Estimating Literacy Levels at a Detailed Regional Level: an Application Using Dutch Data

Bijlsma, Ineke; Brakel, Jan van den; Velden, Rolf van der; Allen, Jim

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 251–274, http://dx.doi.org/10.2478/JOS-2020-0014

Estimating Literacy Levels at a Detailed Regional Level: an Application Using Dutch Data

Ineke Bijlsma[1], *Jan van den Brakel*[1], *Rolf van der Velden*[1], *and Jim Allen*[1]

Policy measures to combat low literacy are often targeted at municipalities or regions with low levels of literacy. However, current surveys on literacy do not contain enough observations at this level to allow for reliable estimates when using only direct estimation techniques. To provide more reliable results at a detailed regional level, alternative methods must be used.

The aim of this article is to obtain literacy estimates at the municipality level using model-based small area estimation techniques in a hierarchical Bayesian framework. To do so, we link Dutch Labour Force Survey data to the most recent literacy survey available, that of the Programme for the International Assessment of Adult Competencies (PIAAC). We estimate the average literacy score, as well as the percentage of people with a low literacy level. Variance estimators for our small area predictions explicitly account for the imputation uncertainty in the PIAAC estimates. The proposed estimation method improves the precision of the area estimates, making it possible to break down the national figures by municipality.

Key words: Literacy; basic skills; municipality; region; small area estimation.

1. Introduction

Research shows that cognitive skills play an important role in individual life chances (Coulombe and Tremblay 2007; Hanushek and Woessmann 2008, 2011). People with high skill proficiency levels earn more, are more often employed, and generally face fewer economic disadvantages. Moreover, they are more often engaged in civic and social activities (Organisation for Economic Co-operation and Development (OECD) 2013a).

Generally, the skill levels in the Netherlands are among the highest in the world. In the Programme for the International Assessment of Adult Competencies (PIAAC) of 2012, the Netherlands ranked third in literacy, just behind Japan and Finland. Even so, there are still around 1.3 million people (11.9%) in the population of 16- to 65-year-olds who do not have the literacy skills necessary to function well in society (Buisman et al. 2013). The cost

[1] Maastricht University, ROA, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Emails: i.bijlsma@maastrichtuniversity.nl, j.vandenbrakel@maastrichtuniversity.nl, r.vandervelden@maastrichtuniversity.nl, and j.allen@maastrichtuniversity.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. This study was made possible by a grant from the Reading and Writing Foundation, which other than funding had no role in this research. The authors are grateful to the unknown referees and the associate editor for reading and commenting on a former draft of this article. We also thank Olivier Marie, Stefa Hirsch, participants of the SAE 2016 conference (17–19 August 2016, Maastricht, The Netherlands) and the third PIAAC International Conference (6–8 November 2016, Madrid, Spain) for useful comments and suggestions.

of low literacy in the Netherlands is estimated to be some 550 million euros per year (PriceWaterhouseCoopers 2013).

As policy aimed at increasing literacy levels is often decentralized, local and regional governments need reliable data on the literacy levels in their particular municipality or region. However, this is usually not available, since most literacy surveys such as PIAAC focus on the national level. To illustrate the problem: The Dutch PIAAC sample contains about 5,000 observations. However, the Netherlands comprises 415 municipalities, and only the four biggest cities in the Netherlands have more than 90 observations in the PIAAC sample, while roughly half of the municipalities have fewer than 20 observations. The use of direct estimators would result in unacceptably large design variances. To increase the precision of municipal estimates, model-based small area estimation (SAE) techniques are applied in this article. These methods assume an explicit statistical model to increase the effective sample size of each separate area.

The basic idea of this regression method is that we assume that our dependent variable, literacy, is closely linked to personal characteristics such as age, gender, education, and labor status, which are also available in large auxiliary data sets. We also make the necessary assumption that the way these characteristics are linked is similar at both the national and detailed regional levels. Therefore, with detailed information for these characteristics at the regional level, it is possible to make more accurate model-based literacy predictions per municipality: a synthetic estimate. Unexplained variation between the areas is modeled with a random component in a multilevel model.

Model-based small area predictors can be expressed as the weighted average between the direct estimates based on PIAAC data and the aforementioned synthetic estimates, where the weights are based on the accuracy measures of the two estimators. If the underlying assumptions hold, this allows us to greatly reduce the variance of the estimates while introducing only limited bias to the estimates.

SAE techniques are widely applied in social and economic sciences to produce reliable statistical information in detailed breakdowns. Taylor et al. (2016) use synthetic estimates to predict expected levels of limiting long-term illnesses. The World Bank (2002) applies a synthetic estimation procedure proposed by Elbers et al. (2003) to estimate poverty and income inequality in developing countries. The U.S. Census Bureau applies an SAE approach based on the Fay and Herriot (1979) model to estimate income at low regional levels. These estimates are used to determine fund allocations to local government units. The National Research Council (2000) also used the method of Fay-Herriot to produce county estimates of poor school-aged children in the United States for the allocation of supporting funds. Statistics Netherlands applies time series SAE methods to calculate official monthly unemployment Figures (Van den Brakel and Krieg 2015). Finally, Tighe et al. (2010) applied hierarchical Bayesian models to obtain reliable estimates for low-incidence groups defined by religion or ethnicity not included in the U.S. Census Bureau.

To the best of our knowledge, SAE techniques in the context of literacy skills have only been applied sparsely, and take a quite different approach than the one we present here. Schmid et al. (2017) use self-assessed literacy from the Demographic and Health Survey in combination with mobile phone data to estimate literacy in Senegal, as a way to use alternative data sources instead of requiring statistics on socio-demographic indicators. Gibson and Hewson (2012) use UK census data and SAE modeling to obtain synthetic estimates of literacy levels in detailed geographical areas. Yamamoto (2014) adopts a similar approach to produce synthetic estimates for the different Canadian provinces.

While these two papers focus on synthetic estimates only, the contribution of this article is the application of SAE techniques to estimate municipalities' literacy levels that are a weighted average of direct and synthetic estimates, with the weights based on the uncertainty measures of both estimates. This approach has the advantage that, in large municipalities with relatively large sample sizes, the direct estimates make a relatively large contribution to the final estimate, whereas in small municipalities, the final estimate is dominated by the synthetic estimator. The PIAAC data setup presents a number of challenges that prevent straightforward estimations. Addressing these challenges is novel in the application of SAE techniques. Respondents were randomly assigned to (parts of) the literacy tests. This requires imputation techniques to account for missing observations. Moreover, the PIAAC tests follow an adaptive design, so that respondents are assigned items that are close to their expected proficiency levels, based on the scores of previous questions. The model follows an item response theory (IRT) approach, which assumes that the scores on the tests are based on a latent construct that cannot be measured directly. Instead, for each respondent, ten plausible values are calculated and several replicate weights are constructed, which can be seen as a form of multiple imputation. This approach allows for the construction of point estimates as well as variance estimates for literacy. We use both a unit-level model (Battese et al. 1988) and an area-level model (Fay and Herriot 1979) and detail how to incorporate this structure into our SAE approach.

Our article is organized as follows. Section 2 covers the definition of literacy, as well as the data description. Section 3 details the techniques of the small area predictors for this application. Section 4 presents the selected models and their fit. Section 5 evaluates the model and presents robustness checks. Section 6 reports the results of our analysis and Section 7 concludes the article.

2. Definition of Literacy and Data Description

2.1. PIAAC – Primary Data Source

The data set we are using is the 2012 PIAAC survey. It is designed to map skills and competencies in developed countries, measuring the numeracy, literacy, and problem solving skills of adults. In addition, it collects a range of information on how often respondents use these skills.

Literacy in PIAAC is defined as "the ability to understand, evaluate, use and engage with written texts to participate in society, to achieve one's goals, and develop one's knowledge and potential" (OECD 2013a, 59). It does not include the ability to write or produce texts, but focuses on the ability of an individual to interact with written text. It is this definition that will be used throughout the article.

Data collection in the Netherlands took place from August 1, 2011, to March 31, 2012, and was undertaken in the respondents' homes. The target population was between 16 and 65 years of age, residing in the country at the time the data were collected. For the Netherlands, 5,170 respondents were randomly selected by one-stage stratified simple random sampling without replacement from the Dutch population register. Strata were formed by municipalities. The sample weights are based on the sampling design. The response rate, as defined by complete cases divided by eligible cases, was 51% (OECD 2013b).

The PIAAC survey used specific data collection modes and procedures to measure skill proficiency levels (for details, see OECD 2013c). For the literacy domain, the questions differed in content, cognitive strategies, and context. A multistage adaptive design was used between the items and an algorithm determined the next item depending on the responses. This survey design was such that different groups of respondents were routed to items with potentially various degrees of difficulty, disallowing direct comparisons between the respondents' test scores. Therefore, the item responses were first fitted to an IRT model. After item calibration, the IRT model was combined with a latent regression model using information from the background questionnaire in a population model to further improve accuracy. From this step, 10 plausible values were drawn on a scale from zero to 500. Lastly, a replication approach (Johnson and Rust 1992) was used to estimate the sampling variability as well as the imputation variance associated with the plausible values. The percentage of respondents in the Netherlands who were unable to complete the questionnaire due to literacy-related issues is 2.3%; no proficiency scores were estimated for this group, but they were included in the weighting (OECD 2013b). The effect of list-wise deletion of these cases is therefore limited.

Variance estimation, taking into account the sample design, the selection process, the weighting adjustment, and the measurement error through imputation, is carried out using a replication approach. For the Netherlands, a paired jackknife estimator was used with 80 replicate weights. To take this survey design into account, we used the Stata module PIAACTOOLS of Pokropek and Jakubowski (2013). A detailed description of the construction of the variance term, as well as the above imputation, can be found in OECD (2013c).

Literacy scores are categorized at multiple levels based on the scoring range. Level 1 literacy starts at a score 176, and every 50 points above indicates an additional level, up to Level 5 (376 points or higher). At Level 1 (range 176–225), one can complete simple forms, understand basic vocabulary, and read continuous texts, but would have trouble making low-level inferences. For reference, Level 3 requires multiple steps to access the correct information and at Level 5 one can work with multiple, dense texts and conflicting information. These levels are described in full in OECD (2013b).

One straightforward method for describing the literacy levels in a region would be to look at the average test score for literacy. This is a good way of providing a quick snapshot of the literacy level. A limitation, however, is that it provides no further information as to how literacy levels are distributed within regions. Another measure would be to look at the proportion of low literates per area. We define someone as *low literate* when that individual has literacy Level 1 or below. This measure would be most important for policy making, as this group would benefit the most from policy interventions. A disadvantage of this measure is that information is lost due to its dichotomous nature. Taken together, both measures – the average score and the proportion of low literates – provide the best picture of the situation concerning literacy levels in a region.

The total number of respondents in PIAAC is 5,170, but for some respondents the municipality is unknown. We are left with 5,073 respondents, whose statistics are given below (see Table 1). The average score across respondents is in the lower half of Level 3 (276–325), with only about 12% at Level 1 or below (225 or below).

Table 1. Summary of the statistics of the target sample (PIAAC).

	Mean	St. Error	Lower Bound	Upper Bound
Average Score	283.94	0.68	282.61	285.27
% Low Literates	12.00	0.46	11.07	12.86

In Section 3, two different small area estimation models are applied. The area level model (Fay and Herriot 1979) use direct estimates for the target variable and their variances at the level of the areas as input for the model. The unit level model (Battese et al. 1988) use the observations of the sampling units as input for the model. Both models are multilevel models and need auxiliary information for the fixed effect part of the model. The area level model can only use auxiliary information that is aggregated at the level of the area (municipality). The unit level model can use both auxiliary information at the level of the sampling units (individuals) and auxiliary information aggregated at the level of the areas. As stated in the introduction, we are interested in both the average literacy score and the percentage of low literacy per municipality. We estimate the literacy score using the unit-level model and low literacy using the area-level model (dichotomous); we expand on the construction of the dependent variables under *Literacy Measures*.

2.2. Labor Force Survey (LFS) – Data Source for Auxiliary Information

SAE requires auxiliary data that include personal characteristics that are closely linked to literacy levels. The Dutch LFS's features (large sample sizes, good overlap in questions about personal characteristics) make it a good choice for auxiliary data.

In our selected timeframe, interviews for the LFS took place face to face and by phone. The weights are calculated in two steps using general regression estimators (Särndal et al. 1992). In the first step, design weights are derived from the sample design and account for differences in selection probabilities. In a second step, the design weights are calibrated to available auxiliary information for which the true population distributions are known from registrations to correct, at least partially, for selective nonresponse.

To ensure sufficient data from each area, we chose to include three years of LFS data: 2010, 2011 and 2012, that is, years close to the data collection period for PIAAC. We apply the same age restriction (between 16 and 65 years old) as in the PIAAC survey.

The LFS is based on a household sample. All household members aged 15 years and older are observed. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household. Households in which one or more of the selected persons do not respond for themselves or in a proxy interview are treated as non-responding households.

The total response and nonresponse numbers can be found in the *Methods and definitions* of the LFS data (Statistics Netherlands 2010; 2011; 2012), with a minimum response of roughly 63% of the approached households. This results in about 41,000 completely responding households on a yearly basis, and thus about 123,000 over three years (with a maximum of eight persons per household).

Since the LFS has a rotating panel design, people were asked multiple times to participate and thus are included multiple times. We weight these people over the number of samples within our selection, so that those who are covered multiple times in the data set are not oversampled. This leaves us with 309,000 unique respondents (with a rough average of 2.5 persons per household).

3. Small Area Estimation

Sample surveys are usually designed to meet minimum precision requirements for sample estimates at national level and at the level of planned domains using standard direct estimators. For other unplanned domains or subpopulations, the sample size is frequently too small to create reliable estimates based on direct estimators. Sample size is restricted by available resources and time and, in many surveys, it is too costly to sample a large number of individuals within each subpopulation of interest. In such cases, model-based inference methods from the literature on SAE can be considered as an alternative. SAE refers to estimation procedures that explicitly rely on a statistical model that increases the effective sample size of a particular domain with sample information from other domains (cross-sectional correlations) or preceding sampling periods (temporal correlations). The extent to which the precision of direct estimates is improved with these methods depends on the availability of auxiliary data contained in register data sets or large surveys, such as the LFS.

A large amount of SAE procedures are available in the literature. See Rao and Molina (2015) for a detailed overview, or Pfeffermann (2013) for a more summarized overview. In this article, we have chosen a multilevel modeling approach. The models are fitted in a *hierarchical Bayesian* (HB) framework. All models, including the model selection measures, were run using the fSAE function in the software program R, available via the hbsae package (Version 1.0, available in the Comprehensive R Archive Network; Boonstra 2015).

It is important to keep some things in mind when interpreting the results from SAE. In particular, model miss-specification can result in biased domain predictions. One important possible bias is due to the assumption that the relations between literacy and personal characteristics at the national level are the same at the regional level. While we do not expect the literacy model to have regional variation, violation of this assumption can lead to large differences between the regional estimations and the true regional literacy.

3.1. Literacy Measures

As stated earlier, we are interested in two measures of literacy per area: the average score and the percentage of low literates. In the first case, the dependent variable y is continuous per individual and area and we assume that y has a linear relation with the chosen covariates X. In this case, we use the basic unit-level model originally proposed by Battese et al. (1988), where the input variables for the model are individual measurements obtained from the sampling units. We go into more detail in the section below on the unit-level model.

In the second case regarding the percentage of low literates, the dependent variable is dichotomous at the individual level, since each plausible value will be binary, equal to one if the score is below the low-literacy cutoff point of 226 and zero otherwise. We decided to model the percentage of low literates with a basic area-level model, as originally proposed by Fay and Herriot (1979), as the *hbsae* package has no support for binary outcome variables that would be necessary for a unit-level model. In the next two sections, we elaborate both the area-level model and the unit-level model. Afterwards, we explain how we incorporated the PIAAC imputation structure in the estimations.

3.2. Area-Level Model

The input for the area-level model is provided by the direct estimates for the areas. Let y_{ia} denote the average of the ten plausible values of an individual i who belongs to municipality a, as observed in the original survey data (PIAAC). Specific for the area-level model, we transform each y_{ia} in a dichotomous value, as described in the above paragraph.

Then, the average of these values is used to construct the area average of literacy, for example, $\bar{y}_a$, using the paired jackknife estimator (see also Section 2). The jackknife is used to estimate the variance of $\bar{y}_a$, denoted Ψ_a^2, and accounts for sampling error, the uncertainty of multiple imputation for missing values, and the uncertainty of the IRT model underlying the adaptive tests for literacy, using both replicate weights and plausible values. Therefore, it takes fully into account the uncertainty resulting from the PIAAC questionnaire design (OECD 2013c). Furthermore, let $\bar{\mathbf{X}}_{\boldsymbol{a}}$ denote the vector with the population means of the auxiliary variables derived from the LFS used for calibration. The sample area means for the auxiliary variables derived from the PIAAC sample are denoted $\bar{\mathbf{x}}_{\boldsymbol{a}}$. Survey errors regarding the estimation of $\bar{\mathbf{X}}_{\boldsymbol{a}}$ from the LFS are assumed to be small enough to be negligible and are not taken into account.

In a first step, direct estimates for the target variable for each area are obtained using the survey regression estimator $\hat{y}_a^{surv}$:

$$\hat{y}_a^{surv} = \bar{y}_a + (\bar{\mathbf{X}}_{\boldsymbol{a}} - \bar{\mathbf{x}}_{\boldsymbol{a}})^{\mathrm{t}}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the vector with regression coefficients from the linear model that describes the relation between the target variable y and the auxiliary variables x. These direct estimates are the input for the area level or Fay–Herriot model:

$$\hat{y}_a^{surv} = \alpha + \bar{\mathbf{X}}_{\boldsymbol{a}}\boldsymbol{\beta} + \mathrm{u}_a + \mathrm{e}_a \tag{1}$$

where α is the intercept, $\bar{\mathbf{X}}_{\boldsymbol{a}}$ the area covariate averages, $\boldsymbol{\beta}$ the vector of coefficients of covariates, and u_a a random effect to take into account area-level variation not explained by the fixed part of the equation. The random effects are assumed to be normally and independently distributed, with an expected value equal to zero and model variance σ^2. Finally, e_a is an independently distributed sampling error that has expected value zero and sampling variance Ψ_a^2. Based on this model, the best linear unbiased predictor (BLUP) estimator for the area means is equal to (Rao and Molina 2015):

$$\hat{y}_a^{BLUP} = \varphi_a\left(\bar{y}_a + (\bar{\mathbf{X}}_{\boldsymbol{a}} - \bar{\mathbf{x}}_{\boldsymbol{a}})^{\mathrm{t}}\hat{\boldsymbol{\beta}}\right) + (1 - \varphi_a)(\bar{\mathbf{X}}_{\boldsymbol{a}}^{\mathbf{t}}\hat{\boldsymbol{\beta}}), \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of fixed effects estimated at the national level and φ_a is a weight between the direct and synthetic estimator given by $\varphi_a = \sigma^2/(\Psi_a^2 + \sigma^2)$. Now, if in Equation (2), the variance of the random area effects σ^2 is replaced by its estimator $\hat{\sigma}^2$, the empirical BLUP (EBLUP) estimator is obtained. Moreover, the sampling variance Ψ_a^2 is assumed to be known; however, in practice, this is not true and, in this application, it is replaced by its estimator obtained with the paired jackknife. The mean squared error

(MSE) of the EBLUP accounts for the additional uncertainty that is introduced, since σ^2 is replaced by its estimator $\hat{\sigma}^2$ but ignores the uncertainty of using an estimator for Ψ_a^2, which is common practice in SAE procedures.

In this article, an HB approach is applied to fit Equation (2). The HB model is based on Equation (1) under the assumption that $e_a \sim N(0, \psi_a^2)$ and $u_a \sim N(0, \sigma^2)$. For $\boldsymbol{\beta}$ and σ^2, a flat prior distribution is assumed. The HB estimates for the area means, including their MSEs, are obtained by the posterior means and posterior variances of the posterior density for the area means μ_a. These estimates can be evaluated using separate one-dimensional numerical integrations.

To obtain stable variances for the survey regression estimates, the variance approximations obtained with the jackknife are pooled using an analysis of variance type pooled estimator:

$$\Psi_a^{2;P} = \frac{1}{N_a} \frac{\Sigma_{a=1}^m (N_a - 1)\Psi_a^2}{\Sigma_{a=1}^m (N_a - 1)},$$

where m is equal to the total number of areas.

Furthermore, it was clear that some municipalities had unrealistically low literates estimates (one was even negative): they were underestimated due to the linearity of the model. Therefore, two post-result changes were implemented. First, we acknowledged that the model had problems estimating the true percentages in areas where the percentage of low literates is very small (<5%), which is further considered in the results. So, during categorization, we marked these municipalities as having a very small percentage (0–5%) of low literates and grouped them together when publishing the results. Second, a choice was made to benchmark the results such that they would add up to the national level as per You et al. (2004), by means of the direct estimate of undercoverage per area and the sampling variances.

Since the dependent variable in the Fay–Herriot model are direct estimates of percentages, we also considered a log odd transformation, that is, Equation (1) applied to $\log\left(\hat{y}_a^{surv}/(1-\hat{y}_a^{surv})\right)$. As shown in the results section, the area level model after applying a log-odds transformation results in more biased domain predictions than the area level model applied to the untransformed estimates. Applying a linear model directly to binary data or percentages might appear rigid at first sight, but similar linear models are used to motivate the general regression estimator that is generally used in survey sampling to estimate sample means or totals of binary or categorical variables. Examples where the area level model is applied to untransformed estimated percentages in the context of SAE are Datta et al. (1999), You et al. (2003) and Arima et al. (2017).

3.3. Unit-Level Model

As before, let y_{ia} denote the average of the 10 plausible values of the literacy proficiency level of an individual i in area a. The true mean is then equal to

$$y_{ia} = \mu_{ia} + e_{ia} = \alpha + \mathbf{x}_{ia}^{\mathbf{t}}\boldsymbol{\beta} + u_a + e_{ia}, \tag{3}$$

where $\mathbf{x}_{ia}$ is a vector with covariates for respondent i from area a and u_a is an area-specific random effect assumed to be independent and identically distributed. We assume e_{ia} is a measurement error for respondent i, with expected value zero and variance σ_e^2. The EBLUP estimator is then equal to

$$\hat{y}_a^{EBLUP} = \varphi_a(\hat{y}_a^{surv}) + (1 - \varphi_a)(\bar{\mathbf{X}}_{\boldsymbol{a}}^{\mathbf{t}}\hat{\boldsymbol{\beta}}),$$

where the weight φ_a, dependent on area size N_a, is given by $\varphi_a = \sigma^2/(\sigma^2 + \sigma_e^2/N_a)$. The HB model is obtained with Equation (3) with the assumption that $e_{ia} \sim N(0, \sigma_e^2)$ and $u_a \sim N(0, \sigma^2)$. Furthermore, flat priors are assumed for $\boldsymbol{\beta}$, σ_e^2, and σ^2. The HB predictors for the area means, for example, $\hat{y}_a^{HB}$, with their MSEs, are computed as the posterior means and posterior variance of the posterior distribution of μ_a in a similar way as for the area-level model. The resulting integrals are solved using numerical integration.

Unlike the area-level model for the percentage of low literates, where the imputation uncertainty is taken into account when constructing $\bar{y}_a$, the unit-level model as described above does not take into account the imputation uncertainty.

Multiple imputation is one way to take into account this imputation uncertainty, combining results by means of Rubin's rules (Rubin, 1996). The plausible values generated with the PIAAC software are used to calculate multiple HB predictions for the areas. Let $\hat{y}_{aj}^{HB}$ denote the HB prediction for area a based on the jth set of plausible values generated for the PIAAC sample and $MSE\left(\hat{y}_{aj}^{HB}\right)$ denote the posterior variance of $\hat{y}_{aj}^{HB}$. The final HB prediction for area a is defined as

$$\hat{y}_a^{imp} = \sum_{j=1}^{k} \frac{\hat{y}_{aj}^{HB}}{k},$$

where k is the total number of plausible values. The total variance V_a^{imp} is equal to

$$V_a^{imp} = W_a + \frac{k+1}{k} B_a,$$

where the within-imputation variability W_a is obtained as the mean over the MSE of the HB small area predictions:

$$W_a = \sum_{j=1}^{k} \frac{MSE(\hat{y}_{aj}^{HB})}{k}.$$

The between-imputation variability B_a is

$$B_a = \sum_{j=1}^{k} \frac{(\hat{y}_{aj}^{HB} - \hat{y}_a^{imp})^2}{k-1}.$$

Note that Rubin's rule for multiple imputation is derived for large samples. It is unclear to what extent the application of this methodology to small area estimation problems introduces additional bias in point estimates and uncertainty measures. This is left for further research.

4. Model Fitting

4.1. Merging of Municipalities

As stated before, in 2012 the Netherlands was comprised of 415 municipalities. However, some municipalities are quite small and we cannot guarantee that their LFS data cover enough respondents to provide an accurate representation of its inhabitants. Therefore, it is necessary to work with municipality clusters instead. We use 40,000 as the minimum number of residents per area to ensure the LFS estimates can be considered reliable, for example, the variance being low enough to be negligible. This minimum value is based on Statistics Netherlands' publication strategy that three year averages of direct LFS estimates are published for municipalities with a minimum of 40,000 residents aged 16 years and over from 2010 onwards. Municipalities with fewer residents are clustered together with adjacent municipalities. During this merging, we made sure that all the areas could still be nested in larger official area aggregates, the COROP regions. This is a 40-area classification based on educational provisions. Finally, 208 municipality clusters are obtained, for which small area estimates about literacy will be made. In the PIAAC sample, the minimum number of observations for these clusters is 6, the maximum is 146, and the median is 20.

4.2. Variable Selection

SAE uses auxiliary variables at the area level for additional predictive power. This means that all data available in the LFS that is also included in the PIAAC questionnaire can be picked for use in our model. The list of auxiliary variables for the full model and descriptive results (averages and standard deviations) are presented in Table 2.

Table 2. Comparison of weighted dataset averages and their standard deviations (in parentheses).

Covariate[1]	PIAAC average[2]	LFS average
Age**[4]	41.0 (14.2)	40.6 (14.1)
Male	49.3% (50.0)	50.2% (50.0)
ISEI08-score***	48.7 (18.4)	46.5 (10.6)
Immigrant status		
1st gen***	12.8% (32.6)	14.0% (34.7)
2nd gen***	3.1% (16.8)	9.4% (29.2)
Employment status		
Student	13.9% (34.4)	13.7% (33.8)
Self-employed	9.1% (28.7)	9.1% (29.8)
Full time employee***	37.5% (48.4)	30.9% (46.2)
Part time employee	22.1% (41.5)	21.6% (41.2)
Unemployed***	2.6% (16.0)	3.5% (18.4)
Education[3]		
Vocational ed.	57.5% (49.4)	57.5% (49.4)
Years of schooling***	13.2 (3.7)	13.4 (3.6)

[1]The full list of interactions considered for the full model are age with gender, ISEI-08 score, immigrant status variables, employment status variables and education variables, plus years of schooling with immigrant status variables, ISEI-08 score and vocational education.

[2]For the Netherlands, the control variables that were used to calibrate weights in PIAAC are: Gender by age (10), origin by generation (5), group of provinces by degree of urbanization (18), household type (5), social status by income (25), term of registration in population registry (2), percentage of high level education by percentage of low level education (18).

[3]The education variables contained slightly more than 1% missing values. For area estimates, missing values are assumed have the same distribution as the known values.

[4]Indicates the level of statistical significance of the t-test between the two datasets. ***$p < 0.001$, **$p < 0.05$, *$p < 0.01$.

There are some statistically significant differences in the distribution of these variables between PIAAC and LFS, although most of these differences in distribution are rather small in nature; our large sample sizes allow even minor differences to be statistically significant. The most notable difference is the percentage of second-generation immigrants in the PIAAC data set, which is significantly lower in the PIAAC data set compared to the LFS data set. Also, there is a (non-significant) larger percentage of fulltime employees, and a lower percentage of unemployed persons. There are some minor differences for age, occupational status and years of schooling where the gap between the means is very small.

In the literature, different methods are proposed for model selection. In this article, optimal models are selected by means of the conditional Akaike information criterion (cAIC) using a stepwise backward variable selection procedure. This method is applied more often in small area estimation (see e.g., Van den Brakel and Buelens 2015). The cAIC, proposed by Vaida and Blanchard (2005), is applicable to mixed models where the focus is on prediction at the level of areas. The penalty (p) on the log likelihood is based on the model complexity. The random part of the model contributes to the number of degrees of freedom p with a value between zero in the case of no area effects (i.e., $\hat{\sigma}^2 = 0$) and the total number of areas m in the case of fixed area effects (i.e., $\hat{\sigma}^2 \rightarrow \infty$). The effective number of degrees of freedom used for the penalty is defined as the trace of the hat matrix H, which maps the observed data to the fitted values, for example $\hat{y} = Hy$, see Hodges and Sargent (2001). The cAIC has a more realistic penalty for the random component of a multilevel model, compared to the standard AIC (where a random effect counts for one degree of freedom). Nevertheless, the cAIC in a stepwise selection procedure might result in complex models that overfit the data. Alternatively, cross-validation is sometimes used as a measure for model selection, see Boonstra et al. (2008). Other authors propose the LASSO (Hastie et al. 2001) as a form of model selection (Thao and Geskus 2019). In this article, the cAIC is used in combination with a backward selection procedure and in the model evaluation it is established that the selected models do not overfit the data.

Covariates were removed one by one until a minimum for the cAIC was reached for the unit-level model on literacy scores. The list of the selected predictors is as follows:

- Age, Age squared,
- Immigrant Status,
- Years of Schooling,
- Area of Study (eight categories),
- Highest level of education is Vocational Education (Dummy); Note that vocational education in the Netherlands can be secondary, upper-secondary and tertiary level,
- Employment Status,
- Occupational Status Measure based on the International Socio-Economic Index (ISEI) of ISCO-08 occupations by Ganzeboom et al. (1992), a continuous variable measuring the socio-economic status of an occupation,
- Two 2-way interaction terms of Years of Schooling with Immigrant Status and Occupational Status, and
- Six 2-way interaction terms of Age with Gender, Vocational Education and Employment Status.

The interaction terms help with estimating effects of variables not captured in our data sets. For example, international knowledge workers would be classified as immigrants, which is generally a negative indicator. By including the interaction effect with years of schooling, we can partially correct for this. For the area level model, we can find a model with a slightly lower cAIC score (ΔcAIC $= 2.9$) by leaving out the self-employed and one dummy regarding the area of study. However, in theory there is no reason why the two sets of literacy measures should have different predictors. Given the small difference in model selection, we opt to use the same model for both predictors. A quick test using the other model reveals that all results lie within the confidence interval of our preferred model.

5. Model Evaluation

The SAE results can differ from the direct results for a number of reasons. The most important reason is why SAE techniques are applied in the first place, namely, to improve the precision of the direct municipality estimates. However, it is important to make sure the differences are not dominated by the bias introduced in the model. Since SAE techniques explicitly rely on statistical models to improve the effective sample size in the separate areas, one must evaluate the underlying assumptions of the models to ensure the bias introduced by the synthetic estimator is small. Model misspecification can easily result in heavily biased area estimates. This section evaluates the normality assumptions underlying the applied models. Furthermore, direct area estimates are compared with model-based small area predictions to assess possible systematic bias. Finally, the improvement in precision is evaluated by comparing the standard errors of both estimators.

5.1. Robustness Checks

The direct estimates at the national level are precise and unbiased, since they do not depend on model assumptions and are based on a large sample. Therefore, the difference between the model-based small area predictions, aggregated at the national level, with the direct estimates at the national level is often used as a measure of bias in SAE.

As noted earlier in Section 3, benchmarking was applied to remove differences between model-based area estimates aggregated at the national level and direct estimates at the national level. Small area estimates for literacy scores and the percentage of low literates at the national level are obtained by calculating the mean over the municipalities weighted by the number of residents in 2012. Table 3 displays the results of the non-benchmarked estimates against the (robust) national results.

Table 3. Estimated aggregated results at higher levels, without benchmarking.

Type	Direct	SAE (*)
Average Literacy	283.9	287.9
% Low Literates	12.0%	12.8%

*indicates the average of the SAE results over municipalities, weighted by the number of residents in 2012.

For both measures of literacy, the SAE scores are slightly overestimated. The average literacy of 287.9 is greater than the upper bound of 285.3 for the direct estimates given in Table 1. The estimate of the percentage of low literates estimates is contained within the 95% confidence interval, but barely. On the basis of these results, we decided to benchmark our estimates.

Before benchmarking, we look at the differences between the direct estimates and the SAE results. Two measures are applied to summarize the differences between the direct and model-based area estimates. The first one is the *mean relative difference* (MRD), in percentages, defined as

$$MRD = \frac{1}{m}\sum\nolimits_{a=1}^{m} \frac{(\hat{y}_a^{direct} - \hat{y}_a^{SAE})}{\hat{y}_a^{direct}} 100,$$

where $\hat{y}_a^{SAE}$ is the unbenchmarked Hierarchical Bayesian SAE estimator. The second one is the *absolute mean relative difference* (AMRD), in percentages, defined as

$$AMRD = \frac{1}{m}\sum\nolimits_{a=1}^{m} \frac{(\hat{y}_a^{direct} - \hat{y}_a^{SAE})}{\hat{y}_a^{direct}} 100.$$

Table 4 gives the MRD and AMRD for the two literacy measures.

The MRD for both estimates is quite small, with roughly 1.7 percentage point for the average literacy and half a percentage point for the low literacy percentage. Since it is negative, the SAE estimators are generally slightly bigger. When we look at the absolute difference, we see a 2.78% mean difference for average literacy, and 0.70% for low literacy.

To interpret the differences between the direct estimates and the domain predictions obtained with the finally selected SAE models in more detail, we compare the distribution of the benchmarked SAE estimates with the distribution of the direct results from PIAAC. Figure 1 shows the tendency of the SAE estimates to tend towards the mean. Regarding the average literacy scores, the scores at the right side of the distribution consist mostly of those for university cities, where the number of students seems to be oversampled. The scores at the left side of the distribution are mostly for small villages, but the worst results are for some municipalities of medium-sized cities.

For the estimated percentage of low literates, the distribution is close to the distribution of the direct estimates; however, note that the SAE results for the average and below-average percentage of low literates are often higher than the direct results. The relatively high proportion of municipalities (over 10%) that perform well in terms of percentage of low literates (with percentages in the range of 0–5%) in the direct estimates could be due to the fact that these municipalities are very small and have few direct observations in

Table 4. Measures of quality of the estimates (%), without benchmarking.

	Average Literacy	% Low Literates
MRD	−1.66	−0.51
AMRD	2.78	0.70

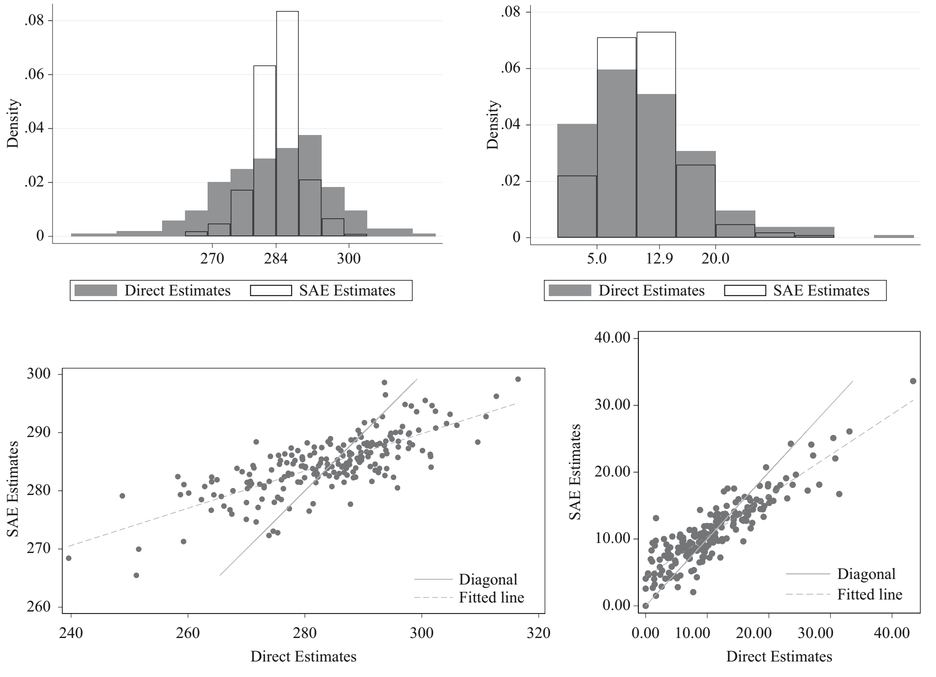

Fig. 1. Histograms and distribution plots of the direct results and the SAE results (left, literacy scores; right, % low literates; the solid line is the diagonal, the dashed line is the linear fit).

PIAAC. Therefore, these differences would be a result of the improved accuracy of the point estimates.

Figure 2 shows the scatter plots of the fitted values of both SAE measures versus the quantiles of the residuals. No pattern can be distinguished within the two graphs, meaning the residuals are well behaved.

Q-Q plots for the estimates, residuals and random effects can be found in the Supplementary materials.

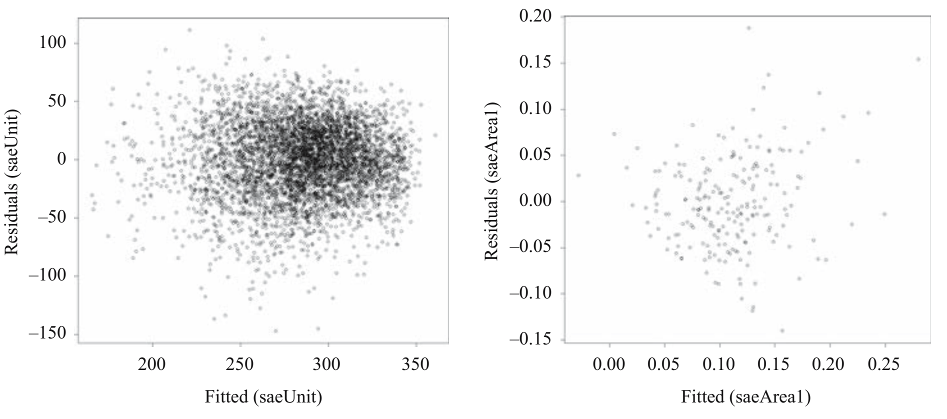

Fig. 2. Fitted values versus the residuals of the unit-level estimates of the estimated literacy scores (left) and the area-level estimates of the percentage of low literates after benchmarking (right).

For the percentage of low literates, a log odds transformation of the dependent variable was also considered and applied. The model under the log odds transformation shrinks in particular the direct domain estimates with small values much stronger to the overall mean, resulting in larger amounts of bias (RMD and ARMD have values of respectively -1.485 and 1.580). Furthermore, the residuals and random effects show stronger deviations from normality. See the Supplementary materials for more details. Therefore, the model applied to the untransformed direct estimates is chosen to be our final model. As explained in Section 3, this is not uncommon in survey sampling and SAE literature.

5.2. *Reduction in Standard Error*

To measure the increase in precision obtained with the SAE techniques, the *mean relative difference in standard errors* (MRDSE) is used. This is defined as the ratio between the standard errors between the direct and the SAE estimator, averaged per area, or in formula form:

$$MRDSE = \frac{1}{m}\sum\nolimits_{a=1}^{m} \frac{(SE_a^{direct} - SE_a^{Bench})}{SE_a^{direct}} * 100$$

The results are shown in Table 5. The MRDSE for average literacy is 67.9%, which, compared to the direct estimates, is a significant reduction. For the percentage of low literates, the reduction measure is 51.2% (31.3%) when compared to the pooled variance) but, as a less powerful model, lower returns are to be expected.

In Figure 3, we look at the number of respondents in PIAAC versus the standard error of the direct estimates, as well as the SAE results for the average literacy scores per municipality. Given the high frequency of respondents numbering between 5 and 20 per municipality, we decided to plot this graph on a logarithmic scale.

For small sample sizes, the SAE results show a large decrease in terms of standard errors compared to the direct estimator, whose margin of error is far too large when it comes to accurate point estimates. As the sample size increases, the difference between the two estimators decreases greatly.

In Figure 4, we look at the standard errors for the percentage of low literates. Here, the standard errors of the direct estimator are much more spread out and sometimes even zero (due to the direct estimator being zero). When compared to the direct estimator with pooled standard errors they are much closer to the SAE results due to the decrease in information compared to the model utilizing literacy scores, but there is still a significant gain in municipalities with low numbers of PIAAC respondents.

Table 5. Measures of the quality of estimates (%), without benchmarking.

	Average Literacy	% Low Literates*
MRDSE	67.9	51.2 (31.3)

*indicates the numbers in parentheses are compared to the standard errors of the pooled variance instead of the direct standard errors.

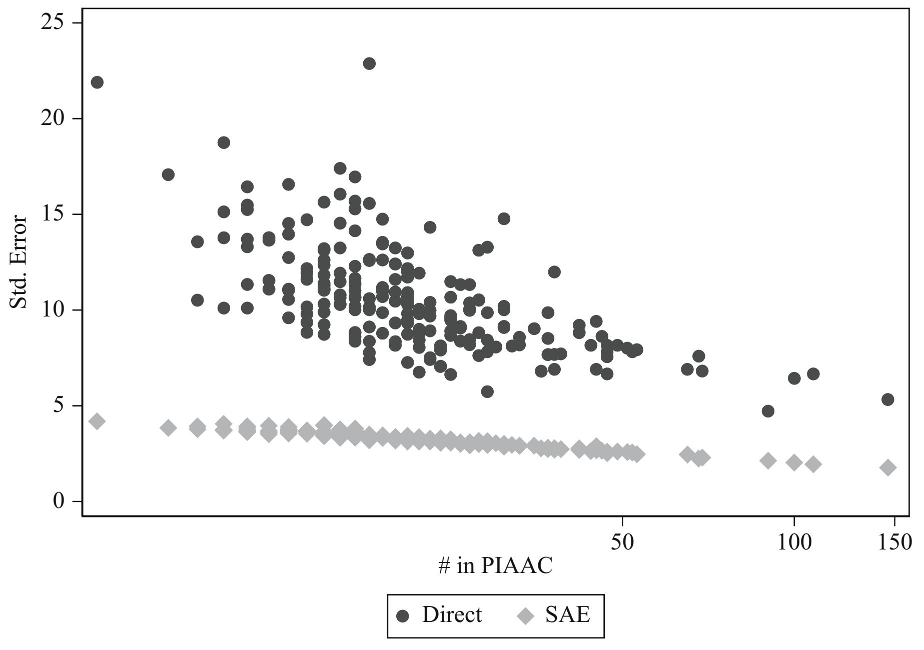

Fig. 3. Standard errors versus the (logarithmic) number of PIAAC respondents for both the direct estimates and the SAE for the estimated literacy scores per municipality.

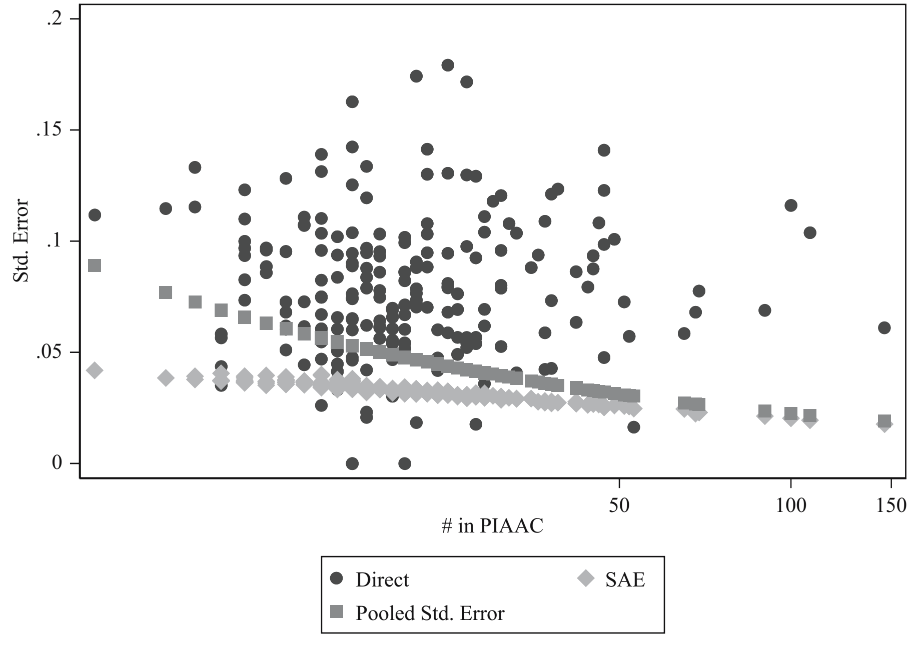

Fig. 4. Standard errors versus the (logarithmic) number of in PIAAC respondents for both estimates and the SAE for the percentage of low literates per municipality.

6. Results

In this section, we present the substantive results graphically, review them, and discuss the differences in results for the two chosen measures of literacy. The full list of results per municipality can be found in the online Supplementary material.

Figure 5 shows the average literacy scores per municipality cluster. Neighbors are rarely in the same category and often differ by multiple categories. Generally, the highest scores for literacy can be found in the center of the country, around the city of Utrecht. Large university cities also do well (Rotterdam being a notable exception). Aside from known problem areas in the western part of the Netherlands, the scores for literacy are low in the peripheral regions.

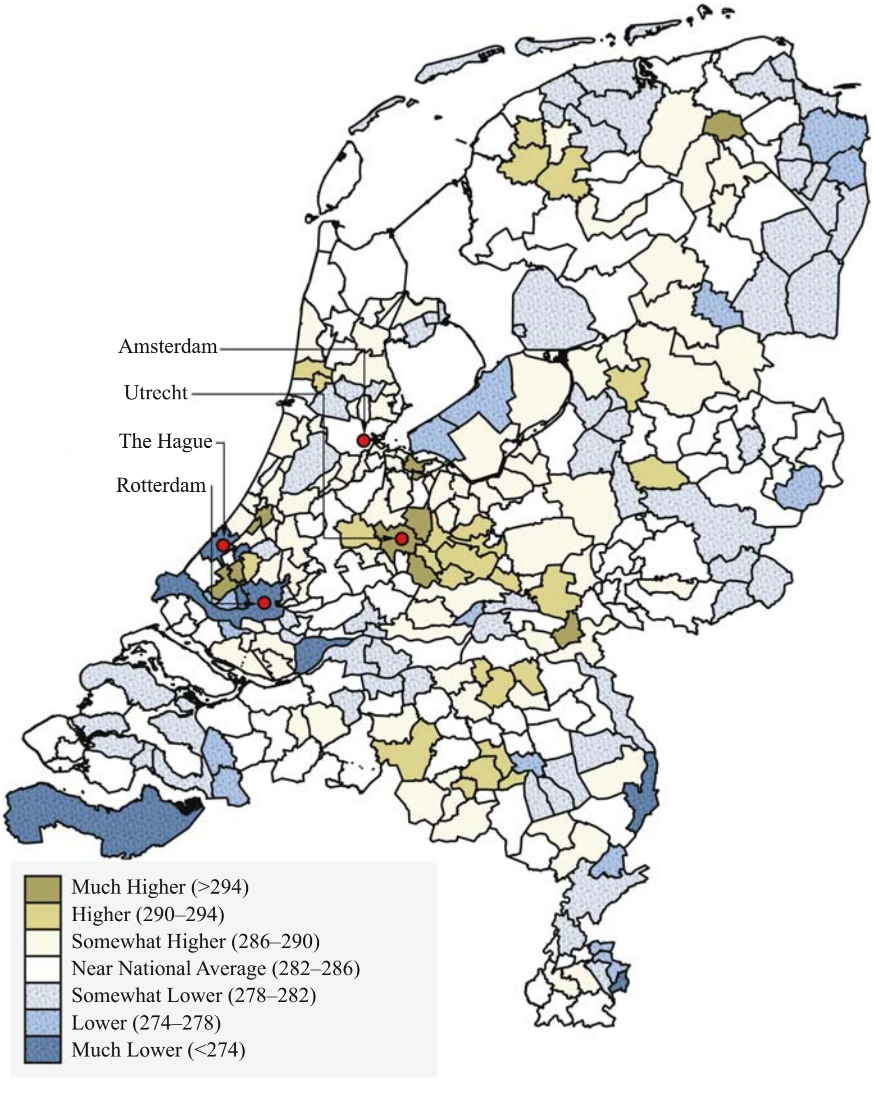

Fig. 5. Estimated average literacy scores per municipality.

Figure 6 shows the regional estimates for the percentage of low literates. There is a similar pattern when we look at areas in terms of the percentage of low literates. The first big notable difference, however, is that, in most cases, large cities do much worse in terms of their percentage of low literates in their population, which underlines the usefulness of having both indicators. Low literacy is mainly found in populations with certain characteristics. The average literacy score could give an idea of the overall situation of a population, but not how it is distributed. Both measures together provide a more complete picture of the literacy within each area.

Next, we give some examples of how SAE estimates for literacy can relate to other outcomes at the regional level. Knowledge of regional differences can be a powerful tool

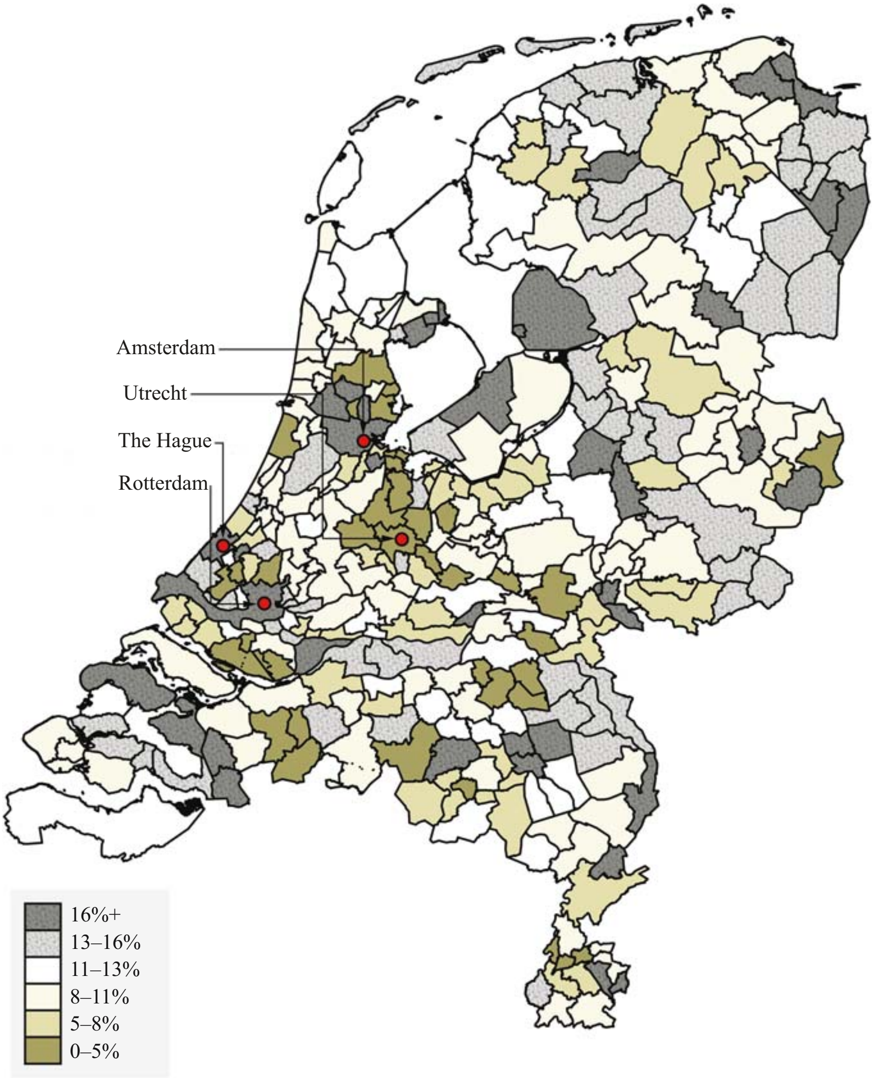

Fig. 6. Estimated percentage of individuals classified as having low literacy proficiency scores per municipality.

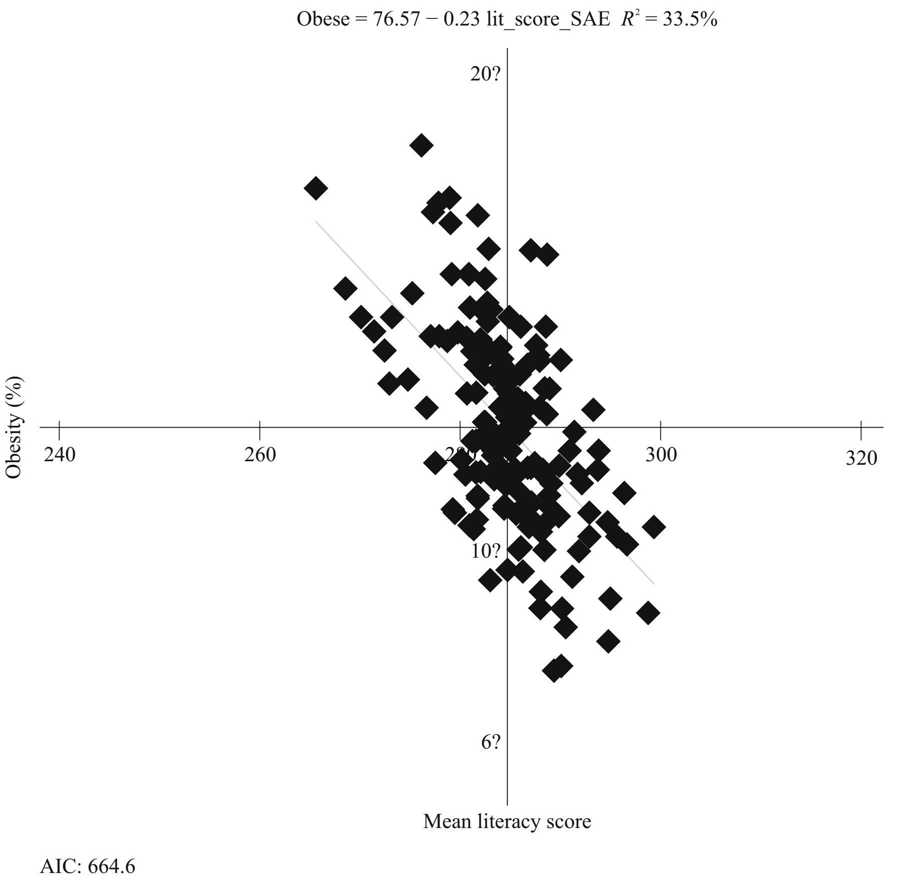

Fig. 7. Linear model of the proportion of obese people (in 2012; Source: Statistics Netherlands) versus the average literacy estimates in that region.

for policy interventions aimed at tackling these problems. This is not simply a matter of identifying areas of low literacy, since this is unlikely to be the sole cause of such problems. Policy makers and professionals responsible for policy implementation have an interest in distinguishing regions in which poor health, and other unwanted outcomes are associated with low literacy from regions in which these problems are driven more by other factors. Such knowledge can greatly improve the cost effectiveness of interventions.

As a simple illustration, in Figure 7, we plot the relation between (low) literacy and one unwanted non-economic problem: obesity. Note that the following is for illustration purposes only. This approach facilitates the implementation of more targeted policy interventions. The idea behind this is the following. Very often problems like low literacy, health problems or socio-economic problems go hand in hand. Policy, therefore, is often aimed at an integral approach, such as a combination of helping to find work, improvement of a healthy lifestyle and improving the literacy proficiency. For policy makers it is helpful to see which combinations of problems occur in their municipality so that they can fine-tune their interventions for the specific group. Our goal is not to ‘explain’ obesity, but to identify areas in which there is an accumulation of both types of problems versus areas where this is not the case.

The relation between the average literacy score and the incidence of obesity is quite strong ($R^2 = 33.5\%$), but also far from perfect. There are areas where the two problems go hand in hand and areas where this is not the case at all. In terms of policy interventions, the position of a given municipality in the graph is indicative of the kind of policy response that could be considered appropriate. There is little incentive to launch literacy-based interventions in the regions in the lower right quadrant, since these are regions with high literacy and a low incidence of obesity. In the lower left and upper right quadrants, literacy-based interventions also do not look promising, at least not to combat obesity, since literacy and obesity do not coincide in these regions. Only in the upper left quadrant do we see a high incidence of obesity together with a low average level of literacy. This finding suggests that literacy could potentially be targeted as a policy lever to tackle the problem of obesity in these regions.

7. Conclusion

In this article, we have combined PIAAC survey data with LFS data to obtain estimates of the literacy levels for municipalities in the Netherlands, both the average literacy scores and the percentage of low literates. These estimations are obtained using SAE models fitted with an HB approach.

Direct estimators only use observations obtained in each specific area to estimate literacy for that area. Results obtained with direct estimators at the regional level, therefore, suffer from small samples sizes for most areas, leading to high standard errors. In this article, we applied model-based estimation procedures to improve the effective sample size in the different areas, resulting in a considerable improvement of the precision of the estimates of literacy levels, even in larger cities of the Netherlands.

We show that we can obtain estimates at a very detailed regional level by using these SAE techniques, with standard errors reduced more than 50%. This is important, since policy to combat low literacy is often targeted at the municipality level. We show that we can obtain reliable estimates for the average literacy level and the percentage of low literates for over 200 municipalities in the Netherlands. The findings show that average literacy levels are higher in big cities than in more rural areas, a finding that is consistent with the literature (e.g., McHenry 2014). However, we also show that large cities cope with higher proportions of low literates, indicating the importance of looking at both measures of literacy.

The estimates can help to determine a more optimal allocation of resources to combat low literacy. We also illustrated that more precise SAE estimates are helpful in establishing relations with other variables more clearly. This approach can be used, for example, to identify municipalities that suffer from multiple problems, such as low literacy and health problems or other social problems. In some municipalities, these problems coincide, and in some municipalities they do not. Identifying the typical mix of problems a municipality is confronted with is key to the development of a successful intervention strategy. The regional estimates for literacy, therefore, give room for policy makers to implement more directed policies at a detailed regional level.

Future research will focus on the estimation of other skills measured in PIAAC, such as numeracy, or by estimating literacy levels in other areas, such as detailed levels of occupation (for an example, see Van der Velden and Bijlsma 2018). By making these kinds of estimates possible, detailed data become available in areas previously inaccessible due to time and budget constraints.

However, there are a number of caveats to keep in mind when interpreting the results. First and foremost, it must be stressed that these methods rely on statistical model assumptions. Careful model selection and evaluation are, therefore, an important and necessary part of SAE. The method assumes that the effects of covariates at the regional level are the same as at the national level, with random effects capturing regional differences. While this should hold in most cases, exceptions can occur. The results should always be viewed with possible local anomalies in mind.

A number of improvements can be made in the estimation of the model. Currently, data used from the LFS are assumed to be the true population means and the corresponding sampling errors are assumed to be negligible. There are ways to properly consider these errors, such as the method of Ybarra and Lohr (2008) for the area-level model and the method of Lohr and Prasad (2003) for the unit-level model. For the percentage of low literates model, a logarithmic model could lead to better estimations between the 0% and 5%, which currently show some bias toward the bottom end of the distribution. Methods such as the standard ratio raking used in Casas-Cordero et al. (2016) are also an option.

8. References

Arima, S., W.R. Bell, G.S. Datta, C. Franco, and B. Liseo. 2017. "Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error." *Journal of the Royal Statistical Society, Series A* 180: 1191–1209 DOI: https://doi.org/10.1111/rssa.12321.

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 401: 28–36. DOI: https://doi.org/10.1080/01621459.1988.10478561.

Boonstra, H.J. 2015. *Package 'hbsae'* (version 1.0). Available at: https://cran.r-project.org/web/packages/hbsae/hbsae.pdf (accessed December 2015).

Boonstra, H.J., J.A. van den Brakel, B. Buelens, S. Krieg, and M. Smeets. 2008. "Towards small area estimation at Statistics Netherlands." *METRON International Journal of Statistics* LXVI: 21–49. Available at: https://EconPapers.repec.org/RePEc:mtn:ancoec:080102 (accessed April 2020).

Buisman, M., J. Allen, D. Fouarge, W. Houtkoop, and R. van der Velden. 2013. *PIAAC: Kernvaardigheden voor werk en leven. Resultaten van de Nederlandse survey 2012*, Den Bosch/Maastricht: ECBO/ROA.

Casas-Cordero, C., J. Encina, and P. Lahiri. 2016. "Poverty mapping for the Chilean Comunas." In *Analysis of Poverty Data by Small Area Estimation*, edited by M. Pratesi, 379–403. Hoboken: Wiley. DOI: https://doi.org/10.1111/j.1467-9787.2007.00538.x

Coulombe, S. and J.F. Tremblay. 2007. "Skills, Education, and Canadian Provincial Disparity." *Journal of Regional Science* 47: 965–991. DOI: https://doi.org/10.2307/2669921.

Datta, G., P. Lahiri, T. Maiti, and K. Lu. 1999. "Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S." *Journal of the American Statistical Association* 448: 1074–1082.

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro estimation of poverty and inequality." *Econometrica* 71: 355–364. DOI: https://doi.org/10.1111/1468-0262.00399.

Fay, R.E. and R.A. Herriot. 1979. "Estimates of income for small places: An application of James-Stein procedures to census data." *Journal of the American Statistical Association* 366: 269–277. DOI: https://doi.org/10.2307/2286322.

Ganzeboom, H.B.G., P.M. de Graaf, and D.J. Treiman. 1992. "A Standard International Socio-Economic Index of Occupational Status." *Social Science Research* 21: 1–56. DOI: https://doi.org/10.1016/0049-089X(92)90017-B.

Gibson, A. and P. Hewson. 2012. "2011 Skills for Life Survey: Small Area Estimation Technical Report." *BIS Research Report* 81C. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/36077/12-1318-2011-skills-for-life-small-area-estimation-technical.pdf (accessed November 2018).

Hanushek, E.A. and L. Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46: 607–668. DOI: https://doi.org/10.3386/w15949.

Hanushek, E.A. and L. Woessmann. 2011. The Economics of International Differences in Educational Achievement. In *Handbook of the Economics of Education*, Vol. 3: 89–200. Amsterdam: North Holland.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning*. Springer: New York.

Hodges, J.S. and D.J. Sargent. 2001. "Counting degrees of freedom in hierarchical and other richly parameterized models." *Biometrika* 88: 367–379. DOI: https://doi.org/10.1093/biomet/88.2.367.

Johnson, E.G. and K.F. Rust. 1992. "Sampling and Weighting in the National Assessment." *Journal of Educational and Behavioral Statistics* 17: 111–129. DOI: https:/doi.org/10.2307/1165165.

Lohr, S. and N. Prasad. 2003. "Small Area Estimation with Auxiliary Survey Data." *The Canadian Journal of Statistics* 31: 383–396. DOI: https://doi.org/10.2307/3315852.

McHenry, P. 2014. "The Geographic Distribution of Human Capital: Measurement of Contributing Mechanisms." *Journal of Regional Science* 54: 215–248. DOI: https://doi.org/10.1111/jors.12067.

National Research Council. 2000. "Small Area Estimates of School-Age Children in Poverty: Evaluation of current methodology." *Committee on National Statisitcs*, edited by C.F. Citro and G. Kalton. Washington, DC: National Academy Press.

OECD. 2013a. *OECD skills outlook 2013: first results from the survey of adult skills*. Paris: OECD Publishing. DOI: https:/doi.org/10.1787/9789264204256-en.

OECD. 2013b. *The Survey of Adult Skills – Reader's Companion*. Paris: OECD Publishing. DOI: https://doi.org/10.1787/9789264204027-en.

OECD. 2013c. *Technical Report of the Survey of Adult Skills (PIAAC)*. Available at: http://www.oecd.org/site/piaac/publications.htm (accessed December 2015).

Pokropek, A. and M. Jakubowski. 2013. *Package 'PIAAC tools'* (version 4.3). Available at: https://ideas.repec.org/c/boc/bocode/s457728.html (accessed September 2016).

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28: 40–68. DOI: https://doi.org/10.1214/12-STS395.

PricewaterhouseCoopers. 2013. *Laaggeletterdheid in Nederland kent aanzienlijke maatschappelijke kosten.* Internal Rapport, PWC, Amsterdam.

Rao, J.N.K. and I. Molina. 2015. *Small Area Estimation, Second Edition*. New York: John Wiley and Sons.

Rubin, D.B. 1996. "Multiple Imputation After 18 + Years." *Journal of the American Statistical Association* 434: 473–489. DOI: https://doi.org/10.2307/2291635.

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer Verlag.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 180: 1163–1190. DOI: https://doi.org/10.1111/rssa.12305Y.

Statistics Netherlands. 2010. "Methoden en definities Enquête Beroepsbevolking 2010." Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/enquete-beroepsbevolking-uitgebreide-onderzoeksbeschrijving-2010 (accessed March 2018).

Statistics Netherlands. 2011. "Methoden en definities Enquête Beroepsbevolking 2011." Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/enquete-beroepsbevolking-uitgebreide-onderzoeksbeschrijving-2011 (accessed March 2018).

Statistics Netherlands. 2012. "Methoden en definities Enquête Beroepsbevolking 2012." Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/enquete-beroepsbevolking-uitgebreide-onderzoeksbeschrijving-2012 (accessed March 2018).

Taylor, J., G. Moon, and L. Twigg. 2016. "Using geocoded survey data to improve the accuracy of multilevel small area synthetic." *Social Science Research* 56: 108–116. DOI: https://doi.org/10.1016/j.ssresearch.2015.12.006.

Thao, L.T.P. and R. Geskus. 2019. "A comparison of model selection methods for prediction in the presence of multiply imputed data." *Biometrical Journal* 61: 343–356. DOI: https://doi.org/10.1002/bimj.201700232.

Tighe, E., D. Livert, M. Barnett, and L. Saxe. 2010. "Cross-Survey Analysis to estimate low-incidence religious groups." *Sociological Methods & Research* 39: 56–82. DOI: https://doi.org/10.1177/0049124110366237.

Vaida, F. and S. Blanchard. 2005. "Conditional Akaike information for mixed effect models." *Biometrika* 92: 351–370. DOI: https://doi.org/10.1093/biomet/92.2.351.

Van den Brakel, J.A. and B. Buelens. 2015. "Covariate selection for small area estimation in repeated sample surveys." *Survey Methodology and Statistics in Transition*, Special issue on *Small Area Estimation,* Vol.16: 523–540. DOI: https://doi.org/10.21307/stattrans-2015-031.

Van den Brakel, J.A. and S. Krieg. 2015. "Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design." *Survey Methodology* 41: 267–296. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14231-eng.pdf (accessed April 2020).

Van der Velden, R. and I. Bijlsma. 2018. "Effective skill: a new theoretical perspective on the relation between skills, skill use, mismatches and wages." *Oxford Economic Papers*, Advance articles. DOI: https://doi.org/10.1093/oep/gpy028.

World Bank. 2002. "How Low Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa." *Journal of African Economies* 11: 169–200. DOI: https://doi.org/10.1093/jae/11.2.169.

Yamamoto, K. 2014. *Using PIAAC Data for Producing Regional Estimates*. Working Paper, Educational Testing Service, Princeton.

Ybarra, L.M.R. and S.L. Lohr. 2008. "Small area estimation when auxiliary information is measured with error." *Biometrika* 95: 919–931. DOI: https:///doi.org/10.1093/biomet/asn048.

You, Y., J.N.K. Rao, and P. Dick. 2004. "Benchmarking Hierarchical Bayes Small Area Estimators in the Canadian Census Undercoverage Estimation." *Statistics in Transition* 6: 631–640. Available at: https://www.semanticscholar.org/paper/BENCHMARKING-HIERARCHICAL-BAYES-SMALL-AREA-IN-THE-You-Rao/efaafa565aa134fe0943f03bbad15278eb228e3a (accessed April 2020).

You, Y., J. Rao, and J. Gambino. 2003. "Model-based unemployment rate estimation for the Canadian Labour Force Survey: A Hierarchical Bayes approach." *Survey Methodology* 29: 25–32. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20030016602 (accessed April 2020).

Received November 2018
Revised August 2019
Accepted January 2020