Structured Additive Distributional Regression applied to the spatio-temporal analysis of soil-plant variability

Regressione Distribuzionale Additiva Strutturata applicata all'analisi spazio-temporale delle relazioni suolo-pianta

Giovanna Jona Lasinio, Alessio Pollice, Thomas Kneib, Stefan Lang, Roberta Rossi, Mariana Amato

Abstract We analyze sensor data describing soil productivity in terms of NDVI and soil physical features measured in terms of electro resistivity. We adopt a Bayesian modeling approach to account for covariates with measurement error combined with regression models for a class of continuous, discrete and mixed univariate response distributions with potentially all parameters depending on a semiparametric structured additive predictor. Estimates are obtained by Markov chain Monte Carlo simulations.

Abstract *Si analizzano dati da sensori, volti a descrivere la produttività di un suolo in termini di NDVI e le sue caratteristiche fisiche in termini di elettro-resistività. Viene utilizzato un approccio modellistico bayesiano al fine di includere covariate affette da errori di misura in un modello di regressione per una classe di distribuzioni univariate continue, discrete e miste i cui parametri possano dipendere tutti o in parte da predittori semiparametrici additivi e strutturati. Le stime sono ottenute mediante simulazioni Markov chain Monte Carlo.*

Key words: structured additive distributional regression, Bayesian semiparametric regression, measurement error, agricultural management.

Giovanna Jona Lasinio
Sapienza Università di Roma, Rome, ITALY, e-mail: giovanna.jonalasinio@uniroma1.it

Alessio Pollice
Università degli studi di Bari Aldo Moro, Bari, ITALY, e-mail: alessio.pollice@uniba.it

Thomas Kneib
Universität Göttingen, Göttingen, GERMANY, e-mail: tkneib@uni-goettingen.de

Stefan Lang
Universität Innsbruck, Innsbruck, AUSTRIA, e-mail: stefan.lang@uibk.ac.at

Roberta Rossi
CREA-ZOE , Bella (PZ), ITALY, e-mail: roberta.rossi@crea.gov.it

Mariana Amato
Università della Basilicata, Potenza, ITALY, e-mail: mariana.amato@unibas.it

1 Introduction

Standard regression theory assumes that explanatory variables are deterministic or error-free, but this assumption is quite unrealistic for many biological processes and replicated observations of covariates are often obtained to quantify the variability induced by the presence of measurement error (ME). The most well known effect of measurement error is the bias towards zero induced by additive i.i.d. measurement error, but under more general measurement error specifications (as considered in this paper), different types of misspecification errors are to be expected [3, 6]. This is particularly true for semiparametric additive models, where the functional shape of the relation between responses and covariates is specified adaptively and therefore is also more prone to disturbances induced by ME. Recent papers advocate the hierarchical Bayesian modeling approach as a natural route for accommodating ME uncertainty in regression models.
In this work we introduce a functional ME modeling approach allowing for replicated covariates with ME within a flexible class of regression models recently introduced by [4], namely *structured additive distributional regression models*. In this modeling framework, each parameter of a class of potentially complex response distributions is modeled by an additive composition of different types of covariate effects, e.g. non-linear effects of continuous covariates, random effects, spatial effects or interaction effects. We allow for quite general measurement error specifications including multiple replicates with heterogeneous dependence structure. From a computational point of view, based on the seminal work [2] for Gaussian scatterplot smoothing and [5] for general semiparametric exponential family and hazard regression models, we develop a flexible fully Bayesian ME correction procedure based on Markov chain Monte Carlo (MCMC) techniques to generate observations from the joint posterior distribution of structured additive distributional regression models. ME correction is obtained by the imputation of unobserved error-free covariate values in an additional sampling step. Our implementation is based on an efficient binning strategy that avoids recomputing the complete design matrix after imputing true covariate values and combines this with efficient storage and computation schemes for sparse matrices.

The main motivation of our investigation comes from a case study on the use of proximal soil-crop sensor technologies to analyze the within-field spatio-temporal variation of soil-plant relationships in view of the implementation of efficient agricultural management practices. More precisely, we analyze the relationship between multi-depth soil information indirectly assessed through the use of high resolution geophysical soil proximal sensing technology and data of forage ground-cover variation measured by a multispectral radiometer within a seven hectares Alfalfa stand in South Italy. Observations of both quantities were made using sensors with very refined spatial resolution: ground-cover data were obtained at four sampling occasions with point locations changing over time, while soil data were sampled only once for three different depth layers. Estimating a functional relation between ground-cover and soil with the data at hand involves addressing several issues, also linked to the spatial and temporal misalignment and the large data size. The nonlinear relation between crop productivity and soil is estimated by additive distributional regression models with structured additive predictor and measurement error correction. While distributional regression allows to deal with the heterogeneity of the response scale at the four sampling occasions, the ME correction is motivated by observations of covariates being replicated along a depth gradient and extends the model proposed by [5], accounting for heterogeneous variances and possibly dependent replicates of the soil covariate.

2 Measurement Error Correction in Distributional Regression

The main motivation for our modeling proposal comes from the need to estimate the nonlinear dependence of ground-cover on soil information by a smooth function, accounting for the heterogeneity in the position and scale of the response due to the sampling time, for the repeated measurements of the soil covariate and for the residual variation of unobserved spatial features.

2.1 Distributional Regression

Assume that independent observations (y_i, ν_i), $i = 1, \dots, n$, are available on the response y_i and covariates ν_i and that the conditional distribution of the responses belongs to a K-parametric family of distributions such that $y_i|\nu_i \sim \mathscr{D}(\vartheta(\nu_i))$ and the K-dimensional parameter vector $\vartheta(\nu_i) = (\vartheta_1(\nu_i), \dots, \vartheta_K(\nu_i))'$ is determined based on the covariate vector ν_i. More specifically, we assume that each parameter is supplemented with a regression specification $\vartheta_k(\nu_i) = h_k(\eta^{\vartheta_k}(\nu_i))$, where h_k is a response function that ensures restrictions on the parameter space and $\eta^{\vartheta_k}(\nu_i)$ is a regression predictor. In our analyses, we will consider one specific special case where $y_i \sim \text{Beta}(\mu(\nu_i), \sigma^2(\nu_i))$, i.e. responses are conditionally beta distributed with regression effects on location and scale. For both parameters $\mu(\nu_i)$ and $\sigma(\nu_i)^2$ of the beta distribution we employ a logit link, since they are restricted to the unit interval.

2.2 Structured Additive Predictor

For each of the predictors, we assume an additive decomposition as $\eta^{\vartheta_k}(\nu_i) = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\nu_i) + \ldots + f_{J_k}^{\vartheta_k}(\nu_i)$, i.e. each predictor consists of a total of J_k potentially nonlinear effects $f_j^{\vartheta_k}(\nu_i)$, $j = 1, \dots, J_k$, and an additional overall intercept $\beta_0^{\vartheta_k}$. The nonlinear effects $f_j^{\vartheta_k}(\nu_i)$ are a generic representation for a variety of different effect types (including nonlinear effects of continuous covariates, interaction surfaces, spatial effects, etc.). Any of these effects can be approximated in terms of a linear combination of basis functions as $f(\nu_i) = \sum_{l=1}^{L} \beta_l B_l(\nu_i) = b_i'\beta$, where we dropped both the function index j and the parameter index ϑ_k for simplicity, $B_l(\nu_i)$ denotes the different basis functions with basis coefficients β_l and $b_i = (B_1(\nu_i), \dots, B_L(\nu_i))'$ and $\beta = (\beta_1, \dots, \beta_l)'$ denote the corresponding vectors of basis function evaluations and basis coefficients, respectively.

Since in many cases the number of basis functions will be large, we assign informative multivariate Gaussian priors $p(\beta|\theta) \propto \exp\left(-\frac{1}{2}\beta' K(\theta)\beta\right)$ to the basis coefficients to enforce certain properties such as smoothness or shrinkage. The specific properties are determined based on the prior precision matrix $K(\theta)$ which itself depends on further hyperparameters θ.

2.3 Measurement Error

In our application we are interested in estimating the nonlinear effect $f(x)$ in one of the predictors of a distributional regression model where instead of the continuous covariate x we observe M replicates $\tilde{x}_i^{(m)} = x_i + u_i^{(m)}, \quad m = 1, \dots, M$, contaminated with measurement error $u_i^{(m)}$. For the measurement error, we consider a multivariate Gaussian model such that $u_i \sim \mathrm{N}_M(\mathbf{0}, \Sigma_{u,i})$, where $u_i = (u_i^{(1)}, \dots, u_i^{(M)})'$ and $\Sigma_{u,i}$ is a known, pre-specified unstructured covariance matrix.

The basic idea in Bayesian measurement error correction is now to include the unknown, true covariate values x_i as additional unknowns to be imputed by MCMC simulations along with estimating the other parameters in the model. This requires that we assign a prior distribution to x_i as well and rely on the simplest version $x_i \sim \mathrm{N}(\mu_x, \tau_x^2)$, where we achieve flexibility by adding a further level in the prior hierarchy via $\mu_x \sim \mathrm{N}(0, \tau_\mu^2)$ and $\tau_x^2 \sim \mathrm{IG}(a_x, b_x)$. To obtain diffuse priors on these hyperparameters, we use $\tau_\mu^2 = 1000^2$ and $a_x = b_x = 0.001$ as default settings.

3 Case study

Given the aim of this work and the data size (ranging from 91438 to 222278 spatial points) , the spatial resolution was downscaled by interpolating samples to a 2574 cells square lattice overlaying the study area. Given the different number of sampled points corresponding to each sampling occasion (NDVI) and survey (ER), we used a proportional nearest neighbors neighborhood structure to compute the downscaled values. At each grid point we calculated the neighbors' means for both NDVI and ER, while neighbors' variances and covariances between depth layers were obtained for ER. Summary measures of the scale and correlation of ER repeated measures at each of the 2574 grid points provide valuable information that enables us to increase the model complexity with no additional costs in terms of parameters, i.e. degrees

Model	DIC	WAIC
M1	-22841.0	-22835.9
M2	-27131.4	-27126.4

Table 1 Model fit statistics for Beta distribution additive mean (M1) and distributional (M2) regression models: deviance information criterion and Watanabe-Akaike information criterion.

of freedom. Such a by-product of the downscaling of the original data is plugged into the model likelihood.

For available NDVI recordings, we consider Beta distributional regression models and specify the two predictors as follows. For $s = 1,\dots,2574$ grid points and $t = 1,\dots,4$ time points, the structured additive predictor of the location parameter is determined as an additive combination of three linear and functional effects: a linear seasonal effect, a tensor product spatial effect and a nonlinear smooth effect of the continuous covariate ER. The linear predictor of the scale parameter is assumed to depend only on the effect of time, thus allowing heteroscedasticity of seasonal NDVI recordings. The Metropolis-Hastings algorithm, implemented using `BayesX` [1], to sample the posteriors of the Beta models, required runs of 50000 iterations with 35000 burnin and thinning by 15. Convergence was reached and checked by visual inspection of the trace plots and standard diagnostic tools. Fine tuning of hyperparameters lead us to 8 equidistant knots for each of the two components of the tensor product spatial smooth.

The additive distributional regression model was compared to standard additive mean regression with the same mean predictor, applying the proposed measurement error correction to both models. By this comparison we show that adding a structured predictor for the scale parameter improves both the in-sample and out-of-sample predictive accuracy. In the following, M1 is an additive regression model with mean predictor, while M2 is an additive distributional regression model with the same mean predictor and scale predictor. As far as additive distributional regression is concerned, DIC, WAIC agree in assessing the proposed model M2 as performing better than a simple additive mean regression M1 (Tab. 1).

Based on the resulting estimated smooth functions (Figure 1, left), two ER cut-offs (at 10 and 20 *Ohm m*) are proposed that can be used to split the field in three areas characterized by a different monotonic soil-plant relationship:

- **Zone i: ER $<$ 10 *Ohm m***, where NDVI grows with ER and very low ER readings correspond to intermediate to high NDVI values (the former correspond to the presence of poorly drained soils and the consequent risk of waterlogging; crop management needs to take into account in-season rain patterns to minimize the risks of waterlogging damages in wet years);
- **Zone ii: 10 *Ohm m* $<$ ER $<$ 20 *Ohm m***, where ER is negatively related to NDVI and soil factors affecting ER act almost linearly and consistently on plant performance (precision management can be applied as a function of ER, i.e. the resistivity map itself can be used as a prescription map in the corresponding areas);

- **Zone iii: ER $>$ 20 *Ohm m***, where despite the large variation in ER there is a limited NDVI-soil responsiveness and NDVI is constantly low (corresponds to the presence of the hardpans and management criteria should differ accordingly).

Each zone conveys information on the shape and strength of the association between soil and crop variability, thus the proposed field zonation helps discerning areas where even a little change in soil properties can affect plant productivity (zone ii) from areas where soil environment is not practically alterable (zone iii) or in-season evaluations are possibly needed (zone i).

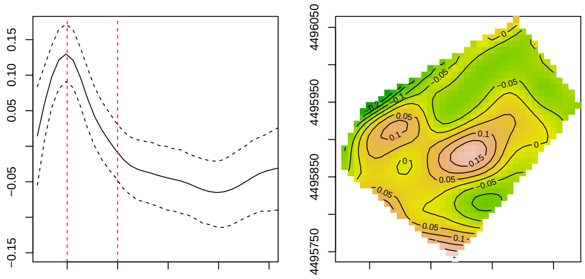

Figure 1 Smooth estimates of ER effects (left) and residual spatial effects (right) for model BM2. Effects are estimated on the logit scale. Dotted red vertical lines locate ER cut-offs corresponding to different monotonic soil-plant relationships.

Acknowledgements The first two authors have been supported by MIUR-PRIN (20154X8K23-SH3) EPHASTAT

References

1. Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2015). *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*. Version 3.0.2.
2. Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, **97**(457), 160–169.
3. Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapmann And Hall, CRC PRESS.
4. Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015a). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **64**(4), 569–591.
5. Kneib, T., Brezger, A., and Crainiceanu, C. M. (2010). Generalized semiparametric regression with covariates measured with error. In T. Kneib and G. Tutz, editors, *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, pages 133–154. Physica-Verlag HD, Heidelberg.
6. Loken, E. and Gelman, A. (2017). Measurement error and the replication crisis. *Science*, **355**(6325), 584–585.