

Research Group: *Econometrics and Statistics*

*December, 2009*

# Using Nonparametric Conditional M-Quantiles to Estimate a Cumulative Distribution Function in a Domain

SANDRINE CASANOVA

# Using nonparametric conditional M-quantiles to estimate a cumulative distribution function in a domain

Sandrine Casanova

*TSE (GREMAQ), Université Toulouse 1 Capitole,  
21, allée de Brienne, 31000 TOULOUSE, France*

E-mail : sandrine.casanova@univ-tlse1.fr

Tel : 33.(0)5.61.12.85.43, Fax : 33.(0)5.61.22.55.63

## *Summary*

Estimating the cumulative distribution function in survey sampling is of interest on the population but also on a sub-population (domain). However, in most practical applications, sample sizes in the domains are not large enough to produce sufficiently precise estimators. Therefore, we propose new nonparametric estimators of the cumulative distribution function in a domain based on M-quantile estimation. The obtained estimators are compared by simulations and applied to real data.

## **1 Introduction**

Sample surveys allow to obtain estimates for characteristics of interest at both population and domain level. A domain can be for example a geographic area or a socio-demographic group. If a domain is large enough, the estimation of interest parameters relies on data from sample units in the domain and the resultant estimates will be of acceptable precision. However in most practical applications, sample sizes are not large enough to produce sufficiently precise estimators. In such situations, the estimation is based on auxiliary information related to the variable of interest and information is "borrowed" from the other domains. The term Small area estimation denotes the set of techniques of estimation for such domains (Rao, 2003). The actual literature focuses on estimation of totals or means in small areas but in many applications the interest parameters are more complex: they can be quantiles, or other non linear parameters derived from the cumulative distribution function of the interest variable. Estimating the cumulative distribution function (cdf) in survey sampling has been widely studied at population level. In a parametric regression frame and a model-based approach, Chambers and Dunstan (1986) propose an estimator of the cdf. They prove the asymptotic normality of their estimator when both the population and the sample sizes tend to infinity. More recently, Breidt, Johnson and Opsomer (2008) propose a non parametric approach using local polynomials to estimate the cdf. Small area estimation methods are used when sample data can not provide acceptably precise direct estimators. The classical technique in order to capture the domain effect is the linear mixed random effects model which includes area-specific random effects to take into account the between-area variation in addition to that explained by the auxiliary information . The between-area variation is measured by a normally distributed

random variable. The estimation of the impact of auxiliary information and the prediction of the area effect are classically made by the Empirical Best Linear Unbiased Prediction (EBLUP) (see Rao, 2003). But this model is very dependent on strong distributional assumptions such as the normality and homoscedasticity of the residuals. In a parametric frame and a model-based approach, Chambers and Tzavidis (2006) propose an estimator of the cdf based on conditional Huber  $M$ -quantiles as follows. The position of the domain is summarized by a mean  $M$ -quantile order and the interest variable for a unit of the domain outside the sample is predicted by the conditional Huber  $M$ -quantile of this mean order and an estimator of the cdf is derived. The outline of this paper is the following. In section 2, we recall the definition of the Huber  $M$ -quantiles as well as their main properties. In section 3 we propose in a first time to extend the Chambers and Tzavidis estimator of the interest variable to a nonparametric frame by using non parametric estimates of the conditional Huber  $M$ -quantiles. In a second time, we propose a new class of estimators also based on Huber  $M$ -quantiles nonparametric estimation. As a matter of fact, nonparametric estimation is more flexible than parametric estimation because we do not make any assumption about the relationship between the interest variable and the auxiliary information. We compare these estimators using simulated data in section 3. Finally in section 4, we apply our methods on a sample that contains measurements on 2802 physicians in the Midi-Pyrénées region of France during 1999 to estimate the cumulative distribution function of the drug prescribing activity in each department taking into account some individual characteristics of the physicians.

## 2 Huber $M$ -quantiles of a distribution

Let  $F$  denote the cumulative distribution function of a random variable  $Y$ . Consider the minimization problem

$$\min_{\theta} \int \rho_q(y - \theta) dF(y) \quad (1)$$

where  $\rho_q$  is a loss function and  $q$  is fixed,  $0 < q < 1$ . Differentiating the objective function in (1) with respect to  $\theta$  leads to the estimating equation

$$\int \psi_q(y - \theta) dF(y) = 0 \quad (2)$$

where  $\psi_q(u) = \delta\rho_q(u)/\delta u$  is called the influence function. When for some  $c > 0$  called the cutoff, we consider the influence function

$$\psi_{q,c}(y) = \begin{cases} -(1-q)c & \text{if } y < -c \\ (1-q)y & \text{if } -c \leq y < 0 \\ qy & \text{if } 0 \leq y \leq c \\ qc & \text{if } y \geq c \end{cases}$$

then the solution to (1) and (2) is called the Huber  $M$ -quantile of order  $q$  of the distribution of  $Y$ . Notice that if  $c$  tends to infinity,  $\theta$  is the  $q$ -expectile and moreover if  $q = 0.5$ ,  $\theta$  is

$q$	0.5	0.6	0.7	0.8	0.9	0.95
$c$	1.345	1.386	1.517	1.758	2.121	2.379

Table 1: Table of the cutoffs of the standard normal distribution

the expectation. When  $c$  tends to 0, then  $\theta$  is the  $q$ -quantile of the distribution of  $Y$ . In addition to it, it can be checked that  $M$ -quantiles are more robust than expectiles and more efficient than quantiles. An equivalent characterization of the Huber  $M$ -quantile is given by the following ratio of expectations:

$$q = \frac{\mathbb{E}(c\mathbb{I}(Y - \theta < -c) + |Y - \theta| \mathbb{I}(-c \leq Y - \theta \leq 0))}{\mathbb{E}(c\mathbb{I}(|Y - \theta| > c) + |Y - \theta| \mathbb{I}(|Y - \theta| \leq c))}. \quad (3)$$

The choice of an accurate cutoff has to be discussed in order to estimate Huber  $M$ -quantiles. It is straightforward that the optimal  $c$  depends on  $q$  and increases with  $q$  being far from 0.5. The value of  $c$  is chosen to ensure a given asymptotic variance at the standard normal distribution. It is well known that if  $c = 1.345$ , the variance of the Huber  $M$ -quantile of order 0.5 at the normal is only 5% larger than that of the empirical mean (see for example Maronna *et al.* (2004)). Similarly an optimal cutoff can be computed in order to obtain an asymptotic relative efficiency of 95% for the Huber  $M$ -quantile estimator with respect to the  $q$ -expectile estimator in the standard gaussian case. Table (1) gives the values of the cutoffs associated different values of  $q$ . In a regression context, the cutoff can be computed by first performing a local regression to the mean and then computing a robust estimation  $\hat{\sigma}$  of the standard deviation of the residuals of the regression. The cutoff is then equal to the product of the "gaussian" cutoff by  $\hat{\sigma}$ . In what follows Huber  $M$ -quantiles will be called  $M$ -quantiles.

### 3 Estimation of the cdf in a domain

In what follows we assume that the population is partitioned into  $m$  domains  $U_i$  of size  $N_i$ ,  $i = 1, \dots, m$ . Let  $s$  a sample of size  $n$  of the population and  $s_i = s \cap U_i$  be a sample of size  $n_i$  of the domain  $U_i$ .  $y_{ij}$  denotes the interest variable  $y$  for the  $j$ -th individual of the domain  $U_i$  and is only available for units in the sample. Auxiliary information at individual level is available at population level through a covariate  $x$ . More precisely,  $x_{ij}$  denotes the value of the covariate measured for the  $j$ th individual of the domain  $U_i$ . Since the sample is possibly small, "borrowing strength" from the other domains will improve the estimation. The cdf on the domain  $U_i$  can be written as:

$$F_i(t) = \frac{1}{N_i} \left( \sum_{j \in s_i} \mathbb{I}(y_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \mathbb{I}(y_{ij} \leq t) \right).$$

The first term of the sum only depends on the sample and the purpose is therefore the prediction of  $\mathbb{I}(y_{ij} \leq t)$  for  $j \in U_i \setminus s_i$  which requires the prediction of  $y_{ij}$ . This

prediction is based on conditional M-quantiles as follows. According to (3), for any point  $(y, x)$ , there exists a real  $q$  so that  $y$  is the  $q$ -M-quantile of the conditional cdf  $F(\cdot|x)$ .  $q$  is called the M-quantile order of the point  $(y, x)$ . Let us denote  $(y_k, x_k)$  the  $k$ -th individual of the sample  $s$ . Using the technique developed in Aragon *et al.* (2005) for the expectile-orders, we estimate non parametrically the M-quantile orders of each point  $(y_k, x_k)$  of the sample  $s$  by :

$$\hat{q}_c(y_k, x_k) = \frac{\sum_{l \in s} (c \mathbb{I}_{(y_l - y_k \leq -c)} + |y_l - y_k| \mathbb{I}_{(-c \leq y_l - y_k \leq 0)}) K\left(\frac{x_k - x_l}{h}\right)}{\sum_{l \in s} (c \mathbb{I}_{(|y_l - y_k| > c)} + |y_l - y_k| \mathbb{I}_{(|y_l - y_k| \leq c)}) K\left(\frac{x_k - x_l}{h}\right)}.$$

$K$  denotes a kernel and  $h$  an adequate bandwidth. The adequate cutoff  $c$  has to be determined. To this aim, we first estimate the quantile-order of each point of  $s$  by the Nadaraya-Watson estimator of the conditional cdf. :

$$\hat{q}^*(y_k, x_k) = \frac{\sum_{l \in s} \mathbb{I}(y_l \leq y_k) K\left(\frac{x_k - x_l}{h}\right)}{\sum_{l \in s} K\left(\frac{x_k - x_l}{h}\right)}.$$

We deduce a average quantile order of the sample  $\hat{q}^* = n^{-1} \sum_{k=1}^n \hat{q}^*(y_k, x_k)$ . The "gaussian" cutoff associated with  $\hat{q}^*$  is obtained from table (1) by interpolation. As we are in a regression context, we perform a local regression to the mean and compute a robust estimation  $\hat{\sigma}$  of the standard deviation of the residuals of the regression. The cutoff  $c$  is then equal to the product of the "gaussian" cutoff by  $\hat{\sigma}$ .

Notice that M-quantile orders are determined at population level. These orders define a conditional ordering of the individual's value relative to the values of other units of the population.

### 3.1 First class of estimators

We adapt the estimator of Chambers and Tzavidis (2006) to the nonparametric framework. The position of the domain  $U_i$  with respect to the population is summarized by a mean M-quantile order  $\hat{q}_{i,c} = \frac{1}{n_i} \sum_{l \in s_i} \hat{q}_c(y_{il}, x_{il})$ . For each domain  $U_i$ , a M-quantile nonparametric model of order  $\hat{q}_{i,c}$  is then fitted. And  $\hat{m}(\hat{q}_{i,c}, x)$  is the local constant estimator of the conditional M-quantile of order  $\hat{q}_{i,c}$  of  $Y$  conditionally to  $x$ . This estimator is the solution of the estimating equation :

$$\sum_{k \in s} \psi(y_k - \theta) K\left(\frac{x - x_k}{h}\right) = 0 \quad (4)$$

with  $\psi = \psi_{\hat{q}_{i,c}}$  and  $h$  is an adequate bandwidth. Notice that all the sample (and not only the sample of the domain) is used to perform the estimation on the domain. Following the technique developed by Chambers and Dustan (1986), the  $n_i$  residuals

$$\hat{\epsilon}_{il} = y_{il} - \hat{m}(\hat{q}_{i,c}, x_{il})$$

allow to build as many predictions  $m(x_{ij}, \hat{q}_{i,c}) + \hat{\epsilon}_{il}$  of  $Y$  with covariate  $x_{ij}$  known in  $U_i$ . The estimator of the cdf of  $Y$  in the domain  $U_i$  is defined by :

$$\hat{F}_i^{CT}(t) = \frac{1}{N_i} \left( \sum_{j \in s_i} \mathbb{I}(y_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} n_i^{-1} \sum_{l \in s_i} \mathbb{I}(\hat{m}(\hat{q}_{i,c}, x_{ij}) + \hat{\epsilon}_{il} \leq t) \right). \quad (5)$$

Kokic *et al.* (1997) proved that if  $F(\cdot|x)$  belongs to a location-scale family of distributions, than the distribution of conditional M-quantile orders does not depend on the covariate. But this distribution depends on  $x$  in the domain. Therefore, a kernel should be added in formula (5) to give more weight to observations whose covariate is close to  $x$ . The following estimator derives from this consideration.

$$\hat{F}_i^{CTK}(t) = \frac{1}{N_i} \left( \sum_{j \in s_i} \mathbb{I}(y_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \frac{\sum_{l \in s_i} K\left(\frac{x_{il} - x_{ij}}{h_1}\right) \mathbb{I}(\hat{m}(\hat{q}_{i,c}, x_{ij}) + \hat{\epsilon}_{il} \leq t)}{\sum_{l \in s_i} K\left(\frac{x_{il} - x_{ij}}{h_1}\right)} \right).$$

$h_1$  denotes an appropriate bandwidth.

### 3.2 Second class of estimators

Instead of summarizing the domain  $U_i$  by a mean M-quantile order, we propose to describe the domain by the set of estimated orders  $\{\hat{q}_{ik} \ k = 1, \dots, n_i\}$  in the sample  $s_i$ . We can think that this technique will give better results than the previous one when there is a large variation of the orders inside the domain. For each non sampled individual  $j$  with covariate  $x_{ij}$  of the domain, there are  $n_i$  possible predictions  $\hat{m}(\hat{q}_{il}, x_{ij})$  of  $y_{ij}$  ( $\hat{m}(\hat{q}_{il}, x)$  solution of the estimating equation (4) with  $q = \hat{q}_{il}$ .) Therefore,  $\mathbb{I}(y_{ij} \leq t)$  will be predicted by the average of the  $\mathbb{I}(\hat{m}(\hat{q}_{il}, x_{ij}) \leq t)$  on the sample  $s_i$ . This estimator is not derived from Chambers and Dunstan (1986) in the sense that we do not add any residual, considering that the deviation between the M-quantile order and 0.5 can be viewed as a residual. The obtained estimator denoted  $\hat{F}_i^C$  is:

$$\hat{F}_i^C(t) = \frac{1}{N_i} \left( \sum_{j \in s_i} \mathbb{I}(y_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} n_i^{-1} \sum_{l \in s_i} \mathbb{I}(\hat{m}(\hat{q}_{il}, x_{ij}) \leq t) \right) \quad (6)$$

If we consider the whole population, the distribution of the  $M$ -quantile orders calculated at the first step is independent on the value of the covariate. But the distribution of these

$M$ -quantile coefficients of a considered domain may depend on the covariate. Taking the previous remark into account, we will give a more important weight to the prediction  $\hat{m}(q_{il}, x_{ij})$  when the value of the covariate  $x_{il}$  is close to  $x_{ij}$ . This leads to the following estimator:

$$\hat{F}_i^{CK}(t) = \frac{1}{N_i} \left( \sum_{j \in s_i} \mathbb{I}(y_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \frac{\sum_{l \in s_i} K\left(\frac{x_{il} - x_{ij}}{h_1}\right) \mathbb{I}(\hat{m}(\hat{q}_{il}, x_{ij}) \leq t)}{\sum_{l \in s_i} K\left(\frac{x_{il} - x_{ij}}{h_1}\right)} \right)$$

Practically, for each order  $\hat{q}_{il}$ , the estimator  $\hat{m}(\hat{q}_{il}, \cdot)$  is evaluated on a grid of 50 points  $x$  regularly spaced in the range of the non sampled points of the domain. The value of the prediction for each non sampled point is then evaluated by interpolation or extrapolation.

## 4 Simulations

One generates a population  $U$  of size  $N = 1000 = N_1 + N_2 = 600 + 400$  following the model  $y_{1j} = 1 + 2(1 + 4x_{1j}) + x_{1j}u$  on the domain  $U_1$  and  $y_{2j} = 1 + 5(1 + 4x_{2j}) + 1.5u$  on the domain  $U_2$  where  $x$  is uniformly distributed on  $[0, 1]$  and  $u$  follows a standard normal distribution. For domain 1, we take a simple random sample without replacement of size  $n_1 = 30$  and for domain 2, we take a simple random sample without replacement of size  $n_2 = 20$ , leading to an overall sample size of  $n = 50$ . The simulation is performed with  $S = 500$  samples. For each simulation, the bandwidth  $h$  is equal to 30% of the range of the covariate in the sample  $s$  and the bandwidth  $h_1$  is equal to 30% of the range of the covariate in the sample  $s_i$ . Figure (1) shows the scatterplot of the data for the simulated domains. Figure (2) represents the estimated conditional  $M$ -quantile orders in the two domains. Let us notice that the  $M$ -quantiles orders depend on  $x$  in the domain  $U_1$  because of the heteroscedasticity. We also remark that the estimated orders vary more in domain  $U_1$  than in domain  $U_2$ .

For each domain  $U_i$ , the different estimators are compared to the Horvitz-Thompson estimator :

$$\hat{F}_i^{HT}(t) = \frac{1}{n_i} \sum_{j \in s_i} \mathbb{I}(y_{ij} \leq t)$$

For each domain  $U_i$ , the estimators of the cdf were computed for a set of  $M = 50$  regularly spaced values  $\{t_1, \dots, t_M\}$  of  $y$  in  $[t_{min,i}, t_{max,i}]$  where  $t_{min,i}$  (resp.  $t_{max,i}$ ) represents the minimum (resp. maximum) of  $y$  in the domain  $U_i$ . Table (2) gives the mean averaged squared error (MASE) of the different estimators in each domain. The MASE is defined by:

$$MASE(F_i^{est,s}) = \frac{1}{S} \frac{1}{M} \sum_{s=1}^S \sum_{j=1}^M \{ \hat{F}_i^{est,s}(t_j) - F_i(t_j) \}^2$$

for  $est$  in the set  $\{CT, C, CTK, CK, HT\}$ . The kernel estimators perform very well and bring an important gain with respect to the Horvitz-Thompson estimator especially in the

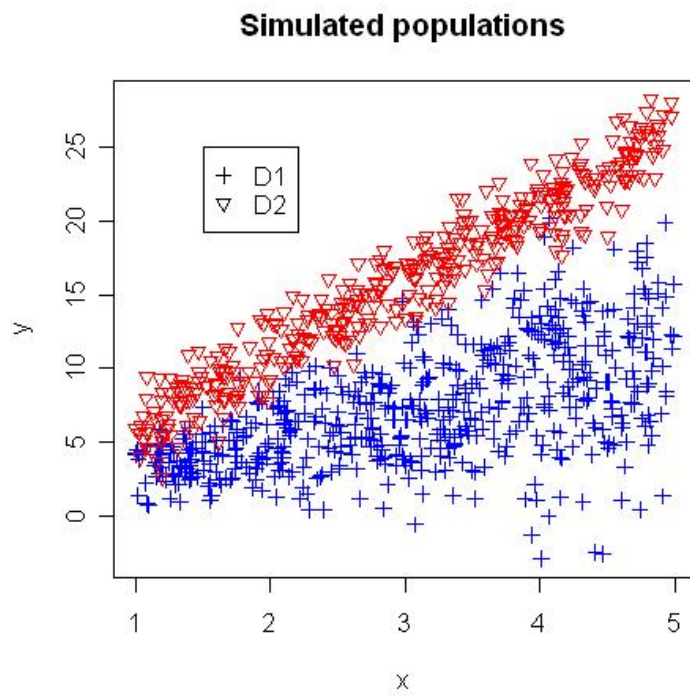


Figure 1: Scatterplot of the simulated data



### M-quantile estimated orders on the two domains

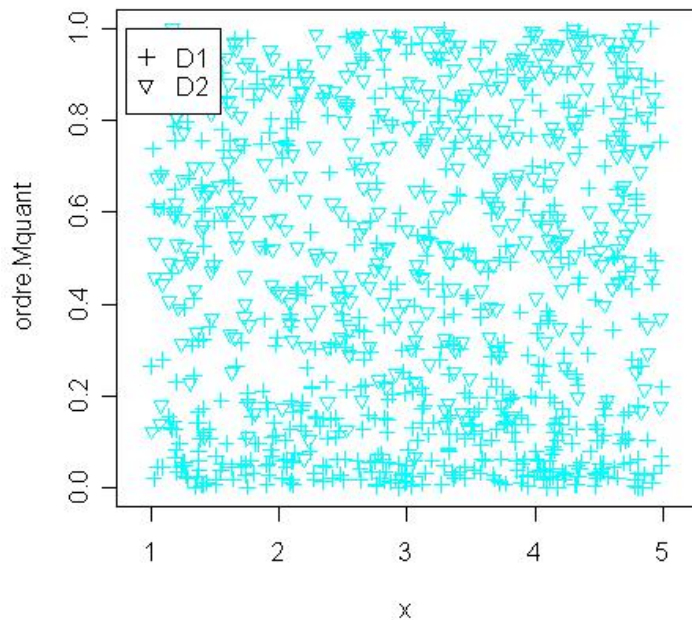


Figure 2: Scatterplot of the estimated conditional M-quantile orders

	HT	CT	C	CTK	CK
$U_1$	13.8393	14.8493	12.0546	9.3250	9.7695
$U_2$	31.5092	37.6701	29.0494	6.9951	7.8160

Table 2: table of the MASE of the 5 estimators for the 2 domains ( $\times 10^{-4}$ )

domain  $U_2$ .

## 5 Real data

We focus on a data set that contains measurements on 2802 physicians in the Midi-Pyrénées region of France during 1999, including most of the general practice physicians in this region. The study variable, denoted  $Y$ , measures the drug prescribing activity of a physician, and is defined as the logarithm of the ratio of the value of drug prescriptions issued by the physician over the year divided by the number of "acts" carried out by the physician over the same period. An act may be a house call or a consultation. 15 covariates are available : physician seniority (years), total practice size, % of practice less

	Ariège	Aveyron	Haute-Garonne	Gers	Lot	Hautes-Pyr.	Tarn	Tarn et Gar.
HT	48.511	19.737	3.8961	46.966	23.952	20.192	14.238	27.070
CT	40.169	16.642	3.4889	38.460	20.752	17.811	12.345	23.156
CTK	39.209	15.270	3.0551	35.881	18.297	15.558	11.209	21.476
C	36.716	15.252	3.1427	34.484	19.093	16.216	11.299	21.824
CK	36.823	15.173	2.8918	33.660	17.682	14.771	10.527	20.755

Table 3: table of the MASE of the 5 estimators for the 8 departments ( $\times 10^{-4}$ )

than 16, % of practice from 60 to 69, % of practice more than 70, % of practice who don't pay medical fees, % of practice who are farm employed, % of practice who are self employed, number of consultations and house calls, proportion of house calls, number of consultations per patient, number of house calls per patient, average fee per patient, age of physician, gender of physician. Since nonparametric regression can become unstable if there are too many covariates, we perform a slice inverse regression (SIR) to reduce the dimension of the covariate space by taking into account the dependence between the covariates and the response variable. SIR provides a few number of synthetic indices. We use the first SIR indice as the covariate in the nonparametric regression fit to the M-quantiles of the value of drug prescription per act. Our aim is to estimate the cdf of  $Y$  in the 8 departments which constitute the Midi-Pyrénées region. Figure (3) shows the scatterplot of the physicians by department. The population sizes are respectively 145, 261, 1195, 190, 171, 275, 361, 204 for Ariège (09), Aveyron (12), Haute-Garonne (31), Gers (32), Lot (46), Hautes-Pyrénées (65), Tarn (81) and Tarn and Garonne (82). Let us notice that the scatterplots are different according to the departments. In Gers, there are physicians for whom the SIR index is extreme (almost equal to -4). In Aveyron, Haute-Garonne, Lot and Tarn, some individuals have extreme values for the prescription and the SIR index, others have extreme values only for the prescription and some have both (in Haute-Garonne and Aveyron). In Hautes-Pyrénées, Ariège and Tarn and Garonne, some physicians have low prescriptions with respect to the others. For each department we take a simple random sample without replacement, leading to an overall sample size  $n = 280$  most of the time. Notice that we perform a design-based simulation where the original dataset is acting as a fixed population. Table (3) shows the MASE of the 9 estimators for the 8 departments. For each method, Haute-Garonne has a widely smaller MASE than the other departments because its sample size is much more important. The kernel methods are more efficient which indicates that the dependence of the response variable with respect to the covariate is different according to the domain. Finally, the Chambers-Tzavidis and the Casanova estimators behave similarly with a light advantage for the latter. Figure (4) shows the boxplots of the ratios  $ASE(EST)/ASE(HT)$  in Gers. For most of the samples, the estimators obtained by small area estimation techniques have a smaller averaged squared error than the corresponding Horvitz-Thompson estimators.

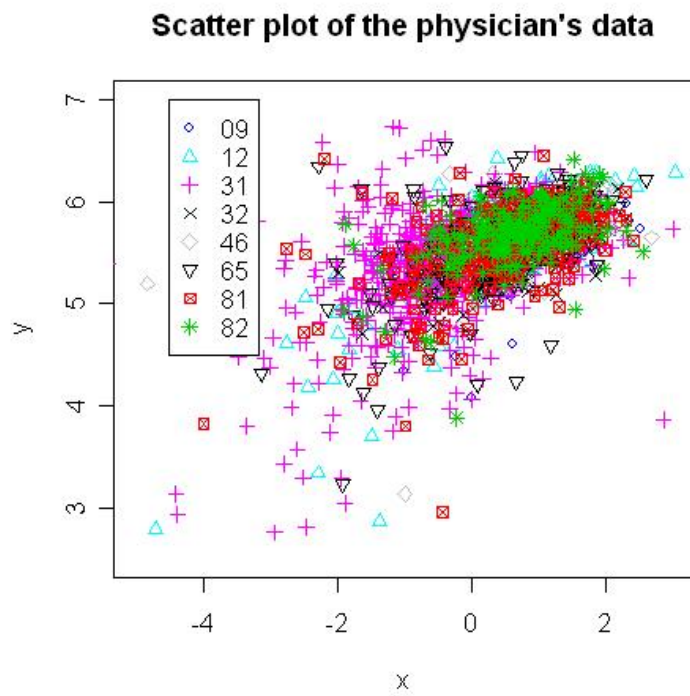


Figure 3: Scatterplots of the data in the 8 departments

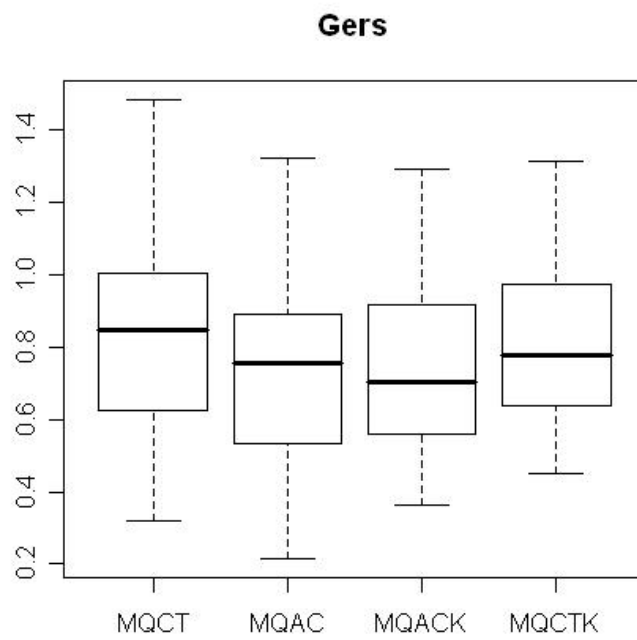


Figure 4: Boxplots of the ratios ASE(EST)/ASE(HT) in Gers

## 6 Discussion

The M-quantile approach provides an efficient alternative to the mixed effects models for small area estimation. Other M-quantiles like Tukey M-quantiles can replace Huber M-quantiles, the main difficulty being the choice of the cutoff. Moreover, the flexibility of nonparametric estimation allows to apply easily the proposed estimators. However, it can be difficult to estimate extreme M-quantiles nonparametrically and a parametric fit can be more appropriate.

## 7 Acknowledgment

The author gratefully thanks Yves Aragon, retired Professor of Statistics of the University of Social Sciences of Toulouse who is at the origin of the paper and actively participated to the applied part of the article.

## 8 References

- Aragon, Y., Casanova, S., Chambers, R. and Leconte, E. (2005). Conditional ordering using nonparametric expectiles, *Journal of Official Statistics*, 21, 617-633.
- Chambers, R.L. and Dunstan R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chambers, R. and Tzavidis, N. (2006). M-quantiles models for small area estimation, *Biometrika*, 93, 255-268.
- Johnson, A.A, Breidt, F.J and J.D Opsomer (2008). Estimating distribution functions from survey data using non parametric regression. *Journal of Statistical Theory and Practice*, 2, 419-431.
- Maronna, R. A., Douglas Martin, R. and Yohai, V. J. (2004). *Robust Statistics. Theory and methods*. Wiley, New-York.
- Rao, J.N.K. (2003). *Small area estimation*. Wiley, New-York.