Mortality Rates by Selected Causes from 1980-2014

Jonathan Stumpf | Maddie Warndorf | Olivia Zarroli

Objective and Introduction to Dataset

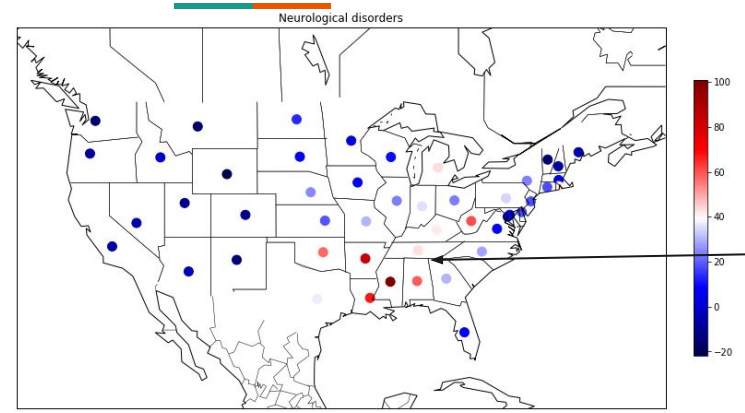
- The dataset:
  - United States Mortality Rates by County 1980-2014
  - Downloaded from: https://www.kaggle.com/IHME/us-countylevel-mortality
- Dataset cleaning:
  - The resolution of the dataset was very high, and offered at a county level for each State. To ensure that each datapoint was statistically significant (avoids counties with small populations), we chose to perform out analysis on a statewide level.
  - The raw dataset also offered metrics on 21 causes of mortality. Four causes were chosen for this analysis

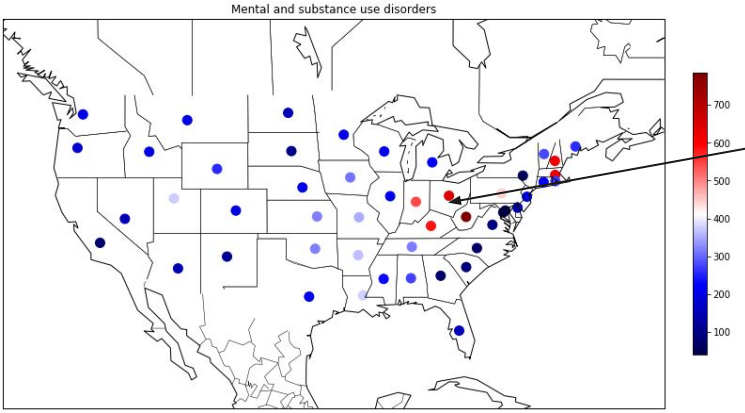
	Location	FIPS	Category	Mortality Rate, 1980*	Mortality Rate, 1980* (Min)	Mortality Rate, 1980* (Max)	Mortality Rate, 1985*	Mortality Rate, 1985* (Min)	Mortality Rate, 1985* (Max)	Mortality Rate, 1990*	...	Mortality Rate, 2005* (Max)	Mortality Rate, 2010*	Mortality Rate, 2010* (Min)	Mortality Rate, 2010* (Max)	Mortality Rate, 2014*	M
0	United States	NaN	Neonatal disorders	9.18	8.83	9.93	6.91	6.73	7.36	6.09	...	4.55	3.75	3.43	3.85	3.32	
1	Alabama	1.0	Neonatal disorders	11.03	10.57	12.00	8.51	8.25	9.12	7.52	...	6.52	5.58	5.16	5.78	5.10	
2	Autauga County, Alabama	1001.0	Neonatal disorders	9.58	8.37	11.02	7.50	6.56	8.58	6.76	...	6.30	4.83	4.19	5.49	4.56	
3	Baldwin County, Alabama	1003.0	Neonatal disorders	8.75	7.86	9.81	6.54	5.88	7.27	5.76	...	5.03	4.02	3.53	4.48	3.68	

Raw Dataset snippet

- Objective:
  - Analyze the following causes (neurological disorders, mental and substance use disorders, nutritional deficiencies, and self-harm and interpersonal violence) to see if the mortality rates are correlated across the United States as well as from 1980-2014.

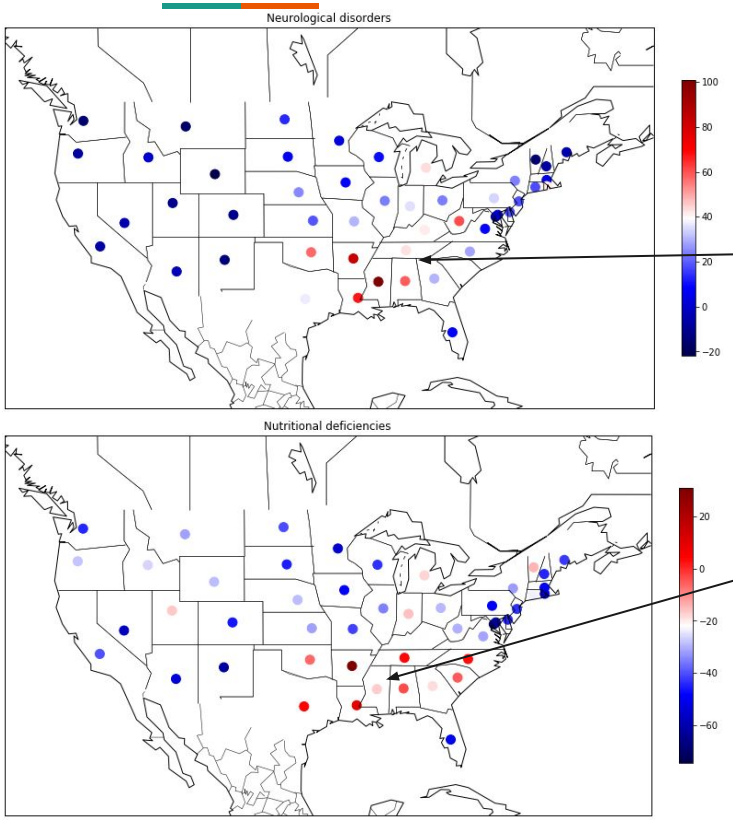
Neuro Disorder vs Mental and Substance Disorder Rate of Change





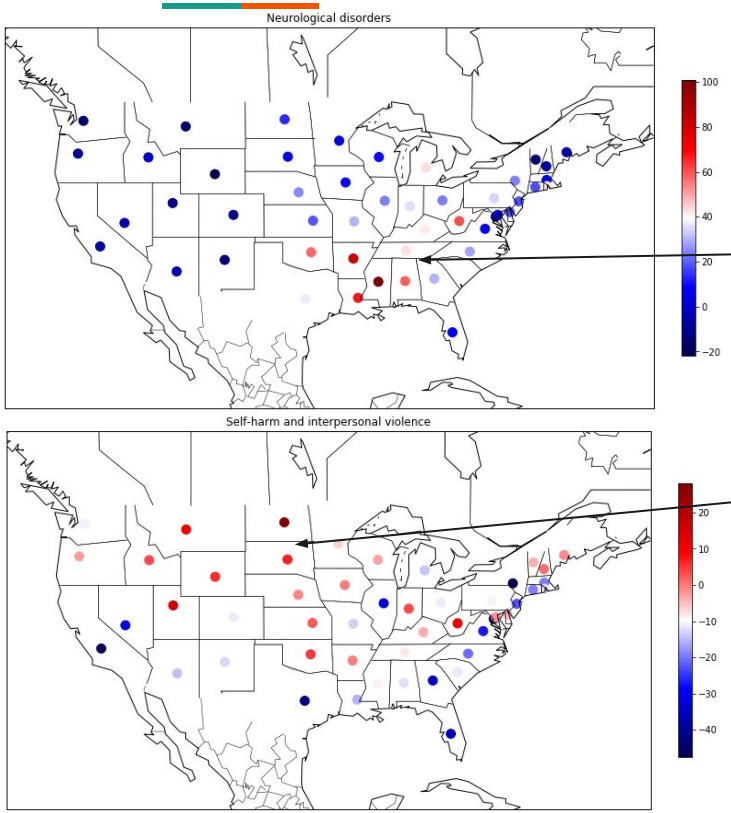
- The top figure is a geo-plot showing the rate of change of mortality caused by Neurological Disorders between 1980-2014
  - The heat map (with respect to the color-scale) shows deaths by neurological disorders increasing in the midwest / southern regions.
- The bottom figure is a geo-plot showing the rate of change of mortality caused by Mental and Substance Use Disorders between 1980-2014
  - The heat map (with respect to the color-scale) shows all states with increased deaths due to mental and substance abuse. Specifically, a pocket of severe increase can be seen between Indiana, Ohio, Kentucky and West Virginia.
- A visual comparison of the plots suggest that a correlation between neuro disease and substance abuse deaths exists

Neuro Disorder vs Nutritional Deficiency Rate of Change



- The top figure is a geo-plot showing the rate of change of mortality caused by Neurological Disorders between 1980-2014
  - The heat map (with respect to the color-scale) shows deaths by neurological disorders increasing in the midwest / southern regions.
- The bottom figure is a geo-plot showing the rate of change of mortality caused by Nutritional Deficiency between 1980-2014
  - The heat map (with respect to the color-scale) shows deaths by nutritional deficiencies increasing in the midwest/southern regions.
- A geographical correlation can be seen between Neurological Disorders and Nutritional Deficiency mortality rate, specifically in the South. This suggest that the two mortality causes may be correlated.

Neuro Disorder vs Self Harm Rate of Change



- The top figure is a geo-plot showing the rate of change of mortality caused by Neurological Disorders between 1980-2014
  - The heat map (with respect to the color-scale) shows deaths by neurological disorders increasing in the midwest / southern regions.
- The bottom figure is a geo-plot showing the rate of change of mortality caused by Self Harm and Interpersonal Violence between 1980-2014
  - The heat map (with respect to the color-scale) shows deaths by nutritional deficiencies increasing the most to the North
- From a visual inspection, it does not seem that nuero disorders and self harm due to mortality are correlated in geography

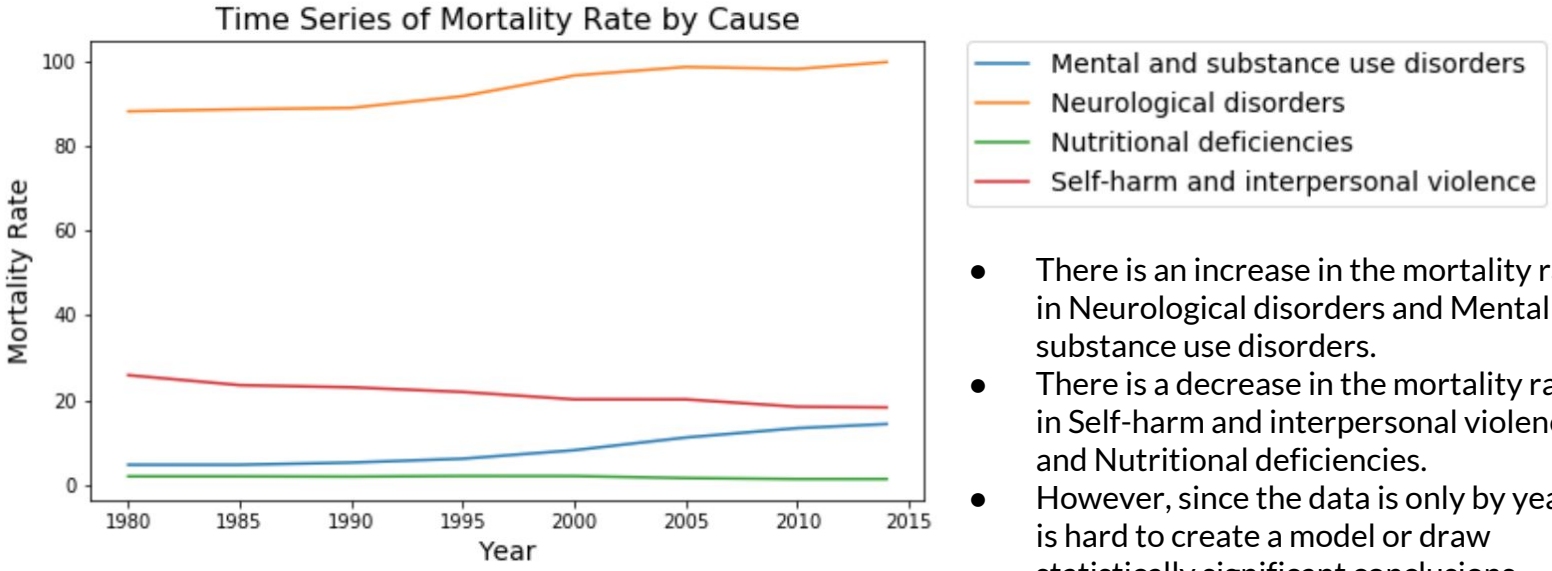
Time Series Analysis

- Had to reshape the data in Excel so that included a “Year” column.
- Added a column containing “Category Number”.
- Focused on analyzing cause of death instead of state, so the data was grouped to find the mean of “Mortality Rate” based on “Category Number” and “Year”
- New dataset was created using “Year”, “Category Number”, “Category”, and “Mortality Rate”.

```
def categoryfun(df):
    "function to add categ
    catlist = ["Mental and
    numlist = []
    for row,rowv in df.ite
        if rowv["Category"
            numlist.append
        elif rowv["Categor
            numlist.append
        elif rowv["Categor
            numlist.append
        else:
            numlist.append

    return(numlist)

catnum = categoryfun(df)

df["Category Number"] = ca
ndf = df.drop(["Unnamed: 6

ndf.head()

grouped = ndf.groupby(["Ca
#print(grouped)
print(grouped.mean())

d = {'Year': [1980,1985,19
     'Category Number': [1
     'Category':["Mental a
     'Mortality Rate': [4.
mdf = pd.DataFrame(data=d)
mdf.head()
```

	Year	Category Number	Category	Mortality Rate
0	1980	1	Mental and substance use disorders	4.693725
1	1985	1	Mental and substance use disorders	4.700000
2	1990	1	Mental and substance use disorders	5.217451
3	1995	1	Mental and substance use disorders	6.128431
4	2000	1	Mental and substance use disorders	8.145490
5	2005	1	Mental and substance use disorders	11.145294
6	2010	1	Mental and substance use disorders	13.351765
7	2014	1	Mental and substance use disorders	14.325294
8	1980	2	Neurological disorders	88.247647
9	1985	2	Neurological disorders	88.724706
10	1990	2	Neurological disorders	89.052549
11	1995	2	Neurological disorders	91.834510
12	2000	2	Neurological disorders	96.715294
13	2005	2	Neurological disorders	98.744706
14	2010	2	Neurological disorders	98.261373
15	2014	2	Neurological disorders	99.884118
16	1980	3	Nutritional deficiencies	1.952157
17	1985	3	Nutritional deficiencies	1.952549
18	1990	3	Nutritional deficiencies	1.896078
19	1995	3	Nutritional deficiencies	2.042745
20	2000	3	Nutritional deficiencies	2.021373
21	2005	3	Nutritional deficiencies	1.547451
22	2010	3	Nutritional deficiencies	1.325294
23	2014	3	Nutritional deficiencies	1.328824
24	1980	4	Self-harm and interpersonal violence	25.881569
25	1985	4	Self-harm and interpersonal violence	23.515392

Time Series Analysis



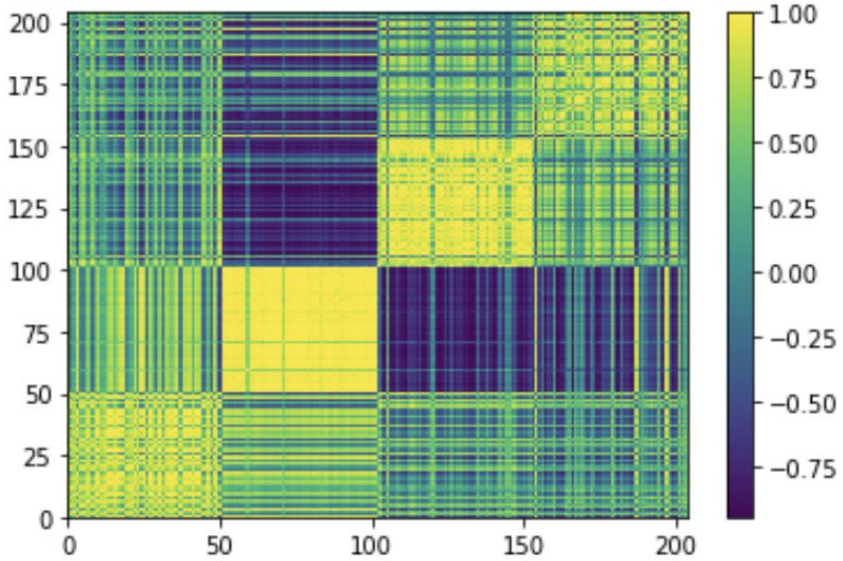
- There is an increase in the mortality rates in Neurological disorders and Mental and substance use disorders.
- There is a decrease in the mortality rates in Self-harm and interpersonal violence and Nutritional deficiencies.
- However, since the data is only by year it is hard to create a model or draw statistically significant conclusions without more data or the data being broken down by MM-YYYY.

Correlation Coefficient

- Using the cleaned data, I was able to enter the data into a matrix
- Parsed the necessary rows and columns so that I was only using the mortality rates from 1980-2010 and not considering any of the percentages
- Used np.corrcoef for the entire matrix
- Put this data into a heat map to make more sense of it..

```
[[ 1.          0.20052748  0.59742211 ... -0.62474542 -0.74171228
  -0.73832292]
 [ 0.20052748  1.          0.7465493  ...  0.46937878  0.41430416
   0.43562871]
 [ 0.59742211  0.7465493   1.         ...  0.16835292  0.04282374
   0.01690195]
 ...
 [-0.62474542  0.46937878  0.16835292 ...  1.          0.97317551
   0.96526096]
 [-0.74171228  0.41430416  0.04282374 ...  0.97317551  1.
   0.99434564]
 [-0.73832292  0.43562871  0.01690195 ...  0.96526096  0.99434564
   1.        ]]
```

Heat Map of Correlation Coefficient



0-50: neurological disease deaths in each state

50-100: mental and substance use disorders

100-150: Nutritional deficiencies

150-200: Self-harm and interpersonal violence

In places where the blocks are most yellow, such as mental and substance use disorders, the trends over time throughout states are closely correlated. Where it is more muffled, as in nutritional deficiencies, the opposite is true.

Final Conclusions

- Visual analysis (geo-location) of rate of mortality change by disease suggest that neurological disorders could be linked to nutritional deficiencies.
- Visual analysis also suggests that substance abuse is increasing the most in a tight geo-centric location (Indiana/Kentucky/West Virginia/Ohio
- Though the time series graph showed changes from 1980 to 2014, it is hard to create a model or draw statistically significant conclusions without more years included or the data being broken down by MM-YYYY.
- We discovered that deaths as a result of mental and substance abuse disorders were the most correlated over time, and neurological diseases were not correlated at all.
- Throughout the entire country, deaths as a result of mental and substance use disorders were most correlated. Self-harm and interpersonal violence was the least correlated cause of death among states.