Multi-Modal Sensing for Estimating U.S. Life Expectancy: Identifying Environmental Determinants and Disparities (2003-2019)

Shisir Ruwali , David Lary * , Samyak Shrestha , Faiz Ahmad

Posted Date: 29 April 2026

doi: 10.20944/preprints202604.1997.v1

Keywords: life expectancy; machine learning; environment; particulate matter; formaldehyde; aerosols

 doi:10.20944/preprints202604.1997.v1

Article

Multi-Modal Sensing for Estimating U.S. Life Expectancy: Identifying Environmental Determinants and Disparities (2003-2019)

Shisir Ruwali, David Lary *, Samyak Shrestha and Faiz Ahmad

Department of Physics, The University of Texas at Dallas, Richardson, TX 75080, USA
* Correspondence: david.lary@utdallas.edu

Abstract

We processed the life expectancy data of age group less than 1 year old from the Institute for Health Metrics and Evaluation (IHME) in contiguous USA and a set of 33 environmental variables (or features) from the European Centre for Medium-Range Forecasts (ECMWF) from the years 2003 through 2019. Visualizing the IHME data we identified the massive disparity in life expectancy in contiguous USA where counties in southern states have relatively less life expectancy compared to counties in northern states. We made use of machine learning to estimate the life expectancy and obtained moderate accuracy as coefficient of determination (R^2) and Root Mean Square Error (RMSE) between the true and estimated values were found to be 0.77 and 1.18 year respectively in an independent test set using only a set of 5 environmental variables. Our key finding shows that apart from well-known pollutants such as particulate matter (PM), ozone, carbonmonoxide, it is essential to reduce pollutants such as formaldehyde, sulphate aerosols, dust aerosols; increase vegetation areas, and good working condition such as lower wet-bulb temperature can potentially increase life expectancy in the US. Future work can include socio-economic variables such as household income, poverty rate and other relevant features to create a comprehensive set of variables to improve the results and livelihood of people.

Keywords: life expectancy; machine learning; environment; particulate matter; formaldehyde; aerosols

1. Introduction

The trend of life expectancy and the factors that affect it have been studied for several years. Despite the growth in economy, massive spending on health care, and medical innovations, life expectancy in the United States compared to other high-income countries falls behind [1]. Life expectancy in the U.S. increased between 1959 and 2016 from 69.9 years to 78.9 years, but it was on a decreasing trend after 2014 for three years [2,3]. Previous studies have also described the disparity in life expectancy between various ethnic groups in the United States [4–6].

The human body interacts with the surrounding and is affected by several factors such as temperature, humidity, altitude, and air pollution. This in turn affects life expectancy. The decrease in concentration of $PM_{2.5}$ has been known to increase life expectancy [7,8]. Cohort studies also show that improved air quality results in good health, leading to a better life expectancy [9,10]. The temperature of the environment significantly affects health. High and low temperature have been associated with years of loss of life, as shown by data from seven locations with low, middle, and high income sites [11]. An increase in the average annual temperature due to climate change could also lead to a decrease in life expectancy [12]. Health hazards are created not only by the temperature condition but also by the humidity. Lower levels of humidity create a favorable condition for the transmission of spread of airborne diseases such as influenza [13,14]. Higher levels of humidity make it harder for the body to sweat, reducing its ability to cool itself.

Formaldehyde is a chemical that is used in building materials and in the production of several household products. Acute and chronic exposure to formaldehyde can have harmful effects on the

 doi:10.20944/preprints202604.1997.v1

respiratory system, urinary system, acute poisoning, irritation, and chest pain [15,16]. The greenness of the environment is also a crucial factor that affects life expectancy. Greenness is known to affect human health by possibly encouraging physical activity [17], promoting mental health [18], and is beneficial in reducing air pollution [19].

In this study, we used data from two different sources: first, life expectancy data from the Institute for Health Metrics and Evaluation (IHME) [20], which provides county wise life expectancy data from 2001 to 2019 of various age groups. We then consider a set of environmental variables from the European Center for Medium-Range Forecasts (ECMWF). From ECMWF, we gather a comprehensive set of environmental variables that could possibly affect life expectancy in the U.S. Using these data sources, our study aims to achieve the following three objectives:

- Identify the disparity in life expectancy in contiguous U.S. considering 17 years of county wise life expectancy data from 2003 through 2019.
- Estimate life expectancy in the contiguous US using a set of environmental variables.
- Identify environmental factors that affect life expectancy.

Instead of particular variables such as temperature, humidity, and air pollution, our study simultaneously considers a comprehensive set of environmental variables that could possibly affect life expectancy. Furthermore, instead of a localized region and small period of time, our study considers 17 years of county wise data of the entire contiguous U.S. We also utilize machine learning to test the accuracy in estimating life expectancy in the U.S. using a set of environmental variables. We have also taken a machine learning approach to identify environmental factors that affect life expectancy in contiguous U.S.

2. Methodology

The methodology we have implemented first involves gathering life expectancy data from IHME and a set of environment variables from ECMWF. We then match the temporal and spatial resolution of the environmental variables to our target variable, which is the life expectancy. We also construct new features based on other features and finally we develop our machine learning model. A brief description of the sources of data is given below:

2.1. IHME Data

The University of Washington houses an independent global health research center called the Institute for Health Metrics and Evaluation (IHME) [20,21]. The IHME creates and supports a data catalog called the Global Health Data Exchange (GHDx). Improving the health of the world's population is a mission of IHME supported by GHDx by providing information related to population health. IHME provides a collection of datasets among which we have used the United States Mortality Rates by Causes of Death and Life Expectancy by County, Race, and Ethnicity 2000-2019 (https://ghdx.healthdata.org/record/ihme-data/united-states-causes-death-life-expectancy-by-county-race-ethnicity-2000-2019, accessed on December 22 2025). Among the several csv files, we have used the "life expectancy estimates 2000-2019, both sexes". This life expectancy was estimated by IHME using the population and deaths data provided by the National Center for Health Statistics [21]. This csv file provides an estimate of life expectancy in the U.S. at county, state, and national level for each of the following:

- Age group: $<$ 1 year, 1 to 4, 5 to 9, 10 to 14, 15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 60 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 85+.
- Racial ethnic group: total population, Hispanic or Latino (Latino), non-Hispanic Black (Black), non-Hispanic White (White), non-Hispanic American Indian Alaska Native, Asian or Pacific Islander (Asian).

Among the 3 different geographic regions, various age and racial ethnic groups, we have considered the life expectancy in county of age group less than 1 year of the entire population instead of a particular ethnic group.

 doi:10.20944/preprints202604.1997.v1

2.2. Environmental Variables from ECMWF

The European Union's Space program has an Earth observation component called Copernicus. ECMWF is a research institute that plays a key role in Copernicus by producing and disseminating weather data. The Copernicus Atmosphere Monitoring Service (CAMS) [22] and the Climate Change Service (C3S) [23] are two of six services provided by Copernicus and implemented by the ECMWF. We gather a set of environmental variables from CAMS and C3S. A brief description of the datasets is given below:

2.2.1. CAMS Data

CAMS is a research institute that delivers reliable global data related to air pollution and health, greenhouse gasses, solar energy, and climate forcing. Among the various datasets provided by CAMS, we have used two of the datasets. The first is the CAMS global reanalysis (EAC4) (https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4?tab=overview, accessed on 22 December 2025) which stands for the ECMWF Atmoshperic Composition Reanalysis 4. EAC4 is the fourth generation global reanalysis of atmospheric composition produced by ECMWF. In the reanalysis procedure, the model data are combined with observations from all over the world. The second dataset that we make use of is the CAMS global reanalysis (EAC4) monthly averaged fields (https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4-monthly?tab=overview, accessed on 22 December 2025).

The CAMS EAC4 and EAC4 monthly averaged data are gridded data that cover the whole world with a horizontal resolution of $0.75^o \times 0.75^o$, and currently cover the years 2003 to 2024. The variables by EAC4 are available every 3 hours, whereas the EAC4 monthly averaged is available every one month. Vertical coverage consists of data at the surface level, total column, various model levels (or vertical distance), and pressure levels. The total column refers to the total amount of a chosen variable within a column of air extending from the Earth's surface up to the top of the atmosphere. Both of these datasets have variables available at different vertical regions of the atmosphere.

From CAMS EAC4 we have taken variables at the surface level and only one total column variable, which is the total column ozone. We have taken variables at the surface level as humans reside most of the life time at the surface. We have included total column ozone because of the large abundance of ozone in the ozone layer located in the stratosphere. Ultraviolet-B are rays in the wavelength band (280-315 nm) that can reach the surface of Earth causing sunburns, DNA damage, and some forms of cancer as well [24]. This ozone layer is essential because it absorbs ultraviolet radiation, protecting living beings from harmful rays. From the CAMS EAC4 monthly averaged data we have taken variables at the surface level. These variables were downloaded from their corresponding website using their Graphical User Interface (GUI). To obtain the surface level values, model level 60 is to be selected in the GUI. Some of the variables available from the CAMS EAC4 data is also available in the CAMS EAC4 monthly averaged, whereas some variables exist only in the CAMS EAC4 monthly averaged.

The list of variables that were used is given in Table 1 below. In Table 1, variables from 1 to 11 except Total column ozone are available only at the surface level. Variables 12 to 28 are multilevel variables, that is, these variables are available at various vertical distances of the atmosphere, but we chose the values at the surface level. Variables 29 to 32 are variables that are only available as monthly averages. Variables 12 to 27 have units kgkg^{-1}, to indicate that the quantity is expressed as the mass mixing ratio, that is, the amount of the variable in a mixture compared to the total amount of all other variables. Leaf area index, high vegetation includes portions of land covered by vegetation of tall heights such as evergreen trees, mixed forest/woodland, deciduous trees, and interrupted forest. Leaf area index, low vegetation includes land covered with vegetation of small heights, such as crops and mixed farming, irrigated crops, tundra, short and long grass, semidesert, marshed and bogs, evergreen and deciduous shrubs, and land and water mixtures. An index of 0 indicates bare land.

Among the list of variables, Figure 1 shows the 2 meter temperature in degrees Celsius at 3:00 PM on 1/1/2003 in the given latitude and longitude box. The boundary of the states of the contiguous U.S.

 doi:10.20944/preprints202604.1997.v1

is overlaid with black boundaries. The resolution is $0.75^o \times 0.75^o$ because of which the image looks grainy. The image shows that the southern region was warmer compared to the northern region in the given time. Most of the northwest regions are cooler than other regions as well.

Table 1. List of variables downloaded from CAMS.

S.N.	Variable	Unit	S.N.	Variable	Unit
1	10m u-component of wind	ms^{-1}	17	Hydroxyl radical	$kgkg^{-1}$
2	10m v-component of wind	ms^{-1}	18	Isoprene	$kgkg^{-1}$
3	2 meter dewpoint temperature	Kelvin	19	Nitric acid	$kgkg^{-1}$
4	2 meter temperature	Kelvin	20	Nitrogen dioxide	$kgkg^{-1}$
5	Mean sea level pressure	Pa	21	Nitrogen monoxide	$kgkg^{-1}$
6	PM_1	kgm^{-3}	22	Ozone	$kgkg^{-1}$
7	$PM_{2.5}$	kgm^{-3}	23	Peroxyacetyl nitrate	$kgkg^{-1}$
8	PM_{10}	kgm^{-3}	24	Propane	$kgkg^{-1}$
9	Surface geopotential	m^2s^{-2}	25	Specific humidity	$kgkg^{-1}$
10	Surface pressure	Pa	26	Sulphate aerosol mixing ratio	$kgkg^{-1}$
11	Total column ozone	kgm^{-2}	27	Sulphur dioxide	$kgkg^{-1}$
12	Carbon monoxide	$kgkg^{-1}$	28	Temperature	Kelvin
13	Dust aerosol (0.03-0.55 µm) mixing ratio	$kgkg^{-1}$	29	Leaf area index, high vegetation	m^2m^{-2}
14	Ethane	$kgkg^{-1}$	30	Leaf area index, low vegetation	m^2m^{-2}
15	Formaldehyde	$kgkg^{-1}$	31	Snow albedo	(0-1)
16	Hydrogen peroxide	$kgkg^{-1}$	32	Snow depth	meter of water equivalent

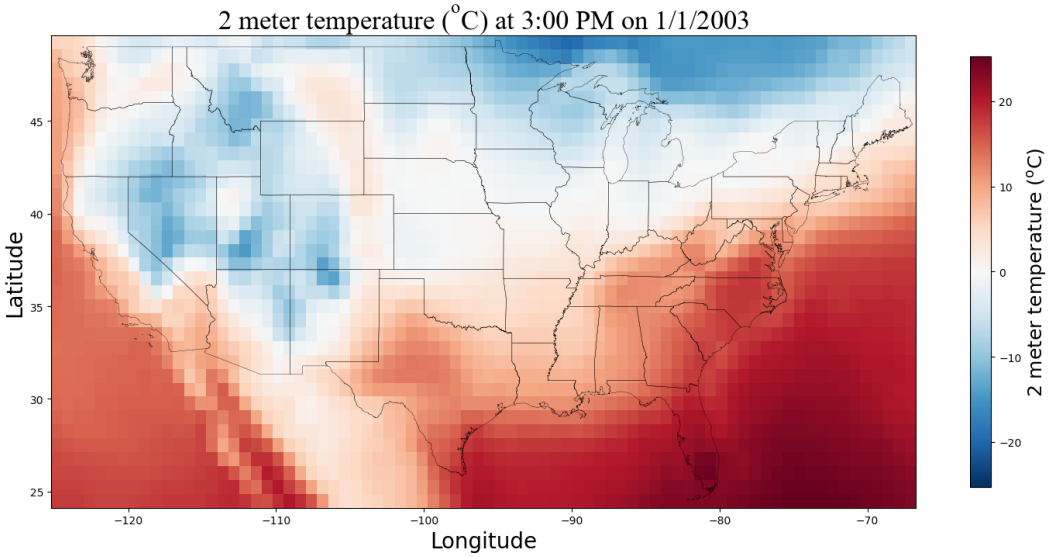


Figure 1. 2 meter temperature (°C) at 3:00 PM on 1/1/2023 with the boundary of contiguous states overlaid. Data source: CAMS EAC4 [22].

2.2.2. ERA5 Data

The ERA5 data are available from 1940 to the present day. The variables are available every hour but were downloaded with a resolution of every 3 hours because of the limitation with the number of

 doi:10.20944/preprints202604.1997.v1

data that can be downloaded at a single time. Among the variables available from ERA5, we have considered relative humidity as it was not available in CAMS EAC4. Figure 2 shows a timestamp of relative humidity data at 3:00 PM on 1/1/2003. The resolution of the data is 0.25°×0.25° because of which the image looks smoother. The figure clearly shows a higher relative humidity in the southern and western parts of the United States compared to other regions at that timestamp.

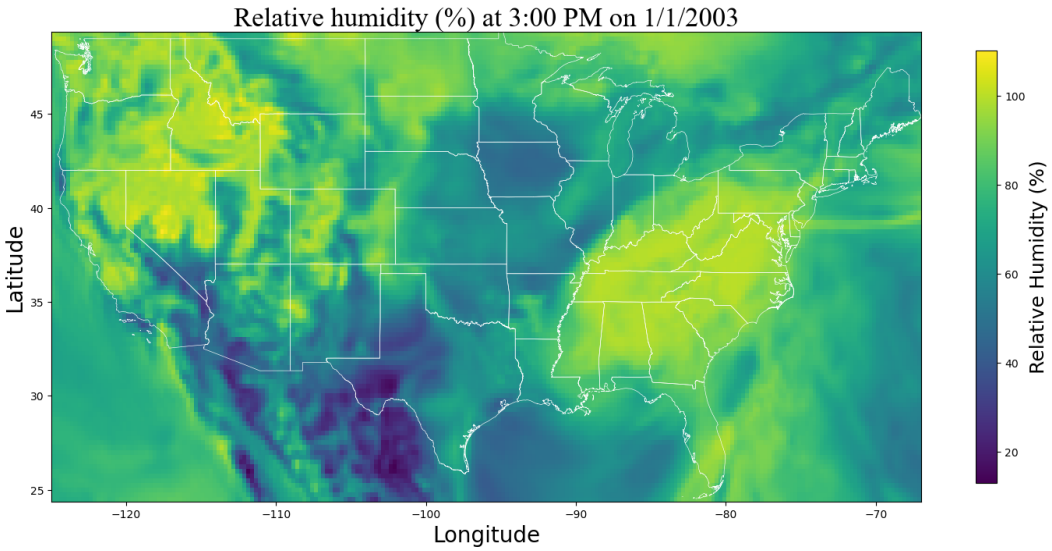


Figure 2. A timestamp of Relative humidity (%) at 3:00 PM on 1/1/2023 with the boundary of contiguous states overlaid. Data source: ERA5 [23].

2.3. Matching Temporal and Spatial Resolution

The data sources we have taken in this study have different temporal and spatial resolutions. A summary of the resolutions is given in Table 2 below.

Table 2. List of sources of data and corresponding temporal and spatial resolution.

Data source	Temporal resolution	Spatial resolution
CAMS EAC4	2003 to 2024, every 3 hour	0.75°×0.75°
CAMS EAC4 monthly averaged	2003 to 2024, 1 month	0.75°×0.75°
ERA5	1940 to present, every hour[1]	0.25°×0.25°
IHME	2000 to 2019, annual	Counties of USA

[1] Data was downloaded every 3 hour because of the limitation of the number of data that can be downloaded at a single time.

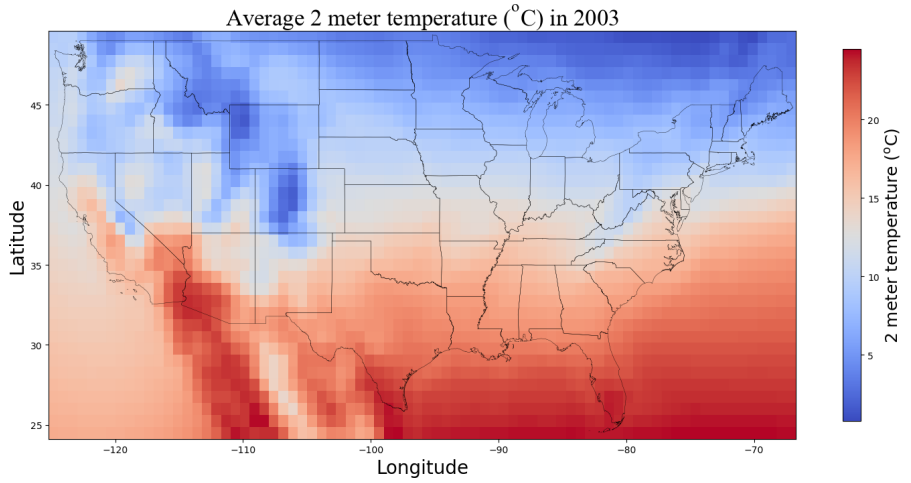
The target variable we are studying is life expectancy data from IHME because of which we match the temporal and spatial resolution of rest of the data to IHME data. The temporal and spatial resolution was resolved in the following way:

- Temporal resolution: The life expectancy data from IHME is annual. We calculated a one year average of each of the variables from CAMS EAC4, CAMS EAC4 monthly averaged, and ERA5. For example: the temporal resolution of the CAMS EAC4 data is 3 hours, so 1 year has a total of (24/3)×365 = 2960 data points. The average of these 2960 data points was calculated.
- Spatial resolution: The CAMS and ERA5 data are raster data, whereas the IHME data are at the county level. To resolve this, multilinear interpolation was performed in which the raster data from each of the sources of the data were interpolated to the boundary points of the counties. These boundary points were accessed from the shapefile of the counties for a particular year. Shapefiles are a set of files that store the location, shape, and attributes of geographic features such as points, lines, and polygons. These shapefiles were downloaded from the United States Census Bureau website (https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.2010.html#list-tab-1556094155, accessed on December 22 2025). The shapefiles

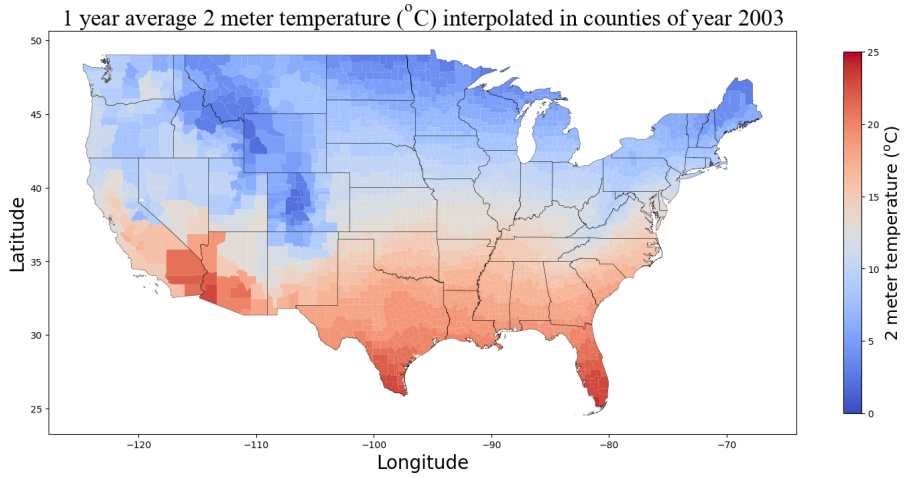
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2026 doi:10.20944/preprints202604.1997.v1

of the years 2003 and 2007 could not be found, so for these years shapefiles of the years 2008 were used. The pair of latitude and longitude points available for a county from a shapefile can vary. For counties that have more than 100 pairs of latitude and longitude available, multilinear interpolation of each of the variables was done on 100 pairs of latitude and longitude and then averaged. For counties where there were less than 100 pairs of longitude and latitude pairs available, multilinear interpolation of each of the variables was done to that many pairs of latitude and longitude and then averaged. The total number of counties in each year, the smallest number of pair of latitude and longitude available among all counties, and the total number of counties where there were less than 100 pair of latitude and longitude pairs available for each year are given in Table A1 in Appendix A.

Figure 3a shows the raster data where the average of 2 meter temperature of the year 2003 was calculated using the data from CAMS EAC4. Figure 3b shows the corresponding values in the counties after the raster data was multilinearly interpolated to 100 or less pairs of latitude and longitude pairs (as available from the shapefile) and then averaged. The calculated values in Figure 3a are almost identical to the interpolated values in Figure 3b, indicating an excellent representation of the variable at the county level.



(**a**)



(**b**)

Figure 3. (**a**) Raster plot of 1 year average of 2 meter temperature of the year 2003 calculated using data from CAMS EAC4 with boundaries of the states overlaid. (**b**) Plot where multilinear interpolation was performed on boundary points of counties and then averaged to find corresponding values of the variable in each county with boundaries of states overlaid.

 doi:10.20944/preprints202604.1997.v1

2.4. Feature Engineering

The data from CAMS EAC4 and CAMS EAC4 monthly averaged (32 variables) and the relative humidity data from ERA5 make up a total of 33 variables. We feature engineer (the process of constructing new features) based on a subset of these 33 features. A set of two feature engineered variables are explained below:

- Wind velocity: The 10m u-component of wind and the 10m v-component of wind are the eastward and northward components of the wind at a height of 10 meters, respectively. Using these two features, we feature engineer wind velocity using the following formula:

$$\text{Wind Velocity} = \sqrt{(\text{10m u-component of wind})^2 + (\text{10m v-component of wind})^2} \quad (1)$$

  This process was done to reduce the number of features. Instead of using two different features: the 10m u-component of wind and 10m v-component of wind, we instead use wind velocity.
- Wet bulb temperature: Wet bulb temperature is measured by wrapping the bulb of a thermometer with a wet muslin. Measurement of this temperature is important because it indicates how much a body can be cooled by the process of evaporation. Dry air can absorb water faster than humid air. Lower levels of humidity lead to a lower wet bulb temperature, and our body can stay cooler through evaporation. However, if the relative humidity is 100%, the wet bulb temperature will be identical to the dry bulb, since the water in the wet muslin cannot evaporate. The wet bulb temperature is calculated as follows [25]:

$$\begin{aligned}\text{Wet bulb temperature} = \text{T}\arctan[0.151977(\text{RH\%} + 8.313659)^{1/2}] + \arctan(\text{T} + \text{RH\%}) \\ - \arctan(\text{RH\%} - 1.676331) \\ + 0.00391838(\text{RH\%})^{3/2}\arctan(0.023101\text{RH\%}) - 4.686035]\end{aligned} \quad (2)$$

  where, air temperature is expressed in ^{o}C and relative humidity is expressed in %. We calculated the 1 year average temperature and 1 year average relative humidity of a each year and used Equation 2 to calculate the Wet bulb temperature for that particular year.

Replacing the 10m u-component of wind and 10m v-component of wind by wind velocity and adding wet bulb temperature in the list of features makes up a total of 33 features.

We also feature engineer further 12 variables to consider extreme cases such as high temperature, low temperature, and fraction of time (FoT) where a county exceeded the threshold of $PM_{2.5}$ standard set by the Environmental Protection Agency (EPA). Exposure to heat can affect the regulation of internal temperature in a human body [26]. Study has shown that extreme high temperatures associated with human mortality during the 2010s in the United States [27]. Short term exposure to extreme cold temperatures below 5 ^{o}C can lead to cold stress, drastic physiological changes such as change in heart rate and blood pressure, and affect the ability to make judgments [28]. Health effects also tend to be prominent among elderly people in high temperature conditions compared to conditions in low temperature [29]. The list of 12 features that were engineered is given in Table 3.

Among the features shown in Table 3, for features 11 and 12, the corresponding threshold was set and the number of times a county was above the threshold was calculated from the 2920 timestamps of the data provided by CAMS EAC4 and then multiplied by 100. The threshold set by EPA for PM_{10} particles is 150 $\mu g/m^3$ for a 24-hour average and the threshold set by EPA for $PM_{2.5}$ is 9.5 $\mu g/m^3$. For the rest of the variables, after the values of the variable in a county were calculated as mentioned in Section 2.3, the 90^{th} or the 10^{th} percentile were calculated among all counties in the particular year. From the set of 2960 timestamps as provided by CAMS EAC4, the number of times the variables were above the 90^{th} percentile (or below the 10^{th} percentile) multiplied by 100 was calculated.

With the set of 12 feature engineered variables, the total number of features we have considered is 45.

 doi:10.20944/preprints202604.1997.v1

Table 3. List of feature engineered variables. Fraction of Time is abbreviated as FoT.

S.N.	Variable	S.N.	Variable
1	FoT formaldehyde above 90th percentile	7	FoT temperature above 90th percentile
2	FoT hydroxyl radical above 90th percentile	8	FoT temperature below 10th percentile
3	FoT isoprene above 90th percentile	9	FoT temperature above 90 ^{o}F
4	FoT Peroxyacetyl nitrate above 90th percentile	10	FoT temperature below 0 ^{o}C
5	FoT ozone above 90th percentile	11	FoT $PM_{2.5}$ above EPA threshold
6	FoT hydrogen peroxide above 90th percentile	12	FoT PM_{10} above EPA threshold

2.5. Machine Learning Model Development

Before we make use of machine learning, we first test if there is a subset of features that are linearly correlated, as we used a total of 45 features. The information carried by one of the features will be almost identical to the rest of the features with which it is highly linearly correlated, and we can choose a single feature from a group of multi collinear features, which will reduce the number of features. To identify the cluster (or groups) of features that are linearly correlated, we make use of a dendrogram. A dendrogram and further details are explained in Appendix B.

From the dendrogram, we can identify a set of 9 clusters (or groups) of features that are multi-collinear, which are shown in Table 4. As each of the features in the group is linearly correlated with each other, we can use any of the features, as they consist of similar information. We have used features that are easier to interpret among these set of features. The feature that was chosen is mentioned first in each of the groups.

Table 4. List of 9 clusters of variables that are multicollinear identified from the dendrogram. A total of 24 features.

S.N.	Variable	S.N.	Variable
1	$PM_{2.5}$, FoT $PM_{2.5}$ above EPA threshold, PM_1, PM_{10}	6	Surface pressure, Surface geopotential
2	Nitrogen dioxide, Nitrogen monoxide	7	Peroxyacetyl nitrate, FoT Peroxyacetyl nitrate above 90th percentile
3	Specific humidity, 2m dew point temperature	8	Isoprene, FoT Isoprene above 90th percentile
4	Wet bulb temperature, FoT temperature below 0 ^{o}C, FoT temperature below 10th percentile, 2m temperature, Temperature	9	Formaldehyde, FoT fromaldehyde above 90th percentile
5	Hydrogen peroxide, FoT hydrogen peroxide above 90th percentile, FoT temperature above 90th percentile		

The remaining set of 21 features are given in Table 5.

This provides us with a total of 30 features, among which 24 features are distributed in 9 different groups, and 21 features are not multicollinear with any other feature.

The final tabular dataset where we make use of machine learning consists of 31 columns which includes 30 features and 1 target variable (the life expectancy), and 52,320 number of rows (or samples of data) from the 17 years of life expectancy data. To estimate the life expectancy we utilize the Random Forest algorithm [30] using scikit-learn [31] for non-linear, multidimensional data. As our dataset is tabular, tree-based models such as Random Forest tend to be more efficient than other models [32,33]. 80% of the dataset were randomly chosen for training and the rest of 20% of the dataset was used as an independent test set. To quantify the accuracy of the estimation, the root mean square error (RMSE), the coefficient of determination (R^2), and adjusted R^2 were calculated between the true and estimated values. The adjusted R^2 takes into account the number of features in a model, whereas the coefficient of determination does not. The adjusted R^2 was calculated using the following formula:

 doi:10.20944/preprints202604.1997.v1

$$1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{3}$$

where
R^2 is the coefficient of determination between the true and estimated values
k is the total number of variables in the regression model
n is the number of samples in the data

Table 5. List of remaining 21 features that are not linearly correlated with any other feature.

S.N.	Variable	S.N.	Variable
1	Carbon monoxide	15	Mean sea level pressure
2	Ethane	16	Leaf area index, high vegetation
3	Propane	17	FoT PM_{10} above EPA threshold
4	Nitric acid	18	Hydroxyl radical
5	Sulphate aerosol mixing ratio	19	FoT hydroxyl radical above 90^{th} percentile
6	Sulphur dioxide	20	Relative humidity
7	Snow depth	21	FoT temperature above 90^{o} F
8	Snow albedo		
9	Total column ozone		
10	Dust aerosol (0.03-0.55 µm) mixing ratio		
11	10m wind speed		
12	Leaf area index, low vegetation		
13	Ozone		
14	FoT ozone above 90^{th} percentile		

To qualitatively assess the goodness of fit, we also plot a scatter plot and a quantile-quantile plot between the true and estimated values of life expectancy.

To identify the features that were the most important in estimating life expectancy, we make use of permutation feature importance and the SHAP (Shapley additive explanations) values [34,35] from the SHAP library. In the permutation feature importance, a single feature is randomly shuffled and the degradation of the model is observed. This process is done for every feature of the model. If the degradation is high after a feature is shuffled, the feature is considered to be of high importance, as there is high dependency between the feature and the target variable. On the other hand, if the degradation is low, then the feature is considered as low importance as the feature has low dependency with the target variable. If any of the features is highly linearly correlated with any other feature, then when the feature is permuted, the model can still have access to the feature from the other feature with which it is linearly correlated. This was also one of the reasons for getting rid of multicollinearity among features using a dendrogram. Further details about permuation feature importance is given in Appendix C. Permutation feature has a major limitation. "Permutation importance does not reflect to the intrinsic predictive value of a feature by itself but how important this feature is for a particular model" [36]. Due to this limitation, we also used SHAP values, which uses a game theory approach to identify the importance of features.

3. Results

We first identify the disparity in life expectancy in contiguous U.S. of age group less than 1 year old from the year 2003 to 2019. Then we estimate the life expectancy in the U.S. using a set

 doi:10.20944/preprints202604.1997.v1

of environmental factors and also identify the environmental factors that were most influential in affecting the life expectancy.

3.1. Identifying Disparity in Life Expectancy

We used county wise life expectancy data of age group less than 1 year old as provided by IHME from the year 2003 through 2019. Table 6 shows the year and the corresponding county having the highest and lowest life expectancy of the year. For each year, the highest life expectancy was between two counties in Colorado and the county with the lowest life expectancy was Oglala Lakota County in South Dakota. In each year, the difference between the highest and lowest life expectancy is greater than 20 years.

Table 6. The highest and lowest life expectancy in the years 2003 to 2019 in contiguous US. The highest life expectancy is between two counties both of which are located in Colorado and the county with lowest life expectancy in each of the year is Oglala Lakota County in South Dakota.

Year	Highest life expectancy (Year)	County	Lowest life expectancy (Year)
2003	87.0	Pitkin County	65.4
2004	87.6	Pitkin County	65.7
2005	87.6	Summit County	65.8
2006	88.2	Pitkin County	65.8
2007	88.7	Summit County	66.0
2008	88.8	Summit County	66.4
2009	89.5	Summit County	66.3
2010	89.6	Summit County	66.6
2011	89.4	Summit County	66.5
2012	89.7	Summit County	66.6
2013	89.9	Summit County	66.2
2014	90.4	Summit County	66.0
2015	90.0	Summit County	65.9
2016	90.7	Summit County	65.3
2017	91.0	Summit County	65.3
2018	91.6	Summit County	65.1
2019	92.2	Summit County	65.4

Figure 4 shows the life expectancy of the age group less than 1 year old in the contiguous US of the years 2003, 2010, 2015 and 2019. The white spaces indicate counties where the life expectancy data do not exist. A gif file showing the life expectancy from 2003 to 2019 is available in the following link: (https://github.com/mi3nts/Life_Expectancy_Complete/blob/main/Figures/True_LE_plots/True_LE_gif.gif, accessed on December 28, 2025). A visual inspection of the gif file shows that the overall map appears to go from dark to light, indicating the overall life expectancy to increase over the years. However, a massive disparity in life expectancy is also observed in each year, where counties in southern states have relatively lower life expectancy compared to counties in northern states.

 doi:10.20944/preprints202604.1997.v1

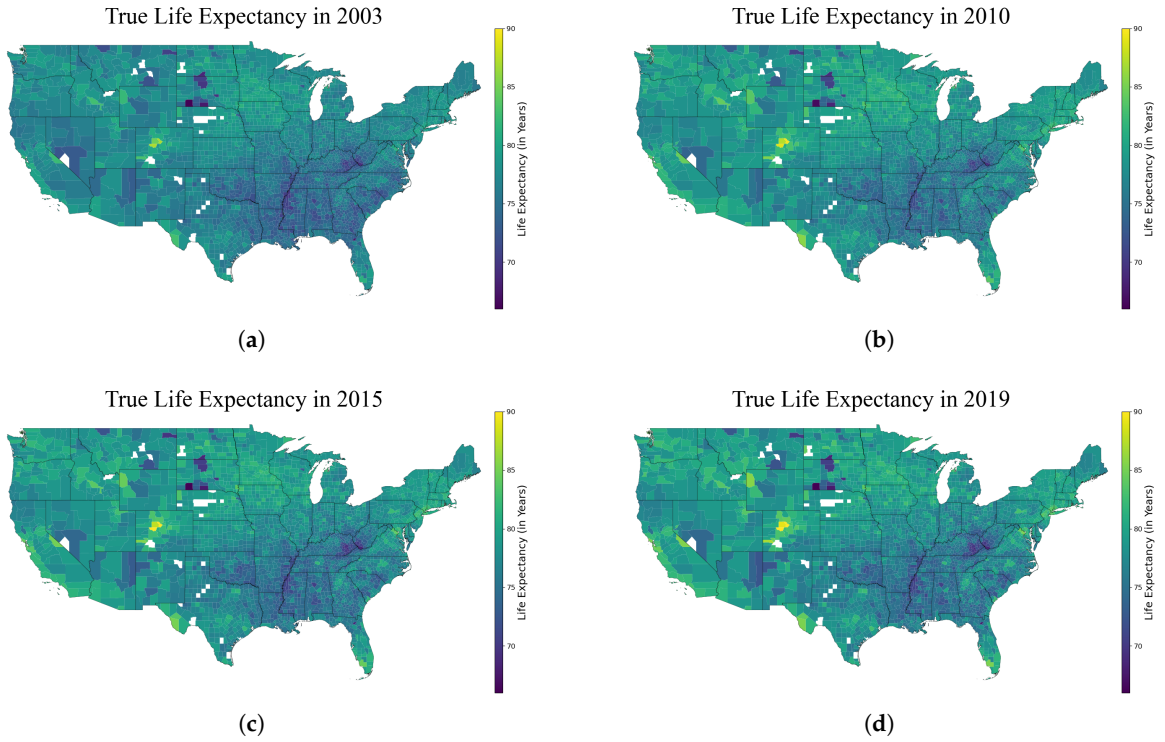


Figure 4. True life expectancy of age group less than 1 year old in the year (**a**) 2003, (**b**) 2010, (**c**) 2015, (**d**) 2019. White spaces indicate counties with no life expectancy data.

3.2. Estimating Life Expectancy in the U.S. Using Environmental Variables and Identifying Environmental Variables Affecting Life Expectancy

We estimated life expectancy in the US using a set of environmental variables and identified environmental factors that affect life expectancy. We first use a set of 35 features and then reduce the number of variables to test if the same accuracy can be obtained.

3.2.1. Using 30 Features

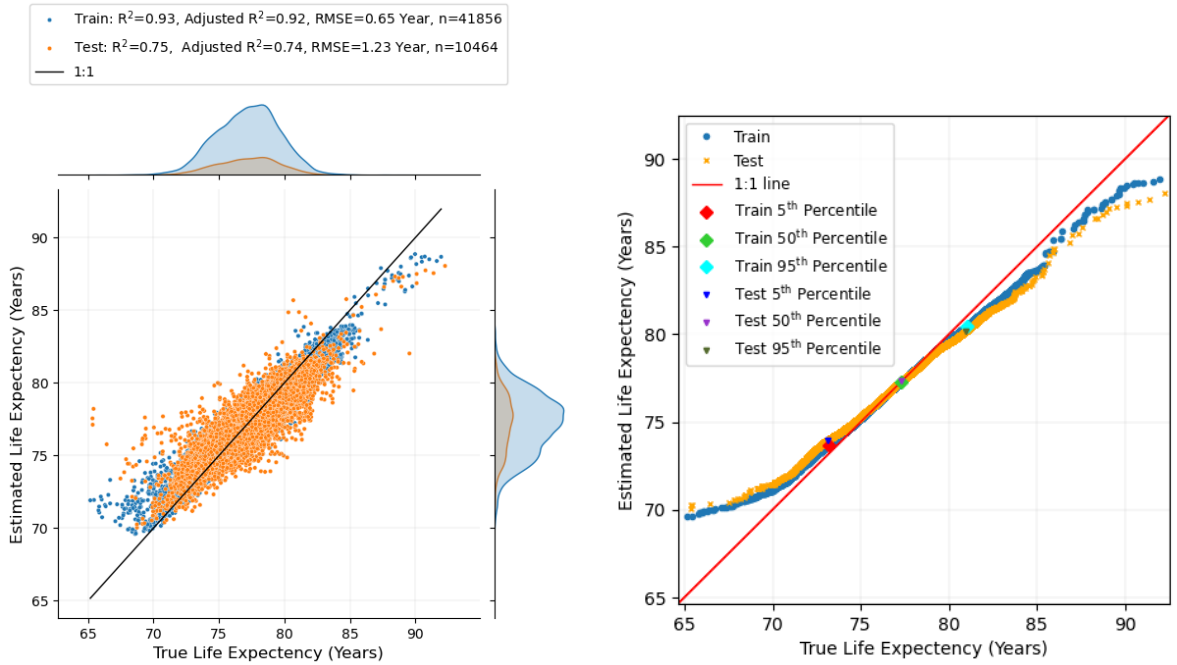
We use a set of 30 features to estimate life expectancy in the US using the random forest algorithm from scikit learn with optimized hyperparameters as explained in Section 2.5. The list of hyperparameters that were optimized and the one selected is given in Table A2 in Appendix D. The list of 9 features is given in Table 4 where we have used the first feature from each of the 9 groups of feature and the remaining 21 features are given in Table 5. The scatter plot and the quantile-quantile plot of estimation using 30 features are given in Figures 5a and 5b, respectively. The quantified metric of goodness of fit is also shown on top of Figure 5a.

The result of the estimation was found to be moderate when we used 30 features. The R^2, adjusted R^2 and RMSE in the training set were found to be 0.93, 0.92 and 0.65 years, respectively. This high accuracy is expected, as this part of the data was used for training. The accuracy in the testing set was found to be moderate as we obtained the R^2, adjusted R^2 and RMSE to be 0.75, 0.74 and 1.23 years respectively.

Figure 5a shows the scatter plot of estimation. The density curves on the top and right sides of the figure indicate that for a large number of counties from 2003 through 2019, the life expectancy was between 70 and 85. As we obtained the R^2 value of 0.93 and the RMSE of 0.65 years in the training set, the data points of this set are closer to the 1:1 line compared to that in the testing set, where we can observe that a large number of data points deviate from the perfect 1:1 line. Figure 5b shows the quantile-quantile plot of the estimation where the quantiles of the true life expectancy and the quantiles of the estimated life expectancy are plotted on the X and Y axes, respectively. The 5^{th}, 50^{th}, and 95^{th} percentiles of each distribution are also overlaid. The plot shows that the estimate was almost close to

 doi:10.20944/preprints202604.1997.v1

the perfect 1:1 line for data points between the 5th and 95th percentiles, whereas the estimation deviates from the 1:1 line below the 5th and above the 95th percentile in each of the training and testing set.



(**a**) (**b**)

Figure 5. (**a**) Scatter plot of true values of life expectancy plotted against the estimated values of life expectancy when 30 features were used for estimating life expectancy. A perfect fit is indicated by the black 1:1 line. (**b**) Quantile-Quantile plot of true values of life expectancy plotted against the estimated values of life expectancy when 30 features were used for estimating life expectancy. A perfect fit is indicated by the red 1:1 line.

As the deviation from the perfect 1:1 line is below the 5th percentile and above the 95th percentile, we also visualize the life expectancy data below the 5th percentile and above the 95th percentile, which is shown in Figure 6.

The following link (https://github.com/mi3nts/Life_Expectancy_Complete/blob/main/Figures/Extreme_LE_plots/Extreme_LE_gif.gif, accessed on December 28 2025) shows a gif file of images from the year from 2003 through 2019 that indicates counties with a life expectancy less than the 5th percentile and greater than the 95th percentile. Each of the figure, shows that most of the counties with life expectancy below the 5th percentile are located in the southern states whereas counties with life expectancy above 95th percentile are scattered throughout the contiguous US.

In order to identify the importance of features that affect life expectancy, we first used the permutation feature importance. Figure 7 shows the feature ranking of 20 features in the training set. A complete feature ranking of 30 features is given in the link: (https://github.com/mi3nts/Life_Expectancy_Complete/blob/main/Results/permutation_importance_30_train.png, accessed on December 28 2025). Each of the features was permuted 10 times, and the mean value of the degradation of the model was calculated. The features are arranged in descending order of the mean value, and the standard deviation is plotted as an error bar.

Figure 7 shows that leaf area index, low vegetation, and specific humidity are among the most influential features that affect life expectancy. Other pollutants such as formaldehyde, sulphate aerosols, dust aerosols, and sulfur dioxide are among the most important features; while pollutants such as $PM_{2.5}$, ozone were in the bottom half of the ranking, indicating there relative importance to be small compared to other variables at the top. As the standard deviation of each of the features is small, the ranking of the features will tend to be the same when the algorithm is rerun. As there were considerable differences in accuracy in estimating life expectancy in the training and testing set, a

 doi:10.20944/preprints202604.1997.v1

feature ranking was also conducted on the testing set. The permutation feature ranking in the testing set is given in Figure A2 of the Appendix E. The ranking in the testing set is almost identical.

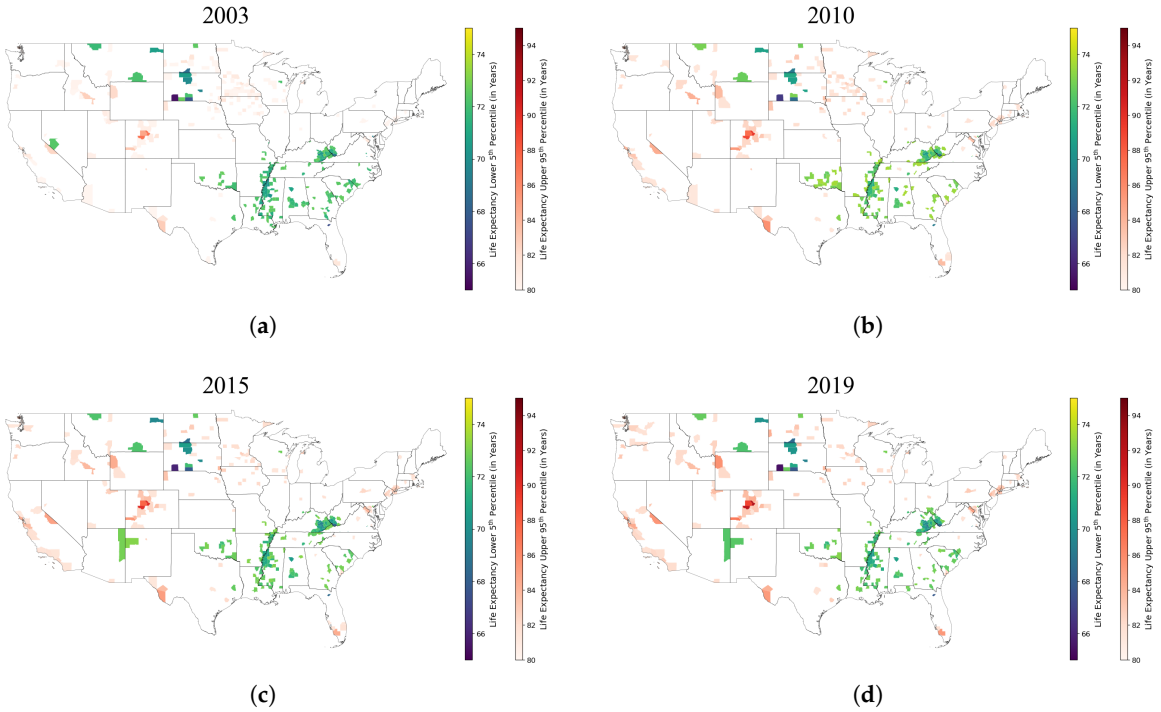


Figure 6. Life expectancy in the year (**a**) 2003, (**b**) 2010, (**c**) 2015, and (**d**) 2019 where the counties below the 5th percentile are represented by color pallette on the left and above 95th percentiles are shown by the color palette on the right.

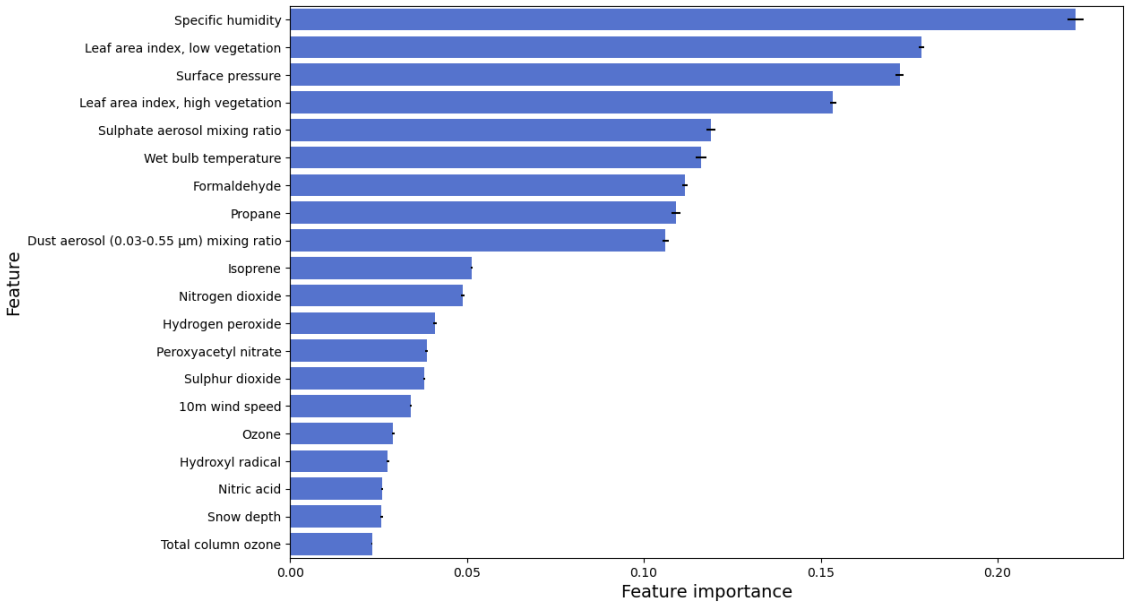


Figure 7. Feature ranking of variables that were most influential to estimate life expectancy according to permutation feature importance in the training set when 30 features were used for estimation. Features are arranged in descending order with features that were most influential at the top. Top 20 of the features is shown.

Figure 8 shows the SHAP values of the 20 top features in the training set that indicate the importance of feature in descending order with the most influential feature at the top. A complete list of features is available in the following link: (https://github.com/mi3nts/Life_Expectancy_Complete/blob/main/Results/shap_30_train.png, accessed on December 28 2025).

 doi:10.20944/preprints202604.1997.v1

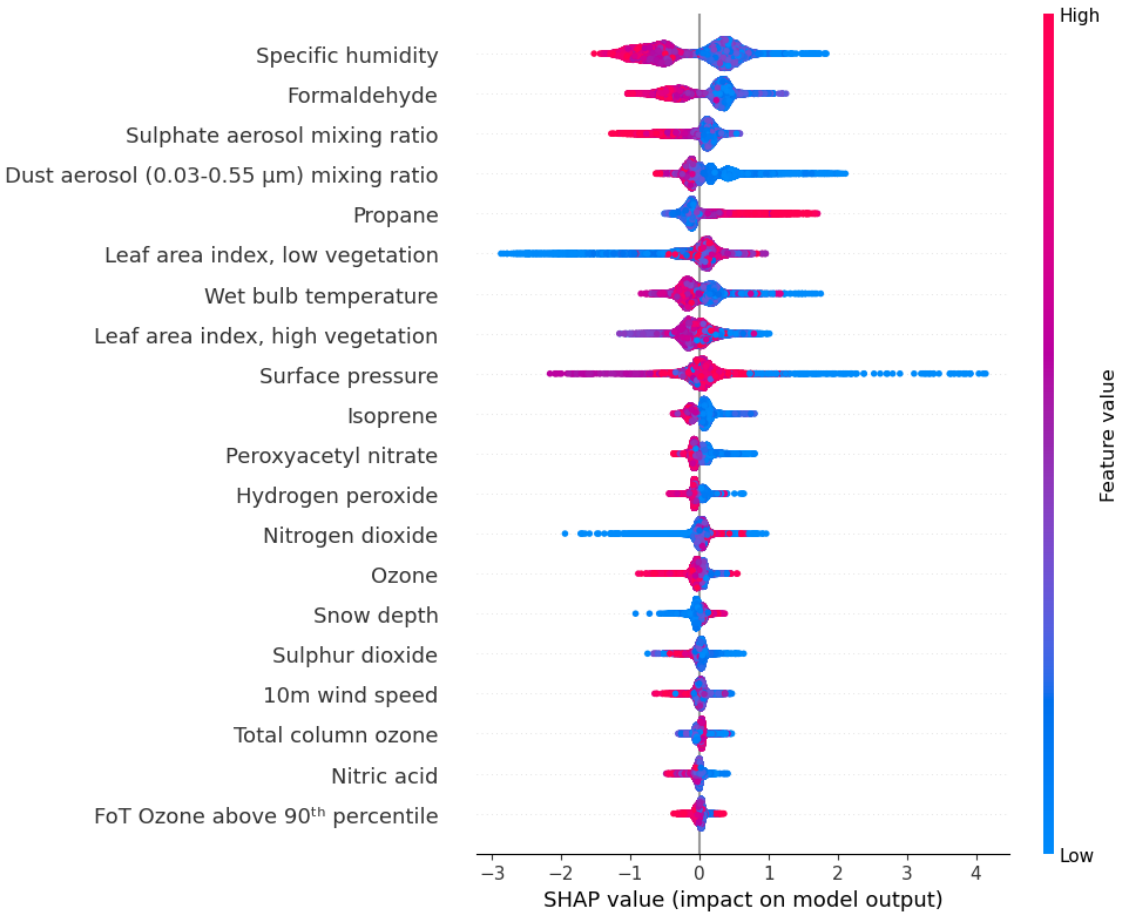


Figure 8. Feature ranking of variables that were most influential to estimate life expectancy according to SHAP values in the training set when 30 features were used for estimating life expectancy. Features are arranged in descending order with features that were most influential at the top. X-axis is in the units of year (unit of life expectancy). Each of the feature has 41,846 circular dots representing the data points of the training set. Higher feature values are represented in red and lower feature values are in blue. Data points of each of the feature having identical SHAP value are stacked on top of each other.

The feature ranking in Figure 7 is not consistent with the permutation feature ranking in Figure 8. However, the main variables remain consistent, such as formaldehyde, leaf area index, low vegetation, sulfate aerosol mixing ratio, specific humidity, wet bulb temperature, indicating the relatively high importance of these variables compared to other variables. These SHAP values also indicate that for pollutants such as formaldehyde, higher values tend to lower the life expectancy whereas lower values tend to increase life expectancy as higher values of formaldehyde (indicated in red) are on the left of the central black vertical line, whereas lower values of formaldehyde (indicated in blue) are on the right side of the vertical central black line. Similarly, for the feature leaf area index: low vegetation, higher index of leaf vegetation (portion of land covered by vegetation of smaller heights) tends to increase life expectancy, whereas lower values tend to decrease life expectancy. However, a mixture of increase and decrease in life expectancy can be seen in the case of leaf area index, high vegetation. As there was a performance difference in the training and testing set, the SHAP values in the testing set were also calculated. The SHAP values in the testing set for the top 20 features and the entire list is identical to those of the training set. The SHAP values in the testing set in given in Figure A3 in Appendix E.

Figure 9 shows the distribution of six variables in the year 2019.

 doi:10.20944/preprints202604.1997.v1

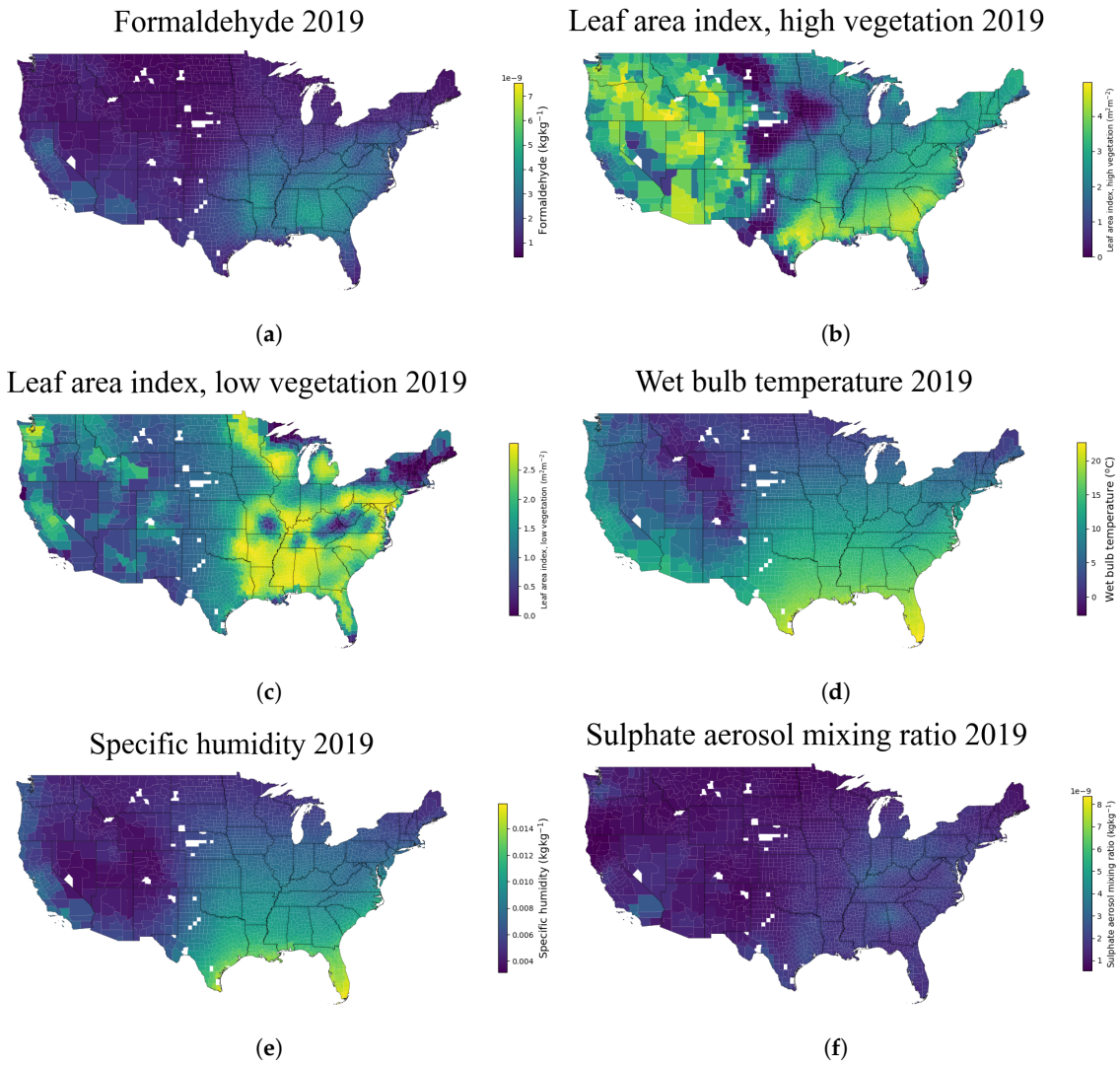


Figure 9. Distribution of (**a**) Formaldehyde, (**b**) Leaf area index, high vegetation (**c**) Leaf area index, low vegetation, (**d**) Wet bulb temperature, (**e**) Specific humidity, and (**f**) Sulphate aerosol mixing ratio in the year 2019 in counties of contiguous USA. White spaces indicate counties where life expectancy data is not available. Formaldehyde, specific humidity, and sulphate aerosol mixing ratio are in the units of $kgkg^{-1}$ which indicates a mass mixing ratio. It is the amount of formaldehyde (or sulphate aerosol mixing ratio or mass of water vapor incase of specific humidity) in a mixture compared to the total amount of all other variables. $kgkg^{-1}$ of formaldehyde and sulphate aerosol mixing ratio can be converted to ppm by multiplying the value in $kgkg^{-1}$ by 10^6. Leaf area index, low vegetation refers to vegetation of small height such as short and tall grass, evergreen shrubs. Leaf area index, high vegetation refers to vegetation of tall heights such as evergreen trees, deciduous trees. Higher the index of these vegetation, higher the portion of the land covered by these vegetation. An index of 0 indicates barren land.

A gif file of these variables from the year 2003 to 2019 is available in the following link (https://github.com/mi3nts/Life_Expectancy_Complete/tree/main/create_gifs, accessed on December 28 2025) in the corresponding folders having the extension .gif.

Figure 9a shows the high concentration of Formaldehyde in the counties of southern regions and in California. As leaf area such as shrubs and trees tends to rarely change throughout the years, changes in leaf area index of both high and low vegetation throughout the years are small. The wet bulb temperature and specific humidity also seems to be high in the southern region of the contiguous US throughout the years. Only a small number of counties have a higher concentration of sulphate aerosols compared to other counties, but it was an important factor affecting life expectancy.

 doi:10.20944/preprints202604.1997.v1

3.2.2. Using Reduced Number of Features

In this section, we estimate the life expectancy using reduced number of features identified by permutation feature importance, the ranking of which was almost similar to the ranking by SHAP values. We used the top 15, 10, 5 and 3 features from the permutation feature importance as shown in Figure 7 and estimated the life expectancy using the Random Forest algorithm from scikit-learn with optimized hyperparameters. The hyperparameters that were optimized and the ones selected as given in Table A2 in Appendix D.

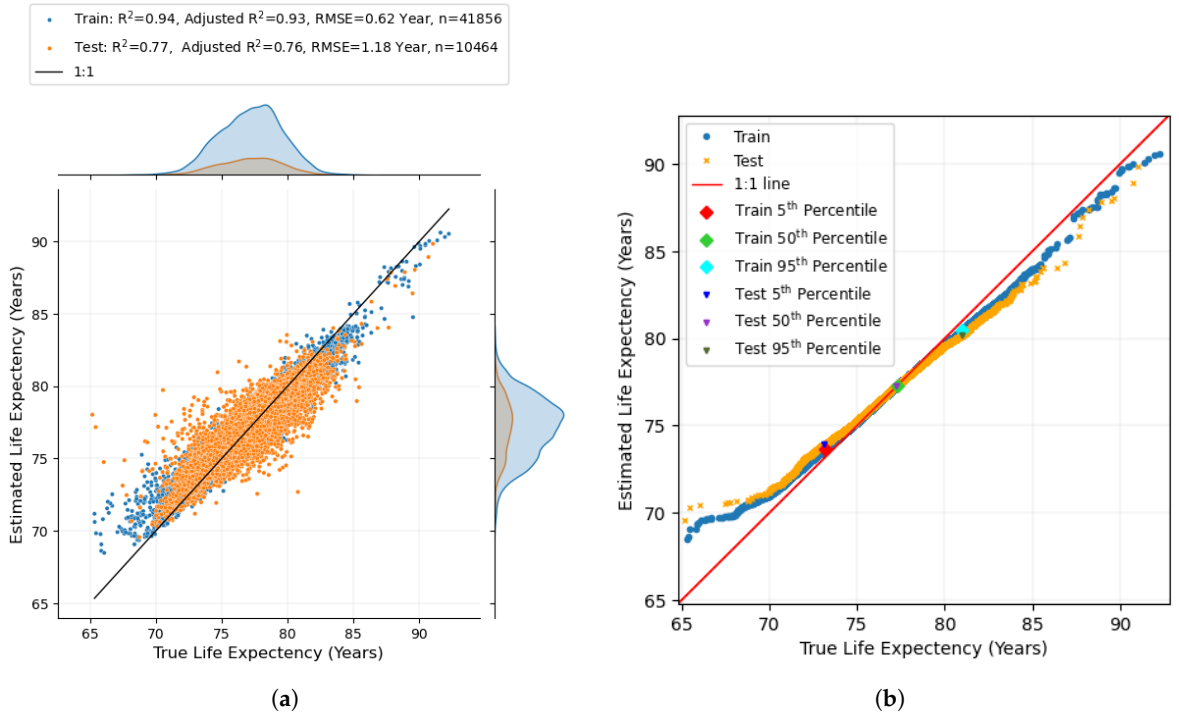
A summary of the results of estimating life expectancy is given in Table 7.

Table 7. Summary of results of estimating life expectancy using 30 features and reduced number of features.

Number of Features	Train RMSE (year)	Test RMSE (year)	Train R^2	Test R^2	Train adjusted R^2	Test adjusted R^2
30	0.64	1.28	0.93	0.72	0.92	0.71
15	0.62	1.22	0.94	0.75	0.93	0.74
10	0.63	1.18	0.94	0.77	0.93	0.76
5	0.62	1.18	0.94	0.77	0.93	0.76
3	0.71	1.16	0.92	0.77	0.91	0.76

Table 7 shows that instead of using 30 set of features we can use a set of just 5 features to better estimate life expectancy. The set of 5 features include: Specific humidity, Leaf area index, low vegetation; Leaf area index, high vegetation; sulphate aerosol mixing ratio and wet bulb temperature.

A scatter plot and quantile-quantile plot when 5 features were used for estimating life expectancy is shown in Figure 10. The structure of scatter plot and quantile-quantile plot is similar to Figure 5a and Figure 5b respectively. As there was improvements in estimation of life expectancy when 5 features were used, the data points are closer to their respective 1:1 line compared to when 30 features were used.



(**a**) (**b**)

Figure 10. (**a**) Scatter plot of true values of life expectancy plotted against the estimated values of life expectancy when 5 features were used for estimating life expectancy. A perfect fit is indicated by the black 1:1 line. (**b**) Quantile-Quantile plot of true values of life expectancy plotted against the estimated values of life expectancy when 5 features were used for estimating life expectancy. A perfect fit is indicated by the red 1:1 line.

 doi:10.20944/preprints202604.1997.v1

4. Discussion

Using life expectancy data for the age group less than 1 year old as provided by IHME, we achieve our first objective, as we have shown the enormous disparity in life expectancy in the contiguous US. As shown in Table 6 the difference between the highest and lowest life expectancy was greater than 20 years from the years 2003 through 2019. More information on the disparity of life expectancy can be found in Figure 4 and the gif of life expectancy from 2003 through 2019 where counties in southern states have relatively lower life expectancy compared to most counties in northern states. Figure 6 which shows the lower 5th and upper 95th percentile also augments the finding as the upper 95th percentile in life expectancy was scattered around the US whereas the lower 5 th percentile was mostly concentrated in the lower counties of the US.

Our study also showed that we can obtain moderate accuracy in estimating life expectancy in the US using a series of environmental variables. As summarized in Table 7, instead of 30 features, we can use a set of 5 features to achieve even better accuracy in estimating life expectancy. As 5 features provided better result, the adjusted R^2 in the test set is higher compared to when 30 features were used, indicating that a small number of features were sufficient. This aligns with the Occams razor principle, which states that simpler models usually generalize well. Furthermore, this also shows that an increase in the number of variables does not necessarily increase the accuracy of the results as well. Our results were likely distorted due to the small number of data points below the 5th percentile and above the 95th percentile, as shown in the quantile-quantile plot in Figures 5b and 10b. As machine learning models require a large number of data points to learn from a training set and then to be tested in an independent test set, due to the small number of data points in this region, the estimation deviated from the perfect 1:1 line. However, the estimation was nearly perfect between the 5th percentile and the 95th percentile, where data points were abundant.

Our key finding is the feature ranking as shown in Figures 7 and 8. These ranking shows formaldehyde to be one of the top factors affecting life expectancy in the US. Formaldehyde, which is found primarily in manufactured wood products, materials used in bulding, household products such as paints, glues, preservatives used in some medicines and cosmetics, and also in fertilizers [37] is known to cause several problems in the human body related to the respiratory system, acute poisoning, and irritation [15,16]. Figure 9a and the corresponding gif file also show the higher concentration of Formaldehyde in the southern states compared to the northern states. The feature ranking in Figure 8 shows that a land area consisting of low vegetation tends to increase life expectancy, whereas there seems to be mixed results in the case of a land area consisting of high vegetation. Greenness has benefits related to the reduction of mental health and air pollution [18,19]. Figures 9b, Figures 9c, and the corresponding gif files show that the southern states have a large low and high vegetation area. This indicates that there are other crucial factors that affect life expectancy in addition to vegetation.

Figure 9d and the corresponding gif file of wet bulb temperature show that counties in southern states have higher wet bulb temperature compared to counties in northern states. Figures 8 show that higher wet bulb temperature tends to decrease the life expectancy. The temperature of the wet bulb, which is an indicator of heat stress if it exceeds 35 ^{o}C, can potentially result in hyperthermia in mammals, including humans, as in this condition dissipating heat becomes impossible [38]. In our study, we have calculated the one year average of the wet bulb temperature where the values are around 25 ^{o}C. However, there may be instances where the wet bulb temperature is much higher. Better working conditions, especially outdoors, where the wet bulb temperature is low, can potentially improve life expectancy. The SHAP value plot of specific humidity in Figure 9a shows a mixture of red and blue dots on either side of the central black line indicating that life expectancy can vary based on the location. Peroxyacetyl nitrate is also one of the main features affecting life expectancy, the acute toxicity is similar to nitrogen dioxide, but higher than SO_2 [39]. More research is needed on the effect of surface pressure on humans. Although we expected pollutant such as $PM_{2.5}$ to be higher in the feature ranking, we did observe pollutants such as sulfate aerosols, dust aerosols, nitrogen dioxide as shown in Figures 7 and 8.

 doi:10.20944/preprints202604.1997.v1

Our results in estimating life expectancy, in which we obtained a moderate accuracy as indicated by RMSE of 1.18 years and R^2 of 0.77 in an independent test set using a set of 5 features, clearly indicate missing variables in our study. Life expectancy is affected not only by environmental variables, but also by several other factors as well. Studies have shown that life expectancy tends to be longer with higher incomes in the United States [40,41] and in 22 European countries as well [42]. Disease related factors such as better cardiovascular health have been associated with longer life expectancy [43]. Education has also been considered a crucial factor in affecting life expectancy with a decrease in life expectancy among people with limited education [44,45]. Future studies will need to include socio-economic variables such as income of people, education rate, poverty rate, and other relevant variables such as access to health care for a comprehensive collection of variables to better estimate life expectancy and identify factors affecting life expectancy in the US to improve the livelihood of people.

Author Contributions: Conceptualization, D.L. and S.R.; methodology, D.L., S.R., S.S. and F.A.; software, S.R.; validation, D.L. and S.R.; formal analysis, S.R. and D.L.; investigation, S.R.; resources, D.L.; data curation, S.R.; writing—original draft preparation, S.R.; writing—review and editing, S.R., D.L., S.S., F.A.; visualization, S.R.; supervision, D.L.; project administration, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The code to produce the results, figures, and the final dataframe consisting the environmental variables and the target variable is available in GitHub (https://github.com/mi3nts/Life_Expectancy_Complete, accessed on April 20 2026). The entire project including the environmental data from ECMWF will be made available in Zenodo after acceptance.

Acknowledgments: The authors would like to thank the Physics department at The University of Texas at Dallas.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IHME	Institute for Health Metrics and Evaluation
ECMWF	European Centre for Medium-Range Forecasts
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
PM	Particulate Matter
GHDx	Global Health Data Exchange
CAMS	Copernicus Atmpshpere Monitoring Service
C3S	Climate Change Service
EAC4	ECMWF Atmoshperic Composition Reanalysis 4
ERA5	ECMWF Reanalysis version5
EPA	Environmental Protection Agency
FoT	Fraction of Time
SHAP	Shapley additive explanations
AOD	Aerosol Optical Depth

 doi:10.20944/preprints202604.1997.v1

Appendix A

Table A1. Year and the corresponding number of counties in contigous USA as available from the shapefiles by United States Census Bureau. Among all the counties for each year, the smallest number of latitude, longitude pair available from the shapefiles and the number of counties with less than 100 pair of latitude and longitude pair is also listed.

Year	Total number of counties	Smallest number of latitude, longitude pair	Number of counties with less than 100 pair of latitude and longitude
2008	3109	129	0
2009	3109	129	0
2010	3109	18	796
2011	3109	129	0
2012	3109	129	0
2013	3109	16	830
2014	3108	14	770
2015	3108	16	772
2016	3108	18	781
2017	3108	14	779
2018	3108	14	784
2019	3108	14	794

Appendix B

A dendrogram is used to identify multicollinear features. A dendrogram shows hierarchical clustering (the number of clusters is not predetermined) on the Spearman rank-order correlations. From the dendrogram, we manually pick a threshold to group features that are multicollinear and we use a single feature among the group of features. A dengrogram to identify multicollinear features is shown in Figure A1.

The vertical axis of the dendrogram represents the similarities between the variables. The shorter the vertical distance, the higher the collinearity between them, whereas the longer distance indicates less collinearity between the features. The threshold for vertical distance we have chosen is 0.10, which is indicated by the horizontal red line in the figure. This threshold was chosen considering the cluster (or group) of temperatures, which is also annotated in the figure. Choosing a higher threshold will include snow depth in the cluster, whereas choosing a lower threshold will include two temperature variables in the cluster. Each of the clusters below this threshold has a unique color, which indicates that these groups of features have high multicollinearity, whereas the rest of the features are not linearly correlated with any other features.

 doi:10.20944/preprints202604.1997.v1

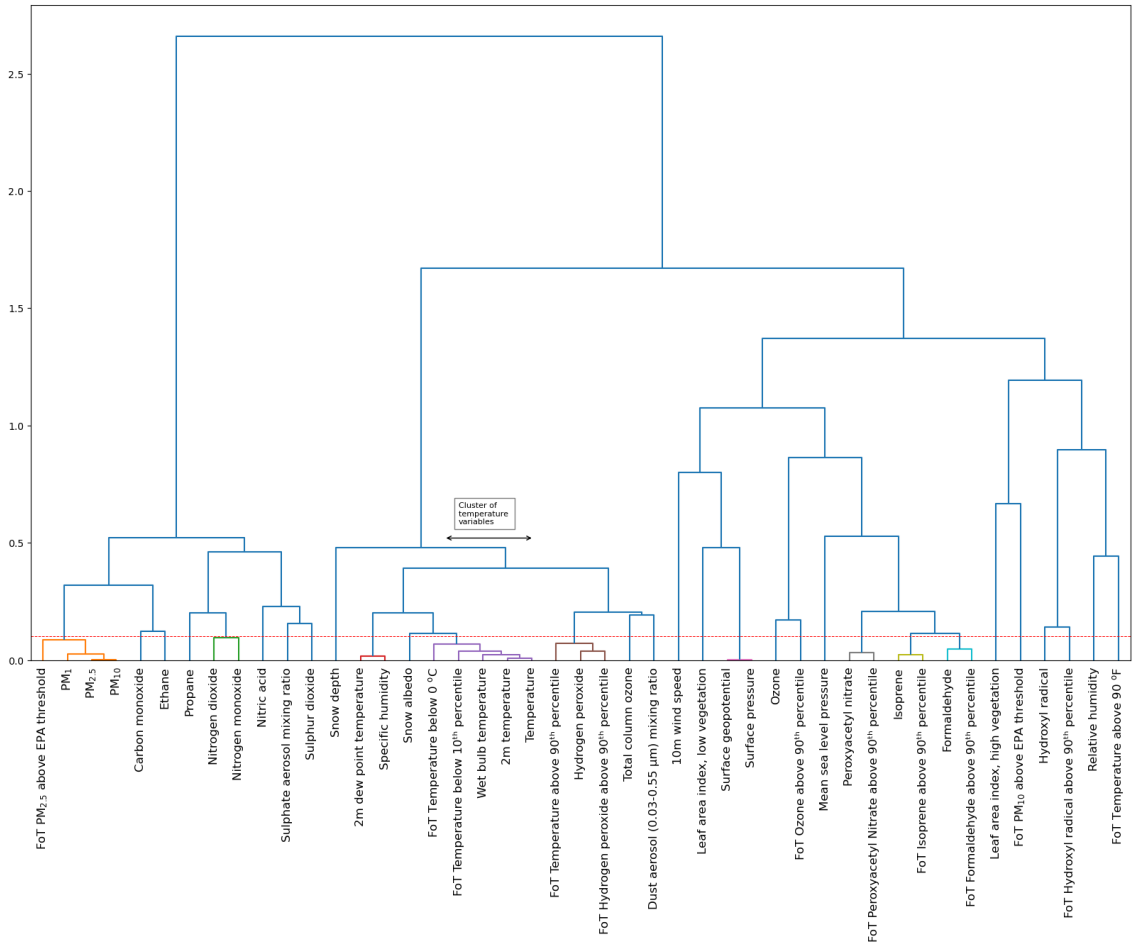


Figure A1. A dendrogram of 45 features to identify multicollinear features. Vertical axis represents the similarities between the variables where smaller distance indicate high similarities. Each unique color below the threshold of 0.10 indicated by the horizontal red line indicates group of multicolinear features. A cluster of temperature variables is also annotated.

Appendix C

A general outline of the calculation of permutation feature importance is given below [36]:

- for a model with a dataset, compute the reference score s, such as R^2 for regression. The target variable is estimated, and R^2 is calculated between the true and estimated values.
- For each feature j, randomly shuffle the data in that feature K times and compute the score $s_{k,j}$ (R^2 value) of the model on the new dataset with only one of the features randomly shuffled and the rest of the feature kept the same.
- The feature permutation i_j for a feature f_j is then calculated as follows:

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \tag{A1}$$

The result is interpreted as follows:

- High positive value of i_j : This is interpreted as a feature of high importance. This occurs when the shuffling of the data for a feature leads to a worse prediction (lower values of $s_{k,j}$). It is so because the way a data is related to a target should actually matter.
- High negative value of i_j: This is interpreted as a feature of low importance. In this case, the shuffling has led to a better prediction (high value of $s_{k,j}$). This means that the data do not represent any dependency with respect to the target variable.

 doi:10.20944/preprints202604.1997.v1

Appendix D

Table A2. Set of 4 hyperparameters that were optimized in Random Forest Models when 30, 15, 10, 5 and 3 features were used for estimating life expectancy in the U.S. GridSearchCV was used for optimizing the hyperparameters with 4-fold cross-validation. The hyperparameter that was selected is also provided.

Number of Features	n_estimators	max_features	max_depth	min_samples_split	Optimized hyperparameter
30	200, 300	10, 20	100, 150	10, 20	300, 20, 100, 10
15	200, 300	10, 20	100, 150	10, 20	300, 20, 100, 10
10	200, 300	10, 20	100, 150	10, 20	300, 20, 100, 10
5	200, 300	10, 20	100, 150	10, 20	300, 20, 100, 10
3	200, 300	10, 20	100, 150	10, 20	300, 10, 150, 10

Appendix E

The result of feature ranking in the test set using permutation feature importance and SHAP value when 30 features were used to estimate life expectancy is shown in Figure A2 and Figure A3 respectively. The complete list of feature ranking using permutation feature importance is in the link (https://github.com/mi3nts/Life_Expectancy_Complete/blob/main/Results/permutation_importance_30_test.png, accessed on December 28 2025).

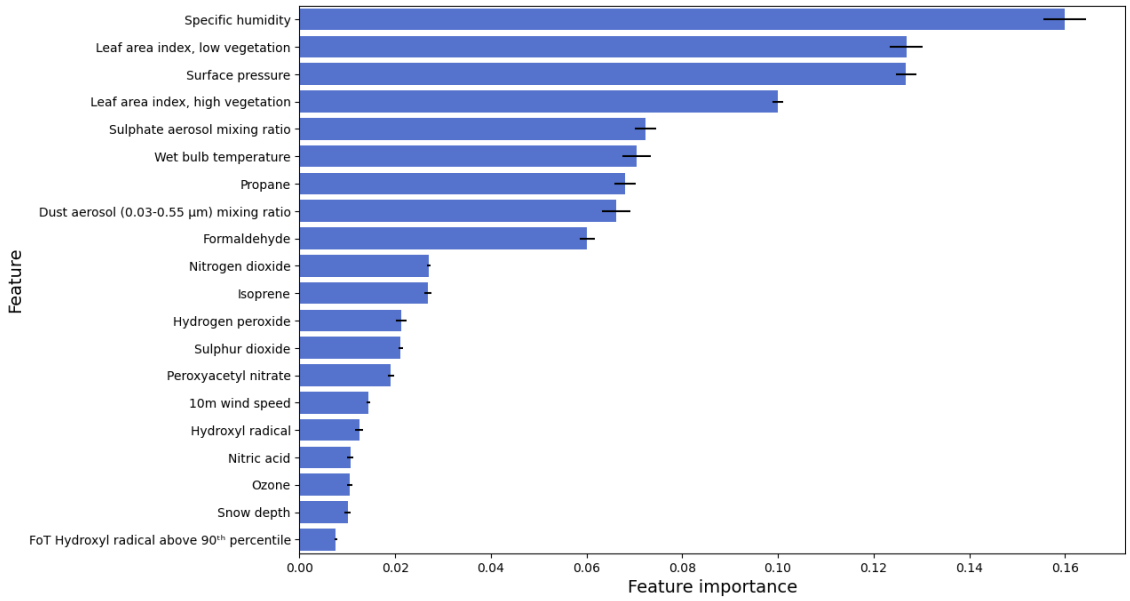


Figure A2. Feature ranking of top 20 variables using permutation feature importance in the test set in a descending order when 30 features were used for estimating life expectancy. Each of the feature was shuffled 10 times and mean value is plotted. The standard deviation is plotted as error bar.

Figure A3 shows the 20 main features in descending order of importance using the SHAP values calculated in the test set when 30 features were used to estimate life expectancy. The complete list of feature ranking using SHAP values is available in the link (https://github.com/mi3nts/Life_Expectancy_Complete/blob/main/Results/shap_30_test.png, accessed on December 28 2025).

 doi:10.20944/preprints202604.1997.v1

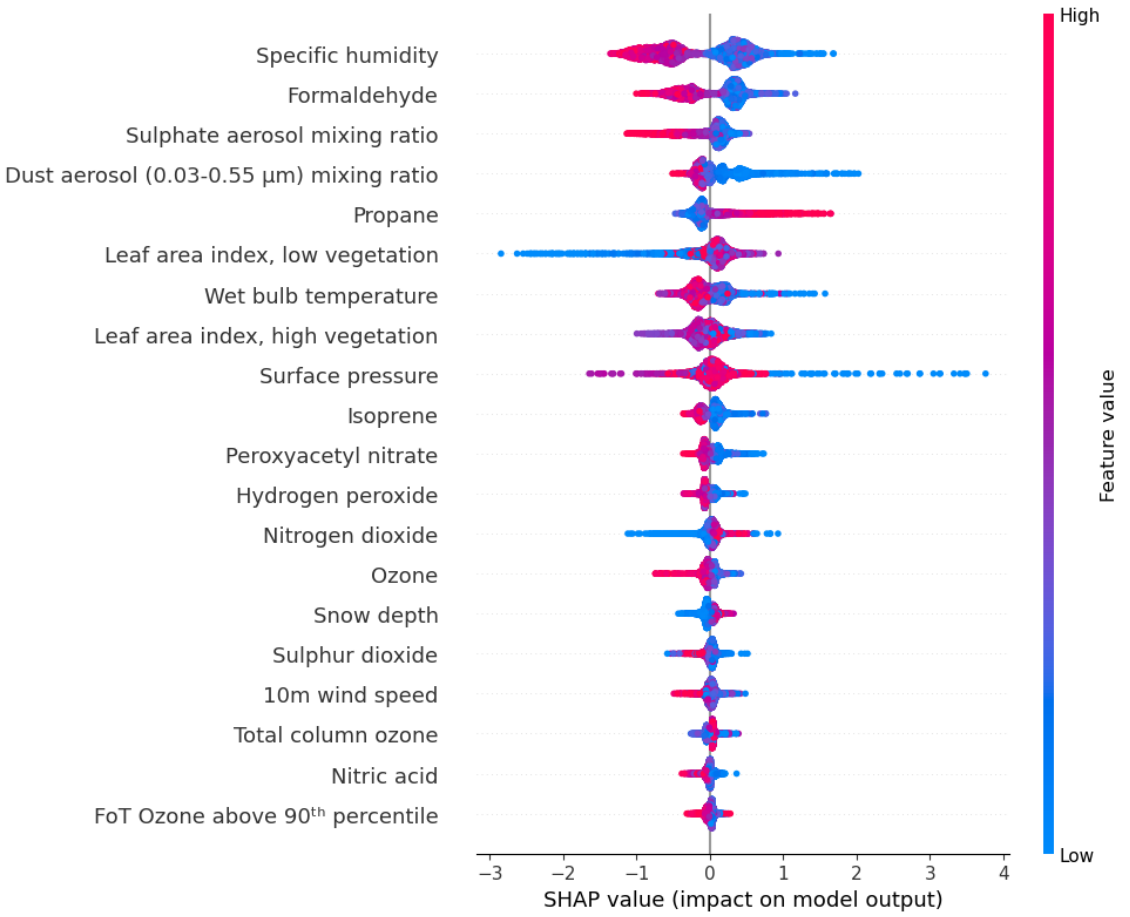


Figure A3. SHAP value beeswarm plot of top 20 features when 30 features were used for estimating life expectancy in test set. The X-axis is in the units of year which is the unit of life expectancy. Each of the feature has 10,464 circular dots which is the number of data points in the test set. Figure shows that high values of formaldehyde tend to decrease the life expectancy as higher values are on the left of the central black line. Similarly, higher values of leaf area index, low vegetation tend to increase the life expectancy as most of the high feature values are on the right side of the central black line.

References

1. Ho, J.Y. Causes of America's lagging life expectancy: an international comparative perspective. *The Journals of Gerontology: Series B* **2022**, *77*, S117–S126.
2. Woolf, S.H.; Schoomaker, H. Life expectancy and mortality rates in the United States, 1959-2017. *Jama* **2019**, *322*, 1996–2016.
3. Harper, S.; Riddell, C.A.; King, N.B. Declining life expectancy in the United States: missing the trees for the forest. *Annual review of public health* **2021**, *42*, 381–403.
4. Harper, S.; Lynch, J.; Burris, S.; Smith, G.D. Trends in the black-white life expectancy gap in the United States, 1983-2003. *Jama* **2007**, *297*, 1224–1232.
5. Arias, E.; Johnson, N.J.; Vera, B.T. Racial disparities in mortality in the adult Hispanic population. *SSM-population health* **2020**, *11*, 100583.
6. Arias, E.; Xu, J.; Jim, M.A. Period life tables for the non-Hispanic American Indian and Alaska Native population, 2007–2009. *American Journal of Public Health* **2014**, *104*, S312–S319.
7. Pope III, C.A.; Ezzati, M.; Dockery, D.W. Fine-particulate air pollution and life expectancy in the United States. *New England journal of medicine* **2009**, *360*, 376–386.
8. Correia, A.W.; Pope III, C.A.; Dockery, D.W.; Wang, Y.; Ezzati, M.; Dominici, F. Effect of air pollution control on life expectancy in the United States: an analysis of 545 US counties for the period from 2000 to 2007. *Epidemiology* **2013**, *24*, 23–31.
9. Laden, F.; Schwartz, J.; Speizer, F.E.; Dockery, D.W. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. *American journal of respiratory and critical care medicine* **2006**, *173*, 667–672.

 doi:10.20944/preprints202604.1997.v1

10. Schwartz, J.; Coull, B.; Laden, F.; Ryan, L. The effect of dose and timing of dose on the association between airborne particles and survival. *Environmental health perspectives* **2007**, *116*, 64.
11. Sewe, M.O.; Bunker, A.; Ingole, V.; Egondi, T.; Åström, D.O.; Hondula, D.M.; Rocklöv, J.; Schumann, B. Estimated effect of temperature on years of life lost: a retrospective time-series study of low-, middle-, and high-income regions. *Environmental Health Perspectives* **2018**, *126*, 017004.
12. Roy, A. A panel data study on the effect of climate change on life expectancy. *PLoS Climate* **2024**, *3*, e0000339.
13. Lowen, A.C.; Mubareka, S.; Steel, J.; Palese, P. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens* **2007**, *3*, e151.
14. Shaman, J.; Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences* **2009**, *106*, 3243–3248.
15. Kim, K.H.; Jahan, S.A.; Lee, J.T. Exposure to formaldehyde and its potential human health hazards. *Journal of Environmental Science and Health, Part C* **2011**, *29*, 277–299.
16. İnci, M.; Zararsız, İ.; Davarcı, M.; Görür, S. Toxic effects of formaldehyde on the urinary system. *Turkish journal of urology* **2013**, *39*, 48.
17. Bedimo-Rung, A.L.; Mowen, A.J.; Cohen, D.A. The significance of parks to physical activity and public health: a conceptual model. *American journal of preventive medicine* **2005**, *28*, 159–168.
18. Grinde, B.; Patil, G.G. Biophilia: does visual contact with nature impact on health and well-being? *International journal of environmental research and public health* **2009**, *6*, 2332–2343.
19. Coleman, C.J.; Yeager, R.A.; Pond, Z.A.; Riggs, D.W.; Bhatnagar, A.; Pope III, C.A. Mortality risk associated with greenness, air pollution, and physical activity in a representative US cohort. *Science of the total environment* **2022**, *824*, 153848.
20. for Health Metrics, I.; (IHME), E. United States Mortality Rates by Causes of Death and Life Expectancy by County, Race, and Ethnicity 2000-2019, 2023. https://doi.org/10.6069/3WQ2-TG23.
21. Dwyer-Lindgren, L.; Kendrick, P.; Kelly, Y.O.; Sylte, D.O.; Schmidt, C.; Blacker, B.F.; Daoud, F.; Abdi, A.A.; Baumann, M.; Mouhanna, F.; et al. Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *The Lancet* **2022**, *400*, 25–38.
22. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.M.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics* **2019**, *19*, 3515–3556. https://doi.org/10.5194/acp-19-3515-2019.
23. Service, C.C.C. ERA5 hourly data on pressure levels from 1940 to present, 2018. https://doi.org/10.24381/CDS.BD0915C6.
24. Karentz, D. Ozone Layer. In *Encyclopedia of Ecology*; Jørgensen, S.E.; Fath, B.D., Eds.; Academic Press: Oxford, 2008; pp. 2615–2621. https://doi.org/https://doi.org/10.1016/B978-008045405-4.00865-X.
25. Stull, R. Wet-bulb temperature from relative humidity and air temperature. *Journal of applied meteorology and climatology* **2011**, *50*, 2267–2269.
26. Crimmins, A.; Balbus, J.; Gamble, J.L.; Beard, C.B.; Bell, J.E.; Dodgen, D.; Eisen, R.J.; Fann, N.; Hawkins, M.D.; Herring, S.C.; et al. The impacts of climate change on human health in the United States: a scientific assessment. *The impacts of climate change on human health in the United States: A scientific assessment* **2016**.
27. Shindell, D.; Zhang, Y.; Scott, M.; Ru, M.; Stark, K.; Ebi, K.L. The effects of heat exposure on human mortality throughout the United States. *GeoHealth* **2020**, *4*, e2019GH000234.
28. Wu, J.; Hu, Z.; Han, Z.; Gu, Y.; Yang, L.; Sun, B. Human physiological responses of exposure to extremely cold environments. *Journal of Thermal Biology* **2021**, *98*, 102933. https://doi.org/https://doi.org/10.1016/j.jtherbio.2021.102933.
29. Schneider, A.; Breitner, S. Temperature effects on health-current findings and future implications. EBioMedicine, 6, 29–30, 2016.
30. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
32. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.
33. Shwartz-Ziv, R.; Armon, A. Tabular Data: Deep Learning is Not All You Need, 2021, [arXiv:cs.LG/2106.03253].

 doi:10.20944/preprints202604.1997.v1

34. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.
35. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2020**, *2*, 2522–5839.
36. Scikit-learn. Permutation feature importance. https://scikit-learn.org/stable/modules/permutation_importance.html, Accessed on Jan 9, 2025.
37. EPA. Facts About Formaldehyde. https://www.epa.gov/formaldehyde/facts-about-formaldehyde, 2025. Accessed: 2026-02-02.
38. Sherwood, S.C.; Huber, M. An adaptability limit to climate change due to heat stress. *Proceedings of the national academy of sciences* **2010**, *107*, 9552–9555.
39. Vyskocil, A.; Viau, C.; Lamy, S. Peroxyacetyl nitrate: review of toxicity. *Human & Experimental Toxicology* **1998**, *17*, 212–220.
40. Braveman, P.A.; Cubbin, C.; Egerter, S.; Williams, D.R.; Pamuk, E. Socioeconomic disparities in health in the United States: what the patterns tell us. *American journal of public health* **2010**, *100*, S186–S196.
41. Cristia, J.P.; DeLeire, A.H.; Iams, H.; Kile, J.; Manchester, J.; Meyerson, N.; Sabelhaus, J.; Waldron, H.; Walker, L.; et al. *The empirical relationship between lifetime earnings and mortality*; Congressional Budget Office Washington, DC, 2007.
42. Mackenbach, J.P.; Stirbu, I.; Roskam, A.J.R.; Schaap, M.M.; Menvielle, G.; Leinsalu, M.; Kunst, A.E. Socioeconomic inequalities in health in 22 European countries. *New England journal of medicine* **2008**, *358*, 2468–2481.
43. Ma, H.; Wang, X.; Xue, Q.; Li, X.; Liang, Z.; Heianza, Y.; Franco, O.H.; Qi, L. Cardiovascular health and life expectancy among adults in the United States. *Circulation* **2023**, *147*, 1137–1146.
44. Meara, E.R.; Richards, S.; Cutler, D.M. The gap gets bigger: changes in mortality and life expectancy, by education, 1981–2000. *Health affairs* **2008**, *27*, 350–360.
45. Jemal, A.; Ward, E.; Anderson, R.N.; Murray, T.; Thun, M.J. Widening of socioeconomic inequalities in US death rates, 1993–2001. *PloS one* **2008**, *3*, e2181.