capturer l'effet domaine est le modèle linéaire mixte (voir Rao, 2003). L'estimation des coefficients de régression et la prédiction des effets aléatoires s'obtiennent par maximum de vraisemblance sous l'hypothèse de normalité des erreurs, conduisant au meilleur prédicteur empirique linéaire sans biais (EBLUP) de la variable d'intérêt. Cependant, ces modèles reposent sur des hypothèses fortes comme la normalité et l'homoscédasticité des erreurs. Pour prendre en compte une relation non-linéaire entre la variable d'intérêt et la variable auxiliaire, Salvati, Chandra, Ranalli et Chambers (2010) ont proposé une version non paramétrique de l'EBLUP pour la moyenne sur un petit domaine en utilisant des splines pénalisées.

Dans un cadre non paramétrique, Salvati, Ranalli et Pratesi (2010) estiment la moyenne sur un petit domaine par des M-quantiles conditionnels basés sur des splines pénalisées. En ce qui concerne l'estimation de la fdr, Chambers et Tzavidis (2006) prédisent la fdr d'un petit domaine à l'aide de M-quantiles conditionnels paramétriques. Casanova (2012) a étendu la technique de Chambers et Tzavidis (2006) au cas non paramétrique en utilisant des estimateurs à noyaux des M-quantiles conditionnels. Enfin, Salvati, Chandra et Chambers (2010) proposent d'estimer la fdr sur un petit domaine par la moyenne des données de l'échantillon du petit domaine pondérée par des poids d'échantillonnage calibrés.

Nous proposons un estimateur non paramétrique de la fdr sur petits domaines dans le cas où la variable d'intérêt est censurée à droite, cas n'ayant, à notre connaissance, jamais été considéré dans la littérature. Après avoir défini les notations dans ce nouveau cadre, nous détaillons dans la section 2 la construction de l'estimateur, qui est une adaptation au cas censuré de la technique de Casanova (2012). Un exemple d'application à des données sur les temps d'accès au premier emploi de jeunes diplômées illustre la méthode dans la section 3. Enfin, la section 4 présente des simulations basées sur le modèle qui comparent cet estimateur avec l'estimateur naïf de Kaplan-Meier et l'estimateur non paramétrique *model-based* de Casanova et Leconte (2015), calculés sur les points échantillonnés du domaine.

2 Estimateurs de la fdr sur un domaine

2.1 Notations

Nous considérons une population finie $\mathcal{P}$ de taille N, partitionnée en m sous-populations — appelées domaines — U_i de taille N_i, $i = 1, \ldots, m$. Soient s un échantillon de $\mathcal{P}$ de taille n et $s_i = s \cap U_i$ un échantillon du domaine U_i de taille n_i. t_{ij} est la variable d'intérêt mesurée pour le j-ème individu du domaine U_i. On suppose que t_{ij} est seulement connu sur s_i et éventuellement censuré à droite par z_{ij}. Avec les notations d'Efron, nous observons, sur l'échantillon s_i, $y_{ij} = \min(t_{ij}, z_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} < z_{ij})$. Nous noterons $y_{i(n_i)} = \max_{j \in s_i} y_{ij}$ le plus grand délai de l'échantillon s_i. Nous disposons en outre d'une information auxiliaire x_{ij}, valeur d'une variable continue X pour l'individu j du domaine i, connue sur toute la population.

Dans le cadre des sondages, la fdr de la variable d'intérêt T sur le domaine U_i s'écrit $F^i(t) = \dfrac{1}{N_i} \sum_{j \in U_i} \mathbb{1}(t_{ij} \leq t)$ que l'on peut décomposer en

$$F^i(t) = \frac{1}{N_i} \left(\sum_{j \in s_i} \mathbb{1}(t_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \mathbb{1}(t_{ij} \leq t) \right). \tag{1}$$

2.2 Les estimateurs qui n'utilisent que les données du domaine

L'estimateur de Kaplan-Meier sur le domaine

Il est bien connu que la fonction de répartition empirique ne fournit pas un estimateur convergent de la fdr en présence de censure. Par contre, le complémentaire à 1 de l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958) de la fonction de survie calculé à partir des points de l'échantillon s_i du domaine U_i converge uniformément presque sûrement vers F^i.

Comme l'estimateur original de Kaplan-Meier n'est pas défini après le plus grand délai $y_{i(n_i)}$ si celui-ci correspond à une censure, afin d'obtenir une fonction de répartition, nous considérons la version d'Efron (1967) définie par :

$$\widehat{F}^i_{\mathrm{KM}}(t) = \begin{cases} 1 - \displaystyle\prod_{j \in s_i} \left\{ 1 - \frac{1}{\displaystyle\sum_{r \in s_i} \mathbb{1}(y_{ir} \geq y_{ij})} \right\}^{\mathbb{1}(y_{ij} \leq t, \delta_{ij} = 1)} & \text{si } \ t < y_{i(n_i)}, \\ 1 & \text{sinon.} \end{cases} \tag{2}$$

L'estimateur de Casanova et Leconte (2015)

On peut sans doute améliorer l'estimation de la fdr F^i en construisant des estimateurs basés sur un modèle, qui utilisent de l'information auxiliaire, en se basant sur la formule (1). Dans ce cadre, un estimateur naturel de F^i peut être obtenu en appliquant l'estimation de la fdr proposée par Casanova et Leconte (2015) au domaine U_i, ce qui revient à remplacer s par s_i dans toutes les formules. Cette approche nécessite de postuler le modèle de superpopulation ξ suivant :

$$t_{ij} = m(x_{ij}) + e_{ij}, \ i = 1, \ldots, m, \ \ j = 1, \ldots, N_i,$$

où les e_{ij} sont des variables i.i.d. de fdr G^i et où $m(x_{ij})$ désigne la médiane conditionnelle de T sachant $X = x_{ij}$. Il est également nécessaire de supposer que le plan de sondage est non informatif, ce qui signifie que le modèle sur lequel nous nous basons est valide à la fois pour l'échantillon et la population. Cela justifie aussi le choix d'un modèle non paramétrique pour lequel le risque de mauvaise spécification est réduit.

L'estimateur qui en résulte est

$$\widehat{F}^i_{\mathrm{M}}(t) = \frac{1}{N_i} \left(n_i \widehat{F}^i_{\mathrm{KM}}(t) + \sum_{j \in U_i \setminus s_i} \widehat{G}^i_{\mathrm{KM}}(t - \widehat{m}(x_{ij})) \right), \tag{3}$$

où $\widehat{m}(x_{ij}) = \widehat{F}^{i,-1}_{\mathrm{SGKM}}(0,5 \mid x_{ij})$ est un estimateur de la médiane conditionnelle $m(x_{ij})$ qui utilise la version lissée $\widehat{F}^i_{\mathrm{SGKM}}$ de l'estimateur de Kaplan-Meier généralisé de la fdr conditionnelle proposée par Leconte *et al.* (2002). Les fenêtres h_T et h_X nécessaires à son calcul doivent être remplacées par des fenêtres adaptées à chaque domaine. Enfin, $\widehat{G}^i_{\mathrm{KM}}$ est l'estimateur de Kaplan-Meier de la fdr des erreurs G^i, calculé à partir des résidus $\widehat{\varepsilon}_{ij} = y_{ij} - \widehat{m}(x_{ij}), \ j \in s_i$.

2.3 Le nouvel estimateur

Quand la taille du domaine est petite, les estimateurs précédents peuvent avoir une grande variance et des méthodes qui utilisent l'information apportée par les autres do-

maines doivent être préférées pour améliorer la précision de l'estimation. Pour ce faire, nous proposons la procédure suivante.

Le premier terme de (1) n'est pas connu à cause de la censure à droite. Comme il peut s'écrire :

$$\frac{1}{N_i}\sum_{j\in s_i}\mathbb{1}(t_{ij}\leq t)=\frac{n_i}{N_i}\left(\frac{1}{n_i}\sum_{j\in s_i}\mathbb{1}(t_{ij}\leq t)\right), \tag{4}$$

nous reconnaissons la fdr basée sur l'échantillon s_i entre les parenthèses, qui peut donc être estimée par l'estimateur de Kaplan-Meier calculé sur les individus de l'échantillon s_i.

En ce qui concerne le second terme de (1), nous utilisons l'information de l'échantillon total s pour prédire la valeur de la variable réponse pour les individus non échantillonnés du domaine U_i. Dans ce cadre, nous devons supposer le modèle de superpopulation ζ :

$$t_{ij}=m(q_i,x_{ij})+\varepsilon_{ij},\quad i=1,\ldots,m,\quad j=1,\ldots,N_i,$$

où les ε_{ij} sont des variables i.i.d. de fdr G'^i ; q_i est un coefficient de $[0,1]$ caractérisant la position du domaine U_i et $m(q_i,x_{ij})$ est le quantile conditionnel d'ordre q_i de T sachant $X=x_{ij}$. Chaque valeur t_{ij} peut en effet être considérée comme le quantile conditionnel de T sachant $X=x_{ij}$ pour un certain ordre noté $q(t_{ij},x_{ij})$. De ce fait, en imitant Chambers et Tzavidis (2006), le coefficient q_i du domaine U_i peut être défini comme la moyenne ou la médiane des ordres-quantiles conditionnels $q(t_{ij},x_{ij})$ des individus j du domaine U_i. On peut noter que les ordres-quantiles conditionnels sont définis au niveau de la population et que nous nous attendons à ce que des individus du même domaine aient des valeurs de leurs ordres-quantiles assez proches si une part de la variabilité des données est expliquée par le domaine.

Les ordres-quantiles conditionnels sont estimés à l'aide de la version lissée de l'estimateur de Kaplan-Meier généralisé (Leconte *et al.*, 2002) calculé sur l'échantillon total s :

$$\widehat{q(t_{ij},x_{ij})}=\widehat{F}_{\text{SGKM}}(y_{ij}\mid x_{ij}).$$

Comme les valeurs y_{ij} peuvent être censurées à droite, les ordres $\widehat{q(t_{ij},x_{ij})}$ le sont aussi. Ainsi, pour estimer l'ordre global q_i du domaine U_i par la moyenne ou la médiane des ordres $\widehat{q(t_{ij},x_{ij})}$, nous devons d'abord estimer leur fdr en tenant compte des observations censurées. Cela peut être facilement réalisé en utilisant une fois encore l'estimateur de Kaplan-Meier. Comme la médiane est plus facile à estimer que la moyenne en présence de censure, nous choisissons donc d'estimer q_i par la médiane $\widehat{q}_i$ des ordres-quantiles, obtenue en inversant l'estimateur de Kaplan-Meier. Comme $E_\zeta\left(\mathbb{1}(t_{ij}\leq t)\right)=P(t_{ij}\leq t)=G'^i(t-m(q_i,x_{ij}))$, $\mathbb{1}(t_{ij}\leq t)$ peut être prédite en estimant $G'^i(t-m(q_i,x_{ij}))$. Un estimateur naturel $\widehat{m}(\widehat{q}_i,x_{ij})$ de $m(q_i,x_{ij})$ est le quantile conditionnel d'ordre $\widehat{q}_i$ sachant x_{ij}, qui est la solution en θ de $\widehat{F}_{\text{SGKM}}\left(\theta\mid x_{ij}\right)=\widehat{q}_i$ et est donc obtenue en inversant $\widehat{F}_{\text{SGKM}}$. On peut remarquer que, là encore, comme pour l'estimation de l'ordre-quantile q_i, l'échantillon tout entier est utilisé pour calculer cet estimateur, ce qui permet d'emprunter de la force

aux autres domaines. Comme cela a été fait pour l'estimateur de Casanova et Leconte (2015), $G'^i(t - m(q_i, x_{ij}))$ peut être estimé par l'estimateur de Kaplan-Meier calculé à partir des résidus censurés $\widehat{\varepsilon}_{ij} = y_{ij} - \widehat{m}(\widehat{q}_i, x_{ij}),\ j \in s_i$. Nous notons cet estimateur $\widehat{G}'^i_{\text{KM}}$ et en déduisons l'estimateur suivant de la fdr de T dans le domaine U_i :

$$\widehat{F}^i_{\text{Q}}(t) = \frac{1}{N_i}\left(n_i \widehat{F}^i_{\text{KM}}(t) + \sum_{j \in U_i \setminus s_i} \widehat{G}'^i_{\text{KM}}\left(t - \widehat{m}(\widehat{q}_i, x_{ij})\right)\right). \tag{5}$$

Cet estimateur est une fdr de façon évidente.

3 Application

Nous avons analysé par nos méthodes des données du Centre d'études et de recherches sur les qualifications (Céreq), en collaboration avec Hélène Couprie, enseignant-chercheur en économie. Le Céreq interroge par sondage les jeunes diplômés sur leur devenir professionnel trois ans après l'obtention de leur diplôme (étude rétrospective sur la situation mensuelle des trois années précédentes). Nous nous sommes restreintes aux 10135 jeunes filles des régions Midi-Pyrénées et Languedoc-Roussillon qui sont sorties de l'enseignement secondaire en 2010. La variable d'intérêt T est le temps d'accès au premier emploi depuis l'obtention du diplôme, censuré à droite pour les jeunes filles qui n'ont pas trouvé d'emploi à la fin de l'enquête (12,5 % dans nos données). Le Céreq cherche à avoir des statistiques selon le niveau et le type de la formation suivie, ce qui partitionne la population en 34 domaines (de taille variant de 7 à 1480 jeunes filles). L'échantillon, d'un effectif total de 306, se partitionne par domaine en sous-échantillons de taille 1 à 37. La variable auxiliaire choisie est le taux de chômage local de la zone d'emploi de l'établissement de fin d'études, qui est significativement et négativement lié à la probabilité de trouver un emploi (p=0,014). La figure 1 présente les fonctions de survie correspondant aux trois estimateurs $\widehat{F}^i_{\text{KM}}$, $\widehat{F}^i_{\text{M}}$ et $\widehat{F}^i_{\text{Q}}$ pour 6 domaines de tailles très différentes.

4 Simulations

Des simulations basées sur le modèle sont en cours pour comparer le nouvel estimateur $\widehat{F}^i_{\text{Q}}$ à l'estimateur de Kaplan-Meier $\widehat{F}^i_{\text{M}}$ et à l'estimateur $\widehat{F}^i_{\text{M}}$ de Casanova et Leconte (2015), qui n'utilisent que les données relatives au domaine. Pour cela, nous simulons un modèle log-linéaire de régression avec un effet aléatoire du domaine : $ln(t_{ij}) = 4 - \nu * x_{ij} + u_i + \varepsilon_{ij}$ où la covariable x_{ij} suit une loi uniforme sur l'intervalle $[1, 4]$. Le terme d'erreur ε_{ij} suit une loi de valeur extrême de façon à ce que les t_{ij} suivent une loi exponentielle. Ce modèle correspond dans chaque domaine à un modèle à risques proportionnels avec un risque relatif égal à ν, ce qui signifie que le rapport des risques de deux individus dont les covariables x diffèrent d'une unité ne dépend pas du temps et vaut $\exp(\nu)$. Les effets u_i

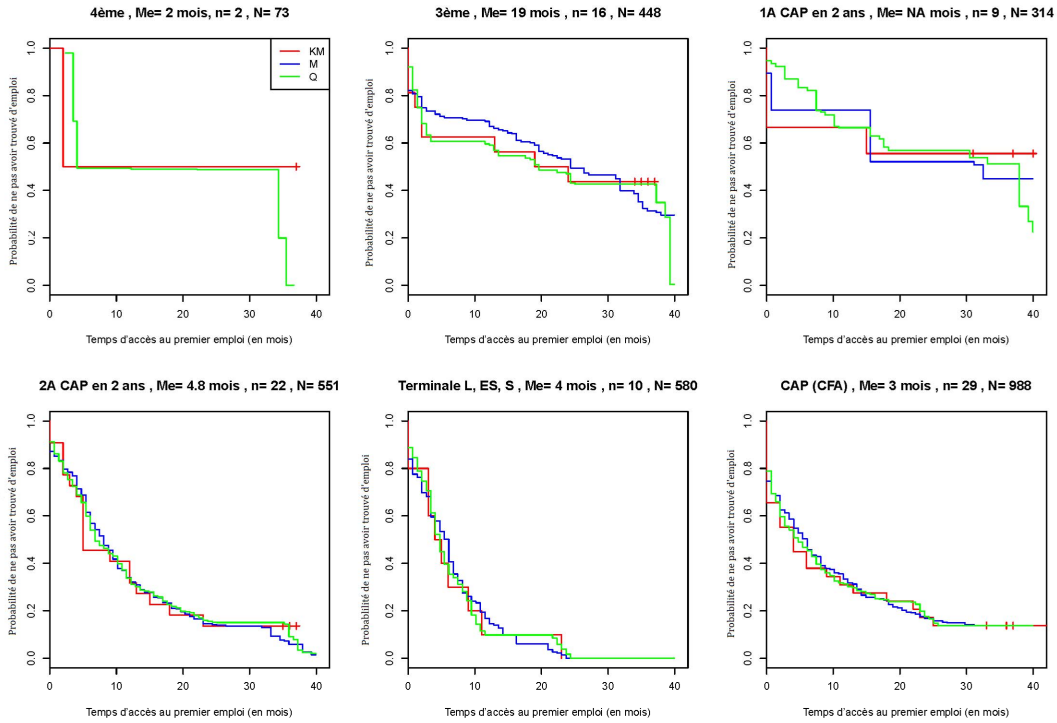

FIGURE 1 – Probabilité de ne pas avoir trouvé de premier emploi, estimée par les trois estimateurs pour six domaines différents correspondant à un niveau et type de formation. La médiane indiquée (Me) est obtenue en inversant l'estimateur naïf de Kaplan-Meier.

des domaines sont générés selon une loi normale $\mathcal{N}(0, \sigma^2)$. De plus, t_{ij} est censuré par c_{ij}, la censure suivant une loi uniforme. Les paramètres qui varient sont le taux de censure, la taille des échantillons des domaines, la force de la relation entre la variable d'intérêt et la variable auxiliaire mais aussi la part de la variabilité due aux domaines.

Les premiers résultats montrent que l'estimateur $\widehat{F}^i_{\mathrm{M}}$ est toujours meilleur que l'estimateur naïf de Kaplan-Meier sur le domaine. Le nouvel estimateur $\widehat{F}^i_{\mathrm{Q}}$ qui emprunte de la force aux voisins semble plus performant que $\widehat{F}^i_{\mathrm{M}}$ quand l'échantillon du domaine est de très petite taille (inférieure à 6 individus) et que la variabilité due aux domaines est faible ou modérée. Quand les domaines diffèrent plus fortement, les deux estimateurs sont sensiblement équivalents.

Bibliographie

Casanova S. (2012). Using nonparametric conditional M-quantiles to estimate a cumulative distribution function in a domain. *Annales d'Economie et de Statistique*, 107/108, 287–297.

Casanova S. et Leconte E. (2015). A nonparametric model-based estimator for the cumulative distribution function of a right censored variable in a finite population. *Journal of Surveys : Statistics and Methodology*, 3, 317–338.

Chambers R. L. et Tzavidis N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268.

Efron B. (1967). The two sample problem with censored data. *Proc. 5th Berkeley Symp.*, 4, 831–853.

Kaplan E. L. et Meier P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53.282, 457–481.

Leconte E., Poiraud-Casanova S. et Thomas-Agnan C. (2002). Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Analysis*, 8, 229–246.

Rao J. N. K. (2003). *Small area estimation.* Wiley, New-York.

Salvati N., Chandra H. et Chambers R. (2010). Model-based direct estimation of small area distributions. *Centre for Statistical and Survey Methodology*, working paper.

Salvati N, Chandra H., Ranalli M. G. et Chambers R. (2010). Small area estimation using a nonparametric model-based direct estimator. *Journal of Computational Statistics and Data Analysis*, 54, 2159–2171.

Salvati N, Ranalli M. G. et Pratesi M. (2010). Small area estimation of the mean using nonparametric M-quantile regression : a comparison when a linear mixed model does not hold. *Journal of Statistical Computation and Simulation*, 1–20.