Using the power of data science for real-time spatial and temporal visualization and modeling of COVID-19

Rachel Moss Carroll, Ph.D.

Assistant Professor of Statistics and Data Science

Department of Mathematics and Statistics

University of North Carolina Wilmington

COVID-19 Background

- Began in Wuhan, China in December of 2019 and quickly evolved into a global crisis
- As of May 18, 2020 in the United States
  - 1,492,822 confirmed cases
  - 89,101 deaths
- This project involves:
  - Data mining and compilation
  - Development of data visualization dashboards
  - Statistical modeling

Data

- Outcomes: Johns Hopkins University Center for Systems Science and Engineering
  - Confirmed cases, active cases, new cases, and deaths in the U.S.

- Predictors:
  - Presence of an international airport (yes vs. no)
  - Day of week (early vs. late)
  - Presence of a state-issued stay at home order (before, during – partial, during – full, and none)
  - Continuous population characteristics represented as a percent of the population: Unemployed, smoker, aged >=65, African American

Data

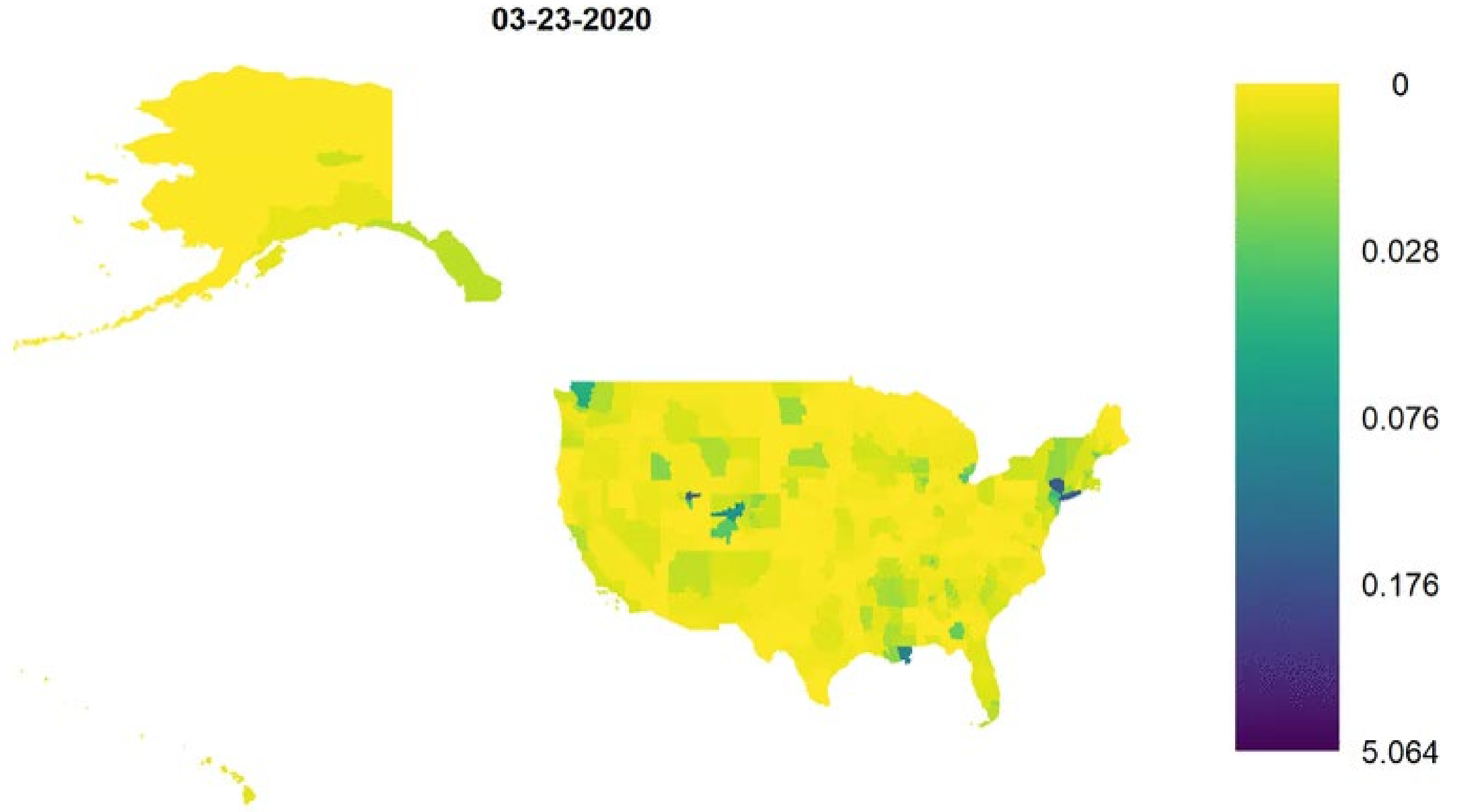
Spatial unit: Health Regions

- Definition - One or more counties serviced by the same health department
- Justification:
  - Zero count of confirmed cases (as of May 18, 2020): 33,206 (U.S. county) to 302 (U.S. Health Region)
  - Spatial dimension: 3,142 (number of U.S. counties) to 389 (number of U.S. Health Regions)

Temporal Unit: Days

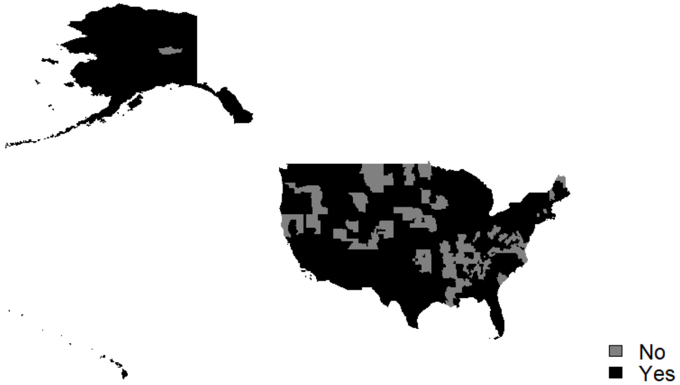
- Limited to the first 8 weeks of recorded data (March 23, 2020 to May 18, 2020)

Data – Confirmed Cases

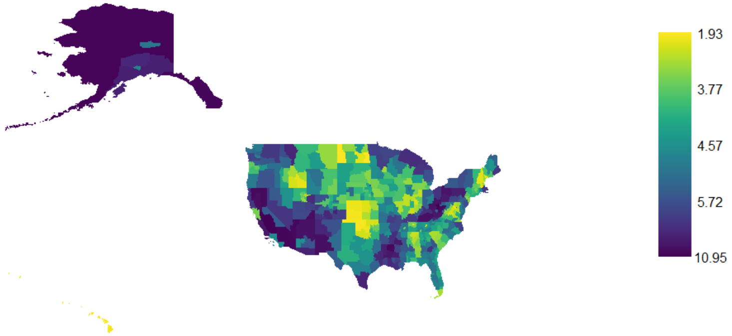


Data – Spatial Predictors

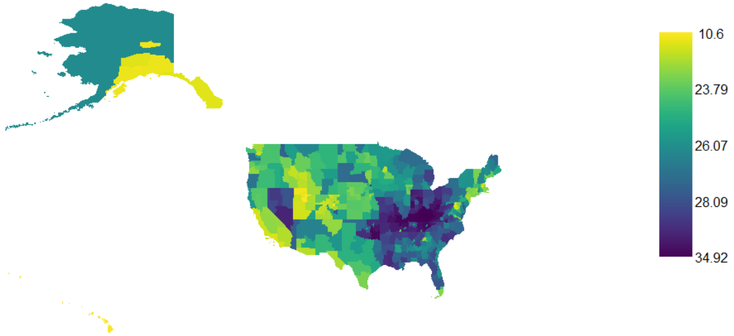
International Airports



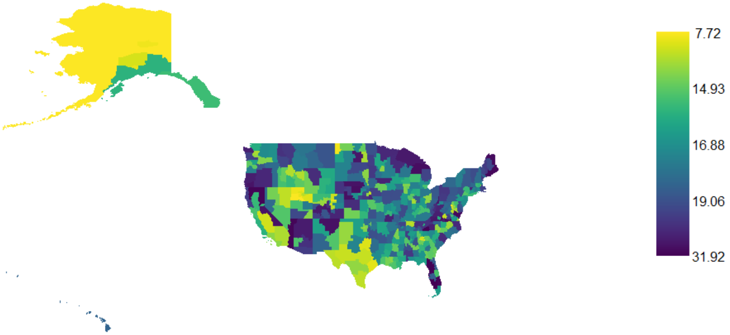
Unemployment



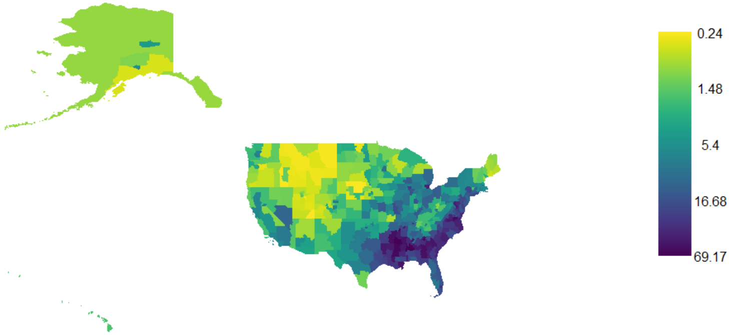
Smoking



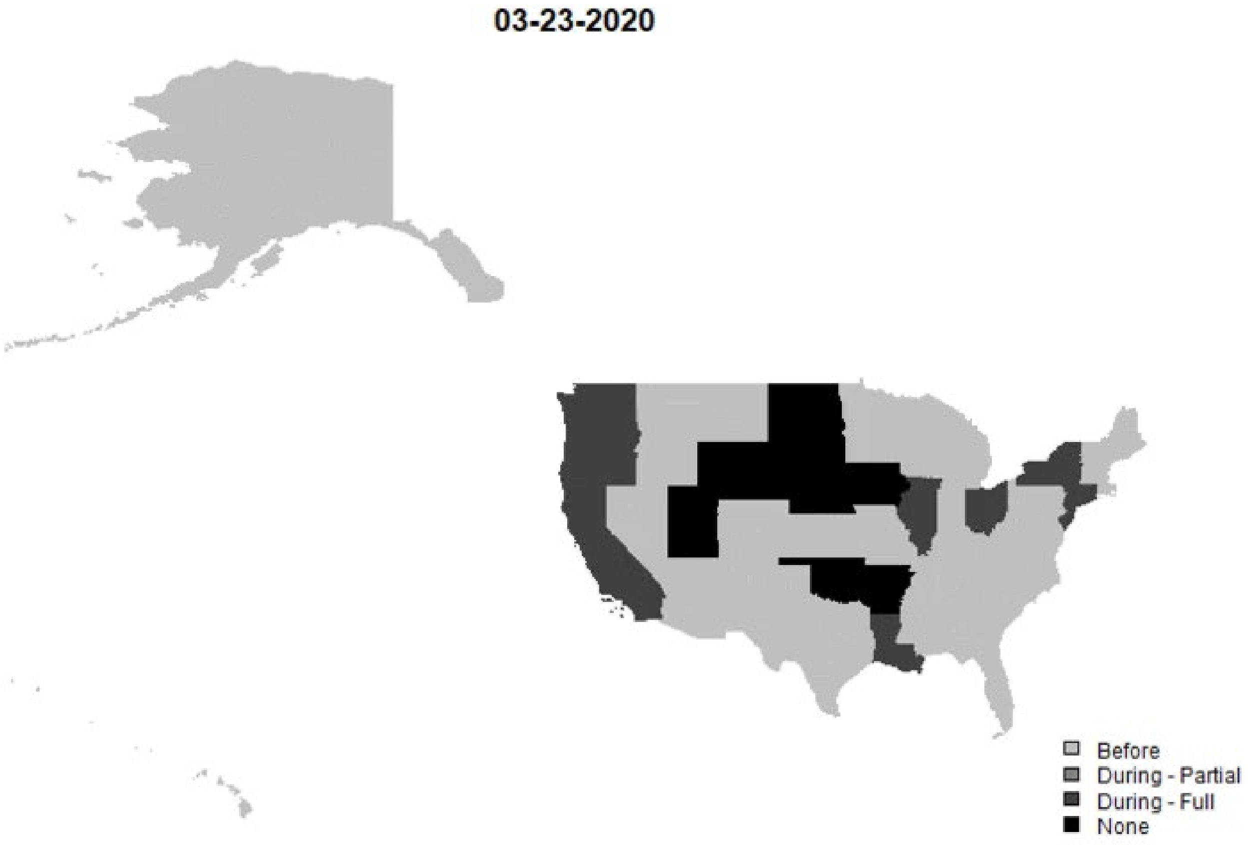
Age >= 65



African American



Data – Stay at Home



Statistical Model

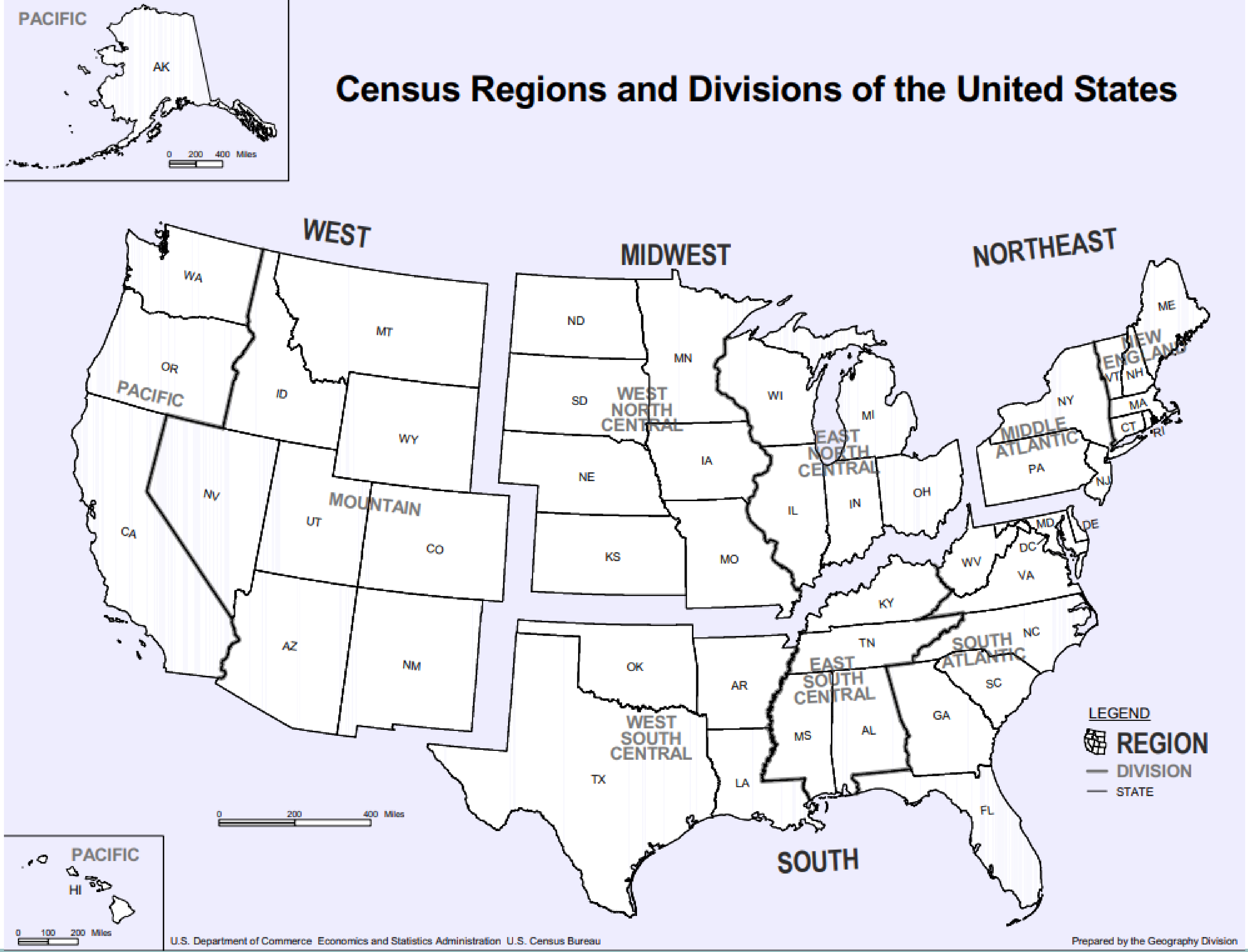
For Health Region i and day j:

$$y_{ij} \sim Pois(\mu_{ij})$$
$$\mu_{ij} = e_{ij}\theta_{ij}$$
$$\log(\theta_{ij}) = X'_{ij}\boldsymbol{\beta} + u_i + \gamma_j$$

where

- y_{ij} is the outcome of interest
- μ_{ij} is the mean of the Poisson model
- e_{ij} is the expected count
- θ_{ij} is the relative risk
- X_{ij} is the design matrix for the predictors
- $\boldsymbol{\beta}$ is the vector of fixed effect estimates
- u_i is the spatial random effect
- γ_j is the temporal random effect

Census Regions and Divisions of the United States



Computational Details

- Data processing, statistical modeling, and plotting
  - R statistical software – rgdal, INLA, fillmap, and shiny

- Data visualization
  - R-shiny
  - Tableau
  - Power BI

Results – [Statistical Modeling]

(click the link)

General Highlights of Results

- Less risk under full stay at home orders
- Varied risk under partial stay at home orders
- Less risk with more unemployment
- More risk with higher African American population in the south
- Less risk of new cases, but higher risk of death with more older population

Conclusions

Data Visualization

- Provides an important, interactive environment for use in academia and the general public

Statistical Model

- Potentially robust to reporting delays through the e_{ij} matrix
- Has become less feasible for real-time assessment
- Highlights important predictor relationships
- Offers interesting spatial and temporal residuals

Future directions

- Predictive modeling

- Temporally dependent coefficients

Questions?

- This work was sponsored in part by the UNCW Center for Social Impact

- Collaborative work by: UNCW M.S. Data Science Students, Mark Lammers, Zachary Williams, and Dylan McNamara

- Additional public dissemination efforts by New Hanover's Cape Fear Collective

- Link to my GitHub

Results – Outcome SIR/SMR

Confirmed cases

Active cases

New cases

Deaths