Chronic Respiratory Disease: Risk Modeling Potential and Limitations

Alexander He
Information Technology and Web Science
Rensselaer Polytechnic Institute
Troy, NY, USA
hea2@rpi.edu

Thilanka Munasinghe
Information Technology and Web Science
Rensselaer Polytechnic Institute
Troy, NY, USA
munast@rpi.edu

~~Alexander He~~
~~Information Technology and Web Science~~
~~Rensselaer Polytechnic Institute~~
~~Troy, NY, USA~~
~~hea2@rpi.edu~~

~~Thilanka Munasinghe~~
~~Information Technology and Web Science~~
~~Rensselaer Polytechnic Institute~~
~~Troy, NY, USA~~
~~munast@rpi.edu~~

Formatted: Number of columns: 2
Formatted: Number of columns: 2
Commented [AH1]: email address or ORCID
Commented [AH2]: email address or ORCID
Formatted: Don't add space between paragraphs of the same style
Formatted: Number of columns: 2
Formatted: Number of columns: 2
Formatted: Justified

Abstract*—Chronic Respiratory Diseases (CRDs), including Chronic Obstructive Pulmonary Disease (COPD) and asthma, are among the leading causes of mortality worldwide, with 545 million prevalent cases in 2017. Symptoms of noninfectious CRDs are often exacerbated by ambient air pollution and changes in temperature and humidity. This study explores a novel application of machine learning to forecast CRD risk and discuss its merits and limitations. We developed, trained, and tested a Random Forest regressor using datasets over the United States during 2000-2016, with mortality rate as the target variable. The final regressor produced R-squared values of 0.7526 and 0.7528 for cross-validation and test dataset prediction, respectively, implying that our model generalizes well out-of-sample. The selected features comprise location and temporal encoders, population density, and net primary production. The study reveals significant potential for modeling CRD risk but highlights setbacks due to the primarily noninfectious nature of CRDs, phenomena only identifiable on finer spatiotemporal scales, and data limitations. We also identify methods that may refine our approach and describe future developments that may improve CRD risk modeling.

Keywords*—chronic respiratory disease, air pollution, climate variables, machine learning, random forest

I. Introduction

According to the Global Burden of Diseases Study (GBD) reports for 2017, there were an estimated 545 million prevalent cases and 62 million incident cases (new cases) of Chronic Respiratory Diseases (CRDs), the vast majority of which are of Chronic Obstructive Pulmonary Disease (COPD) and asthma [1]. In terms of mortality, there were 3.2 million deaths from COPD and 495,000 deaths from asthma [2].

Tobacco, air pollutants, airborne allergens, and occupational exposures are the main risk factors associated with a lifetime increase of respiratory diseases and symptoms, according to reviews [3,][4]. Previous studies have also established that temperature and humidity play a role in respiratory disease exacerbation [5,][6].

Furthermore, the World Health Organization emphasizes the potentially wide-ranging health impacts of climate change, with many being indirect [7], such as increases in wildfire emissions and longer pollen seasons. However, due to limited datasets on a national, county-level scale, we are constrained to environmental factors, particularly air pollutants and climate variables, provided as datasets constructed with remote sensing.

The present study investigates the potential and limitations for a novel application of machine learning analyses with climate variables and indices, fire emissions, and particulate matter to predict mortality rates attributed to Chronic Respiratory Diseases (CRDs), specifically Chronic Lower Respiratory Diseases (CLRDs), in the contiguous United States during 2000-2016. CLRDs include emphysema, asthma, bronchiectasis, and other COPDs [8]. We chose CLRDs over asthma since the asthma data is overly subject to data suppression constraints affecting lower death counts. We chose them over respiratory diseases as a whole in order to exclude infectious respiratory diseases, such as influenza, pneumonia, and other respiratory infections [8]. However, despite being not primarily infectious in etiology, many CRDs have some aspects of their pathogenesis influenced by infectious organisms [9]. This is an active area for investigation and outside the scope of the present study.

For model development, we performed regression with random forest, performed feature selection and hyperparameter tuning with cross-validation, and measured model performance with the coefficient of determination (R-squared, R^2) as the scoring metric. Finally, we draw insights from our approach, discuss its limitations, and describe future developments that may benefit the modeling of CRD risk.

II. Materials

The datasets included at the time of the present study overlap for the years 2000-2016, thus determining the period of interest. All datasets we describe are available in monthly temporal resolution unless otherwise specified.

A. Underlying Cause of Death – Chronic Lower Respiratory Diseases

The Underlying Cause of Death data available on CDC WONDER is county-level national mortality and population data, beginning from 1999. The data compilates death certificates for U.S. residents, stratified by cause-of-death and demographics [10]. The data is subject to suppression constraints that omit all sub-national data representing less than 10 deaths.

We extracted death counts for monthly, county-level deaths caused by Chronic Lower Respiratory Diseases (CLRDs) from the request form for Underlying Cause of Death [11], and calculated mortality rates per 100,000 people. To interpolate the suppressed data for counties in each state and month, we subtract the sum of reported county deaths from the state total, then calculate the mortality rate adjusted by county population. In doing so, we apply the assumption that mortality rates are equal between the counties with suppressed data.

B. Fine Particulate Matter

Global estimates of fine particulate matter (PM2.5) concentrations, available at [12], are developed using advances in satellite observations, chemical transport modeling, and ground-based monitoring [13]. We extracted the data files for monthly North American Regional Estimates (V4.NA.03).

C. Global Fire Emissions Database

The Global Fire Emissions Database (GFED), available at [14], estimates monthly burned area and fire emissions using remote sensing data of fire activity and vegetation productivity [15]. The fourth version has several modifications from the previous version and uses higher quality input datasets. A notable upgrade is the inclusion of contributions from small fires. The GFED layers included in the present study are fraction of area that burned (burned_frac), fraction of total emissions stemming from small fires (smallf_frac), total emissions measured in carbon (C) and dry matter (DM), net primary production (NPP), heterotrophic respiration (Rh), and fire emissions (BB). The GFED authors use 'biosphere fluxes' as the umbrella term referring to NPP, R_h, and BB.

~~*D. Air Quality Index*~~

~~Air Quality Index (AQI) data for the United States is available at [16]. We opted to use AQI instead of individual criteria pollutants, such as ground-level ozone; AQI was less spatially and temporally sparse and better for interpolation. For each site, we linearly interpolated dates with missing data, then averaged by month. We then interpolated and rasterized AQI at 0.01° resolution using SciPy's 'scipy.interpolate.griddata' function [17][18], before aggregating to county shapes. To choose the best interpolation parameters, we performed cross-validation with R^2 to test the number of consecutive missing dates to interpolate between existing dates and the spatial interpolation method (linear, nearest, or cubic). Finally, we use zeros to replace months with incomplete interpolation, which was the best imputation strategy based on cross-validation.~~

~~Due to the sparsity of the data, we considered the use of AQI experimental, as naive spatiotemporal interpolation is inherently erroneous. For this reason, we expected to eliminate AQI during feature selection.~~

~~*G.*~~*D. Population and Median Income*

Yearly estimates for population by county are available at ~~[19]~~[17] under the "Population Estimates" filter. Population estimates of each year are for July 1st. We opted to apply the July 1st estimates for the entire month of July, then linearly interpolate the months in between.

Yearly estimates for median income by county are available at ~~[20]~~[18]. We applied the yearly values for all months in the calendar year, as no month or day is specified. Given this loss of temporal resolution, the use of median income was considered experimental. Thus, we expected to eliminate median income during feature elimination, though we may gain some insights during collinearity analysis.

~~*H.*~~*E. Climate Variables and Indices*

Monthly data for climate variables and indices is available from NOAA's Climate Divisional Database ~~[21]~~[19]. We extracted data for temperature, precipitation, and drought indices to best represent environmental factors described in [5,~~]~~[6,~~]~~[~~22~~]20,~~[23~~21]. The included drought indices are Palmer Drought Severity Index ~~[24~~[22] (PDSI) and monthly Standardized Precipitation Index ~~[25~~[23] (monthly SPI; SP01). Unfortunately, the database does not have data for humidity, and we could not find humidity data that can be rasterized elsewhere, so the drought indices were considered a proxy for humidity. The data for drought indices is stratified by climate divisions. U.S. climate divisions do not follow county boundaries, so we rasterized the data at 0.01° resolution based on the climate divisions shapefile provided in the File Transfer Protocol (FTP) repository, then aggregated by county.

~~*I.*~~*F. County Boundaries and Area*

The shapefile for U.S. County boundaries is available at ~~[26~~[24]. The file also includes county land and water area data, which we used to calculate county-level population density in population per square kilometer. Since there are

several changes to counties within the period of this study (2000-2016), we processed all county-level datasets to the most current boundaries based on changes described in [~~27~~25].

We used the shapefile data to aggregate all gridded datasets at 0.01° resolution and adjusted by the total area of 0.01° grid cells, which we calculated with spherical trigonometry. To simplify area calculation for complex polygons, we approximated area by including an entire cell for a polygon if it contains the center of the cell.

III. Methods

A. Model Development

Our study used SciKit Learn's function for random forest regression [~~28~~26].~~–~~ Similar to the use of monthly lagged values of climate variables in [~~29~~27], we incorporated 1- and 2-month lagged values for the appropriate features in our model development. We used 10-fold cross-validation with the training dataset and scoring with R^2 for feature selection, hyperparameter tuning, and overall evaluation of model performance. Predicting on an unseen test dataset excluded from training and feature selection was used to evaluate the generalization capability of the model. Specifically, we used a 70:30 split between the training and testing datasets.

We conducted feature selection with SciKit Learn's Recursive Feature Elimination and Cross-Validated selection (RFECV) function [~~30~~28]. A positive contribution to model performance equates to a higher mean R^2 from cross-validation when a feature is included in the model versus when the feature is not included. When using SciKit Learn's RFECV function, strange behavior occurs with more than 15 starting features, where the cross-validation R^2 values are incorrect. Thus, we executed RFECV with random combinations of starting features. For each combination, an output array from the function call ranks features based on cross-validation R^2 immediately before each feature is eliminated (when the feature is included but is the least important) [~~30~~28,][~~31~~29]. We averaged rankings between each combination of starting features. ~~–~~Up to half of the features that ranked worst on average are removed. We then used the remaining features to rerun RFECV. These steps were performed iteratively until RFECV no longer eliminates any features. Furthermore, we did not include lagged values until after the second iteration; for the scope of the present study, we deemed this will increase efficiency while minimally affecting feature selection.

We performed hyperparameter tuning with SciKit Learn's 'GridSearchCV' function [~~32~~30], which tests specified combinations of values for hyperparameters. The tuned hyperparameters for random forest regression are listed in Table I and are described in the documentation [~~28~~26]. For the present study, we performed hyperparameter tuning before and after feature selection.

B. Spearman Rank Correlation Coefficients

To assist in feature selection, we performed collinearity analysis using Spearman rank correlation with SciPy's 'scipy.stats.spearmanr' function [16,[~~33~~31][~~18~~]. The function also calculates statistical significance as two-sided p-values where the null hypothesis is that two datasets are uncorrelated. We used the returned correlation matrix (~~Fig. 9~~Fig. 6) to identify instances of collinearity. We considered absolute pairwise correlations to be high in collinearity if they are above a threshold ($|r| > 0.7$), as suggested by [~~34~~32].

IV. Results

After feature selection and hyperparameter tuning, our model produced an R^2 of 0.7526 during cross-validation and 0.7528 for test dataset prediction. When tuning hyperparameters solely to maximize R^2 for test dataset prediction, the model produced an R^2 of 0.7533 for test dataset prediction.

A. Feature Selection

The Recursive Feature Elimination and Cross-Validated selection (RFECV) function was executed iteratively with random combinations of starting features, as described previously. For each combination, an output array from the RFECV function call ranks features based on cross-validation R^2 immediately before each feature is eliminated. To characterize each iteration of RFECV, we averaged the cross-validation R^2 for each rank between all combinations tested in the iteration (Fig. 3).

Among lagged/unlagged GFED layers, ~~AQI,~~ PM2.5, and climate variables/indices, the 1-month lagged net primary production (NPP) was the only feature not eliminated during feature selection. The other selected features are the location encoders for state and county (STATEFP and GEOID), temporal encoders for month and months from the start of the period (month and months_from_start), and land-area population density (popuDensity_ALAND_km2). Fig. 1 shows the Feature Importances for the model trained on these features.

By order of elimination, 1- and 2-month lagged temperature (temp_F) were the last and second last, respectively, 1-month lagged PM2.5 was the third last, followed by unlagged temperature, median income, and 2-month lagged Palmer Drought Severity Index (PDSI) (Fig. 2).

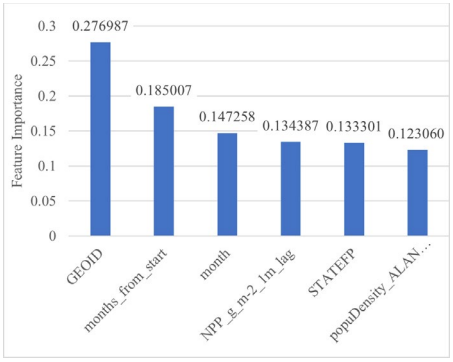


Fig. 1. Impurity-based feature importances for the model with the final selected features and optimal hyperparameters based on cross-validation.

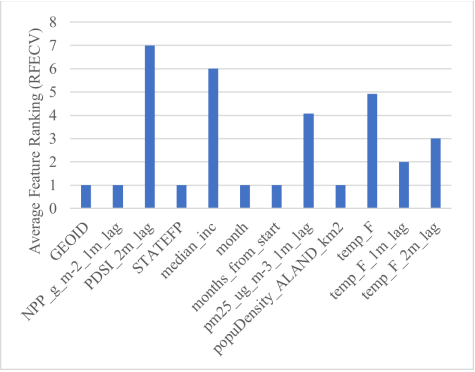


Fig. 2. Average feature rankings during Iteration 4 of feature selection. Features are ranked by cross-validation R^2 immediately before each feature is eliminated. For this iteration, selected features were the same across all Recursive Feature Elimination and Cross-Validated selection (RFECV) function calls, thus having an average ranking of 1.

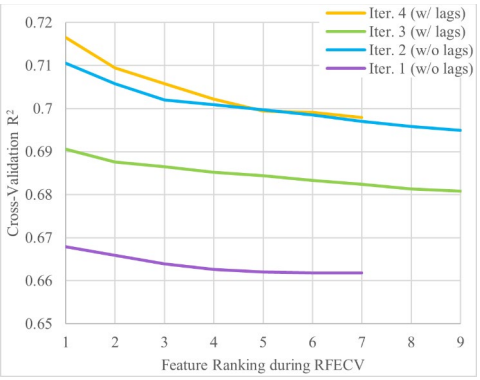


Fig. 3. Cross-validation R^2 by feature ranking during RFECV, regardless of feature. Features are ranked by cross-validation R^2 immediately before each feature is eliminated. For any iteration, selected features are assigned rank 1.

B. Collinearity Analysis

We observed many cases of high collinearity ($|r| > 0.7$) between the lagged and unlagged values of the same variable, especially for Palmer Drought Severity Index (PDSI), net primary production (NPP), heterotrophic respiration (R_h), and temperature (temp_F) (~~Fig. 9~~Fig. 6). We also observed high collinearity between the GFED layers, with correlations equal or nearly equal to 1 between fire emissions (BB), carbon (C), and dry matter (DM), and correlations of 0.867 to 0.988 between the other pairs of GFED layers. Unlagged NPP and R_h are strongly correlated with temperature (0.869 and 0.753, respectively) and satisfy the threshold for high collinearity (~~Fig. 9~~Fig. 6). Out of these features, only 1-month lagged NPP is among the final selected features.

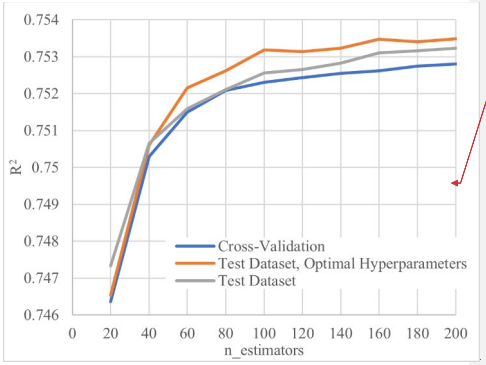
C. Random Forest Hyperparameter Tuning

We listed the tuned hyperparameters and their optimal tested values in Table I. We performed hyperparameter tuning before and after feature selection, as described previously. To create manageable runtime during feature selection, we set max_samples to 0.1 beforehand.

~~Increasing the number of estimators (n_estimators) only increased R^2 with diminishing returns (Fig. 4) and increased runtime. We determined 140 estimators to be sufficient.~~

~~Increasing min_impurity_decrease generally decreases R^2. However, there is significant noise when testing values below 5.0E-7 intervals (Fig. 5). This noise poses difficulty when selecting an optimal value, though for our model, simply using zero was near optimal. Thus, we tuned the remaining hyperparameters with min_impurity_decrease set to zero, then tuned min_impurity_decrease afterward.~~

~~Tuning the remaining hyperparameters displayed the effects of "smoothing" the model and controlling overfitting, as reflected by the maximums observed for cross-validation R^2 (Fig. 6). However, the improvement in R^2 is small for some hyperparameters and significant for others, with max_samples providing the largest improvement.~~



~~Fig. 1. R^2 by max_samples. The other hyperparameters are each at their optimal tested value (except for min_impurity_decrease = 0). "Optimal Hyperparameters" refers to the hyperparameters achieving the highest R^2 for test dataset prediction.~~

Formatted: Indent: First line: 0"

Formatted: Indent: First line: 0.2"

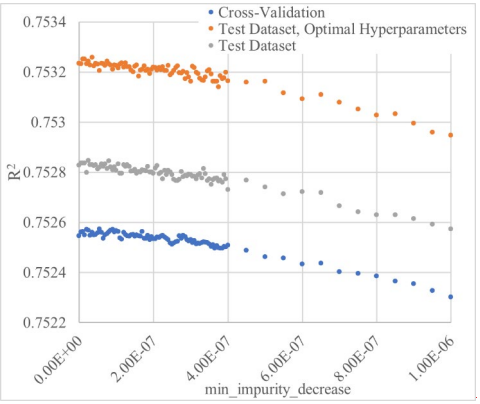
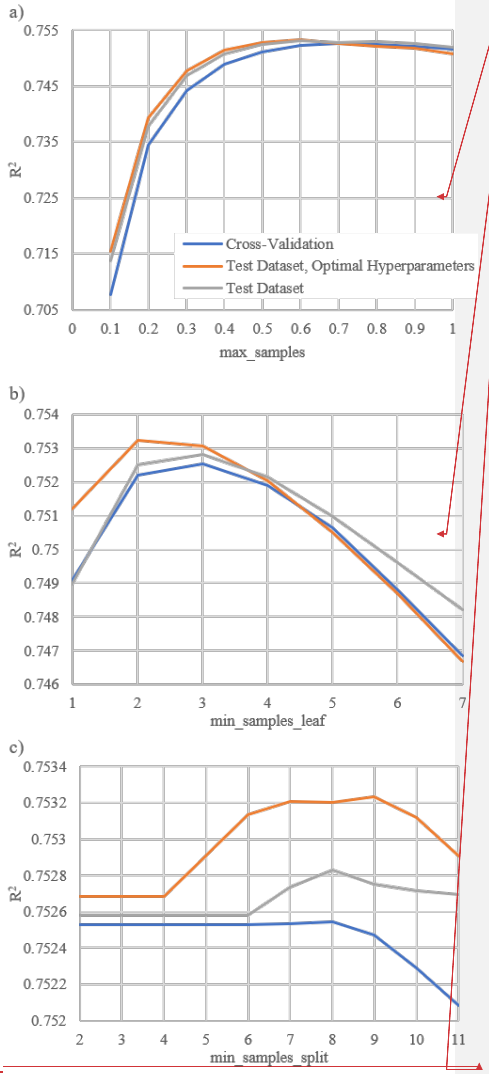


Fig. 3. R^2 by tested values for min_impurity_decrease. We tuned the other hyperparameters with min_impurity_decrease = 0, then tuned min_impurity_decrease afterwards. There is noise in R^2 when testing values for min_impurity_decrease below 5.0E-7 intervals. To narrow down values to test, we tested values in 5.0E-9 intervals for values less than 4.0E-7. "Optimal Hyperparameters" refers to the hyperparameters achieving the highest R^2 for test dataset prediction.

TABLE IV. TABLE I. OPTIMAL TESTED HYPERPARAMETER VALUES

Name	Before RFECV (based on cross-validation R^2)	After RFECV (based on cross-validation R^2)	After RFECV (based on prediction R^2)
n_estimators	140	140	140
max_samples	0.1	0.7	0.6
min_impurity_decrease [a]	0	5.5E-8	3.5E-8
min_samples_leaf	2	3	2
min_samples_split	4	8	9

[a] There is noise in R^2 when testing values for min_impurity_decrease below 5.0E-7 intervals. Simply using 0 is very close to optimal. We tuned with n_estimators = 140 and min_impurity_decrease = 0, then tuned min_impurity_decrease at the final step.



[a]

Fig. 4. R^2 by tested values for the hyperparameters a) max_samples, b) min_samples_leaf, and c) min_samples_split. We display each hyperparameter's cross-validation and test dataset prediction graph when other hyperparameters are each at their optimal tested value (except for min_impurity_decrease = 0). "Optimal Hyperparameters" refers to the hyperparameters achieving the highest R^2 for test dataset prediction.

E.D. *Test Dataset Prediction*

Similar trends in R^2 were observed for cross-validation and test dataset prediction during both feature selection (Fig.

Formatted: figure caption, Left, Space After: 0 pt, Line spacing: single, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: table footnote, Left, Space After: 0 pt, Line spacing: single, Border: Top: (No border), Bottom: (No border), Left: (No border), Right: (No border), Between : (No border)

Formatted: Font: Font color: Black

~~7~~Fig. 4) and hyperparameter tuning (~~Fig. 8~~Fig. 5), suggesting that our model generalizes well for unseen data. Notably, R^2 tends to be slightly higher for test dataset prediction.

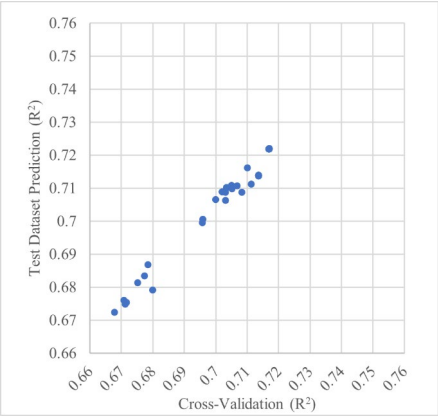


Fig. 4. Comparison of R^2 between cross-validation and test dataset prediction for features with rank 1 across all RFECV executions (before hyperparameter tuning).

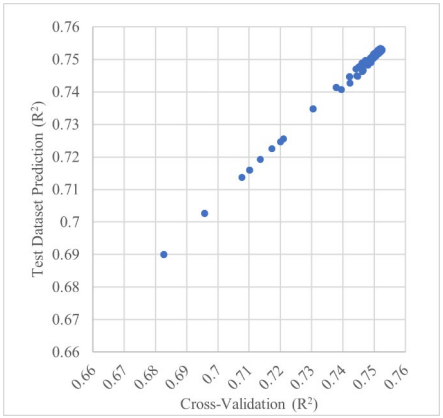
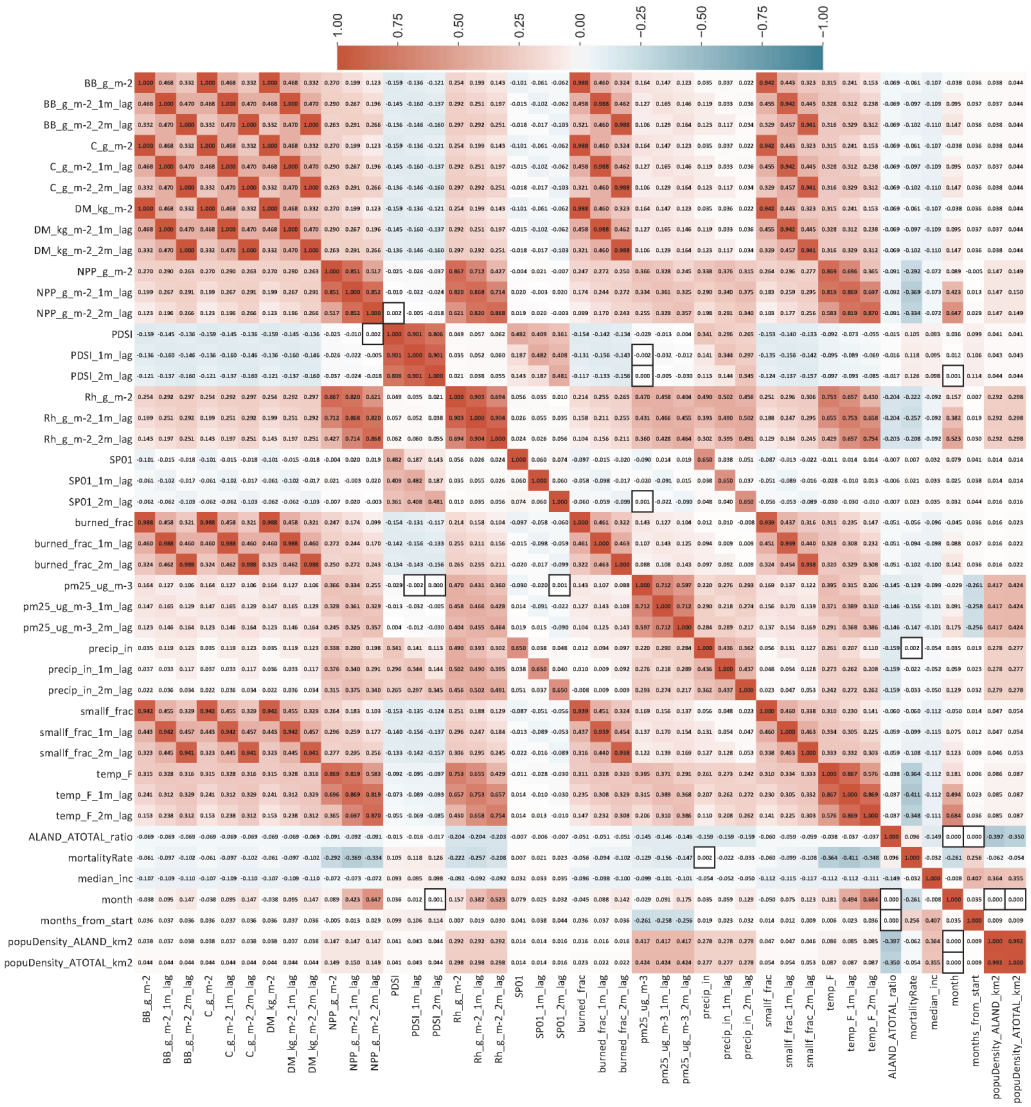


Fig. 5. Comparison of R^2 between cross-validation and test dataset prediction for hyperparameter tuning after feature selection, with min_impurity_decrease = 0 and n_estimators = 140.


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
BB_g_m-2
BB_g_m-2_1m_lag
BB_g_m-2_2m_lag
C_g_m-2
C_g_m-2_1m_lag
C_g_m-2_2m_lag
DM_kg_m-2
DM_kg_m-2_1m_lag
DM_kg_m-2_2m_lag
NPP_g_m-2
NPP_g_m-2_1m_lag
NPP_g_m-2_2m_lag
PDSI
PDSI_1m_lag
PDSI_2m_lag
Rh_g_m-2
Rh_g_m-2_1m_lag
Rh_g_m-2_2m_lag
SPI01
SPI01_1m_lag
SPI01_2m_lag
burned_frac
burned_frac_1m_lag
burned_frac_2m_lag
pm25_ug_m-3
pm25_ug_m-3_1m_lag
pm25_ug_m-3_2m_lag
precip_in
precip_in_1m_lag
precip_in_2m_lag
smallf_frac
smallf_frac_1m_lag
smallf_frac_2m_lag
temp_F
temp_F_1m_lag
temp_F_2m_lag
ALAND_ATOTAL_ratio
mortalityRate
median_inc
month
months_from_start
popuDensity_ALAND_km2
popuDensity_ATOTAL_km2

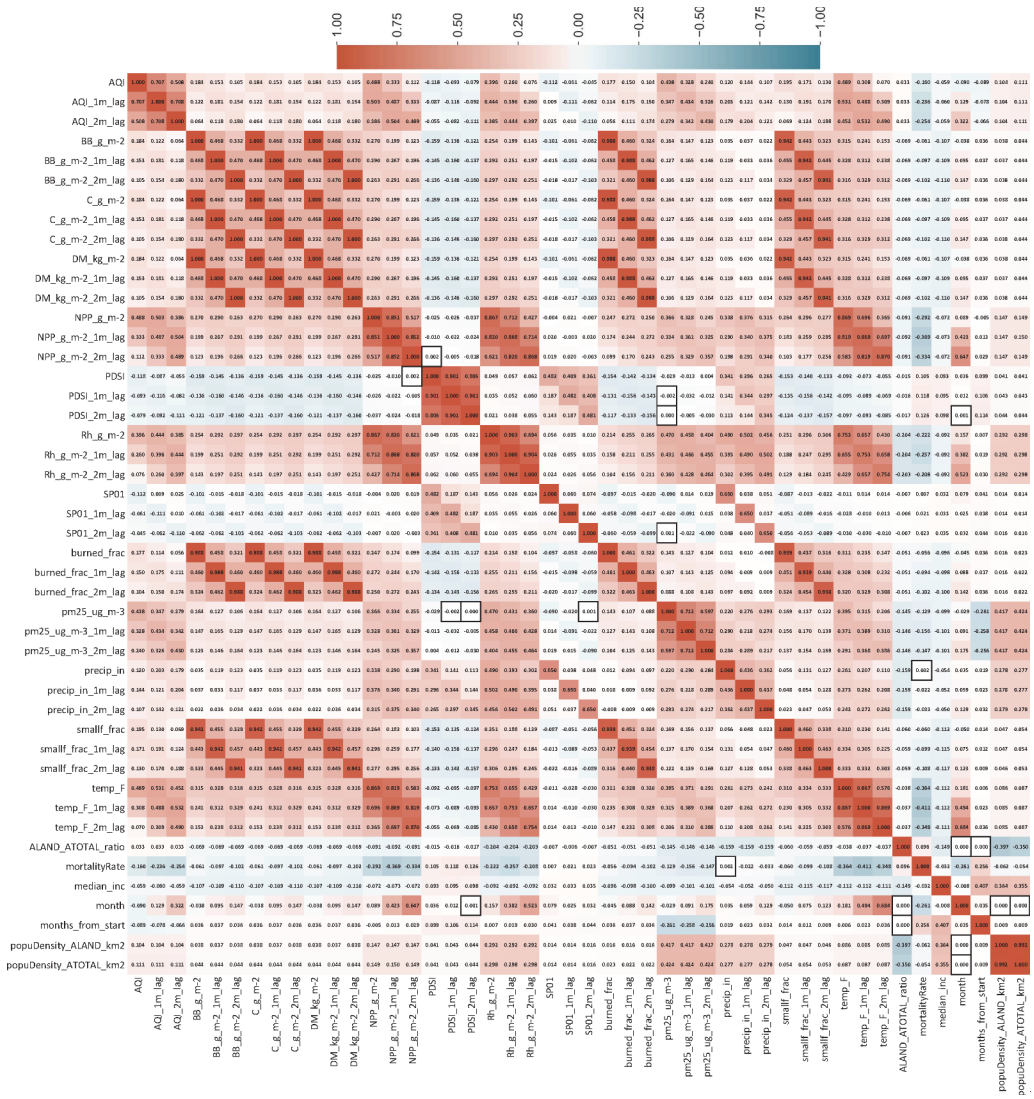


Fig. 6. Spearman rank correlation matrix for all essential climate variables and lagged values considered in the present study. Values with significant p-values (<0.05) are borderless. Statistical significance is calculated as two-sided p-values where the null hypothesis is that two datasets are uncorrelated [16,~~33~~31][~~18~~].

V. Discussion

A. Random Forest Feature Performance Analysis

The high feature importance of the county encoder (GEOID) suggests there are relatively strong county-specific patterns in the form of complex relationships between features. The model likely splits the decision trees at intermediate stages so that these patterns are handled separately. Meanwhile, the state encoder (STATEFP) could be useful in handling patterns found on a lower spatial resolution. It may also improve model performance by creating a divide-and-conquer effect, as counties within the same state will have the most similar patterns.

Iteration 4 incorporates lagged values and produced the highest R^2 values, while Iteration 2 does not incorporate lagged values yet produced the next highest R^2 values (Fig. 3). This suggests that incorporating lagged values for air pollutants, carbon emissions, and climate variables does not offer a substantial increase in our model's performance. We attributed this to the many cases of high collinearity between lagged and unlagged values of the same variable (~~Fig. 9~~Fig. 6).

B. *Importance Measures*

Impurity-based feature importances for random forests and other tree-based models suffer from two flaws. Firstly, they are biased towards high cardinality features (features with many unique values, such as continuous variables). Secondly, they are computed on statistics derived from the training dataset. Thus, they do not necessarily describe which features are the most important to make good predictions in unseen test data ~~[35~~[33].

Permutation importance is another feature importance measure commonly used for random forest models that do not suffer from the flaws associated with impurity-based importance. However, SciKit Learn does not provide a recursive feature elimination function for permutation importance. Thus, we did not incorporate permutation importance in our model development. The close resemblance of R^2 between cross-validation and test dataset prediction during feature selection (~~Fig. 7~~Fig. 4) suggests that the flaws of impurity-based importance were not pronounced in our model. This may compensate for our exclusion of permutation importance.

C. *Climate Variables and Drought Indices*

During feature selection, 1- and 2-month lagged temperature were the last and second last, respectively, to be eliminated. This indicates that, among all eliminated features, the lagged temperature values have the lowest negative impact on model performance.

Humidity and temperature have synergistic effects on the symptoms of COPD patients, where high humidity enhances the risk of COPD due to low temperature [6]. However, we used the drought indices as a proxy to represent humidity. Furthermore, if we found suitable data for humidity, it likely cannot represent indoor conditions.

Lagged/unlagged values for Palmer Drought Severity Index (PDSI) have a weakly positive correlation with mortality rate (0.105 to 0.126), and temperature's lagged/unlagged values have a moderately negative correlation with mortality rate (-0.348 to -0.411) (~~Fig. 9~~Fig. 6). In other words, higher relative soil moisture (higher PDSI value) and lower temperatures are correlated with higher mortality rates. However, higher relative soil moisture does not always coincide with higher humidity ~~[36~~[34,]~~[37~~35]. This may explain the weak correlations between the lagged/unlagged PDSI values and mortality rate (~~Fig. 9~~Fig. 6).

D. *Emissions, Biosphere Fluxes, and PM2.5*

Biosphere fluxes refer to net primary production (NPP), heterotrophic respiration (R_h), and fire emissions (BB). We observed a significant negative correlation between lagged/unlagged NPP and mortality rate (-0.369 to -0.292) (~~Fig. 9~~Fig. 6). NPP is the net carbon gained by vegetation, calculated as the carbon gained by photosynthesis minus the carbon released by plant respiration ~~[38~~[36]. In other words, higher NPP implies less carbon is released into the atmosphere, and this is correlated with a lower mortality rate.

Emissions for carbon (C), dry matter (DM) and fire (BB), and PM2.5 concentrations~~, and AQI~~ have a weak negative correlation with mortality rate (~~Fig. 9~~Fig. 6). This seems to contradict the commonly known exacerbating effects of air pollution on respiratory disease ~~[39~~[37,]~~[40~~38,]~~[41~~39]. An explanation may be that pollutants in the form of fine particulate matter and carbon emissions do not trump climate variables as predictors of mortality.

E. *Relationship Between Drought and Other Features*

Elevated ozone and PM2.5 levels have been attributed to increasing drought ~~[22]~~[20], but the drought indices have a near-zero correlation with PM2.5 ~~and AQI~~ (~~Fig. 9~~Fig. 6). Estimates from ~~[23~~[21] imply a reduction in global NPP due to droughts during 2000-2009, with increased NPP over the Northern Hemisphere offset by decreased NPP over the Southern Hemisphere. However, the present study is limited to the contiguous U.S., and a near-zero correlation between NPP and the drought indices (PDSI and SP01) was observed (~~Fig. 9~~Fig. 6).

Correlation coefficients cannot describe the relationship between drought and these features. Aside from eliminating the drought indices during RFECV, the impact of including drought indices on model performance is inconclusive.

F. *Median Income*

Based on our Spearman rank correlation analysis, median income has a weak negative correlation with all GFED layers and, PM2.5~~, and AQI~~. This may be relevant for literature on environmental inequality, especially North American studies that indicate areas where low-socioeconomic-status communities dwell experience higher concentrations of pollutants ~~[42~~[40].

G. *Data Limitations*

Emergency Room Visits (ERVs) were used as a measure of disease exacerbation in the global study ~~[43~~[41]. However, data is limited for it when stratified by county. Thus, we used mortality provided as death counts by CDC WONDER for every state and county for the present study. However, due to data suppression constraints [10], we interpolated the unreported counts as described previously. Furthermore, deaths only capture the most acute cases of exacerbation and do not necessarily reflect non-mortality cases, not to mention self-treated cases that do not manifest in an ERV.

Not to confuse with total-column ozone, ground-level ozone data is available as annual gridded mean concentrations, which are estimates simulated by an ensemble of five chemical transport models ~~[44~~[42]. These

annual mean concentrations were used to estimate global asthma ERVs attributable to ozone [43][41]. Due to the lack of a monthly ground-level ozone dataset, we did not include ground-level ozone for the present study.

Data for daily AQI, ground-level ozone, and other pollutants from EPA Air Data is available in coordinate-specific format. This format is more beneficial for case studies on a localized scale, such as in [41][39], where monitoring stations distributed within a city can more closely capture the spatial variation of air pollution and subsequent health outcomes between different areas within the city. AQI was spatially interpolated between coordinates with available data and was eliminated in the RFECV as we expected. We can attribute this to the naive interpolation applied to AQI.

H. Other Limitations

Datasets on county-level and monthly resolutions cannot represent phenomena present in finer spatiotemporal resolutions, which is a significant source of limitations for the present study. Existing studies have analyzed various exacerbating effects on a more localized scale, such as within individual cities [41][39]. Thus, we may characterize these limitations.

A multi-city case-crossover study [45][43] found that ERVs for respiratory diseases increase after days with higher concentrations of pollutants, even where pollutant concentrations are relatively low. However, the lagging effect on ERVs lasts only up to several days. For PM2.5 and ERVs for COPD, positive results were observed with lags of 1-8 days. Thus, the use of monthly predictive variables is likely a limitation of this study.

Excluding the predictors tested in this study, Chronic Respiratory Diseases (CRDs) are also attributable to other factors, particularly tobacco smoking and other airborne allergens [46][3,4,44][3][4], but also infectious microorganisms [9].

Exacerbation of asthma and COPD also stems from sudden changes in temperature associated with the overuse of cold air conditioning during warmer months [5]. Meanwhile, in cold and temperate climates, due to more time spent indoors, indoor pollutants are likely more important than outdoor ones [47][45].

VI. Conclusions

This study has created some insights on the application of recursive feature elimination and random forest as a machine learning technique for regression. It also demonstrates the difficulties in modeling Chronic Respiratory Disease (CRD) risk on a national, county-level scale due to data availability limitations, factors not captured on the monthly timescale, and the nature of many CRDs being not primarily infectious in etiology [9]. It highlights the caveats of using machine learning methods to generating predictions from complex relationships between correlated variables. We can naively test for features that positively contribute to model performance, but the mechanisms behind their relationships remain unclear.

A similar approach has been effectively used in a recent study to classify cholera outbreaks in India using essential climate variables, also in monthly temporal resolution, with 89.5% of outbreaks correctly identified in the unseen test dataset [29][27]. Pathogenic *Vibrio cholerae* bacteria are responsible for human cholera. Infection occurs through ingesting contaminated seafood and water and during recreational activities in contaminated waters. Strong spatiotemporal relationships between essential climate variables and distribution of *V. cholerae* bacteria allow for accurate predictions of cholera outbreaks.

Meanwhile, the role of microorganisms in "noninfectious" lung diseases, including COPD, bronchiectasis, and asthma, is an active area for investigation. More sophisticated methods and analyses to investigate the interactions between the immune system, microbiota, and inflammatory pathways are needed to improve our understanding of the pathogenesis of these diseases [9].

In the global study [43][41], health impact functions were used to estimate asthma ERVs and incidence attributable to each pollutant, stratified by country and age group. The health impact functions rely on relative risks (RRs) extracted from meta-analyses of epidemiological studies. However, they do not include climate variables such as temperature, which affect asthma and COPD rates, as existing studies indicate. The effects of climate are somewhat captured since climates tend to be country-specific and thus influence a country's baseline rate of ERVs/deaths, but this becomes an issue for larger countries. Thus, incorporating accessible data of climate variables may improve estimates of ERVs attributable to specific pollutants. Contributions of other factors, such as sudden changes in temperature and infectious determinants, remain open areas for future studies.

Replacing our use of grid-search hyperparameter optimization with randomized or gradient-based [48][46,][4947] may provide immediate improvements to the present study, but our methodologies leave room for greater modifications.

Any asymmetry of importance in target domain values (mortality rates) were not addressed. Thus, imbalanced regression techniques can be used, such as the SMOGN algorithm [50][48], which involves pre-processing with random over-/under-sampling. This also requires additional evaluation metrics, such as F1 scores. Predicting extreme values can be an additional objective for imbalanced regression to address. The squared error-relevance area (SERA) metric [51][49] can be used.

We can explore the model sensitivity to different machine learning techniques, such as neural networks. This may also further validate our results. We can also train with datasets from alternative sources that may develop their datasets differently.

Code Availability

The code developed for this study is available via GitHub at [52][50]. Downloaded datasets, and descriptions of sources for datasets too large for a GitHub repository, are included.

REFERENCES

[1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017," Lancet, vol. 392, no. 10159, pp. 1789–1858, 2018.

[2] GBD 2017 Causes of Death Collaborators, "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017," Lancet, vol. 392, no. 10159, pp. 1736–1788, 2018.

[3] G. Viegi, F. Pistelli, D. L. Sherrill, S. Maio, S. Baldacci, and L. Carrozzi, "Definition, epidemiology and natural history of COPD," European Respiratory Journal, vol. 30, no. 5, pp. 993–1013, 2007.

[4] S. Baldacci, S. Maio, S. Cerrai, G. Sarno, N. Baïz, M. Simoni, I. Annesi-Maesano, and G. Viegi, "Allergy and asthma: Effects of the exposure to particulate matter and biological allergens," Respiratory medicine, vol. 109, no. 9, pp. 1089–1104, May 2015.

[5] M. D'Amato, A. Molino, G. Calabrese, L. Cecchi, I. Annesi-Maesano, and G. D'Amato, "The impact of cold on the respiratory tract and its consequences to respiratory health," Clin. Transl. Allergy, vol. 8, no. 20, May 2018.

[6] Z. Mu, PL. Chen, FH. Geng, L. Ren, WC. Gu, JY. Ma, L. Peng, and QY. Li, "Synergistic effects of temperature and humidity on the symptoms of COPD patients," Int J Biometeorol, vol. 61, pp. 1919–1925, 2017.

[7] A. Prüss-Ustün, J. Wolf, C. Corvalán, R. Bos, and M. Neira, "Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks," World Health Organization, Genève, Switzerland, 2016.

[8] "ICD-10 Version:2010," Who.int. [Online]. Available: https://icd.who.int/browse10/2010/en.

[9] M. E. Fitzpatrick, S. Sethi, C. L. Daley, P. Ray, J. M. Beck, and M. R. Gingo, "Infections in 'noninfectious' lung diseases," Ann. Am. Thorac. Soc., vol. 11 Suppl 4, no. Supplement 4, pp. S221-6, 2014.

[10] "Underlying cause of death 1999-2019," Cdc.gov. [Online]. Available: https://wonder.cdc.gov/wonder/help/ucd.html.

[11] "Underlying cause of death, 1999-2019 request," Cdc.gov. [Online]. Available: https://wonder.cdc.gov/ucd-icd10.html.

[12] "Surface PM2.5," Wustl.edu. [Online]. Available: https://sites.wustl.edu/acag/datasets/surface-pm2-5/.

[13] M. S. Hammer, A. van Donkelaar, C. Li, A. Lyapustin, A. M. Sayer, N. C. Hsu, et al., "Global estimates and long-term trends of fine particulate matter concentrations (1998-2018)," Environ. Sci. Technol., vol. 54, no. 13, pp. 7879–7890, 2020.

[14] "Global fire emissions database," Globalfiredata.org. [Online]. Available: https://www.globalfiredata.org/.

[15] G. R. van der Werf, J. T. Randerson, L. Giglio, T. T. van Leeuwen, Y. Chen, B. M. Rogers, et al., "Global fire emissions estimates during 1997–2016," Earth Syst. Sci. Data, vol. 9, no. 2, pp. 697–720, 2017.

~~[16] "Pre-Generated Data Files," US EPA. [Online]. Available: https://aqs.epa.gov/aqsweb/airdata/download_files.html.~~

~~[17] "scipy.interpolate.griddata," SciPy.org. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html.~~

~~[18]~~[16] P. Virtanen, R. Gommers, et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nat. Methods, vol. 17, no. 3, pp. 261–272, 2020.

~~[19]~~[17] US Census Bureau, "Datasets," Census.gov. [Online]. Available: https://www.census.gov/data/datasets.html.

~~[20]~~[18] US Census Bureau, "SAIPE datasets," Census.gov. [Online]. Available: https://www.census.gov/programs-surveys/saipe/data/datasets.html.

~~[21]~~[19] R. S. Vose, S. Applequist, M. Squires, I. Durre, M. J. Menne, C. N. Williams Jr., C. Fenimore, K. Gleason, and D. Arndt, "NOAA's Climate Divisional Database (nCLIMDIV)." NOAA National Climatic Data Center, 2014.

~~[22]~~[20] Y. Wang, Y. Xie, W. Dong, Y. Ming, J. Wang, and L. Shen, "Adverse effects of increasing drought on air quality via natural processes," Atmos. Chem. Phys., vol. 17, no. 20, pp. 12827–12843, 2017.

~~[23]~~[21] M. Zhao and S. W. Running, "Drought-induced reduction in global terrestrial net primary production from 2000 through 2009," Science, vol. 329, no. 5994, pp. 940–943, 2010.

~~[24]~~[22] W. C. Palmer, Meteorologic Drought. US Department of Commerce, Weather Bureau, Research Paper No. 45, p. 58., 1965.

~~[25]~~[23] T. B. McKee, N. J. Doesken, and J. Kleist, "The Relationship of Drought Frequency and Duration to Time Scales," in 8th Conference on Applied Climatology, 1993.

~~[26]~~[24] US Census Bureau, "Cartographic boundary files," Census.gov. [Online]. Available: https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html.

~~[27]~~[25] US Census Bureau, "Changes to counties and county equivalent entities: 1970-present," Census.gov. [Online]. Available: https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html.

~~[28]~~[26] "sklearn.ensemble.RandomForestRegressor," Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

~~[29]~~[27] A. M. Campbell, M.-F. Racault, S. Goult, and A. Laurenson, "Cholera risk: A machine learning approach applied to essential climate variables," Int. J. Environ. Res. Public Health, vol. 17, no. 24, p. 9378, 2020.

~~[30]~~[28] "sklearn.feature_selection.RFECV," Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html.

~~[31]~~[29] "sklearn.feature_selection.RFE," Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html.

~~[32]~~[30] "sklearn.model_selection.GridSearchCV," Scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

~~[33]~~[31] "scipy.stats.spearmanr," Scipy.org. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html.

~~[34]~~[32] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, et al., "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," Ecography, vol. 36, no. 1, pp. 27–46, 2013.

~~[35]~~[33] G. Louppe, "Understanding random forests: From theory to practice," arXiv [stat.ML], 2014.

~~[36]~~[34] R. R. Heim Jr., "A review of twentieth-century drought indices used in the United States," Bull. Am. Meteorol. Soc., vol. 83, no. 8, pp. 1149–1166, 2002.

~~[37]~~[35] W. M. Alley, "The Palmer drought severity index: Limitations and assumptions," J. Clim. Appl. Meteorol., vol. 23, no. 7, pp. 1100–1109, 1984.

~~[38]~~[36] F. S. Chapin and V. T. Eviner, "Biogeochemistry of terrestrial net primary production," in Treatise on Geochemistry, H. D. Holland and K. K. Turekian, Ed. Elsevier, 2007, pp. 1–35.

~~[39]~~[37] K. J. Foreman, N. Marquez, A. Dolgert, K. Fukutaki, N. Fullman, M. McGaughey, et al., "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories," Lancet, vol. 392, no. 10159, pp. 2052–2090, 2018.

~~[40]~~[38] K. Ito, G. D. Thurston, and R. A. Silverman, "Characterization of PM2.5, gaseous pollutants, and meteorological interactions in the context of time-series health effects models," J. Expo. Sci. Environ. Epidemiol., vol. 17 Suppl 2, no. S2, pp. S45-60, 2007.

[41][39] L. Chen, K. Mengersen, and S. Tong, "Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia," Sci. Total Environ., vol. 373, no. 1, pp. 57–67, 2007.

[42][40] A. Hajat, C. Hsia, and M. S. O'Neill, "Socioeconomic disparities and air pollution exposure: A global review," Curr. Environ. Health Rep., vol. 2, no. 4, pp. 440–450, 2015.

[43][41] S. C. Anenberg, D. K. Henze, V. Tinney, P. L. Kinney, W. Raich, N. Fann, et al., "Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits," Environ. Health Perspect., vol. 126, no. 10, p. 107004, 2018.

[44][42] S. Galmarini, B. Koffi, E. Solazzo, T. Keating, C. Hogrefe, M. Schulz, et al., "Technical note: Coordination and harmonization of the multi-scale, multi-model activities HTAP2, AQMEII3, and MICS-Asia3: simulations, emission inventories, boundary conditions, and model output formats," Atmos. Chem. Phys., vol. 17, no. 2, pp. 1543–1555, 2017.

[45][43] M. Szyszkowicz, T. Kousha, J. Castner, and R. Dales, "Air pollution and emergency department visits for respiratory diseases: A multi-city case crossover study," Environ. Res., vol. 163, pp. 263–269, 2018.

[46][44] J. L. López-Campos, W. Tan, and J. B. Soriano, "Global burden of COPD: Global burden of COPD," Respirology, vol. 21, no. 1, pp. 14–23, 2016.

[47][45] J. Heinrich, "Influence of indoor factors in dwellings on the development of childhood asthma," Int. J. Hyg. Environ. Health, vol. 214, no. 1, pp. 1–25, 2011.

[48][46] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, "Forward and reverse gradient-based hyperparameter optimization," arXiv [stat.ML], 2017.

[49][47] D. Maclaurin, D. Duvenaud, and R. P. Adams, "Gradient-based hyperparameter optimization through reversible learning," arXiv [stat.ML], 2015.

[50][48] P. Branco, L. Torgo, and R. P. Ribeiro, "SMOGN: a Pre-processing Approach for Imbalanced Regression," in *1st International Workshop on Learning with Imbalanced Domains - Theory and Applications*, 2017.

[51][49] R. P. Ribeiro and N. Moniz, "Imbalanced regression and extreme value prediction," Machine Learning, vol. 109, pp. 1803–1835, Sep. 2020.

[52][50] "Research-Spring2021," GitHub. [Online]. Available: https://github.com/Unusuala1l2e3x4/Research-Spring2021.