adjusting population data from yearly to monthly

Yearly data from US Census Bureau website are estimates of population count for July 1st of that year, for each county. Thus, with date in ‘yyyymm’ format as the headers, ‘mm’ for the given data is ‘07’. For the remaining dates (months, for each county, the difference between each date in the original data for is taken, then linearly applied by dividing by 12 and adding the result to each subsequent month.

The PM2.5 is the limiting factor for the range of dates. Thus, the period for analysis is from January 2000 to December 2018.

For missing data

Let lowercase symbols imply the deaths data is suppressed for that county and month

Let state deaths unsuppressed county total deaths

Let state population unsuppressed county total population

Let rate of deaths (for suppressed deaths)

county deaths (suppressed)

4/13/2021

For random forest, burned\_fraction data only available up to 201612; still used to train RF albeit range of dates limited even further

Aggregation of gridded data to counties is only accurate to . Should I decrease it (increase resolution) even further to more accurately capture smaller counties?

Imputation of missing values for AQI: Used pandas interpolation function (wrapper for scipy interpolation functions)

Interpolation and Imputation strategies:

* Interpolation
  + For each site, linearly (method = ‘linear’) interpolate between dates, strictly between dates with existing values (limit\_area=’inside’)
  + Test different values to interpolate in any direction in time (in number of days) (test different values (days) for ‘limit’ parameter, i.e. limit = 0,4,8,12,20,24)
    - [pandas.DataFrame.interpolate](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.interpolate.html)
  + Aggregate daily values to months. If a month is fully interpolated for a site (no missing day values), take the average of the daily values
    - [pandas.DataFrame.groupby](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.groupby.html)
  + Test strategies (method = ‘linear’ / ‘nearest’ / ‘cubic’) for interpolating grid cells between each site, for every month
    - [scipy.interpolate.griddata](https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html)
  + Aggregate each resulting grid (from each limit + method combination) to county shapes.
* Impute missing values
  + Test each grid with imputation strategies for the remaining missing values. Missing values for a given site + month either:
    - did not make the cut for aggregating by month (days in month were not fully interpolated due to limit parameter), or
    - was outside in