The Survey Statistician, 2024, Vol. 90, 30–37.

The Use of New Data Sources in Small Area Estimation of Attitudes towards Climate Change

Camilla Salvatore[1] **and Angelo Moretti**[2]

[1]Utrecht University, The Netherlands, c.salvatore@uu.nl
[2]Utrecht University, The Netherlands, a.moretti@uu.nl

Abstract

Climate change is a global problem that has a significant impact on the world's economy and society. To effectively address climate change and other societal challenges, policymakers often require reliable estimates of relevant variables at a sub-national level. Nationally representative surveys are not often designed for this purpose. In this study we propose to use small area estimation techniques to obtain reliable estimates of the proportion of people very and extremely worried about climate change at regional level. A novel aspect of our approach is that we include non-traditional auxiliary information, specifically web data, into our model. For the data used in this paper, our results show that incorporating web data yields more reliable estimates than the model without them. Finally, we also acknowledge and address certain limitations associated with the use of web data in small area estimation.

Keywords: Digital Trace Data, Data Integration, Attitudes, Fay-Herriot model

1 Introduction

Climate change is one of the greatest challenges of the present century, with consequences for ecosystems, the economy, and society [Lee et al., 2023]. This global issue has promoted collaborative cross-national efforts, such as the Paris Agreement, which provides a common framework to combat climate change and mitigate its impact. Furthermore, in 2020, the European Union (EU) approved the European Green Deal, a comprehensive strategy with the goal of achieving climate neutrality by 2050.

As governments worldwide engage in policies concerning climate change and sustainable development, the necessity for high-quality data to monitor these processes and the public opinion becomes increasingly important. Beyond environmental metrics, understanding societal attitudes and behaviours towards climate change is necessary for developing effective strategies taking into account the perspectives of the local communities [Prakash and Bernauer, 2020].

Large-scale nationally representative survey data are the base for the construction of such indicators. They are designed to produce precise and accurate estimates for large population domains, e.g., at country-level. However, policymakers and researchers are often interested in sub-national indicators, i.e., at regional or province-level. Direct estimates obtained for these areas may return large variability

to small sample sizes [Rao and Molina, 2015]. A popular approach to derive reliable estimates at sub-national level is Small Area Estimation (SAE).

The key idea of SAE models is to "*borrow strength*" from the other areas and auxiliary information from nationally representative surveys, administrative data or non-traditional data sources.

In recent years, the integration of digital trace data (e.g. from websites, social media, google trends) with survey data has gained importance [Salvatore, 2023]. In SAE the use of these non-traditional sources is very promising because this data can provide additional and relevant information that characterize the small-area of interest. Some examples in SAE include the use of Twitter (now X) data to improve the estimation of food consumption expenditure and the use of Google trends data to estimate relative changes in rates of household Spanish-speaking in the United States [Marchetti and Schirripa Spagnolo, 2024, Marchetti et al., 2015, 2016, Porter et al., 2014].

In this study, our focus is to estimate attitudes towards climate change at a regional level (NUTS2). We present some preliminary results on one country only, Spain, with plans to extend the analyses to more countries in future work. To do that, we utilise the European Social Survey (ESS) data and we apply a SAE model, named Fay-Herriot (FH). As an innovative methodological element in our analyses, we consider auxiliary information in the model, from both traditional data sources (Eurostat archive) and non-traditional ones (web data from Booking.com).

The aim of this paper is twofold. Firstly, given the importance of climate change in the current political and social debate, the aim of this paper is to show how reliable indicators of attitudes towards climate change at a local level can be obtained through SAE. Secondly, we study whether the use of new data sources can be beneficial in the estimation process.

The remainder of this article is structured as follows. In Section 2 we present the data and the modeling strategy. In Section 3 we discuss the results and we draw conclusions in Section 4.

2 Data and Methods

2.1 The European Social Survey

We employ data from ESS round 10. ESS is a nationally representative European cross-national survey that has been running every two years since 2001 [European Social Survey, 2022]. The survey collects data for a large umber of countries in Europe, however, in this article, we focus on Spain only. We consider the NUTS2 level of classification, which pertains to 17 autonomous regions and 2 autonomous cities (Melilla and Ceuta). The latter areas are from our analyses since they did not have all the auxiliary variables available. This results in the selection of 17 autonomous regions for a total sample size of 2214 units. The first row in Table 1 presents descriptive statistics for the regional sample sizes. It is evident that some regions present very small sample sizes, thus, they do not allow for reliable direct estimates. Indeed, the ESS is not designed to produce accurate and precise estimates at the sub-national level [Moretti and Whitworth, 2020, Santi and Moretti, 2021], hence SAE methods are needed.

We focus on a specific indicator, i.e., the proportion of individuals who are very and extremely worried about climate change. The survey participants were asked to answer this question: "How worried are you about climate change?" on a 5-point scale (1 not at all worried, 2 not very worried, 3 somewhat worried, 4 very worried, and 5 extremely worried). We dichotomies the variable following the median split method (median = 4). Thus, we re-code the variable as "very and extremely worried" (score equal to 4 or 5) versus "other" (score from 1 to 3). The weighted proportion at the national level of people very concerned about climate change is 56.40%. The second and third rows in Table 1 show descriptive statistics of the direct estimates and the coefficient of variation (CV) across the 17

regions. As a rule of thumb, a CV larger than 16% is considered non-reliable, hence, SAE is needed [Marchetti and Schirripa Spagnolo, 2024].

Table 1: Summary statistics of regional sample sizes, direct regional estimates and their CVs

	Min.	Median	Mean	Max.
Sample Size	21	73	130	368
Direct estimate	41.66	57.37	54.97	69.58
CV%	6.71	26.27	16.02	31.20

2.2 The auxiliary data

In the FH model, we use auxiliary data from both traditional sources, i.e., official statistics, and also from non-traditional data sources, specifically web data.

In terms of traditional variables, we consider the following ones: proportion of people with tertiary education, long term unemployment rate, robbery rate and population density. These variables show a good spatial heterogeneity at regional level in Europe, and they were used in public attitudes context before [Santi and Moretti, 2021, Moretti and Whitworth, 2020]. These variables can be obtained by Eurostat data archive[1].

As web data, we consider information about sustainable hotels available on Booking.com. Until recently, the *Travel Sustainable*[2] programme included four different labels that could be assigned to hotels that satisfied some sustainability requirements: namely, from the lowest to the highest, Level 1, Level 2, Level 3 and Level 3+. However, in March 2024 the programme has undertaken significantly changes and now only one label, named *Sustainability certification*, is available. The data we use for our analysis was scraped prior to the change, thus, in this article we use the old classification system. In the concluding remarks, we discuss this issue further. The rational for using the proportion of hotels with a specific label is the following. The number of hotels with an environmental certification can give valuable insights into the spatial socio-environmental context of different areas, in particular with respect to the environmental awareness and sustainability practices. Thus, this can help in estimating attitudes towards climate change at a local level.

In order to obtain the data we perform web-scraping using R [R Core Team, 2021] and the *rvest* package [Wickham, 2024]. Our research query relates only 1 room for 1 adult. For each region, we gather data on the total number of hotels and we calculate the proportions of hotels categorized under different sustainability levels. Due to fluctuations in hotel availability over time, we randomly select 84 dates, roughly 7 days per month, spanning form February 1st 2024 and January 31st 2025. Thus, we proceed by averaging the proportions across the 84-day period for each region. Across all regions, the proportion of hotels with label Level 1 is 17.2%, with Level 2 is 9.26%, with Level 3 is 2.74% and with Level 3+ is 2.11%.

2.3 The Fay-Herriot model

We assume that we have a finite population P with dimension N partitioned into $d = 1, ..., D$ disjoint small areas. N_d is the population dimension in area d, thus $\sum_{d=1}^{D} N_d = N$. From P a random sample s with dimension n is selected, $\sum_{d=1}^{D} n_d = n$, where n_d denotes the sample size in area d. We are interested in estimating the mean of a variable denoted by Y, denoting worry about climate change, for area d, and this is denoted by $\bar{Y}_d$. A direct estimator for this is $\hat{\bar{Y}}_d^{DIR} = \sum_{i=1}^{n_d} y_{di} w_{di} / \sum_{i=1}^{n_d} w_{di}$,

[1] https://ec.europa.eu/eurostat/web/regions/database

[2] see https://news.booking.com/en/bookingcom-celebrates-one-year-of-travel-sustainable-with-new-product-features-for-accommodations-rental-cars-and-flights/

where w_{di} denotes the survey weight for unit i in area d. However, in case of small area sample sizes this will be unreliable [Rao and Molina, 2015]. In article, we apply the Fay-Herriot model in order to provide accurate and precise estimates of our study phenomena [Fay and Herriot, 1979]. This model consists in a sampling model:

$$\hat{\bar{Y}}_d^{DIR} = \bar{Y}_d + e_d, d = 1, ..., D \tag{1}$$

where e_d is the sampling error of the direct estimator, and a linking model:

$$\bar{Y}_d = \bar{\mathbf{X}}_d^T \beta + u_d, d = 1, ..., D, \tag{2}$$

and combining the two we obtain the following:

$$\hat{\bar{Y}}_d^{DIR} = \bar{\mathbf{X}}_d^T \beta + u_d + e_d, d = 1, ..., D, \tag{3}$$

where $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_{e_d}^2)$, with $\sigma_{e_d}^2$ (variance of the direct estimates) is assumed to be known. $\bar{\mathbf{X}}_d$ are the auxiliary variables (e.g., area level means) for area d. The Empirical Best Linear Unbiased Predictor (EBLUP) of $\bar{Y}_d$ under model 3 is given by [Fay and Herriot, 1979]:

$$\hat{\bar{Y}}_d^{EBLUP,FH} = \hat{\gamma}_d \hat{\bar{Y}}_d^{DIR} + (1 - \hat{\gamma}_d) \bar{\mathbf{X}}_d^T \hat{\beta}, \tag{4}$$

where $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_d}^2}$ is the shrinkage factor. Thus, when the n_d is small more weight will be given to the direct estimates, since they will reliable, and vice-versa. This optimises the trade-off between large variance of the direct estimator $\hat{\bar{Y}}_d^{DIR}$ and bias of the synthetic estimator $\bar{\mathbf{X}}_d \hat{\beta}$. The Mean Squared Error (MSE) of 4 can be estimated following Prasad and Rao [1990]. In this article, we use the Maximum Likelihood (ML) estimation technique to obtain the FH model estimates [Marhuenda et al., 2014].

3 Results

3.1 Modelling strategy

In order to provide model-based estimates of the proportion of individuals who is very worried and extremely worried about climate change, based on the FH model, we consider two scenarios: 1) only traditional variables from the Eurostat archive are available, and 2) additional variables are derived from web data. We address model selection and estimation issues in both scenarios separately and then we compare the results obtained in the two scenarios. We begin with a full model that includes all available covariates and we use a stepwise algorithm[3] to identify which covariates include in the FH model. To compare models, we employ the Kullback symmetric divergence (KICb2) criterion. This information criterion, developed for FH models, is used to assess dissimilarities between two statistical models and the model with the lowest KICb2 value should be preferred [Marhuenda et al., 2014]. Thus, the model with the lowest KICb2 value is selected for analysis in both scenarios.

3.1.1 Scenario 1: only traditional data

In this scenario, only the traditional data from the Eurostat data archive are available (see Section 2.2). Table 2 shows the results of model selection. The model with the lowest KICb2 value is selected, i.e., the one with long term unemployment (LTU) and proportion of people with tertiary education (t.edu).

[3] see the step function in the emdi R package Harmening et al. [2023]

Table 2: Model selection results based on the KICb2 for Scenario 1 (traditional data only, i,e, ESS and Eurostat Archive).

Predictors	**KICb2**
t. edu., LTU, robbery, pop. density	-8.24
t. edu., LTU, pop. density	-16.46
LTU, t. edu.	**-21.63**

3.1.2 Scenario 2: traditional and web data

In Scenario 2, in addition to the traditional variables included in Scenario 1, we also consider web data. The web data refers to the proportion of hotels with a specific environmental label level from Booking.com (see Section 2.2). Table 3 shows the values of the KICb2 metrics. We select the model with the lowest KICb2 value, which is the model that includes long term unemployment (LTU), population density (pop. density), and the proportion of hotels with level 1 and level 3 labels.

Table 3: Model selection results based on the KICb2 for Scenario 2 (traditional data, i.e, ESS and Eurostat data supplemented by web data from Booking.com).

Predictors	**KICb2**
t. edu., LTU, robbery, pop. density, level 1, level 2, level 3, level 3 plus	-10.42
t. edu., LTU, pop. density, level 1, level 2, level 3, level 3 plus	-14.25
LTU, pop. density, level 1, level 2, level 3, level 3 plus	-17.81
LTU, pop. density, level 1, level 3, level 3 plus	-22.05
LTU, pop. density, level 1, level 3	**-25.26**

3.2 Diagnostics

In order to evaluate whether we introduce bias in the final regional estimates produced by the FH models, we perform some diagnostics measures in both scenarios. We implement the Brown test [Brown et al., 2001] in order to evaluate the quality of the EBLUPs. For this the Wald statistics is used, with null hypothesis being the EBLUP estimates do not differ significantly from the direct estimates:

$$W = \sum_{d=1}^{D} \frac{(\hat{\bar{Y}}_d^{DIR} - \hat{\bar{Y}}_d^{EBLUP,FH})^2}{\hat{var}(\hat{\bar{Y}}_d^{DIR}) + \hat{MSE}(\hat{\bar{Y}}_d^{EBLUP,FH})} \tag{5}$$

This is approximately distributed as a χ^2 with D (number of areas) degrees of freedom, under the null hypothesis.

In Scenario 1 the correlation between synthetic part and direct estimator is 0.19, however, the EBLUP estimates do not differ significantly from the direct estimates (W=9.38, df=17 and p-value=0.93). In Scenario 2 the correlation between synthetic part and direct estimator is larger, i.e., equal to 0.62 and the EBLUP estimates do not differ significantly from the direct estimates (W=8.36, df=17 and p-value=0.96). According to these results, Scenario 2 should be preferred given the higher correlation between the two sets of estimates.

3.3 Comparing results

Figure 1 shows the the CV% and percentage Relative Root Mean Squared Error (RRMSE) % of the direct estimates and EBLUPs, respectively, across the regions. The RRMSE is defined as the ratio between the squared root of the MSE and the EBLUP. Here, we compare the direct estimates (Direct)

to the EBLUPs with (EBLUP with web data - Scenario 2) and without the (EBLUP with Traditional Data - Scenario 1) use of web data.

We can see that the use of auxiliary information from web data helps reducing the RRMSE% of the estimates considerably, for the areas with small sample size (i.e. less than 200 units). This gain is larger compared to the use of the model with traditional data only.

It can be seen that, in case of large regional sample sizes the RRMSE% estimates of the EBLUP with traditional data only are similar to the CV% of the direct estimates. On the contrary, the RRMSE% of the EBLUP obtained using web data is higher than the CV% of the direct estimates. However, in this case, the direct estimates are reliable due to very small CV%, thus, these should be used in practice.

Given the RRMSE results discussed above and the diagnostics results presented in the previous section, we consider Scenario 2 when mapping the estimates.

In Figure 2 we map the regional estimates for NUTS2 level in Spain obtained via the EBLUP approach and the following auxiliary variables LTU, pop. density, level 1, and level 3. It can be seen that areas with a larger presence of tourists and especially coastal regions, show larger level of worries about climate change. This is also valid for Madrid region. The region with the lowest value of the indicator is Balearic islands, followed by Castile-La Mancha and Castile-León.

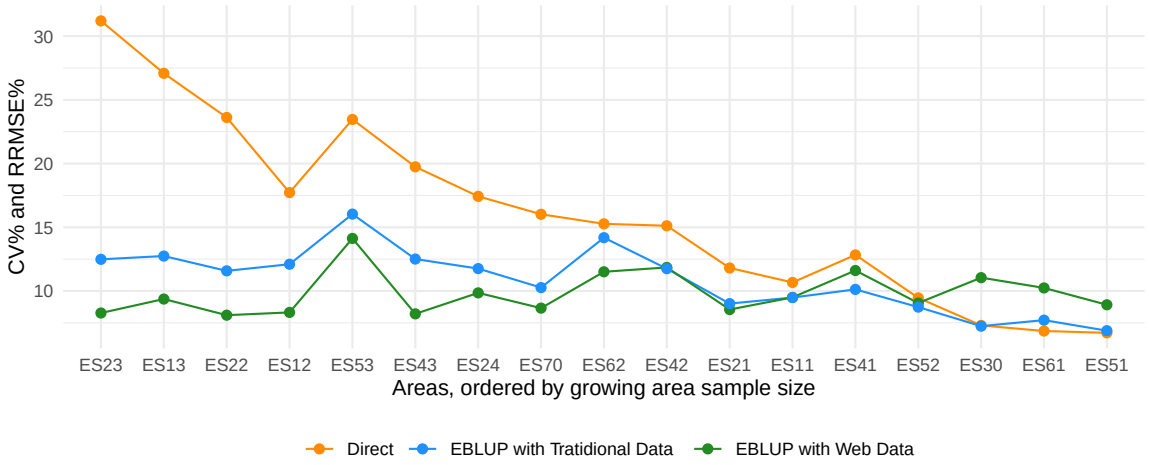

Figure 1: CV% (for the direct estimates) and RRMSE% (for the EBLUPs considering the two scenarios) of the regional estimates of worry about climate change for NUTS2 level in Spain.

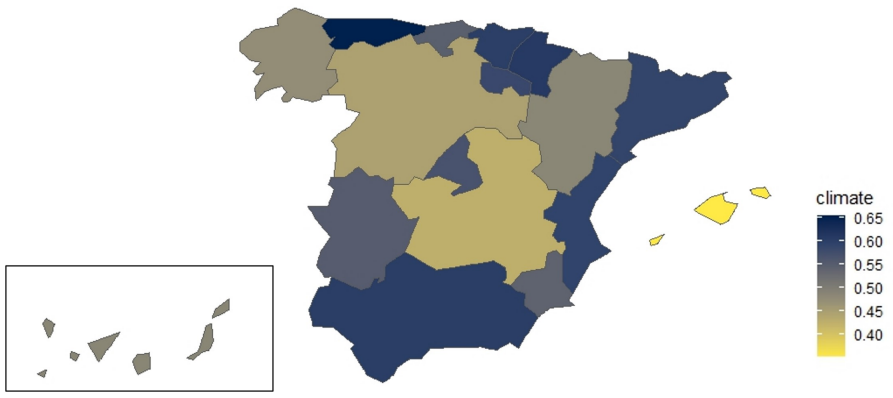

Figure 2: Regional Estimates NUTS2 level Spain of worry about climate change.

4 Concluding remarks

To effectively address climate change and other societal challenges, policymakers often require reliable estimates of relevant indicators at a sub-national level. Nationally representative surveys are not designed for this purpose. In this article, we discuss how SAE can address this issue. We construct a SAE model to estimate the proportion of people very and extremely worried about climate change at regional level in Spain. We employ both traditional and non-traditional (web data) variables as auxiliary information. Our results demonstrate that incorporating web data yields more reliable estimates than the ones produced by the model without those variables. Thus, our empirical analyses highlight the opportunity of using non-traditional data in SAE.

Future work will investigate the problem discussed in this article with a larger number of countries in the ESS, which means that the number of regions will also be larger. This will possibly show more evident gains in efficiency in the model-based small area estimates. Furthermore, users will be able to carry out comparisons between countries.

In addition, it is crucial to acknowledge and address certain drawbacks associated with the use of digital trace data for survey research. While these data sources offer valuable and innovative auxiliary information, their quality may be questionable. For instance, in our analysis, the availability of hotel data depends on scraping timing and request parameters, introducing selection biases. Additionally, web data are subject to volatility; changes in company policies, such as the *Travel Sustainable* programme in our case, can alter what we can measure with web data.

Research about the use of digital trace data in small area estimation, and more in general, in survey research is expanding. However, there remains a necessity for additional empirical evaluations and deeper exploration into the quality aspects of these data to fully understand the benefits and limitations of incorporating them in statistical models [Keusch and Kreuter, 2021].

References

Gary Brown, Ray Chambers, Patrick Heady, and Dick Heasman. Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. In *Proceedings of statistics Canada symposium*, volume 2001, pages 1–10. Statistics Canada, 2001.

European Social Survey. *European Social Survey round 10*, 2022. URL `https://www.europeansocialsurvey.org/news/article/round-10-data-now-available`.

Robert E Fay and Roger A Herriot. Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

Sylvia Harmening, Ann-Kristin Kreutzmann, Sören Schmidt, Nicola Salvati, and Timo Schmid. A framework for producing small area estimates based on area-level models in r. *R Journal*, 15(1), 2023.

Florian Keusch and Frauke Kreuter. Digital trace data: Modes of data collection, applications, and errors at a glance. In *Handbook of Computational Social Science, Vol 1*. Taylor & Francis, 2021.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barret, et al. Ipcc, 2023: Climate change 2023: Synthesis report, summary for policymakers. Technical report, IPCC, Geneva, Switzerland., 2023. URL `https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_FullVolume.pdf`.

Stefano Marchetti and Francesco Schirripa Spagnolo. Social big data to enhance small area estimates. *The Survey Statistician*, 89:59–67, 2024.

Stefano Marchetti, Caterina Giusti, Monica Pratesi, Nicola Salvati, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, Luca Pappalardo, and Lorenzo Gabrielli. Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2):263–281, 2015.

Stefano Marchetti, Caterina Giusti, Monica Pratesi, et al. The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. *AStA. Wirtschafts- und sozialstatistisches Archiv*, pages 1–15, 2016.

Yolanda Marhuenda, Domingo Morales, and María del Carmen Pardo. Information criteria for fay–herriot model selection. *Computational statistics & data analysis*, 70:268–280, 2014.

Angelo Moretti and Adam Whitworth. European regional welfare attitudes: a sub-national multidimensional analysis. *Applied Spatial Analysis and Policy*, 13(2):393–410, 2020.

Aaron T Porter, Scott H Holan, Christopher K Wikle, and Noel Cressie. Spatial fay–herriot models for small area estimation with functional covariates. *Spatial Statistics*, 10:27–42, 2014.

Aseem Prakash and Thomas Bernauer. Survey research in environmental politics: why it is important and what the challenges are. *Environmental Politics*, 29(7):1127–1134, 2020.

NG Narasimha Prasad and Jon NK Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171, 1990.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

John NK Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015.

Camilla Salvatore. Inference with non-probability samples and survey data integration: a science mapping study. *Metron*, 81(1):83–107, 2023.

Caterina Santi and Angelo Moretti. Carbon risk premium and worries about climate change. *Available at SSRN 3942738*, 2021. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3942738`.

Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2024. URL `https://rvest.tidyverse.org/`. R package version 1.0.4, https://github.com/tidyverse/rvest.