Reproducible Geoprocessing of Agricultural, Climate, and Land Use Data at Scale with R

Nicholas A. Potter	Email:	nicholas.a.potter@wsu.edu
Economics PhD Candidate	Twitter:	@econpotter
Washington State University	Follow along:	github repository

TWEEDS 2020

October 30, 2020

Initial Idea

- We start out as bright-eyed researchers

- But spatial data processing is difficult and slow, especially in the beginning

- How can we set up a workflow to minimize frustrations and errors?

Genesis

What is the relationship between irrigated agriculture and climate?

Genesis

Irrigation adds complexity:

Genesis

Irrigation adds complexity:

- Decouples growing season from precipitation
  - Conditional on water supply, warmer winters may allow shifting the growing season to avoid extreme heat

Genesis

Irrigation adds complexity:

- Decouples growing season from precipitation
    - Conditional on water supply, warmer winters may allow shifting the growing season to avoid extreme heat
- Nonlocal climate effects become relevant: reduced snowpack in distant mountains affects water supply

Genesis

Irrigation adds complexity:

- Decouples growing season from precipitation
  - Conditional on water supply, warmer winters may allow shifting the growing season to avoid extreme heat

- Nonlocal climate effects become relevant: reduced snowpack in distant mountains affects water supply

- Enables a diverse array of crop choices

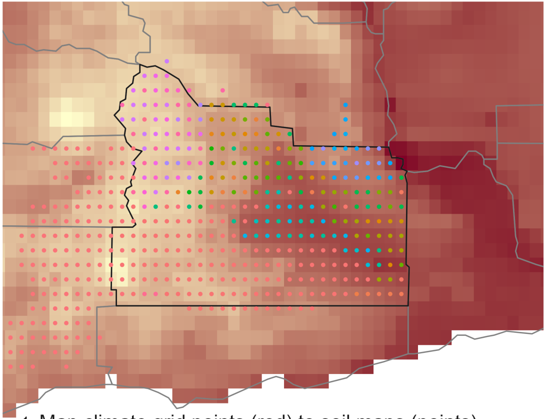
Diverse Data Needs

- Climate: gridMet/MACA daily climate records and projections (4km resolution)

- Land Use: USDA's Cropland Data Layer (30m resolution, 1.5GB per year zipped)

- Soil: USDA/USGS SSURGO

- County-level agriculture and water use: USDA-NASS Quick Stats, USGS water withdrawals, Census demographics



1. Map climate grid points (red) to soil maps (points)
2. Summarize by county

142 counties in 2012, 131 counties in 2040, 2060, 2080

Diverse Data Problems

- Long processing times and large storage needs

Diverse Data Problems

- Long processing times and large storage needs
- Mistakes and changes are easy to make and costly

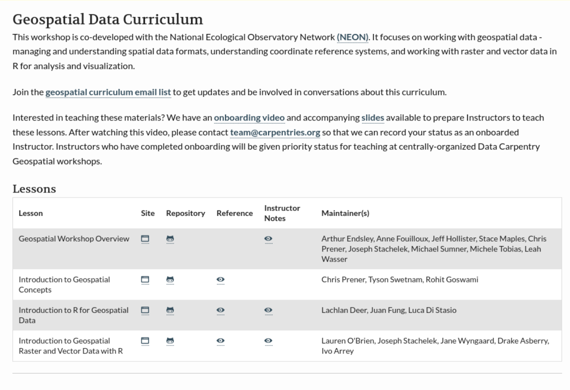
Diverse Data Problems

- Long processing times and large storage needs
- Mistakes and changes are easy to make and costly
- Differences in software versions add complications

What can I contribute?

There's already loads of info on geoprocessing and research computing best practices:

- Data and Software Carpentry's geospatial lessons and reproducible research lessons
- Grant McDermott's environmental economics and data science course
- Robin Lovelace's Geocomputation with R



Geospatial Data Curriculum

This workshop is co-developed with the National Ecological Observatory Network (NEON). It focuses on working with geospatial data - managing and understanding spatial data formats, understanding coordinate reference systems, and working with raster and vector data in R for analysis and visualization.

Join the geospatial curriculum email list to get updates and be involved in conversations about this curriculum.

Interested in teaching these materials? We have an onboarding video and accompanying slides available to prepare Instructors to teach these lessons. After watching this video, please contact team@carpentries.org so that we can record your status as an onboarded Instructor. Instructors who have completed onboarding will be given priority status for teaching at centrally-organized Data Carpentry Geospatial workshops.

Lessons

Lesson	Site	Repository	Reference	Instructor Notes	Maintainer(s)
Geospatial Workshop Overview	🗔	🐱		👁	Arthur Endsley, Anne Fouilloux, Jeff Hollister, Stace Maples, Chris Prener, Joseph Stachelek, Michael Sumner, Michele Tobias, Leah Wasser
Introduction to Geospatial Concepts	🗔	🐱	👁		Chris Prener, Tyson Swetnam, Rohit Goswami
Introduction to R for Geospatial Data	🗔	🐱	👁	👁	Lachlan Deer, Juan Fung, Luca Di Stasio
Introduction to Geospatial Raster and Vector Data with R	🗔	🐱	👁	👁	Lauren O'Brien, Joseph Stachelek, Jane Wyngaard, Drake Asberry, Ivo Arrey

What can I contribute?

I don’t really want to only talk about all the ways I messed up for the past 4+ years

What even are projection coordinates?

Rerun the climate processing on the HPC again?!?

R crashes when I try to upscale from 30m to 4km?

Be brave

What can I contribute?

Two things:

- rnassqs, an R package for accessing USDA-NASS data
- A scalable spatial data science environment geared toward “medium” data in the GB-TB range

Accessing USDA-NASS data with rnassqs

- rnassqs, provides access to USDA-NASS Quick Stats data via API

- Good for recurring queries and for reproducibility

- Other options:
  - tidyusda: easier for mapping, perhaps less flexible
  - direct ftp download of entire dataset

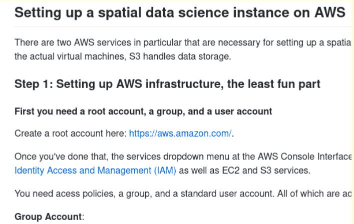
```
# First define a base parameter list to modify for each new query
base_params <- list(
  sector_desc = "ECONOMICS",
  commodity_desc = "AG LAND",
  agg_level_desc = "COUNTY",
  unit_desc = "ACRES",
  statisticcat_desc = "AREA",
  domain_desc = "TOTAL",
  domaincat_desc = "NOT SPECIFIED",
  year = 2012
)

# List of parameters that vary for each query
param_list <- list(
  ag_land_other = list(
    class_desc = "(EXCL CROPLAND & PASTURELAND & WOODLAND)"),
  ag_land_irr = list(
    prodn_practice_desc = "IRRIGATED",
    class_desc = "ALL CLASSES"),
  ag_woodland = list(
    class_desc = "WOODLAND"),
```

Aside: a software review process (kudos to rOpenSci) is extremely helpful

A Scalable Spatial DS Environment

Incomplete guide: Setting up an AWS instance with spatial data science software

- RStudio Server or Jupyter - can work from anywhere in the same software environment
- Scalable (and cheap!) computation and storage
- consistent software environment
- Why not just use an HPC? HPCs are great! This allows scalable easy use with complete control and easier debugging. Good for medium level needs.



Setting up a spatial data science instance on AWS

There are two AWS services in particular that are necessary for setting up a spatia
the actual virtual machines, S3 handles data storage.

Step 1: Setting up AWS infrastructure, the least fun part

First you need a root account, a group, and a user account

Create a root account here: https://aws.amazon.com/.

Once you've done that, the services dropdown menu at the AWS Console Interface
Identity Access and Management (IAM) as well as EC2 and S3 services.

You need acess policies, a group, and a standard user account. All of which are ac

Group Account:

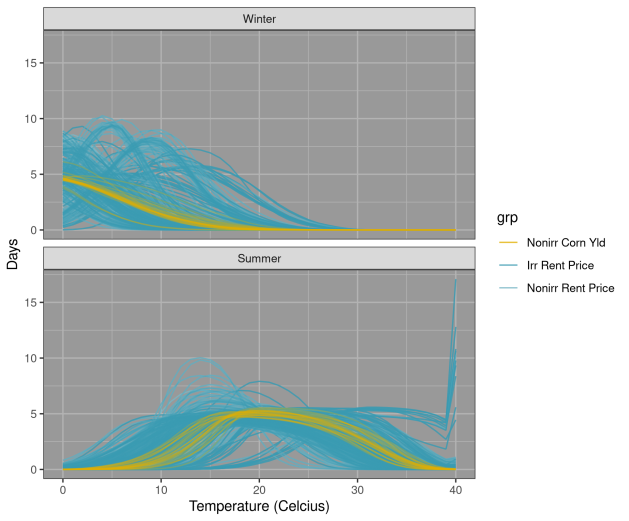
Actual Results

Seasonal time-at-temperature to detect thresholds at which heat is harmful

- Worked for Schlenker and Roberts (2009) using nonirrigated corn yields

- We replicate their result for CO, WY, and NM counties

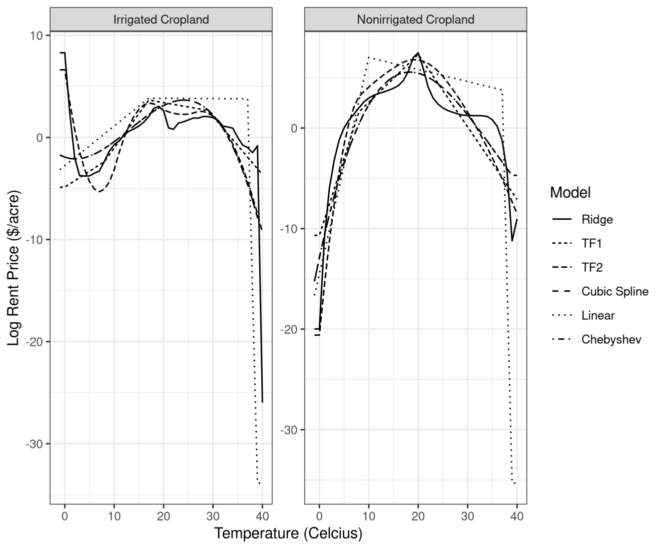
- But climate in counties with rental prices in the Mountain West is significantly different



Actual Results

- Some indication of a decline in rental price with more time spent above 30°C if we exclude warm winter counties

- Counties with warm winters and water may be able to mitigate heat impacts by shifting growing seasons to avoid extreme heat.



Minimizing Time Costs: Mortal Sins and Guiding Lights

Guiding light: minimum viable unit

- DRY versus Premature abstraction / optimization
- Big data is usually decomposable into small or medium data, work with smallest useful size and scale afterward
- Profile as needed (took processing 30 years of climate data from 3 days to 3 hours)
- Data assertions and tests (don't have to be in a formal testing framework)