Expectile regression for multicategory outcomes with application to small area estimation of labour force participation

James Dawber

Southampton Statistical Sciences Research Institute, University of Southampton, UK

Nicola Salvati

Department of Economics and Management, University of Pisa, Italy

Enrico Fabrizi

DISES and DSS, Università Cattolica del S. Cuore, Italy

Nikos Tzavidis

Southampton Statistical Sciences Research Institute, University of Southampton, UK

Abstract. In many applications of small area estimation, dichotomous or categorical outcomes are the targets of statistical inference. For example, in the analysis of labour markets, proportions of working-age people in the various labour market statuses are of interest. In this paper, in line with recent literature, we consider a classification with more than three statuses and estimate related population parameters for 611 local labour market areas using data from the 2012 Italian Labour Force Survey, administrative registers and the 2011 Census. As for the methodology, we propose multinomial expectile regression models. These models provide a means to utilise M-quantile type approaches, which have been shown to be a useful alternative to mixed model approaches when parametric assumptions on the distribution of random effects cannot be met. Via a large scale simulation study, we show how this novel approach is much faster and provides reliable results when compared to multinomial mixed model approaches, and works for any number of categories rather than just a small number of categories as is more commonly the case with existing methods. Furthermore, the proposed approach potentially provides a framework for developing other methods for prediction with multicategory outcomes.

Keywords: M-quantile estimation; categorical data analysis; multinomial logistic regression

1. Introduction

In the analysis of survey data, estimates of relevant descriptive quantities may be needed for small sub-populations defined by geography or other classification criteria. When the samples specific to most sub-populations (labelled as areas) are too small to allow the direct application of survey weighted estimators, small area estimation (SAE) methodologies are used instead.

These methodologies are often based on, or assisted by, models that link survey data to auxiliary information available from external sources (see Pfeffermann, 2013; Rao and Molina, 2015, for a general introduction to the topic). More specifically, auxiliary variables are used to predict the target variable, using linear or generalised linear models depending on the nature of the variable being predicted. In most of the cases, there is a need to improve predictions accounting for area-specific unobserved heterogeneity. This can be done in different ways.

The most common solution is provided by mixed models in which random effects, in most cases random intercepts, are used to model unobserved area level heterogeneity (see Rao and Molina, 2015, Chapter 5). M-quantile models allow for an alternative approach to measuring area heterogeneity. The impact of auxiliary variables is captured by means of a linear model (or a generalisation of it) in which M-estimation is used also as an approach to ensuring outlier robustness. Unobserved heterogeneity is characterised by using quantile coefficients or q-scores, which can be viewed as pseudo-random effects. The idea, in the linear case, is that each sample data point lies on a regression plane associated with a specific quantile. If there is area-level variation beyond that predicted by auxiliary variables, these unit-specific q-scores tend to be similar (clustered) within the same area; area specific averages of these q-scores are used to summarise unobserved area-level heterogeneity. See Chambers and Tzavidis (2006) for a more detailed description of how q-scores are defined in the linear case and Section 4 of this paper for their definition in the context of the analysis of multi-category responses.

In many applications of small area estimation, dichotomous or categorical variables are targeted. For example, in the analysis of labour markets we can be interested in estimating proportions or numbers of people in different labour market statuses at the small area level using labour force surveys. Traditionally, a three-status classification is considered: employed, unemployed and economically inactive. Recent literature highlights how a larger number of labour market statuses should be considered to better describe flexible employment and the *grey area* between unemployed and inactive people (Brandolini et al., 2006; Barbieri and Scherer, 2009).

In this paper we analyse data from the 2012 Italian Labour Force Survey (ILFS) with the aim of estimating the proportions of working age people in respect to their labour market status split into three and six categories for the 611 Local Labour Market Areas (LLMAs) using auxiliary information from administrative registers and the Census 2011. The LLMAs are unplanned domains of the ILFS obtained as clusters of municipalities in which the bulk of the labour force lives and works, and where establishments can find the largest amount of the labour force necessary to occupy the offered jobs. Several LLMAs are characterised by a very small sample size that hinders reliability of direct estimates.

In the small area literature, at present, multicategory responses have received some attention. The challenge is to develop a model that can accurately capture the category-specific unobserved heterogeneity across the small areas. For example, some areas can have higher proportions of unemployed than predicted using auxiliary variables while others can be characterised by unexpected high proportions of economically inactive people. A good model should be flexible enough to capture the area-specific characteristics in order to produce accurate estimates of finite population parameters e.g. area-specific proportions. Several attempts at developing such models for SAE have been proposed.

Dealing with a three-status labour market classification, Molina et al. (2007) proposed a multinomial logistic mixed model where the same random effect is shared across multinomial categories for each domain. In other words, the random effects are univariate despite the response variable being multivariate. Saei and Taylor (2012) and López-Vizcaíno et al. (2013, 2015) propose a multinomial logistic mixed model for small area estimation of a multicategory response that allows for category-specific (multivariate) area random effects. The methods proposed by Molina et al. (2007) and Saei and Taylor (2012) were specifically for three categories and the generalisation to more categories is computationally not trivial. They used categorical explanatory variables (e.g., sex-age by area counts), so individual values within each area can be aggregated by category and predictions needed to obtain small area estimates do not require access to census microdata provided that the full cross-classification structure and population sizes of the cells are available. However, these models are unit-level in a strict sense. Whenever the auxiliary information includes one or more unit-level covariate with values that vary within a category by area group, access to census microdata would be required. The model by Molina et al. (2007) is specified at the area level. However, this can also be expressed as a unit-level model under certain covariate specifications. For example, this is the case when all the model covariates are categorical and the population counts of the full cross-classification defined by these covariates are available or when unit-level continuous covariates are available. In contrast, López-Vizcaíno et al. (2013) propose an area-level small area approach that imposes an assumption of independence between the multivariate random effects but it is unclear how well the method works with more than three categories. As noted by Verbeke et al. (2014), computational problems emerge for multivariate random-effects models due to increased dimensionality. Random-effects models for multicategory responses are not as computationally efficient, as they are for continuous or even binary outcomes (see James, 2017, for further discussion on computational difficulties).

SAE methods based on M-quantile regression models have been developed as viable alternatives and have been extended to the analysis of non-continuous responses, such as binary or count data (Chambers et al., 2014; Tzavidis et al., 2015; Chambers et al., 2016) but not to multicategory data. We propose a regression model based on a type of M-quantile called the expectile (Newey and Powell, 1987), as an alternative to multinomial mixed models when interest is in predicting finite population parameters e.g. the probability of belonging to a category within certain geographic sub-populations. Regression expectiles allow for straightforward prediction with multicategory response data. These multicategory expectiles exploit an elegant isotonic relationship between the probability and the expectile under a multinomial assumption. This allows for simple and computationally efficient predictions of q-scores and hence of small area proportions and totals. Importantly, our proposed method also ensures that all predictions are invariant to the choice of the reference category. Although this is not a problem for multinomial models, empirical work we did with multinomial random-effects models indicate that changing the reference category can impact on the small area estimates.

More specifically, the methodology we introduce in this article considers as a starting point the work by Manski and Thompson (1989). These authors discuss several predictors for binary responses under alternative loss functions, including tilted loss functions of the type used in expectile regression (later shown in equation (4)). We extend the proposal of Manski and Thompson (1989) for binary data to both the regression case and to multicategory response variables. While expectile regression has been widely explored for continuous responses (Sobotka and Kneib, 2012; Sobotka et al., 2013; Yang and Zou, 2015), expectile regression for multicategory responses is a contribution of the current research.

The rest of the paper is organised as follows. In Section 2 we introduce the ILFS dataset that is the basis of the motivating application in this paper. In Section 3 we review M-quantile, quantile and expectile regression and present the proposed approach for Bernoulli and multinomial expectile regression. In Section 4 we review the current mixed model approach to SAE for multicategory data. In Section 5 we show how multinomial expectile models can be used for

SAE and further propose a bootstrap estimator of the mean-squared error (MSE) of the small area predictor. Section 6 includes results from a simulation exercise that was designed to assess the performance of the proposed SAE methodology compared to existing methods. In Section 7 we use SAE methodologies to estimate labour market characteristics in the 611 LLMAs in Italy in 2012. Section 8 summarises the main findings. The programs that were used to analyse the data can be obtained from
`http://wileyonlinelibrary.com/journal/rss-datasets`.

2. The Italian Labour Force Survey data

The ILFS is designed to obtain quarterly estimates of the main aggregates regarding the labour market which are important both at the local and the central government levels for the development of labour market policies. A two-stage municipality-household sampling design is used to collect data. The methodological notes released by the Italian National Institute of Statistics (ISTAT) state that the institute has sound reasons to evaluate the LFS design as non-informative with respect to unemployment. Primary sampling units are stratified by province (LAU1) and population size. Secondary sampling units are selected with equal probabilities. All individuals with usual residence in the dwelling are interviewed. See De Vitiis et al. (2018) and Eurostat (2022) for details.

The ILFS provides quarterly estimates of the main aggregates for the labour market, such as employment status, type of work and work experience by gender, age and region (NUTS2). Direct estimates are reliable for large areas such as administrative regions, but not for those we target in this application. The data set we consider in this article contains data from the first quarter of 2012 which consist of measurements taken on 93,217 units aged 15-65 and distributed in 453 local labour market areas (LLMAs). They refer to unplanned domains obtained as clusters of municipalities in which the bulk of the labour force lives and works, and where establishments can find the largest amount of the labour force necessary to occupy the offered jobs. Several LLMAs are characterised by a very small sample size: the sample size ranges between 13 (Acqui Terme, Piedmont Region) and 3,301 (Milan, Lombardy Region). The mean and the median values are equal to 205.8 and 122, respectively.

The labour market status variable can be defined in different ways, with a varying number of categories, offering a more or less refined classification: (i) three categories: employed, unemployed and inactive; (ii) six categories: employed, unemployed $\underline{a}$ (in active job search, previous job experiences, formerly employed), unemployed $\underline{b}$ (in active job search, previous job experiences, formerly inactive), unemployed $\underline{c}$ (in active search of their first job), reserve workers, inactive.

Focusing on the six labour market categories we see that the precision for most direct estimates and especially those related to unemployment (rates for unemployed $\underline{a}$, $\underline{b}$, $\underline{c}$) is poor. Statistics Canada (2007) suggests that estimates with a coefficient of variation (CV) less than 16.6% are sufficiently reliable for general use and those with a CV between 16.6% and 33.3% can be published but accompanied by a warning to users whereas those with an even larger CV should be deemed completely unreliable and not published. This is not a good criterion when the estimated proportions are very small (Eurostat, 2013, p. 13). For this reason the CVs in Table 1 are not computed for areas with direct estimates below 0.02. Note that for a sizeable

Table 1. Number of small areas with values of CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for the direct estimators of the proportion of workers classified in six categories. CV for estimated proportions below 0.02 are not computed (the number of proportions not computed because of 0 count in the category within brackets).

CV	employed	unemployed a	unemployed b	unemployed c	reserve workers	inactive
< 16.7%	317	9	0	3	62	417
16.7% – 33.3%	125	105	8	22	175	36
> 33.3%	11	219	127	116	173	0
not computed	0	120 (42)	318 (107)	312 (130)	43 (7)	0
	453	453	453	453	453	453

fraction of these latter areas, the direct estimates are exactly 0 despite these areas having units in the sample (in this case standard errors are not computable with elementary methods).

The recourse to small area estimation methods is motivated by the large number of areas characterised by overly imprecise direct estimates as displayed in Table 1 as well as the presence of many out-of-sample LLMAs.

The following two explanatory variables are also available for fitting the models: age-sex, a categorical variable with six categories corresponding to female or male (F/M) by three age groups (15-24, 25-34 and 35-65) and U-count, a discrete variable measuring the number of unemployed in a given gender-age group for each LLMA according to the 2011 Census. Note that both explanatory variables used in this working model are categorical variables for which age-sex by area counts are available. This corresponds to the situation that was considered by Molina et al. (2007), and so model fitting does not require access to census microdata.

3. M-quantile, quantile and expectile estimation

Breckling and Chambers (1988) first introduced M-quantiles, which provided a general framework for the pre-established quantiles and expectiles. For a univariate, continuous variable Y with density $f(\cdot)$, the M-quantile, M_q, for a pre-selected q is defined by:

$$E\left[\psi_q\left(y - M_q\right)\right] = \int_{-\infty}^{\infty} \psi_q\left(y - M_q\right) f(y)dy = 0, \tag{1}$$

where $\psi_q(\cdot)$ is an influence function defined by:

$$\psi_q(u) = \begin{cases} 2(1-q)\psi(u) & u \leqslant 0 \\ 2q\psi(u) & u > 0, \end{cases} \tag{2}$$

and $\psi(\cdot)$ is assumed to be a function that is bounded and monotone non-decreasing over the real line with $\psi(0) = 0$. This M-quantile specification provides a general framework that includes the mean when $q = 0.5$ and $\psi(u) = u$ and the median when $q = 0.5$ and $\psi(u) = \mathrm{sgn}(u)$. For general q, these two influence functions also determine the expectile and the quantile respectively (Newey and Powell, 1987; Koenker and Bassett, 1978). While these two estimators are scale invariant, not all specifications of $\psi(\cdot)$ ensure this, hence generally a nuisance scale parameter should be added to equation (1). One such function of $\psi(\cdot)$ is the Huber function: $\psi_k(u) = uI(|u| \leq k) + k\mathrm{sgn}(u)I(|u| > k)$, where $k \in (0, +\infty)$ is the value of the tuning constant, and is the suggested choice for the M-quantile approach to SAE in most publications (Chambers and Tzavidis, 2006; Chambers et al., 2014, 2016).

To extend to the regression case, the argument inside the influence function is replaced by standardised residuals. For a linear M-quantile regression model, with $M_q(\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_q$, estimators of the M-quantile regression coefficients $\boldsymbol{\beta}_q$ can be obtained by solving the estimating equations:

$$\sum_{i=1}^{n} \psi_q \left(\frac{y_i - M_q(\boldsymbol{x}_i)}{\sigma_q} \right) \boldsymbol{x}_i = 0, \tag{3}$$

where y_i is the continuous response variable of interest, $\boldsymbol{x}_i$ represents a p-dimensional vector of observed covariates, $i = 1, \ldots, n$, and σ_q is a scale parameter, which must also be estimated. Again, note that if $q = 0.5$ and $\psi(u) = u$, then the estimating equations reduce to ordinary least squares. For general q and $\psi(u) = u$, equation (3) allows for estimating regression coefficients for the conditional expectiles rather than mean. Solutions to the estimating equations can be found simply using iteratively reweighted least squares (IRLS), the Newton-Raphson algorithm or linear programming (Bianchi et al., 2018; Koenker and Bassett, 1978).

3.1. Expectile regression models for binary data

In this section we focus specifically on the expectile estimator under a Bernoulli assumption. For binary data, we take a result from Manski and Thompson (1989) as our starting point.

THEOREM 1. *Let Y be a Bernoulli random variable with probability π, the expectile q of Y is defined as:*

$$\mu_q = \frac{q\pi}{(1-q)(1-\pi) + q\pi}. \tag{4}$$

PROOF. If $Y \sim Bernoulli(\pi)$, the expectile q for a binary variable can be derived by solving equation:

$$E[\psi_q(Y - \mu_q)] = 0,$$

where $\psi_q(\cdot)$ is the influence function for the expectile with $\psi(u) = u$. Then:

$$\begin{aligned}
\sum_{y=0}^{1} \psi_q(y-\mu_q)p(y) &= \sum_{y=0}^{1} \left\{2(1-q)(y-\mu_q)I_{y \le \mu_q} + 2q(y-\mu_q)I_{y > \mu_q}\right\} p(y) \\
&= 2\left\{(1-q)(-\mu_q)p(y=0) + q(1-\mu_q)p(y=1)\right\} \\
&= 2\left\{-\mu_q(1-q)(1-\pi) + (1-\mu_q)q\pi\right\} = 0,
\end{aligned}$$

where $p(y)$ is the probability mass function. This formula can be rearranged to make μ_q the subject.

Though this result is simple, it shows an elegant relationship between the probability (π) and the expectile (μ_q) of any Bernoulli variable for a given q. Note that $\mu_q \in (0, 1)$ for any given q or π, and of course, $\mu_{0.5} = \pi$, the variable expectation. So if $\hat{\pi}$ is a maximum likelihood estimate then with the invariance property any monotone function of $\hat{\pi}$, such as the one in equation (4), must be the maximum likelihood estimate of that function. Hence estimates of the binary expectiles can be calculated directly from estimates of the probability and vice-versa:

$$\pi = \frac{\mu_q(1-q)}{\mu_q(1-2q) + q}. \tag{5}$$

Equation (5) lends itself naturally as the basis for an isotonic model of the binary response, rather than a weaker single-crossing model which assumes a latent variable (see Manski and Thompson, 1989).

Equation (4) can be easily extended to the regression case. For instance, by performing a logistic regression on the binary response, estimates of π conditional on $\boldsymbol{x}_i$ can be obtained and used to estimate μ_q. For regression covariates $\boldsymbol{x}_i$ and regression parameter vector $\boldsymbol{\beta}$, the probability can be modelled as:

$$\pi(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}. \tag{6}$$

So given $\pi(\boldsymbol{x}_i)$, then the regression expectile $\mu_q(\boldsymbol{x}_i)$ can be found by substituting into equation (4) giving:

$$\begin{aligned}\mu_{qi} &= \frac{q\pi(\boldsymbol{x}_i)}{(1-q)(1-\pi(\boldsymbol{x}_i)) + q\pi(\boldsymbol{x}_i)} \\ &= \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta} + \log \frac{q}{1-q})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} + \log \frac{q}{1-q})}.\end{aligned}$$

Hence there is a simple relationship between the binary expectile and probability regression parameters:

$$\boldsymbol{x}_i^T \boldsymbol{\beta}_q = \boldsymbol{x}_i^T \boldsymbol{\beta} + \log \frac{q}{1-q}, \tag{7}$$

where $\boldsymbol{\beta}_q$ is the expectile regression coefficient vector for a given q. This shows that the binary expectile regression coefficient $\boldsymbol{\beta}_q$ is just an intercept adjustment of the logit of q from the logistic regression coefficient $\boldsymbol{\beta}$. This property ensures that the regression expectiles will not cross over, a common problem with any quantile-like regression model. To visualise this property, Figure 1 (left panel) presents binary expectile regression lines in a simple example.

Chambers et al. (2016) show that their proposed binary expectile regression estimator is a version of the asymmetric maximum likelihood estimator proposed by Efron (1992). This regression estimator is exactly the same as the proposed estimator, except without the Bernoulli distribution assumption. This stricter assumption provides the relationship between the binary probabilities and expectiles that make the generalisation to multicategory data relatively straightforward.

3.2. *Expectile regression models for multicategory data*

Since it has been shown that binary expectiles are merely transformations of the probability of a Bernoulli variable, the extension to multicategory response from binary response follows. By estimating the probability vector of the multicategory response under a multinomial distribution assumption the expectiles can be calculated similarly to the Bernoulli case.

THEOREM 2. *Let $\boldsymbol{Y}$ be a Multinomial$(1, \boldsymbol{\pi})$ random variable for G distinct categories represented by an $n \times G$ matrix with $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_G)$, the multinomial expectile is given by:*

$$\mu_{qg} = \frac{q\pi_g}{(1-q)(1-\pi_g) + q\pi_g}, \quad g \in 1, 2, \ldots, G. \tag{8}$$

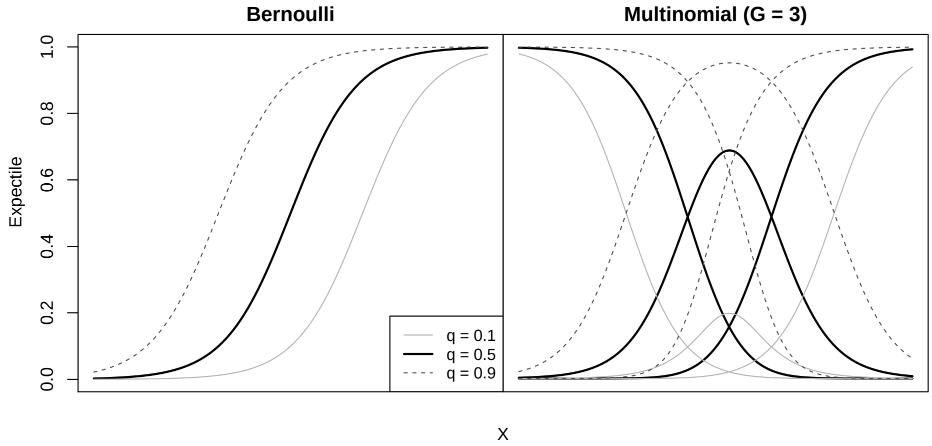

Figure 1. Bernoulli and multinomial expectiles from a simple simulated data example with $q = 0.1, 0.5, 0.9$. In these examples, X is generated from a uniform distribution, and $\beta = 2$ for the first plot, and $\beta_1 = [-2\ 2]^T$ and $\beta_2 = [-7\ 4]^T$ for the second plot.

PROOF. The multinomial expectile ($\boldsymbol{\mu}_q$) can be derived from the multinomial probability $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_G)$ by solving $E[\psi_q(\boldsymbol{Y} - \boldsymbol{\mu}_q)] = \mathbf{0}$ in much the same way as in the Bernoulli case.

Due to this direct relationship between expectiles and probabilities, the multinomial expectile remains the same for any $\boldsymbol{Y} \sim Multinomial(m, \boldsymbol{\pi})$, regardless of $m > 0$, hence the relationship holds generally. For simplicity, the methods in the remainder of the article will focus on the simple case of $m = 1$. Extensions to the general multinomial response with $m > 1$ are outlined in Section S.1 of the supplementary material.

The extension from binary to multicategory expectile regression first requires regression predictions of $\boldsymbol{\pi}(\boldsymbol{x}_i)$, which can be found using multinomial logistic regression. Let $\boldsymbol{\beta}$ be the $p \times (G-1)$ matrix of the multinomial logistic regression coefficients, then $\boldsymbol{\pi}(\boldsymbol{x}_i)$ can be expressed as:

$$\pi_g(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_g)}{1 + \sum_{g=1}^{G-1} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_g)}, \quad g = 1, \ldots, G-1, \tag{9}$$

$$\pi_G(\boldsymbol{x}_i) = \frac{1}{1 + \sum_{g=1}^{G-1} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_g)}, \tag{10}$$

where $\boldsymbol{\beta}_g$ is the g-th column of $\boldsymbol{\beta}$, and $\pi_G(\boldsymbol{x}_i)$ represents the probabilities of the reference category, which is arbitrarily chosen to be the G-th or last category. Expressing the probabilities like this ensures that the probabilities must sum to one which is obviously an important requirement. This follows the standard multinomial logistic regression model.

With estimates of $\boldsymbol{\pi}(\boldsymbol{x}_i)$ we can then calculate the estimates of the multinomial expectile vector $\boldsymbol{\mu}_q(\boldsymbol{x}_i)$ using equations (8), (9) and (10), from which we obtain:

$$\mu_{qg}(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_g + \log \frac{q}{1-q})}{1 + \sum_{g=1}^{G-1} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_g + \log \frac{q}{1-q})}, \quad g = 1, \ldots, G-1, \tag{11}$$

$$\mu_{qG}(\boldsymbol{x}_i) = \frac{1}{1 + \sum_{g=1}^{G-1} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_g + \log \frac{1-q}{q})}. \tag{12}$$

Note that this simplifies down to the binary case when $G = 2$, as required. The proposed method is then computationally very simple due to the estimates coming from the multinomial logistic regression, and then the predictions coming from the q-score derivations. To visualise these multicategory expectiles Figure 1 (right panel) shows fitted lines in an example. Based on this derivation, the choice of the reference category does not affect $\boldsymbol{\mu}_{qg}(\boldsymbol{x}_i)$.

It is worth noting that, unlike probabilities that sum to one, multinomial expectiles do not sum up to a fixed constant. The sum of the expectiles does not even sum to a constant dependent solely on q, but instead it is also dependent on $\boldsymbol{x}_i$ hence this sum cannot simply be universally integrated into the model structure, as is done for multinomial logistic regression.

Expectiles are often criticised for their lack of robustness and interpretability which is why quantiles are the more natural option for continuous response. This criticism is relevant for continuous variables but much less so for binary and multicategory response variables. In fact, with multicategory responses we cannot talk of heavy tails over-influencing parameter estimation, as observations can only be a vector with elements of 0 or 1. Observations can still have high leverage however, which should be considered and assessed as with any regression model. In this sense, robust estimates can still be sought by first robustly estimating the probability. Cantoni and Ronchetti (2001) developed an approach to robust GLMs based on quasi-likelihood estimation, naturally including binary responses. This approach down-weights observations that have large Pearson residuals and high leverage. Hence predictions of binary expectiles can be made from the robust predictions of the probabilities. This could similarly be done using robustly estimated multinomial regression models such as that proposed by Tabatabai et al. (2014).

As for interpretability, with respect to ordinary quantiles, we note that there is no obvious interpretation for the binary or multicategory case. As Manski and Thompson (1989) show, the isotonic form of Bernoulli quantiles are either 0 or 1 for all q determined simply by $I(\pi \geq 1-q)$. This may be somewhat intuitive, though it is essentially trivial and not very interpretable. The single-crossing models of binary quantiles, with a latent variable assumption, provide a more useful conceptualisation for binary quantiles with a range within $(0, 1)$. Bernoulli expectiles naturally have this range even with the simpler isotonic model, where the higher q is, the closer the expectile is to one. The simple relationship, and the fact the model can remain isotonic suggests that the expectile is perhaps more natural with binary and multicategory response data, since the isotonic quantile is trivial and single-crossing quantiles have added complexity. M-quantile models using a Huber influence function can also be expressed isotonically, which is in Section S.2 of the supplementary material.

With the Bernoulli and multinomial expectile regression models now presented, we next provide an introduction to SAE and how these models can be applied.

4. Small area estimation for multicategory data

SAE models 'borrow strength' from the population (e.g. country) to infer about the sub-population (e.g. regions or provinces), when sample sizes are too small for precise direct estimation (Rao and Molina, 2015). We broadly focus on two general approaches to SAE models, the mixed model approach and the M-quantile approach, see Dawber and Chambers (2019) for a recent review and comparison of these two approaches. In this section we first present the mixed model approach to SAE for multicategory data, before presenting the application of multinomial expectile regression using the M-quantile approach.

To extend small area estimators to multicategory data it is necessary to extend the binary response approaches to the multicategory case. In particular, to extend the small area estimators to multicategory data requires a multinomial random-effects model. Hartzel et al. (2001) unified multinomial logistic random-effects model methods and presented a model for multicategory response data. Suppose there were G categories and J small areas and let $\boldsymbol{y}_{ij}$ be the G-vector of multicategory responses for unit i in area j, $i = 1, \ldots, N_j$ and $j = 1, \ldots, J$ and N_j is the population size of small area j, with probabilities $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \ldots, \pi_{ijG})$. Also let $\boldsymbol{x}_{ij}$ be the explanatory fixed-effect variables for each unit and $\boldsymbol{z}_{ij}$ be the random-effect design matrix, let $\boldsymbol{\beta}_g$ be the fixed-effect parameter vector associated with the g-th category and γ_{jg} be a component of the random-effect parameter vector $\boldsymbol{\gamma}_j$, which is dependent on both the area and the category. Finally, let $h(\cdot)$ be a known invertible link function, $\boldsymbol{\mu}_{ij} = E[\boldsymbol{y}_{ij}|\boldsymbol{\gamma}_j]$ such that $h(\boldsymbol{\mu}_{ij})$ can be expressed as a linear function of both $\boldsymbol{\beta}_g$ and $\boldsymbol{\gamma}_j$. Specifically, if category G is arbitrarily set to be the baseline category then the multinomial logistic random-effects model for the g-th category is given by:

$$h(\mu_{ijg}) = \log\left(\frac{\pi_{ijg}}{\pi_{ijG}}\right) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_g + \boldsymbol{z}_{ij}^T \boldsymbol{\gamma}_j, \tag{13}$$

where $\boldsymbol{\gamma}_j \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{j(G-1)})$ and unconstrained covariance matrix as recommended by Hartzel et al. (2001). The multinomial distribution has categories that are correlated, and hence the random-effects structure should capture that correlation in the $\boldsymbol{\Sigma}$ parameter. The random effects should be able to capture the between-category correlation, as well as the different within-category variation across areas. This is why the random effects structure must be considered multi-dimensionally and should not be constrained too much.

Small area estimates for the category g are then obtained by:

$$\hat{y}_{jg}^{CEP} = N_j^{-1}\left[\sum_{i \in s_j} y_{ijg} + \sum_{i \in r_j} \hat{\mu}_{ijg}\right], \; g = 1, \ldots, G, \tag{14}$$

where y_{ijg} is the observed value for unit i in small area j and for category g, s_j denotes the set of n_j units sampled in area j, and r_j the $N_j - n_j$ remaining (i.e. non-sampled) units in this area. Here $\hat{\mu}_{ijg} = h^{-1}(\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_g + \hat{\gamma}_{jg})$, $\hat{\boldsymbol{\beta}}_g$ and $\hat{\gamma}_{jg}$ are the estimates of the model coefficients and predicted values of the random area effects, respectively. This small area estimator is known in the literature as the conditional expectation predictor (CEP). See Chambers et al. (2016) for details. One difficulty with such a model is ensuring that $\hat{y}_{jg}^{CEP}$ is the same regardless of the choice of reference category. Our understanding is that changing the reference category will affect $\hat{\gamma}_{jg}$ because they are derived relative to that category. This then results in changes to $\hat{\mu}_{ijg}$, which is not a desirable property. This is confirmed by our empirical investigations (see Section 5), and implies that the small area predictions vary depending on the choice of the reference category. To the best of the authors' knowledge this problem has not been solved for SAE using multinomial mixed models.

Molina et al. (2007) used a multinomial logistic random-effects model for SAE on labour force status. The model they used specifies that the random-effect variance across areas is the same for each logit, which is a rather restrictive and unrealistic constraint. A less restrictive random-effects structure was utilised in SAE by Scealy (2010) and Saei and Taylor (2012). Fewer restrictions on the correlation were used in these cases, and were generally shown to yield improved small area estimates compared to the constrained method by Molina et al. (2007). However, the methods presented focussed on the particular case of three categories, and did not provide methods more generally with more than three, presumably due to the added complexity of the random effects structure. López-Vizcaíno et al. (2013) also utilised a multinomial logistic random-effects regression for SAE but with constrained independence between the categories of the random effects, which is not a realistic assumption.

Estimates are always accompanied by a measure of accuracy, the most common one being the mean squared error (MSE). Molina et al. (2007) proposed an analytical MSE estimator based on Taylor expansion based on methods by Prasad and Rao (1990) and a bootstrap for MSE estimation on finite populations. This method works by generating bootstrap populations from the multinomial model with probabilistic properties similar to the original model but conditional on the initial sample, and then extracting samples from these populations. Saei and Taylor (2012) proposed a MSE estimator based on an analytical approximation approach presented in Saei and Chambers (2003).

The problem with this mixed model approach with multicategory response is twofold. First the computation can be unstable and time-consuming due to the multi-dimensional integrals that are calculated, and second the random-effects structure is difficult to realistically specify and check in the model. Hence with mixed model approaches not offering a viable general solution to SAE with a multicategory response, we turn to the alternative M-quantile approaches.

5. SAE using binary and multicategory expectile regression models

The M-quantile approach to SAE relies on so-called 'q-scores' that are calculated for each observed sample unit. Before the expectile small area estimator can be introduced for the binary and multicategory responses we first introduce how the q-scores are derived for continuous response variables (Chambers and Tzavidis, 2006). Mixed effects models assume that variability associated with the conditional distribution of y given x can be at least partially explained by a prespecified hierarchical structure, such as the small areas of interest. However, an alternative approach to modelling the variability in this conditional distribution is via M-quantile regression, which does not depend on a hierarchical structure. In this approach the q-score conditional on $\boldsymbol{x}_{ij}$, is the value of q that yields regression predictions equal to the response y_{ij}. More formally the q-score, q^*_{ij} for the i-th unit in area j is such that $y_{ij} = M_{q^*_{ij}}(\boldsymbol{x}_{ij}) = \boldsymbol{x}^T_{ij}\boldsymbol{\beta}_{q^*_{ij}}$. Hence, if quantile regression is used, and the response is directly on the median regression line, then the q-score will be $q^*_{ij} = 0.5$. If the response is in the upper tail, then the q-score will be close to one, and if the response is in the lower tail, then the q-score will be close to zero.

Generally, the objective of a q-score is to assign each observation a score in $(0, 1)$ where q-

scores close to 0.5 are considered 'typical' and q-scores close to 0 or 1 are considered 'extreme' at each end of the conditional distribution. Furthermore, these q-scores should ideally distribute across the full range of $(0, 1)$, so the heterogeneity of each unit is suitably captured. Note that an appropriately fitted quantile regression will yield q-scores that are approximately uniformly distributed. An alternative way of conceptualising q-scores is to consider them like a residual, except instead of being distributed on the real line about zero as is the case for Pearson residuals, they are distributed in $(0, 1)$, with the centre being close to 0.5 for symmetric distributions.

As stated above, the q-scores variability reflects the heterogeneity at the unit level. If clustering exists, population units in the same cluster (or small area) will have similar q-scores and these will be different from those of units that belong to other clusters (or areas). An area-specific q-score can be then defined as $q_j^* = E[q_{ij}^*|j]$, where the expectation is conditional on the distribution of q_{ij}^* within area j. Then, in mixed models the random part effectively accounts for residual between-area variation beyond that explained by x. In contrast, a linear M-quantile modelling approach captures this residual between-area variation by the deviation of the area-specific M-quantile regression coefficient $\boldsymbol{\beta}_{q_j^*}$ from the 'median' M-quantile coefficient $\boldsymbol{\beta}_{0.5}$. This allows Chambers and Tzavidis (2006) to write the small area predictor obtained fitting the linear M-quantile regression model in a form that mimics that achieved fitting mixed effects model, where $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{0.5}$ represents the fixed effect part of the predictor and $\bar{\boldsymbol{x}}_j^T(\boldsymbol{\beta}_{q_j^*} - \boldsymbol{\beta}_{0.5})$ can be interpreted as a pseudo-random effect for area j with $\bar{\boldsymbol{x}}_j$ denoting the vectors of average values of $\boldsymbol{x}_{ij}$ of area j. Therefore these q-scores are best thought of as M-quantile analogues to the random effects in a mixed model approach.

The q-score definition for continuous response cannot be simply extended to the binary response case. This is because binary responses of 0 and 1 will always be outside the $(0, 1)$ support of binary expectiles, hence $y_{ij} \neq \mu_{q_{ij}^*}(\boldsymbol{x}_{ij})\ \forall q_{ij}^*$. As such, different methods must be used to acquire appropriate q-scores for these response types. For binary response, Chambers et al. (2016) suggested three q-score definitions (the authors refer to q-scores as M-quantile coefficients) and they use in the application their *Definition 3* where the estimated q-score for unit i in area j is $\hat{q}_{ij}^*$, where $\hat{M}_{\hat{q}_{ij}^*}(\boldsymbol{x}_{ij}) = \{\hat{M}_{0.5}(\boldsymbol{x}_{ij}) + y_{ij}\}/2$. However, we propose a new q-score definition, which can be used for binary and multicategory responses. To explain why, first consider what desirable properties a binary q-score should have:

- The support of the q-scores should be $(0, 1)$.
- For a fixed $\boldsymbol{x}_{ij}$, the q-score, q_{ij}^*, should be greater when $y_{ij} = 1$ than $y_{ij} = 0$.
- If q_{ij}^* is the q-score derived from y_{ij} and $\boldsymbol{x}_{ij}$, then $1 - q_{ij}^*$ should be the q-score if $1 - y_{ij}$ is used as response instead of y_{ij}. This symmetry allows q-scores to be equidistant to 0.5, regardless of how y_{ij} is defined.
- For binary response, 'typical' observations occur in two scenarios: when $y_{ij} = 0$ with π_{ij} close to 0, and also when $y_{ij} = 1$ with π_{ij} close to 1. Then q_{ij}^* should be defined so that $q_{ij}^* \cong 0.5$ for 'typical' observations. Hence, when $y_{ij} = 0$, then $q_{ij}^* \rightarrow 0.5$ as $\pi_{ij} \rightarrow 0$, and similarly when $y_{ij} = 1$, then $q_{ij}^* \rightarrow 0.5$ as $\pi_{ij} \rightarrow 1$.
- The ensemble of the q-scores should be as close to a uniform distribution on $(0, 1)$ as possible.

The latter requirement is based on the uniformity of q-scores of continuous response derived from quantile regression, and provides a degree of variability that will obtain q-scores that capture the required unit-level heterogeneity. Although a binary variable cannot simply be transformed into a uniform variable, and since expectile rather than quantile regression is used, it nevertheless provides useful properties for q-scores in practice. To 'uniformalise' a Bernoulli variable we refer to randomised quantile residuals presented by Dunn and Smyth (1996).

THEOREM 3. *Let $F(y_{ij};\pi_{ij})$ be the distribution function of a Bernoulli variable, and $U_{ij} \sim \mathcal{U}(a_{ij}, b_{ij})$, where $a_{ij} = \lim_{y\to y_{ij}^-} F(y;\pi_{ij})$ and $b_{ij} = F(y_{ij},\pi_{ij})$, then $U_{ij} \sim \mathcal{U}(0,1)$.*

PROOF. We have that $U_{ij}|(y_{ij} = 0) \sim \mathcal{U}(a_{ij} = 0, b_{ij} = 1 - \pi_{ij})$, and $U_{ij}|(y_{ij} = 1) \sim \mathcal{U}(a_{ij} = 1 - \pi_{ij}, b_{ij} = 1)$. The marginal distribution of U_{ij} will be:

$$\begin{aligned} f_{U_{ij}}(u) &= (1-\pi_{ij}) f_{U_{ij}|y_{ij}=0}(u) + \pi_{ij} f_{U_{ij}|y_{ij}=1}(u) \\ &= (1-\pi_{ij})(1-\pi_{ij})^{-1}\mathbf{1}_{(0,1-\pi_{ij}]} + \pi_{ij}(\pi_{ij})^{-1}\mathbf{1}_{(1-\pi_{ij},1]} = \mathbf{1}_{(0,1]}, \end{aligned}$$

where $\mathbf{1}_{(0,1]}$ is the density function of the standard uniform distribution.

Based on Theorem 3, we define q^*_{ij} conditional on $y_{ij}, \boldsymbol{x}_{ij}$ any number generated from $\mathcal{U}(a_{ij}, b_{ij}]$. It can easily be shown that this definition of a binary q-score also captures all the previously listed properties.

In empirical analyses π_{ij} would be replaced by an estimate $\hat{\pi}_{ij}$. Moreover to obtain a replicable estimator we replace the uniformly distributed number in the definition of q^*_{ij} with the midpoint between $\hat{a}_{ij}$ and $\hat{b}_{ij}$ defined replacing π_{ij} with $\hat{\pi}_{ij}$ in the definitions of a_{ij}, b_{ij}. The empirical q-score $\hat{q}^*_{ij}$ will then be defined as

$$\hat{q}^*_{ij} = \frac{1 - \hat{\pi}_{ij} + y_{ij}}{2}. \tag{15}$$

In fact if $y_{ij} = 0$, then $\hat{q}^*_{ij} = (1 - \hat{\pi}_{ij})/2$, and if $y_{ij} = 1$ then $\hat{q}^*_{ij} = (2 - \hat{\pi}_{ij})/2$; both can be written as (15). We note that the $\hat{q}^*_{ij}$ score will not be exactly uniform, but will provide a reasonable approximation to uniform in practice.

Consider a simple example, suppose $\hat{\pi}_{ij} = 0.99$ and $y_{ij} = 1$ is observed which is very much expected, then the corresponding q-score would be $\hat{q}^*_{ij} = 0.505$. Conversely, if the observation was actually $y_{ij} = 0$ then this would not typically occur hence the q-score should be close to 0, which it is with $\hat{q}^*_{ij} = 0.005$. Further intuitive properties arise from this definition also, suppose $\hat{\pi}_{ij} = 0.5$ then if $y_{ij} = 0$ or 1 then the q-score is either 0.25 or 0.75, an intuitive outcome. Also note that on average the q-score will be 0.5. So this proposed derivation of the q-scores offers a new way of defining binary q-scores, and in practice estimates can be made very easily based on estimates of π_{ij}.

The three binary q-score definitions introduced by Chambers et al. (2016) do not meet all of the proposed properties, are more complex and hence are not as adequate as the proposed q-score. Chambers et al. (2016) selected the third definition as most preferable, which we compare to the proposed q-score approach in Figure 2 using straightforward simulated data. It is clear how the proposed approach includes q-scores distributed all across $(0,1)$ and are very close to 0.5 when the probability of the observed response is very high. The current approach by Chambers et al. (2016) has no q-scores close to 0.5, and none even within the range of $(0.4, 0.6)$.

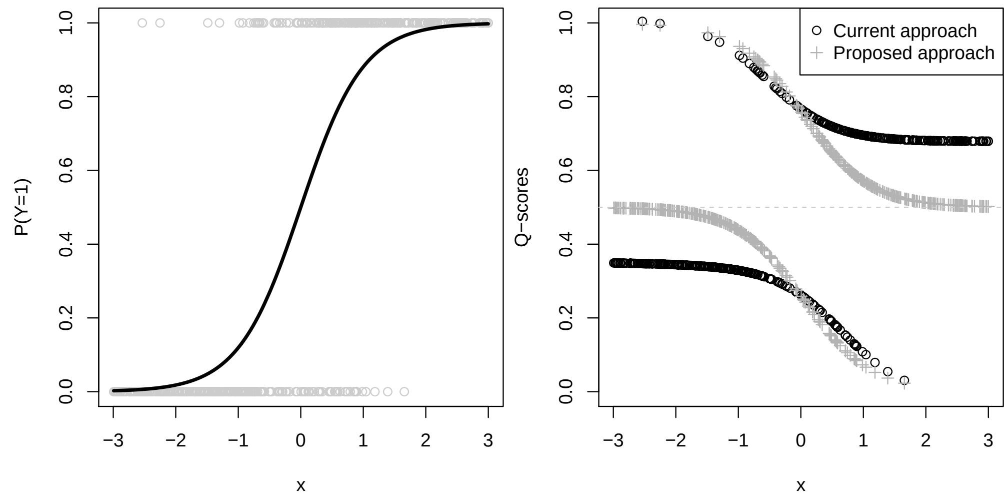

Figure 2. Simulated $Y \sim Bernoulli(\pi)$ data on the left figure where $\pi = (1+exp(-2x))^{-1}$ fitted line shown. On the right are the corresponding q-scores for the current and proposed approaches.

For binary response variables, individual units can be represented by a single q-score, where the q-scores increase with the response for given covariates. To extend binary q-scores to the multicategory case requires first conceptualising a multivariate q-score. In such cases, a q-score vector is required to capture each dimension. So multicategory response variables with $G > 2$ categories cannot be expressed by a single valued q-score, but instead a q-score vector of length G. The same occurs with the random-effects models, where each unit requires a random-effects vector to specify how the unit varies across categories.

Calculating the q-score vector for each unit can be done very similarly to the binary case. Suppose $\hat{\boldsymbol{q}}^*_{ij} = \left(\hat{q}^*_{ij1}, \ldots, \hat{q}^*_{ijG}\right)$ is the empirical q-score vector for the i, j-th multinomial response $\left(y_{ij1}, \ldots, y_{ijG}\right)$, from one trial, then the g-th element of the vector can be defined similarly to equation (15): $\hat{q}^*_{ijg} = ((1 - \hat{\pi}_{ijg}) + y_{ijg})/2$. A useful property of this q-score definition is that the sum of the q-score vector will be fixed, equal to $G/2$. Note that in the special binary case this ensures the symmetric property of the q-scores, as is one of the listed properties. The idea behind this method is that $\hat{q}^*_{ijg}$ can now be inserted back into the multinomial expectile regression equations (11) and (12), providing a conditional multinomial expectile prediction.

For this multicategory case a ‘typical’ q-score vector will have all elements close to 0.5, which will occur when all marginal probabilities are close to the corresponding observation. For example, suppose $G = 3$, then the q-score vector of a ‘typical’ unit would be close to $(0.5, 0.5, 0.5)$ when $\hat{\boldsymbol{\pi}}_{ij} = (0.01, 0.98, 0.01)$ and $\boldsymbol{y}_{ij} = (0, 1, 0)$.

With the q-scores now defined for the binary and multicategory case, as well as the expectile regression estimators, the SAE estimators can now also be defined. Chambers et al. (2016) defined the M-quantile SAE estimator for binary data which remains unchanged for the proposed expectile SAE estimator. The binary small area estimator based on expectile (E) regression model for area j is calculated using:

$$\hat{p}_j^E = N_j^{-1}\left[\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\hat{\mu}_{\hat{q}_j^*}(\boldsymbol{x}_{ij})\right] \tag{16}$$

where $\hat{q}_j^*$ is an aggregate of all q-scores in area j (typically the average) and $\hat{\mu}_{\hat{q}_j^*}(\boldsymbol{x}_{ij}) = \exp(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\hat{q}_j^*})/(1+\exp(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\hat{q}_j^*}))$. So essentially the estimate $\hat{\mu}_{\hat{q}_j^*}(\boldsymbol{x}_{ij})$ serves as an estimate of the proportion at area j given $\boldsymbol{x}_{ij}$. The further the aggregated q-score $\hat{q}_j^*$ is from 0.5, the greater the area effect.

Multicategory q-score vectors can be simply applied to calculate the expectile-based small area estimates. This is very similar to the binary case in equation (16), except done marginally. In the binary case the expectile estimator is $\hat{\mu}_{\hat{q}_j^*}(\boldsymbol{x}_{ij})$. This expectile estimate represents the proportion at area j. In the multicategory case, expectile q-scores are a vector where each category represents an element, hence the g-th element of $\hat{\boldsymbol{q}}_j^*$ is aggregated within each category, e.g. for the mean q-scores of area j and category g: $\hat{q}_{jg}^* = n_j^{-1}\sum_{i=1}^{n_j}\hat{q}_{ijg}^*$. Hence, the expectile estimate for the i-th unit in category g and area j is $\hat{\mu}_{\hat{q}_{jg}^*g}(\boldsymbol{x}_{ij})$ using equation (11). The multinomial expectile (ME) small area estimates of the 'pseudo-proportion' for area j in category g can then be calculated using:

$$\hat{p}_{jg}^{ME'} = N_j^{-1}\left[\sum_{i\in s_j} y_{ijg} + \sum_{i\in r_j}\hat{\mu}_{\hat{q}_{jg}^*g}(\boldsymbol{x}_{ij})\right],\ g = 1,\dots,G, \tag{17}$$

where $\hat{\mu}_{\hat{q}_{jg}^*g}(\boldsymbol{x}_{ij}) = \frac{\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{\hat{q}_{jg}^*g})}{1+\sum_{g=1}^{G-1}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{\hat{q}_{jg}^*g})}$, $g = 1,\dots,G-1$ and $\hat{\mu}_{\hat{q}_{jG}^*G}(\boldsymbol{x}_{ij}) = \frac{1}{1+\sum_{g=1}^{G-1}\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{\hat{q}_{jg}^*g})}$, $g = G$. The value $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{\hat{q}_{jg}}$ can be written as $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g + \boldsymbol{x}_{ij}^T(\hat{\boldsymbol{\beta}}_{\hat{q}_{jg}^*} - \hat{\boldsymbol{\beta}}_g)$. Note that, in view of (11), (12), the difference $\hat{\boldsymbol{\beta}}_{\hat{q}_{jg}^*} - \hat{\boldsymbol{\beta}}_g$ is a vector whose first entry is $\log\frac{\hat{q}_{jg}^*}{1-\hat{q}_{jg}^*}$ while all the rest is 0; $\log\frac{\hat{q}_{jg}^*}{1-\hat{q}_{jg}^*}$ can then be thought as a pseudo random effect specific to both area and category.

Moreover, note that the marginal residual $\boldsymbol{x}_{ij}^T(\hat{\boldsymbol{\beta}}_{\hat{q}_{jg}^*} - \hat{\boldsymbol{\beta}}_g)$ can be decomposed in two pseudo-effects: (i) an area pseudo-random effect of area j defined as $\bar{\boldsymbol{x}}_j^T(\hat{\boldsymbol{\beta}}_{\hat{q}_{jg}^*} - \hat{\boldsymbol{\beta}}_g)$ and (ii) an individual pseudo-random effect of unit i in area j denoted as $(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)^T(\hat{\boldsymbol{\beta}}_{\hat{q}_{jg}^*} - \hat{\boldsymbol{\beta}}_g)$. This ME predictor is very similar to the CEP in equation (14), where the only difference is how μ_g is estimated. The CEP uses random-effects to characterise area-level differences, while the ME predictor uses the proposed expectile q-scores. Chambers et al. (2016) develop some empirical evidence for the relationship between the random effects of a generalised linear mixed models (GLMM) and the pseudo-random effects obtained by M-quantile models for binary data conducting a simulation experiment. The results suggest that estimated pseudo-random effects are comparable with predicted area effects computed by using standard GLMM fitting procedures as far as capturing intra-area (domain) variability is concerned.

From a practical implementation point of view, we note that all the $\boldsymbol{x}_{ij}$, $i \in r_j$ need to be known in order to compute (17). The information requirement is decreased when the covariates are categorical: in this case only the totals at the area level need to be known.

The pseudo-proportions $\hat{p}_{jg}^{ME'}$ cannot act as an estimate of a proportion because they are not guaranteed to sum to one. Hence the multinomial expectile proportion estimates $\hat{p}_{jg}^{ME}$ for each category g within area j should be normalised such that the pseudo-proportions sum to one across categories. An important property of the proposed method is that the choice of reference category does not affect the predictions. In contrast, the choice of the reference category is an issue for the random-effects models, since the multivariate random-effects structure would have to be constrained to ensure consistency. We have run a small simulation study and it shows that the random-effects models produce predictions that change considerably when the reference category changes. On average it changes by 6% and for some instances up to 39%. Using the expectile regression models there are no differences in predictions changing the reference category.

5.1. Bootstrapped estimates of mean squared error

A bootstrap-based method for estimating the MSE of estimator $\hat{p}_{jg}^{ME}$ can be implemented following Chambers et al. (2016). This is based on the random-effects block (REB) bootstrap of Chambers and Chandra (2013). The method is a robust alternative to the parametric bootstrap for clustered data. The REB bootstrap is free of both the distribution and the independence assumptions of the parametric bootstrap and is consistent when the mixed model assumption is valid. In particular, it preserves area effects by bootstrap resampling within areas. Here we adapt this procedure to estimate the distribution of the predictor based on expectile regression models. The steps in the REB bootstrap are as follows.

Step 1: Calculate $J \times (G-1)$ vectors of marginal residuals $\boldsymbol{r}_{jg}^{E} = (r_{ijg}^{E}) = \boldsymbol{x}_{ij}^{T}(\hat{\boldsymbol{\beta}}_{\hat{q}_{ijg}^{*}} - \hat{\boldsymbol{\beta}}_{g})$, $i = 1,\dots,n_j$, $j = 1,\dots,J$, $g = 1,\dots,G-1$, where $\hat{\boldsymbol{\beta}}_{\hat{q}_{ijg}^{*}}$ is the estimate of $\hat{\boldsymbol{\beta}}_{g}$ with intercept adjusted by $\log \frac{\hat{q}_{ijg}^{*}}{1-\hat{q}_{ijg}^{*}}$ as per equation (11). Rescale the elements of the vector $\boldsymbol{r}_{jg}^{E}$ so they have mean equal to 0.

Step 2: Construct the individual bootstrap errors for the N_j population units in area j for category g as $\boldsymbol{r}_{jg}^{E*} = (r_{ijg}^{E*}) = srswr(\boldsymbol{r}_{f(j)g}^{E}, N_j)$ where $f(j) = srswr(\{1,\dots,J\},1)$. Here we use the notation $srswr(A,m)$ to denote sampling with replacement m times from the set A.

Step 3: Generate a bootstrap population U^* of N independent bootstrap multinomial realisations made up of J areas with area j of size N_j, and with bootstrap multinomial realisation y_{ijg}^{*} in area j taking the value 1 with probability

$$\pi_{ijg}^{*} = \frac{\exp(\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}_g + r_{ijg}^{E*})}{1+\sum_{g=1}^{G-1}\exp(\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}_g + r_{ijg}^{E*})}, \quad g = 1,\dots,G-1 \tag{18}$$

$$\pi_{ijG}^{*} = \frac{1}{1+\sum_{g=1}^{G-1}\exp(\boldsymbol{x}_{ij}^{T}\boldsymbol{\beta}_g + r_{ijg}^{E*})}, \quad i = 1,\dots,N_j. \tag{19}$$

Step 4: Calculate the bootstrap population parameters p_{jg}^{*}, $j = 1,\dots,J$ and $g = 1,\dots,G$.

Step 5: Extract a sample s^* of size n from the bootstrap population U^* by using the same sample design as that used to obtain the original sample and calculate the bootstrap expectile predictor $\hat{p}_{jg}^{ME^*}$, $j = 1,\dots,J$ and $g = 1,\dots,G$.

Step 6: Repeat steps $2 - 5$ B times. In the bth bootstrap replication, let $p_{jg}^{*(b)}$, be the quantity of interest for area j and category g and let $\hat{p}_{jg}^{ME^*(b)}$ be its corresponding estimate.

Step 7: The REB bootstrap estimator of the MSE of $\hat{p}_{jg}^{ME}$ is

$$mse^{REB}(\hat{p}_{jg}^{ME}) = B^{-1}\sum_{b=1}^{B}(\hat{p}_{jg}^{ME^*(b)} - p_{jg}^{*(b)})^2. \tag{20}$$

The performance of the REB bootstrap has been evaluated in a simulation study in the next section.

6. Simulation study for multicategory data in SAE

The purpose of this simulation study is to assess how well multinomial expectile regression works in SAE of multicategory proportions and how it compares to the method proposed by Molina et al. (2007) that assume common random effects across categories and to that proposed by Saei and Taylor (2012) with unconstrained random effects. For simplicity, these will respectively be referred to as the constrained and unconstrained random-effects (RE) methods respectively. To highlight the benefits of modelling area-level unobserved heterogeneity by means of random effects or q-scores, we consider also SAE predictions based on a fixed effects multinomial logistic regression (FE). Moreover, the performance of the small area predictors based on multinomial M-quantile estimation presented in Section S.2 of the supplementary material is also considered. Altogether, the following six methods will be evaluated: (a) Direct estimator (sample proportion); (b) FE method; (c) Constrained RE method; (d) Unconstrained RE method; (e) Expectile method; (f) M-quantile method.

For the simulation study, a population of $N = 5,000$ is generated in $J = 50$ small areas each with population size $N_j = 100, j = 1, \ldots, J$. Values for the single explanatory variable x_{ij} are simulated from $\mathcal{U}(-1, j/4), j = 1, \ldots, J, i = 1, \ldots, N_j$, and $\boldsymbol{x}_{ij} = (1, x_{ij})$. A multicategory response variable with $G = 3$ categories was simulated from $Multinomial(n_i = 1, \boldsymbol{\pi}_{ij})$ where $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \pi_{ij2}, \pi_{ij3})$ and $\pi_{ijg} = \exp(\eta_{ijg})/(1 + \sum_{g=1}^{2}\exp(\eta_{ijg}))$ when $g = 1, 2$. For the reference category $g = 3$, $\pi_{ij1} = 1/(1 + \sum_{g=1}^{2}\exp(\eta_{ijg}))$, and $\eta_{ijg} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g + \gamma_{jg}$ where $\boldsymbol{\beta}_g$ is the column of $\boldsymbol{\beta}$ corresponding with category g, and $\boldsymbol{\beta} = \left(\begin{smallmatrix} 0 & -0.5 \\ 1 & -0.5 \end{smallmatrix}\right)$.

The small area effects γ_{jg} are simulated in two different ways. In one case they are in line with the assumptions in Molina et al. (2007) with γ_{j1} generated from $\mathcal{N}(0, 0.15)$ and $\gamma_{j1} = \gamma_{j2}$. In the other, based on the assumptions of Hartzel et al. (2001) random effects are simulated from a bivariate normal $MVN(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 0.15 & -0.05 \\ -0.05 & 0.15 \end{smallmatrix}\right)$. These settings for the variance components are close to those obtained as estimates in the application of the Molina et al. (2007) and Saei and Taylor (2012) models in the ILFS application with three categories. We also note that the expected values of the target variable are close to the prevalence of the labour force statuses with three categories (0.73, 0.17 and 0.10) in the application of Section 7.

Moreover, as the magnitude of area-specific heterogeneity is rather small using these variance components, we consider an alternative setting with values increased to 0.5 in the constrained scenario and to $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 0.5 & -0.2 \\ -0.2 & 0.5 \end{smallmatrix}\right)$ in the unconstrained one. We label this latter scenario as *large* variance components, as opposed to *small* for the previous one.

After a population of $N = 5,000$ is generated, samples of size equal to either $n_j = 10$ or $n_j = 20$ are drawn by simple random sampling from each area , yielding a total sample size of $n = 500$ and $n = 1,000$ respectively.

Moreover, a second scenario with samples perturbed by a combination of misclassification and measurement errors, denoted MM, is considered. Specifically, 1% of the overall sample are randomly selected and the x_{ij} values replaced with 20 and the corresponding $\boldsymbol{y}_{ij}$ values set to be from category $g = 3$. This category is highly improbable for large values of x_{ij}. This scenario is added to assess the performance of the different estimators when model assumptions are violated.

To sum up, the experimental factors we consider are: random effects generation (unconstrained, constrained), area-specific sample sizes (10, 20), variance components (small, large), presence of misclassification and measurement errors (no, yes). Each scenario is replicated $R = 1000$ times and the predicted small area proportions $\hat{p}_{jg}$ are estimated using the different estimating methods. The performance of the estimators are evaluated for each of the three categories within each of the 50 small areas with the bias, $R^{-1}\sum_{r=1}^{R}(\hat{p}_{jgr} - p_{jgr})$, and root MSE (RMSE), $\sqrt{R^{-1}\sum_{r=1}^{R}(\hat{p}_{jgr} - p_{jgr})^2}$, where $\hat{p}_{jgr}$ is the estimate of the proportion of category g in small area j at iteration r and p_{jgr} is the corresponding true value.

Results for all the scenario with *small* variance components are shown in Table 2, while those with *large* variance components can be found in section S3 of the supplementary material. As the data is simulated to fit the constrained RE model it is expected that it should perform well without misclassification and measurement error. However, in regard to both bias and RMSE for all models without misclassification and measurement error, all modelling methods perform very similarly. While there are no notable differences in bias, except for the M-quantile predictor, only on very close inspection can it be seen that the FE method generally performs the best in terms of RMSE, with the expectile method having the only instance where one of the three categories is slightly better than the FE method. In this scenario all methods, except the M-quantile method, as expected, have negligible bias, and generally perform comparably without misclassification and measurement error.

When misclassification and measurement error are introduced, larger differences among the methods emerge. The FE and constrained RE method are generally the most adversely affected by the presence of misclassification and measurement error in regard to bias. The M-quantile method performs best with respect to bias while the expectile and M-quantile methods generally have the lowest RMSE when misclassification and measurement error are present, but again the FE method performs quite well in comparison to constrained and unconstrained RE.

The results of the unconstrained $\boldsymbol{\Sigma}$ simulation are also shown in Table 2. In the scenarios with no misclassification and measurement error all the methods again have small and comparable biases, except the M-quantile method that shows larger bias values. As for the RMSE, the unconstrained RE, expectile and M-quantile are noticeably performing better than the other methods. This is expected for the unconstrained RE method since the simulated data have random effects for each category in concordance with this method. We also note that there is a substantial difference between the RMSE for the direct estimates compared to the five model-based estimates which is expected.

In the unconstrained $\boldsymbol{\Sigma}$ simulation, the unconstrained RE and expectile methods also perform well in the presence of misclassification and measurement error in terms of both bias and RMSE, but the M-quantile approach shows the best results especially in terms of efficiency. It

Table 2. Results for constrained and unconstrained $\mathbf{\Sigma}$ simulations: Bias and RMSE of three-category predicted small area proportions.

Median values of Bias×1000												
	Constrained $\mathbf{\Sigma}$ simulation						Unconstrained $\mathbf{\Sigma}$ simulation					
	No MM			MM			No MM			MM		
$n_j = 10$	g_1	g_2	g_3	g_1	g_2	g_3	g_1	g_2	g_3	g_1	g_2	g_3
Direct	0.6	-0.4	0.6	0.2	-1.3	1.1	0.1	-0.6	-0.2	-0.6	-0.6	-0.1
FE	0.2	-0.0	0.1	0.6	-32.5	31.5	0.5	-0.5	0.2	1.0	-33.1	31.7
Constrained RE	-0.0	-0.0	0.0	0.2	-32.0	31.5	0.3	-0.4	0.3	0.6	-32.6	31.8
Unconstrained RE	-0.0	-0.2	0.0	0.9	-22.6	21.2	0.4	-0.3	-0.0	0.3	-18.3	18.0
Expectile	0.6	-1.1	0.4	1.5	-28.1	26.9	0.9	-1.3	0.7	1.5	-28.0	26.8
M-quantile	-5.3	11.0	-5.8	-5.3	-1.1	8.1	-5.5	11.6	-6.0	-5.4	-0.8	7.3
$n_j = 20$												
Direct	0.1	-0.4	0.1	-0.1	-0.7	0.9	0.1	0.6	-0.5	-0.4	-0.3	-0.2
FE	-0.0	-0.1	0.1	0.3	-28.4	27.9	0.7	-0.2	-0.4	0.8	-28.8	27.6
Constrained RE	0.4	-0.2	0.1	0.2	-28.4	28.0	0.4	-0.2	-0.4	0.6	-28.2	27.6
Unconstrained RE	0.3	-0.3	0.1	1.3	-15.9	15.0	0.6	-0.1	-0.5	0.0	-12.8	12.6
Expectile	0.4	-0.8	0.3	0.9	-24.2	23.9	0.6	-0.6	-0.2	1.1	-24.0	23.4
M-quantile	-4.8	10.1	-5.5	-4.9	0.1	7.0	-4.6	10.8	-6.1	-4.8	0.4	6.3
Median values of RMSE×1000												
	Constrained $\mathbf{\Sigma}$ simulation						Unconstrained $\mathbf{\Sigma}$ simulation					
	No MM			MM			No MM			MM		
$n_j = 10$	g_1	g_2	g_3	g_1	g_2	g_3	g_1	g_2	g_3	g_1	g_2	g_3
Direct	102.0	120.7	79.7	102.0	120.7	79.8	101.9	122.5	80.7	101.9	122.5	80.7
FE	45.5	40.4	25.2	45.6	55.5	42.1	38.0	50.2	35.1	38.2	61.7	48.8
Constrained RE	45.7	40.2	25.4	46.1	54.8	42.2	39.9	49.8	35.8	40.5	61.4	49.3
Unconstrained RE	44.8	42.7	28.7	45.1	53.9	43.7	38.6	50.6	36.4	39.6	59.2	47.6
Expectile	44.4	40.7	26.5	44.8	53.5	41.6	39.4	48.3	34.1	40.1	57.3	45.7
M-quantile	45.2	42.8	27.5	45.6	43.8	31.9	40.7	50.7	35.1	41.2	50.1	37.6
$n_j = 20$												
Direct	69.7	82.5	51.1	69.7	82.6	51.1	68.7	80.9	53.6	68.7	80.9	53.6
FE	40.8	36.7	22.9	40.9	48.9	37.4	34.7	44.7	31.2	34.8	54.8	43.4
Constrained RE	39.2	35.7	22.6	39.4	47.2	37.1	35.7	43.9	32.0	35.8	53.9	43.8
Unconstrained RE	39.4	37.1	24.2	39.8	46.4	35.7	35.0	42.7	30.7	35.7	48.0	38.3
Expectile	38.3	35.7	23.2	38.4	45.4	35.2	34.5	41.7	29.7	34.7	49.1	38.6
M-quantile	38.9	37.7	24.1	39.1	37.5	26.5	35.6	43.8	30.4	35.8	41.9	31.6

Table 3. Computation time in seconds for one iteration of the constrained and unconstrained $\boldsymbol{\Sigma}$ simulations.

Time (s)	Constrained RE	Unconstrained RE	Expectile
Constrained $\boldsymbol{\Sigma}$	10.50	12.95	4.53
Unconstrained $\boldsymbol{\Sigma}$	9.10	17.08	4.65

is evident that in some iterations the constrained RE and expectile methods return very high bias when misclassification and measurement error are present. In general this shows that the unconstrained RE, but especially expectile and M-quantile methods are both viable options in regard to performance in bias and RMSE when $\boldsymbol{\Sigma}$ is unconstrained.

The good performance of the FE, especially for constrained $\boldsymbol{\Sigma}$ simulations, can be explained by the small values of the variance components in this setting, even if they are realistic as proved in the ILFS application. Actually, if we consider results with *large* variance components (results are reported in Section S.3 of the supplementary material) we find patterns similar to those displayed in table 2 but, as expected, results show that the efficiency of the FE decreases as the values of the variance components increase.

It is important to point out that in practice it might be difficult to determine whether $\boldsymbol{\Sigma}$ should be assumed to be constrained or not. Hence, with this uncertainty the expectile method provides a desirable choice, since it performs very close to the estimator that is most suitable under both conditions. Both the RE and expectile methods assume a multinomial distribution, but the expectile method has no additional distributional assumptions, whereas the RE methods must impose an assumption on the distribution of the random effects and whether they are univariate (constrained) or multivariate (unconstrained).

The computation times for the models are also assessed based on the scenarios without misclassification and measurement error. The estimation time refers to optimised R codes on a laptop with a 2.6 GHz Intel Core i7 and 32 Gb RAM. There are considerable differences in computation times between methods, as shown in Table 3 that shows the time in seconds to compute one iteration for the four scenarios (two sample sizes, and both the constrained and unconstrained $\boldsymbol{\Sigma}$ simulations). Hence in practice the computation time would be approximately a quarter of these times, if using a similar data set. The expectile method was the fastest and took less than 5 seconds to carry out one iteration in both simulations. The constrained RE method was the second fastest and took over 9 seconds for both simulations. This left the unconstrained RE method as the slowest method with 17 seconds that is 4 times longer than the expectile method. Hence the expectile approach turns out to be much faster than the constrained RE and unconstrained RE approaches suggesting it is preferable with larger datasets or when there are time constraints. Note that the starting points for the estimation of constrained RE and unconstrained RE were selected fairly close to the true values of the parameters used in the simulation experiments. For this reason we have also compared the computational time of the three methods in the application of Section 7.

To assess the performance of the bootstrap MSE estimator proposed in the previous section we use the same simulation scenario (variance components) with $J = 50$ and sample sizes $n_j = 10$, but with small differences. The small area effects γ_{jg} are generated as in the scenario simulated for the evaluation of the performance of the small area estimators: the first approach is based on Molina et al. (2007) where γ_{j1} is simulated from $\mathcal{N}(0, 0.15)$ and then $\gamma_{j1} = \gamma_{j2}$. The second approach based on Hartzel et al. (2001) where the random effects are simulated

Table 4. Results for constrained and unconstrained $\mathbf{\Sigma}$ approaches in the two scenarios (no MM and MM): Relative Bias ($\times 100$) and empirical coverage rates for nominal 95% confidence intervals (CR95) of REB for the $G = 3$ categories.

Median values of Relative Bias×100						
	No MM			MM		
	$g = 1$	$g = 2$	$g = 3$	$g = 1$	$g = 2$	$g = 3$
constrained $\mathbf{\Sigma}$	-14.1	-1.8	-3.8	-7.1	11.6	17.5
unconstrained $\mathbf{\Sigma}$	-6.4	-12.0	-11.4	-5.9	14.7	17.5
Median values of CR95×100						
	No MM			MM		
	$g = 1$	$g = 2$	$g = 3$	$g = 1$	$g = 2$	$g = 3$
constrained $\mathbf{\Sigma}$	91.2	94.7	95.4	93.5	97.4	95.9
unconstrained $\mathbf{\Sigma}$	93.0	91.8	92.3	93.4	97.2	95.8

from a bivariate normal $MVN(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} = \left(\begin{smallmatrix} 0.15 & -0.05 \\ -0.05 & 0.15 \end{smallmatrix}\right)$.

In this case, $R = 100$ Monte Carlo populations were generated and for each generated population a simple random sample without replacement of size n_j was drawn from each area j, which was then used to calculate the small area predictors and their REB bootstrap MSE estimators. The performance of these MSE estimators for each scenario is presented in Table 4 where we show the medians of the area-specific relative bias and of the empirical coverage rates ($\times 100$) for nominal 95% confidence intervals (CR95). In this case, the intervals were defined by the small area estimates plus or minus twice the value of the square root of MSE estimators of equation (20). Examination of the results in Table 4 shows that both MSE estimation method tends to be biased low, especially in the unconstrained case, but all generate nominal 95% confidence intervals with acceptable coverage.

Simulation experiments were also replicated with different variance-covariance matrices and the results show that the performance of the MSE estimators improve when the variance and/or the covariance values decrease. To be concise, they are not reported here, but are available from the authors upon request. A conditional MSE estimator based on the linearisation approach that was set out in Chambers et al. (2016) could be developed and compared with the proposed REB bootstrap estimator. This is an aim for future research.

7. Application to Italian labour force data

The aim of this section is to present estimates of proportions of working age people in various labour market statuses at the LLMA level, considering either the classification into three and six categories as illustrated in section 2.

Small area estimates for the three categories classification are computed for the constrained random effects (RE), unconstrained RE and multinomial expectile methods, all using age-sex and number of registered unemployed individuals as explanatory variables. The multinomial expectile estimation method is in this case compared in terms of estimates and computation times with the methods by Molina et al. (2007) with common category or constrained random effects and Saei and Taylor (2012) with unconstrained random effects to estimate the proportions of employed, unemployed and inactive in each LLMA in Italy. For the six categories classification of labour market statuses the comparison is limited to the Molina et al. (2007) method as its extension to more than three categories is straightforward because the model is based on common random effects across categories. In contrast, the extension of Saei and Taylor (2012) is complex. In particular, as this model uses unconstrained random effects, in the case of six response categories the model requires the estimation of five different random effects for each small area and fifteen variance components.

To assess the quality of predictions, we used a set of diagnostic tools based on the requirement that model-based small area estimates should be coherent with, in the sense of being close to, the corresponding direct estimates, albeit more precise. Figure 3 shows the differences between the unweighted direct estimates and the model-based estimates of the total number of unemployed for LLMA in the three categories scenario. We report these values for the unemployed category because it represents the most difficult category to predict with the expected proportion of unemployed across areas approximately equal to 0.1. We note that the multinomial expectile-based estimates show high correlation with the direct estimates, with an average coefficient of 0.79. The corresponding correlations between the direct estimates and the estimates obtained by Molina et al. (2007) and Saei and Taylor (2012) are lower.

It can be noted that, none of the three predictors being considered make use of sampling weights so, although converging to the (unweighted) direct estimator, they are not design consistent unless the design is self-weighting within areas. The extension of the expectile regression model for multicategory to include sampling weights and the derivation of a design consistent multinomial expectile-based estimator is an area of current research.

To assess the potential gain in precision we obtain by using the proposed expectile approach, we compare in Figure 4 the empirical cumulative density functions (ECDFs) of the estimated coefficients of variation (CV) of both estimators for the three categories (employed, inactive, unemployed). Consistently with the computation of direct estimates, also the CVs are computed using a standard, unweighted formula. In each row, the first panel uses CVs from all areas, while the second (third) one focuses on small areas with sample sizes smaller (larger) than 100. In the first panel, the ECDFs corresponding to expectile-based approach almost always dominates the one for the direct estimates, highlighting that CV values for the former approach are lower than those estimated with the latter. This is more evident in the second panel related to the presence of some areas with a small sample size ($n_i < 100$). The plots of the empirical cumulative density functions (ECDFs) of the estimated coefficients of variation (CV) of the predictor based on the constrained random effect model (Molina et al., 2007) are reported in the Supplementary Material. The squared root of the MSE of the predictor based on the constrained random effect model for each area has been computed using the parametric bootstrap proposed in Molina et al. (2007). The ECDFs corresponding to expectile-based approach in Figure 4 almost always dominate the one for the estimates obtained by the constrained random effect model reported in Figure S.3 confirming the results of the simulation experiments.

With respect to the estimation exercise with six categories, Figure 5 maps the estimated levels of employed, unemployed a, unemployed b, unemployed c, reserve workers, inactive for LLMAs in Italy in 2012 by using expectile approach. The patterns of employment, unemployment and inactive produced by the proposed approach are consistent with those obtained by all the other methods (Marino et al., 2019). As expected, relatively larger values for unemployment incidences are mainly located in the south of Italy and in the islands. These results confirm the existence of the so-called North-South divide in Italy. Although not clearly apparent from the maps (Figure 3) results confirm that the regional divide is largest for unemployment c (those actively searching for their first job) and reserve workers; both facts are in line with the

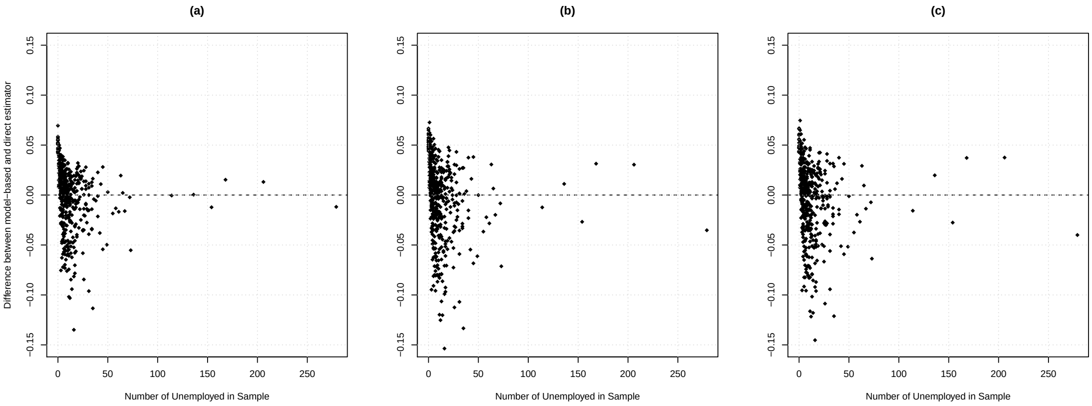

Figure 3. Difference values between direct estimates and of model-based estimates of the total number of unemployed for LLMA in Italy in 2012: (a) multinomial expectile-based estimates, (b) constrained RE-based estimates, (c) unconstrained RE-based estimates.

difficult conditions of the labour market in Italian Southern regions, especially for the young (Mussida and Sciulli, 2018). From Figure 5 we also note how the patterns of the different rates are not spatially smooth with large variations even between neighbouring LLMAs. This is in line with insights from the economic literature: differences between *core* metropolitan centres and peripheral areas can be as big as long range regional divides (Ghignoni and Verashchagina, 2016).

We note that the estimates obtained with the expectile approach appear to closely correlated with the direct estimates, with the average correlation between the two sets of estimates being 0.78. The corresponding correlation between the direct estimates and the constrained RE estimates is 0.40. The gain in terms of correlation obtained by the expectile method is high especially in the categories with small proportions such as for unemployed $\underline{b}$, unemployed $\underline{c}$, reserve workers. This is confirmed in Figure 6 where the differences of the direct and of expectile-based estimates are reported.

To evaluate the potential gains in precision from using expectile-based estimates instead of the direct estimates, we examine the distribution of the ratios of the estimated CVs of the direct and the expectile-based estimates for the ILFS data for six categories. A value greater than 1 for this ratio indicates that the estimated CV of the model-based estimate is smaller than that of the direct estimate. Figure 7 shows the relationship between these ratios and the number of employed, unemployed $\underline{a}$, unemployed $\underline{b}$, unemployed $\underline{c}$, reserve workers, inactive in the ILFS sample in each LLMA. Figure 7 shows that the estimated CVs of the expectile-based estimates are generally much lower than those of the direct estimates and the differences between the direct and the multinomial expectile-based estimates become more evident as the number of observations for each category in the sample decreases.

The time to estimation refers to optimised R codes on a laptop with a 2.6 GHz Intel Core i7 and 32 Gb RAM. The expectile method is the fastest as it takes less than 5 seconds to produce small area estimates, whereas the Molina et al. (2007) and Saei and Taylor (2012) method take 105.3 and 1171.1 seconds, respectively. The difference in computation time increases when the methods are used to estimate the proportions of workforce participation for six categories.

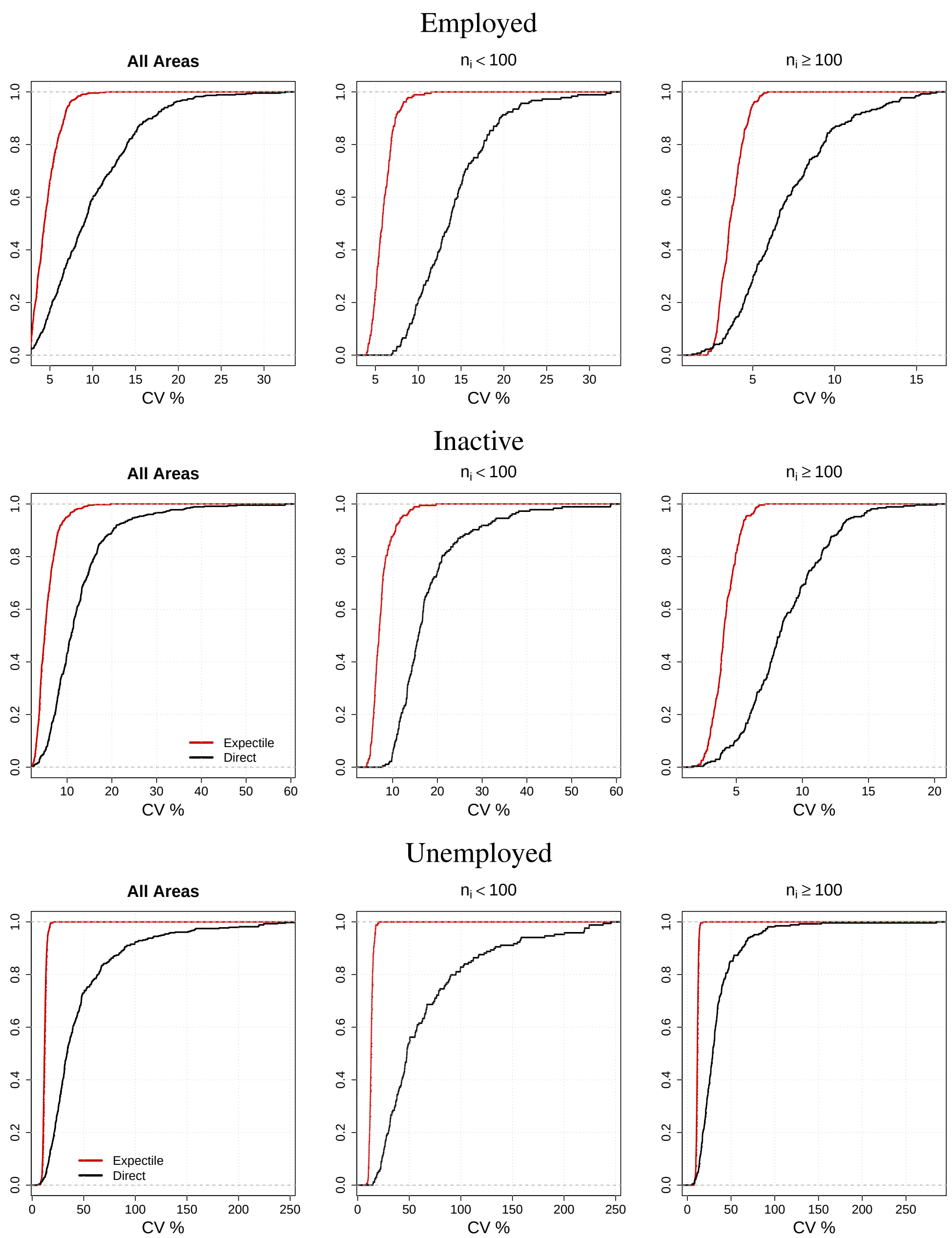

Figure 4. CVs empirical cumulative density functions for the multinomial expectile-based estimator and the direct estimator for the three categories (employed, inactive, unemployed).

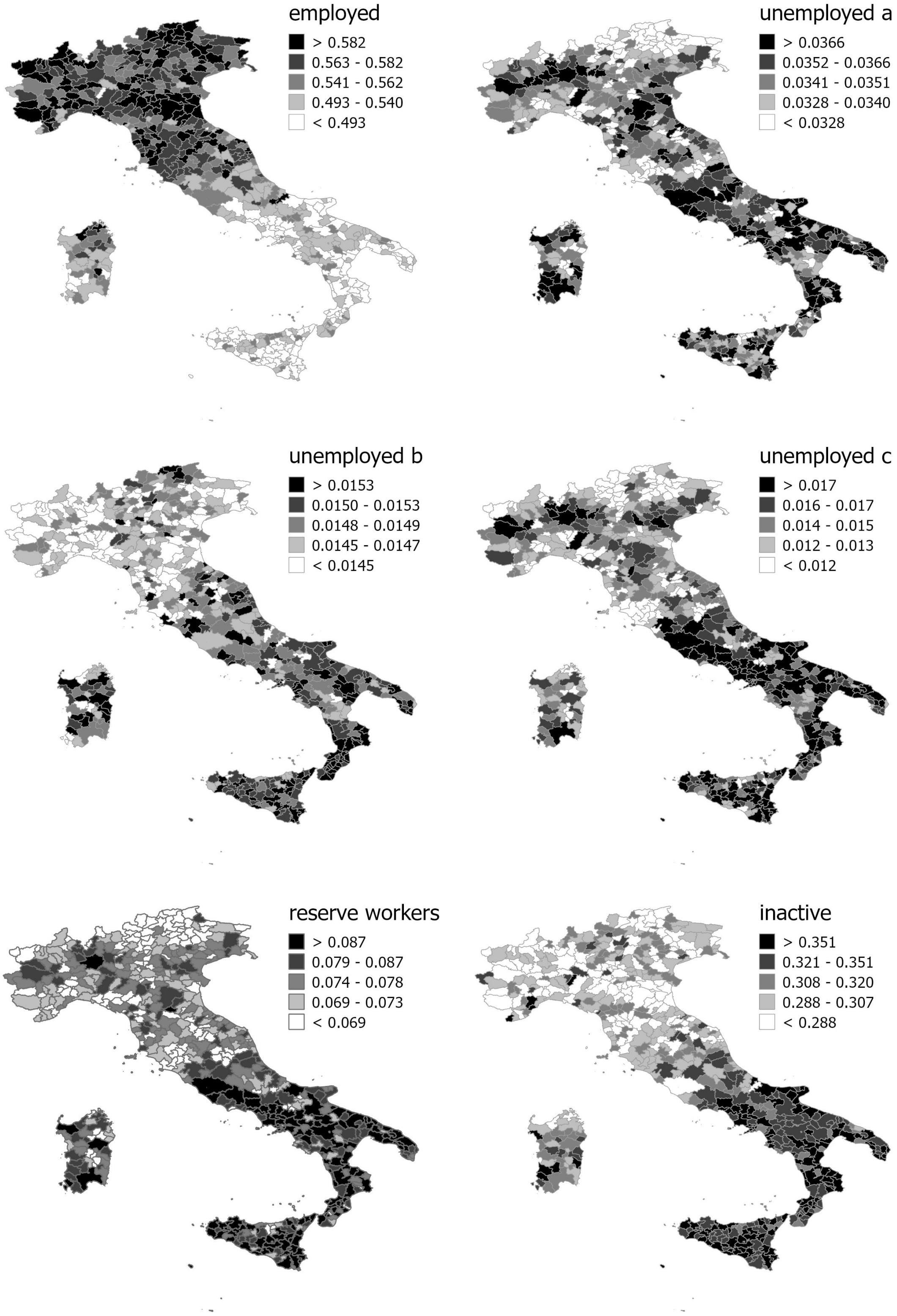

Figure 5. Maps of estimated levels of employed, unemployed a (in active job search, previous job experiences, formerly employed), unemployed b (in active job search, previous job experiences, formerly inactive), unemployed c (in active search of their first job), reserve workers, inactive for each LLMA in Italy in 2012.

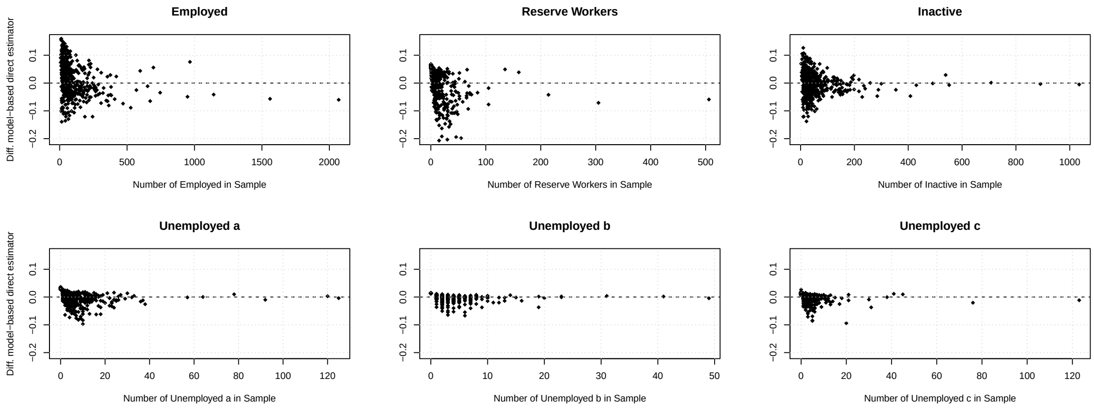

Figure 6. Difference between direct and expectile-based proportion estimates of working-age people in the six-category labour market statuses for LLMA in Italy in 2012.

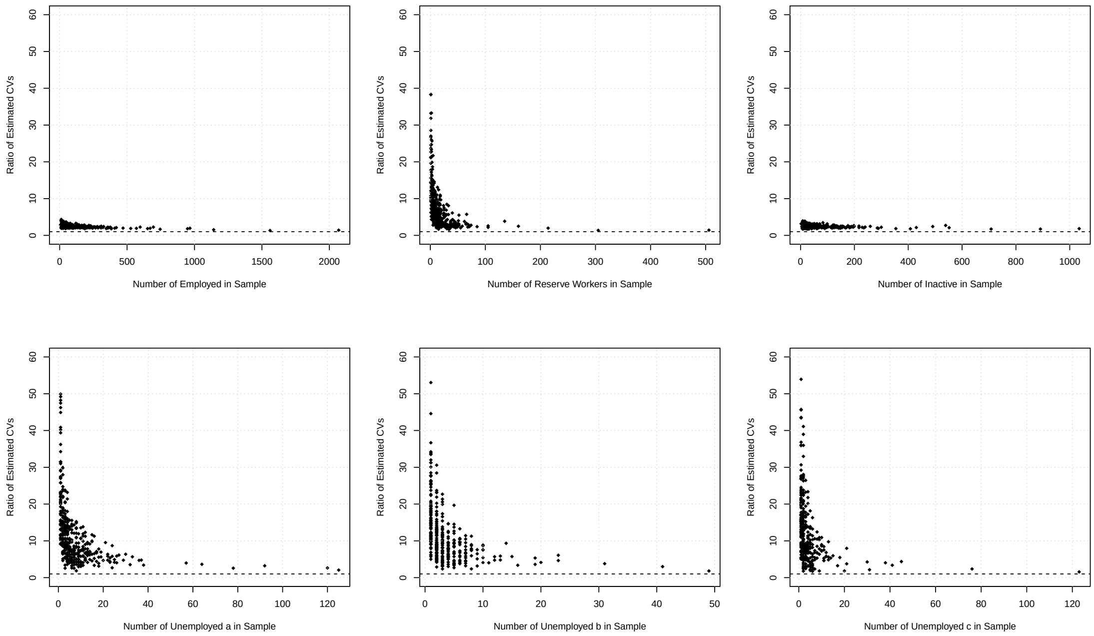

Figure 7. Ratio of estimated CVs for direct estimates and expectile-based proportion estimates of working-age people in the six-category labour market statuses for LLMA in Italy in 2012.

The expectile method remains the fastest taking only 7.1 seconds while the Molina et al. (2007) method needs approximately 873 seconds to complete small area estimation, that is a 125 times longer operation time than that needed by the expectile method. In this application the expectile approach is much faster than the constrained RE and unconstrained RE approaches suggesting that, when working with larger data sets this advantage can become very relevant from a practical point of view.

According to the simulation results in Section 6 of the paper, the expectile approach is a potentially useful approach as: (i) it performs better than the other considered predictors in terms of bias and efficiency; (ii) it shows that estimates are generally consistent with the direct estimates and more accurate; and (iii) it decreases the computational complexity of the estimators under the constrained and unconstrained RE.

8. Final remarks

Constructing SAE models for a multicategory response has proven to be a challenge, with major shortfalls with pre-established methods based on mixed models. However, by exploiting a simple, yet elegant, relationship between the binary expectile and the probability, a SAE approach based on expectile regression can be constructed. Such an approach provides a method that is straightforward to generalise to multicategory responses. This provides the first quantile-like estimator for multicategory response data known to the authors.

A new approach to binary q-scores was also introduced which we argue has better properties than those defined by Chambers et al. (2016). And again this q-score method can easily be extended to the multicategory case. Therefore we suggest that the proposed q-score derivation be used for not only the proposed models in this article but any M-quantile model with binary or multicategory responses.

This novel approach to SAE using multinomial expectile regression and q-scores is the first of its kind. Consequentially, M-quantile approaches to SAE can now be applied to multicategory data. The simulation studies showed that the multinomial expectile SAE models perform relatively well compared to pre-existing methods that utilise multinomial random-effects models. Even under ideal conditions for these other models, the expectile method performed comparably or only slightly worse. More importantly, the simulation studies showed that the expectile method can perform better than random-effects models when the assumptions of multinomial random-effects models are violated, when the random-effects structure is unknown, and additionally requires considerably less computation time. Thus making this proposed expectile method an effective approach to SAE with multicategory response data.

The bootstrap method proposed for estimating the MSE of the expectile predictor provides acceptable bias and coverage performance in our simulations, but it is computationally slow. For this reason, a MSE estimator based on a linearisation approach is an area of current research. We also suggest extending the proposed expectile model to ordinal responses, perhaps under an ordered logit model with proportional odds this may be feasible. However it would require a new derivation for ordinal q-scores that account for the ordinal rather than nominal categorisation.

A limitation of our proposal is that it does not make use of survey weights. This choice is consistent with the one underlying the competing methods we consider and allows us to illustrate our ideas in a very simple form. Nonetheless, for official statistics agencies the property of design consistency is fundamental because it is a general purpose form of protection against model failures, as it guarantees that, at least for large domains, estimates make sense even if the assumed model fails. The expectile-based estimators using multinomial expectile models do not make use of survey weights and, in general, the derived estimator is not design consistent unless the sampling design is self-weighting within area. In the presence of complex and potentially non-ignorable (informative) sampling designs, ignoring the sampling weights could lead to biased estimates (Pfeffermann, 1993). Extensions of the expectile approach to account for this e.g. following Fabrizi et al. (2014) is an area of current research.

Finally, we identify the potential for the multinomial expectile SAE models for analysing political polling data. The proposed method is similar to models using multilevel regression and post-stratification and a comparison between the two methods would be interesting, specifically in applications to national elections with sub-regional constituencies.

Acknowledgment: The authors thank Dr Natalia Tejedor Garavito of the WorldPop group at the University of Southampton for her expert input mapping the small area estimates.

References

Barbieri, P. and Scherer, S. (2009) Labour market flexibilization and its consequences in Italy. *European sociological review*, **25**, 677–692.

Bianchi, A., Fabrizi, E., Salvati, N. and Tzavidis, N. (2018) Estimation and testing in M-quantile regression with applications to small area estimation. *International Statistical Review*, **86**, 541–570.

Brandolini, A., Cipollone, P. and Viviano, E. (2006) Does the ILO definition capture all unemployment? *Journal of the European Economic Association*, **4**, 153–179.

Breckling, J. and Chambers, R. (1988) M-quantiles. *Biometrika*, **75**, 761–771. URL: `http://www.jstor.org/stable/2336317`.

Cantoni, E. and Ronchetti, E. (2001) Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022–1030. URL: `http://www.tandfonline.com/doi/abs/10.1198/016214501753209004`.

Chambers, R. and Chandra, H. (2013) A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, **22**, 452–470.

Chambers, R., Dreassi, E. and Salvati, N. (2014) Disease mapping via negative binomial regression M-quantiles. *Statistics in Medicine*, **33**, 4805–4824. URL: `http://dx.doi.org/10.1002/sim.6256`.

Chambers, R., Salvati, N. and Tzavidis, N. (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **179**, 453–479. URL: `http://dx.doi.org/10.1111/rssa.12123`.

Chambers, R. and Tzavidis, N. (2006) M-quantile models for small area estimation. *Biometrika*, **93**, 255–268. URL: `http://www.jstor.org/stable/20441279`.

Dawber, J. and Chambers, R. (2019) Modelling group heterogeneity for small area estimation using M-quantiles. *International Statistical Review*, **87**, S50–S63. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12284.

De Vitiis, C., Di Consiglio, L. and Falorsi, S. (2018) Studio del disegno campionario per la nuova rilevazione continua sulle forze di lavoro. *Tech. rep.* URL: https://www.istat.it/it/files//2018/07/200506.pdf.

Dunn, P. K. and Smyth, G. K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Efron, B. (1992) Poisson overdispersion estimates based on the method of asymmetric maximum likelihood. *Journal of the American Statistical Association*, **87**, 98–107. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475180.

Eurostat (2013) Handbook on precision requirements and variance estimation for ESS households surveys. URL: dx.doi.org/10.2785/13579.

— (2022) Labour force survey in the eu, efta and candidate countries. URL: https://ec.europa.eu/eurostat/documents/7870049/14295950/KS-FT-22-001-EN-N.pdf

Fabrizi, E., Salvati, N., Pratesi, M. and Tzavidis, N. (2014) Outlier robust model-assisted small area estimation. *Biometrical Journal*, **56**, 157–175. URL: http://dx.doi.org/10.1002/bimj.201200095.

Ghignoni, E. and Verashchagina, A. (2016) Added worker effect during the great recession: evidence from Italy. *International Journal of Manpower*, **37**, 1264–1285.

Hartzel, J., Agresti, A. and Caffo, B. (2001) Multinomial logit random effects models. *Statistical Modelling*, **1**, 81–102. URL: http://dx.doi.org/10.1177/1471082X0100100201.

James, J. (2017) MM algorithm for general mixed multinomial logit models. *Journal of Applied Econometrics*, **32**, 841–857.

Koenker, R. and Bassett, Gilbert, J. (1978) Regression quantiles. *Econometrica*, **46**, 33–50. URL: http://www.jstor.org/stable/1913643.

López-Vizcaíno, E., Lombardía, M. J. and Morales, D. (2013) Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, **13**, 153–178. URL: http://dx.doi.org/10.1177/1471082X13478873.

— (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society, Series A*, **178**, 535–565.

Manski, C. F. and Thompson, T. (1989) Estimation of best predictors of binary response. *Journal of Econometrics*, **40**, 97 – 123. URL: http://www.sciencedirect.com/science/article/pii/0304407689900328.

Marino, M., Ranalli, M., Salvati, N. and Alfó, M. (2019) Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, **13**, 1166–1197.

Molina, I., Saei, A. and José Lombardía, M. (2007) Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**, 975–1000. URL: http://dx.doi.org/10.1111/j.1467-985X.2007.00493.x.

Mussida, C. and Sciulli, D. (2018) Labour market transitions in Italy: The case of the NEET. In *European Youth Labour Markets*, 125–142. Springer.

Newey, W. K. and Powell, J. L. (1987) Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819–847. URL: http://www.jstor.org/stable/1911031.

Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 317–337.

— (2013) New important developments in small area estimation. *Statistical Science*, **28**, 40–68.

Prasad, N. G. N. and Rao, J. N. K. (1990) The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10475320.

Rao, J. N. and Molina, I. (2015) *Small area estimation*. John Wiley & Sons.

Saei, A. and Chambers, R. (2003) Small area estimation under linear and generalized linear mixed models with time and area effects. *Tech. rep.* URL: http://eprints.soton.ac.uk/id/eprint/8165.

Saei, A. and Taylor, A. (2012) Labour force status estimates under a bivariate random components model. *Journal of the Indian Society of Agricultural Statistics*, **66**, 187–201.

Scealy, J. (2010) Small area estimation using a multinomial logit mixed model with category specific random effects. *Australian Bureau of Statistics*. URL: https://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.029.

Sobotka, F., Kauermann, G., Waltrup, L. S. and Kneib, T. (2013) On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, **23**, 135 – 148.

Sobotka, F. and Kneib, T. (2012) Geoadditive expectile regression. *Computational Statistics & Data Analysis*, **56**, 755 – 767. URL: http://www.sciencedirect.com/science/article/pii/S0167947310004433.

Tabatabai, M., Li, H., Eby, W., Kengwoung-Keumo, J., Manne, U., Bae, S., Fouad, M. and Singh, K. (2014) Robust logistic and probit methods for binary and multinomial regression. *Journal of Biometrics & Biostatistics*, **5**.

Tzavidis, N., Ranalli, M., Salvati, N., Dreassi, E. and Chambers, R. (2015) Robust small area prediction for counts. *Statistical Methods in Medical Research*, **24**, 373–395.

Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian, M. (2014) The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, **23**, 42–59. URL: https://doi.org/10.1177/0962280212445834. PMID: 22523185.

Yang, Y. and Zou, H. (2015) Nonparametric multiple expectile regression via er-boost. *Journal of Statistical Computation and Simulation*, **85**, 1442–1458. URL: https://doi.org/10.1080/00949655.2013.876024.