SIS | 2022

51st Scientific Meeting of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

SIS Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

800 1222·2022 ANNI

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Stata

sas

Università degli Studi del Sannio Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Data-driven design-based mapping of forest resources

Mappatura di risorse forestali in approccio da disegno

S. Franceschi, R.M. Di Biase, L. Fattorini, M. Marcheselli and C. Pisani

Abstract The mapping of forest resources in a study region is approached when the region is partitioned into spatial units by means of a completely data driven, design-based sampling strategy. When auxiliary variables are available for all the units, the prediction of the densities of an interest attribute can be performed by using an assisting model. Under these circumstances, the model residuals are interpolated using the inverse distance weighting interpolator with a data-driven smoothing parameter selection, and the density of the attribute for each unit is obtained by summing prediction and interpolated error. Finally, densities are rescaled to match the traditional total estimate with the sum of mapped values. The uncertainty is accounted for by a bootstrap procedure. A simulation study is performed and a case study is presented.
Abstract *La mappatura di risorse forestali per le unità spaziali che compongono una regione di studio viene affrontata con un approccio data-driven basato sul disegno. Se sono disponibili variabili ausiliarie, la previsione del valore dell'attributo di interesse può essere effettuata con strategie assistite da modello. I residui del modello vengono previsti utilizzando l'interpolatore inverse distance weighting dove il parametro di smorzamento viene selezionato con una procedura data-driven e le densità dell'attributo vengono ottenute sommando le previsioni e gli errori interpolati. Inoltre le densità vengono riscalate in modo tale che la stima del totale ottenuta sommando i valori previsti coincida con quella ottenuta con approcci tradizionali. L'incertezza delle stime è valutata attraverso una procedura bootstrap. Uno studio di simulazione viene effettuato e viene presentato un caso di studio.*

Key words: selection rule, smoothing parameter, regression estimator, harmonization, pseudo-population bootstrap

Sara Franceschi, Lorenzo Fattorini, Marzia Marcheselli and Caterina Pisani
Dipartimento di Economia Politica e Statistica e-mail: lorenzo.fattorini@unisi.it,e-mail: sara.franceschi@unisi.it,e-mail: marzia.marcheselli@unisi.it,e-mail: caterina.pisani@unisi.it

Rosa Maria Di Biase
Dipartimento di Sociologia e Ricerca Sociale, Università di Milano-Bicocca, e-mail: rosamaria.dibiase@unimib.it

1 Introduction

Wall-to-wall maps of forest attributes are essential information sources for forest management and most of the methodologies applied for mapping rely on model-based inference, owing to the impossibility to estimate values at non-sampled units without any assumption.

We present an alternative approach to forest mapping, in which an assisting model is used, but uncertainty stems from the adopted sampling scheme, in accordance to a model-assisted perspective [1]. However, any interpolator adopted for mapping can achieve statistical soundness only if it is proven to be design-based asymptotically unbiased and consistent (DBAU&C). Conditions ensuring DBAU&C, derived by [3] for finite populations of spatial units when the sole spatial information is available, are extended to allow for the exploitation of auxiliary sources of information, such as those provided by satellites and aircraft-based sensors, available for each spatial unit (e.g., pixels) partitioning the study area.

In particular, we propose a completely data-driven, design-based strategy for mapping forest attributes. The first step is the selection of a model using an Akaike-type criterion for choosing an appropriate set of available auxiliary variables. Using the least-squares method, the predictions of the attribute densities are obtained for each spatial unit. Then, the inverse distance weighting (IDW) interpolator, which relies on Tobler's first law of geography [9], is adopted for interpolating the residuals for non-sampled units on the basis of the residuals that are known for each sampled unit. Following [5], the leave-one-out cross-validation is used to select the smoothing parameter adopted in the IDW interpolator. Subsequently, the interpolated values at non-sampled units are obtained by summing model predictions and interpolated residuals.

Finally, we rescaled the resulting densities in such a way that the total obtained from the interpolated values matches traditional total estimates [7]. Moreover, a bootstrap technique is used to estimate relative root mean squared errors of the predicted values.

2 Data-driven mapping

Consider a study area partitioned into N spatial units and suppose to be interested in the reconstruction of the whole map of the spatial population, that is the estimation of the density of an interest attribute for each spatial unit, by using a suitably selected sample of units. It is worth noting that usually the size of the spatial units are known and therefore the estimation of the density is theoretically equivalent to the estimation of the total amount in each spatial unit. Furthermore, suppose that a large set of covariates is available, as it is common in forest inventories. To select the assisting model for obtaining predictions, we performed a variable selection using an Akaike-type criterion, as suggested by [2]. Indeed, strongly correlated auxiliary variables can induce instability in the model while they can exhibit a poor prediction capability when weakly correlated to the interest variable.

Let f_j and $\mathbf{x}_j$ respectively be the density of the interest attribute and the vector of selected covariates for spatial unit j, in such a way that densities can be expressed as $f_j = \boldsymbol{\beta}^t \mathbf{x}_j + e_j(\boldsymbol{\beta})$. The choice of $\boldsymbol{\beta}$ can be made using a least-square criterion to minimize the residual sum of squares. Denoting by $\mathbf{b}$ the ordinary least-squares solution for $\boldsymbol{\beta}$, if $\mathbf{b}$ was known, the residuals for each sampled unit would be $e_j(\mathbf{b}) = f_j - \mathbf{b}^t \mathbf{x}_j$, while the residuals for non-sampled units could be obtained using the IDW interpolator. In particular, the j-th interpolated residual would be given by

$$\hat{e}_j(\mathbf{b}) = Z_j e_j(\mathbf{b}) + (1 - Z_j) \sum_{i=1}^{N} w_{i,j}(\alpha) e_j(\mathbf{b})$$

where Z_j is equal to 1 when the j-th unit is sampled and 0 otherwhise and $w_{ij}(\alpha) = Z_i d_{i,j}^{-\alpha} / \sum_{l=1}^{N} Z_l d_{l,j}^{-\alpha}$, with $d_{i,j}$ denoting the distance between the centroid of unit i and unit j, and $\alpha > 2$ is the smoothing parameter. As consequence, the resulting interpolated value for the density for unit j would be $\hat{f}_j(\mathbf{b}) = \mathbf{b}^t \mathbf{x}_j + \hat{e}_j(\mathbf{b})$.

Unfortunately, the least-square solution $\mathbf{b}$ is unknown, involving the knowledge of the density for each unit. Nevetherless, as suggested by [8], an estimate $\hat{\mathbf{b}}$ can be obtained as a function of Horvitz-Thompson estimators. Then, IDW interpolation can be performed with $\mathbf{b}$ replaced by its sampling estimator $\hat{\mathbf{b}}$ on the basis of the observed residuals $e_j(\hat{\mathbf{b}}) = f_j - \hat{\mathbf{b}}^t \mathbf{x}_j$ so that the interpolated densities are given by $\hat{f}_j(\hat{\mathbf{b}}) = \hat{\mathbf{b}}^t \mathbf{x}_j + \hat{e}_j(\hat{\mathbf{b}})$.

As to the asymptotic properties, $\hat{f}_j(\hat{\mathbf{b}})$ is consistent because, as the extent of the spatial units partitioning the study area decreases and their number increases, $\hat{\mathbf{b}}$ converges to $\mathbf{b}$ and, consequently, $\hat{f}_j(\hat{\mathbf{b}})$ converges to $\hat{f}_j(\mathbf{b})$, which in turn is DBAU&C under conditions derived in [3]: i) the existence of a Riemann integrable function giving the density of the interest attribute at any point of the study area; ii) some sort of regularity in the shape of the spatial units; iii) the use of an asymptotically balanced spatial sampling scheme. Commonly adopted sampling schemes ensuring DBAU&C are simple random sampling without replacement, one-per-stratum stratified sampling and systematic sampling.

Moreover, the IDW interpolator depends on the smoothing parameter α determining the roughness of the surface of interpolated residuals. As suggested by [5], the value of α can be selected by means of the leave-one-out cross-validation and the corresponding IDW interpolator is proven to remain DBAU&C.

Finally, it is worth noting that, in a design-based setting, the total of the interest attribute is commonly estimated by traditional estimators, such as the regression estimator, which necessarily give rise to different total estimates with respect to the one achieved by summing the interpolated values for all the spatial units. To obtain non-discrepant results, a harmonization of the estimated map can be achieved by rescaling density estimates in analogy with [7].

As to the evaluation of precision, each step of the proposed mapping strategy (Akaike-type selection of the assisting model, choice of the smoothing parameter and harmonization of maps with the total estimate) should be considered, since all of them are sample dependent. The use of bootstrap seems to be the sole way for facing the complexity of this data-driven mapping procedure. More precisely, the map of the interpolated values is taken as a pseudo-population from which bootstrap samples are drawn using the same sampling scheme adopted to produce the original sample, as suggested by [5] and [4]. Under the conditions ensuring DBAU&C, the estimated map converges to the true map, so that the bootstrap distribution should converge to the true distribution, also providing consistent estimators of its mean squared errors.

3 Simulation study

The performance of the proposed data-driven mapping strategy was empirically assessed by means of a simulation study performed on a real survey region located in Calabria (Southern Italy). The values of several auxiliary variables and the value of growing stock volume (interest attribute) were available for each pixel partitioning the study region.

In order to check the improvement as the number of pixels partitioning the study area increases and their size become smaller, the study area was partitioned considering three different pixels sizes and the values of interest and auxiliary attributes were aggregated within those pixels.

From each of the three partitions, $10,000$ samples were independently selected by means of one-per-stratum sampling (OPSS) with a constant sampling fraction of 4%. For each sample, selection of auxiliary variables was performed, the growing stock volume estimate for each pixel was derived by using the IDW interpolator, where the value of the smoothing parameter was obtained by means of leave-one-out cross-validation and, finally, harmonization was implemented. Furthermore, for each simulation run, $1,000$ bootstrap samples were selected from the estimated map by the same OPSS scheme adopted for extracting the original sample, then bootstrapped maps were used to achieve the bootstrap root mean squared error estimates. From the resulting Monte Carlo distributions, several perfomance indexes were computed. Results show that usually 4 or 5 covariates are selected out of the 11 originally available, suggesting that the selecting rule is likely to choose parsimonious models. Also for the choice of the smoothing parameter, small values are the most commonly selected, probably owing to the smoothness of the error surfaces to be interpolated (see Fig. 1). Furthermore, relative bias values and relative root mean squared errors quickly decrease in minima, means and maxima as the spatial grain decreases, with a balance between underestimation and overestimation, showing also a relevant spatial autocorrelation of negative and positive values (see Tab. 1).

As to the bootstrap root mean squared error estimator, underestimation seems to be prevalent. Nevertheless, its tendency to be conservative is more apparent as the spatial grain decreases. Indeed, the number of pixels in which the ratio between the expectation of the bootstrap root mean squared error estimator and the true root

Table 1 Minima, means and maxima of the absolute bias (AB), root mean squared errors (RMSE) and its ratio with the expectation of bootstrap root mean squared error (RAT) achieved for the three populations considered in the simulation study.

	AB			RMSE			RAT		
POP	MIN	MEAN	MAX	MIN	MEAN	MAX	MIN	MEAN	MAX
POP1	0.381	27.400	86.010	28.86	58.59	133.80	0.348	0.652	0.919
POP2	0.038	22.770	81.215	12.93	35.99	92.29	0.221	0.592	1.100
POP3	0.003	14.243	65.035	6.00	23.43	73.88	0.263	0.730	1.809

mean squared error is greater than 1 increases as the spatial grain decreases. It is probable that with a thinner partition conservativeness over the whole area can be achieved.

4 Case study

The proposed mapping strategy was applied to provide the map of densities of wood volume in the forest estate of Rincine (Central Italy), an area partitioned into square cells of $23 \times 23m^2$. For each cell, the values of several auxiliary variables were available. For sampling purposes, the area was partitioned into 50 blocks of cells and, following OPSS, one cell was randomly selected within each block, determining a sample of 50 units. The wood volume of each tree in the selected cells was reckoned to achieve the wood volume density by three size classes: small, medium and large trees.

For each class, variable selection was performed using the procedure adopted in the simulation. Mapping was performed for each size class harmonizing the total of the interpolated values with the regression estimates of the total over the whole survey area. Subsequently, the three estimated maps were taken as pseudo-populations

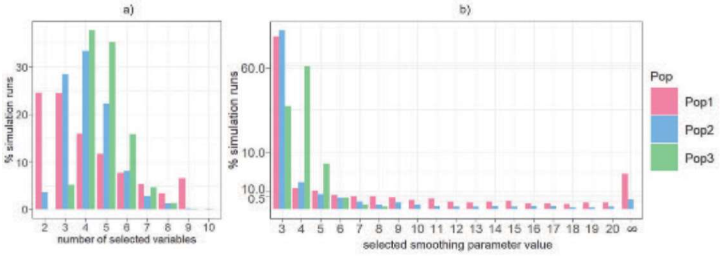

Fig. 1 For each of the three considered populations a) frequencies (expressed as percentage of the number of simulation runs) of the number of variables adopted in the predicting model; b) frequencies (expressed as percentage of the number of simulation runs) of the selected smoothing parameter.

from which $10,000$ bootstrap samples were selected by the OPSS scheme adopted to select the original sample. From each bootstrap sample, the three bootstrap maps were achieved for each class repeating the usual steps and the total map was achieved by sum. Then, bootstrap root mean squared errors were computed for each of the four maps. Bootstrap samples were also adopted to achieve the bootstrap root mean squared error of the regression estimator, incorporating uncertainty determined from the variable selection.

The maps of estimated densities within cells show a relevant level of uncertainty for small trees that decreases for medium trees and becomes satisfactory for large trees and for the total. High uncertainty for small trees is mainly due to the high variability in the number of small trees (many of which from scattered forest natural regeneration) within the selected cells, with many cells with 0 small trees and few with more than 150 trees. Consequently, as expected (see e.g. [6]), the precision is deteriorated by the highly clustered spatial pattern.

In order to highlight the improvement when exploiting auxiliary sources of information, those maps were compared to those obtained using the sole spatial information and selecting the value of the smoothing parameter from data as in [5]. From this comparison it is evident that the proposed mapping strategy reduces the excessive smoothing produced by IDW interpolation based on the sole spatial information.

5 Final remarks

Despite being data-driven, the proposed mapping strategy is based on several subjective choices, such as the choice of the linear model for performing regression and the choice of the criterion for implementing the leave-one-out cross-validation. However, it should be also noticed that these choices only impact on the sample statistics adopted for mapping, while the precision of the map is objectively determined by the sampling scheme used for selecting the units partitioning the study area.

A further issue is that two-phase sampling schemes are usually adopted in forest inventories, whereas one-phase OPSS was here proposed. However, [4] proved the consistency of the IDW interpolator if the two-phase scheme continues to provide spatially balanced samples. As consequence, the procedure can be applied also in two-phase large-scale forest inventories.

The proposed design-based, data-driven strategy for mapping forest resources exploiting auxiliary variables, guides the user from the preliminary choice of the assisting model to the final map and the estimation of its precision, through model selection, exploitation of selected covariates for mapping, choice of the smoothing parameter for IDW interpolation, harmonization with traditional estimates of totals and bootstrap resampling from the estimated map. Moreover, this strategy seems to be suitable also for mapping environmental attributes besides forestry.

References

1. Breidt, F.J., Opsomer, J.D.: Model-assisted survey estimation with modern prediction techniques. Stat. Sci. (2017) doi: 10.1214/16-STS589
2. Burman, P., Nolan, D.: A general Akaike-type criterion for model selection in robust regression. Biometrika (1995) doi: 10.2307/2337352
3. Fattorini, L., Marcheselli, M., Pratelli, L.: Design-based maps for finite populations of spatial units. J. Am. Stat. Assoc. (2018) doi: 10.1080/01621459.2016.1278174
4. Fattorini, L., Franceschi, S., Marcheselli, M., Pisani, C., Pratelli, L.: Two-phase sampling strategies for design-based mapping of continuous spatial populations in environmental surveys. Ann. Appl. Stat. (2021) doi: 10.1214/20-AOAS1392
5. Fattorini, L., Franceschi, S., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based spatial interpolation with data driven selection of the smoothing parameter. Submitted (2022)
6. Gregoire, T. G., Valentine, H. T.: Sampling Strategies for Natural Resources and the Environment. Chapman & Hall, New York (2008)
7. Marcelli, A., Fattorini, L., Franceschi, S.: Harmonization of design-based mapping for spatial populations. Stoch. Environ. Res. Risk Assess. (2022) doi: 10.1007/s00477-022-02186-2
8. Särndal, C. E., Swensson, B., and Wretman, J.: Model Assisted Survey Sampling. Springer, Berlin (1992)
9. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. (1970) doi: 10.2307/143141