Constrained Functional Regression of National Forest Inventory Data over Time Using Remote Sensing Observations

Author Contributions Checklist Form

Data

Abstract

The data analyzed in this article has two components: (a) the remote sensing data on Tasseled Cap (TC) components from Landsat 7 ETM+ imagery, and (b) the field data on basal area measurements from FIA plots. Under the data policy of 2008, granting unrestricted access to the entire USGS archive of Landsat, (a) is publicly available. Access to (b) has been provided exclusively to the authors by USDA Forest Service Northern Research Station.

Availability (Mandatory)

- The TC dataset, used in Appendix A.1.3 in the supplementary materials as well as in Section 4.1 of the article, is derived from the publicly available archive of Landsat imagery and can be downloaded from DataVerse at https://doi.org/10.7910/DVN/CZLKCK.
- The basal area dataset, used in Sections 4.2, 4.3 and 4.4 of the article, cannot be shared since the USDA Forest Service cannot, by law, release the FIA plot coordinates due to data confidentiality and integrity concerns. However, we simulated

a hypothetical basal area dataset and included it in the Inputs_and_Outputs folder within the Codes.zip compressed file in the Supplementary Materials so that interested readers can implement the proposed model on that data using the codes we have provided. Please see the discussion in the item 2 of the following section about this simulated dataset.

- Additionally, it is useful to know that researchers may request access to the actual plot coordinates only by entering into a materials transfer agreement (MTA) with the USDA Forest Service. They should contact the Spatial Data Services (SDS) group of USDA Forest Service for such requests. More information on SDS can be found here: https://www.fia.fs.fed.us/tools-data/spatial/index.php. Alternatively, there is also a publicly available version of the FIA Database, containing plots with fuzzed coordinates and some swapped records that can be found at: https://apps.fs.usda.gov/fia/datamart/datamart.html.

Description (Mandatory if data available)

1. Remote Sensing Data:

   File Format: R data file (.Rdata)

   Metadata: The data file "TC_features_with_missing_values.Rdata" contains a list named pixel_wise_TC_data that consists of two components:

   (a) The first component is a 2560000×2 location matrix with its rows representing coordinates of the pixel centers from the study area described in Section 2 of the article. The coordinates are given in terms of the Albers Equal Area Conic projection with origin at 23°N and 96°W.

(b) The second component contains the spatiotemporal data in form of a $2560000 \times 120 \times 3$ array where first and second dimensions correspond to pixels and time points (months), respectively. Along the third dimension, it contains values for TC1 (Brightness), TC2 (Greenness) and TC3 (Wetness). The occurrences of Zero (0) within this array indicate missing data at those space-time combinations.

Version information: 3.0

2. Hypothetical Basal Area Data:

File Format: CSV data file (.csv)

Data: The data file named hypothetical_basal_area_data contains the simulated live tree basal area data. The data is in form of a matrix with five columns where first two entries in each row represents the coordinates of the field plot of live tree basal area data, next two entries represents the time point with year and month number (a number from 1 to 12) of the corresponding year, respectively, and the last entry shows the corresponding the live tree basal area measurement.

This hypothetical dataset was simulated under the proposed model (using a specified set of parameter values) without utilizing any information about the field plot coordinates in the actual basal area data. The sole purpose of providing this dataset is that, in absence of the actual basal area dataset, the users can explore the attached codes by running them on this hypothetical dataset. Thus, it is useful to remember that,

- the figures and the tables obtained by running the attached codes on this dataset

may exhibit patterns different from the corresponding figures and tables in Sections 4.2, 4.3 and 4.4 of the article.

- this dataset should never be used for validation of the proposed approach (or for comparison against other methods) since it was generated under the proposed approach.

Code

Abstract

Within the "Codes.zip" compressed file in the Supplementary Materials, there are three folders named TC_Codes, FIA_Codes as well as Inputs_and_Outputs, a .Rproj file named Codes.Rproj and a .pdf file named README.pdf that contains the detail instructions to use these codes sequentially. The Inputs_and_Outputs folder contains a R data (.Rdata) and a CSV file (.csv) named parameters_used_to_simulate_basal_area_data and hypothetical_basal_area_data, respectively, and is intended to store the data as well as outputs from these codes. The R codes are provided in two separate folders: (i) TC_Codes: related to the models in Appendix A.1.1 in the supplementary materials is to produce the outputs of Appendix A.1.3 in the supplementary materials as well as of Section 4.1 (additionally, Figure 2 in Section 2) of the article and (ii) FIA_Codes: related to the models in Section 3 is to produce the outputs of Sections 4.2, 4,3 as well as 4.4 of the article. First, run all codes from TC_Codes folder and then go to the FIA_Codes folder since the latter uses the output files generated by the former.

Description

The attached codes are implemented using R (R Core Team, 2019, Version 3.6.1) within RStudio (RStudio Team, 2019, Version 1.2.5033). All codes are provided as R script files with extension .R. The authors plan to post the codes in the JASA ACS Github Repository after acceptance of the manuscript.

Optional Information

Supporting software requirements: the R packages needed to be installed are:

- abind (Version 1.4-5)
- ape (Version 5.2)
- coda (Version 0.19-2)
- fields (Version 9.6)
- ggplot2 (Version 3.1.0)
- gridBase (Version 0.4-7)
- lattice (Version0.20-38)
- ltsa (Version 1.4.6)
- matrixStats (Version 0.54.0)
- mvtnorm (Version 1.0-8)
- proj4 (Version 1.0-8)
- RColorBrewer (Version 1.1-2)
- sp (Version 1.3-1)
- e1071 (Version 1.7-3)
- gbm (Version 2.1.5)
- gam (Version 1.16.1)

Instructions for Use

Reproducibility

The following instructions can only reproduce Table A.1.1 as well as Figure A.1.1 of Appendix A.1.3 in the supplementary materials and Figures 2-4 of the article, that only use the dataset on TC variables. Since we cannot provide the actual basal area data due to data confidentiality, we provide a simulated dataset of hypothetical basal area measurements so that one can produce outputs equivalent to (but not same as) Table 1-2 and Figures 5-8 of the article. In addition, we provide the R code and necessary parameters used to simulate basal area data.

How to reproduce analyses:

1. Download the Codes.zip folder and unzip it. This folder consists of three sub-folders named TC_Codes, FIA_Codes as well as Inputs_and_Outputs, a .Rproj file named Codes.Rproj and a .pdf file named README.pdf that contains the detail instructions to use these codes. The Inputs_and_Outputs folder contains a R data (.Rdata) and a CSV file (.csv) named parameters_used_to_simulate_basal_area_data and hypothetical_basal_area_data, respectively, and is intended to store the data as well as the outputs from these codes.

2. Go to https://doi.org/10.7910/DVN/CZLKCK to download and save the data on TC variables in Inputs_and_Outputs folder.

3. Install Rstudio. Then, open Codes.Rproj file using File/Open Project in RStudio or double-click the Codes.Rproj file. In case one wants to use the standalone R, the

Codes folder needs to be set as the working directory of R using setwd("dir") where dir should be replaced by the path of the Codes folder.

4. Before sourcing any of these R codes, install the R packages: abind, ape, coda, fields, ggplot2, gridBase, lattice, ltsa, matrixStats, mvtnorm, proj4, RColorBrewer, sp, e1071, gbm, gam.

5. Now, start with TC_Codes folder and run Aggregation.R to aggregate every adjacent 16×16 pixels as described in Section 2 of the article. (Runtime: $\sim$ 21 minutes.)

6. Source Figure_2.R to reproduce Figure 2. (Runtime: Instantaneous.)

7. Then source TC_Model_NUMBER.R where NUMBER should be replaced by any one of the five candidate model numbers from Appendix A.1.1 in the supplementary materials − I, II, IIIA, IIIB and IV. This code runs the particular model on the entire set of available TC observations and fills in the missing values. In addition, it calculates the predictive uncertainty, log likelihood (LL), BPIC and Moran's I statistic. (Runtime: For 60,000 iterations, Model I takes $\sim$ 24 minutes, Model II takes $\sim$ 1.1 hours, Model IIIA takes $\sim$ 3 hours, Model IIIB takes $\sim$ 7 hours, and Model IV takes $\sim$ 8 hours. Additionally, for Moran's I computation, $\sim$ 10 minutes is required for each of these codes.)

8. Run TC_cv_positions_selection.R to randomly select the positions that will be considered as test data in the cross validation. (Runtime: Instantaneous.)

9. Then source TC_Model_NUMBER_cv.R where NUMBER should be replaced by any one of the five candidate model numbers from Appendix A.1.1 in the supplementary materials − I, II, IIIA, IIIB and IV. This code performs cross validation by holding

out a subset of available TC observations as test set. (Runtime: For 36 test sets, per 60,000 iterations, Model I takes $\sim$ 24 minutes, Model II takes $\sim$ 1.1 hours, Model IIIA takes $\sim$ 3 hours, Model IIIB takes $\sim$ 7 hours, and Model IV takes $\sim$ 8 hours.)

10. Now, to reproduce Table A.1.1 and Figure A.1.1 of Appendix A.1.3 in the supplementary materials, run Table_A.1.1_Figure_A.1.1.R. (Runtime: Instantaneous.)

11. Source Figure_3_4.R to reproduce Figure 3 and Figure 4. (Runtime: Instantaneous.)

12. Run TC_best_model_output.R to merge the complete sets of TC1, TC2 and TC3 values, after filling in the missing ones using predictions from the best candidate model. (Runtime: Instantaneous.)

Note: The rest of the figures and tables in the article are not reproducible due to non-availability of the actual basal area data. However, a simulated dataset of hypothetical basal area measurements is provided to produce equivalent outputs.

13. Now, go to the FIA_Codes folder and run TC_best_output_merge_basal_area.R to merge the output of TC_best_model_output.R with basal area data. (Runtime: Instantaneous.)

14. Then run Model_for_y1.R, Unconstrained_y2_model.R and Constrained_y2_model.R. (Runtime: For 60,000 iterations, Model for $y^{(1)}$ takes $\sim$ 1 minute, Unconstrained Model for $y^{(2)}$ takes $\sim$ 1 minute, and Constrained Model for $y^{(2)}$ takes $\sim$ 5 hours.)

15. Now source Basal_area_cv_positions_selection.R. The outputs of this .R file are two .Rdata files where one contains the random positions of both zero, non-zero live tree basal area data and the other contains only the random positions of non-zero live

tree basal area data those will be considered as test data in holdout cross validation method. (Runtime: Instantaneous.)

16. Run Model_for_y1_cv.R, Unconstrained_y2_model_cv.R, Constrained_y2_model_cv.R, Nonparametric_Models_for_y1_cv.R and Nonparametric_Models_for_y2_cv.R. These codes perform cross validation using holdout method. (Runtime: For 36 test sets, per 60,000 iterations, Model for $y^{(1)}$ takes $\sim$ 1 minute, Unconstrained Model for $y^{(2)}$ takes $\sim$ 1 minute, and Constrained Model for $y^{(2)}$ takes $\sim$ 5 hours. The cross validation of Nonparametric Models for $y^{(1)}$ and $y^{(2)}$ takes $<$ 1 minute each.)

17. Source Table_1_2_Figure_5.R to produce equivalents of Table 1, Table 2 and Figure 5 of the article. (Runtime: Instantaneous.)

18. Run Basal_area_model_cv.R to produce the cross validation outputs of combined model for B_{sk} (unconstrained $y^{(1)}$ and constrained $y^{(2)}$ model) equivalent to those specified in Section 4.2 of the article. (Runtime: For 36 test sets, per 60,000 iterations, the basal area model takes $\sim$ 5 hours.)

19. Finally, produce equivalents of Figure 6, Figure 7 and Figure 8 in the article by sourcing Figure_6_7_8.R. (Runtime: $\sim$ 3 minutes.)

Notes:

- In case one wants to simulate a hypothetical basal area data different from what is already included in Inputs_and_Outputs folder, the code file Basal_area_data_simulation.R needs to be run, with a different seed and/or specifying a different set of model parameters, after Step 12.

- Runtimes, reported for the steps above, are as observed while implementing the codes using a single 3.79GHz processor in a Windows machine with 128GB RAM.

References

R Core Team (2019), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2019), *RStudio: integrated development environment for R*, RStudio, Inc., Boston, MA.