Assessing potential environmental justice implications of U.S. food production

A Capstone Project submitted in partial satisfaction of the requirements for the degree of Master of Environmental Data Science for the Bren School of Environmental Science & Management

By Connor Flynn, Mia Forsline, Scout Leonard, and Alex Vand

June 3, 2022

Assessing potential environmental justice implications of U.S. food production

As developers of this Capstone Project documentation, we archive this documentation on the Bren Schools' website such that the results of our research are available for all to read. Our signatures on the document signify our joint responsibility to fulfill the archiving standards set by the Bren School of Environmental Science & Management

Connor Flynn

Mia Forsline

Scout Leonard

Alex Vand

The Bren School of Environmental Science & Management produces professionals with unrivaled training in environmental science and management who will devote their unique skills to the diagnosis, assessment, mitigation, prevention, and remedy of the environmental problems of today and the future. A guiding principle of the School is that the analysis of environmental problems requires quantitative training in more than one discipline and an awareness of the physical, biological, social, political, and economic consequences that arise from scientific or technological decisions.

The Capstone Project is required of all students in the Master of Environmental Data Science (MEDS) Program. The project is a six-month-long activity in which small groups of students contribute to data science practices, products or analyses that address a challenge or need related to a specific environmental issue. This MEDS Capstone Project Technical Documentation is authored by MEDS students and has been reviewed and approved by:

Dr. Benjamin Halpern

Dr. Allison Horst

Date

Contents

Chapter 1

Abstract

As the human population grows, global food production must increase to satisfy the rising demand for food. Absent from much of the research to address this developing challenge is recognition of how food production systematically exacerbates environmental inequities. In short, food production is a complex yet often overlooked pillar in environmental justice discourse.

New data from the Client of this project, the National Center for Ecological Analysis & Synthesis (NCEAS) Environmental Impact and Sustainability of Global Food Systems working group maps, at high resolution, nearly all reported global food production and associated environmental stressors. This data offers a novel opportunity to examine how the spatial distribution of these stressors may differentially impact human welfare.

This project seeks to harmonize the Client's footprint of food production data with human welfare data from the County Health Rankings & Roadmaps program to explore county-level patterns in the United States. Such identified spatial patterns are used to explore potential environmental justice implications regarding the human health consequences of U.S food production. This work will inform future research by the Client which will guide more equitable and sustainable food production policies and management. Lastly, a reproducible, analytical workflow is constructed to visualize findings and support future NCEAS research publications.

Chapter 2

Executive Summary

As human populations and demand for food grow, increasing global food production to meet this demand strains finite natural resources [Conjin et al., 2018], generates a significant environmental footprint [Bene et al., 2019], and is the focus of a wide range of environmental sustainability initiatives [Campi et al., 2021, Sanyé-Mengual et al., 2018]. Absent from much of this research and policy is recognition of the ways food production exacerbates and perpetuates environmental inequities. For example, the global agricultural sector contributes to thousands of air-pollution-caused premature deaths every year [Domingo et al., 2021], and air pollution has been shown to disproportionately burden vulnerable communities of color [Jbaily et al., 2022]. Food production is a key, complex, and often overlooked pillar in environmental justice discourse [Menton et al., 2020].

Previous studies of food production typically focus on singular environmental stressors from singular systems [Jbaily et al., 2022, Domingo et al., 2021]. However, newly available data from the National Center for Ecological Analysis & Synthesis (NCEAS) Environmental Impact and Sustainability of Global Food Systems working group maps, at high resolution, nearly all reported global food production and four primary associated environmental stressors: greenhouse gas emissions, nutrient pollution, spatial disturbance, and water consumption.

This data offers a novel opportunity to examine how the spatial distribution of these stressors may differentially impact human welfare of different communities.

This project harmonizes NCEAS global footprint of food production data with human welfare data from the University of Wisconsin Population Health Institute's County Health Rankings & Roadmaps program to explore county-level patterns in the United States. Key components of this work include the preparation of the NCEAS dataset for analysis at the U.S. county level and preliminary exploration of potential environmental justice implications regarding the human health consequences of U.S. food production.

The NCEAS global environmental footprint of food production data is synthesized by the National Center for Ecological Analysis & Synthesis (NCEAS) Environmental Impact and Sustainability of Global Food Systems working group (hereafter referred to as the Client). Upon publication of future NCEAS research, supplementary materials about the collection of the food production data will be made available.

This project's processing of the NCEAS global environmental footprint of food production data calls on 472 raw geoTIFF files from the NCEAS server Aurora. These files spatially represent the environmental pressures of food globally and are stored on the NCEAS private Aurora server.

Environmental Pressures of Food Production Variables

Dataset from the NCEAS Environmental Impact and Sustainability of Global Food Systems working group

Pressure Variable	Variable Name	Units
Greenhouse Gas Emissions	ghg	tons CO2 equivalent per year
Nutrient Pollution	nutrient	tons nitrogen (N) and Phosphorus (P) per year
Spatial Disturbance	distrubance	square kilometers equivalent
Water Consumption	water	cubic meters of water per year

A reproducible, analytical workflow was constructed to process this data for analysis at the U.S. county level. These patterns were visualized in R to archive and convey findings as well as produce preliminary visualizations to support future NCEAS research, such as global analyses, of these environmental justice implications. Moving forward, identified spatial patterns and anomalies can also inform more equitable and sustainable food production policies and management.

Using this data, the products from these analyses include:

1. Intermediate processed datasets
2. Maps of environmental pressures from food production distributed among U.S. counties
3. Correlation plots visualizing correlations between environmental pressures from food productions and geographic regions
4. Variable importance plots based on random forest regressions to identify relevant input variables

The MEDS (Masters of Environmental Data Science) Capstone team acknowledges those whose support was instrumental to the completion and success of this Capstone project. These supportive collaborators and mentors include, but are not limited to:

- Our faculty advisors:
  - Dr. Bejamin Halpern, NCEAS Executive Director;
  - Dr. Allison Horst, Bren School of Environmental Science and Management, UC Santa Barbara;
- The Halpern Lab:

  - Gage Clawson, NCEAS;
  - Haley Epperly, NCEAS;
  - Dr. Melanie Frazier, NCEAS;
  - Dr. Caitlin Fong, NCEAS;

- NCEAS support staff:

  - Nick Outin, NCEAS;

- and the NCEAS Environmental Impact and Sustainability of Global Food Systems working group.

Chapter 3

Problem Statement

Producing food more sustainably is of paramount importance to feed the growing global population without degrading the planet [Conjin et al., 2018]. Currently, global food production uses approximately 50% of habitable land and 4% of sea area, accounts for about 70% of global freshwater withdrawal, and is responsible for 26% of all anthropogenic greenhouse gas (GHG) emissions [Kuempel et al., 2020]. Thus, global food systems are crucial to consider when discussing salient environmental issues, such as mitigating the detrimental effects of climate change.

Food production also impacts human welfare, though not always evenly. Environmental impacts from human activities are often felt disproportionately by underserved or marginalized communities [Boyce, 1994]. This realm of environmental justice is beginning to interact with the interdisciplinary field of *food justice*—the study of inequalities of race, class, and gender within food systems. Food justice research has skyrocketed in the past decade and explores a wide variety of topics centering around *just sustainability*, or the intersection of ecological sustainability and social justice [Glennie and Alkon, 2018]. To answer many key environmental justice questions regarding food production, there is a need for information about where and how diverse people experience different environmental pressures. The recent development of computationally robust methods to map and visualize the cumulative pressures and impacts of all food types across the production process can help achieve these goals [Kuempel et al., 2020].

However, missing from this body of research is an examination of how the total environmental footprint of food production results in environmental pressures that affect human health and well-being. While maps of individual environmental pressures from specific food sectors exist, aggregate maps of all foods and all key environmental pressures from these foods are lacking. This knowledge gap has largely persisted due to lack of data. Until now, researchers have not possessed systematically studied, spatially-mapped, and comprehensive environmental pressure data for global food systems.

With the novel data collected by the National Center for Ecological Analysis & Synthesis (NCEAS) Environmental Impact and Sustainability of Global Food Systems working group, this project's goal is to process this data for use in exploring how exposure to environmental pressures relate to the risk of detrimental human welfare variables and to conduct preliminary explorations of such dynamics. Understanding the spatial relationships between environmental pressures and human welfare can inform environmentally just policies, management, and research to improve global food systems moving forward.

Chapter 4

Specific Objectives

Ultimately, this project seeks to explore and visualize spatial patterns relating human welfare indicators with one or more of four environmental pressures resulting from food systems:

1. Greenhouse gas emissions,
2. Nutrient pollution,
3. Spatial disturbance, and
4. Water consumption.

To achieve this, there are four main objectives:

1. Process Client data for use in exploring spatial patterns of four individual food system pressures in the United States at the County level;
2. Create choropleth maps to visualize spatial patterns of four individual food systems pressures in addition to cumulative food systems pressure in the United States at the county-level;
3. Produce correlation plots to visually examine correlation of food systems pressures and various human welfare indicator variables in different counties and geographic regions in the United States; and
4. Use random forest regressions to develop a workflow which quantifies variable importance and identifies which variables contribute the most to variance in a human health outcome at the U.S. county-level.

Chapter 5

Summary of Solution Design

5.1 Approach and Methods

To explore and visualize spatial patterns between human welfare indicators and environmental pressures resulting from food production in the United States, the following tasks were completed in R version 4.1.0 (2022-05-24):

5.1.1 Establish collaborative workflow

First, we obtained access the NCEAS private Aurora server to store and retrieve both data and outputs.

Next, we created a GitHub organization for collaboration on repositories related to the project. Within the GitHub organization, we set up three GitHub repositories:

1. A repository for the main code, analysis and workflow documentations, and analysis outputs and visualizations;
2. A repository for GitHub issues to facilitate project management and documentation of resources, troubleshooting, research, and internal communication about completion of deliverables and analytical findings; and
3. A repository for the Technical Documentation, rendered in `Bookdown` [Xie, 2021].

5.1.2 Prepare NCEAS datasets

We examined the NCEAS working group's environmental pressures of global food production data and extract relevant data for the scope of the exploration using R. This included subsetting the global datasets to just the United States and calculating zonal statistics for U.S. counties to examine county-level differences in environmental pressures from different categories of food production. More details can be found in the Processed Dataset section of the User Documentation.

5.1.3 Prepare human welfare dataset

Next, we examined and cleaned county-level human welfare indicator data from the County Health Rankings and Roadmaps. The 2021 County Health Rankings and Roadmaps Data includes both Ranked Measures Variables and Additional Measures variables. We:

- Joined CHR Ranked Measures and CHR Additional Measures into one dateframe and save as a CSV of all desired variables
- Selected variables of interest with the support of the Client's domain knowledge of food production and environmental justice
- Inspected NA distribution and value distributions

5.1.4 Explore patterns in distribution of human welfare indicators and environmental pressures

- First, we calculated what proportion of each environmental pressure burdens each county. In other words, what proportion of each environmental pressure is generated by each county?
  - We visualized the proportions of pressure burdens for counties across the U.S. using choropleth maps
- Then we conducted correlation analyses between environmental pressure burdens from food production and health variables to provide a local-scale comparison to future global scale analysis being conducted by the NCEAS Food Systems Working Group
- Finally, we conducted preliminary multivariate work, including methods for reducing multicollinearity among CHR variables (e.g., random forest models) and simple and multivariate regressions to explore estimated relationships between health variables and food production pressures

5.1.5 Data testing

See Summary of Testing

5.1.6 Visualize findings

See How to Use the Product

5.1.7 Communication

We conveyed our findings and discussed broader implications via:

- Public oral presentations,
- Visualizations such as maps, and
- this Technical Documentation.

5.2 Data Management Plan

5.2.1 Data description and standards

Via GitHub and the NCEAS Aurora server, the Client has provided access to raw and processed datasets of environmental pressures of global food production. Raw data are spatial in nature, formatted as 472 geoTIFF files stored on the Aurora server. These geoTIFF files globally map all major food projections and the four environmental stressors. In GitHub, the Client has RMarkdown (.Rmd) files to read the geoTIFF rasters in Aurora, resample the spatial data, and merge the geoTIFF data with a map of US FIPS county codes.

From this merged data, we will created .csv files with summary statistics for food system stressors per U.S. county. Human welfare data was downloaded from the County Health Rankings & Roadmaps website, in .csv format, and contains 257 human welfare variables.

In addition to data provided by the Client, this project also utilized multiple online data sources such as the United States Census Bureau.

5.2.2 Metadata

The main README.md file of the MEDS Capstone team GitHub repository will contain written context for the overall project (e.g., file and directory organization, project goals, authors and collaborators, etc.) to provide detailed documentation for the use, preservation, and comprehension of the data files. Each .Rmd file will also contain file-specific contextual information such as the script's objectives, workflow, and data sources.

5.2.3 Data sharing, access, intellectual property, and re-use

See the Archive Access section.

5.2.4 Data archiving and preservation

See Archive Access section.

5.2.4.1 Code

Final R code for analyses, metadata, and visualizations will be preserved in the GitHub organization repository, which will be made public after publication of the research. Relevant exploratory visualizations will be archived in the final report project documentation and on the Bren School website.

5.2.4.2 Data

Upon publication of the Client's related work, the environmental pressures of food production dataset will be archived with metadata in an online, open-access repository such as Knowledge Network for Biocomplexity (KNB).

5.3 Proposed Software and Tools

5.3.1 Software

- **Github:** Project management with Issues and Project Board, organization with Organizations, versioning with Repositories, metadata documentation with README.md
- **Google Workspace:** Capstone project meeting notes, managing tasks and deliverables, shared documents and presentations
- **NCEAS Aurora server:** Data access and storage; image storage
- **R (version 4.1.0) and R Studio (2022.2.1.461):** Programming language and environment
- **Slack:** Virtual communication and project management
- **Zoom:** Virtual communication
- **Zotero:** Citations management for documentation of background research

5.3.2 R Packages

The R packages and the specific versions used in the project have been archived using the `renv`[Ushey, 2022] package and can be found in the `renv` directory in the `foodjustice-main` repository.

Chapter 6

Products and Deliverables

6.1 Academic Deliverables

As per Master's of Environmental Data Science (MEDS) academic requirements for Capstone courses EDS 411A/B, final products include:

1. a Design and Implementation Plan;
2. a Faculty Review Presentation of the Design and Implementation Plan;
3. Technical Documentation;
4. a project repository on GitHub;
5. a final oral presentation;
6. and, upon completion of NCEAS working group research publications, the project data and metadata.

6.2 Client Deliverables

For the Client, we produced a reproducible and sustainable workflow, including R scripts and documentation of methods to manipulate and explore the NCEAS environmental pressures of food production dataset. The Client will have access to any intermediate products, including data visualizations of exploratory analyses.

Specifically, the workflow:

1. Cleans, transforms, and creates intermediate data structures to perform analyses.
2. Produces choropleth maps of the four primary environmental pressures of food production and cumulative environmental pressure throughout the United States;

3. Generates correlation plots to visualize correlations between various environmental pressures from food production and County Health Rankings & Roadmaps human welfare indicator variables;
4. And determines random-forest-generated rankings of predictor variables (including food systems pressures) to predict which variables contribute the most to variance in a County Health Rankings & Roadmaps outcome variables.

Chapter 7

Summary of Testing

This project utilized myriad tests throughout the workflow to allow current and future users to inspect intermediate data structures being built, identify common problems at weak points in the code, know where to anticipate breaks in the workflow, and verify expected outputs.

This section summarizes common tests used, which are exemplified by five notable examples, and discusses the user testing strategies implemented to ensure code reproducibility.

7.1 Common Tests

This project primarily utilizes four types of unit testing:

1. Checking transformed data values (see Example 1)
2. Checking county FIPS codes (see Example 2)
3. Checking for correct dataframe dimensions after merging (see Example 3)
4. Checking for expected data class (see Example 4)
5. Checking for the expected coordinate reference system (see Example 5)

7.2 Examples of Testing

7.2.1 Check normalized food pressure values

Source file: `data_prep_pressures_summary.Rmd`

To normalize county-level environmental pressure burdens, we followed the following steps:

1. Raw environmental pressure values were grouped by FIPS code and food group

2. Raw environmental pressure values per county were originally in the unit of each environmental pressure. For example, greenhouse gas raw environmental pressure values were in CO^2 equivalents/year

3. Each of the four pressures were summed for every U.S. county
4. Each county's individual pressure sum was divided by the national sum pressure to return pressure burden ratios per county for each of the four pressures
5. Next, we tested that:
   (a) each pressure's values for all U.S counties sum to 1, and
   (b) the cumulative pressure for all U.S counties sum to 4 to ensure we calculated the proportions accurately

```
#read in data
all_df <- vroom(here("/",
                     "home",
                     "shares",
                     "foodjustice",
                     "data",
                     "pressures_summary.csv")) %>%
  subset(pressure == "cumulative")

#check that the disturbance_sum_rescaled column adds up to approximately 1
y <- sum(all_df$pressure_sum_rescaled)

if (y >= 4.0000000001 | y <= 3.9999999990){
   stop("All rescaled pressure values should sum up to 4")
  }
```

7.2.2 Check FIPS codes

7.2.2.1 Check for misaligned FIPS codes

Source file: raw_summaries_prep.Rmd

1. The County Health Rankings data must be checked for misaligned FIPS codes by comparing FIPS codes included in the County Health Rankings dataset to FIPS codes in the U.S. county shapefile (us_county_sf)

```
#read in county health rankings data
chr_2021 <- read_csv(file = here("/",
                                 "home",
                                 "shares",
                                 "foodjustice",
                                 "data",
                                 "2021_us_county_health_counties.csv")
```

```
                     )
#identify FIPS codes for U.S. territories such as Puerto Rico
#territory FIPS codes are not of interest to this project
us_county_sf %>%
  filter(geo_id %in% missing_fips2) %>%
  select(geo_id, name) %>%
  distinct()
```

2. Compare FIPS codes between the County Health Rankings data and the U.S. Census counties shapefile. If this code chunk returns 0, then there are no discrepancies between the counties contained in the two files, and the US Census shapefile data has all the counties needed for further analyses.

```
# What fips codes are missing from the CHR dataset?
missing_fips <- setdiff(unique(chr_2021$fips), unique(us_county_sf$geo_id))

print(missing_fips)
```

3. Then, the code chunk below does the reverse - comparing FIPS in the US Census shapefile to FIPS in the County Health Rankings data and returning all the unmatched FIPS. The code chunk writes a list of these unmatched FIPS, which should be of length 91. This list includes counties from Puerto Rico, a territory that is not included in this analysis, and so the missing FIPS are not of concern for this workflow.

```
# What FIPS codes are missing from us_county_sf
missing_fips2 <-
  setdiff(unique(us_county_sf$geo_id),
          unique(chr_2021$fips))

length(missing_fips2)

#identify FIPS codes for U.S. territories such as Puerto Rico
#territory FIPS codes are not of interest to this project
us_county_sf %>%
  filter(geo_id %in% missing_fips2) %>%
  select(geo_id, name) %>%
  distinct()
```

7.2.2.2 Fix broken FIPS codes

Source file: cumulative_corr_plots_all_states.Rmd

When loading data, FIPS codes that begin with a zero often lose the leading zero. Thus, data processing ensures that each FIPS code contains 5 digits.

```
#create the data frame
pressures_wide <- summary_pressures_df %>%
  pivot_wider(names_from = pressure,
              values_from = pressure_sum_rescaled) %>%
  #ensure fips code has the correct number of digits
  mutate(fips = if_else(str_length(fips) == 4,
                        str_c("0",
                              as.character(fips)),
                        as.character(fips))) %>%
  rename(state_code = state)

#check if counties do not have FIPS codes with 5 characters
pressures_wide$county[nchar(pressures_wide$fips) != 5]
```

7.2.3 Check dimensions after merging dataframes

Source file: data_prep_chr_data_2021.RMD

After reading in and merging the County Health Rankings data, the dataframe dimensions must be checked for accuracy using the `dim()` function. The dataframe should account for 3,142 counties in the US (rows) and 257 health variables (columns).

```
health_extra <- read_csv(here("/",
                              "home",
                              "shares",
                              "food-systems",
                              "social_justice",
                              "_raw_data",
                              "US_county_health",
"2021_County_Health_Rankings_Additional_Measure_Data.csv"))

health <- read_csv(here("/",
                        "home",
                        "shares",
                        "food-systems",
                        "social_justice",
                        "_raw_data",
                        "US_county_health",
"2021_County_Health_Rankings_Data.csv"))

county <- left_join(health %>% filter(!is.na(county)),
                   health_extra %>% filter(!is.na(county)),
                   by = c("fips", "county", "state")
                   )
```

```
dim(county)
```

7.2.4 Check for the correct object class

Source file: data_prep_rgn_raw_summary.Rmd

After reading in or creating new data objects, we checked to ensure the new object was of the correct class. This is important to ensure that the new data object can be used properly in subsequent code chunks.

```
us_county_raster <- fasterize(us_county_sf,
            template_raster,
            field = "geo_id")

#us_county_raster should be a raster obeject
class(us_county_raster)
```

7.2.5 Check coordinate reference system (CRS) for rasters

Source file: data_prep_rgn_raw_summary.RMD

When working with geospatial data, we implemented checks such as examining the CRS of rasters to ensure the CRS is appropriate for the US.

```
file_list <- list.files(path = file.path(here("/",
                                              "home",
                                              "shares",
                                              "food-systems",
                                              "Food_footprint",
                                              "all_food_systems",
                                              "analysis",
                                              "raw")),
                        full.names = TRUE)

template_raster <- raster(file_list[2])

#check template raster CRS
print(crs(template_raster))

#check census shape file CRS
print(crs(us_county_sf))
```

7.3 User Testing

To test code reproducibility, the MEDS Capstone team engaged itself and the Client NCEAS in several iterations of user testing for the `foodjustice-main` repository. First, each member of the MEDS Capstone team executed the .Rmd files in their entirety to check for code breakages due to common errors such as missing files or misnamed variables. The .Rmd files in the `data_prep` directory were run first followed by .Rmd files in the `analysis` directory. During this process, scripts were run both on Mac and PC computers.

After initial internal checks and code revision as necessary, the MEDS Capstone team submitted the following .Rmd files to the NCEAS team for a second check:

- `data_prep_rgn_raw_summary.Rmd`
- `data_prep_pressures_summary.Rmd`
- `cumulative_corrplots_all_states.Rmd`
- `flex_dashboard.Rmd`
- `mapping_food_pressures.Rmd`
- `random_forest_cumulative_pressure.Rmd`
- `random_forest_health_predictors.Rmd`
- `random_forest_variable_importance_food_type.Rmd`

The NCEAS team encountered some issues in `random_forest_cumulative_pressure.Rmd` due to missing files, which the MEDS Capstone team had previously reshuffled and renamed, in Aurora. Nevertheless, the NCEAS team successfully ran every other script. The identified issues were remedied by the MEDS Capstone team.

Chapter 8

User Documentation

This section details how to use, contribute to, and maintain the products and deliverables produced from the *Assessing potential environmental justice implications of U.S. food production* Capstone project.

8.1 Intended Audience

This code, workflow, and data are intended for use primarily by the National Center for Ecological Analysis & Synthesis (NCEAS) Environmental Impact and Sustainability of Global Food Systems working group and other NCEAS-affiliated teams or individuals working on food systems research. After publication, these products will also be accessible to entities outside of NCEAS that interested in the work and/or implementing similar analyses.

8.2 Data Infrastructure and Organization

This project calls on raw and processed data formats. Data are stored in Aurora or pulled from open-access online sources in the code. Data structural and organizational details are as follows:

8.2.1 Raw Data

- **Environmental pressures of food production**: This data is owned by NCEAS and is stored in the `food-systems` directory of the Aurora server at the end of the file path: `/home/shares/food-systems/Food_footprint/all_food_systems/analaysis/raw`. In this directory, the data is stored in folders in raw and intermediate files. This project calls on 472 geoTIFFs. Each geoTIFF file describes the global spatial distribution of one of the four pressures of environmental food production for a different food system type.

- **County Health Rankings (CHR) and Roadmaps 2021:** This data is stored in the `social-justice` directory of the Aurora server at the file path `/home/shares/food-systems/social_justice/_raw_data` . This directory contains two .csv files from the County Health Rankings and Roadmaps website:
  - **County Health Rankings Data:** CHR data contains 30+ metrics for measuring the health of each U.S. county's respective residents. Observations are at the county level, containing 3,142 observations for every U.S. county.
  - **Additional Measures Data:** 111 additional human welfare metrics to supplement the main CHR data measures.
- **U.S. Census shapefile:** This file is used to calculate zonal statistics for counties with food pressures geoTIFF. The U.S. Census provides shapefiles from the Census Bureau's MAF/TIGER geographic database. The `raw_summaries_prep.Rmd` workflow file downloads a linked zip file from the U.S. Census website, which provides the 2019 counties shapefile, then deletes the zip file.
- **U.S. Census Geo Regions:** The U.S. Census defines region and division groups to states proximate to one another. The defined regions are in PDF format. A .csv (`us_state_regions.csv`) was downloaded from GitHub and cross referenced with the PDF resource from the U.S. Census to ensure it was an accurate tabular form of the U.S. Census data. The .csv file is stored in the `foodjustice` data directory in Aurora.

8.2.2 Processed Data

- **County Health Rankings (CHR) and Roadmaps 2021 Joined Data (`chr_data_2021`):** This .csv contains the joined CHR datasets with geo-regions added from the raw U.S. Census Geo Regions data. It contains 260 columns and 3,142 observations. The .csv file is stored in the `foodjustice` data directory in Aurora.
- **Food pressures by location and food system (`rgn_raw_summary.csv`):** This .csv contains raw food pressure values (not rescaled from 0 - 1) for each US county organized by food-related variables such as category, organism, and system. It contains 10 columns and 1,516,536 observations. The .csv file is stored in the `foodjustice` data directory in Aurora.
- **Rescaled food pressures summarized by location and food group (`pressures_summary.csv`):** This .csv contains rescaled (aka proportions of) environmental food pressures due to food production for each US county. Values range from 0 - 1. It contains 6 columns and 227,265 observations. The .csv file is stored in the `foodjustice` data directory in Aurora.

8.2.3 GitHub

The foodjustice GitHub organization contains three repositories:

8.2.3.1 A) foodjustice-main

This is the primary repository and houses the data preparation, analysis, and workflow. It is divided into three main directories:

1. `data_prep`: data cleaning, transformations, and other manipulations to prepare raw data and create intermediate datasets before analysis. For more detailed information, see Processed Datasets.

- `us_county_health`:
  - `data_prep_chr_data_2021.Rmd`: reads in, joins, and manipulates County Health Rankings & Roadmaps welfare indicator data.
- `us_food_pressures`:
  - `data_prep_rgn_raw_summary.Rmd`: reads in global environmental footprint of food production data and stacks it with a U.S. counties raster file from the US Census because we are only interested in the environmental pressures from food production in the U.S.
  - `data_prep_pressures_summary.Rmd`: creates processed and ready-to-use data in .csv format that summarizes data from the NCEAS environmental footprint of food production dataset; summarizes environmental pressure values as separate observations for all pressures for all food systems for all U.S. counties.
  - `data_prep_rescale_values_us.Rmd`: normalizes all four food systems pressures to proportions ranging from 0 - 1

2. `analysis`: mapping, correlation plots, flex dashboard, random forest regressions

- `choropleth_maps`: For more detailed information, see Choropleth Maps.
  - `mapping_cumulative_pressure_by_food_system.Rmd`: creates choropleth maps to visualize the spatial distribution of environmental food pressures of certain food systems across U.S. counties
  - `mapping_food_pressures.Rmd`: creates choropleth maps to visualize the spatial distribution of environmental food pressures across U.S. counties
- `corrplots`: For more detailed information, see Correlation Plots.
  - `cumulative_corrplots_all_states.Rmd`: creates correlation plots of correlation values between health outcome variables and cumulative pressure values from the environmental footprint of food production in each state
- `flex_dashboard`:

    - `flex_dashboard.Rmd`: creates the flexdashboard used to visualize random forest results

- `random_forests`: For more detailed information, see Random Forest.

    - random_forest_health_predictors.Rmd:
    - random_forest_variable_importable_food_type.Rmd:
    - Your_forest_explained.html:

3. `renv`: stores data archiving R and R package versions and contains `activate.R` script to activate the preserved environment

More details about each directory, subdirectory, and .Rmd file can be found on GitHub in README.md files.

8.2.3.2 B) foodjustice-issues

This repository is used to create the GitHub project board for internal project management and organizational purposes.

8.2.3.3 C) foodjustice_tech_doc

This repository creates the `bookdown` project to build and update the Technical Documentation (TD). The title page is created using the `index.Rmd` file. Each chapter of the TD can be found in a separate .Rmd file that follows the naming convention: `#-section-name.Rmd` (e.g., `01-abstract.Rmd`). TD sections can be updated in the appropriate .Rmd file, built out in RStudio using the bookdown package [Xie, 2021], then output as a gitbook or pdfbook file.

The `_book` directory contains the html files used to create the gitbook, images, and data used to create tables displayed in the final PDF output. The images and data are not hosted on GitHub and are stored locally. The images are found at the following file path: `_book/_main_files/figure-html/`, where the `figure-html` directory contains four subdirectories.

1. `choro_images`: PNG files of choropleth map outputs
2. `corr_impages`: PNG files of corrplot outputs
3. `data`: .csv files used to create data tables
4. `rf_images`: PNG files of random forest variable importance plot outputs

Images and data are stored locally and not on GitHub.

8.2.4 Aurora server

The NCEAS' private Aurora server contains much of the data and scripts used by various NCEAS research teams. The products from this Capstone utilize two main directories in the `/home/shares/` directory of Aurora:

8.2.4.1 `/home/shares/foodjustice/`

This folder contains data and images too large to be pushed to and stored on Github. Data and images within the `foodjustice` directory are all products of the foodjustice MEDS Capstone team. The folder infrastructure is as follows:

- `data`: .csv files of raw and intermediate data commonly used in analyses
- `visualizations`: final visualization outputs

8.2.4.2 `/home/shares/food-systems/`

This folder in Aurora contains many raw and intermediate data files and analysis files associated with NCEAS' various data analysis teams. Much of the MEDS Capstone student team's work and file organization drew on preliminary NCEAS analysis scripts stored in the `social_justice` directory, which contains the following sub-directories:

- `_analysis`
- `_data_layers`
- `_prep_files`
- `_raw_data`

A more detailed description of the `social_justice` directory's file organization can be found on the NCEAS OHI_Science GitHub.

8.3 Code Maintenance

All R code is hosted on GitHub with access provided to the Client to make changes and updates moving forward after the Capstone project has concluded. GitHub version control gives current and future code managers access to prior iterations of the code in order to inform code maintenance in the future.

The `renv` package was used to document and archive the specific versions of R and R packages used during this project. This information can be found in the `renv` directory and `renv.lock` files in the `foodjustice-main` and `foodjustice-tech_doc` GitHub repositories.

In the future, it is likely that R packages and/or the version of R will need to be updated.

8.4 Data & Metadata Access: Raw Data

8.4.1 Environmental Pressures of Food Data

8.4.1.1 Internal Client access

Prior to the release of a NCEAS research publication utilizing the environmental pressures of global food systems data, the data will remain privately stored on the NCEAS Aurora server under the directory `/home/shares/food-systems/Food_footprint/all_food_systems/` directory. The Aurora server is accessible to NCEAS staff and affiliates. Similarly, until publication, the code will remain privately stored on GitHub, but will become publicly accessible upon publication.

For more information, see Archive Access.

8.4.1.2 External access

After publication, final data and metadata will become publicly available on an online, open-access data repository such as the Knowledge Network for Biocomplexity (KNB). Similarly, the code will become publicly available on GitHub. Intermediate data produced from this Capstone project will likely not be formally archived in an online repository.

For more information, see Archive Access.

8.4.2 County Health Rankings and Roadmaps Data

Access to County Health Rankings variables used as our human welfare indicators at the county level are publicly available through the County Health Rankings and Roadmaps 2021 Measures web page. We used data from 2021, which, after future iterations of CHR datasets, will be archived with other previous years' data.

8.5 Data & Metadata Access: Processed Data

8.5.1 Environmental Pressures of Food Data

8.5.1.1 Internal Client access

Processed datasets used in workflows developed as a part of the foodjustice MEDS capstone project are available in the NCEAS server, Aurora, and private to NCEAS until publication of research using the raw environmental pressures of food production data. The processed data can be found in the `foodjustice` Aurora directory, which can be found in the file path `/home/shares/foodjustice/data/pressures_summary.csv`. Documented processing scripts are accessible via the private repository's `data_prep` folder, in the file, `/data_prep/us_food_pressures/data_prep_pressures_summary.Rmd`.

For more information, see Archive Access.

8.5.1.2 External Access

Access to the processed food pressures by U.S. county datasets will be made available upon publication by the Client.

8.5.2 County Health Rankings Data

8.5.2.1 Internal Client Access

The processed CHR dataset developed as a part of the foodjustice MEDS capstone project are available in the NCEAS server, Aurora, and private to the Client. The processed data can be found in the `foodjustice` Aurora directory, which can be found in the file path `/home/shares/foodjustice/data/chr_data_2021.csv`. Documented processing scripts are accessible via the private repository's `data_prep` folder, in the file, `/data_prep/us_food_pressures/us_county_health_data_prep.Rmd`.

For more information, see Archive Access.

8.6 Product Development/Reproducibility

Learn more about product development in the Solution Design section. Learn more about product reproducibility in the User Testing section.

8.7 How to use the product

8.7.0.1 Processed Datasets

To complete exploratory analyses of the Client's environmental footprint of food production data set using county-level calculations, we developed processed datasets that are instrumental to the overall workflow.

This project contributes two main processed datasets:

A) Raw Summary Data

This data is processed in the `data_prep_rgn_raw_summary.Rmd` in the project repository directory `/data_prep/us_food_pressures/data_prep_rgn_raw_summary.Rmd`.

It calls on 472 raw geoTIFF files from the Client server Aurora in the directory `/home/shares/food-systems/Food_footprint/all_food_systems/analysis/raw`. A for loop loops through each file, which maps the environmental footprint of food productions for a unique combination of environmental pressure and food system type and creates a row in a dataframe with zonal statistics for every county in the United States.

The resulting data contains 10 variables, including location character variables, food production classification variables, and the summed pressure for that place and food system in a variable called "sum."

As an example, the first five observations are shown in the table below:

county	fips	state	category	origin	organism	system	product	pressure	sum
Autauga County	01001	AL	farm	land	bana	crop	produce	disturbance	0
Baldwin County	01003	AL	farm	land	bana	crop	produce	disturbance	0
Barbour County	01005	AL	farm	land	bana	crop	produce	disturbance	0
Bibb County	01007	AL	farm	land	bana	crop	produce	disturbance	0
Blount County	01009	AL	farm	land	bana	crop	produce	disturbance	0

Metadata for the processed dataset is included at the bottom of the data processing .Rmd file in Markdown text—including the column names, their class, distribution or distinct values, and a description of the variable represented in that column.

B) Food Pressures Summary

This data is processed in the R Markdown file `data_prep_pressures_summary.Rmd` in the project repository directory `/data_prep/us_food_pressures/data_prep_pressures_summary.Rmd`.

This file aggregates observations from the raw summaries data into one observation per "supra group" defined by different combinations of food production classification variables in the raw summaries data. Each county in the U.S. gets an observation for all 5 environmental pressures of food production for every supra group food production type. The values are rescaled in this dataset so that every pressure value represents the county's pressure burden, or the county's proportion of the total pressure exerted by food systems across the U.S. In other words, each pressure was normalized as a proportion to range from 0 - 1. Each pressure was then summed to create the cumulative food pressure values, which ranged from 0 - 4 for each county.

As an example, the first five observations are shown in the table below:

county	fips	supra_group_legend	state	pressure	pressure_sum_rescaled
Abbeville County	45001	Backyard Pig Meat	SC	disturbance	0.0e+00
Abbeville County	45001	Backyard Pig Meat	SC	ghg	1.0e-07
Abbeville County	45001	Backyard Pig Meat	SC	nutrient	2.3e-06
Abbeville County	45001	Backyard Pig Meat	SC	water	0.0e+00
Abbeville County	45001	Cereals and Grains	SC	disturbance	0.0e+00

Metadata for the processed dataset is included at the bottom of the data processing Rmarkdown file in Markdown text—including the column names, their class, distribution or distinct values, and a description of the variable represented in that column.

8.7.0.2 Choropleth Maps

To visualize the spatial distribution of the four main environmental pressures and cumulative pressure from food production, we created 5 choropleth maps of the United States.

Source file: `mapping_food_pressures.Rmd`

- read in the intermediate `pressures_summary.csv` processed datatset
- create one choropleth map per each of the 4 primary environmental food production pressures and one map for cumulative environmental food production pressure using `plot_usmap()` from the `usmap` package [Di Lorenzo, 2022]
- winsorize values at the 90th, 95th, and 99th percentiles to better visualize differences in pressure values among counties using a custom function
- saves choropleth map outputs into the `visualization` sub-directory in Aurora

```r
#winsorize at the 90th, 95th, and 99th percentiles

#identify the quantiles of interest
quantile(cumulative_df$cumulative_sum_rescaled, c(0.90)) #90th percentile:
quantile(cumulative_df$cumulative_sum_rescaled, c(0.95)) #95th percentile:
quantile(cumulative_df$cumulative_sum_rescaled, c(0.99)) #99th percentile:

#create a function to  winsorize and
#to create numeric vectors for each level of winsorization
fun <- function(x, y){
    quantiles <- quantile(x, y) #choose the percentile
    x[ x > quantiles[1] ] <- quantiles[1]
    x
    }

cumulative_winsor_90 <- fun(x = cumulative_df$cumulative_sum_rescaled,
                            y = 0.90)
cumulative_winsor_95 <- fun(x = cumulative_df$cumulative_sum_rescaled,
                            y = 0.95)
cumulative_winsor_99 <- fun(x = cumulative_df$cumulative_sum_rescaled,
                            y = 0.99)

#add winsorized values to the dataframe to plot
cumulative_winsor <- cumulative_df %>%
  mutate(ninety = cumulative_winsor_90,
         ninety_five = cumulative_winsor_95,
         ninety_nine = cumulative_winsor_99)
```

8.7.0.3 Correlation Plots

The correlation plots were built to explore all at once the correlation coefficients between many environmental pressures of food production variables and human welfare indicator variables using county-level observations. When considering how environmental pressures from food production may impact public health and well being, visualizing basic patterns of correlation between food production and human welfare variables is helpful. Correlation patterns can help us select variables to include when building and selecting models.

This builds on similar work that the Client has previously done at the global scale, but uses different human welfare indicator variables that are more appropriate for a county-level analysis based in the United States.

The correlation plots workflow can be found in the `analysis` directory of the project GitHub repository, under `analysis/corrplots/cumulative_corrplots_all_states.Rmd`. The workflow is documented in the .Rmd and described in detail below:

Workflow: Libraries

The workflow for building correlation plots uses the following packages:

- `tidyverse` [Wickham, 2021b],
- `ggcorrplot` [Kassambara, 2019],
- `ggplot2` [Wickham et al., 2021],
- `here` [Müller, 2020],
- `janitor` [Firke, 2021], and
- `forcats` [Wickham, 2021a].

Workflow: data

The correlation plots workflow uses the `pressures_summary.csv` dataset and the `chr_data_2021.csv` dataset. Both are stored in Aurora.

Workflow: Data processing

Before building the correlation plots, the data must be processed to include both human welfare indicator data from the County Health Rankings dataset and processed environmental pressures of food production data. The data are at the county level.

Data processing is documented in `cumulative_corrplots_all_states.Rmd` in the project repository directory `/analysis/corr_plots`.

This workflow includes...

- ...pivoting the pressure values wider in from the pressures summary data so that each pressure becomes a column with the pressure values per county as observations

- ...selecting variables of interest from the County Health Rankings dataset. The variables utilized in this workflow were first selected by the MEDS Capstone team using prior knowledge about food systems and public health and well being. Variables were then approved by the Client. The variables selected include:
  - **population**
  - **food environmental index**
  - **life expectancy**
  - **median household income**
  - **deaths**
  - **years of potential life lost**
  - **percent fair poor health**
  - **average number of physically unhealthy days**
  - **percent low birth weights**
  - **air pollution average daily pm 2.5**
  - **percent child poverty**
  - **child mortality rate**
  - **infant mortality rate**
  - **percent frequent physical distress**
  - **percent frequent mental distress**
  - **percent adults with diabetes**
  - **percent food insecure**
  - **percent limited access to healthy food**
  - **percent uninsured adults**
- ... join food pressures data with pressure values as columns and County Health Rankings data with selected variables of interest by FIPS code to create a data frame with county level observations for both food pressures and human welfare indicator variables.

To learn more about the each variable, please visit the County Health Rankings website.

Workflow: Correlation matrices

The joined data are then used to create correlation matrices to visualize the calculated correlation values and signs between every combination of each environmental footprint of food production variable and each human welfare variable of interest.

The correlation matrix uses the cumulative pressure from food production by state on the X axis and human welfare variables on the Y axis. This choice was made because it allowed the Client to visualize regional differences in correlation values between cumulative environmental pressure from all varieties of food production with each human welfare variable across all states.

First, using the function `cor()` a correlation matrix is generated with correlation values for every combination of variables, including inverse and 1:1 pairings of variables. In addition, `cor_pmat()` creates a matrix of p-value for the correlation matrix generated:

```
#correlation matrix
correlation_df <- cor(all_data_fips_cumulative[,3:73],
#subset for all the pressure + CHR vars
                      method = "spearman",
                      use = "pairwise.complete.obs")

#corresponding p-values
p.mat <- cor_pmat(all_data_fips_cumulative[,3:73],
                  method = "spearman",
                  use = "pairwise.complete.obs",
                  sig.level = "0.9",
                  exact = FALSE)
```

Next, the correlation data frame is subsetted to include only the correlation pairings of interest. For example, the fixed data frame includes the correlation coefficient for cumulative pressure in Massachusetts and low birth weight rate, but not cumulative pressure in Massachusetts and cumulative pressure in Massachusetts or the inverse pairing, low birth weight rate and cumulative pressure in Massachusetts. The p-value matrix is also fixed:

```
#fix dataframe
correlation_df_fix <- correlation_df[1:19, 20:71]

#fix p-value dataframe
p.mat_fix <- p.mat[1:19, 20:71]
```

Workflow: Plot the correlation matrix

Using the `ggcorrplot()` function, which utilizes `ggplot2()`, the workflow visualizes the fixed correlation matrix and corresponding p-value matrix for the Client's correlation coefficients of interest.

```
ggcorrplot(correlation_df_fix,
           method = "square",
           colors = c("#0086a8", "white", "#a00e00"),
           digits = 2,
           lab = TRUE,
           lab_size = 3,
           p.mat = p.mat_fix,
           pch = 4,
           insig = "pch",
           outline.color = "white")
```

The correlation matrix output is visualized as a colored grid with environmental pressure variables on the X axis and human welfare indicator variables on the Y axis. Grids where variables from the two axes meet give the following information about the correlation between the two variables:

- **Correlation value:** indicated by a number inside the grid where the two variables meet
- **Correlation sign:** indicated by the color of the grid, where positive correlation values are green and negative correlation values are red
- **Correlation magnitude:** indicated by the color hue of the grid, with darker greens representing more positive correlations than lighter greens, less positive correlations
- **Statistical significance:** using a significance level of 0.1, is the returned correlation statistically significant? An X in the grid indicates no.

Final results of corrplots of value to the Client can be viewed here and are stored in Aurora at the end of the workflow in the `/home/shares/foodjustice/visualizations/corrplots_all_states` directory.

8.7.0.4 Random Forest

Random forest regressions were used to compare how influential the cumulative pressure of food production and how influential pressures from distinct food systems (such as soy, eggs, or livestock) are in predicting a human welfare outcomes when compared to other welfare predictors.

It is important to note that random forest models are used to investigate *predictions* and not *relationships*. Random forests are an effective method for this Capstone project because collinearity of predictor variables does not negatively impact random forests' predictive performance. Because of the exploratory and novel nature of this project, many possible predictors were included in each model, and Random Forest uses any of the correlated features.

However, it is still difficult to be sure of which correlated features are causal. This is important to keep in mind when interpreting the model results because the model can incorrectly rank variable importance and drop the actually important collinear feature. The risk of this is mitigated by the random selection of variables at each node of a random forest but is still an important consideration in the methods used.

The scripts can be found in the `analysis/random_forests` directory:

- `random_forest_health_predictors.Rmd`: uses `randomForest()` and `explain_forest()` with the `pressures_summary.csv` and `chr_data_2021.csv` datasets to compare how cumulative pressures from food production ranks with other predictor variables in predicting six human health outcome and explains the workflow for future users tor replicate with other predictors and response variables of interest

- random_forest_variable_importance_food_type.Rmd: uses randomForest() with the pressures_summary.csv and chr_data_2021.csv datasets to compare how from district types of food production (ie soy, eggs, livestock) ranks with other predictor variables in predicting two human health outcome and explains the workflow for future users tor eplicate with other predictors and response variables of interest

- Your_forest_explained.html: a graphical summary of one random forest outputs from the random_forest_health_predictors.Rmd created using the randomForestExplainer package [Paluszynska et al., 2020].

Workflow: Execution

For both random forest .Rmd files in the analysis directory, the investigation of each model follows this general pattern to understand how pressures from food production in counties ranks in predicting the response variable of interest, which is the part of the model that changes throughout the document.

1) **Process data**: Import/join/clean data containing county health variables and measures of the environmental pressures from food production by food type. Convert characters to factors so that the data are compatible with the randomForestSRC [Ishwaran and Kogalur, 2022] package.

2) **Split data**: Next, for each model, data is split into training and testing sets so that the results of the random forest model predictions based on the training data can be compared to the testing data.

3) **Model data**: A random forest model is built with the RandomForest() function to explore possible predictors of different human welfare outcomes. An example looks like this (note that the code chunk below contains pseudo-code for the myriad variables input into each model):

```
#random forest model for the response variable of % low birthweight
rf_per_low_birthweight <- randomForest(percent_low_birthweight ~
                 human_welfare_indicator_variables +
                 food_type_variables +
                 cumulative_food_pressure_variable +
                 racial_demographic_variables,
  data = trainset_lbw,
  importance = TRUE,
  na.action = na.omit)
```

4) **Outputs**: The outputs of the model built using the training data help us understand how well pressures from food production and County Health Rankings variables predict a human welfare outcome (in this case, percent of infants born with low birthweights). Different output values include the percent of the response variable's variation explained, the residuals, and variable importance plots.

5) **Measure Prediction Performance:** The testing data is used to evaluate the model's ability to predict the human welfare outcome of interest for a county. The predicted model outputs are compared to the actual outputs in the testing dataset. A mean squared prediction error and it's 95% confidence interval are calculated.

8.8 Results

8.8.0.1 Processed intermediate datasets

Creating intermediate datasets that contains normalized pressure values allows us to more easily comprehend and compare different pressures across counties.

To read more about the processed data and observe example tables from the datasets, please see the Data Infrastructure and Organization and Data & Metadata Access sections.

8.8.0.2 Choropleth Maps

First, we mapped raw values for all four environmental pressures and cumulative pressure from food production.

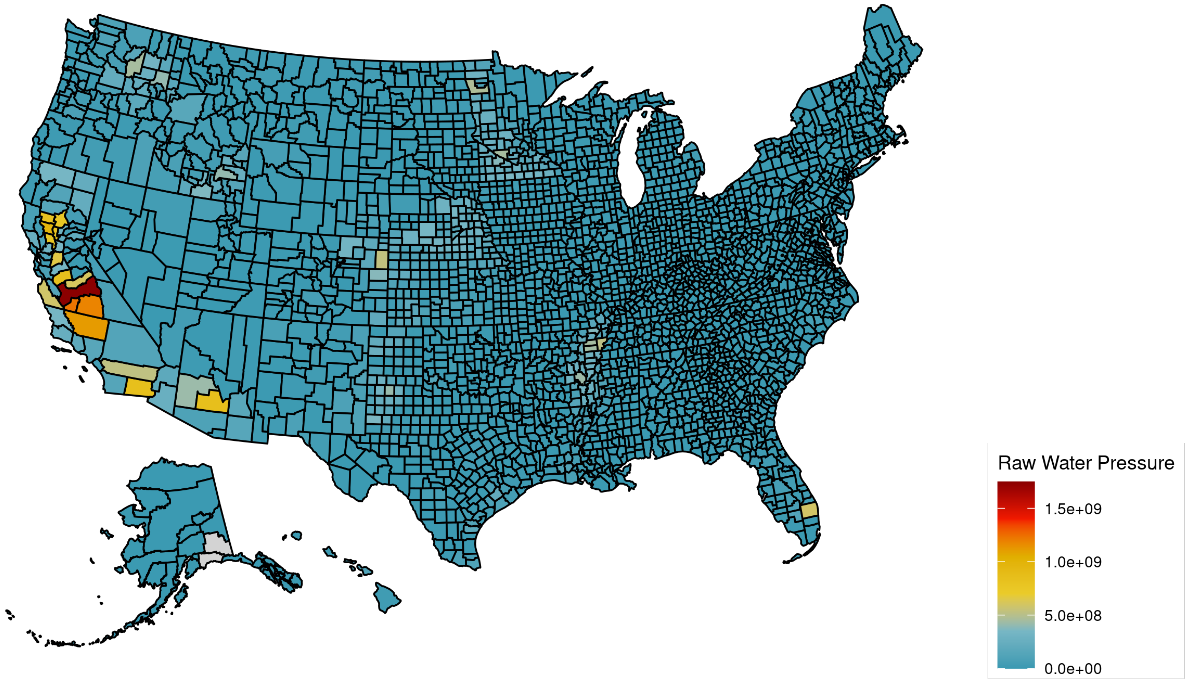


Figure 8.1: Raw values of environmental pressure due to water consumption before normalization.

Then we normalized pressure values from 0 - 1.

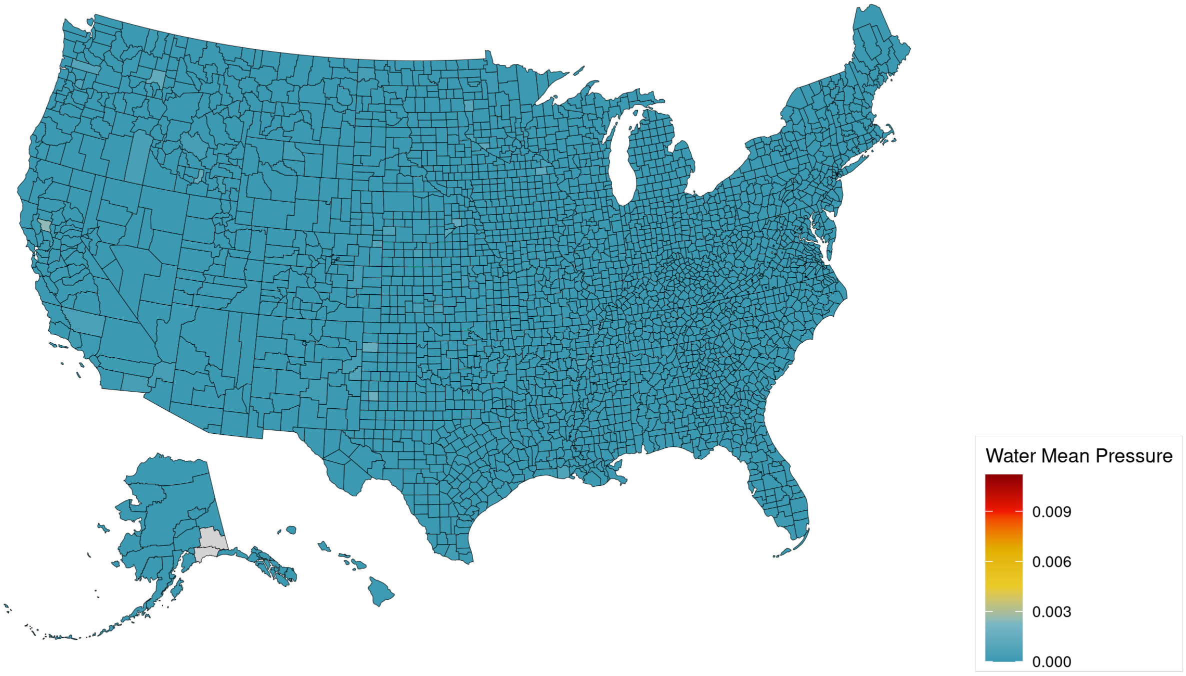


Figure 8.2: Normalized mean values of environmental pressure due to water consumption.

Due to outlier data points, differences in environmental pressures between counties were difficult to visualize. Thus, we winsorized each food pressure and the cumulative pressure at the 99th, 95th, and 90th percentile.

After winsorization, it became easier to distinguish which counties contain relatively higher environmental and cumulative pressures. However, few distinct patterns arise. For example, for greenhouse gas pressures due to food production, counties all across the country demonstrate high values depending on the winsorization level.

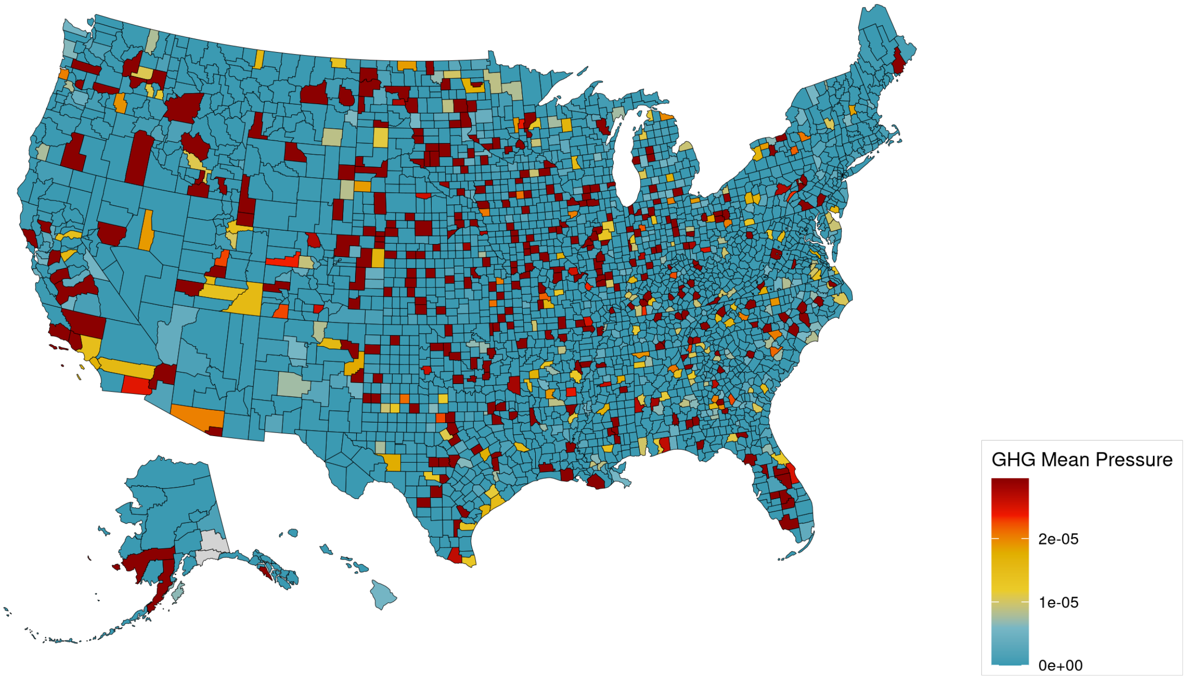


Figure 8.3: Greenhouse gas mean pressure winsorized at the 90th percentile.

To further examine the spatial environmental pressures from food production at the U.S. county level, we also created supplementary visualizations which map the cumulative pressure of food production for each of the 21 types of food production in the Client's dataset. They are available in the `choropleth_maps_by_food_system` folder of the `foodjustice visualizations` directory of the Client's Aurora server.

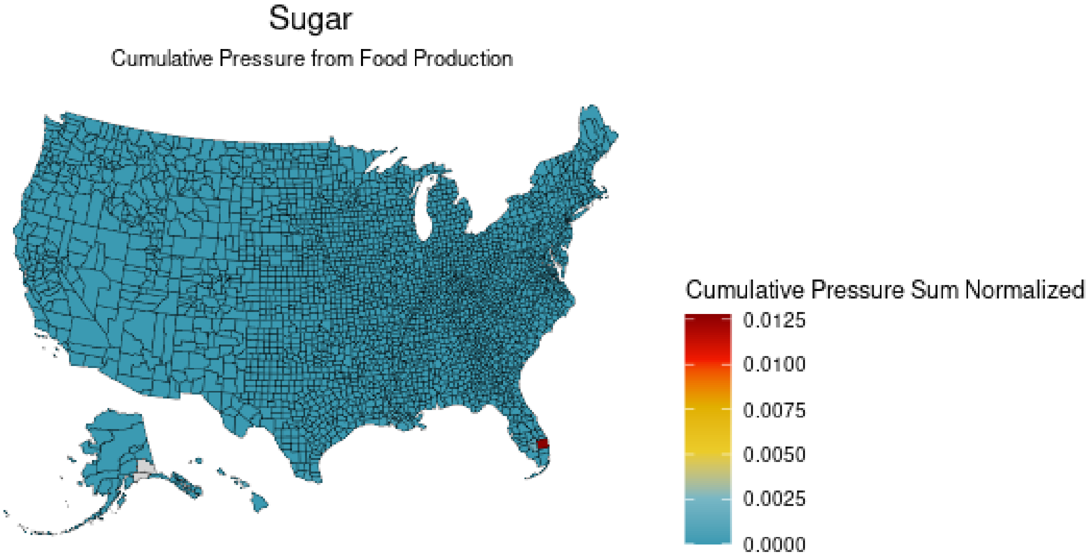
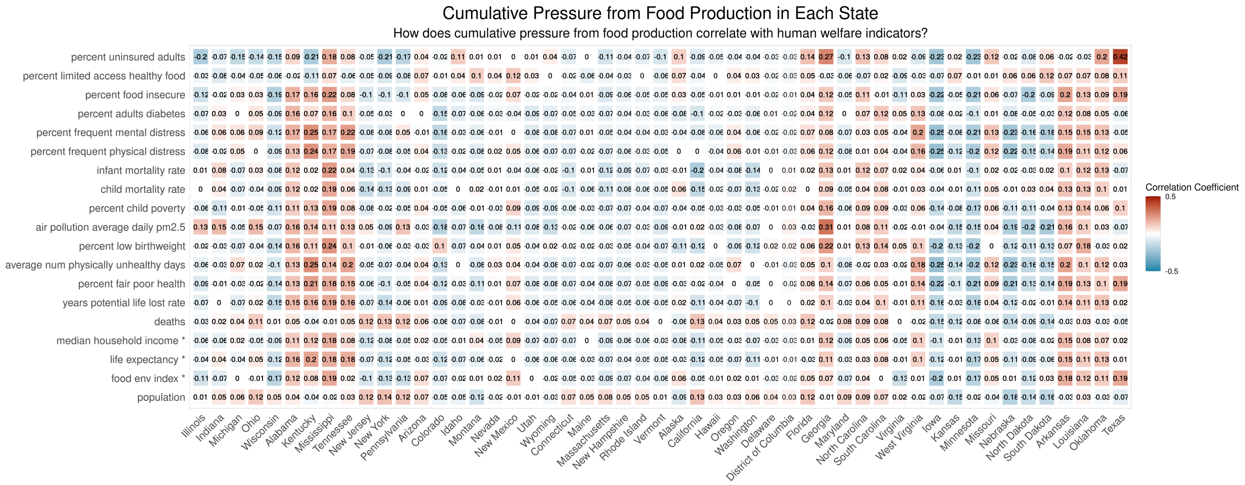

Sugar
Cumulative Pressure from Food Production
Cumulative Pressure Sum Normalized
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000

Figure 8.4: Normalized cumulative pressure from sugar production.

8.8.0.3 Correlation Plots



Cumulative Pressure from Food Production in Each State

How does cumulative pressure from food production correlate with human welfare indicators?

Human warfare variables with asterisk* indicate values multiplied by -1 so that correlation coefficient sign represent outcomes similar in meaning to the sign of other variables (ie poor health outcomes are correlated with high pressures).

Figure 8.5: Correlation plot for all 50 states and 19 human health variables.

The final correlation plot developed from this workflow, developed after many iterations of regional scales used on the X axis, allows the Client to visualize regional differences between the correlation of cumulative pressure in a given state and the human welfare outcomes in that state consistently across the selected human welfare indicator variables included.

While the correlation between pressure magnitude across a state and health outcomes is not meaningful for assigning cause or identifying environmental injustice, this exploration is a foundation to future Client work to identify patterns of environmental food production in the U.S.

To further examine regional differences in human welfare outcomes and their co-occurence with environmental pressures from food production, supplementary products include additional correlation plots with the cumulative pressure values filtered to a single type of food system for each plot. This allows the Client and future users to examine georegional differences in, for example, cumulative pressure from egg production. These additional visualizations can be viewed in the `foodjustice visualizations` directory of the Aurora server in the `food_systems_corrplots` folder.

8.8.0.4 Random Forest

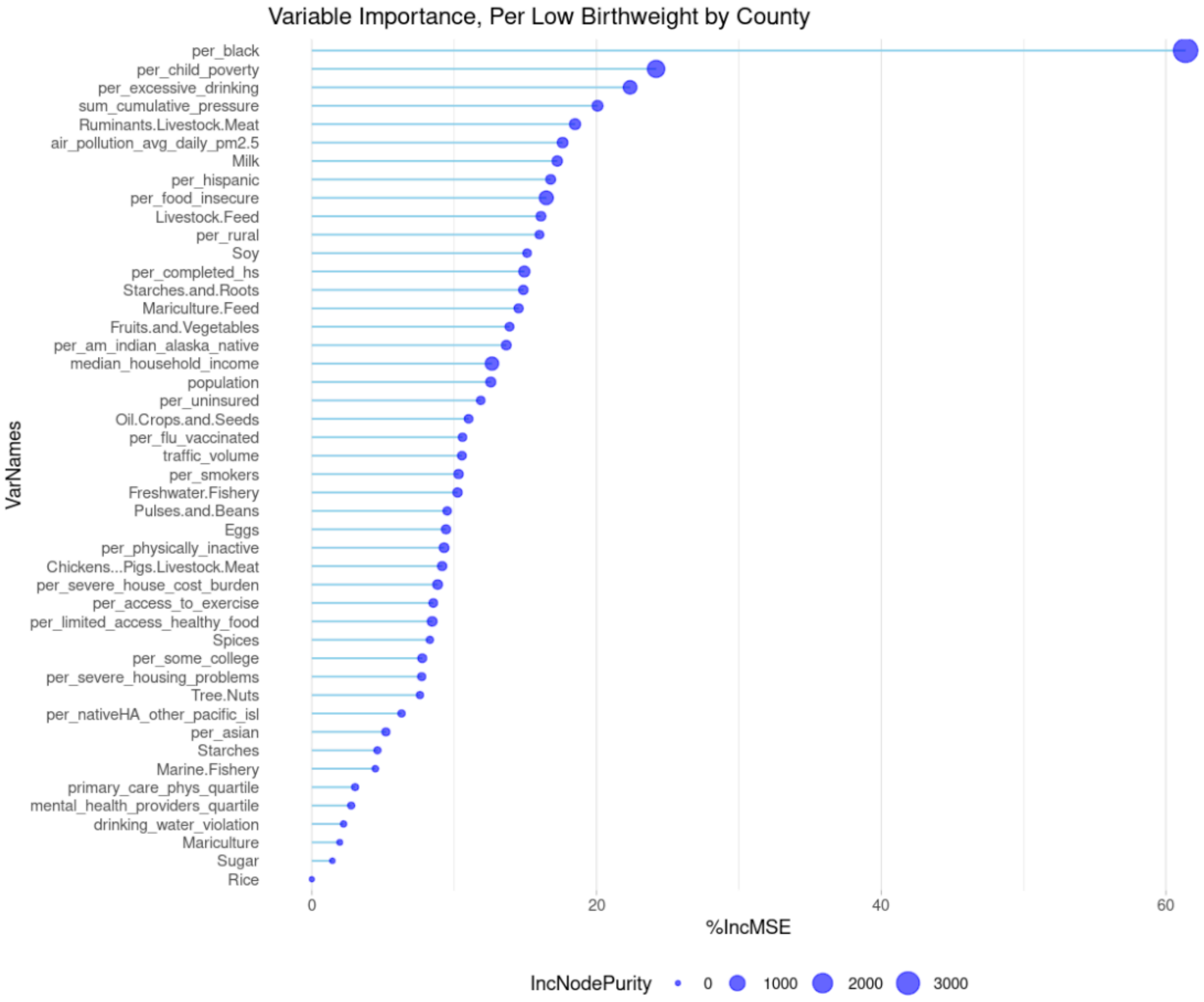


Figure 8.6: Random forest variable importance plot. In this iteration of the model, percent of Black individuals in a county was the top predictor for the response variable of percent of infants with low birthweights per county.

The random forest analysis offers outputs for 8 different random forest models. The .Rmd files also guide the Client and future users in building more models based on their own research questions and knowledge of U.S. food production environmental stressors.

The models that include `sum_cumulative_pressure` (our measurement of the environmental footprint of food production in a county including: spatial disturbance, excess nutrients, water use, and green house gas emissions) as a predictor show that the `sum_cumulative_pressure` variable's importance varies as random forest models are run. Despite this variation, predictive accuracy was tested using a testing dataset. The mean squared prediction error was small for most models and their testing data, except for the model with `years_potential_life_lost_rate` as a response variable. The models that include pressure variables from distinct types of food production (e.g., `eggs`, `chicken_pigs_livestock_meat`, or `milk`) did not show individual types of food production type pressures as important variables in predicting health outcomes compared to other predictors from the human welfare data.

Chapter 9

Archive Access

This documentation is archived on the Bren School of Environmental Science & Management's website, where this project's proposal and deliverables will be available.

In alignment with the mission of the Client, the data and analysis of this Capstone project is intended to be as public and open-access as possible. However, due to the unpublished status of this research, the code and Client data cannot be made publicly available until after publication. After the completion of this project, intermediate data and metadata will be archived in the Client server, Aurora. Final data, metadata, and all related documentation emerging from this project will be available for public viewing and use upon publication of NCEAS research using this data.

To gain access to the Aurora server, you may first contact NCEAS' Executive Director, Dr. Ben Halpern (halpern[at]nceas[dot]ucsb[dot]edu), for permissions before reaching out to NCEAS' System Administrator, Nick Outin (outin[at]nceas[dot]ucsb[dot]edu), to arrange SSH access and further instructions.

Human welfare indicator data comes from open access and publicly available sources. The MEDS Capstone team will have the rights to redistribute the raw data as well as any intermediate and transformed versions.

The foodjustice GitHub organization can be found here.

Bibliography

Christopher Bene, Steven Prager, Harold Achiconoy, Patricia Alvarex Toro, Lea Lamotte, Camila Bonilla, and Brendan Mapes. Global map and indicators of food system sustainability. *Sci Data*, 6(279), 2019.

James Boyce. Inequality as a cause of environmental degradation. *Ecological Economics*, 11 (3):169 – 178, 1994.

Cercedes Campi, Marco Dueñas, and Giorgio Fagiolo. Specialization in food production affects global food security and food systems sustainability. *World Development*, 141, 2021.

J Conjin, P Bindraban, J Schroder, and R Jongshaap. Can our global food system meet food demand within planetary boundaries? *Agriculture, Ecosystems and Environment*, 251:244 – 256, 2018.

Paolo Di Lorenzo. *usmap: US Maps Including Alaska and Hawaii*, 2022. URL https://CRAN.R-project.org/package=usmap. R package version 0.6.0.

Nina Domingo, Srinidhi Balasubramanian, and Sumil Thakrar. Air quality–related health damages of food. *PNAS*, 118(20), 2021.

Sam Firke. *janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2021. URL https://github.com/sfirke/janitor. R package version 2.1.0.

Charlotte Glennie and Alison Hope Alkon. Food justice: cultivating the field. *Environmental Research Letters*, 13(7), 2018.

H. Ishwaran and U.B. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2022. URL https://cran.r-project.org/package=randomForestSRC. R package version 3.1.0.

Abdulrahman Jbaily, Xiaodan Zhou, Jie Liu, Ting-Hwan Lee, Leila Kamareddine, and Stephane Verguet. Air pollution exposure disparities across us population and income groups. *nature*, 601(7):228 – 233, 2022.

Alboukadel Kassambara. *ggcorrplot: Visualization of a Correlation Matrix using ggplot2*, 2019. URL http://www.sthda.com/english/wiki/ggcorrplot. R package version 0.1.3.

Caitlin Kuempel, Melanie Frazier, Kristy Nack, Nis Sand Jacobsen, David Williams, Julia Blanchard, Richard McIntyre, Daniel Moran, Lex Bouwman, Halley Froelich, Jessica Gephart, Marc Metian, Johannes Tobben, and Benjamin Halpern. Integrating life cycle and impact assessments to map food's cumulative environmental footprint. *One Earth*, 3 (1):65 – 78, 2020.

Mary Menton, Carlos Larrea, Sara Latorre, Joan Martinez-Alier, Mika Peck, Leah Temper, and Mariana Walter. Environmental justice and the sdgs: From synergies to gaps and contradictions. *Sustainability Science*, 15:1621 – 1636, 2020.

Kirill Müller. *here: A Simpler Way to Find Your Files*, 2020. URL https://CRAN.R-project.org/package=here. R package version 1.0.1.

Aleksandra Paluszynska, Przemyslaw Biecek, and Yue Jiang. *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*, 2020. URL https://CRAN.R-project.org/package=randomForestExplainer. R package version 0.10.1.

Esther Sanyé-Mengual, Francesco Orsini, and Giogio Gianquinto. Revisiting the sustainability concept of urban food production from a stakeholders' perspective. *Sustainability*, 10 (7):2175, 2018.

Kevin Ushey. *renv: Project Environments*, 2022. URL https://rstudio.github.io/renv/. R package version 0.15.4.

Hadley Wickham. *forcats: Tools for Working with Categorical Variables (Factors)*, 2021a. URL https://CRAN.R-project.org/package=forcats. R package version 0.5.1.

Hadley Wickham. *tidyverse: Easily Install and Load the Tidyverse*, 2021b. URL https://CRAN.R-project.org/package=tidyverse. R package version 1.3.1.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2021. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.3.5.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2021. URL https://CRAN.R-project.org/package=bookdown. R package version 0.24.