Computers & Geosciences 166 (2022) 105161

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

A scalable method for the estimation of spatial disaggregation models

Arthur Nicolaus Fendrich [a,*], Elias Salomão Helou Neto [b], Lucas Esperancini Moreira e Moreira [b], Durval Dourado Neto [a]

[a] University of São Paulo, Luiz de Queiroz College of Agriculture, Avenida Pádua Dias, 11, 13418-900, Piracicaba, SP, Brazil
[b] University of São Paulo, Institute of Mathematics and Computer Science, Avenida Trabalhador Sãocarlense, 400, 13566-590, São Carlos, SP, Brazil

ARTICLE INFO

Keywords:
Disaggregation
Inverse problem
Spatially-explicit
Geoprocessing
Flexible regression

ABSTRACT

Gaining information about detailed processes using aggregation information is a frequent challenge in research involving geospatial data, with examples in different fields of knowledge such as agronomy, soil science, meteorology, public health, epidemiology, and others. Analyses using aggregated data lead to distorted conclusions since they disregard local patterns, and such a problem has motivated different approaches for reconstructing the information in a finer resolution from the aggregated data. However, most existing methods focus on the particular case where the volume of data does not exceed the amount of memory available for computations, a situation that has become increasingly less frequent with the fast pace of data generation nowadays. In practice, this problem limits either spatial resolution or coverage of applications, thus precluding their use in a more general context. In this paper, we address the problem of disaggregation of spatial data with huge datasets by proposing a scalable method to estimate the parameters of a well-established model. We propose an iterative scheme for model estimation and prove its convergence to a critical point of the likelihood function derived. To test the method, we provide a controlled simulation and a real example for sugarcane production in Brazil. In the simulation, the results indicate a successful reconstruction of 1 million pixels from 90 block areas. In the real example, the results had a compatible match with the agronomic literature, indicating a reasonable prediction of sugarcane production in a 100 m spatial resolution (i.e., approx. 5×10^8 pixels) from 5,565 block-areas. Compared to the most similar previous work, scalability allowed us to use a nearly 100 times higher resolution, which corresponds to 10,000 times more pixels. With our methods, we expect to assist researchers from different fields in disaggregating spatial information to larger areas or higher resolutions.

1. Introduction

Analyzing spatial data poses significant challenges by nature. While the geographical space and several processes are continuous, numerous datasets represent only a summarizing function (e.g., the sum or the average) of these underlying processes over large areas that partition the space. The definition of boundaries depends on the problem under consideration, with census data being likely the most common setting. In this case, the administrative frontiers of a district, state, or country (Hawley and Moellering, 2005; Buchin et al., 2012) delimit the areas, such as in the county-level public health data of Goovaerts (2006). However, other partitions can also be used, such as farm boundaries in soil science (Orton et al., 2012), a coarse-scale grid in meteorology (Park, 2013), and others. The procedure of converting data from a finer to a coarser resolution is called aggregation, and the collection of aggregated data can be motivated by technical, administrative, and other reasons (Steinbuch et al., 2019). It has, for example, the advantage of being cheaper and ensuring anonymity (Armstrong et al., 1999).

Despite the benefits during data collection, the aggregation procedure can potentially mislead the results of analyses. For example, in the 1970s, Openshaw and Taylor (1979) found that the relationship between two variables could change not only when they were aggregated but also when different boundaries were used to partition the space. Such a result coined the term modifiable areal unit problem (MAUP), which became later a well-known limitation for the analysis of aggregated data (Wong, 2009). A similar effect of reaching wrong conclusions about individual characteristics from grouped data models was also detected in the epidemiology field, where such a phenomenon is called "ecological fallacy" (Piantadosi et al., 1988; Wakefield, 2005). Besides, in some voting systems, politicians explore the aggregation procedure's bias by manipulating the spatial boundaries of administrative regions to favor some specific outcome in the resulting areas.

* Corresponding author.
E-mail address: arthur.fendrich@usp.br (A.N. Fendrich).

https://doi.org/10.1016/j.cageo.2022.105161
Received 14 September 2021; Received in revised form 28 March 2022; Accepted 2 June 2022
Available online 18 June 2022

Such a practice is known as "gerrymandering" and dates from the 19th century (Friedman and Holden, 2008).

The potential drawbacks of conclusions taken from models for aggregated data have motivated the search for methods to recover the original un-clustered information from block-area observations. Such a reverse procedure is sometimes called spatial downscaling or disaggregation, and many applications exist in the literature of different disciplines, including agriculture (You and Wood, 2006), epidemiology (Utazi et al., 2019), pedology (Vincent et al., 2018), economics (Monteiro et al., 2018), political science (King, 1997), disease risk mapping (Li et al., 2012), digital soil mapping (Møller et al., 2019), among others. In all cases, the models generate a set of information at a higher spatial resolution from the data at a lower spatial resolution, and the approaches vary in complexity and the assumptions made.

Several approaches to solving the disaggregation problem exist in the literature, and they can be split into two broad categories: those who do not make use of auxiliary data and those who do (Comber and Zeng, 2019). On the first group, the simplest example is the area weighting method, in which the aggregated information is distributed to the finer scale proportionally to the size of the overlapping areas (Monteiro et al., 2018). Such a method implicitly assumes that the variable is homogeneously distributed, which might be unlikely in most settings. Also on the first group, the pycnophylatic algorithm (Tobler, 1979) generates a smooth surface on the finer resolution by: (i) iteratively calculating the value on a pixel from the value of its neighbors; and (ii) ensuring so-called 'mass-preserving' property, where the reaggregation of the finer level data reproduces the observed coarse-level information. In the pycnophylatic method, the assumption of a continuous surface for the target level might not be suitable in some situations where discontinuities such as rivers, roads, or waterbodies divide the landscape (Comber and Zeng, 2019).

The second category of methods relies on incorporating external information to explain the distribution of the variable at the disaggregated level. They emerged as a possible alternative to inconsistencies that may appear when using methods from the first category, such as the estimation of a population value in uninhabited regions (Comber and Zeng, 2019). A wide variety of existing methods fall under this category, and a review can be found in Comber and Zeng (2019) and Monteiro et al. (2018). The subcategory of statistical and geostatistical methods has recently attracted increasing attention. In these cases, ancillary data is combined with statistical techniques to extract relationships between the fine and coarse-level data, and the spatial autocorrelation can be taken into account during the modeling process. Applications of such methods can now be found in different fields of knowledge, such as species distribution modeling (Niamir et al., 2011; Keil et al., 2012; Barwell et al., 2014), disease risk mapping (Sturrock et al., 2014; Weiss et al., 2019), climate sciences (Poggio and Gimona, 2015), hydrology (Wang et al., 2020) and others. Sturrock et al. (2014), for example, used a hierarchical Bayesian model to disaggregate Malaria incidence to a 1×1 km spatial grid using several satellite-derived variables, land cover, and distance to rivers as auxiliary information. Even without taking spatial autocorrelation into account, a validation procedure showed that the model correctly indicated zero incidences in 84 out of the 101 observed facilities at the finer level. In one significant methodological advance, Utazi et al. (2019) incorporated the spatial dependence into the linear predictor of a hierarchical model. The authors developed a controlled simulated example, and based on the good results (i.e., correlation with the true data varying from 0.66 to 0.98 at the finer level), they disaggregated vaccination coverage of measles and diphtheria-tetanus-pertussis for Afghanistan and Pakistan at a 5×5 km pixel level.

Actual applications may be interested on large spatial domains, fine spatial resolutions, or both. In these cases, as noted by Li et al. (2012), the complexity of the problems implies computationally demanding and memory-intensive calculations, especially when the spatial correlation is strong. For some problems, such computational costs may even be prohibitive. A common point among existing approaches is that they focus on the particular case where the volume of data can be hold in a computer's memory. To the best of our knowledge, the problem of disaggregating datasets that are too big for a computer's memory remains unexplored. Nowadays, it is well recognized that the generation of spatial data has been increasing quickly, generating new challenges and demanding advances in methods for analysis and prediction (Tang, 2020). This means that most existing disaggregation techniques still limit either spatial resolution or coverage of applications. Thus, for example, they cannot handle the use of high-resolution information (e.g., Hengl et al. (2017), Hulley et al. (2015), Chen et al. (2015)), which have achieved a better representation of the continuous nature of the territory but at the cost of bigger and more computationally demanding datasets (Lasinio et al., 2012).

In this paper, we address the problem of disaggregating spatial data with huge datasets. Our contribution consists of proposing a method to estimate the parameters of a disaggregation model when the amount of information at the fine scale exceeds the computer's memory. Hereinafter, we refer to such method property as 'scalability' (Wang et al., 2016). First, we show the existing regression model at the disaggregated level and describe how its characteristics cover many practical applications. Next, we propose an iterative parallel method to estimate its parameters from the aggregated data and prove the convergence to a critical point of the objective function derived. The method is based on dividing the huge dataset into smaller chunks that can be efficiently processed in parallel, a technique known as domain decomposition (Armstrong and Densham, 1992). To test the method, we set up a simulation study to reconstruct 10^6 pixels from 90 areas and compare the functions estimated with those used to generate the data. We also apply the method for the real example of sugarcane production in Brazil, where we reconstruct a 100 m raster (approx. 5×10^8 pixels) from 5,565 areas using auxiliary data such as remotely sensed land cover, soil characteristics, and climate information. Finally, to check the consistency of the results obtained in the absence of ground-truth data, we compare the functions estimated from the data with the patterns described in the agronomic literature.

2. Methodology

The Methodology section is organized as follows. In Section 2.1, we formalize the disaggregation problem and describe the constraints that most practical applications face, and in Section 2.2, we describe the proposed solution in three parts.

First, in Section 2.2.1, we describe the existing model at the disaggregated level, derive an objective function for the estimation of its parameters at the aggregated level, and present an iterative method for optimizing the objective function based on the Laplace method. Second, in Section 2.2.2, we show how the proposed model relates to the constraints described in Section 2.1, accounting for smooth functions of explanatory variables and recent advances to make spatial and spatiotemporal modeling computationally feasible. This subsection shows how model calculations can be performed in a completely parallel setting and refer to a mathematical proof of convergence (Appendix). Third, in Section 2.2.3, we show how to exploit the matrices' structure to improve the method's practical implementation.

Then, Section 2.2.2 provides further implementation details, and Section 2.3 and Section 2.4 present the case studies.

2.1. Problem description

Let $\boldsymbol{y} = [y_{1,1}, y_{1,2}, \dots, y_{1,n_1}, y_{2,1}, y_{2,2}, \dots, y_{2,n_2}, \dots, y_{N,1}, y_{N,2}, \dots, y_{N,n_N}]^\top$ denote the $n \times 1$ (where $n = n_1 + n_2 + \cdots + n_N$) vector of the unobserved outcome $y_{i,j}$ at the jth pixel belonging to block area i ($i = 1, 2, \dots, N$). What we observe is $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{y}) = [Y_1, Y_2, \dots, Y_N]^\top$, the aggregation of $\boldsymbol{y}$ into N block areas. The problem is to find $y_{i,j}$ respecting the following constraints:

1. all $y_{i,j}$ are constrained to some interval, i.e., $a < y_{i,j} < b, \quad \forall i, j, a, b \in \mathbb{R}$;
2. $\boldsymbol{y}$ is a (possibly non-linear) function of one or more (i.e., m) explanatory variables $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m]^\top$;
3. $\boldsymbol{y}$ is potentially spatially-correlated, i.e., closer pixels are more likely to have similar values than farther pixels;
4. the aggregation constraint is at least nearly respected (i.e., $\boldsymbol{Y} \approx \boldsymbol{h}(\boldsymbol{y})$);
5. the approach needs necessarily to be scalable, as n can be huge.

2.2. Proposed solution

2.2.1. The model and estimation

We consider the non-linear mixed model at the disaggregated level:

$$\boldsymbol{y} = g\left(\mathrm{X}\boldsymbol{\beta}\right) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Omega_\lambda), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2), \tag{1}$$

where g is a non-linear differentiable and invertible link function, X is a $n \times k$ model matrix constructed from the explanatory variables (i.e., $\boldsymbol{x}$) at the pixel level, $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots \beta_k]^\top$ are the normally-distributed parameters with mean $\mathbf{0}$ and positive semidefinite variance–covariance matrix Ω_λ which may depend on some parameters λ, and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \dots, \epsilon_n]^\top$ are the normally-distributed residuals with mean $\mathbf{0}$ and variance–covariance matrix $\sigma^2\mathbf{I}$. Through the text, we denote the generalized inverse of Ω_λ as Ω_λ^{-1}.

Next, we follow Proietti (2006) and denote the aggregation procedure in matrix notation as:

$$\boldsymbol{Y} = \mathrm{A}\boldsymbol{y} = \mathrm{A}g\left(\mathrm{X}\boldsymbol{\beta}\right) + \mathrm{A}\boldsymbol{\epsilon}$$

where A is a $N \times n$ sparse binary matrix that points each pixel to a block area. This construction of A allows the incorporation of any linear mapping, the most relevant of them being the sum, mean and weighted mean, which cover most practical applications. Here, we avoid unnecessary notation by making no distinctions between the random variable and its corresponding observations.

The non-linearity introduced by the link function precludes the derivation of a closed-form expression for $p(\boldsymbol{Y})$ (Lindstrom and Bates, 1990). At this point, we can take advantage on the fact that the multiplication by A is a linear operation to represent the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{\beta}$ as:

$$\boldsymbol{Y}|\boldsymbol{\beta} \sim \mathcal{N}\left(\mathrm{A}g\left(\mathrm{X}\boldsymbol{\beta}\right), \quad \mathrm{AA}^\top\sigma^2\right), \qquad \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \Omega_\lambda\right).$$

Now, we could follow any of the approaches for nonlinear mixed model estimation described by Demidenko (2013), such as two-stage estimation, the Lindstrom–Bates estimator (Lindstrom and Bates, 1990), and the Laplace approximation. We opt for the latter, which has the great advantages of representing the likelihood already as a function of the predicted random effects, and avoiding the calculation of determinants for $N \times N$ matrices. Recalling that $p(\boldsymbol{Y}, \boldsymbol{\beta}) = p(\boldsymbol{Y}|\boldsymbol{\beta})p(\boldsymbol{\beta})$, and denoting by $\hat{\boldsymbol{\beta}}_\lambda$ the mode of $p(\boldsymbol{Y}, \boldsymbol{\beta})$, we have:

$$\begin{aligned}
p(\boldsymbol{Y}) &= \int p(\boldsymbol{Y}, \boldsymbol{\beta}) d\boldsymbol{\beta} = \int \exp\left(\log p(\boldsymbol{Y}, \boldsymbol{\beta})\right) d\boldsymbol{\beta} \\
&\approx \int \exp\left(\log p(\boldsymbol{Y}, \hat{\boldsymbol{\beta}}_\lambda) + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)^\top \frac{\partial^2 \log p(\boldsymbol{Y}, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^\top}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)\right) d\boldsymbol{\beta} \\
&\approx p(\boldsymbol{Y}, \hat{\boldsymbol{\beta}}_\lambda) \int \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)^\top \{\mathrm{X}^\top \hat{\mathrm{G}}_\lambda^\top \mathrm{A}^\top (\mathrm{AA}^\top\sigma^2)^{-1} \mathrm{A}\hat{\mathrm{G}}_\lambda \mathrm{X} + \Omega_\lambda^{-1}\}\right. \\
&\quad \left. \times (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)\right) d\boldsymbol{\beta} \\
&= p(\boldsymbol{Y}, \hat{\boldsymbol{\beta}}_\lambda) \frac{(2\pi)^{k/2}}{\det(\mathrm{X}^\top \hat{\mathrm{G}}_\lambda^\top \mathrm{A}^\top (\mathrm{AA}^\top\sigma^2)^{-1} \mathrm{A}\hat{\mathrm{G}}_\lambda \mathrm{X} + \Omega_\lambda^{-1})^{1/2}},
\end{aligned} \tag{2}$$

where $\hat{\boldsymbol{\beta}}_\lambda = \operatorname{argmin}_\beta \left\{ \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\boldsymbol{\beta}\right)\right)^\top \left(\mathrm{AA}^\top\sigma^2\right)^{-1} \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\boldsymbol{\beta}\right)\right) + \boldsymbol{\beta}^\top \Omega_\lambda^{-1}\boldsymbol{\beta} \right\}$, and $\hat{\mathrm{G}}_\lambda$ is the Jacobian matrix, i.e., a $n \times n$ diagonal matrix with $\operatorname{diag}\left(\hat{\mathrm{G}}_\lambda\right) = g'\left(\mathrm{X}\hat{\boldsymbol{\beta}}_\lambda\right)$. Following Demidenko (2013), the resulting likelihood is approximated, as we consider a quadratic expansion in the second line, and assume that the term containing second-order derivatives of $g\left(\mathrm{X}\boldsymbol{\beta}\right)$ with respect to $\boldsymbol{\beta}$ in the third line tends to zero. Besides, the last line comes from recognizing the integrand as a part of a multivariate normal probability density function that must integrate to one. The profile log-likelihood obtained with the Laplace approximation thus becomes:

$$\begin{aligned}
\ell(\lambda, \sigma^2) = &-\frac{1}{2}\left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\hat{\boldsymbol{\beta}}_\lambda\right)\right)^\top \left(\mathrm{AA}^\top\sigma^2\right)^{-1} \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\hat{\boldsymbol{\beta}}_\lambda\right)\right) \\
&- \frac{1}{2}\hat{\boldsymbol{\beta}}_\lambda^\top \Omega_\lambda^{-1} \hat{\boldsymbol{\beta}}_\lambda \\
&- \frac{1}{2}\log\det\left(\mathrm{AA}^\top\sigma^2\right) - \frac{1}{2}\log\det\left(\Omega_\lambda\right) - \frac{N}{2}\log(2\pi) \\
&- \frac{1}{2}\log\det\left(\mathrm{X}^\top \hat{\mathrm{G}}_\lambda^\top \mathrm{A}^\top (\mathrm{AA}^\top\sigma^2)^{-1} \mathrm{A}\hat{\mathrm{G}}_\lambda \mathrm{X} + \Omega_\lambda^{-1}\right).
\end{aligned} \tag{3}$$

The standard numerical optimization for Eq. (3) would demand a two-step procedure: (i) for an initial guess λ, calculates $\hat{\boldsymbol{\beta}}_\lambda$ via penalized least squares; (ii) then, with $\hat{\boldsymbol{\beta}}_\lambda$ fixed, updates the estimates for λ. While repeating these procedures several times can be practical for small datasets, it becomes very inefficient and even prohibitive for larger datasets because we cannot store the matrix X nor the product $\mathrm{X}\boldsymbol{\beta}$ when n is huge. That means that for each guess for λ, we would have to iterate over chunks of the whole dataset several times in the first step to solve the penalized least squares and then again in the second step to calculate the last log-determinant term. All strategies for nonlinear mixed models described by Demidenko (2013) would result in this same problem. Ideally, we would prefer a more efficient method to iterate over the whole dataset as little as possible, saving time and computation.

To overcome this issue, we combine the Laplace approximation to the likelihood with a first-order approximation to the nonlinear least squares using the Levenberg–Marquardt algorithm (see Golden (2020)). As we show next, this strategy replaces the penalized nonlinear least squares by a problem with a closed-form solution, resulting in a much more efficient scheme that requires only one iteration over the whole dataset per each optimization step. This can be achieved by modifying Eq. (3), linearizing the nonlinear least squares around some initial guess $\bar{\boldsymbol{\beta}}^{[\eta]}$ and avoiding large steps with a quadratic penalty for the difference between the initial guess and the solution $\bar{\boldsymbol{\beta}}^{[\eta+1]}$. Here, the superscript η denotes the iteration number, and the dependence of $\bar{\boldsymbol{\beta}}^{[\eta+1]}$ on λ is omitted to avoid clutter:

$$\begin{aligned}
\ell(\lambda, \sigma^2) \approx &-\frac{1}{2}\left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]} - \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta+1]}\right)^\top \\
&\times \left(\mathrm{AA}^\top\sigma^2\right)^{-1} \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]} - \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta+1]}\right) \\
&- \frac{1}{2}\rho^{[\eta]}\left(\bar{\boldsymbol{\beta}}^{[\eta+1]} - \bar{\boldsymbol{\beta}}^{[\eta]}\right)^\top \left(\bar{\boldsymbol{\beta}}^{[\eta+1]} - \bar{\boldsymbol{\beta}}^{[\eta]}\right) \\
&- \frac{1}{2}\bar{\boldsymbol{\beta}}^{[\eta+1]\top}\Omega_\lambda^{-1}\bar{\boldsymbol{\beta}}^{[\eta+1]} - \frac{1}{2}\log\det\left(\mathrm{AA}^\top\sigma^2\right) \\
&- \frac{1}{2}\log\det\left(\Omega_\lambda\right) - \frac{N}{2}\log(2\pi) \\
&- \frac{1}{2}\log\det\left(\mathrm{X}^\top \bar{\mathrm{G}}^{[\eta]\top} \mathrm{A}^\top (\mathrm{AA}^\top\sigma^2)^{-1} \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X} + \Omega_\lambda^{-1}\right),
\end{aligned} \tag{4}$$

where $\rho^{[\eta]}$ is the problem-specific weight for the Levenberg–Marquardt-type penalty term, and

$$\begin{aligned}
\bar{\boldsymbol{\beta}}^{[\eta+1]} = \operatorname*{argmin}_{\boldsymbol{\beta}} \Bigg\{ &\left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]} - \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\boldsymbol{\beta}\right)^\top \\
&\times \left(\mathrm{AA}^\top\sigma^2\right)^{-1} \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right)\right. \\
&\left. + \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]} - \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\boldsymbol{\beta}\right) \\
&+ \rho^{[\eta]}\left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{[\eta]}\right)^\top \left(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{[\eta]}\right) + \boldsymbol{\beta}^\top\Omega_\lambda^{-1}\boldsymbol{\beta} \Bigg\}.
\end{aligned}$$

Differentiating and setting to zero, we arrive at the following closed-form solution to update $\bar{\boldsymbol{\beta}}^{[\eta]}$ at each iteration:

$$\bar{\boldsymbol{\beta}}^{[\eta+1]} = \left(\mathrm{X}^\top \bar{\mathrm{G}}^{[\eta]\top} \mathrm{A}^\top (\mathrm{A}\mathrm{A}^\top \sigma^2)^{-1} \mathrm{A}\bar{\mathrm{G}}^{[\eta]} \mathrm{X} + \Omega_\lambda^{-1} + \rho^{[\eta]} \mathbf{I}\right)^{-1} \times \left[\mathrm{X}^\top \bar{\mathrm{G}}^{[\eta]\top} \mathrm{A}^\top \left(\mathrm{A}\mathrm{A}^\top \sigma^2\right)^{-1} \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \mathrm{A}\bar{\mathrm{G}}^{[\eta]} \mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \rho^{[\eta]} \bar{\boldsymbol{\beta}}^{[\eta]}\right]. \tag{5}$$

Therefore, we consider the following iterative method for model estimation:

1. First, start a vector of trial values $\bar{\boldsymbol{\beta}}^{[\eta]}$ and the weight $\rho^{[\eta]}$.
2. Then, repeats for $\eta = 1, 2, \ldots$ until the difference between successive likelihood evaluations is small:
   (a) Calculates all the necessary aggregated matrices in parallel;
   (b) Finds $\hat{\boldsymbol{\lambda}}^{[\eta]}$ and $\hat{\sigma}^{2[\eta]}$ by maximizing Eq. (4) numerically, which also returns $\bar{\boldsymbol{\beta}}^{[\eta+1]}$ (Eq. (5));
   (c) Updates the weight $\rho^{[\eta]}$ based on some rule, for example:
      - If step (b) increases the likelihood, then updates $\rho^{[\eta+1]} = 0.75\rho^{[\eta]}$ and continues
      - Otherwise, sets $\rho^{[\eta]} = 2\rho^{[\eta]}$ and returns to step (b);
   (d) $\eta = \eta + 1$.

A more detailed explanation of how the proposed solution relates to the constraints described in Section 2.1 follows.

2.2.2. Further details

In the present subsubsection, we describe how the well-known formulation shown in Section 2.2.1 addresses the five constraints described in Section 2.1. A special focus on the last one is given, since it corresponds to the novelty of our work.

For constraint (1), we model $g^{-1}(\boldsymbol{y})$ instead of $\boldsymbol{y}$ at the disaggregated level. The link function maps the possibly unbounded product between X and its coefficients to the desired (a, b) interval. Choosing a proper link function is a problem-specific decision, but some examples include the identity function when $\boldsymbol{y}$ is unbounded (i.e., $g^{-1}(\boldsymbol{y}) = \boldsymbol{y}$), the exponential function when $\boldsymbol{y}$ is positive, or the generalized logistic function for other intervals.

For constraint (2), we consider all the smooth functions $s(x)$ of (a possibly multivariate) variable x that can be represented as the product between spline basis and coefficients and whose wiggliness can be controlled by a quadratic penalty on these coefficients. In other words, we focus on smoothers that can be represented as $\mathrm{X}\boldsymbol{\beta}$ with one or more $\lambda\boldsymbol{\beta}^\top \mathrm{S}\boldsymbol{\beta}$ penalty terms, for some properly chosen and known model matrix X and penalty matrix S, and unknown smoothing parameter λ. This category covers several options, including P-splines, cubic splines, thin plate regression splines, tensor product interactions, Gaussian Markov Random Fields, among others (Stasinopoulos et al., 2017). The construction of proper model and penalty matrices is beyond the scope of this work, and the reader may refer to Wood (2017) for further methodological details. For our work, it is sufficient to note that the quadratic term of Eq. (3) shows exactly a penalty on the inverse of the variance–covariance matrix adopted for $\boldsymbol{\beta}$ in Eq. (1), which means that we must parameterize our model so that $\Omega_\lambda = \mathrm{S}^{-1}$. Such a parameterization corresponds to the Bayesian interpretation of having an informative prior distribution on the vector of coefficients, $\boldsymbol{\beta}$ (Kimeldorf and Wahba, 1970; Silverman, 1985; Wood, 2017). This view of the process allows the estimation using a likelihood approach and as the regularization parameter $\rho^{[\eta]}$ approaches zero after several iterations (Demidenko, 2013), the linearized likelihood (Eq. (4)) tends to the likelihood of a linear mixed model. This means that at convergence, the posterior distribution of $\boldsymbol{\beta}$ will be approximately Gaussian with its mean described by Eq. (5) and variance–covariance matrix V (Eq. (6)) (Wood, 2017). We can take advantage of this result to generate credible intervals for the smoothers, represented here as linear combinations of $\boldsymbol{\beta}$. It is precisely this link between smoothers and mixed-models that allows the proposed formulation to generate nonlinear functions between the response variable and its predictors.

$$\mathrm{V} = \left(\mathrm{X}^\top \bar{\mathrm{G}}^{[\eta]\top} \mathrm{A}^\top (\mathrm{A}\mathrm{A}^\top \sigma^2)^{-1} \mathrm{A}\bar{\mathrm{G}}^{[\eta]} \mathrm{X} + \Omega_\lambda^{-1} + \rho^{[\eta]} \mathbf{I}\right)^{-1}. \tag{6}$$

Constraint (3), i.e., representing the spatial or spatiotemporal structure of the data, is a challenging part of the disaggregation model. Standard geostatistical approaches demand the storage and computation of dense covariance matrices (Wikle, 2010), leading to the so-called "big n problem" (Lasinio et al., 2012). These approaches are only adequate for applications in a small-data setting, and some lower-dimensional approach must be used to allow the incorporation of larger datasets. Fortunately, Eq. (1) can also be used to accommodate structures that account for spatial and spatiotemporal variation on the systematic part of the model (Stasinopoulos et al., 2019), which opens up the possibility to incorporate some more computationally attractive options such as kernel methods, stochastic partial differential equations, splines and wavelets (Wikle, 2010; Lindgren et al., 2011; Simpson et al., 2012). While the link between some of these approaches may not seem straightforward at first, the work of Miller et al. (2020) is helpful in showing how they converge with the idea presented in Eq. (1). Thus, because the estimation method proposed is general, it can incorporate different branches of the spatial and spatiotemporal modeling literature.

To address constraint (4), we represent the aggregation procedure in matrix form as a part of the model formulation. The likelihood derived (i.e., Eq. (3)) naturally represents the error between observed and predicted aggregated variables in its first term. One novelty of our work is the demonstration we provide in the Appendix, which shows that the linearization procedure of Eq. (4) is guaranteed to converge to a critical point of the likelihood. This means that each iteration can be seen as an improvement of our initial random guess for the coefficients. After several steps, predicted values become closer and closer to the aggregated data, until stagnation at convergence.

The major novelty of our work consists of addressing the computational aspect of constraint (5), allowing the estimation method to be scaled. If no special treatment is adopted, then the storage of n-dimensional matrices and vectors could easily cause memory problems. To overcome this issue, we propose the generation of aggregated matrices on-the-fly, which can be done by reading and storing only a part of the X matrix at a time. This is often possible in platforms that deal with huge datasets, and allows us to compute the N-dimensional aggregated matrices (i.e., $\mathrm{A}g(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]})$, $\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}$, $\mathrm{A}\mathrm{A}^\top$ and $\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}$) without having to store any of the n-dimensional matrices of the model (i.e., $\mathrm{A}_{N\times n}$, $\mathrm{X}_{n\times k}$, $g(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]})_{n\times 1}$, and $\bar{\mathrm{G}}^{[\eta]}_{n\times n}$).

The technique consists of splitting the multiplication of n-dimensional matrices and vectors into independent processes that can run concomitantly in parallel. To illustrate with the vector $\mathrm{A}g(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]})$, what we propose is the division of the n-dimensional vectors and matrices into ω non-overlapping n_u-dimensional subsets such that $\sum_{u=1}^{\omega} n_u = n$. Denoting each subset with an underscore, calculation can be performed as:

$$\begin{aligned}\mathrm{A}g(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}) &= \left[\mathrm{A}_1 \ldots \mathrm{A}\omega\right] g\left(\begin{bmatrix}\mathrm{X}_1\\ \vdots \\ \mathrm{X}_\omega\end{bmatrix}\bar{\boldsymbol{\beta}}^{[\eta]}\right)\\ &= \left[\mathrm{A}_1 \ldots \mathrm{A}\omega\right]\begin{bmatrix} g(\mathrm{X}_1\bar{\boldsymbol{\beta}}^{[\eta]})\\ \vdots \\ g(\mathrm{X}_\omega\bar{\boldsymbol{\beta}}^{[\eta]})\end{bmatrix}\\ &= \mathrm{A}_1 g(\mathrm{X}_1\bar{\boldsymbol{\beta}}^{[\eta]}) + \cdots + \mathrm{A}_\omega g(\mathrm{X}_\omega\bar{\boldsymbol{\beta}}^{[\eta]}).\end{aligned} \tag{7}$$

It should be clear from Eq. (7) that each of the terms in the summation can be calculated efficiently and independently, thus alleviating the burden of performing n-dimensional products. For the other vectors,

$\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}$, $\mathrm{AA}^\top$ and $\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}$, the logic is analogous, and extra efficiency can be gained by exploiting the diagonal structure of $\bar{\mathrm{G}}^{[\eta]}$. This implies that we do not even have to store the $n_u \times n_u$ block $\bar{\mathrm{G}}_u^{[\eta]}$, but only its diagonal vector. For the example of $\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}$, we denote the element-wise product as $\odot$ to obtain:

$$\begin{aligned}
\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]} &= \begin{bmatrix}\mathrm{A}_1 \dots \mathrm{A}_\omega\end{bmatrix}\begin{bmatrix}\bar{\mathrm{G}}_1^{[\eta]} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \bar{\mathrm{G}}_\omega^{[\eta]}\end{bmatrix}\begin{bmatrix}\mathrm{X}_1\\ \vdots\\ \mathrm{X}_\omega\end{bmatrix}\bar{\boldsymbol{\beta}}^{[\eta]}\\
&= \begin{bmatrix}\mathrm{A}_1 \dots \mathrm{A}_\omega\end{bmatrix}\begin{bmatrix}\bar{\mathrm{G}}_1^{[\eta]}\mathrm{X}_1\\ \vdots\\ \bar{\mathrm{G}}_\omega^{[\eta]}\mathrm{X}_\omega\end{bmatrix}\bar{\boldsymbol{\beta}}^{[\eta]}\\
&= \mathrm{A}_1\bar{\mathrm{G}}_1^{[\eta]}\mathrm{X}_1\bar{\boldsymbol{\beta}}^{[\eta]} + \cdots + \mathrm{A}_\omega\bar{\mathrm{G}}_\omega^{[\eta]}\mathrm{X}_\omega\bar{\boldsymbol{\beta}}^{[\eta]}\\
&= \mathrm{A}_1\left[\mathrm{diag}\left(\bar{\mathrm{G}}_1^{[\eta]}\right)\odot\left(\mathrm{X}_1\bar{\boldsymbol{\beta}}^{[\eta]}\right)\right] + \cdots\\
&\quad + \mathrm{A}_\omega\left[\mathrm{diag}\left(\bar{\mathrm{G}}_\omega^{[\eta]}\right)\odot\left(\mathrm{X}_\omega\bar{\boldsymbol{\beta}}^{[\eta]}\right)\right].
\end{aligned} \tag{8}$$

The number of subsets, ω, must be defined based on the number of computational resources available and the size of the problem under consideration. The use of Eqs. (7) and (8) reduces the storage demand from n-dimensional to N-dimensional matrices and vectors, which is generally a huge gain in spatial disaggregation problems that make problems computationally feasible.

2.2.3. Practical improvements

Even though Eqs. (4)–(6) are theoretically correct, using these formulas directly for numerical optimization can become expensive in complex applications. In this section, we present some simple improvements based on Wood (2017) that can simplify and stabilize these expressions.

Let us consider first the decomposition of $\mathrm{F}^{-1} = (\mathrm{AA}^\top)^{-1} = \mathrm{L}^\top\mathrm{L}$, which is output from the parallel calculations described in Section 2.2.2. Since F will be diagonal with positive only elements, its Cholesky factor L will also be a diagonal matrix, with $L_{ii} = \sqrt{1/F_{ii}}$. Because L can be computed directly, a full decomposition of F is unnecessary and only the diagonal of F (and of L) must be stored. Then, we can simplify part of the likelihood calculation by multiplying the other outputs of the parallel calculation prior to the numerical optimization: $\tilde{\mathrm{X}} = \mathrm{L}\left(\mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\right)$ and $\tilde{Y} = \mathrm{L}\left(\boldsymbol{Y} - \mathrm{A}g(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}) + \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]} - \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta+1]}\right)$. These products can be performed efficiently by simply multiplying each of their columns by the diagonal of L. Now, the long log-determinant term in Eq. (4), for example, becomes simply $\log\det(\tilde{\mathrm{X}}^\top\tilde{\mathrm{X}}/\sigma^2 + \Omega_\lambda^{-1})$.

In fact, even the multiplication of N-dimensional matrices can be potentially costly depending on the amount of block areas. To overcome this problem, we can take the QR decomposition of the $\tilde{\mathrm{X}}$ matrix into the product of a $\mathrm{Q}_{N\times k}$ orthogonal and a $\mathrm{R}_{k\times k}$ upper triangular matrix: $\tilde{\mathrm{X}} = \mathrm{QR}$. Because Q is orthogonal, the N-dimensional product $\tilde{\mathrm{X}}^\top\tilde{\mathrm{X}}$ can be replaced by the k-dimensional product $\mathrm{R}^\top\mathrm{R}$, and all other N-dimensional products can be replaced analogously. If we: (i) calculate $\mathrm{Y} = \mathrm{Q}^\top\tilde{Y}$ and $\mathrm{r} = \|\tilde{Y}\| - \|\mathrm{Y}\|$, (ii) use the general property that $\log\det(\sigma^2\cdot\mathrm{V}_{n\times n}) = n\log(\sigma^2) + \log\det(\mathrm{V})$, and (iii) remove constant terms; then the log-likelihood of Eq. (4) rewrites to:

$$\begin{aligned}
\ell(\lambda,\sigma^2) = &-\frac{\|\mathrm{Y} - \mathrm{R}\bar{\boldsymbol{\beta}}^{[\eta+1]}\|^2 + \mathrm{r}}{2\sigma^2} - \frac{1}{2}\bar{\boldsymbol{\beta}}^{[\eta+1]\top}\Omega_\lambda^{-1}\bar{\boldsymbol{\beta}}^{[\eta+1]}\\
&+ \frac{1}{2}\rho^{[\eta]}\left(\boldsymbol{\beta}^{[\eta+1]} - \bar{\boldsymbol{\beta}}^{[\eta]}\right)^\top\left(\boldsymbol{\beta}^{[\eta+1]} - \bar{\boldsymbol{\beta}}^{[\eta]}\right)\\
&+ \frac{1}{2}\log\det(\Omega_\lambda^{-1}) - \frac{1}{2}\log\det(\mathrm{R}^\top\mathrm{R}/\sigma^2 + \Omega_\lambda^{-1}) - \frac{N}{2}\log(\sigma^2).
\end{aligned} \tag{9}$$

The same modifications also simplify the closed-form expressions for Eqs. (5) and (6). As $\rho^{[\eta]}$ approaches zero:

$$\begin{aligned}
&\boldsymbol{\beta}|\boldsymbol{Y} \sim \mathcal{N}\left((\mathrm{R}^\top\mathrm{R}/\sigma^2 + \Omega_\lambda^{-1} + \rho^{[\eta]}\mathbf{I})^{-1}(\mathrm{R}^\top\mathrm{Y}/\sigma^2 + \rho^{[\eta]}\boldsymbol{\beta}^{[\eta]}),\right.\\
&\left.(\mathrm{R}^\top\mathrm{R}/\sigma^2 + \Omega_\lambda^{-1} + \rho^{[\eta]}\mathbf{I})^{-1}\right).
\end{aligned} \tag{10}$$

The numerical optimization of Eq. (9) can be made using any implementation available in specialized software. The preference for gradient-based method with proper analytical derivatives can accelerate the process, as the derivatives can also be expressed as a function of R and Y. Since σ^2 and λ are necessarily positive quantities, it is easier to construct an auxiliary vector $\boldsymbol{\zeta}$ to work on a logarithm scale, with $\sigma^2 = \exp(\zeta_1)$ and $\lambda_j = \exp(\zeta_j)$. The derivatives can then be calculated as in Eq. (11) and Eq. (12):

$$\begin{aligned}
\frac{\partial\ell}{\partial\zeta_1} = &\frac{\|\mathrm{Y} - \mathrm{R}\bar{\boldsymbol{\beta}}^{[\eta+1]}\|^2 + \mathrm{r}}{2\exp(\zeta_1)} - \frac{N}{2} + \frac{\mathrm{tr}\left[\left(\mathrm{R}^\top\mathrm{R}/\exp(\zeta_1) + \Omega_\lambda^{-1}\right)^{-1}\mathrm{R}^\top\mathrm{R}/\exp(\zeta_1)\right]}{2}\\
&- \frac{\partial\bar{\boldsymbol{\beta}}^{[\eta+1]}}{\exp(\zeta_1)}\left[\frac{\mathrm{R}^\top\mathrm{R}\bar{\boldsymbol{\beta}}^{[\eta+1]} - \mathrm{R}^\top\mathrm{Y}}{\exp(\zeta_1)} + \Omega_\lambda^{-1}\bar{\boldsymbol{\beta}}^{[\eta+1]} + \rho^{[\eta]}\left(\bar{\boldsymbol{\beta}}^{[\eta+1]} - \bar{\boldsymbol{\beta}}^{[\eta]}\right)\right]
\end{aligned} \tag{11}$$

$$\begin{aligned}
\frac{\partial\ell}{\partial\zeta_j} = &-\frac{1}{2}\bar{\boldsymbol{\beta}}^{[\eta+1]\top}\frac{\partial\Omega_\lambda^{-1}}{\partial\exp(\zeta_j)}\bar{\boldsymbol{\beta}}^{[\eta+1]} + \frac{1}{2}\mathrm{tr}\left[\Omega_\lambda^{-1}\frac{\partial\Omega_\lambda^{-1}}{\partial\exp(\zeta_j)}\right]\\
&- \frac{1}{2}\mathrm{tr}\left[\left(\mathrm{R}^\top\mathrm{R}/\exp(\zeta_1) + \Omega_\lambda^{-1}\right)^{-1}\frac{\partial\Omega_\lambda^{-1}}{\partial\exp(\zeta_j)}\right]\\
&- \frac{\partial\bar{\boldsymbol{\beta}}^{[\eta+1]}}{\exp(\zeta_j)}\left[\frac{\mathrm{R}^\top\mathrm{R}\bar{\boldsymbol{\beta}}^{[\eta+1]} - \mathrm{R}^\top\mathrm{Y}}{\exp(\zeta_1)} + \Omega_\lambda^{-1}\bar{\boldsymbol{\beta}}^{[\eta+1]} + \rho^{[\eta]}\left(\bar{\boldsymbol{\beta}}^{[\eta+1]} - \bar{\boldsymbol{\beta}}^{[\eta]}\right).\right]
\end{aligned} \tag{12}$$

When evaluating Eqs. (11) and (12), the derivatives of Ω_λ^{-1} with respect to ζ_j will vary according to the structures adopted for the disaggregated model. Besides, the right term in the second lines is exactly what we had to set to zero to arrive at Eq. (5) (and, consequently, Eq. (10)). This means that, in both Equations, such term will always equal zero.

2.3. Simulation study

The model performance was assessed with a simulation study (Fig. 1). We created a $1,000\times1,000$ response variable at the pixel level (i.e., $n = 10^6$) by summing a non-linear function of a known explanatory variable and a non-linear joint function of the spatial coordinates. Then, we aggregated the exponential of the response variable over 90 randomly generated block areas (i.e., $N = 90$) with different sizes and shapes.

Denoting the spatial coordinates as $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and the explanatory variable as $\boldsymbol{x}_3$, the hypothetical response variable $\boldsymbol{y}$ was created using:

$$\begin{aligned}
\boldsymbol{y} = &\exp\left[-4.5 - 4\sin(\boldsymbol{x}_3)/(1+\boldsymbol{x}_3^2) + 3(\boldsymbol{x}_1 - 20)^2(\boldsymbol{x}_2 + 15)\right.\\
&\left. - 2(\boldsymbol{x}_2 + 25)^2(\boldsymbol{x}_1 - 5)\right] + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 = 0.01)
\end{aligned}$$

Fig. 1 shows the explanatory variable $\boldsymbol{x}_3$ (top left), an uncorrelated variable $\boldsymbol{x}_4$ (bottom left), and the 90 block areas with the corresponding exponential aggregation of the simulated response variable over them (right).

The disaggregated level model adopted a tensor product between two cubic splines for the joint function of the spatial coordinates (Wood, 2006) and cubic splines for the univariate variables. In both cases, we adopted the so-called "shrinkage" smoothers (Marra and Wood, 2011), which replace the zero eigenvalues of the penalty matrices for small real numbers, so non-significant predictors are naturally removed during the model fit. To evaluate the shrinkage effect and to test the model's sensitivity to specification mistakes, we also included an uncorrelated spatial variable (x_4) to the formulation:

$$\boldsymbol{y} = \exp\left[\alpha + s(\boldsymbol{x}_1, \boldsymbol{x}_2) + s(\boldsymbol{x}_3) + s(\boldsymbol{x}_4)\right] + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2).$$

The basis dimensions adopted were 20 for all cubic splines. The initial trial value adopted, $\boldsymbol{\beta}^{[1]}$, was randomly sampled from a Gaussian distribution with mean 0 and variance 0.01, and the weight of the Levenberg–Marquardt algorithm was started as $\rho^{[1]} = 10$.

Then, we compared our results with the potential conclusions that could be taken from the aggregated data. We first averaged $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ over the block areas and plotted against the logarithm of the aggregated

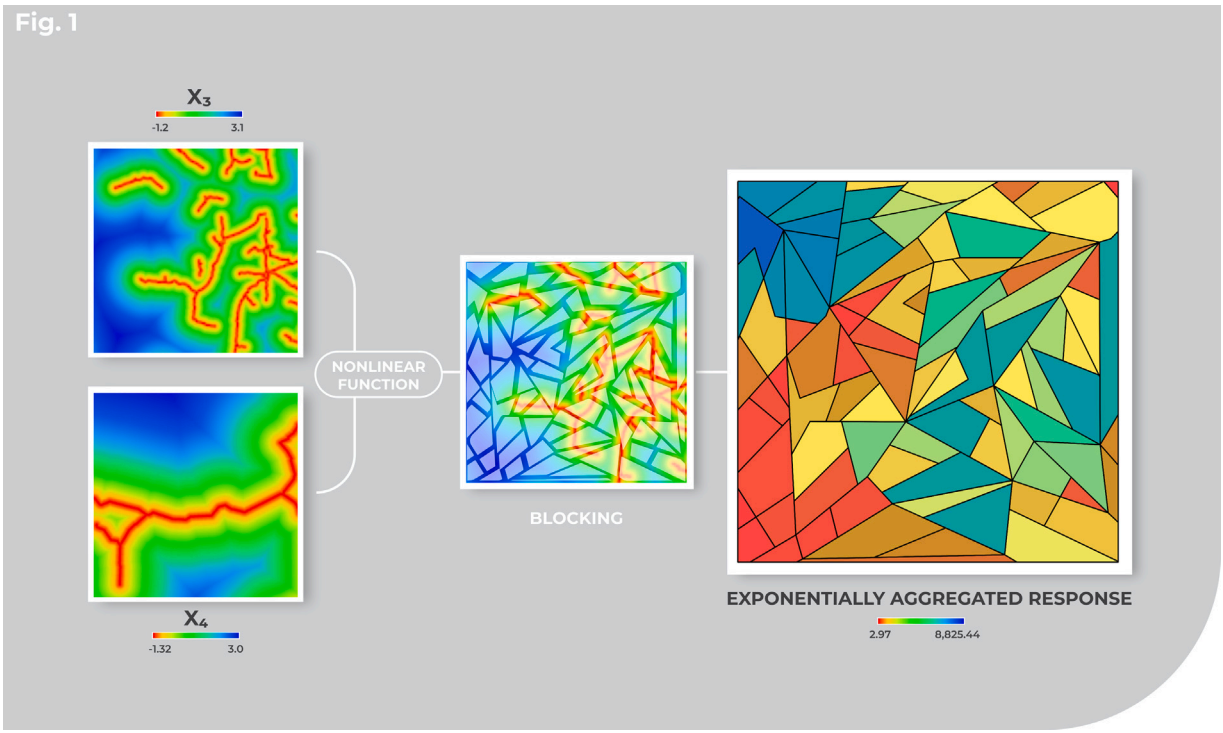

Fig. 1. The simulation procedure adopted. x_3 (top left), and x_4 (bottom left). The right image shows the exponentially aggregated response variable in a red–yellow–blue palette ranging from red: 2.97 to blue: 8,825.44. The continuous lines represent the boundaries of block areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

response variable (i.e., $\log \boldsymbol{Y}$). A LOESS curve was also added to show the trend.

Codes were written in R, version 3.5.1 (R. Core Team, 2020). The implementation of the smoothers was taken from the mgcv (Wood, 2017). Code parallelization was achieved with the Rmpi (Yu, 2002), snow (Tierney et al., 2016), and parallel (R. Core Team, 2020) packages, and the manipulation of spatial data was performed using the raster (Hijmans, 2017) package. All source codes used are available as Supplementary Materials.

2.3.1. Sensitivity analysis

One important question for the simulation is how well the model performs under different conditions, for example, a higher variability on the simulated relationships and a smaller number of block areas. To clarify this question, we performed a sensitivity analysis by replicating the same problem but with different combinations of σ^2 and N. We generated 25 combinations by setting $\sigma = 0.1, 0.3, 0.5, 0.7, 0.9$ and N = $10, 30, 50, 70, 90$. For each combination, ten simulations were run with different initial guesses. Then, the median values of the estimated standard deviation (i.e., $\hat{\sigma}$) and the sum of squared errors between predictions and true values were reported.

2.4. Application: Brazilian sugarcane

The application consisted of trying the algorithm in a real setting where the pixel-level data is unknown. In this exercise, we attempted to downscale Brazilian agricultural data gathered at a municipality (roughly equivalent to the United States' counties) level.

Brazil is one of the world's top producers and exporters of agricultural products and one of the world's largest country, with nearly 850 million hectares (Mha). It is administratively divided into 5,565 municipalities (i.e., $N = 5,565$), each of which records the production of its agricultural outcomes annually. Since the 1980s, Brazil has had the world's largest harvested area and production of sugarcane. In 2017, croplands covered a territorial area of 65 Mha, of which 9 Mha being devoted to sugarcane production (Fig. 2). Despite such a leading position, the country's average yield of sugarcane in 2018 was 74.4 ton/ha, far below the range of 100 to 120 ton/ha achieved by some of the world's most productive countries (Food and Agriculture Organization, 2020). Thus, research on sugarcane, including the generation of new information, is an active area that might aid the search for better use of natural resources combined with increases in agricultural productivity.

Since sugarcane production is a non-negative quantity, we adopted the exponential link function and assumed a sum between a non-linear joint function of the projected spatial coordinates (i.e., $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in meters) and non-linear functions of the explanatory variables. These include physical and chemical soil properties (i.e., the mass fraction of clay content in %, $\boldsymbol{x}_3$, the soil organic carbon content in g/kg, $\boldsymbol{x}_4$, and the cation exchange capacity in cmolc/kg $\boldsymbol{x}_5$ (Hengl et al., 2017)), terrain and climatic characteristics (i.e., terrain slope in %, $\boldsymbol{x}_6$ (USGS, 2018), average precipitation per month in mm, $\boldsymbol{x}_7$, and average maximum temperature per month in degrees Celsius, $\boldsymbol{x}_9$ (Abatzoglou et al., 2018)), and an infrastructure variable (i.e., distance from roads and railroads in kilometers, $\boldsymbol{x}_8$ (IBGE, 2018; DNIT, 2018)). All explanatory variables were aligned to a 100 m resolution grid, and the disaggregated-level model considered for sugarcane production ($\boldsymbol{p}$) was:

$$\boldsymbol{p} = \exp\left[\alpha + s(\boldsymbol{x}_1, \boldsymbol{x}_2) + s(\boldsymbol{x}_3) + s(\boldsymbol{x}_4) + s(\boldsymbol{x}_5) + s(\boldsymbol{x}_6) + s(\boldsymbol{x}_7) + s(\boldsymbol{x}_8) + s(\boldsymbol{x}_9)\right] + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$$

In this application, we adopted thin plate regression smoothers for both the multivariate and univariate variables (Wood, 2003). The dimension adopted for the spatial smoother was 500, and 10 for the univariate smoothers. Again, we adopted the shrinkage version of the smoothers, and the random trial value, $\boldsymbol{\beta}^{[1]}$, was sampled from a Gaussian distribution with mean 0 and variance 0.01. The initial Levenberg–Marquardt weight was $\rho^{[1]} = 2,500$. We reasonably assumed that sugarcane production only occurs in pixels marked as 'cropland' in

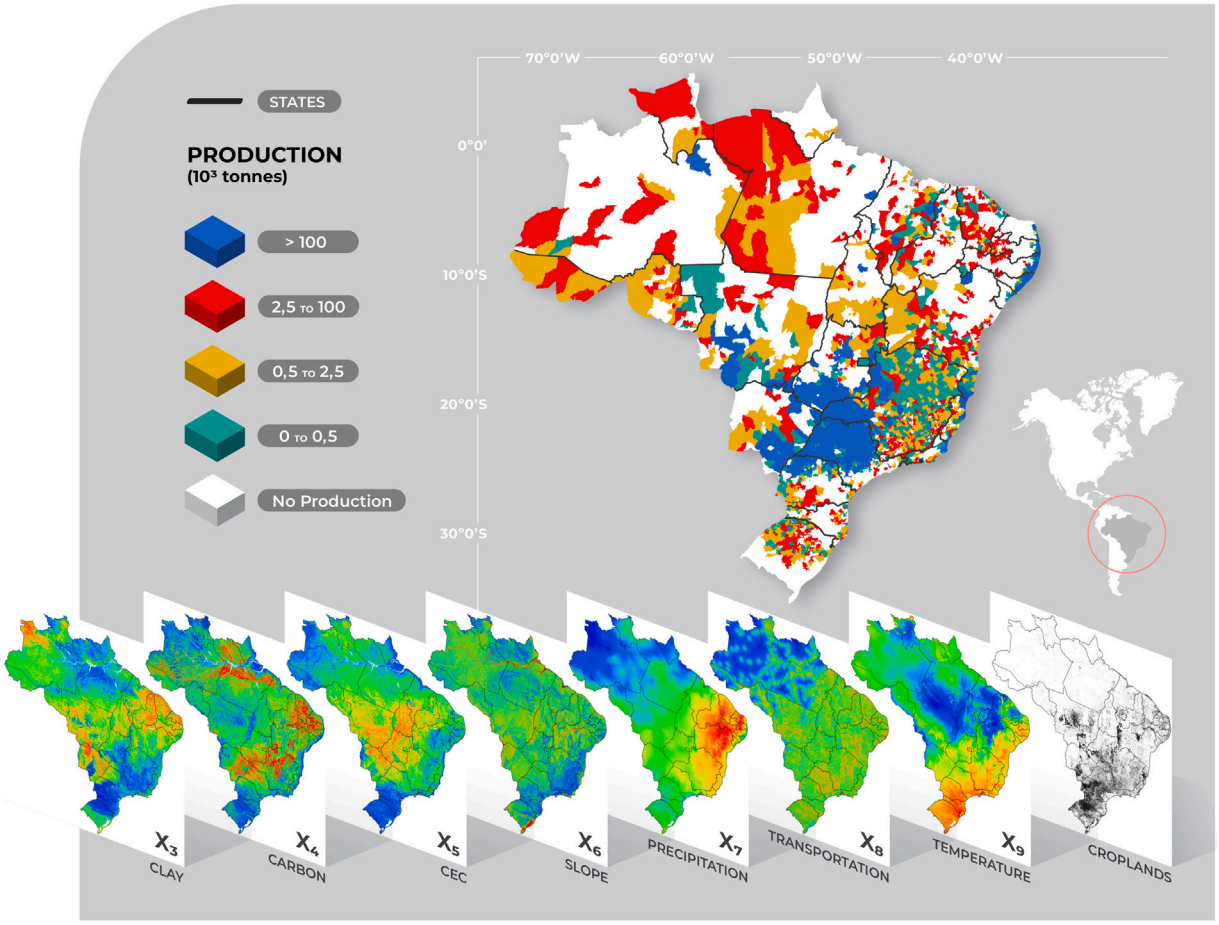

Fig. 2. Sugarcane production (top, with legend), and cropland areas plus predictor variables (bottom): x_3 is the mass fraction of clay content, x_4 is the soil organic carbon content, x_5 is the cation exchange capacity, x_6 is the terrain slope, x_7 is the average precipitation per month, x_8 is the distance from roads and railroads, and x_9 is the average maximum temperature per month. Each covariate has a different range, represented in a red–green–blue gradient (lower to upper values). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the land cover data (Fig. 2, bottom right). Such class was constructed by joining a broad 'agriculture' class with some more detailed classes with a local coverage (e.g., 'annual and perennial crop' and 'semi-perennial crop') (MapBiomas, 2019a; Souza et al., 2020; MapBiomas, 2019b). We aligned all available information to a 100 m pixel resolution, which results in approximately $n = 5 \cdot 10^8$ cropland pixels.

2.4.1. Validation

The validation strategy adopted for our example consisted of a qualitative and a quantitative evaluation. In the qualitative evaluation, we compared the smoothers estimated against the agronomic literature about the theoretical effects of different factors on sugarcane production. It is reasonable to assume that the disaggregated information will only be reliable if the smoothers have a physical meaning coherent with the specialized literature. The choice of a quantitative model evaluation is directly affected by the fact that the municipality level dataset used to fit the model contains Brazil's most detailed agricultural information. An important implication is that no finer agricultural production information (i.e., at the pixel level) is available to validate our results. To overcome this problem, we recall that the 'agriculture' class inputted into the model, as previously described in Section 2.4, was constructed by merging different detailed land-cover classes (e.g., permanent and temporary crops) into a single class. One of such classes is named "semi-perennial crops" (Souza et al., 2020; MapBiomas, 2019b), which can be assumed to be sugarcane, as done by similar studies with the same database (do Nascimento Bendini et al., 2019; Pavinato et al., 2020; Portinho et al., 2021). While this opens up the possibility of validating the spatial allocation, adopting the spatial database as our 'ground truth' could also be potentially problematic, so extra caution is needed.

Of the total 100.4 Mha of croplands in Brazil, the sugarcane area in the spatial database is 8.7 Mha (i.e., 8.66%), while that reported in the national statistics equals 10.2 Mha (i.e., 10.15%) (MapBiomas, 2019b; IBGE, 2020). Such difference is more pronounced when evaluated at the municipality level. A cross-check of information indicates that 3,431 Brazilian municipalities have sugarcane reported either in the spatial database or national statistics. Of these, 2,424 (i.e., 70.67%) have sugarcane according to only one of the two databases and not the other. In the remaining 1,006 municipalities, the relative difference in municipal sugarcane area reported in the two databases ranges from −99.998% to +4,400%, with a mean value and standard deviation of 1.72% and 197.11%, respectively. To ensure a fair model evaluation under these circumstances, we opted to restrict the analysis to only those municipalities whose relative difference of the values reported in both databases is less than 10%. Such a procedure resulted in the selection of 207 municipalities (6.03% of the 3,431 mentioned above). These are relevant for sugarcane for containing 28.50% of the production and from 27.49% to 31.94% of the total sugarcane area of the country according to the spatial database and national statistics, respectively (MapBiomas, 2019b; IBGE, 2020). In these 207 municipalities, the sugarcane area represents 40.20% of the total croplands area, with the shares per municipality ranging from 0.32% to 92.93%, and mean and standard deviation equal to 44.37% and 23.72%, respectively (MapBiomas, 2019b).

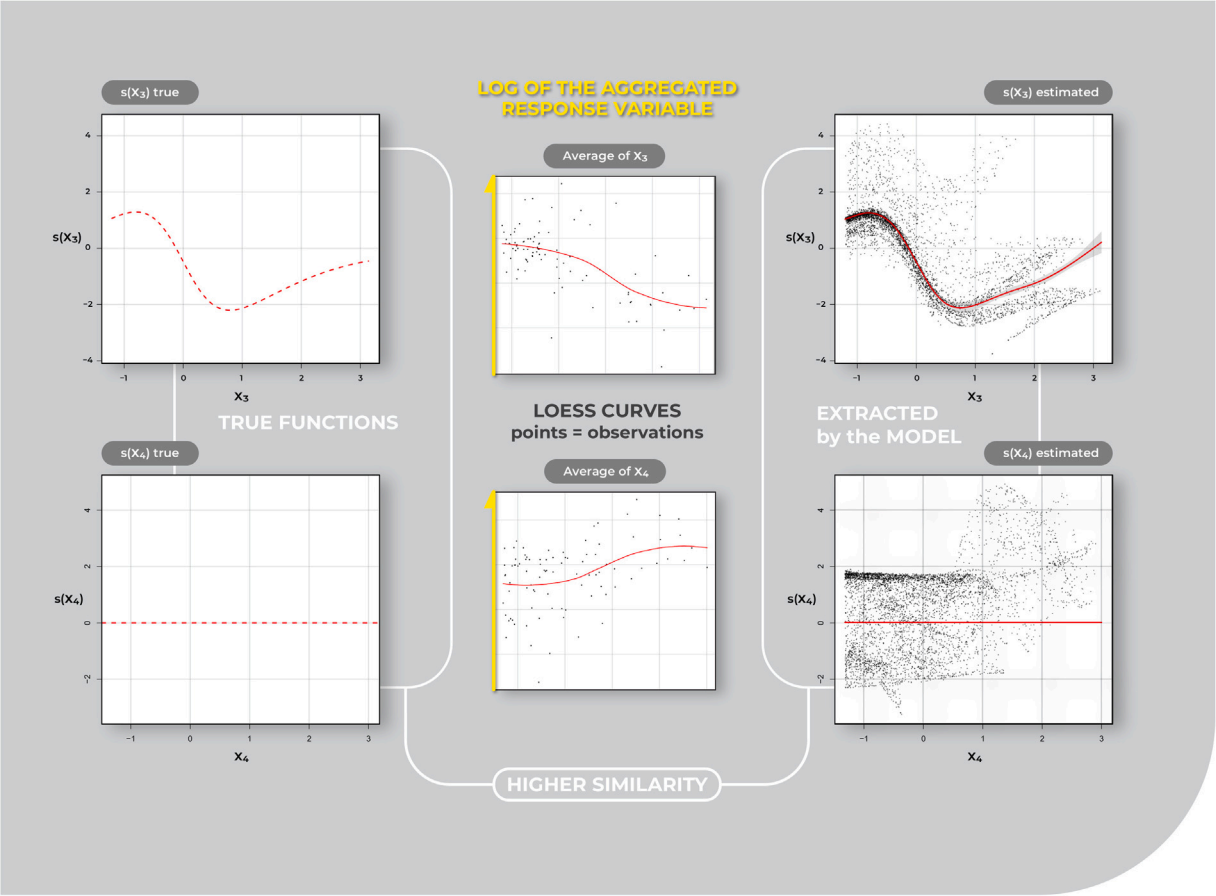

Fig. 3. Results from the simulation study. The left column shows the actual functions used to generate the data. The center column shows the relationship between the response and explanatory variables at the aggregated level, with observations represented as black points. The right column shows the relationship extracted by the model and the observations at the pixel level (which were unknown to the model during the fit).

The additional assumptions introduced do not enable the validation of the predicted production values themselves but allow assessing the accuracy of the spatial allocation of model predictions. Under this view, the model can be interpreted as a classifier: the predicted production being an indirect indicator of the chance of observing sugarcane on a given pixel and with crop presence only in pixels where the predicted values exceed some threshold. Because of the significant role played by the threshold in this evaluation, we followed the standard procedure for binary classification problems to construct a model curve to summarize the existing tradeoffs. The most common option is the Receiver Operating Characteristic (ROC), which balances between true and false-positive error (Maimon and Rokach, 2010; Swets, 1988). However, because our dataset is imbalanced due to the low share of sugarcane among all crops, the ROC curve may unrealistically overestimate model performance. We constructed the Precision–Recall (PR) curve to overcome this problem, which yields a more realistic model assessment for the condition described. The quantities represented, Precision and Recall, do not include the inflated true negative class and correspond to: (i) the probability that sugarcane is indeed present when the model predicts sugarcane; and (ii) the probability that the model predicts sugarcane in locations where there is indeed sugarcane present, respectively (Sofaer et al., 2019; Saito and Rehmsmeier, 2015). Precision and Recall are also known as user and producer accuracy, respectively, in some fields of knowledge. The best threshold for the model was defined as the point closest to the optimal theoretical threshold for the PR curve, and the overall accuracy of the spatial allocation was also reported.

3. Results and discussion

3.1. Simulated study

Fig. 3 showing hows analyzing aggregated data without accounting for the spatial variability can potentially mislead conclusions about the true relationship between the response and explanatory variables. While the true function of $\boldsymbol{x}_3$ (top left) resembles a periodic function with a maximum nearly −1 and a minimum nearly 1, the LOESS curve extracted from the aggregated data (top center) shows only a curvy decrease. For $\boldsymbol{x}_4$, the true function indicates no correlation with the response variable (bottom left), while the LOESS curve (bottom center) suggests an increasing relationship. In both cases, the curve estimated using the proposed method (right column) strongly resembles the original. As the disaggregation model explicitly addresses the spatial dependence, the smoothers (right column) have a relatively narrow standard deviation even with large variability in the data points. These results reinforce that wrong conclusions about the actual process would be taken if a simple correlation analysis was performed.

The model predictions at the unobserved disaggregated level (i.e., for $\boldsymbol{y}$) shown a high similarity with the corresponding true components. The left group of Fig. 4 compares the actual $\boldsymbol{y}$ (left), which were assumed unknown during model fit, and the model predictions (right). The right group row shows the original spatial term (i.e., $s(\boldsymbol{x}_1, \boldsymbol{x}_2)$, left), also assumed unknown during model fit, and the corresponding reconstruction predicted by the smoother adopted (right). In the response variable, the reconstruction is nearly indistinguishable from the

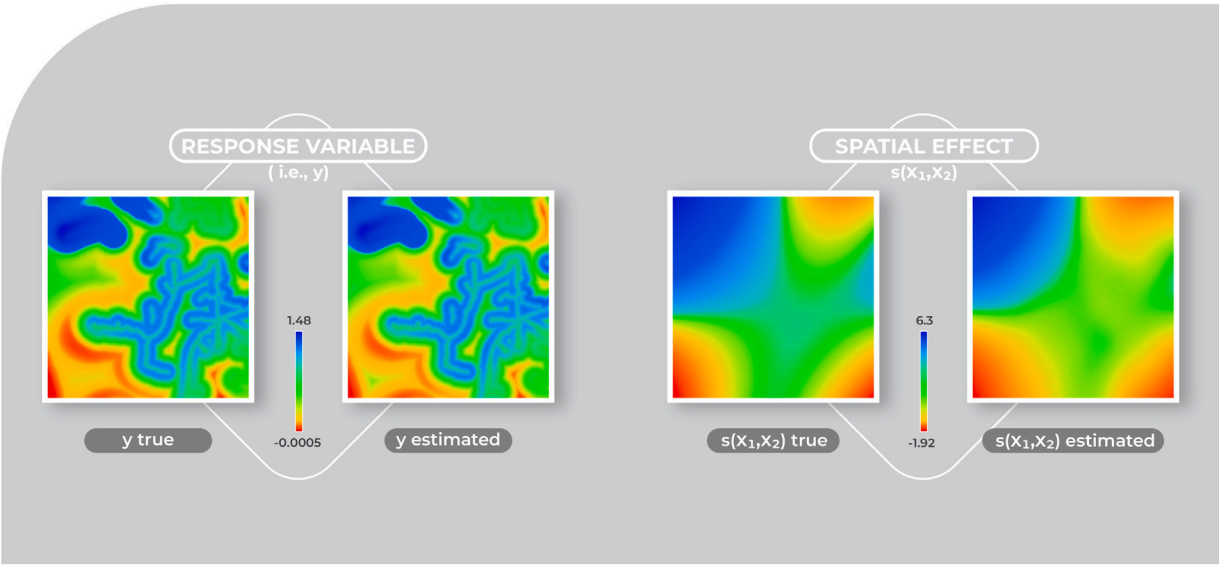

Fig. 4. Results from the simulation study. The left group shows the response variable (i.e., y): true and estimated (left and right, ranging from red: −0.0005 to blue: 1.48); the right group shows $s(x_1, x_2)$: true and estimated (left and center, ranging from red: −1.92 to blue: 6.3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

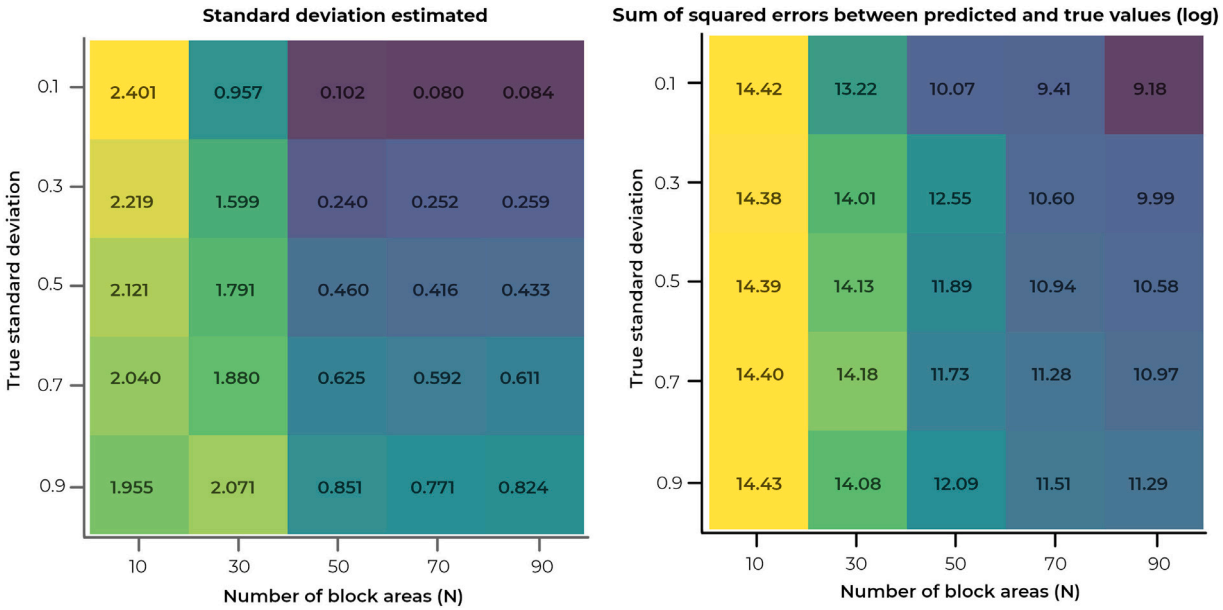

Fig. 5. Results from the sensitivity analysis. The left map shows the effect of true standard deviation and sample size on the standard deviation estimates, while the right map shows the effect on the sum of squared errors between predictions and true values.

original component. For the spatial effect, the middle region is slightly different, which can be explained by the discontinuities of the original function in this region. In general, the results indicate a successful approximation of the true functions by the model.

The sensitivity analysis evaluated the method's ability to reconstruct the original from the aggregated data under different settings. Fig. 5 shows the variance estimated (left) and the sum of squared errors (right). For the former quantity, the results show the considerable influence of the number of block areas, suggesting that having a sufficiently large sample size is crucial for a proper application of the model. A quick degradation of method abilities for this particular case is observed for N < 50 block-areas. In this situation, the study is likely to be underpowered, and the method does not have enough information to distinguish between the systematic and random patterns of the data. The results indicate a joint influence of the true standard deviation and the sample size for the latter. In all calculations where the method has enough power (i.e., N ≥ 50), increasing the standard deviation in the data simulation procedure also increases the total error of model predictions compared to the true values. These results indicate that heterogeneous datasets may demand more block-areas for its estimation, and suggest that the threshold of 50 block-areas cannot be generalized for all applications, as it tends to vary according to problem-specific characteristics such as the variability of the phenomenon modeled, the spatial configuration of block-areas, among others.

3.2. Real example

For the real example, the smoothers obtained are shown in Fig. 6. In the qualitative evaluation of the results, most of the curves were

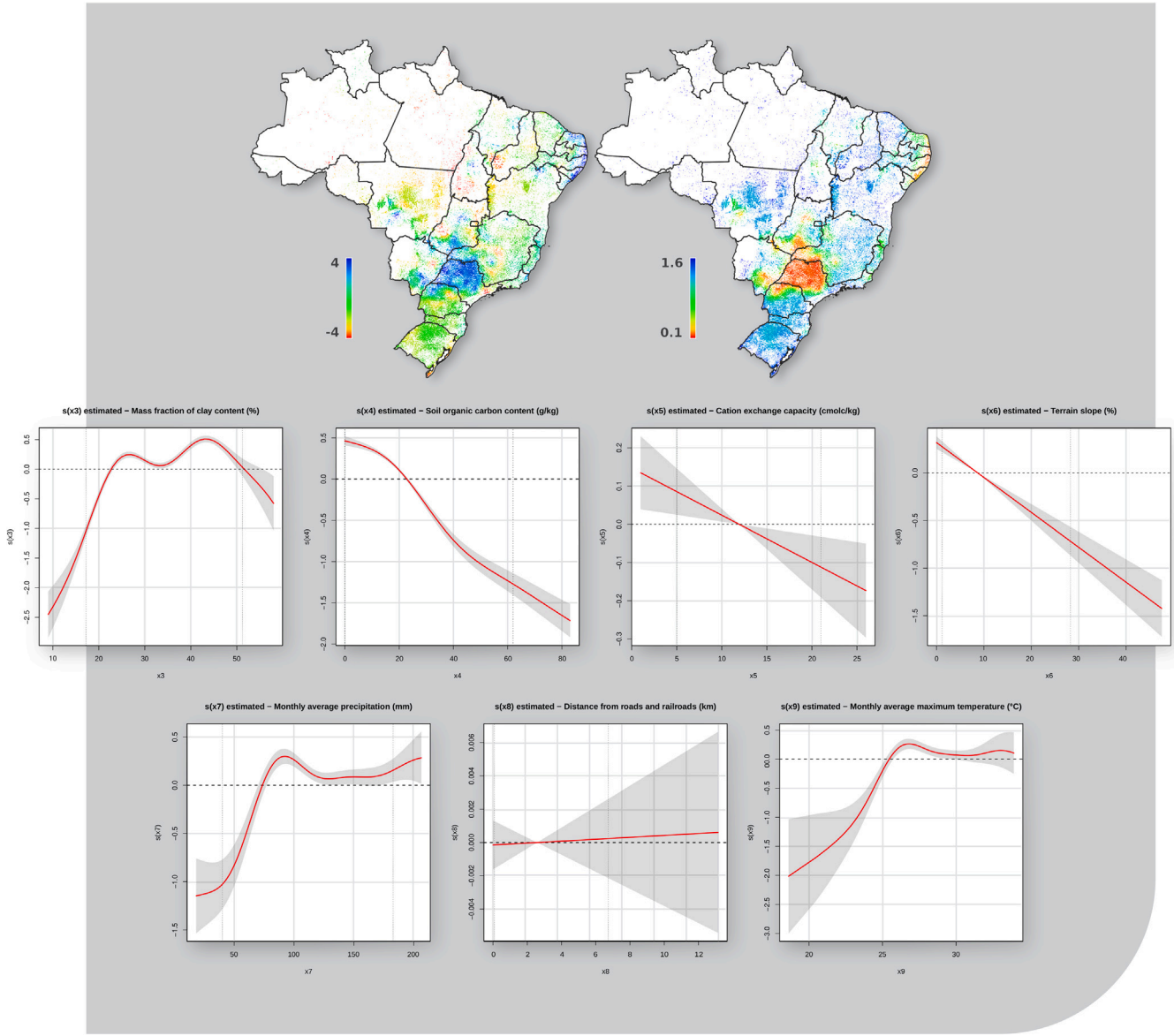

Fig. 6. Results from the actual application. In the top row, the left map shows the estimated spatial effect with the corresponding standard deviations in the right. Both maps are in a red–green–blue gradient (lower to upper values). The frames in the two bottom rows show the estimated smoothers (red) with the corresponding standard deviations (gray). The horizontal dashed line crossing zero allows the interpretation that regions above it contribute to the increase of agricultural production, while regions below it have the opposite effect. Two vertical dotted lines indicate the 5% and 95% percentiles of the data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

found to be coherent with the agronomic literature. For soil conditions, we expected an ideal clay content to be neither too small (i.e., sandy) so that infiltrated water is not available to plants, nor too large (i.e., clayey) so the roots have a low growth rate. While sandy soils have low water retention that impedes root development, overly clayey soils may show salt accumulation and poor aeration, which implies low oxygenation when humid or high physical resistance to root growth when dry (Lal and Shukla, 2004). We also expected the cation exchange capacity to be positively correlated to sugarcane productivity, as its increase represents a greater ability of the soil to retain more nutrients to nourish the plants better (De Boodt, 1972). For the soil organic carbon, we expected a decreasing relationship. The conversion from forest to cropland is known to lead to the depletion of the soil organic carbon pool, and the effect can be worse depending on the management practices adopted. Since different mechanisms such as oxidation, mineralization and erosion can reduce the soil organic carbon with time (Clark et al., 2017; Lal, 2002), lower concentrations may coincide with places where the agricultural activity is older and more consolidated. The estimated smoothers agree with this prior knowledge: the clay smoother, $s(\boldsymbol{x}_3)$, is actually an inverted U-shape with an ideal mass fraction of clay content around 22 to 51%, and the soil organic carbon smoother, $s(\boldsymbol{x}_4)$, contains a curvy decreasing pattern. While our model extracted an inverse relationship for the cation exchange capacity, it was a relatively weak (i.e., with a small magnitude) and highly uncertain effect when compared to the others.

For precipitation, we expected its curve to resemble an inverted parabola. If low precipitation leads to low availability of water and nutrients for the plants, its excess could also harm production as cloudy skies tend to reduce the incidence of photosynthetically active radiation. For the temperature, we expected the curve to indicate that warmer places favor higher sugarcane productions. The estimated smoothers seem to agree with this interpretation. Despite the uncertainty in the extremes, the patterns of Fig. 6 indicate a positive effect of rainfall on sugarcane production, $s(\boldsymbol{x}_7)$, for values above

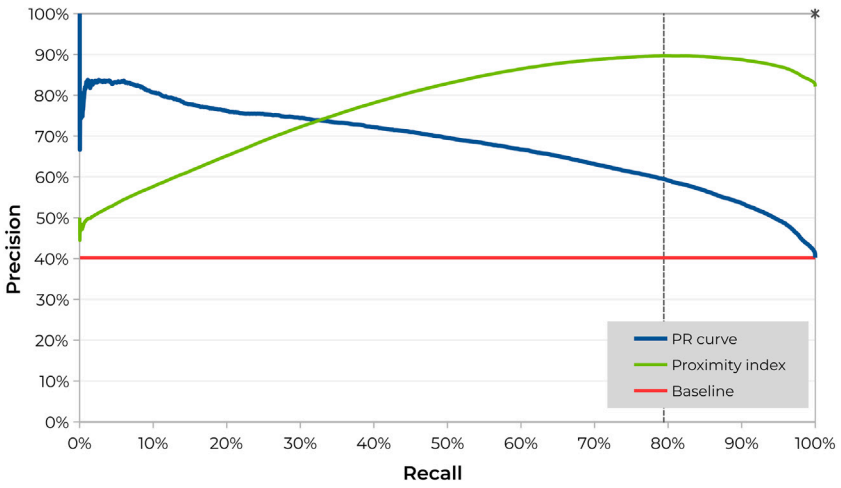

Fig. 7. Precision–Recall (PR) curve for the actual application model when interpreted as a classifier for the existence of sugarcane. The asterisk is the optimal point of Precision = Recall = 100%.

70 mm/month, and with a peak ending at around 125 mm/month. Such a range overlaps with the amount described by the Food and Agriculture Organization (Food and Agriculture Organization, 2021) of 125 to 208 mm/month. For the temperature, the curve $s(\boldsymbol{x}_9)$ points to increasing production for values above 26 °C, with larger uncertainty above 30 °C. These values are slightly higher than the optimum growth interval defined by the Food and Agriculture Organization as between 22 and 30 °C (Food and Agriculture Organization, 2021). This is expected since we used the 'maximum monthly temperature' variable, which is greater than the 'average monthly temperature' used to define the optimum growth interval. Our interval is also very close to the optimum range between 25° and 33 °C described by dos Santos Almeida et al. (2008) for sugarcane in Brazil.

The slope variable also matched with the prior knowledge. We expected it to be negatively correlated with sugarcane production since steep terrains hinder the mechanization of agricultural areas. However, the estimated curve $s(\boldsymbol{x}_6)$ showed that sugarcane tends to be more productive in flat cropland areas, with a slope lower than 9.5%. The only variable with an unexpected pattern was the distance from transportation infrastructures. We expected it to be negatively correlated with sugarcane production because distant areas can be argued to have higher costs and greater losses. However, its curve, $s(\boldsymbol{x}_8)$, tended to zero during estimation, indicating very little influence of this variable. As the model used only the information for cropland areas during the fit, one possible explanation is that the distance variable range (from 0 to 14 km) is already narrow enough to exclude isolated areas and places where the transportation costs are prohibitive.

We assessed the model's ability to correctly allocate sugarcane on the cropland pixels in the quantitative evaluation of predictions. Fig. 7 shows the PR curve (blue) for the model when interpreted as a classifier, along with a baseline (red) representing the performance of random allocation, that is, $h = 40.20\%$ as described in Section 2.4.1. Ideally, we would like our model to maximize both quantities simultaneously so that the optimal point would be located at the upper right corner of the plane, represented with an asterisk in Fig. 7. The green curve represents a normalized measure of proximity (i.e., one minus the Euclidean distance) between points on the PR curve and the asterisk, and the black dashed vertical curve indicates its maximum. Using the threshold indicated by the dashed curve, the chance of observing sugarcane in a pixel where the model predicts sugarcane (i.e., precision or user accuracy) equals 59.59%, the chance of having the model predicting sugarcane where sugarcane is indeed present (i.e., recall or producer accuracy) equals 79.18%, and the overall accuracy is 70.05%.

The values found are, in general, lower than those found in remote sensing papers for the same crop and in the same country. For example, dos Santos Luciano et al. (2019) built Random Forest models for eight sites of 389,000 ha each in the São Paulo state, Southeastern Brazil, and found overall accuracies ranging from 82% to 95%. Vieira et al. (2012) classified sugarcane presence in three Brazilian municipalities also from the Southeastern region and found a global accuracy of 93.99%. (Zheng et al., 2021) detected sugarcane in the 14 Brazilian states that produce 98% of the sugarcane in the country, with the results per state varying from 80.70% to 93.10% (producer accuracy), 85.70% to 100.0% (user accuracy), and 88.60% to 95.68% (overall accuracy). Several reasons could explain the reason for such difference. On the one hand, the example presented here could be expected to perform worse for: (i) having the prediction of yield as its primary goal, with the presence–absence prediction being a subproduct; (ii) calculating the accuracy from another existing database, not from points collected on the field or by visual interpretation; (iii) not using any post-processing techniques as usual in classification maps (Simoes et al., 2020). On the other hand, it could be expected the current application to perform better for (i) predicting over an agricultural mask previously generated and validated, and not the whole territory (i.e., MapBiomas (2019a)); (ii) using national statistics as an input data for calculations and not as for the validation of the results obtained.

However, the central point is that the results could be improved while still under the method proposed. One significant remark is that the model and the inputs used in our example are very simplistic for the phenomenon of sugarcane production, but the same proposed linear structure of Eq. (1) allows for greater complexity in real applications. For sugarcane predictions and other environmental applications, a particularly relevant possibility is the so-called functional data analysis, where the set of explanatory variables can be made up not only by point observations as in our example but of continuous functions (Ramsay and Silverman, 2005; Kokoszka and Reimherr, 2017). In this case, it would be possible, for example, to replace the average monthly temperature with each pixel's annual history of daily temperatures, their Normalized Difference Vegetation Index (NDVI) curve, which is widely used for being correlated with crop yield (Moriondo et al., 2007; Wall et al., 2008), or any other relevant time series. Formulating an actual application with these settings would have better results and greater physical realism.

Fig. 8 shows the predicted production for sugarcane in Brazil. In this map, the background was darkened to improve contrast, and pixels from municipalities with a zero sugarcane production were removed, even though their information was used during the model fit. Two frames were added on the right to show zooms over different municipalities with intra-municipality variations. When fitting the model, we divided the original input by 10^4 to avoid numerical problems. In this

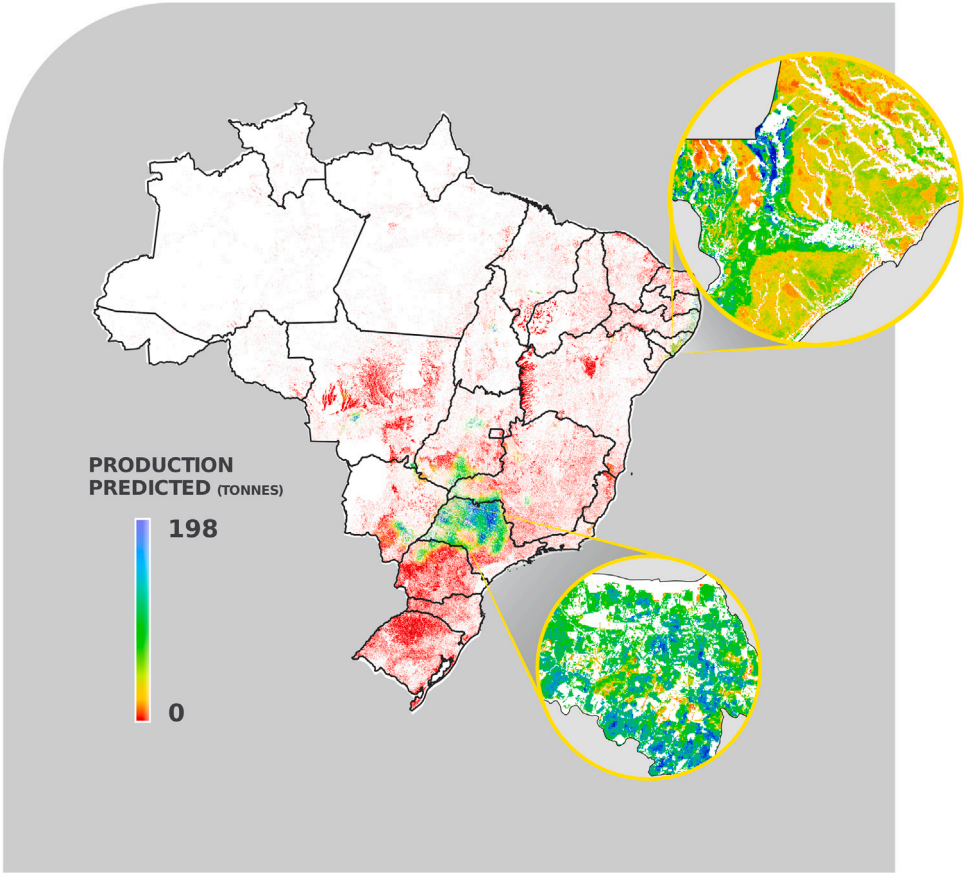

Fig. 8. Predictions at the disaggregated level for the actual application. Values are presented in a red–green–blue gradient, ranging from red: 0 tonnes to blue: 198 tonnes [maximum value]. The two frames in the right show zooms in two different municipalities, and illustrate the variation of predictions within each block-area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

scale, which converts 150 t/ha to 0.015 $\cdot(10^4$t/ha) for example, the standard deviation was estimated as $\sigma = 0.077$. Such a large magnitude is unsurprising since the process of predicting about a much finer scale from aggregated information is unavoidably uncertain by nature. Adds to that the fact that the model considered is relatively simple, with the spatial effect being only a proxy for more complex effects, such as technologies used, agricultural practices adopted on the field, and other biophysical variables not included explicitly. Such effects may not appear much on the secondary pixel-level information adopted in our example, and primary information such as satellite-gathered data are likely the only indirect source possible for such variability. Therefore, another possible solution to reduce uncertainties beyond functional data analysis is to explore the incorporation of satellite information as predictors, which could enrich the inputs in real uses of the example presented here. Similar idea have already been implemented in small scale studies by Mello et al. (2014), Khan et al. (2010) and Walker and Mallawaarachchi (1998). Another more guaranteed approach could be the relaxing the constant variance assumption of Eq. (1) by explicitly modeling the variation as a function of the covariates, as in Stasinopoulos et al. (2017). This has the disadvantage of demanding changes in the estimation method.

Since the pixel adopted in this application has a 100 m spatial resolution, whose area corresponds to 1 hectare (ha), the results obtained can be compared directly to other annual productivity values previously reported in the literature. Our predicted values were lower than 198 tonnes (t), which can be compared to the work of Waclawovsky et al. (2010). The authors reported the maximum commercial yield of sugarcane in Australia, Colombia, and South Africa as 148 t/ha.year, and 240 t/ha.year in Brazil. Also, in Southeastern Brazil, Pinheiro et al. (2010) collected data over 14 years and found a productivity that ranged from 50 to 150 t/ha. Such reference values indicate that our predictions are within the range of actual values. One positive point of the model adopted is that it naturally makes unsuitable pixels tend to zero, which suggests additional post-processing steps could be added.

The match between the results obtained and the literature is a significant indication that the method converged successfully. At this point, it is possible to see that the general modeling approach adopted presents several characteristics that can contribute to other methods previously published. For example, the statistical model can be adapted to different situations and disciplines without much additional effort, thus overcoming the difficulty that arises when adapting top-down models created for specific application or regions of the world, e.g., Szyniszewska (2020). Another example is that the approach did not demand explicit upper bounds for the unobserved variables. The exponential link function adopted is completely unbounded, and the values predicted resulted from a balance between data fit and the wiggliness of smoothers during estimation. Besides, for the estimation method specifically, the objective function minimizes a quadratic error between aggregated predicted values and the data observed instead of forcing the equality as in Malone et al. (2012, 2017). This decision brings extra reliability to the method in the sense that it does not assume that the observed data is perfect but may contain measurement errors. The application for the Brazilian agricultural production data illustrates the importance of this assumption since the database is made of educated guesses declared by several independent agents around the country (IBGE, 2020).

The potential of scaling calculations becomes more evident when the results are compared to other works previously published about

the same problem and in the same study area. The most prominent work is that of You and Wood (2006), who disaggregated maize, wheat, rice, sorghum, potato, cassava, bean, and soybean production to a 9 km grid in Brazil. Our methodological proposal opens up the possibility to far exceed the resolution adopted by the authors, thus increasing our ability to input finer datasets and produce more detailed outputs. Our results presented have a nearly 10^2 times higher resolution, which converts to nearly 10^4 times more pixels used for the disaggregation. Such a gain in the resolution of simulated crop datasets can potentially enrich several applications, including the identification of hotspots of land-use change (Kuemmerle et al., 2016), the projection of future food security scenarios (Nelson et al., 2010), the assessment of nutrient flow (Liu et al., 2010), spatial zoning, and other planning applications. The attempt to quantify how much the production of a single crop comes from illegal deforestation practices by Rajão et al. (2020) also highlights how much applications and studies could benefit from a high resolution simulated crop dataset.

Beyond the example provided for an agricultural application, other areas that can benefit from the proposed method include epidemiology (Utazi et al., 2019), demography (Wardrop et al., 2018), climatology (Görner et al., 2020), ecology (Tamis and Van 't Zelfde, 1998), economics (Monteiro et al., 2018), remote sensing (Zhan et al., 2013), land system science (Fendrich et al., 2020), and others. Furthermore, the method can also be applied in cases where the disaggregated-level areas are not continuous but discrete. As long as the disaggregated-level areas can be assigned to a single block-level area, Bastiani et al. (2018) show that the same structure of Eq. (1) can accommodate an intrinsic autoregressive model for areal data. This fact increases even more the possible applications of the disaggregation method, especially for economic data where many phenomena follow political boundaries.

4. Conclusions

In this work, we presented a method to estimate the parameters of a known disaggregation model from the aggregated data in parallel. The technique is computationally efficient and opens up the possibility of working with high-resolution data and large geographical areas, situations that have not been addressed in the literature so far. By exemplifying with a very general model for the data at the disaggregated level, our method is suitable to be readily transferred to other fields of research, thus providing another option researchers and practitioners.

Our results for a simulation study indicate that the method could successfully predict 1 million pixels using only the information of 90 block areas for the fit. The success was not affected by an uncorrelated variable purposely inserted, which indicates some robustness against misspecification. The sensitivity analysis shows that model performance tends to degrade as the variance of the true phenomenon at the finer scale increases and the number of block-areas decreases. In our particular example, there seemed to exist a threshold of $N = 50$ block-areas above which the method's performance was satisfactory as judged by its ability to recover the true standard deviation parameter and regenerate the true functions.

For the real example of sugarcane production in Brazil, the values predicted were mostly within the range of actual values found in the literature, and the smoothers obtained were generally coherent with previous knowledge. When considered as a classifier and validated against known locations of sugarcane in Brazil, the model had user, producer, and overall accuracies of 59.59%, 79.18%, and 70.05%, respectively. While these values are not as high as specific remote-sensing studies for the same phenomenon in the same study area, this can be reasonably explained by the quite simplistic formulation adopted as an example of the method's functioning. In actual applications, our estimation method can be used for models with more detailed formulations and datasets, such as the usage of time-series of climate or satellite-derived variables for explaining sugarcane production. Compared to the most successful previous work published that disaggregated crop information in Brazil, the scalability brought by our work allowed us to work on a nearly 100 times higher resolution, thus with nearly 10,000 times more pixels. Such a gain in resolution makes the results more adequate for use in practical applications and studies.

Despite the advances, further work can still improve the method. Incorporating the ability to model the variance parameter as a function of one or more covariates can improve predictions' uncertainty. Such gain comes at the cost of changing the derived objective function and consequently, the estimation procedures. While methods for this problem exist in the literature, the possibility of scaling their calculations in a disaggregation setting still needs further investigation. Also, developing measures to evaluate model fit is necessary to assess all the assumptions made.

Computer code availability

Name: "Disaggregation".

Developer: Arthur Nicolaus Fendrich. University of São Paulo, Luiz de Queiroz College of Agriculture. Corresponding address: Avenida Pádua Dias, 11. Postal Code 13418-900, Piracicaba, SP, Brazil. E-mail address: arthfen@gmail.com. Telephone: +55 19 3417 2100.

Year first available: 2021

Hardware required: 2 GB RAM, 100MB storage.

Software required: All codes were written in R, version 3.5.1 (R. Core Team, 2018). The implementation of the smoothers was taken from the `mgcv` (Wood, 2017). Code parallelization was achieved with the `Rmpi` (Yu, 2002), `snow` (Tierney et al., 2016), and `parallel` (R. Core Team, 2018) packages, and the manipulation of spatial data was performed using the `raster` (Hijmans, 2017) package.

Program language: R, version 3.5.1.

Program size: 7.7 MB.

Details on how to access the source code: All source codes used for the simulated example are available at: https://github.com/arthfen/disaggregation.

CRediT authorship contribution statement

Arthur Nicolaus Fendrich: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Elias Salomão Helou Neto:** Conceptualization, Methodology, Formal Analysis, Investigation, Writing – original draft, Supervision. **Lucas Esperancini Moreira e Moreira:** Methodology, Software, Validation, Investigation, Writing – original draft. **Durval Dourado Neto:** Methodology, Validation, Investigation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The research was carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0). The authors thank Dr. Júlio Hoffimann for the valuable comments prior to the submission, and Rafael Corrêa Leme for editing the images.

Appendix

In the present appendix we prove convergence of our proposed algorithm. Because we linearize the likelihood of $\boldsymbol{Y}$ with relation to $\boldsymbol{\beta}$ at each iteration and we add a quadratic penalty term to the resulting objective function (for stabilization purposes), our method resembles the Levenberg–Marquardt iteration but it is not really the same algorithm. In our method, the optimization of each problem is performed over the (σ^2, λ) variables and $\boldsymbol{\beta}$ is actually computed as a function $\boldsymbol{\beta}(\sigma^2, \lambda)$, which may be nonlinear on (σ^2, λ). Therefore, a mathematical convergence analysis is necessary in order to provide a characterization of the outcome of the method.

We start our analysis by considering an abstract method for optimization and showing that it has satisfactory convergence properties. Afterwards, we consider how the algorithm used in the paper fits the mathematical development.

Proposition 1. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be Lipschitz-continuous and differentiable. Consider the algorithm*

$$\zeta^{[\eta+1]} = \underset{\zeta}{\operatorname{argmin}} \quad f(\zeta) + h^{[\eta]}(\zeta, \zeta^{[\eta]}), \tag{13}$$

where $h^{[\eta]} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ are such that

$$|h^{[\eta]}(\zeta, \zeta^{[\eta]})| \leq E\|\zeta - \zeta^{[\eta]}\|^2$$

for some large constant $E \geq 0$. Then, if $\|\zeta^{[\eta+1]} - \zeta^{[\eta]}\| \to 0$, every limit point of the sequence $\{\zeta^{[\eta]}\}$ is a critical point of f.

Proof. Notice that the hypothesis lead to

$$f(\zeta) - C\|\zeta - \zeta^{[\eta]}\|^2 \leq f(\zeta) + h^{[\eta]}(\zeta, \zeta^{[\eta]}) \leq f(\zeta) + C\|\zeta - \zeta^{[\eta]}\|^2 \tag{14}$$

for $C = LE$, where L is a Lipschitz constant for f.

Now, suppose that $\|\zeta^{[\eta+1]} - \zeta^{[\eta]}\| \to 0$, and assume that there is an accumulation point $\overline{\zeta}$ of $\{\zeta^{[\eta]}\}$ that is not a critical point of f and denote $\{\zeta^{[\ell_\eta]}\}$ a subsequence converging to $\overline{\zeta}$.

It is a well known fact in optimization that given $\overline{\zeta}$ with $\nabla f(\overline{\zeta}) \neq \mathbf{0}$ and fixed $\alpha \in (0, 1)$, then there exists $\Phi > 0$ such that for $\phi \in [0, \Phi]$ we have

$$f\left(\overline{\zeta} - \phi\nabla f(\overline{\zeta})\right) \leq f(\overline{\zeta}) - \phi\alpha\|\nabla f(\overline{\zeta})\|^2.$$

By denoting $\zeta(\phi) := \overline{\zeta} - \phi\nabla f(\overline{\zeta})$ and $K := \alpha\|\nabla f(\overline{\zeta})\|^2$ the above can be more simply stated as

$$f\left(\zeta(\phi)\right) \leq f(\overline{\zeta}) - \phi K. \tag{15}$$

Furthermore, notice that

$$\|\zeta(\phi) - \overline{\zeta}\| \leq \phi M, \quad \text{where} \quad M = \|\nabla f(\overline{\zeta})\|. \tag{16}$$

We then use the second inequality in (14) with $\zeta = \zeta(\phi)$ and η replaced by $\ell_\eta - 1$ to get

$$f\left(\zeta(\phi)\right) + h^{[\ell_\eta - 1]}\left(\zeta(\phi), \zeta^{[\ell_\eta - 1]}\right) \leq f\left(\zeta(\phi)\right) + C\|\zeta(\phi) - \zeta^{[\ell_\eta - 1]}\|^2.$$

Application of (15) assuming a fixed $\phi \in (0, \Phi)$ now gives

$$\begin{aligned} f\left(\zeta(\phi)\right) + h^{[\ell_\eta - 1]}\left(\zeta(\phi), \zeta^{[\ell_\eta - 1]}\right) &\leq f(\overline{\zeta}) + C\|\zeta(\phi) - \zeta^{[\ell_\eta - 1]}\|^2 - \phi K \\ &= f(\overline{\zeta}) + C \parallel \zeta(\phi) - \overline{\zeta} + \overline{\zeta} - \zeta^{[\ell_\eta]} + \zeta^{[\ell_\eta]} \\ &\quad - \zeta^{[\ell_\eta - 1]} \parallel^2 - \phi K \\ &\leq f(\overline{\zeta}) + C(\|\zeta(\phi) - \overline{\zeta}\| \\ &\quad + \|\overline{\zeta} - \zeta^{[\ell_\eta]}\| \\ &\quad + \|\zeta^{[\ell_\eta]} - \zeta^{[\ell_\eta - 1]}\|)^2 - \phi K. \end{aligned}$$

Taking (16) into consideration we get

$$\begin{aligned} &f\left(\zeta(\phi)\right) + h^{[\ell_\eta - 1]}\left(\zeta(\phi), \zeta^{[\ell_\eta - 1]}\right) \\ &\leq f(\overline{\zeta}) + C(\phi M + \|\overline{\zeta} - \zeta^{[\ell_\eta]}\| + \|\zeta^{[\ell_\eta]} - \zeta^{[\ell_\eta - 1]}\|)^2 - \phi K. \end{aligned}$$

Because the above arguments hold for every $\phi \in (0, \Phi)$ we can assume that $\phi < K/(9CM^2)$ and because $\zeta^{[\ell_\eta]} \to \overline{\zeta}$ and $\|\zeta^{[\eta]} - \zeta^{[\eta+1]}\| \to 0$ we can assume that η is large enough so that $\|\overline{\zeta} - \zeta^{[\ell_\eta]}\| \leq \phi M$ and $\|\zeta^{[\ell_\eta]} - \zeta^{[\ell_\eta - 1]}\| \leq \phi M$. Therefore

$$\begin{aligned} f\left(\zeta(\phi)\right) + h^{[\ell_\eta - 1]}\left(\zeta(\phi), \zeta^{[\ell_\eta - 1]}\right) &\leq f(\overline{\zeta}) + \phi^2 9CM^2 - \phi K \\ &= f(\overline{\zeta}) + \phi(\phi 9CM^2 - K) \\ &\leq f(\overline{\zeta}) - \phi\gamma, \end{aligned} \tag{17}$$

where $\gamma = K - \phi 9CM^2 > 0$. On the other hand, using the first inequality in (14) with $\zeta = \zeta^{[\ell_\eta]}$ and $\overline{\zeta} = \zeta^{[\ell_\eta - 1]}$ we have

$$\begin{aligned} f(\zeta^{[\ell_\eta]}) + h^{[\ell_\eta - 1]}(\zeta^{[\ell_\eta]}, \zeta^{[\ell_\eta - 1]}) &\geq f(\zeta^{[\ell_\eta]}) - C\|\zeta^{[\ell_\eta]} - \zeta^{[\ell_\eta - 1]}\|^2 \\ &= f(\overline{\zeta}) - f(\overline{\zeta}) + f(\zeta^{[\ell_\eta]}) \\ &\quad - C\|\zeta^{[\ell_\eta]} - \zeta^{[\ell_\eta - 1]}\|^2. \end{aligned}$$

Recalling that f is continuous, then, again because $\zeta^{[\ell_\eta]} \to \overline{\zeta}$ and $\|\zeta^{[\eta]} - \zeta^{[\eta+1]}\| \to 0$ we can assume that η is large enough so that $|f(\overline{\zeta}) - f(\zeta^{[\ell_\eta]})| \leq \phi\gamma/8$ and $\|\zeta^{[\ell_\eta]} - \zeta^{[\ell_\eta - 1]}\| \leq \phi/(8C)$. Therefore

$$f(\zeta^{[\ell_\eta]}) + h^{[\ell_\eta - 1]}(\zeta^{[\ell_\eta]}, \zeta^{[\ell_\eta - 1]}) \geq f(\overline{\zeta}) - \frac{\phi\gamma}{4}. \tag{18}$$

Comparing (17) with (18) we see that

$$f\left(\zeta(\phi)\right) + h^{[\ell_\eta - 1]}\left(\zeta(\phi), \zeta^{[\ell_\eta - 1]}\right) \leq f(\zeta^{[\ell_\eta]}) + h^{[\ell_\eta - 1]}(\zeta^{[\ell_\eta]}, \zeta^{[\ell_\eta - 1]}),$$

which contradicts (13). □

In order to understand how our proposed method fits the above convergence analysis, we will start by observing that at iteration η we linearize the likelihood of $\boldsymbol{Y}$ in a way that

$$\ell^{[\eta]}(\zeta) = \ell(\zeta) + O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^{[\eta]}\|^2).$$

Let us simplify the notation by using $\zeta = (\sigma^2, \lambda)$ and we will assume that the sequence of functions $\boldsymbol{\beta}^{[\eta]}(\zeta)$ given by

$$\begin{aligned} \boldsymbol{\beta}^{[\eta+1]}(\zeta) = &\left(\mathrm{X}^\top \bar{\mathrm{G}}^{[\eta]\top} \mathrm{A}^\top (\mathrm{A}\mathrm{A}^\top \sigma^2)^{-1} \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X} + \Omega_\lambda^{-1} + \rho^{[\eta]}\mathbf{I}\right)^{-1} \\ &\times \left[\mathrm{X}^\top \bar{\mathrm{G}}^{[\eta]\top} \mathrm{A}^\top \left(\mathrm{A}\mathrm{A}^\top \sigma^2\right)^{-1} \left(\boldsymbol{Y} - \mathrm{A}g\left(\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \mathrm{A}\bar{\mathrm{G}}^{[\eta]}\mathrm{X}\bar{\boldsymbol{\beta}}^{[\eta]}\right) + \rho^{[\eta]}\bar{\boldsymbol{\beta}}^{[\eta]}\right] \end{aligned}$$

satisfies

$$\|\boldsymbol{\beta}^{[\eta]}(\zeta) - \boldsymbol{\beta}^{[\eta]}(\boldsymbol{\omega})\| \leq L\|\zeta - \boldsymbol{\omega}\|$$

for some large enough constant $L \geq 0$. In such a case, we have

$$\ell^{[\eta]}(\zeta) = \ell(\zeta) + O(\|\zeta - \zeta^{[\eta]}\|^2)$$

and, therefore, because the quadratic penalty term is also $O(\|\zeta - \zeta^{[\eta]}\|^2)$, our proposed algorithm is an instance of the general method analyzed above with $f = \ell$ and $h^{[\eta]} = \ell^{[\eta]} - \ell$.

References

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Sci. Data 5, 170191. http://dx.doi.org/10.1038/sdata.2017.191.

Armstrong, M.P., Densham, P.J., 1992. Domain decomposition for parallel processing of spatial problems. Comput. Environ. Urban Syst. 16 (6), 497–513. http://dx.doi.org/10.1016/0198-9715(92)90041-o.

Armstrong, M.P., Rushton, G., Zimmerman, D.L., 1999. Geographically masking health data to preserve confidentiality. Stat. Med. 18 (5), 497–525.

Barwell, L.J., Azaele, S., Kunin, W.E., Isaac, N.J.B., 2014. Can coarse-grain patterns in insect atlas data predict local occupancy? Divers. Distrib. 20 (8), 895–907. http://dx.doi.org/10.1111/ddi.12203.

Bastiani, F.D., Rigby, R.A., Stasinopoulous, D.M., Cysneiros, A.H., Uribe-Opazo, M.A., 2018. Gaussian Markov random field spatial models in GAMLSS. J. Appl. Stat. 45 (1), 168–186. http://dx.doi.org/10.1080/02664763.2016.1269728.

Buchin, K., Buchin, M., van Kreveld, M., Löffler, M., Luo, J., Silveira, R.I., 2012. Processing aggregated data: The location of clusters in health data. GeoInformatica 16 (3), 497–521.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al., 2015. Global land cover mapping at 30 m resolution: A POK-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27.

Clark, K.M., Boardman, D.L., Staples, J.S., Easterby, S., Reinbott, T.M., Kremer, R.J., Kitchen, N.R., Veum, K.S., 2017. Crop yield and soil organic carbon in conventional and no-till organic systems on a claypan soil. Agron. J. 109 (2), 588–599. http://dx.doi.org/10.2134/agronj2016.06.0367, URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.2134/agronj2016.06.0367.

Comber, A., Zeng, W., 2019. Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. Geogr. Compass 13 (10), http://dx.doi.org/10.1111/gec3.12465.

De Boodt, M., 1972. Improvement of soil structure by chemical means. In: HILLEL, D. (Ed.), Optimizing the Soil Physical Environment Toward Greater Crop Yields. Academic Press, pp. 43–55. http://dx.doi.org/10.1016/B978-0-12-348540-3.50012-6, URL http://www.sciencedirect.com/science/article/pii/B9780123485403500126.

Demidenko, E., 2013. Mixed Models : Theory and Applications with R. Wiley, Hoboken.

DNIT, 2018. VGEO. Online: http://servicos.dnit.gov.br/vgeo. (Accessed 12 July 2018).

do Nascimento Bendini, H., Fonseca, L.M.G., Schwieder, M., Körting, T.S., Rufin, P., Sanches, I.D.A., Leitão, P.J., Hostert, P., 2019. Detailed agricultural land classification in the Brazilian cerrado based on phenological information from dense satellite image time series. Int. J. Appl. Earth Obs. Geoinf. 82, 101872. http://dx.doi.org/10.1016/j.jag.2019.05.005.

dos Santos Almeida, A.C., Souza, J.L., Teodoro, I., Barbosa, G.V.S., Filho, G.M., Júnior, R.A.F., 2008. Desenvolvimento vegetativo e produção de variedades de cana-de-açúcar em relação à disponibilidade hídrica e unidades térmicas. Ciência E Agrotecnol. 32 (5), 1441–1448. http://dx.doi.org/10.1590/s1413-70542008000500013.

dos Santos Luciano, A.C., Picoli, M.C.A., Rocha, J.V., Duft, D.G., Lamparelli, R.A.C., Leal, M.R.L.V., Maire, G.L., 2019. A generalized space-time OBIA classification scheme to map sugarcane areas at regional scale, using landsat images time-series and the random forest algorithm. Int. J. Appl. Earth Obs. Geoinf. 80, 127–136. http://dx.doi.org/10.1016/j.jag.2019.04.013.

Fendrich, A.N., Barretto, A., de Faria, V.G., de Bastiani, F., Tenneson, K., Guedes Pinto, L.F., Sparovek, G., 2020. Disclosing contrasting scenarios for future land cover in Brazil: Results from a high-resolution spatiotemporal model. Sci. Total Environ. 742, 140477. http://dx.doi.org/10.1016/j.scitotenv.2020.140477, URL http://www.sciencedirect.com/science/article/pii/S0048969720339991.

Food and Agriculture Organization, 2020. FAOSTAT - Crops. Online: http://www.fao.org/faostat/en/#data/QC. (Accessed 05 April 2020).

Food and Agriculture Organization, 2021. Sugarcane - land & water. Online; http://www.fao.org/land-water/databases-and-software/crop-information/sugarcane/en/. (Accessed 07 January 2021).

Friedman, J.N., Holden, R.T., 2008. Optimal gerrymandering: Sometimes pack, but never crack. Amer. Econ. Rev. 98 (1), 113–144.

Golden, R., 2020. Statistical Machine Learning : A Unified Framework. CRC Press, Boca Raton, FL.

Goovaerts, P., 2006. Geostatistical analysis of disease data: Accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. Int. J. Health Geogr. 5 (1), 52. http://dx.doi.org/10.1186/1476-072x-5-52.

Görner, C., Franke, J., Kronenberg, R., Hellmuth, O., Bernhofer, C., 2020. Multivariate non-parametric euclidean distance model for hourly disaggregation of daily climate data. Theor. Appl. Climatol. http://dx.doi.org/10.1007/s00704-020-03426-7.

Hawley, K., Moellering, H., 2005. A comparative analysis of areal interpolation methods. Cartogr. Geogr. Inf. Sci. 32 (4), 411–423.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: GLobal gridded soil information based on machine learning. PLoS One 12 (2), http://dx.doi.org/10.1371/journal.pone.0169748.

Hijmans, R.J., 2017. Raster: Geographic data analysis and modeling. R package version 2.6-7, URL https://CRAN.R-project.org/package=raster.

Hulley, G.C., Hook, S.J., Abbott, E., Malakar, N., Islam, T., Abrams, M., 2015. The ASTER global emissivity dataset (ASTER GED): Mapping earth's emissivity at 100 meter spatial scale. Geophys. Res. Lett. 42 (19), 7966–7976.

IBGE, 2018. Bases cartográficas contínuas. Online; https://www.ibge.gov.br/geociencias/cartas-e-mapas/bases-cartograficas-continuas/15759-brasil.html?edicao=16034&t=downloads. (Accessed 08 June 2018).

IBGE, 2020. Produção agrícola municipal - PAM - Conceitos e métodos. Online; https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9117-producao-agricola-mtunicipal-culturas-temporarias-e-permanentes.html?=&t=conceitos-e-metodos. (Accessed 14 October 2020).

Keil, P., Belmaker, J., Wilson, A.M., Unitt, P., Jetz, W., 2012. Downscaling of species distribution models: A hierarchical approach. Methods Ecol. Evol. 4 (1), 82–94. http://dx.doi.org/10.1111/j.2041-210x.2012.00264.x.

Khan, M., de Bie, C., van Keulen, H., Smaling, E., Real, R., 2010. Disaggregating and mapping crop statistics using hypertemporal remote sensing. Int. J. Appl. Earth Obs. Geoinf. 12 (1), 36–46. http://dx.doi.org/10.1016/j.jag.2009.09.010, URL http://www.sciencedirect.com/science/article/pii/S0303243409000816.

Kimeldorf, G.S., Wahba, G., 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Stat. 41 (2), 495–502. http://dx.doi.org/10.1214/aoms/1177697089.

King, G., 1997. A Solution To the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton.

Kokoszka, P., Reimherr, M., 2017. Introduction To Functional Data Analysis. CRC Press, Boca Raton, FL.

Kuemmerle, T., Levers, C., Erb, K., Estel, S., Jepsen, M.R., Müller, D., Plutzar, C., Stürck, J., Verkerk, P.J., Verburg, P.H., Reenberg, A., 2016. Hotspots of land use change in europe. Environ. Res. Lett. 11 (6), 064020. http://dx.doi.org/10.1088/1748-9326/11/6/064020.

Lal, R., 2002. Soil carbon dynamics in cropland and rangeland. Environ. Pollut. 116 (3), 353–362. http://dx.doi.org/10.1016/S0269-7491(01)00211-1, URL https://www.sciencedirect.com/science/article/pii/S0269749101002111.

Lal, R., Shukla, M.K., 2004. Principles of Soil Physics. CRC Press.

Lasinio, G.J., Mastrantonio, G., Pollice, A., 2012. Discussing the "big n problem". Stat. Methods Appl. 22 (1), 97–112. http://dx.doi.org/10.1007/s10260-012-0207-2.

Li, Y., Brown, P., Gesink, D.C., Rue, H., 2012. Log Gaussian cox processes and spatially aggregated disease incidence data. Stat. Methods Med. Res. 21 (5), 479–507. http://dx.doi.org/10.1177/0962280212446326.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498.

Lindstrom, M.J., Bates, D.M., 1990. Nonlinear mixed effects models for repeated measures data. Biometrics 673–687.

Liu, J., You, L., Amini, M., Obersteiner, M., Herrero, M., Zehnder, A.J.B., Yang, H., 2010. A high-resolution assessment on global nitrogen flows in cropland. Proc. Natl. Acad. Sci. 107 (17), 8035–8040. http://dx.doi.org/10.1073/pnas.0913658107, URL https://www.pnas.org/content/107/17/8035.

Maimon, O., Rokach, L. (Eds.), 2010. Data Mining and Knowledge Discovery Handbook. Springer US, http://dx.doi.org/10.1007/978-0-387-09823-4.

Malone, B.P., McBratney, A.B., Minasny, B., Wheeler, I., 2012. A general method for downscaling earth resource information. Comput. Geosci. 41, 119–125.

Malone, B.P., Styc, Q., Minasny, B., McBratney, A.B., 2017. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91–99.

MapBiomas, 2019a. Cobertura e uso do solo, v. 3.1. Online; http://mapbiomas.org. (Accessed 01 June 2019).

MapBiomas, 2019b. Collection 3.1 of Brazilian land cover & use map series [Dataset, Version 3]. http://dx.doi.org/10.5281/ZENODO.2655065, URL https://zenodo.org/record/2655065.

Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Comput. Statist. Data Anal. 55 (7), 2372–2387.

Mello, M.P., Atzberger, C., Formaggio, A.R., 2014. Near real time yield estimation for sugarcane in Brazil combining remote sensing and official statistical data. In: 2014 IEEE Geoscience and Remote Sensing Symposium. pp. 5064–5067. http://dx.doi.org/10.1109/IGARSS.2014.6947635.

Miller, D.L., Glennie, R., Seaton, A.E., 2020. Understanding the stochastic partial differential equation approach to smoothing. J. Agric. Biol. Environ. Stat. 25 (1), 1–16.

Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Iversen, B.V., Greve, M.H., Minasny, B., 2019. Improved disaggregation of conventional soil maps. Geoderma 341, 148–160. http://dx.doi.org/10.1016/j.geoderma.2019.01.038.

Monteiro, J., Martins, B., Pires, J.M., 2018. A hybrid approach for the spatial disaggregation of socio-economic indicators. In. J. Data Sci. Anal. 5 (2–3), 189–211.

Moriondo, M., Maselli, F., Bindi, M., 2007. A simple model of regional wheat yield based on NDVI data. Eur. J. Agron. 26 (3), 266–274. http://dx.doi.org/10.1016/j.eja.2006.10.007.

Nelson, G.C., Rosegrant, M.W., Palazzo, A., Gray, I., Ingersoll, C., Robertson, R., Tokgoz, S., Zhu, T., Sulser, T.B., Ringler, C., et al., 2010. Food Security, Farming, and Climate Change to 2050: Scenarios, Results, Policy Options, Vol. 172. International Food Policy Research Institute.

Niamir, A., Skidmore, A.K., Toxopeus, A.G., Muñoz, A.R., Real, R., 2011. Finessing atlas data for species distribution models. Divers. Distrib. 17 (6), 1173–1185. http://dx.doi.org/10.1111/j.1472-4642.2011.00793.x.

Openshaw, S., Taylor, P.J., 1979. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In: Wrigley, N. (Ed.), Statistical Applications in the Spatial Sciences. Pion, London, pp. 127–144.

Orton, T.G., Saby, N.P.A., Arrouays, D., Walter, C., Lemercier, B., Schvartz, C., Lark, R.M., 2012. Spatial prediction of soil organic carbon from data on large and variable spatial supports. I. 0.167em inventory and mapping. Environmetrics 23 (2), 129–147. http://dx.doi.org/10.1002/env.2136.

Park, N.W., 2013. Spatial downscaling of TRMM precipitation using geostatistics and fine scale environmental variables. Adv. Meteorol. 2013, 1–9. http://dx.doi.org/10.1155/2013/237126.

Pavinato, P.S., Cherubin, M.R., Soltangheisi, A., Rocha, G.C., Chadwick, D.R., Jones, D.L., 2020. Revealing soil legacy phosphorus to promote sustainable agriculture in Brazil. Sci. Rep. 10 (1), http://dx.doi.org/10.1038/s41598-020-72302-1.

Piantadosi, S., Byar, D.P., Green, S.B., 1988. The ecological fallacy. Am. J. Epidemiol. 127 (5), 893–904.

Pinheiro, E.F.M., Lima, E., Ceddia, M.B., Urquiaga, S., Alves, B.J.R., Boddey, R.M., 2010. Impact of pre-harvest burning versus trash conservation on soil carbon and nitrogen stocks on a sugarcane plantation in the Brazilian Atlantic forest region. Plant Soil 333 (1–2), 71–80. http://dx.doi.org/10.1007/s11104-010-0320-7.

Poggio, L., Gimona, A., 2015. Downscaling and correction of regional climate models outputs with a hybrid geostatistical approach. Spatial Stat. 14, 4–21. http://dx.doi.org/10.1016/j.spasta.2015.04.006.

Portinho, J.L., Gomes, A.C.C., Koga-Vicente, A., Milani, F.C.C., Pentean, R.B., Manzatto, C.V., Spinelli-Araujo, L., Vicente, L.E., 2021. The pathways influence of agricultural expansion on water quality of fish farming in Ilha Solteira reservoir, São Paulo, Brazil. Aquaculture 536, 736405. http://dx.doi.org/10.1016/j.aquaculture.2021.736405.

Proietti, T., 2006. On the estimation of nonlinearly aggregated mixed models. J. Comput. Graph. Statist. 15 (1), 18–38.

R. Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

R. Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Rajão, R., Soares-Filho, B., Nunes, F., Börner, J., Machado, L., Assis, D., Oliveira, A., Pinto, L., Ribeiro, V., Rausch, L., Gibbs, H., Figueira, D., 2020. The rotten apples of Brazil's agribusiness. Science 369 (6501), 246–248. http://dx.doi.org/10.1126/science.aba6646, URL https://science.sciencemag.org/content/369/6501/246.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10 (3), e0118432. http://dx.doi.org/10.1371/journal.pone.0118432.

Silverman, B.W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J. R. Stat. Soc. Ser. B Stat. Methodol. 47 (1), 1–52, URL http://www.jstor.org/stable/2345542.

Simoes, R., Picoli, M.C.A., Camara, G., Maciel, A., Santos, L., Andrade, P.R., Sánchez, A., Ferreira, K., Carvalho, A., 2020. Land use and cover maps for Mato Grosso state in Brazil from 2001 to 2017. Sci. Data 7 (1), http://dx.doi.org/10.1038/s41597-020-0371-4.

Simpson, D., Lindgren, F., Rue, H., 2012. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. Environmetrics 23 (1), 65–74.

Sofaer, H.R., Hoeting, J.A., Jarnevich, C.S., 2019. The area under the precision-recall curve as a performance metric for rare binary events. Methods Ecol. Evol. 10 (4), 565–577. http://dx.doi.org/10.1111/2041-210x.13140.

Souza, C.M., Shimbo, J.Z., Rosa, M.R., Parente, L.L., Alencar, A.A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., Ferreira, L.G., Souza-Filho, P.W.M., de Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti, F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Neto, L.C.F., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V., Azevedo, T., 2020. Reconstructing three decades of land use and land cover changes in Brazilian biomes with landsat archive and earth engine. Remote Sens. 12 (17), 2735. http://dx.doi.org/10.3390/rs12172735.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Bastiani, F.D., 2019. Comments on: Modular regression—a Lego system for building structured additive distributional regression models with tensor product interactions. TEST 28 (1), 52–54. http://dx.doi.org/10.1007/s11749-019-00634-w.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible Regression and Smoothing: Using GAMLSS in R. Chapman and Hall/CRC.

Steinbuch, L., Orton, T.G., Brus, D.J., 2019. Model-based geostatistics from a Bayesian perspective: Investigating area-to-point kriging with small data sets. Math. Geosci. 52 (3), 397–423. http://dx.doi.org/10.1007/s11004-019-09840-6.

Sturrock, H.J., Cohen, J.M., Keil, P., Tatem, A.J., Menach, A.L., Ntshalintshali, N.E., Hsiang, M.S., Gosling, R.D., 2014. Fine-scale malaria risk mapping from routine aggregated case data. Malar. J. 13 (1), http://dx.doi.org/10.1186/1475-2875-13-421.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240 (4857), 1285–1293. http://dx.doi.org/10.1126/science.3287615.

Szyniszewska, A.M., 2020. CassavaMap, a fine-resolution disaggregation of cassava production and harvested area in Africa in 2014. Sci. Data 7 (1), http://dx.doi.org/10.1038/s41597-020-0501-z.

Tamis, W., Van 't Zelfde, M., 1998. An expert habitat suitability model for the disaggregation of bird survey data: Bird counts in the netherlands downscaled from atlas block to kilometre cell. Landsc. Urban Plan. 40 (4), 269–282. http://dx.doi.org/10.1016/S0169-2046(97)00092-3, URL http://www.sciencedirect.com/science/article/pii/S0169204697000923.

Tang, W., 2020. High Performance Computing for Geospatial Applications. Springer, Cham, Switzerland.

Tierney, L., Rossini, A.J., Li, N., Sevcikova, H., 2016. Snow: Simple Network of Workstations. R package version 0.4-2, URL https://CRAN.R-project.org/package=snow.

Tobler, W.R., 1979. Smooth pycnophylactic interpolation for geographical regions. J. Amer. Statist. Assoc. 74 (367), 519–530. http://dx.doi.org/10.1080/01621459.1979.10481647.

USGS, 2018. EarthExplorer. Online; https://earthexplorer.usgs.gov/. (Accessed 08 June 2018).

Utazi, C., Thorley, J., Alegana, V., Ferrari, M., Nilsen, K., Takahashi, S., Metcalf, C., Lessler, J., Tatem, A., 2019. A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. Stat. Methods Med. Res. 28 (10–11), 3226–3241.

Vieira, M.A., Formaggio, A.R., Rennó, C.D., Atzberger, C., Aguiar, D.A., Mello, M.P., 2012. Object based image analysis and data mining applied to a remotely sensed landsat time-series to map sugarcane over large areas. Remote Sens. Environ. 123, 553–562. http://dx.doi.org/10.1016/j.rse.2012.04.011.

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2018. Spatial disaggregation of complex soil map units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142.

Waclawovsky, A.J., Sato, P.M., Lembke, C.G., Moore, P.H., Souza, G.M., 2010. Sugarcane for bioenergy production: An assessment of yield and regulation of sucrose content. Plant Biotechnol. J. 8 (3), 263–276. http://dx.doi.org/10.1111/j.1467-7652.2009.00491.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7652.2009.00491.x.

Wakefield, J., 2005. Health-exposure modeling and the ecological fallacy. Biostatistics 7 (3), 438–455. http://dx.doi.org/10.1093/biostatistics/kxj017.

Walker, P.A., Mallawaarachchi, T., 1998. Disaggregating agricultural statistics using NOAA-AVHRR NDVI. Remote Sens. Environ. 63 (2), 112–125. http://dx.doi.org/10.1016/S0034-4257(97)00130-2, URL http://www.sciencedirect.com/science/article/pii/S0034425797001302.

Wall, L., Larocque, D., Léger, P.M., 2008. The early explanatory power of NDVI in crop yield modelling. Int. J. Remote Sens. 29 (8), 2211–2225. http://dx.doi.org/10.1080/01431160701395252.

Wang, C., Chen, M.H., Schifano, E., Wu, J., Yan, J., 2016. Statistical methods and computing for big data. Stat. Interface 9 (4), 399–414. http://dx.doi.org/10.4310/sii.2016.v9.n4.a1.

Wang, Q., Huang, J., Liu, R., Men, C., Guo, L., Miao, Y., Jiao, L., Wang, Y., Shoaib, M., Xia, X., 2020. Sequence-based statistical downscaling and its application to hydrologic simulations based on machine learning and big data. J. Hydrol. 586, 124875. http://dx.doi.org/10.1016/j.jhydrol.2020.124875.

Wardrop, N.A., Jochem, W.C., Bird, T.J., Chamberlain, H.R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., Tatem, A.J., 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. Proc. Natl. Acad. Sci. 115 (14), 3529–3537. http://dx.doi.org/10.1073/pnas.1715305115, URL https://www.pnas.org/content/115/14/3529.

Weiss, D.J., Lucas, T.C.D., Nguyen, M., Nandi, A.K., Bisanzio, D., Battle, K.E., Cameron, E., Twohig, K.A., Pfeffer, D.A., Rozier, J.A., Gibson, H.S., Rao, P.C., Casey, D., Bertozzi-Villa, A., Collins, E.L., Dalrymple, U., Gray, N., Harris, J.R., Howes, R.E., Kang, S.Y., Keddie, S.H., May, D., Rumisha, S., Thorn, M.P., Barber, R., Fullman, N., Huynh, C.K., Kulikoff, X., Kutz, M.J., Lopez, A.D., Mokdad, A.H., Naghavi, M., Nguyen, G., Shackelford, K.A., Vos, T., Wang, H., Smith, D.L., Lim, S.S., Murray, C.J.L., Bhatt, S., Hay, S.I., Gething, P.W., 2019. Mapping the global prevalence, incidence, and mortality of plasmodium falciparum, 2000–17: A spatial and temporal modelling study. Lancet 394 (10195), 322–331. http://dx.doi.org/10.1016/s0140-6736(19)31097-9.

Wikle, C.K., 2010. Low-rank representations for spatial processes. In: Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (Eds.), Handbook of Spatial Statistics. CRC Press, Boca Raton, pp. 107–118.

Wong, D., 2009. The modifiable areal unit problem (MAUP). In: Fotheringham, A.S., Rogerson, P.A. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, London, pp. 105–124, Chapter 7.

Wood, S.N., 2003. Thin-plate regression splines. J. R. Stat. Soc. (B) 65 (1), 95–114.

Wood, S.N., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62 (4), 1025–1036. http://dx.doi.org/10.1111/j.1541-0420.2006.00574.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00574.x.

Wood, S.N., 2017. Generalized Additive Models. Chapman and Hall/CRC, http://dx.doi.org/10.1201/9781315370279.

You, L., Wood, S., 2006. An entropy approach to spatial disaggregation of agricultural production. Agric. Syst. 90 (1–3), 329–347.

Yu, H., 2002. Rmpi: Parallel statistical computing in R. R News 2 (2), 10–14, URL http://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf.

Zhan, W., Chen, Y., Zhou, J., Wang, J., Liu, W., Voogt, J., Zhu, X., Quan, J., Li, J., 2013. Disaggregation of remotely sensed land surface temperature: Literature survey, taxonomy, issues, and caveats. Remote Sens. Environ. 131, 119–139. http://dx.doi.org/10.1016/j.rse.2012.12.014, URL http://www.sciencedirect.com/science/article/pii/S0034425712004804.

Zheng, Y., dos Santos Luciano, A.C., Dong, J., Yuan, W., 2021. High-resolution map of sugarcane cultivation in Brazil using a phenology-based method. Earth Syst. Sci. Data http://dx.doi.org/10.5194/essd-2021-88.