Robust Small Area Prediction for Counts

Nikos Tzavidis[*] M. Giovanna Ranalli[†] Nicola Salvati[‡]
Emanuela Dreassi[§] Ray Chambers[¶]

Abstract

A new semiparametric approach to model-based small area prediction for counts is proposed and used for estimating the average number of visits to physicians for Health Districts in Central Italy. The proposed small area predictor is based on defining an M-quantile model for count data by extending the ideas in **?** and **?**. This predictor can be viewed as an outlier robust alternative to the more commonly used Empirical Plug-in Predictor that is based on a Poisson generalised linear mixed model with Gaussian random effects. Results from the real data application and from a simulation experiment confirm that the proposed small area predictor has good robustness properties and in some cases can be more efficient than alternative small area approaches.

Keywords: bootstrap; generalized linear models; health survey; M-quantile regression; non-normal outcomes; robust inference.

1 Introduction

The Health Conditions and Appeal to Medicare Survey (HCAMS) is a national, multistage sample survey conducted periodically in Italy by the National Institute of Statistics. The 2012-13 survey is currently running with previous surveys having been conducted in 1999-2000 and in 2004-05. The survey provides information about the health condition and health

[*] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, UK, n.tzavidis@soton.ac.uk

[†] Dipartimento di Economia, Finanza e Statistica, Universtà degli Studi di Perugia, Via Pascoli - I 06123 Perugia, Italy giovanna@stat.unipg.it

[‡] Dipartimento di Economia e Management, Università di Pisa, Via Ridolfi, 10 - I 56124 Pisa, Italy, salvati@ec.unipi.it

[§] Dipartimento di Statistica, Informatica, Applicazioni (DiSIA), Università di Firenze, Viale Morgagni, 59 - I 50134 Firenze, Italy, dreassi@disia.unifi.it

[¶] National Institute for Applied Statistics Research Australia, University of Wollongong, New South Wales 2522, Australia, ray@uow.edu.au

care use of the non-institutionalized population of Italy. The questionnaire comprises of items on basic health condition of individuals such as perceived health status and dietary habits that are also surveyed annually by the Multipurpose Everyday Life Survey. In addition, the survey covers specialized health topics on chronic and acute diseases and number of visits to physicians and general practitioners.

The HCAMS is a multistage survey in which municipalities are the primary sampling units (PSUs) and households are the secondary sampling units (SSUs). The 1999-2000 survey has about 1,449 PSUs (out of 8,102) and 52,332 households with approximately 120,000 individuals. Although the HCAMS is designed to provide reliable direct estimates at the level of Administrative Region (NUTS2), there is also a need for estimates at further levels of geographic disaggregation. This is true in general for National Surveys, but particularly relevant for surveys that collect health related information since in Italy health is managed mainly locally at the level of NUTS2. In particular, policies are endorsed by Administrative Regions by allocating resources and funds to Health Districts (HDs) that are in charge for local implementation. HDs are defined by groups of contiguous municipalities and are not planned domains in the HCAMS. A fairly large number of HDs have very small sample sizes and, as a result, direct estimation using only the survey data is inappropriate as it yields estimates with unacceptable levels of precision. The HDs represent, therefore, the *small areas* of interest in this paper.

In particular, in this paper we are interested in producing estimates of the mean number of visits to physicians within the past four weeks among people aged 65 or more for 60 HDs in three Administrative Regions in Italy: Liguria, Toscana and Umbria. These are neighbouring Regions located in the central part of Italy that have policies in place for assuring the quality of health services for the elderly. Ageing of the population is a great concern in Italy given that this country has the largest proportion of people aged 65 or more in Europe (20.3% in 2011 according to the latest available figure). Liguria, Toscana and Umbria are three of the regions with the highest proportion of elderly people in Italy with proportions of 26.7%, 23.3% and 23.1%, respectively.

The increasing demand for reliable estimates of various parameters at small area level has led to the development of a number of efficient model-based small area estimation (SAE) methods (see **?**, for a review of such methods). For example, the empirical best linear unbiased predictor (EBLUP) based on a linear mixed model (LMM) is often recommended when the target of inference is the small area average of a continuous response variable (**?**). Robust SAE inference under the LMM has recently attracted some interest (**??**). An alternative approach to SAE that automatically allows for robust inference is to use M-quantile models (**?**) to characterise between area heterogeneity (**?**).

Most of the variables in the HCAMS are binary or take the form of a count and are therefore not suited to standard SAE methods based on LMMs. Working within a frequentist paradigm, one can follow **?** who propose an empirical best predictor (EBP) for a binary response, or **?** who extends these results to generalized linear mixed models (GLMMs). Nevertheless, use of EBP can be computationally challenging (**?**). Despite their attractive properties as far as modelling non-normal outcomes is concerned, fitting GLMMs requires numerical approximations. In particular, the likelihood function defined by a GLMM can involve high-dimensional integrals which cannot be evaluated analytically (see **???**). In such cases numerical approximations can be used, as for example in the `R` function `glmer` in the package `lme4`. Alternatively, estimation of the model parameters can be obtained by using an iterative procedure that combines Maximum Penalized Quasi-Likelihood (MPQL) and REML estimation (**?**). Furthermore, estimates of GLMM parameters can be very sensitive to outliers or departures from underlying distributional assumptions. Large deviations from the expected response as well as outlying points in the space of the explanatory variables are known to have a large influence on classical maximum likelihood inference based on generalized linear models (GLMs).

Following a Bayesian paradigm, **?** also consider the estimation of parameters related to the number of visits to physicians using the American National Health Interview Survey. They focus on the proportion of the population with at least one visit in the past twelve months and use a Hierarchical Bayesian model in which a logistic model relates the individual's probability of a doctor visit to his/her characteristics and, then, small area parameters are modelled with respect to area specific covariates. **?** describes a Hierarchical Bayes approach to fitting a GLMM based on an outlier-robust normal mixture prior for the random effects and uses this model for SAE. **?** proposes robust estimation of the fixed effects and the variance components of a GLMM, using a Metropolis algorithm to approximate the posterior distribution of the random effects.

In this paper we present a new approach to SAE for counts based on M-quantile modelling. The proposed approach does not depend on strong distributional assumptions nor on a predefined hierarchical structure, and outlier robust inference is automatically allowed for. Following **?** and **?** we extend the existing M-quantile approach for continuous data to the case where the response is a count. As with M-quantile modelling of a continuous response (**?**) random effects are avoided and between area variation in the response is characterised by variation in area-specific values of quantile-like coefficients. In Section 2, we define the notation and briefly review SAE using GLMMs. In Section 3 we motivate the use of M-quantile regression for estimating the mean number of visits to physicians with some exploratory analysis of the HCMAS data. In particular, model diagnostics indicate departures from the model assumptions and the use of robust estimation methods in this case may prove beneficial for small area prediction. In Section 4, after reviewing M-quantile SAE for a continuous response, we show how the approach for robust inference for GLMs proposed by **?** can be extended for fitting an M-quantile GLM. Approaches for defining the M-quantile coefficients, which play the role of pseudo-random effects in this framework, are discussed in Section 5 alongside the definition of small area predictors and corresponding Mean Squared Error (MSE) estimators. In Section 6 we report the results from the application of the proposed methodology for deriving estimates of the number of visits in primary health care outlets for HDs in Italy. Results from a model-based simulation study aimed at empirically assessing the performance of the proposed small area predictors are presented in Section 7. Section 8 concludes the paper with some final remarks and proposals for further work.

2 Small area prediction using GLMMs

Let us now briefly review small area prediction using a GLMM. Let U denote a finite population of size N which can be partitioned into D domains or small areas, with U_d denoting population on small area d, $d = 1, ..., D$. The small area population sizes N_d, for $d = 1, ..., D$ are assumed known. Let y_{dj} be the value of the outcome of interest, for the purposes of this paper a discrete or a categorical variable, for unit j in area d, and let $\mathbf{x}_{dj}$ denote a $p \times 1$

vector of unit level covariates (including an intercept). It is assumed that the values of $\mathbf{x}_{dj}$ are known for all units in the population, as are the values $\mathbf{z}_d$ of a $q \times 1$ vector of area level covariates. We will see that the first requirement can be relaxed to some extent when there are no continuous variables among the $\boldsymbol{x}$'s. In the presence of categorical covariates, an equivalent alternative representation of the assumed data structure is in the form of a cross-tabulation. The aim is to use the sample values of y_{dj} and the population values of $\mathbf{x}_{dj}$ and $\mathbf{z}_d$ to estimate a proportion or a count of a characteristic in the small area $d = 1, ..., D$.

For discrete outcomes, model-based SAE conventionally employs a GLMM for $\mu_{dj} = E[y_{dj}|\mathbf{u}_d]$ of the form

$$g(\mu_{dj}) = \eta_{dj} = \mathbf{x}_{dj}^T\boldsymbol{\beta} + \mathbf{z}_d^T\mathbf{u}_d, \tag{1}$$

where g is a link function. When y_{dj} is a count outcome the logarithmic link function is commonly used and the individual y_{dj} values in area d are assumed to be independent Poisson random variables with

$$\mu_{dj} = E[y_{dj}|\mathbf{u}_d] = \exp\{\eta_{dj}\} \tag{2}$$

and Var$[y_{dj}|\mathbf{u}_d] = \mu_{dj}$. The q-dimensional vector $\mathbf{u}_d$ is generally assumed to be independently distributed between areas according to a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_u$. $\boldsymbol{\Sigma}_u$ depends on parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$, which are referred to as the variance components and $\boldsymbol{\beta}$ in (1) is the vector of fixed effects. If the target of inference is the small area d mean, $\bar{y}_d = N_d^{-1}\sum_{j\in U_d} y_{dj}$ and the Poisson-GLMM (1) is assumed, the approximation to the minimum mean squared error predictor of $\bar{y}_d$ is $N_d^{-1}[\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \mu_{dj}]$. Since μ_{dj} depends on $\boldsymbol{\beta}$ and $\mathbf{u}_d$, a further stage of approximation is required, where unknown parameters are replaced by suitable estimates. This leads to the Empirical Plug-in Predictor (EPP) for the area d proportion $\bar{y}_d$ under (2),

$$\hat{\bar{y}}_d^{\text{EPP}} = N_d^{-1}\Big\{\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \hat{\mu}_{dj}\Big\}, \tag{3}$$

where $\hat{\mu}_{dj} = \exp\{\hat{\eta}_{dj}\}$, $\hat{\eta}_{dj} = \mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_d^T\hat{\mathbf{u}}_d$, $\hat{\boldsymbol{\beta}}$ is the vector of the estimated fixed effects and $\hat{\mathbf{u}}_d$ denotes the vector of the predicted area-specific random effects (see **????**). In (3) s_d and r_d denote the set of sampled (of size n_d) and non-sampled (of size $N_d - n_d$) units in small area d, respectively.

3 The HCAMS data and challenges

In this section we describe the data available for performing SAE and also present diagnostics for the Poisson GLMM. These diagnostics allow us to motivate the use of the alternative semiparametric approach we propose in this paper.

The data we consider in this paper are coming from the 1999-2000 HCAMS. We are interested in producing estimates of the average number of visits to physicians within the past four weeks for the elderly (aged 65 and above) in the 60 HDs of Toscana, Liguria and Umbria. The total sample size for the three Regions is $n =$ 4,021. Figure 1 presents a map of the three regions of interest; HDs are color coded according to the sample size. Note that 5 HDs in Toscana and 1 in Umbria are out of sample areas, i.e. they have zero sample size.

The application of small area methodologies requires the estimation of a working model using the survey data. At the second stage the estimated model parameters are combined with census/administrative data. For this reason among the variables available from the survey, we focus on those that are also available at the population level. One possible source of these data is the Population Census run in Italy in 2001. The alternative is given by administrative registers held at a municipality level and updated annually. Then, from all potential covariates available in the survey data we select the following: 5-year age groups (65-69, 70-74, 75-79, 80-84, 85 and above), gender, marital status and region. An important property of the Poisson model is that it allows for the analysis of individual or grouped data (using an offset term) with equivalent results due to the fact that the sum of independent Poisson random variables is also Poisson. This is useful when we have groups of individuals with identical covariate values as it is the case with the present data.

Table 1 presents the results of the analysis of deviance from fitting a Poisson GLMM with Normal random effects to the sample data with random intercepts specified at the level of HD i.e. the target small areas. We note that age class and gender are significant. Region appears to be non-significant, however, since here we are interested in prediction and regions are closely related to HDs, we decided to leave the regional effect in the model. On the other hand, marital status and the 2-way interaction terms are not significant. A Likelihood Ratio Test (LRT) for the significance of the variance component has been also conducted. The value of the test statistic is 33.695, with a p-value $3.22\text{e}{-}09$, which provides evidence of significant between HD variability. Given that here we are testing whether the variance component is zero, i.e. a value on the boundary of its parameter space, the p-value has been determined using a 50:50 mixture between a χ^2_0 and a χ^2_1 distribution.

Figure 2 presents two plots of Pearson residuals from the Poisson GLMM (age class, gender and region). The histogram clearly shows that the distribution of the residuals is positively skewed and has some fairly large values. This is confirmed by the second plot, representing the distribution of the residuals by HD: some HDs contain a number positive residuals.

The skewed distribution of the residuals (Figure 2) indicates that the problem of overdispersion may arise here. Overdispersion is a common phenomenon in Poisson modelling, and the Negative Binomial model is frequently used to account for overdispersion; see **?**, **?** and **?**. We have tested for overdispersion comparing Poisson versus Negative Binomial GLM (age class, gender and region) using the LRT and using the P_B test by **?** on Poisson model. The results indicate significant overdispersion: the LRT is 834.07 (p-value $<$ 2.2e-16) and the Dean's test P_B is 43.42 (p-value $<$ 2.2e-16). According to the literature, overdispersion could arise from misspecification of the model, i.e. unobserved covariates. When we introduce in the model (age class, gender and region as fixed effects) and a set of area random effects for the Health Districts (HDs), some overdispersion persists. In fact, the ratio between the sum of squared Pearson residuals (7700.401) and the residual degrees of freedom (4021-7= 4014) is greater than one, which suggests the presence of overdispersion (see **?**).

Finally, Figure 3 plots the raw residuals against the fitted values for the number of visits to physicians. The x-axis ranges between 0 and 20 in order to show clearly over 98% of the observations. In the x-range between 8 and 20, there is no obvious pattern. However, for x between 0 and 8 we see a pattern, which suggests higher variability and the presence of a larger number of negative residuals when the predicted number of visits is small. These diagnostics, showing the presence of potential model misspecification, suggest that the use of an alternative to the Poisson GLMM may be justified in this case.

Before concluding this section we refer to the availability of population auxiliary information from a Census or an administrative source, which is crucial for SAE. The information needed for performing SAE in this case is given by the population sizes for groups defined by the age, gender, region, HD cross-classification. The population sizes can be combined with the estimated model parameters to produce estimates of the target parameter for each small area (HD). As mentioned above, one possible source of this data is the Population Census in Italy in 2001, which is the one closest to the time of the survey we use in this paper. An alternative source of population level auxiliary information is offered by administrative registers in municipalities, which can be used for updating the small area estimates in the intercensal period. In Section 6 the survey and the 2001 Census information are used for producing small area estimates of the average number of visits to physicians in HDs.

4 M-quantile regression

In this Section we present an extension of linear M-quantile regression to count data following **?** and **?**. We start by providing a fairly detailed presentation of M-quantile regression for continuous outcomes before focusing on the case of count outcomes. In this Section we drop subscript d for ease of notation.

4.1 M-quantile regression for a continuous response

The classic regression model summarises the behaviour of the mean of a random variable y at each point in a set of covariates x. This provides a rather incomplete picture, in much the same way as the mean gives an incomplete picture of a distribution. Quantile regression summarises the behaviour of different parts (e.g. quantiles) of the conditional distribution of y at each point in the set of the x's. In the linear case, quantile regression leads to a family of hyper-planes indexed by a real number $q \in (0, 1)$. For a given value of q, the corresponding model shows how the q-th quantile of the conditional distribution of y varies with x. For example, if $q = 0.5$ the quantile regression hyperplane shows how the median of the conditional distribution changes with x. Similarly, for $q = 0.1$ the quantile regression hyperplane separates the lower 10% of the conditional distribution from the remaining 90%.

Suppose $(\mathbf{x}_j^T, y_j),\ j = 1, \dots, n$ denotes the values observed for a random sample consisting of n independent observations from a population, where $\mathbf{x}_j^T$ are row p-vectors of a known design matrix $\mathbf{X}$ and y_j is a scalar response variable corresponding to a realisation of a continuous random variable with unknown continuous cumulative distribution function F. A linear regression model for the q-th conditional quantile of y_j given $\mathbf{x}_j$ is

$$Q_y(q|\mathbf{x}_j) = \mathbf{x}_j^T \boldsymbol{\beta}_q. \tag{4}$$

An estimate of the q-th regression parameter $\boldsymbol{\beta}_q$ is obtained by minimizing

$$\sum_{j=1}^{n} |y_j - \mathbf{x}_j^T \boldsymbol{\beta}_q| \{(1-q)I(y_j - \mathbf{x}_j^T \boldsymbol{\beta}_q \leq 0) + qI(y_j - \mathbf{x}_j^T \boldsymbol{\beta}_q > 0)\}.$$

Solutions to this problem are usually obtained by linear programming methods (**?**) and algorithms for fitting quantile regression are now available in standard statistical software, for example the library `quantreg` in `R` (**?**), the command `qreg` in `Stata`, and the procedure `quantreg` in `SAS`.

Quantile regression can be viewed as a generalization of median regression. In the same way, expectile regression (**?**) is a 'quantile-like' generalization of mean (i.e. standard) regression. M-quantile regression (**?**) integrates these concepts within a framework defined by a 'quantile-like' generalization of regression based on influence functions (M-regression). The M-quantile of order q for the conditional density of y given the set of covariates $\boldsymbol{x}$, $f(y|\boldsymbol{x})$, is defined as the solution $MQ_y(q|x;\psi)$ of the estimating equation $\int \psi_q\{y - MQ_y(q|x;\psi)\} f(y|\boldsymbol{x})dy = 0$, where ψ_q denotes an asymmetric influence function, which is the derivative of an asymmetric loss function ρ_q. A linear M-quantile regression model y_j given $\mathbf{x}_j$ is one where we assume that

$$MQ_y(q|\mathbf{x}_j;\psi) = {\mathbf{x}_j}^T \boldsymbol{\beta}_q. \tag{5}$$

That is, we allow a different set of p regression parameters for each value of $q \in (0,1)$. Estimates of $\boldsymbol{\beta}_q$ are obtained by minimizing

$$\sum_{j=1}^{n} \rho_q(y_j - {\mathbf{x}_j}^T \boldsymbol{\beta}_q). \tag{6}$$

Different regression models can be defined as special cases of (6). In particular, by varying the specifications of the asymmetric loss function ρ_q we obtain the expectile, M-quantile and quantile regression models as special cases. When ρ_q is the squared loss function we obtain the linear expectile regression model if $q \neq 0.5$ (**?**) and the standard linear regression model if $q = 0.5$. When ρ_q is the loss function described by (**?**) we obtain the linear quantile regression.

Setting the first derivative of (6) equal to zero leads to the following estimating equations

$$\sum_{j=1}^{n} \psi_q(r_{jq})\mathbf{x}_j = \mathbf{0}, \tag{7}$$

where $r_{jq} = y_j - \mathbf{x}_j^T\boldsymbol{\beta}_q$, $\psi_q(r_{jq}) = 2\psi(s^{-1}r_{jq})\{qI(r_{jq} > 0) + (1-q)I(r_{jq} \leq 0)\}$ and $s > 0$ is a suitable estimate of scale. For example, in the case of robust regression, $s = \text{median}|r_{jq}|/0.6745$, and we use the Huber Proposal 2 influence function, $\psi(u) = uI(-c \leq u \leq c) + c \cdot \text{sgn}(u)I(|u| > c)$. Provided that the tuning constant c is strictly greater than zero, estimates of $\boldsymbol{\beta}_q$ are obtained using iterative weighted least squares (IWLS).

4.2 M-quantile regression for count data: A Quasi-likelihood approach

The use of M-quantile regression with discrete outcomes is challenging as in this case there is no agreed definition of an M-quantile regression function (**??**). A popular approach for modelling the mean of a discrete outcome as a function of predictors is via the use of GLMs by assuming that the response variable follows a distribution that is a member of the exponential family of distributions using an appropriate link function.

In the same way that we impose in the linear specification (4) the continuous case, we impose an appropriate continuous (in q) specification on $MQ_y(q|\mathbf{X};\psi)$ for count data (**??**). The most obvious specification for count data is the log-linear specification. That is, we replace (5) by

$$MQ_y(q|\mathbf{x}_j;\psi) = t_j \exp(\mathbf{x}_j^T\boldsymbol{\beta}_q), \tag{8}$$

where t_j is an offset term. Alternative parametric specifications such as the use of a Negative Binomial model to capture excess dispersion in the data or a Zero Inflated Poisson model is work in progress. For estimating $\boldsymbol{\beta}_q$, following **??**, we consider extensions of the robust version of the estimating equations for GLMs by **?** to the M-quantile case. In particular, **?** propose a robust version of the estimating equations for GLMs and consider two popular

GLMs namely, the binomial and the Poisson models. Estimating equations are defined by

$$\Psi(\boldsymbol{\beta}) := n^{-1} \sum_{j=1}^{n} \Big\{ \psi(r_j) w(\mathbf{x}_j) \frac{1}{\sigma(\mu_j)} \mu_j' - a(\boldsymbol{\beta}) \Big\} = \mathbf{0}, \tag{9}$$

where $r_j = \sigma(\mu_j)^{-1}(y_j - \mu_j)$ are Pearson residuals, $E[Y_j] = \mu_j$, μ_i' is its derivative with respect to $\boldsymbol{\beta}$, $\text{Var}[Y_j] = \sigma^2(\mu_j)$, and $a(\boldsymbol{\beta}) = n^{-1} \sum_{j=1}^{n} E[\psi(r_j)] w(\mathbf{x}_j) \mu_j' / \sigma(\mu_j)$ ensures the Fisher consistency of the estimator. The bounded ψ function is introduced to control deviation in y-space, whereas weights $w(\cdot)$ are used to down-weight the leverage points. When $w(\mathbf{x}_j) = 1,\ j = 1, \ldots, n$ **?** call the estimator the Huber quasi-likelihood estimator. Notice that when ψ is the identity function we obtain the classic quasi-likelihood estimator for GLMs.

For M-quantile regression the estimating equations (9) can be re-written as

$$\Psi(\boldsymbol{\beta}_q) := \frac{1}{n} \sum_{j=1}^{n} \Big\{ \psi_q(r_{jq}) w(\mathbf{x}_j) \frac{1}{\sigma(MQ_y(q|\mathbf{x}_j;\psi))} MQ_y'(q|\mathbf{x}_j;\psi) - a(\boldsymbol{\beta}_q) \Big\} = \mathbf{0}, \tag{10}$$

where $r_{jq} = \sigma(MQ_y(q|\mathbf{x}_j;\psi))^{-1}(y_j - MQ_y(q|\mathbf{x}_j;\psi))$, $\sigma(MQ_y(q|\mathbf{x}_j;\psi)) = = MQ_y(q|\mathbf{x}_j;\psi)^{1/2}$, $MQ_y'(q|\mathbf{x}_j;\psi) = MQ_y(q|\mathbf{x}_j;\psi)\mathbf{x}_j^T$ and $a(\boldsymbol{\beta}_q)$ is a correction term to obtain unbiased estimators, which is defined following the arguments in **?**,

$$a(\boldsymbol{\beta}_q) = n^{-1} \sum_{j=1}^{n} 2 w_q(r_{jq}) w(\mathbf{x}_j) \Big\{ cP(Y_j \geqslant i_2 + 1) - cP(Y_j \leqslant i_1) +$$

$$\frac{MQ_y(q|\mathbf{x}_j;\psi)}{\sigma(MQ_y(q|\mathbf{x}_j;\psi))} [P(Y_j = i_1) - P(Y_j = i_2)] \Big\} MQ_y(q|\mathbf{x}_j;\psi)^{1/2} \mathbf{x}_j^T,$$

with

- $i_1 = \lfloor MQ_y(q|\mathbf{x}_j;\psi) - c\sigma(MQ_y(q|\mathbf{x}_j;\psi)) \rfloor$,
- $i_2 = \lfloor MQ_y(q|\mathbf{x}_j;\psi) + c\sigma(MQ_y(q|\mathbf{x}_j;\psi)) \rfloor$ and
- $w_q(r_{jq}) = [qI(r_{jq} > 0) + (1-q)I(r_{jq} \leqslant 0)]$.

When $w(\mathbf{x}_j) = 1, j = 1, \ldots, n$ a Huber quasi-likelihood estimator is again obtained. An alternative simple choice for $w(\mathbf{x}_j)$ suggested by robust estimation in linear models is $w(\mathbf{x}_j) = \sqrt{1 - h_j}$ where $h_j = \mathbf{x}_j^T (\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T)^{-1} \mathbf{x}_j$, i.e. the jth diagonal element of the hat matrix.

The solution to the estimating equations (10) can be obtained numerically by using a

Fisher scoring procedure. Note that (9) can be obtained as special case of (10) for specific choices of q. In particular, when $q = 0.5$ we obtain (9). Moreover, linear M-quantile regression is a special case of (10) if the the linear link function $MQ_y(q|\mathbf{x}_j; \psi) = \mathbf{x}_j^T \boldsymbol{\beta}_q$ is used and c tends to infinity. `R` routines for fitting M-quantile regression for count data are available from the authors.

4.3 Quantiles, M-quantiles and Expectiles: Linking the alternative estimation approaches

The problem with estimating conditional quantiles of counts is caused by the combination of a non-differentiable sample objective function with a discrete outcome. In this section we theoretically link the proposed approach to modelling M-quantiles of counts to alternative estimation approaches that they have been proposed in the literature.

An alternative approach to modelling conditional location parameters for counts was proposed by **?**. In particular, **?** proposed using asymmetric maximum likelihood (AML) estimation. Starting with the Poisson deviance, the AML estimate $\hat{\boldsymbol{\beta}}_w$ for $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_w = \arg\max_{\mathbf{b}} n^{-1} \sum_{j=1}^{n} [y_j \log(y_i/\mu_j(\mathbf{b})) - (y_j - \mu_j(\mathbf{b}))] w^{I\{y_j > \mu_j(\mathbf{b})\}}, \tag{11}$$

where $\mu_j(\mathbf{b}) = t_j \exp(\mathbf{x}_j^T \mathbf{b})$. From (11), by vector differentiation with respect to $\mathbf{b}$, the following estimating equation is obtained:

$$n^{-1} \sum_{j=1}^{n} \left[(y_j - \mu_j(\mathbf{b}))\mathbf{x}_j^T \right] w^{I\{y_j > \mu_j(\mathbf{b})\}} = \mathbf{0}. \tag{12}$$

Efron's approach results in estimates of conditional location parameters for counts that are similar to the conditional expectiles proposed by **?**. As **?** pointed out, asymmetric maximum likelihood estimation can be seen as the result of smoothing the objective function used to define the quantile regression estimator.

The approach we propose in this paper for estimating M-quantile regression also uses an objective function that has a degree of smoothness. In particular, the smoothness can be increased by setting the tuning constant in the influence function equal to a large value in which case estimates of the model parameters from our approach are those obtained by Efron's asymmetric maximum likelihood estimation for a specific choice of w. In particular, setting the tuning constant equal to a large value, (10) can be written as:

$$\Psi(\boldsymbol{\beta}_q) := n^{-1} \sum_{j=1}^{n} \left\{ (y_j - MQ_y(q|\mathbf{x}_j;\psi)) w_q(r_{jq}) \mathbf{x}_j^T \right\} = \mathbf{0}, \tag{13}$$

where $w_q(r_{jq})$ can be also written as $w_q(r_{jq}) = \left[\left(\frac{q}{1-q} \right) I\{y_j > MQ_y(q|\mathbf{x}_j;\psi)\} + I\{y_j \leqslant MQ_y(q|\mathbf{x}_j;\psi)\} \right]$. Setting $w = \frac{q}{(1-q)}$ in Efron's estimating equation (11) results in estimates that are equivalent to those obtained from our proposed estimating equation (13). At this point, a comment about the use of the Poisson deviance is needed. As Efron (1992) pointed out, the Poisson assumption enters the calculations only in that the fitting algorithm uses the Poisson deviance function. Efron's asymmetric maximum likelihood approach does not assume a specific form for overdispersion. Nevertheless, it gives reliable estimates of the conditional percentiles even in presence of overdispersion, when the Poisson assumption is incorrect.

A further alternative approach for estimating conditional quantiles for counts has been proposed in the literature by **?** and **?**. For overcoming the lack of smoothness, the authors propose the use of jittering. In particular, smoothness is achieved by adding to the count outcome noise generated, for example, from a Uniform$(0, 1)$. The quantiles of the resulting continuous outcome are then directly modelled by using the Asymmetric Laplace distribution. Modelling the conditional quantiles of counts in this way is possible because there is a one-to-one relation between the conditional quantiles of the count outcome and those of the jittered outcome. Quantiles have a more natural interpretation compared to M-quantiles and expectiles. However, since different types of location parameters are used here only for characterizing the variability in the data and subsequently for prediction, all three approaches can be used for developing the methodology presented in the next section. Nevertheless, the focus of this paper will be on M-quantiles.

5 Robust prediction for small area counts

5.1 The M-quantile small area population model and point estimation

Linear mixed effects models and GLMMs include random area effects to account for between-area variation. The M-quantile approach avoids parametric specification of the random effects, allowing between area differences to be characterised by the variation of area-specific M-quantile coefficients. To start with, the population model is specified at the unit level. Define q_{dj} such that $y_{dj} = MQ_y(q_{dj}|\mathbf{x}_{dj}; \psi)$. Under the log-linear specification, the population model is defined by

$$MQ_y(q_{dj}|\mathbf{x}_{dj}; \psi) = \exp(\mathbf{x}_{dj}^T \boldsymbol{\beta}_{q_{dj}}).$$

? used the term M-quantile coefficients for q_{dj}. The variability in q_{dj} reflects variability at the unit level. If a hierarchical structure does explain part of the variability in the population data, units within areas are expected to have similar M-quantile coefficients. An area-specific M-quantile coefficient is then defined as $\theta_d = E[q_{dj}|d]$.

Estimation of the parameters in linear mixed models and in GLMMs is implemented by means of parametric assumptions such as that the random effects are normally distributed. Efficient prediction of random effects is crucial due to their central role in SAE. Although for linear models closed form solutions exist, for GLMMs this is not the case. For GLMMs and from a frequentist perspective predicted random effects are obtained by using approximations to the likelihood, for example via first or second order penalised quasi-likelihood, or numerical methods such as Gaussian quadrature. Hence, for GLMMs outlier robust prediction of random effects becomes more challenging and the use of semi-parametric methods may offer a simpler solution to outlier robust estimation.

As discussed at the start of this section, a key concept in the application of M-quantile methods to data with group structure is the identification of a unique 'M-quantile coefficient' associated with each observed datum. These coefficients are then averaged, in some suitable way, over observations making up the group to define a group level M-quantile coefficient, which can be used to characterise the distribution of $y|\boldsymbol{x}$ within the group in very much the same way as a random group effect. In the continuous y case, the M-quantile coefficient for observation j is simply defined as the unique solution q_j to the equation $y_j = \widehat{MQ}_y(q_j|\mathbf{x}_j; \psi)$.

However, for count data the equation $y_j = \widehat{MQ}_y(q_j|\mathbf{x}_j;\psi)$ does not have a solution when $y_j = 0$. To overcome this problem we use the definition by **?**:

$$\widehat{MQ}_y(q_j|\mathbf{x}_j;\psi) = \begin{cases} k(\mathbf{x}_j) & y_j = 0 \\ y_j & y_j = 1, 2, \dots \end{cases}$$

A possibility is $k(\mathbf{x}_j) = \widehat{MQ}_y(q_{\min}|\mathbf{x}_j;\psi)$ where $q_{\min}$ denotes the smallest q-value in the grid of q-values used to determine the q_j values of the observed units. However, this implies that $q_j = q_{\min}$ whenever $y_j = 0$, irrespective of the value of $\mathbf{x}_j$, which does not appear to be appropriate. One way to tackle this is by following the same line of argument that **?** used in motivating the definition of q_j for the Bernoulli case. This implies that an observation with value $y_1 = 0$ corresponds to a smaller q-value than another with value $y_2 = 0$ when $\widehat{MQ}_{y_1}(0.5|\mathbf{x}_1;\psi) > \widehat{MQ}_{y_2}(0.5|\mathbf{x}_2;\psi)$. A way to define this is by setting $k(\mathbf{x}_j) = \min\{1-\epsilon, [\widehat{MQ}_y(0.5|\mathbf{x}_j;\psi)]^{-1}\}$, where $\epsilon > 0$ is a small positive constant. Then the M-quantile coefficient for unit j is q_j, where

$$\widehat{MQ}_y(q_j|\mathbf{x}_j;\psi) = \begin{cases} \min\Big\{1-\epsilon, \dfrac{1}{\exp(\mathbf{x}_j^T\hat{\boldsymbol{\beta}}_{0.5})}\Big\} & y_j = 0 \\ y_j & y_j = 1, 2, \dots \end{cases} \tag{14}$$

For a detailed discussion see **??**.

Provided there are sample observations in area d, an area d specific M-quantile coefficient, $\hat{\theta}_d$ can be defined as the average value of the sample M-quantile coefficients in area d, otherwise we set $\hat{\theta}_d = 0.5$. Following **?**, the M-quantile predictor of the average count $\bar{y}_d$ in small area d is then

$$\hat{\bar{y}}_d^{MQ} = N_d^{-1}\Big\{\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \widehat{MQ}_y(\hat{\theta}_d|\mathbf{x}_{dj};\psi)\Big\}, \tag{15}$$

where $\widehat{MQ}_y(\hat{\theta}_d|\mathbf{x}_{dj};\psi) = \exp\{\mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}}_{\hat{\theta}_d}\}$.

5.2 Mean squared error estimation

The Mean Squared Error of the predictor $\hat{\bar{y}}_d^{MQ}$ is defined as

$$\text{MSE}(\hat{\bar{y}}_d^{MQ}) = E[(\hat{\bar{y}}_d^{MQ} - \bar{y}_d)^2]. \tag{16}$$

Following **?** we propose bootstrap-based estimator of the MSE of the $\hat{\bar{y}}_d^{MQ}$. For developing the bootstrap procedure we express the linear predictor of the M-quantile regression model in a form that mimics the mixed effects model form,

$$y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta}_{0.5} + \mathbf{x}_{dj}^T(\boldsymbol{\beta}_{\theta_d} - \boldsymbol{\beta}_{0.5}). \tag{17}$$

Averaging the last term on the right-hand side of (17) for each small area results in a term u_d^{MQ} which can be interpreted as a pseudo-random effect for area d in that it quantifies an average difference of the area-specific M-quantile fit from the median fit.

The steps of the bootstrap procedure are summarized below:

- (Step 1) Using sample s, fit (8) and obtain predictors $\hat{\bar{y}}_d^{MQ}$. For each small area compute the pseudo-random effect $\hat{u}_d^{MQ}$ by computing the $E(\mathbf{x}_{dj}^T(\boldsymbol{\beta}_{\theta_d} - \boldsymbol{\beta}_{0.5}))$ for each area. It is convenient to re-scale the elements $\hat{\mathbf{u}}^{MQ}$ so that they have mean exactly equal to zero.
- (Step 2) Construct the vector $\hat{\mathbf{u}}^{MQ*} = \{\hat{u}_1^{MQ*}, \ldots, \hat{u}_D^{MQ*}\}^T$, whose elements are obtained by extracting a simple random sample with replacement of size D from the set $\{\hat{u}_1^{MQ}, \ldots, \hat{u}_D^{MQ}\}^T$.
- (Step 3) Generate a bootstrap population U^* of size $N = \sum_{d=1}^D N_d$, by generating values from a Poisson distribution with

$$\mu_{dj}^* = \exp\{\mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}_{0.5} + \hat{u}_d^{MQ*}\},\ j = 1, \ldots, N_d$$

  and calculate the bootstrap population parameters $\bar{y}_d^*$, $d = 1, \ldots, D$.
- (Step 4) Extract a sample s^* of size n from the bootstrap population U^* using the assumed sampling design and compute small area estimates with the bootstrap sample $\hat{\bar{y}}_d^{MQ*},\ d = 1, \ldots, D$.

- (Step 5) Repeat steps 2-4 B times.
- (Step 6) Denoting by $\hat{\bar{y}}_d^{MQ*(b)}$ the M-quantile predictor in the b-th bootstrap replication and by $\bar{y}_d^{*(b)}$ the corresponding population value in the b-th bootstrap population, a bootstrap estimator of MSE is

$$\text{MSE}(\hat{\bar{y}}_d^{MQ}) = B^{-1} \sum_{b=1}^{B} \left(\hat{\bar{y}}_d^{MQ*(b)} - \bar{y}_d^{*(b)} \right)^2. \quad (18)$$

The proposed bootstrap is not the only approach to MSE estimation. An alternative approach would have been to use the random effects block bootstrap (**?**), which is free both of the distribution and the dependence assumptions of the usual parametric bootstrap. **?** adapted the block bootstrap for estimating the MSE of the M-quantile small area predictor in the case of a Bernoulli outcome. A similar approach can be used in the case of a count outcome. A comparison between the alternative approaches to MSE estimation will be discussed in future work.

6 Application

In this section we present the results from the application of the SAE methods to the HCAMS data for estimating the average number of visits to physicians among the elderly (aged 65 and above) and corresponding MSE estimates for HDs in three Italian regions. The following small area predictors are being considered: (i) the direct estimator, which is defined as a ratio estimator using the calibration weights available in the survey; (ii) the EPP in (3) based on the Poisson GLMM with random intercepts specified at the level of HD and fixed effects including age class, gender and region (see the results in Table 1); (iii) the M-quantile predictor (15) based on the M-quantile model (see Section 4.2) with the same fixed effects as in the case of the Poisson GLMM. The Poisson GLMM and the M-quantile model are fitted using the aggregate data with appropriate offset terms.

For the GLMM using the aggregate data does not impact upon the results due to the equivalence between the the individual and the aggregate level analysis (with appropriate offset terms). In fact, an aggregate level analysis is useful when we have groups of individuals with identical covariate values as it is the case with the HCAMS data. For M-quantile regression and for the AML estimator by Efron (1992), the model cannot be estimated for values of q below the proportion of 0's in the sample. It is easy to see this since $I\{y_j \leqslant \exp(\mathbf{x}_j^T \hat{\boldsymbol{\beta}}_q)\}$ is necessarily equal to 1 when $y_j = 0$ (see **?**, for details). In the HCAMS data there are 2298 units with the value of visits equal to 0 (57%), so the M-quantile regression model cannot be fitted for values of $q \leqslant 0.57$. This can in turn create problems in the application of the small area M-quantile methodology and in particular in the estimation of the M-quantile coefficient associated with each datum and hence in the computation of the area-specific M-quantile coefficients, θ_d. In order to enable a comparison between the M-quantile and GLMM methodologies, we have decided to use aggregate data in particular, the data defined by the cross-classification of the sample according to 5-year age groups (65-69, 70-74, 75-79, 80-84, 85 and above), gender and HD.

For the M-quantile model the ψ function is set to be the Huber Proposal 2 with the tuning constant $c = 1.6$ (**?**). The estimates of the model parameters from Efron's (1992) method (using the `vgam` function with family equal to `amlpoisson` in `R`) and those obtained by M-quantile model, when setting Huber's tuning constant equal to a large value, are equivalent, which is a confirmation of the theoretical link described in Section 4.3.

Model selection is carried out via a robust stepwise procedure based on the Huber quasi-deviance at $q = 0.5$ (**?**). The analysis of deviance reported in Table 2 shows that the auxiliary variables age, gender and region, added sequentially, are highly significant on the basis of their deviance value. Interactions between pairs of these variables are again nonsignificant. Table 3 reports the estimated $\boldsymbol{\beta}_q$ coefficients at $q = 0.5$, alongside corresponding standard errors obtained by using the results in **?** and p-values. Moreover Table 3 reports the estimated $\boldsymbol{\beta}$ coefficients, their standard errors and p-values for the Poisson GLMM. The results confirm what we expected: Controlling for the effects of gender and region, the rate of visiting physicians increases as people grow older, particularly for those over 75 years old. Controlling for the effects of age and region, women visit physicians more often than men. This result can be explained by the the fact that women are more prone to some types of chronic diseases such as osteoporosis and varicose veins that can potentially require more intensive health care. Other things being equal, Umbria and Toscana show an overall higher rate of visits to physicians than Liguria.

Efficient estimates of area effects are necessary for SAE via GLMMs. Similarly, estimation of M-quantile coefficients is necessary for SAE using the M-quantile model proposed in this paper. Figure 4 shows how the standardized M-quantile coefficients estimated with (14) are related to the standardized area effects estimated using the `glmer` function in `R`. Figure 4 shows that the relationship between the estimated area effects and the estimated M-quantile coefficients is strong. The correlation between the estimated area effects and the estimated M-quantile coefficients is 0.91. This result suggests that M-quantile coefficients are comparable to estimated area effects obtained by using standard GLMM fitting procedures as far as capturing intra-area (domain) variability is concerned.

For comparing the performance of the different small area estimators we must use a set of diagnostics. Such diagnostics are suggested in **?**. Model-based estimates should be (i) 'close' to the direct estimates and (ii) more precise than direct estimates. The first diagnostic is based on the idea that if model-based estimates are 'close' to the small area value of interest, then unbiased direct estimates are considered like random variables whose expected value corresponds to the value of the model-based estimates. In other words, the model-based estimates should be unbiased predictors of direct estimates. To validate the reliability of the model-based small-area estimates, therefore, we use the goodness of fit (GoF) diagnostic and the values of the coefficient of variation (CV). The former inversely weights the squared difference between model-based and direct estimates by their variance and sums over all areas. This sum gives more weight to differences from more reliable direct estimates than from less reliable ones. The sum is tested against a χ^2 distribution to provide a significance test of bias of the model-based estimates relative to their precision (**?**).

Overall, the correlation between the model-based estimates and the direct estimates is high and positive, which indicates that the model-based estimates are close to the direct estimates (Direct/M-quantile correlation is 0.87 and Direct/EPP correlation is 0.95). This result for the M-quantile estimates is confirmed by Figure 5 where the direct estimates are plotted against the M-quantile estimates: we note that M-quantile estimates appear to be close to the direct estimates of the total number of visits to physicians.

The GoF diagnostic is based on the null hypothesis that the model-based estimates are equal to the expected values of the direct estimates; the statistic has a χ^2 distribution with degrees of freedom equal to the number of small areas. The GoF diagnostic is computed using the following Wald statistic for every model based estimator

$$W = \sum_{d} \Big\{ \frac{(\hat{y}_d^{\text{direct}} - \hat{y}_d^{\text{model}})^2}{[\widehat{\text{var}}(\hat{y}_d^{\text{direct}}) + \widehat{\text{mse}}(\hat{y}_d^{\text{model}})]} \Big\}.$$

The value from the test statistic W is compared against the value from a χ^2 distribution with $D = 54$ degrees of freedom. In our case, this value is 72.15 at 5% level of significance. For estimating the MSE of the M-quantile estimates, we use the bootstrap procedure outlined earlier in this paper and for the EPP estimates we use the bootstrap mean squared error estimator proposed by **?**. Variance estimates for the direct estimator have been computed by taking into account the complex two stage design employed for HCAMS. In particular, variance estimates for the estimate of the average number of visits to physicians has been computed separately for each of the three regions to provide a first estimate of the design effect. Then, following **?**, Section 2.6, the design effects have been recomputed to account for the fact that the elderly constitute a sub-domain that cuts across PSUs (the final deff values for each small area range between 0.9 and 1.5). Using the derived MSE estimates, the values of the GoF are 27.3 for the M-quantile predictor and 15.9 for EPP. These results indicate that all model-based estimates are not statistically different from the direct estimates but are potentially more efficient.

Figure 6 shows the distribution, across HDs, of the estimated CVs (expressed in percentage terms) of the direct (solid black line) and model-based estimates (blue denotes M-quantile estimates and red denotes EPP estimates). The estimated gains of the model-based predictors over the direct estimator are large, particularly for HDs with small sample sizes. Generally, the M-quantile estimates have a smaller estimated CV than the corresponding EPP estimates.

Moreover, in order to evaluate the precision of the M-quantile predictor, in Table 4 we report the number of HDs with values of CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for direct, EPP and MQ estimators for three groups of areas formed according to area-specific sample sizes. These values of CV are suggested by **?** to provide quality level guidelines for publishing tables: estimates with a coefficient of variation less than 16.6% are considered reliable for general use. Estimates with coefficient of variation between 16.6%

and 33.3% should be accompanied by a warning to users. Estimates with coefficients of variation larger than 33.3% are deemed to be unreliable. Table 4 shows that the M-quantile estimator provides reliable estimates. Only for two areas with sample size less than 24 the values of CV are between 16.6% and 33.3%. The MQ estimates are more reliable than those of the direct and the EPP when the sample sizes are small. For large sample sizes ($n_d > 100$) the M-quantile and EPP predictors appear to be equivalent.

In Figure 7 we compare the maps obtained for the average number of visits to physicians in HDs of Liguria, Toscana and Umbria in 2000 as estimated by the direct, the M-quantile and the EPP based estimators. Note that we have used the same cut-points to depict the three maps. Most HDs have similar levels of average number of visits to physicians, but there are some areas deviating from the bulk of the distribution in both directions.

As anticipated by the aforementioned analyses, the estimates in different maps are all comparable. However, those maps based on the two model-based predictors are more alike and show a smoother pattern as opposed to the map of direct estimates. In addition, model-based methods allow for estimates for those areas with zero sample sizes for which the direct estimator cannot be computed.

7 Simulation study

The purpose of this simulation experiment is to compare the performance of the M-quantile predictor against that of the EPP predictor and the direct estimator and to evaluate the performance of the bootstrap mean squared estimator (18) proposed in Subsection 5.2. The simulated data are generated by using the $\mathbf{x}_{dj}^T$ values, the estimated $\hat{\boldsymbol{\beta}} = (-0.44, 0.05, 0.28, 0.27, 0.29, -0.1$ and the estimated variance component $\hat{\varphi} = 0.192$ obtained by fitting the GLMM to the real data of Section 6. In each run of the simulation aggregated y values are generated for the groups given by the cross-classification of gender by age group for each of $D = 54$ small areas for which we have sample values. In total, we have ten groups for each small area. The value of the y variable for each cell y_{dk} ($d = 1, \ldots, D,\ k = 1, \ldots, 10$) is calculated as Poisson(μ_{dk}) with $\mu_{dk} = N_{dk} \exp\{\eta_{dk}\}$ and $\eta_{dk} = \mathbf{x}_{dk}^T \hat{\boldsymbol{\beta}} + u_d$, where u_d are independently drawn from a normal distribution with mean 0 and variance $\hat{\varphi}$. Here, $T = 1,000$ populations are generated and the true values of the average number of visits for each of the 54 sampled

HDs of Liguria, Toscana and Umbria, $\bar{y}_d = N_d^{-1} \sum_{k=1}^{10} y_{dk}$, $d = 1, \dots, D$, of the synthetic populations are then computed.

For each population, sample values y_{dk} for each cell are generated from a Poisson(μ^*_{dk}) with $\mu^*_{dk} = n_{dk} \exp\{\mathbf{x}_{dk}^T \hat{\boldsymbol{\beta}} + u_d\}$, where u_d is the value of random effect drawn previously to create the population, and according to two scenarios:

- (0) - No outliers.
- (M) - Measurement-type error: $2\%, 5\%, 10\%$ of randomly chosen response values has been changed from y_{dk} to $y_{dk} = y_{dk} + 10$.

For each sample, the M-quantile, the EPP and the direct estimator are used to estimate the small area average $\bar{y}_d,\ d = 1, \dots, D$. The performance of different small area estimators are evaluated with respect to two criteria: the bias and the root mean squared error (RMSE). Empirical values of the bias and of the mean squared error for a small area estimator are obtained as $T^{-1} \sum_{t=1}^{T} (\hat{\bar{y}}_{dt} - \bar{y}_{dt})$ and $T^{-1} \sum_{t=1}^{T} (\hat{\bar{y}}_{dt} - \bar{y}_{dt})^2$, respectively. Here $\bar{y}_{dt}$ denotes the actual area d value at simulation t and the predicted value is denoted by $\hat{\bar{y}}_{dt}$. The median and maximum value of the absolute Bias and the median value of RMSE over small areas are presented in Table 5. The results confirm our expectations regarding the behaviour of the estimators: under the (0) scenario the EPP performs better than the M-quantile in terms of RMSE, whereas there is no noticeable difference between the three estimators in terms of bias. The bias of the direct estimator can be explained by the approach we used to simulate the data. In particular, in this case not only the population is simulated at each Monte-Carlo run, but in addition the sample is generated under the Poisson assumption. This process may introduce some technical bias that affects the performance of the direct estimator. Furthermore, relatively larger values for the bias are displayed for those areas with a smaller sample size. This is expected because we are using a ratio type estimator whose bias can be non-negligible when the sample size is very small. The M-quantile predictor is the best in terms of bias and RMSE under the (M) scenarios and it is clearly superior to alternative estimators as contamination increases.

Regarding the second purpose of the simulation study, i.e. the evaluation of the performance of the bootstrap MSE estimator (18) proposed in Section 5.2, we use the data generated for scenarios (0) and (M)-10% and a subset of small areas: $D = 14$, the HDs of the

Region Liguria. The results of the MSE estimator, based on 500 bootstrap iterations, for each scenario are shown in Table 6 where we report the median values (over areas and simulations) of the bias and the root mean squared error, expressed in relative terms $(\%)$. The MSE estimator shows small bias and a good stability under both scenarios. In particular, under scenario (M)-10% MSE estimates tend to be biased up and the Relative RMSE increases but not considerably compared to the no-contamination scenario. Table 6 also shows that the proposed MSE estimator generates nominal 95 per cent confidence intervals with some under coverage.

8 Final remarks

An M-quantile model for count data is proposed and used for small area prediction. This presents a semiparametric approach to small area prediction that reduces the need for parametric assumptions and allows for outlier robust estimation. The results from the model-based simulation and the real data application indicate that the proposed method provides a reasonable and useful alternative to existing methods specifically when the assumptions of parametric models are not valid.

Despite the fact that the proposed methodology provides encouraging results, further research is necessary. To start with, the estimation of the area quantile coefficients is challenging and alternative approaches should be investigated. Developing analytic estimators of the MSE, although challenging, is something that can be potentially very useful as it will reduce the computational burden. Taking into account the presence of overdispersion in data by using alternative approaches for example, a Negative Binomial M-quantile model (see **?**) might offer some efficiency gains in SAE. Finally, developing design-consistent small area estimators under the M-quantile model is a topic of interest especially for those working in survey sampling from a design-based or a model-assisted perspective.

Table 1: Analysis of deviance table from fitting Poisson GLMMs to the HCAMS dataset.

Covariates	Resid. df	df	χ^2 value	p-value
Null	4019			
age	4015	4	59.590	1.5e-11
age, gender	4014	1	12.816	0.0003
age, gender, marital status	*4011*	*3*	*2.277*	*0.5170*
age, gender, region	4012	2	4.196	0.1227
age, gender, region, age × gender	*4008*	*4*	*1.165*	*0.8838*
age, gender, region, region × gender	*4010*	*2*	*0.746*	*0.6888*
age, gender, region, region × age	*4004*	*8*	*8.962*	*0.3455*

Table 2: Analysis of quasi-deviance table for the M-quantile model at $q = 0.5$.

Covariates	Resid. df	df	χ^2 value	p-value
Null	506			
age	502	4	52.473	1.0e-10
age, gender	501	1	10.375	0.0013
age, gender, region	499	2	10.991	0.0041

Table 3: Estimated M-quantile coefficients ($\boldsymbol{\beta}_q$) at $q = 0.5$ and Poisson GLMM coefficients and corresponding standard errors at . The baseline for age is group 65-69, for variable gender is female and for region is Liguria.

Covariates	$\boldsymbol{\beta}_q$			$\boldsymbol{\beta}$		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
Intercept	-0.411	0.047	1.2e-09	-0.444	0.073	1.2e-09
age 70-74	0.029	0.052	0.2854	0.055	0.050	0.2731
age 75-79	0.245	0.051	1.1e-06	0.274	0.049	2.4e-08
age 80-84	0.211	0.069	0.0011	0.267	0.064	3.1e-05
age >84	0.256	0.064	3.1e-05	0.293	0.060	1.4e-06
gender	-0.125	0.038	0.0005	-0.130	0.080	0.0003
region Toscana	0.110	0.044	0.0067	0.164	0.080	0.0407
region Umbria	0.149	0.046	0.0007	0.142	0.094	0.1327

Table 4: Number of HDs with values of CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for direct estimator, EPP and MQ predictor grouped by area sample sizes.

Estimator	CV	n_i			Total
		<24	24-100	101-556	
Direct	$< 16.6\%$	0	10	9	19
	$16.6\% - 33.3\%$	2	26	2	30
	$> 33.3\%$	3	2	0	5
EPP	$< 16.6\%$	2	34	11	47
	$16.6\% - 33.3\%$	3	4	0	7
	$> 33.3\%$	0	0	0	0
MQ	$< 16.6\%$	3	38	11	52
	$16.6\% - 33.3\%$	2	0	0	2
	$> 33.3\%$	0	0	0	0

Table 5: Model-based simulation results: performances of predictors of small area counts. Scenarios (0) and (M), Contamination: $2\%, 5\%, 10\%$, $D = 54$.

Predictor/Scenario	(0)	(M) 2%	(M) 5%	(M) 10%
	Median values of Absolute Bias			
EPP	0.0024	0.0360	0.0863	0.1814
M-quantile	0.0085	0.0095	0.0299	0.0699
Direct	0.0147	0.0318	0.0828	0.1785
	Maximum value of Absolute Bias			
EPP	0.0809	0.1050	0.1826	0.3837
M-quantile	0.0112	0.0775	0.1356	0.1974
Direct	0.0970	0.1431	0.2891	0.5553
	Median values of RMSE			
EPP	0.0973	0.1217	0.1624	0.2548
M-quantile	0.1072	0.1088	0.1163	0.1410
Direct	0.1272	0.1562	0.1966	0.2790

Table 6: Performance of the bootstrap MSE estimator (18). Scenarios (0) and (M)-10%, $D = 14$.

Indicator/Scenario	(0)	(M) 10%
Relative bias	-3.05	1.61
Relative RMSE	27.44	31.09
Coverage rate (95% nominal)	90%	86%

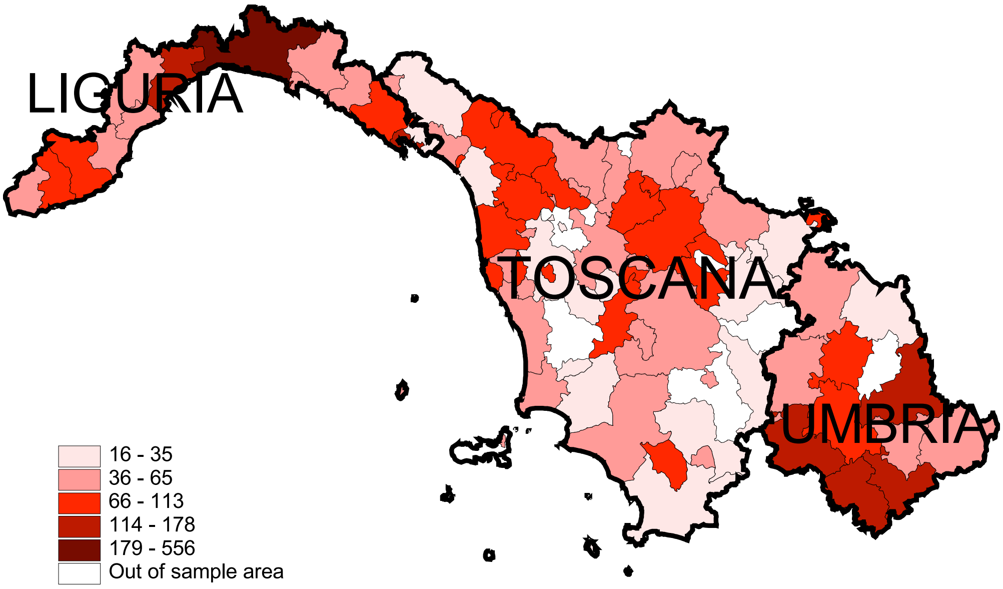


Figure 1: Sample sizes in Health Districts of Liguria, Toscana and Umbria in 2000.

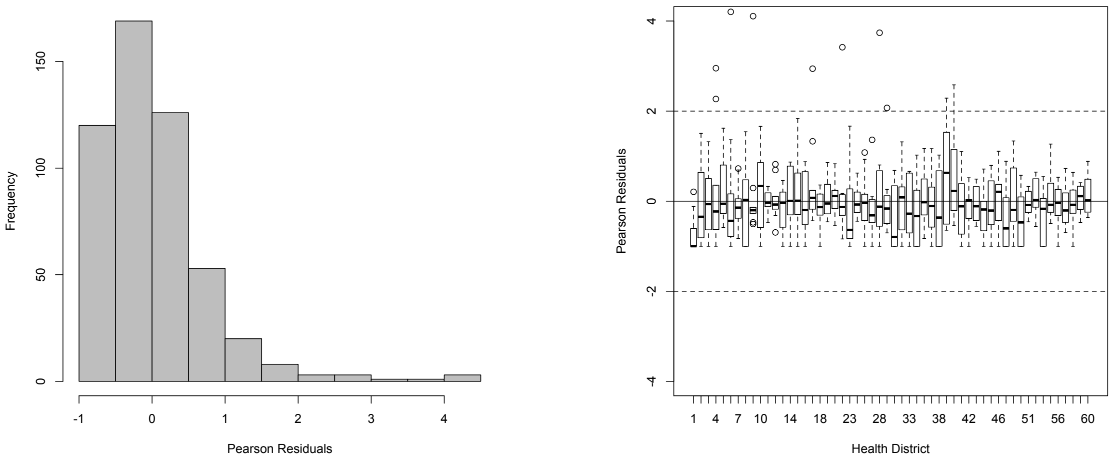


Figure 2: Model fit diagnostics for the Poisson GLMM: histogram of Pearson residuals (left) and box-plots of Pearson residuals by Health District (right).

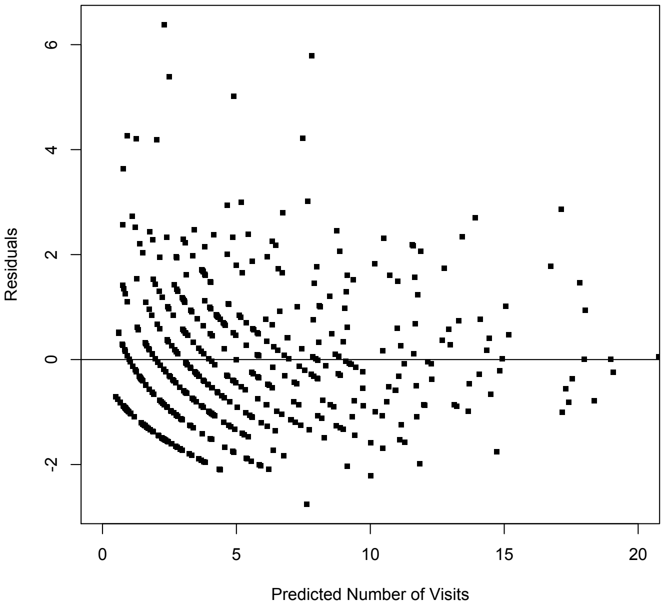


Figure 3: Model fit diagnostics for the Poisson GLMM: raw residuals Vs. predicted values.

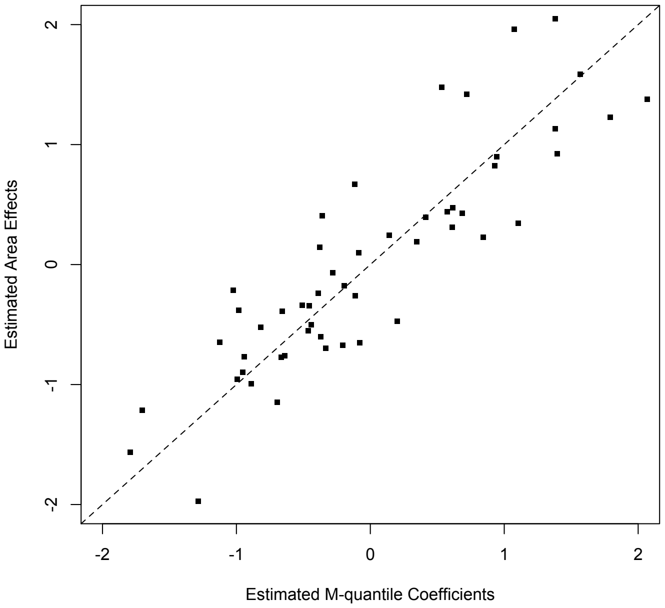


Figure 4: Estimated M-quantile coefficients vs. predicted random area effects (standardized values).

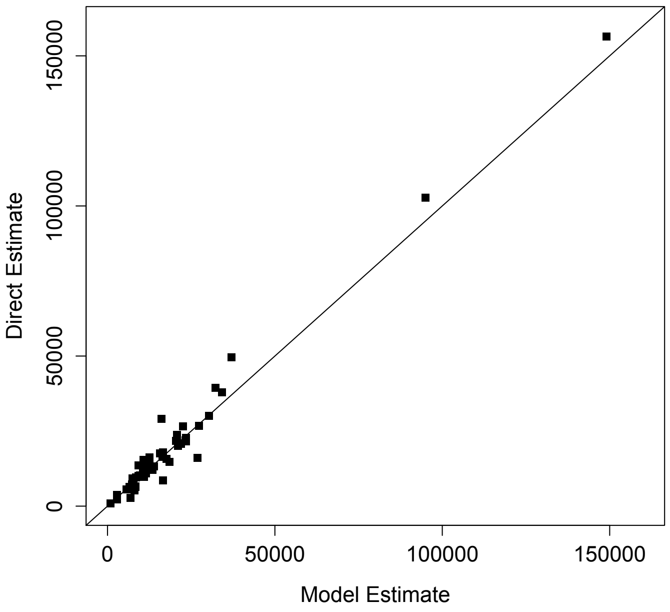


Figure 5: Total number of visits to physicians in Health Districts of Liguria, Toscana and Umbria in 2000: Model-based M-quantile estimates versus corresponding direct estimates.

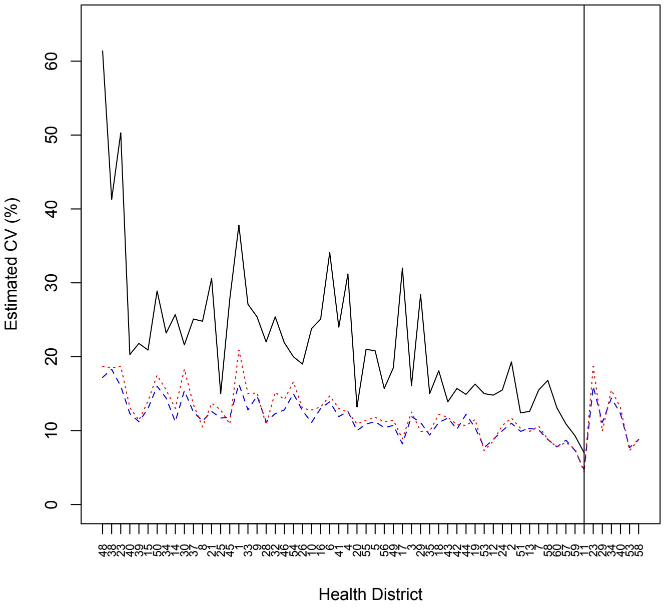
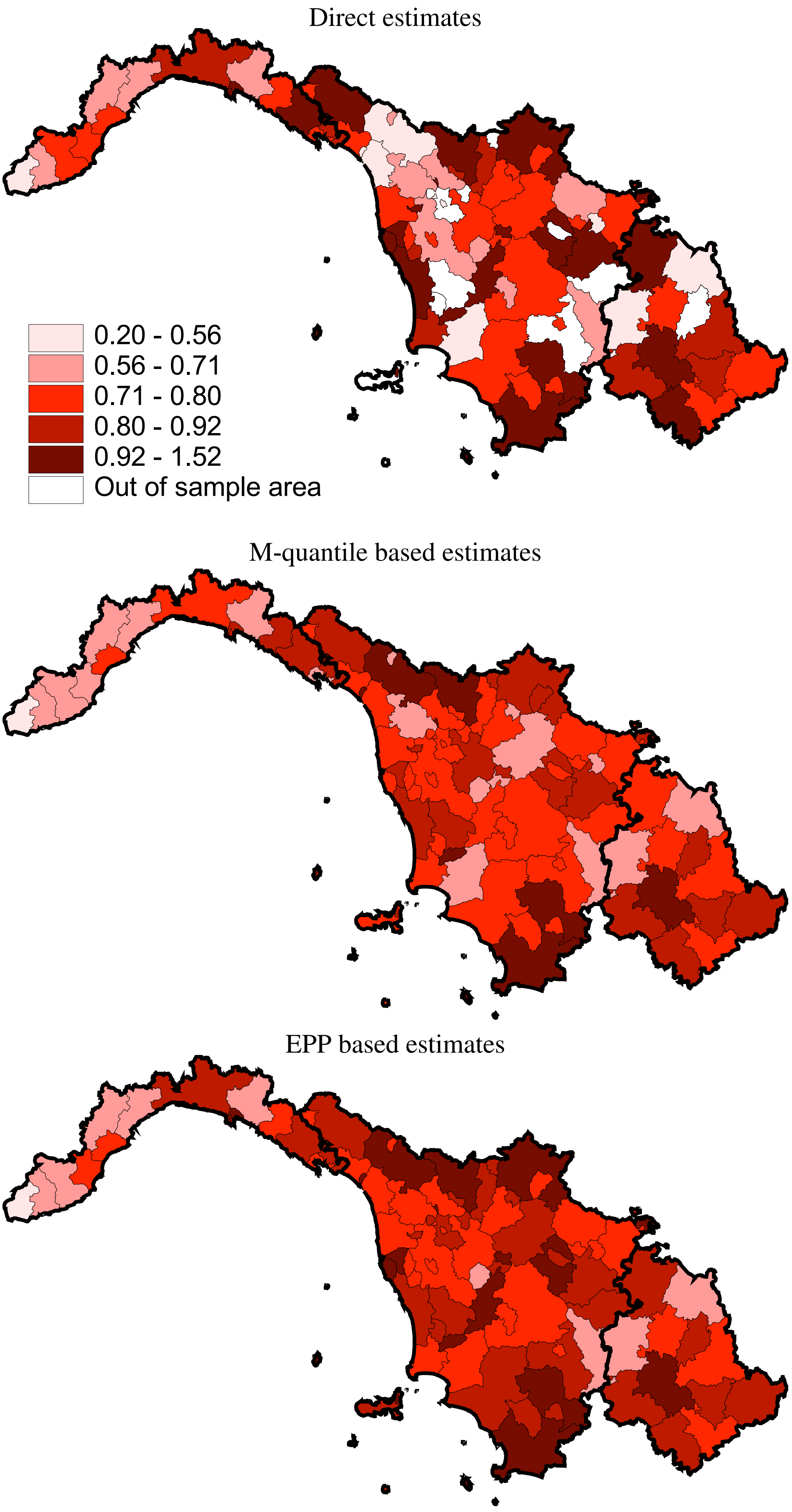


Figure 6: Estimated CVs for direct estimates (solid line) and model-based estimates. Estimated CVs for the M-quantile predictor are represented by the dashed blue line and estimated CVs for the EPP are represented by the dashed red line. HDs are ordered by increasing sample size. The last six areas are out of sample areas.



Direct estimates

M-quantile based estimates

EPP based estimates

Figure 7: Maps of the direct, M-quantile and EPP based estimates of the average number of visits to physicians in Health Districts of Liguria, Toscana and Umbria in 2000.