

Part 1: Data

- ☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).
- ☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

The reliable prediction of crop production is a crucial task for policymakers as well as individual planners to make decisions related to agricultural risk management. This study builds up a model for corn yield prediction at large spatial scales based on meteorological variables, temperature trajectories and precipitation, which have significant impacts on agricultural productivity. The data consists of the county-level annual corn yield data (measured in bushels per acre) from a part of the Corn Belt in the United States, including Illinois, Indiana, Iowa, Kansas, and Missouri, along with county-level climate information, daily maximum and minimum temperature trajectories, and annual precipitation, from 1999 to 2020. We model the daily temperature trajectories as bivariate functional predictors and the annual precipitation as a scalar covariate.

Availability

- ☒ Data **are** publicly available.
- ☐ Data **cannot be made** publicly available.

Publicly available data

- ☐ Data are available online at:
- ☒ Data are available as part of the paper’s supplementary material.
- ☐ Data are publicly available by request, following the process described here:
- ☐ Data are or will be made available through some other mechanism, described here:

Organized data is available by loading “MidwestData.RData”. The county-level annual corn yield data (measured in bushels per acre) is obtained from the National Agricultural Statistics Agency (<https://quicksstats.nass.usda.gov/>) and daily meteorology measurements for each county are collected from the National Climatic Data Center (<https://www.ncdc.noaa.gov/data-access>).

Description

“MidwestData.RData” consists of four datasets.

- CountyI.info: state-county information with unique id (“CountyI”) assigned to each of them. The abbreviated state name is followed by the county name, for example, “IL-ADAMS” indicates county Adams in Illinois state. We use a total of 403 counties from five states in our analysis.
- dist.mat: distance matrix (km) among 403 counties specified in CountyI.info with the matched order.
- regdat: data frame including information of year (“Year”), state (“State”), county (“County”), unique county id (“CountyI”) introduced in CountyI.info, annual corn yield per acre (“Yield”), annual precipitation (“avgPRCP”), and the size of harvest land (“Area”).
- fundat: centered trajectories of maximum (columns 1 to 365) and minimum temperatures (columns 366 to 730), matched to the observation in the regdat data.

File format(s)

- ☐ CSV or other plain text.
- ☒ Software-specific binary format (.Rda, Python pickle, etc.): .RData
- ☐ Standardized binary format (e.g., netCDF, HDF5, etc.):

☐ Other (please specify):

Data dictionary

- ☐ Provided by authors in the following file(s):
- ☒ Data file(s) is(are) self-describing (e.g., netCDF files)
- ☐ Available at the following URL:

Part 2: Code

Abstract

We provide the code to reproduce the corn yield analysis results in Section 5.3 of the manuscript by fitting the Bayesian spatially varying functional model (BSVFM) to the whole data. It includes source files including all help functions necessary to run the model. The main code contains the MCMC algorithms to perform the multivariate spatial FPCA, estimate the model parameters, and implement the spatial model selection.

Description

The provided code folder includes four R files; (i) **MFDLMBasisEstimation.R**, (ii) **GenSamples.R**, (iii) **FitMidWestData.R**, and (iv) **ModelEst.R**

The first two files, **MFDLMBasisEstimation.R** and **GenSamples.R**, are source files containing all necessary functions to implement the MCMC:

- **MFDLMBasisEstimation.R**: help functions to estimate spatial FPC of multivariate functional predictors. This code is borrowed from a part of Supplementary Material of Kowal et al. (2017) “A Bayesian Multivariate Functional Dynamic Linear Model”, JASA.
- **GenSamples.R**: help functions to estimate model parameters based on posterior distributions derived in Supplementary Material.

FitMidWestData.R is the main code to fit the model for the corn yield data analysis. It includes descriptions about data preparation, choice of initial values for MCMC algorithms, and MCMC loop to reproduce the results in the manuscript. As our MCMC implementation depends on randomness on the setting of start values as well as on each iteration, we set the seed number to reproduce the exactly same results in the manuscript. However, different seed number is also okay. Each iteration of the MCMC algorithm takes, approximately, 25 seconds using Intel(R) Core(TM) i-7 with 4 cores.

ModelEst.R includes the code to estimate model parameters based on samples obtained by running **FitMidWestData.R**.

Code format(s)

- ☒ Script files
 - ☒ R
 - ☐ Python
 - ☐ Matlab
 - ☐ Other:
- ☒ Package
 - ☒ R
 - ☐ Python
 - ☐ MATLAB toolbox
 - ☐ Other:
- ☐ Reproducible report
 - ☐ R Markdown
 - ☐ Jupyter notebook

- ☐ Other:
- ☐ Shell script
- ☐ Other (please specify):

Supporting software requirements

R

Version of primary software used

R version 4.0.2

Libraries and dependencies used by the code

- fda version 5.1
- splines version 4.0.2
- mvtnorm version 1.1.1
- truncnorm version 1.0.8
- MCMCpack version 1.4.9
- KFAS version 1.3.7

Parallelization used

- ☒ No parallel code used
- ☐ Multi-core parallelization on a single machine/node
 - Number of cores used:
- ☐ Multi-machine/multi-node parallelization
 - Number of nodes and cores used: 3 nodes, 243 cores

License

- ☒ MIT License (default)
- ☐ BSD
- ☐ GPL v3.0
- ☐ Creative Commons
- ☐ Other: (please specify below)

Additional information (optional)

Scope

The provided workflow reproduces:

- ☐ Any numbers provided in text in the paper
- ☐ All tables and figures in the paper
- ☒ Selected tables and figures in the paper, as explained and justified below:

We implement the MCMC algorithms for a total of 15,000 iterations and obtain a posterior sample of size 2500 by using the first 5000 iterations as burn-in and thinning the remaining 10,000 by a factor of 4. Figure 4-7 can be reproduced based on estimated coefficients or parameters calculated from MCMC iterations.

Workflow

Format(s)

- ☒ Single master code file
- ☐ Wrapper (shell) script(s)
- ☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach

- ☐ Text file (e.g., a readme-style file) that documents workflow
- ☐ Makefile
- ☐ Other (more detail in *Instructions* below)

Instructions

1. Load the data: **MidwestData.RData**.
2. Load source files: **MFDLMBasisEstimation.R** and **GenSamples.R**.
3. Run the main code, **FitMidWestData.R**, following each step following the provided order.
4. Run the code for parameter estimation, **ModelEst.R**, based on saved objects from the main code in step 3.

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

- ☐ < 1 minute
- ☐ 1-10 minutes
- ☐ 10-60 minutes
- ☐ 1-8 hours
- ☒ > 8 hours
- ☐ Not feasible to run on a desktop machine, as described here: