

Exploring patterns and determinants of agricultural carbon footprint in Northern Italy through small area estimation

Riccardo Pajno¹, Riccardo Borgoni¹, Felicetta Carillo²
Paolo Maranzano¹

¹ Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milano, 20126, Italy

² Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria - Centro di ricerca Politiche e Bioeconomia, Via Giacomo Venezian, 26, 20133 Milano, Italy

Abstract

In this paper, we aim to investigate the spatial dynamics of the carbon footprint associated with farms in the Po Valley (northern Italy), one of the most polluted areas in the world, where intensive livestock farming plays a crucial role in driving high levels of ammonia and particulate matter in the atmosphere. Specifically, we investigate the relationship between the agricultural carbon footprint, economic and techno-productive characteristics of farms (e.g., number and type of livestock or available farmland, land value), and landscape characteristics (e.g., orography) for 2020 using small-area spatial models belonging to the Fay and Herriot (1979) framework and which leverage on several spatially-varying auxiliary information. The goal is to provide accurate and credible estimates of the environmental impacts of agricultural activities at a high level of spatial granularity, but at the same time to enable the development of robust agricultural policies based on data-driven evidence and able to capitalize on the natural complementarity between institutional data sources and sample estimates.

Key Words: Agricultural Carbon Footprint; Small Area Estimation; Spatial Fay-Herriot model; CAMS Ammonia Emissions.

1. Introduction

The agricultural sector is undergoing a period of rapid transformation, driven by the interrelated and pervasive impacts of climate change, demographic shifts, and disparities in economic growth across territories and populations worldwide. It is incumbent upon governments and local authorities to address these challenges in order to safeguard vulnerable socio-economic systems and territories and to facilitate positive change in rural economic development. This necessitates the undertaking of small-scale studies and an analysis of the economics of the agricultural sector and the businesses operating within it. In particular, the actions of policymakers should be guided by the implementation of territorial policies that must necessarily take into account local conditions and specificities, rather than being limited to considerations at the supply chain level.

The availability of accurate, up-to-date, and relevant data is of critical importance for this understanding, which has often constituted the primary obstacle to rigorous research and, ultimately, effective policymaking in the agricultural sector. The scarcity of high-quality data necessary to examine the evolution of agricultural systems at the integrated and micro-area levels necessitates the implementation of techniques with technical solutions that yield more precise estimates from available data and more effectively utilize complementarities between traditional and alternative data sources.

The principal aim of this paper is to provide accurate and reliable estimates of the environmental impact of agricultural activities at a high level of spatial resolution, while simultaneously facilitating the development of evidence-based agricultural policies. In particular, our objective is to examine the spatial dynamics of the carbon footprint associated with agricultural farm holdings in the Po Valley (Northern Italy), which is one of the most polluted areas in the world and where intensive livestock farms play a significant role in driving high levels of ammonia (Fassò et al., 2023) and atmospheric particulate matter (Otto et al., 2024; Rodeschini, Fassò, Finazzi, & Fusta Moro, 2024). In the Italian context, the most appropriate observational units are agrarian regions, a territorial subdivision defined by the Italian Ministry of Agriculture that aggregates municipalities into internally homogeneous areas with similar agricultural characteristics. Notable examples at the international scene are Oltrepo Pavese, renowned for its viticulture, and Lomellina, celebrated for its rice production. The entire area under consideration, that is, the Po Valley, is officially divided into 254 distinct areas.

This study examines the relationship between agricultural carbon footprint, economic and technoproductive characteristics of farms (e.g., number and type of farm animals or available agricultural areas, soil quality), and landscape features (e.g., orography) for the year 2020. The study is based on a comparison between direct estimates of the carbon footprint and small area spatial models belonging to the family of the Fay and Herriot (1979) framework, which utilize a variety of spatially varying covariates. Small area models for Italian agrarian regions have been previously presented in other empirical studies, including those by Maranzano, McConville, Otto, and Carillo (2023) on the relationship between manure and pollution in Lombardy and Carillo, Maranzano, Marcis, Pagliarella, and Salvatore (2024) on the estimation of key economic dimensions in subregional agricultural systems.

2. Data and area of interest

The data used in this paper comes from several institutional sources, namely the Farms Accountancy Data Network (FADN) survey, the 2020 Italian National Census of Agriculture, and the National Veterinary Registry Office Database. We also consider some cartographic and orographic information related to the area of interest for this study and incorporated it into our analysis.

2.1 Area of interest: the Italian Po Valley

As mentioned above, the area of interest for this study is referred to as the Po Valley throughout the paper (left panel of Figure 1). This term denotes the region traversed by the Po River, which is geographically defined as the union of four Italian administrative regions¹, namely Lombardy, Piedmont, Veneto, and Emilia Romagna.

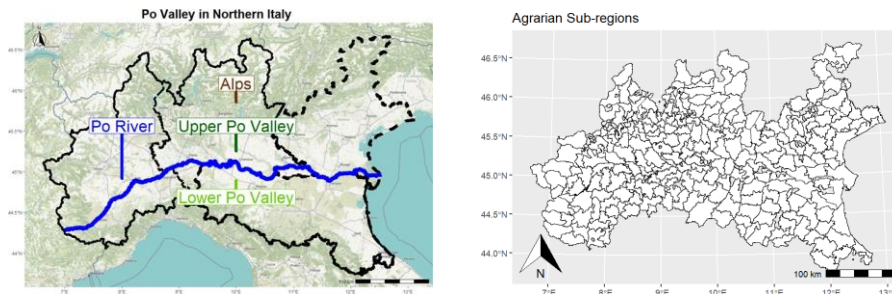


Figure 1: The Po Valley and agrarian sub-regions

¹ To reference countries' regions for statistical purposes, the European Union has developed a classification known as NUTS, Nomenclature of Territorial Units for Statistics (Eurostat, 2023), that divide each EU country into several geographically nested levels.

The final goal of this study is to provide estimate of the carbon footprint for small geographical areas called agrarian Sub-Regions (ASR). ASR are groups of municipalities belonging to the same province and created according to territorial contiguity rules, that can be considered homogeneous according to natural and agrarian characteristics. The total number of ASRs in the study area is 254, covering 3,572 municipalities out of the approximately 7,900 across the whole country. The surface area of the ASRs ranges from a minimum of 6,863,901 m² in Montagna del Verbano Occidentale to a maximum of 983,896,043 m² in Alto Taro, with an average surface area of 350,702,335 m². The right panel of 1 shows the map of the Po Valley decomposed into the 254 ASR.

This area is known to be the economic engine of the country. In particular, intensive crops and breeding are spread all across the lowlands providing food both for domestic consumption and export. This intensive farming activity (Fassò et al., 2023), together with its geographical position between two mountain chains (Raffaelli et al., 2020) and its high level of industrialization and urbanization (Regional Statistical Yearbook, 2017) makes the Po Valley one of the most polluted areas in the world.

2.2 The RICA dataset

Farm-level data on carbon footprint and techno-economic specialization are provided by the Farms Accountancy Data Network (FADN) survey, which collects detailed structural and accountancy data on EU farms and is the only harmonized European statistical source on the economic management of farms. At the Italian level, the surveys are conducted by the Italian Council for Research in Agricultural Economics (CREA) through the RICA (Italian acronym standing for “Information Network for Agricultural Accounting”) database. This database has been created in accordance with Law CEE 79/56 of the European Union in order to collect data about the European agriculture. RICA data are harmonized with the European FADN (Farm Accountancy Data Network) that includes a database for each member of the Community. The data include information from a random sample of agrarian businesses selected from the entire population of Italian farms with a Standard Output² of no less than 8.000 euro per year. The sample encompasses 37% of the Italian businesses and 90% of the overall Italian Standard Output. The sample is stratified according to the following three criteria: geographical location (regions), type of farming and economic size.

Type of farming refers to the business’s productive orientation and it is articulated into three levels: general, main and specific. Economic size refers to the economic dimension according to the Standard Output. RICA collects unit-level information on various aspects of the farm, including structural, geographical, productive, environmental, economic, and financial factors (Banca Dati RICA, 2024).

The RICA sample considered in our study includes 2.806 farms spread across the four regions as follows: 858 farms in Emilia Romagna, 674 farms in Lombardia, 600 farms in the Piemonte and 674 farms in the Veneto. Farms are stratified according to three typologies of activity: crops, breeding, and mixed. The majority of farms are crop farms (1.934 farms) followed by the breeding farms (763 farms), while only a few are mixed farms (109 farms). The prevalence of crop businesses is particularly significant in Emilia-Romagna and Lombardy, while breeding farms are more common in Piedmont and Veneto. Regarding the size of the farms, farms are classified based on whether their standard output was below €50,000 (small farms) or above €50,000 (large farms). We observed nearly equal numbers of small and large farms (1,526 versus 1,280). Small farms are more prevalent in Emilia-Romagna, while larger farms are concentrated in Veneto.

² Standard Output measures the potential gross product evaluated using the potential production of yields and the potential prices

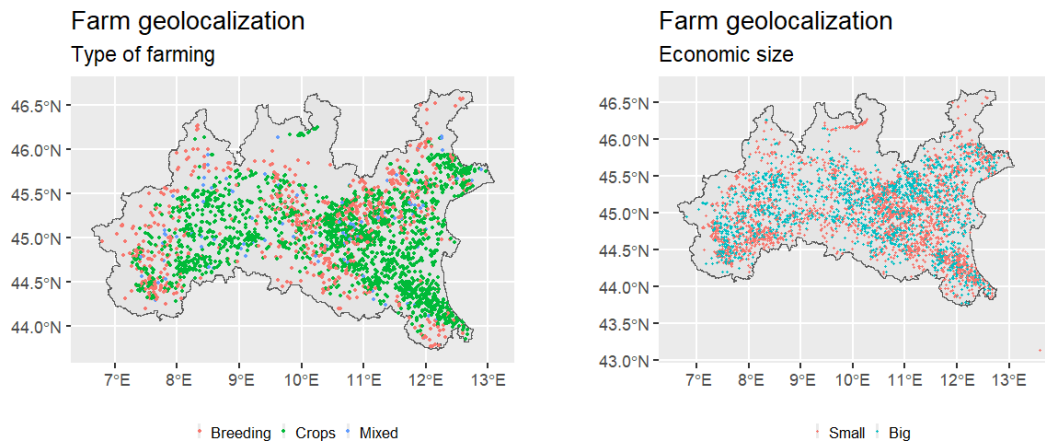


Figure 2: Farm geolocalization by type of farming and economic size.

Figure 2 illustrates the spatial distribution of the farms in the sample, revealing a non-uniform distribution across the region with most farms concentrated in the fertile, flat areas of the central Po Valley. In particular, the northern and southern mountainous regions mostly feature small-scale livestock operations. Additionally, the eastern part of the Po Valley shows a clear predominance of crop farming over livestock, whereas the western part exhibits a more balanced distribution between the two types of agriculture.

A farm specific emission indicator of greenhouse gases (GHGs) intensity, namely the carbon footprint indicator, is included in the Italian FADN dataset. This variable has been estimated starting from micro data on the production techniques adopted on the farm and using conversion rates of emissions and of absorption of GHG, taken from specific literature. In details, to reconstruct a GHG farm balance, it was adapted the Intergovernmental Panel on Climate Change methodology at the farm level, using activity data connected to the main agricultural activities (Coderoni, Bonati, Longhitano, Papaleo, & Vanino, 2013). Methane, nitrous oxide, and carbon dioxide (CO₂) emissions are estimated from the following source categories: livestock production, crops, land use, fuel and fertilizers. These different farm-level GHG emissions are then summarised into a unique indicator using each GHG's Global Warming Potential (GWP). The conversion factors updated over time by the IPCC are used. GHG emissions are expressed in CO₂ equivalent and quantify the Carbon Footprint at farm level (Baldoni, Coderoni, & Esposti, 2017).

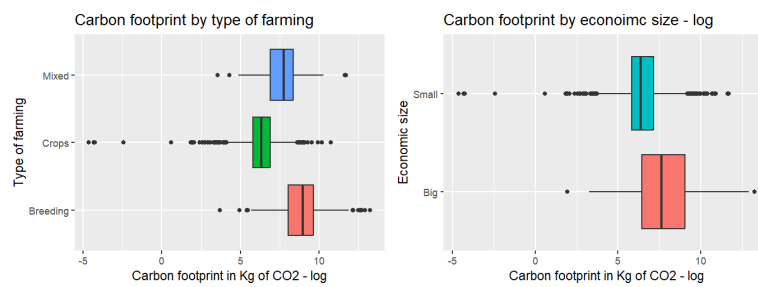


Figure 3: Carbon footprint distribution conditioned to type of farming and economic size.

Figure 3 displays the distribution of the carbon footprint on the log scale by type of farming and economic size. As expected, livestock farming is significantly more polluting than crop farming, and larger farms have a higher carbon footprint than smaller ones. Also, the boxplot reveals a wide

range and a strong positive skewness, with several large values corresponding to large, intensive breeding farms.

2.3 The Italian National Census of Agriculture

The Italian National Census of Agriculture involves the entire population of agrarian businesses in the Italian territory. It occurred every 10 years and the last sweep, which is considered in this paper, took place in 2020. In 2020, there were 235,366 farms in the Po valley (46,893 farms in Lombardy, 51,703 in Piedmont, 83,017 in Veneto, and 53,753 in Emilia Romagna), showing a significant decrease compared to the 314,331 farms recorded in the previous census in 2010. This change reflects a percentage decrease of approximately 24.4%. This decreasing trend is consistent across all four regions. In 2010, there were 54,333 farms in Lombardy, 67,148 in Piedmont, 119,384 in Veneto, and 73,466 in Emilia Romagna. Note that there were no significant changes in the Utilized Agricultural Area (UAA) over the decade, as UAA decreased slightly from 3,873,000 hectares in 2010 to 3,829,000 hectares in 2020 (ISTAT, 2020). As a result, we have witnessed an expansion in the size of businesses in the Po Valley, driven by capital inflows and the entry of investment funds. Although family-managed farms still account for 90% of the total, there has been growth in the number of corporate farms, which are typically larger in size. For example, in Lombardy, the average UAA of a corporate farm is 40 hectares, compared to 16 hectares for a family-managed farm (ISTAT, 2020).

2.4 The National Veterinary Registry Office Database

The National Veterinary Registry Office Database (BDN, Banca Dati Nazionale Anagrafe Veterinaria) offers a detailed amount of information about all the farm animals living in the entire Italian territory. This database is maintained by the Italian Ministry of Health for sanitary purposes. Animals are categorized into four groups: bovines, swine, ovine, and equines. However, in this study, we focus only on bovines and swine due to their significant impact on pollution. In the area considered, the total number of bovines is roughly 3.68 million, while the total number of swine is about 7.45 million, with significant differences across various ASR.

2.5 Orography

As mentioned earlier, the Po Valley has peculiar orographic characteristics. Therefore, we collected data from the ISTAT about the average altitude and soil slope in order to better explain the feature of the areas and their farms. The altitude standard deviation is considered a reasonable measure of soil slope. agricultural sector is undergoing a period of rapid transformation, driven by the interrelated and pervasive impacts of climate change, demographic shifts, and disparities in economic growth across territories and populations worldwide. It is incumbent upon governments and local authorities to address these challenges in order to safeguard vulnerable socio-economic systems and territories and to facilitate positive change in rural economic development. This necessitates the undertaking of small-scale studies and an analysis of the economics of the agricultural sector and the businesses operating within it. In particular, the actions of policymakers should be guided by the implementation of territorial policies that must necessarily take into account local conditions and specificities, rather than being limited to considerations at the supply chain level.

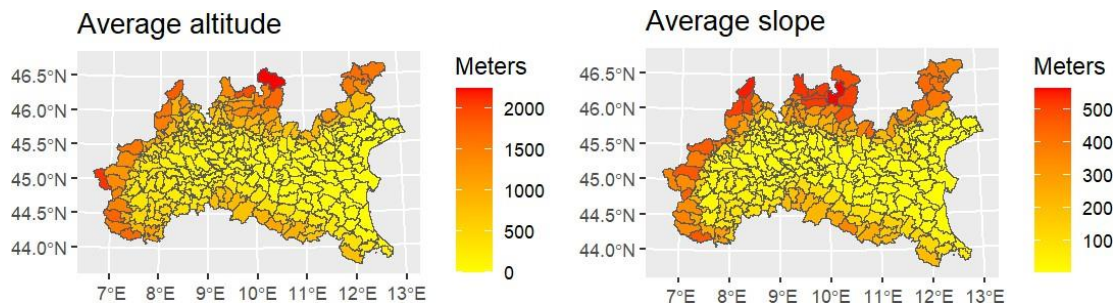


Figure 4: Altitude and soil slope spatial distribution

3. Small area Estimation of Carbon Footprints for Agrarian Sub-Regions: the model

3.1 Direct estimation

We assume that the population, denoted as U , consists of N units, divided into m disjoint small areas. In our case, the population consists of $N = 235,360$ farms operating in the Po Valley, and the areas are the $m = 254$ ASRs. The population is split into $n = 2806$ sampled farms and $N - n$ non-sampled farms.

We use the subscript i below to indicate the restriction to area i , hence n_i and N_i denote the sample size and the population size in area i , respectively. Let y_{ik} denote the variable of interest, that is, the Carbon Footprint of farm k in area i . An estimator for the population total of the Carbon Footprint τ_i in area i is given by

$$\hat{\tau}_i^{direct} = \sum_{i=1}^{n_i} \frac{y_{ik}}{\pi_{ik}} = \sum_{i=1}^{n_i} y_{ik} w_{ik} \quad (1)$$

where w_{ik} is the probability of including unit i in area k in the sample and the unit weight satisfies

$$\pi_{ik} = \frac{1}{w_{ik}}$$

The sample is selected from the population using the sampling design described above, and the inclusion probabilities of each single unit are implicitly defined by the sampling stratification. However, the sampling was not designed at the desired ASR level; instead, it was conducted at a national level.

As we are interested in estimating the agricultural carbon footprint at this finer spatial scale, a proper weighting scheme is necessary to compute reliable small area estimates. For this reason, we post-stratified the sample using census information, meaning we calculated weights for the sample units after sampling, using auxiliary information on the agrarian region. More specifically, the 2020 Census contains information on the number of farms by economic size and technical specialization at the municipal level. Since each municipality belongs to one and only one agrarian sub-region, we were able to aggregate Census information at the ASR level.

Then we calculate the joint frequencies of farm holdings by economic size (small vs. large farms), technical specialization (crop specialized, livestock specialized, and mixed farming), and agrarian sub-region. We then calculate a new proportional weighting scheme for each farm k in the RICA sample located in the generic i -th ASR as:

$$\tilde{w}_{istk} = \frac{N_{ist}}{n_{ist}}$$

Where $i = 1, \dots, m = 254$ indicates the considered ASR, $s = 1,2$ represents the economic size, and $t = 1,2,3$ stands for the farming typology. Once the new weights have been obtained, a direct estimate using the new stratification scheme can be calculated for each ASR using equation (1) where the w -weights are replaced by the new \tilde{w} scheme, namely

$$\tilde{\tau}_i^{direct} = \sum_{s=1}^2 \sum_{t=1}^3 \sum_{k=1}^{n_{ist}} y_{istk} \tilde{w}_{ik} \quad (2)$$

3.2 The Spatial FH model

To obtain small area estimates of the Carbon Footprint for the ASRs in the region of interest, the Fay-Herriot method (Fay & Herriot, 1979) has been applied.

The area level model proposed by Fay and Herriot (the FH model) links the direct estimates with area level covariates. The FH model is based on two stages: the sampling model (first stage)

$$\tilde{\tau}_i^{direct} = \tau_i + \epsilon_i$$

and the linking model (second stage)

$$\tau_i = x_i\beta + u_i$$

Where x_i and β denote the $p \times 1$ vectors of area level covariates and regression parameters respectively. The sampling errors are assumed to be 0-mean homoscedastic normally distributed and independent random variables. The random effects u_i are also assumed to be independently normally distributed. For additional details we refer to Rao and Molina (2015). The combination of both models leads to an area level linear mixed model.

Given the expected spatial dependency across geographically close regions, a simultaneous autoregression (SAR) (Whittle, 1954) random effect was incorporated to account for potential interactions between neighboring areas (Pratesi & Salvati, 2008). In matrix form the models writes:

$$CF = \mathbf{X}\beta + \mathbf{Z}u + \epsilon$$

where CF is the column vector of the 254 direct estimates of the ASR Carbon Footprint, and \mathbf{X} is a 254×6 design matrix that includes the values of the measured covariates described in Section 2, that is

- the utilized Agrarian Surface
- the number of standard working days in agriculture
- the total bovine heads
- the total swine heads
- average altitude
- altitude standard deviation

and β is the corresponding vector of regression coefficients.

In the random effect term, \mathbf{Z} is a 254×254 matrix of known positive constants, and u is a column vector of error terms. In the SAR model, it is assumed that

$$u = \rho\mathbf{W}u + v$$

where ρ is the parameter representing the strength of the spatial dependence, v is a vector of i.i.d. Gaussian random variables, and \mathbf{W} is a row-standardized 254×254 contiguity matrix, defined as $w_{ij} = \frac{I(i \sim j)}{\#\partial(i)}$, where $I(i \sim j)$ indicates that region i is adjacent to region j , $\partial(i)$ is the set of neighbors of region i , and $\#$ represents the size of the set. Hence, this can equivalently be rewritten as $u = (\mathbf{I} - \rho\mathbf{W})^{-1}v$, where \mathbf{I} is the 254×254 identity matrix. Therefore, the model can be rewritten as:

$$CF = \mathbf{X}\beta + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}v + \epsilon$$

It is possible to create higher order SAR models where the spatial term incorporates different neighborhood structure in the form $\rho_1W_1 + \rho_2W_2 + \dots + \rho_lW_l$ with W_b specifying neighborhoods of different order (Haining, 1993). However, most SAR applications consider a single spatial

parameter, since increasing the number of spatial parameters complicates the estimation process. Estimation has been carried out via REML using the R package *sae* (Molina & Marhuenda, 2015).

The model allows one to smooth the direct estimates of Carbon Footprint in a certain ASR d by means of a set of covariates. Then, we shrink this estimate towards the corresponding direct estimate by computing the Empirical Best Linear Predictor (EBLUP)

$$\widehat{CF}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{\epsilon_i}^2} \hat{t}_i + \frac{\sigma_{\epsilon_i}^2}{\hat{\sigma}_u^2 + \sigma_{\epsilon_i}^2} x_i' \hat{\beta}$$

(Rao & Molina, 2015) where

$$\sigma_{\epsilon_i}^2 = \sum_{j=1}^3 \sum_{k=1}^2 \frac{N_{djk}^2}{\widehat{N}_{djk}} \sum_{i=1}^{n_{djk}} w_{djk} (1 - w_{djk}) (y_i - \bar{y}_{djk})^2 \quad (3)$$

where $\widehat{N}_{djk} = \sum_{j=1}^3 \sum_{k=1}^2 w_{djk}$ (Morales, Esteban, Pérez, & Hobza, 2021). These estimates are a convex combination of direct and model estimates and coincide with the latter when there are no observations in a given area.

4. Small area Estimation of Carbon Footprints for Agrarian Sub-Regions: Empirical results

The empirical analysis of the agricultural carbon footprint in the Po Valley commences with the estimation of the total footprint for each ASR, which is achieved through the utilization of the direct Horvitz–Thompson estimator. The resulting numerical estimates are presented in graphical form in Figure 5.

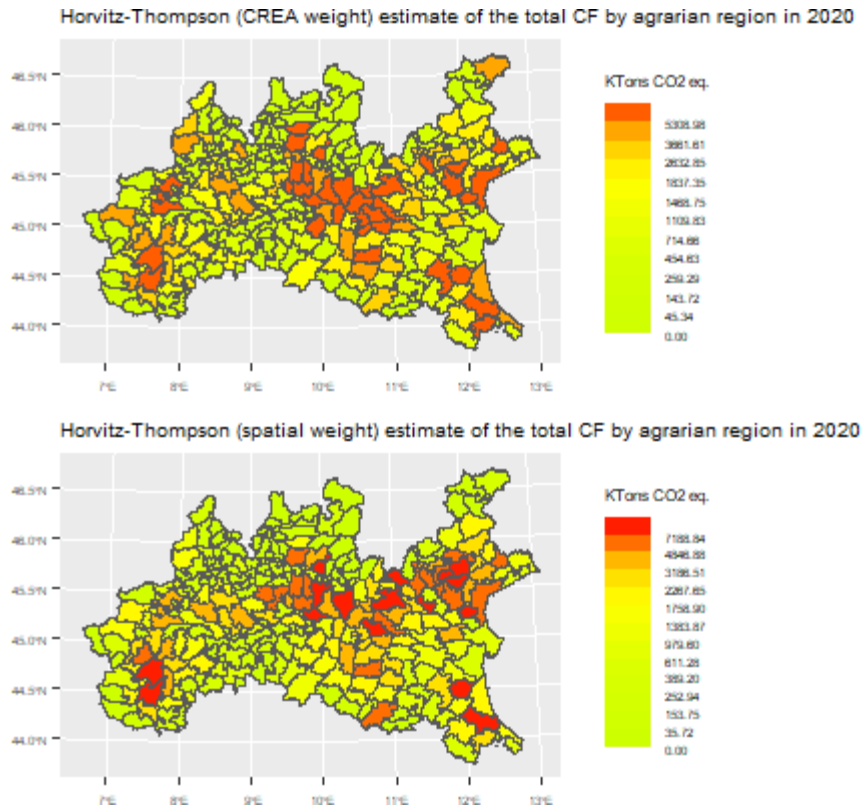


Figure 5: Direct estimates for Agro-Carbon Footprint

Notwithstanding their pertinent inferential attributes (e.g., unbiasedness), direct estimates have been observed to only partially encapsulate the spatial distribution of the agro-carbon footprint. In particular, Horvitz-Thompson estimates tend to assume values for the total of the variable of interest that are either very low or very high. Furthermore, it is notable that certain regions within the Alps, which are presumed to be relatively unpolluted, exhibit remarkably elevated levels of carbon footprint.

The findings confirm that the direct estimator has several inherent limitations. First, the Horvitz-Thompson estimator is based on the assumption that the sampled farms are completely representative of the population and does not rely on any additional information. In the extreme case where the sample size is null but the census size is not, we obtain null values for the total. In particular, the RICA sample is not representative at the ASR level, as the stratified design pertains to a different and less granular profile (i.e., regional stratification). Secondly, the scarcity of the sample may have resulted in the generation of unreliable results, including negative estimates of the carbon footprint. To address these challenges, we employed a post-stratification strategy, as detailed in Section 3.1. The bottom panel of Figure 5 depicts the post-stratified HT numerical estimates. The incorporation of new spatially representative weights has led to notable enhancements in the map, with a discernible increase in smoothness and the absence of outliers in the mountainous regions in the north.

Subsequently, a model-based estimation of the agro-carbon footprint was obtained through the implementation of the classical Fay-Herriot model (FH) and the spatial Fay-Herriot model (SFH), which accounted for the auxiliary information previously outlined. Numerical estimates are represented in Figures 6 and 7, respectively.

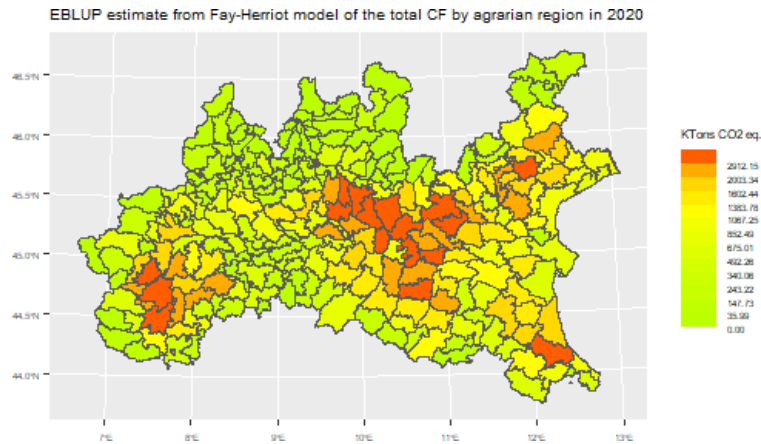


Figure 6: Fay-Herriot EBLUP for agro carbon footprint

As illustrated in Figure 6, the EBLUP estimates of the agro-carbon footprint are more smoothly delineated than their direct estimate counterparts. This is due to the fact that each EBLUP not only utilizes information from its respective region but also exploits the auxiliary variables, which provide information about the entire group of small areas included in the parameters β and σ^2 in common to all the EBLUPs.

Variable	Estimate	SE	t-value	p-value
Intercept	-70.4157	76.5902	-0.9194	0.3579
Utilized agricultural surface	0.0077	0.0044	-1.7336	0.0830
Standard working days in agriculture	0.0028	0.0003	8.7939	0.0000
Total bovine heads	0.0263	0.0065	4.0259	0.0001
Total swine heads	0.0024	0.0024	1.0216	0.3070
Average altitude	0.1293	0.1633	0.7917	0.4285
SD altitude	-0.4232	0.5097	-0.8393	0.4064

Table 1: Table of coefficients for the Fay-Herriot model.

Table 1 presents the numerical estimates of the regression coefficients for the FH model. With the exception of the coefficients for SD altitude and utilized agricultural surface, all the estimates are positive. This is to be expected, given that the agricultural carbon footprint is expected to be positively correlated with the extension of the utilized surface and the amount of number of working days and the number of livestock heads are expected to have a positive correlation, while the altitude and the slope are expected to have a negative correlation. However, the marginal effect of the utilized surface is negative, while the average altitude is positive. This discrepancy may be caused by multicollinearity between some variables. Moreover, the variables associated with the labor market and the quantity of bovines are highly significant, while the variable linked to economic size is marginally significant. The remaining variables are not statistically significant.

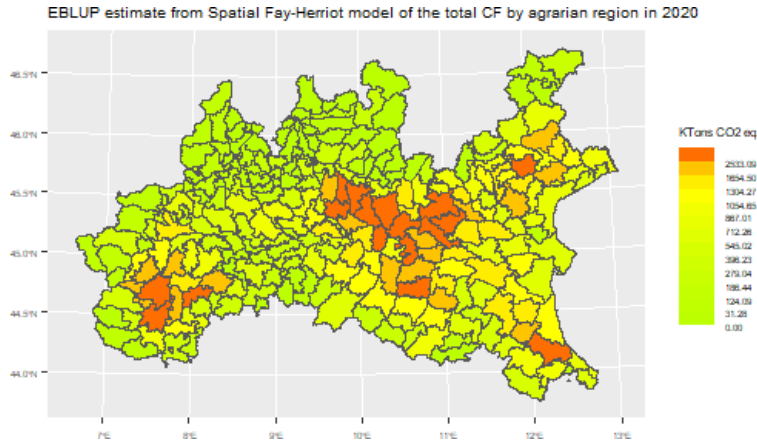


Figure 7: Spatial Fay-Herriot EBLUP for agro carbon footprint

As illustrated in Figure 7, the spatial EBLUP estimates exhibit a noticeable degree of similarity to their classical counterparts. However, a discernible level of spatial persistence is also evident, whereby adjacent agrarian sub-regions tend to display greater similarity in their estimated agro-carbon footprints.

Variable	Estimate	SE	t-value	p-value
Intercept	-39.9614	62.5573	-0.6388	0.5230
Utilized agricultural surface	-0.0058	0.0035	-1.6408	0.1008
Standard working days in agriculture	0.0024	0.0002	9.8119	0.0000
Total bovine heads	0.0102	0.0051	1.9991	0.0456
Total swine heads	0.0059	0.0019	3.1075	0.0019
Average altitude	0.0979	0.0906	1.0798	0.2803
SD altitude	-0.3434	0.2999	-1.1451	0.2522

Table 2: Table of coefficients for the spatial Fay-Herriot model.

In Table 2 we report the numerical estimates of the regression coefficients for the SFH model. We can notice that signs and magnitudes of the spatial Fay-Herriot model are consistent with the traditional model one due to a positive medium level spatial persistence ($\hat{\rho} = 0.6164$). However, we have a great amount of the significance of the total of swine heads, while the bovines have a higher p-value than before: they are still significant at a 5% level but not anymore at a 1% level. Moreover, the labor is still very significant while we observe a loss in significance for the UUA.

To assess the efficacy of the small area estimators, we undertake a comparison of their root mean squared error (RMSE) at the area level. While the variance of Horvitz-Thompson estimators can be calculated with certainty, the variance of the EBLUPs must be determined through a parametric bootstrap, as proposed by Rao & Molina (2015). We expect that variances will decrease as the model becomes more complex especially if the survey rate is small. Maps for the four RMSE are available in Figure 8.

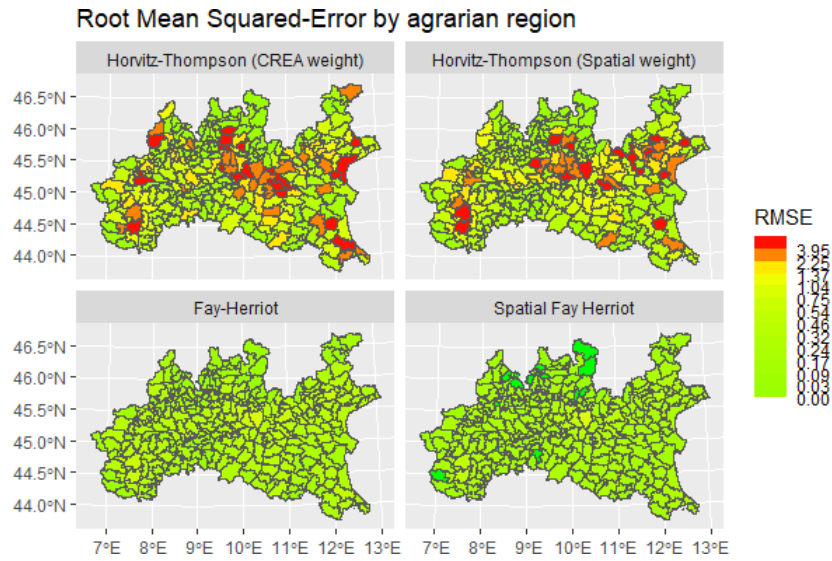


Figure 8: RMSE for different estimates

As can be seen from the map, there is a notable reduction in variability when moving from direct estimation to the model-based approach. Furthermore, a slight yet noteworthy enhancement is evident even when the spatial correlation is incorporated into the model. It is also noteworthy that the largest values of the root mean square error (RMSE) are observed to correspond to the largest values of the punctual estimates across all four models.

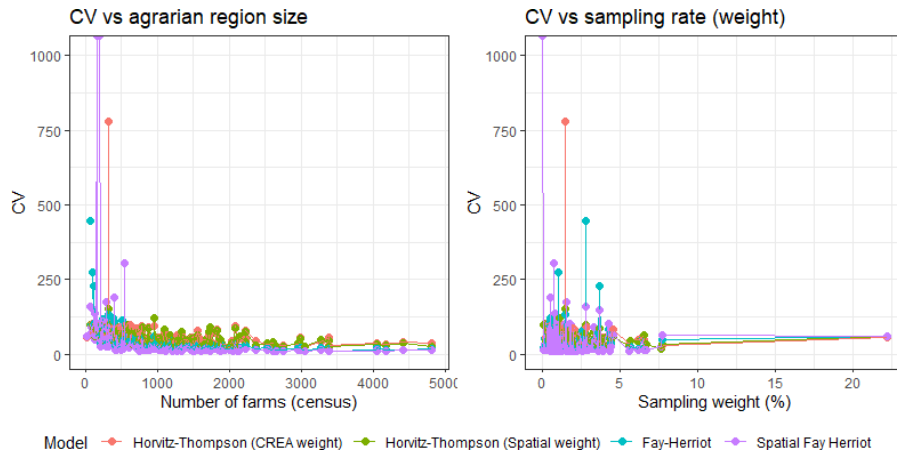


Figure 9: Coefficients of variation vs census and sample size

Figure 9 illustrates the behavior of the coefficient of variation as the sampling rate and census size increase. The plot corroborates the benefits of a model-based approach, as evidenced by the lower coefficients of variability compared to their direct estimates. The discrepancy is particularly pronounced when census size and sampling rate are minimal. Conversely, as the census size increases, the variability of all four estimates decreases, and their difference approaches zero.

5. Small area Estimation of Carbon Footprints for Agrarian Sub-Regions: Validation using satellite data

Once we obtained the numerical estimates of the carbon footprint of farms at the agricultural sub-regional level, we computed a measure of the goodness of our predictions in order to validate them. The validation is done by comparing the area-level carbon footprint with the ammonia emissions provided by the Copernicus-CAMS satellite program. The Copernicus Atmosphere Monitoring Service³ (CAMS), managed by ECMWF, is one of the most recent global databases of emissions and concentrations from anthropogenic sources. The CAMS datasets provide annual emission inventories for many atmospheric compounds from 2000 to 2020. These inventories are based on a combination of existing datasets and newly updated information, describing anthropogenic emissions from fossil fuel use on land, natural emissions from vegetation, soil and more. Anthropogenic land-based emissions are further disaggregated by sector of activity (e.g., transportation, agriculture). Pollutant emissions data are provided by the CAMS anthropogenic emissions dataset, which contains monthly global anthropogenic and natural emissions from 36 sources, again divided into 20 sectors (including agriculture and livestock) on a regular grid with a spatial resolution of $0.1^\circ \times 0.1^\circ$. The gridded data are then aggregated at the area level by summing the monthly emissions of the cells falling within the area boundaries.

In the following, we show the preliminary results of the validation experiment, in which we fitted the following linear regression model

$$NH3_i = \theta_0 + \theta_1 \widehat{CF}_i + \varepsilon_i \quad i = 1, \dots, m = 254 \quad (4)$$

where $NH3_i$ is the quantity of emissions estimated by the CAMS satellite for area i and \widehat{CF}_i is the amount of carbon footprint estimated using one of the small area models discussed above. The goodness-of-fit measure we consider for validating the estimates is the R^2 index. In particular, we consider the small area estimates to be consistent as the R^2 index tends to increase.

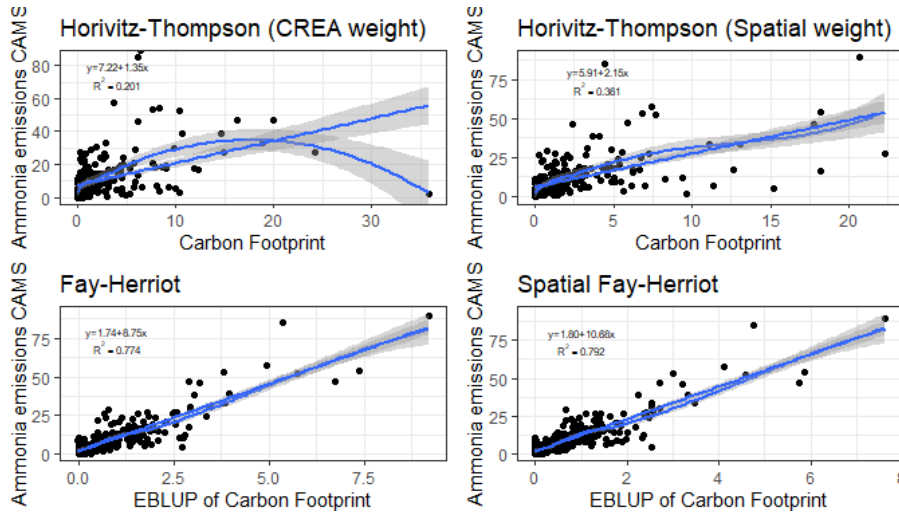


Figure 10: Consistency evaluation using linear models and their R^2 indices

In Figure 10 we show the scatterplot and the estimated linear relationship between the satellite and small area estimates for the four survey estimators used to obtain the values. From the scatter plot we can see that the R^2 index increases as the model used to obtain the estimates becomes more complex. In particular, the plot shows a large gap between the direct estimates and the model-based

³<https://ads.atmosphere.copernicus.eu/cdsapp!/dataset/cams-global-emission-inventories?tab=overview>

techniques, providing strong evidence in favor of the latter. The blue curves represent linear and non-linear smoothed regression functions. For the model-based estimators, we see that the linear and nonlinear curves tend to overlap, while a difference in trend can be seen for the direct HT estimates. Finally, in Figure 11 we show the maps of the numerical difference between the satellite data and the small area estimates. As already noted, increasing complexity in the regression models leads to a decrease in the discrepancy.

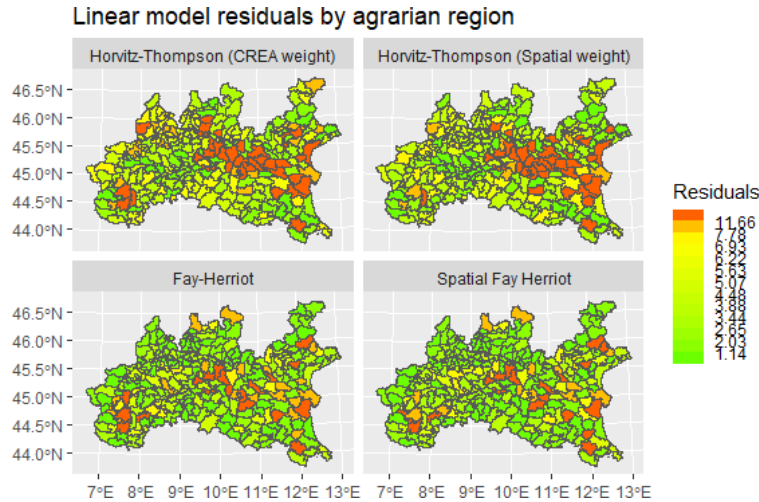


Figure 11: Difference between actual and fitted values

6. Conclusions and final remarks

In this paper, we investigated the spatial dynamics of the carbon footprint associated with farms in the Po Valley, northern Italy, using small-area methods that exploit several spatially variable covariates. In particular, we used linear mixed regression models to provide reliable estimates of the agricultural carbon footprint by exploiting the economic and techno-productive characteristics of farms and landscape features. Data on carbon footprint and techno-economic specialization at the farm (unit) level were provided by the FADN program, the European sample survey system that collects detailed structural and accounting data on agriculture and livestock. Among others, a farm-specific greenhouse gas (GHG) emission indicator, the carbon footprint indicator, was used to represent the environmental impact of agricultural activities.

Empirical results have clearly shown that the small area model-based approach significantly outperforms direct estimators that do not use auxiliary information. The carbon footprint was found to be very heterogeneous among the agricultural sub-regions, while remaining coherent with the economic and environmental configuration of Northern Italy. In particular, the carbon footprint is very high in the plain areas, where intensive livestock farming is widespread. To ensure consistency between sample and institutional estimates of pollution, we developed a validation analysis in which small area estimates were correlated with satellite data on ammonia emissions. Again, modelling makes the estimates much more consistent with the institutional estimates.

References

- Baldoni, E., Coderoni, S., & Esposti, R. (2017). The productivity and environment nexus with farm-level data. The Case of Carbon Footprint in Lombardy FADN farms. *Bio-based and Applied Economics*, 6(2), 119-137.
- Carillo, F., Maranzano, P., Marcis, L., Pagliarella, M. C., & Salvatore, R. (2024). The spatio-temporal Fay-Herriot model using the state-space method: an application to Italian Lombard agrarian sub-regions. In *Book of Short Papers - 2nd Italian Conference on Economic Statistics (ICES 2024) - Statistical Analysis of Complex Economic Data: Recent Developments and Applications* (pp. 66-69). University of Florence - Department of Statistics, Computer Science, Applications "Giuseppe Parenti" (DiSIA): Casa Editrice Bonechi, Via Scipione Ammirato, 100 - 50136 Firenze (FI), info@bonechibooks.it, www.bonechi.it.
- Coderoni, S., Bonati, G., Longhitano, D., Papaleo, A., & Vanino, S. (2013). Impronta carbonica aziende agricole italiane.
- Eurostat. (2023). NUTS - Nomenclature of territorial units for statistics. Retrieved from <https://ec.europa.eu/eurostat/web/nuts/background>
- Fassò, A., Rodeschini, J., Moro, A. F., Shaboviq, Q., Maranzano, P., Cameletti, M., . . . Otto, P. (2023). Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data*, 10(1), 143. doi:10.1038/s41597-023-02034-0
- Fay, R. E. I., & Herriot, R. A. (1979). 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. doi:10.1080/01621459.1979.10482505
- Haining, R. P. (1993). *Spatial data analysis in the social and environmental sciences*: Cambridge university press.
- Maranzano, P., McConville, K. S., Otto, P., & Carillo, F. (2023). A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy. In F. M. Chelli, M. Ciommi, S. Ingrassia, F. Mariani, & M. C. Recchioni (Eds.), *Book of the Short Papers SIS 2023* (pp. 200-205): Pearson.
- Molina, I., & Marhuenda, Y. (2015). sae: an R package for small area estimation. *R J.*, 7(1), 81.
- Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). A course on small area estimation and mixed models. *Methods, theory and applications in R*.
- Otto, P., Fusta Moro, A., Rodeschini, J., Shaboviq, Q., Ignaccolo, R., Golini, N., . . . Fassò, A. (2024). Spatiotemporal modelling of $\text{PM}_{2.5}$ concentrations in Lombardy (Italy): a comparative study. *Environmental and Ecological Statistics*. doi:10.1007/s10651-023-00589-0
- Pratesi, M., & Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17(1), 113-141. doi:10.1007/s10260-007-0061-9
- Raffaelli, K., Deserti, M., Stortini, M., Amorati, R., Vasconi, M., & Giovannini, G. (2020). Improving Air Quality in the Po Valley, Italy: Some Results by the LIFE-IP-PREPAIR Project. *Atmosphere*, 11(4), 429.
- Rao, J. N., & Molina, I. (2015). *Small area estimation*: John Wiley & Sons.
- Regional Statistical Yearbook, R. (2017). Regional Statistical Yearbook of Lombardia in Europe 2017/2018.
- Rodeschini, J., Fassò, A., Finazzi, F., & Fusta Moro, A. (2024). Scenario analysis of livestock-related PM_{2.5} pollution based on a new heteroskedastic spatiotemporal model. *Socio-Economic Planning Sciences*, 102053. doi:<https://doi.org/10.1016/j.seps.2024.102053>
- Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3-4), 434-449. doi:10.1093/biomet/41.3-4.434