



|  |  |
| --- | --- |
| TO | Neal Fann, Ali Kamal, and Jennifer Sellers, U.S. Environmental Protection Agency (USEPA) |
| FROM | William Raich, Hayley Kunkle, and Henry Roman, Industrial Economics, Incorporated (IEc) |
| SUBJECT | Supporting Near-Source Health Benefits Analyses Using Fine-Scale Incidence Rates |

MEMORANDUM | March 21, 2024

A primary goal of air quality management is to reduce the rate of adverse human health outcomes attributable to air pollution, including premature deaths and hospitalizations for respiratory and cardiovascular illnesses. USEPA uses the BenMAP-CE tool to quantify changes in these health outcomes expected to result when ambient pollutant concentrations change. For BenMAP analyses conducted nationally, USEPA typically relies on data sets at either the 12-kilometer grid resolution (population and air quality) or at the county level (baseline incidence). While suitable for national analyses, these datasets may be poorly suited for quantifying health impacts associated with “near source” emissions, such as major roadways or industrial point sources. In these cases, pollutant concentration gradients are spatially attenuated at scales finer than those typically modeled in national analyses, and therefore health benefits analyses require more highly resolved characterizations of baseline health status in neighborhoods near emissions sources.

This memorandum summarizes recent work conducted by IEc at the request of USEPA to identify and assess methods for better characterizing variation in baseline mortality incidence at fine geographic resolutions. The bulk of this document describes our efforts to derive incidence rates using census tract life tables published by the U.S. Small-Area Life Expectancy Estimates Project (USALEEP). The decision to focus on the USALEEP study was based on the merits of the study, described below, and how readily incidence rates could be derived from its results. Prior to our work with the USALEEP data, we conducted an initial literature review in 2020 focused on methodologies for estimating fine-scale incidence rates, with the goal of identifying existing methodologies or studies that could be readily applied or adapted to USEPA’s needs. Below, we first describe select studies identified in this search. We then present our methods for transforming USALEEP life tables into BenMAP-CE-ready incidence data. Finally, we describe validation efforts conducted by IEc to evaluate the reliability of these mortality rates and limitations to consider with this research.

# Methods for Deriving Mortality Incidence Rates

As noted above, USEPA currently relies on county-level mortality incidence data. These incidence rates were derived using county-level death counts reported in the Centers for Disease Control (CDC) WONDER database.[[1]](#footnote-2) These data, aggregated for the years 2012-2014, are stratified by age and cause of death and subsequently projected into future decades. The procurement and processing of these data highlight two challenges with assessing mortality incidence at fine geographic scales. First, publicly available databases do not typically report deaths at geographic resolutions finer than county. The underlying death certificates, reported to the CDC and maintained in the National Vital Statistics System, are confidential and require special access to use. Second, data masking procedures aimed at ensuring individual privacy often limit the usefulness of mortality data, even at the county level. CDC WONDER does not report death counts below 10 for any demographic cell. Estimation of rates at finer geographic resolutions exacerbates this concern and necessitates significant imputation of mortality incidence rates to replace suppressed values.[[2]](#footnote-3)

Three studies highlight how access to death records can be used to derive mortality estimates at a fine geographic resolution. Dwyer-Lindgren et al. (2017) use death records in King County, WA to estimate life expectancy, cause-specific mortality rates, and years of life lost (YLL) for 152 causes of death and 397 census tracts between 1990 and 2014. The authors employ Bayesian mixed-effects regression models to estimate mortality for each census tract. Similarly, Bennett et al. (2015) use Bayesian statistical models to estimate and project age-specific mortality and life expectancy for 375 of England and Wales’ districts. This model, fit using over 17 million death records between 1981 and 2012, is used to estimate mortality through 2030. Finally, USALEEP uses U.S. death records to estimate abridged life tables for U.S. census tracts. This study is described in greater detail below.

Notably, the three studies described above were made possible through access to confidential death records. Additionally, the studies conducted advanced statistical modeling that is beyond of the scope of this work assignment. Following consultation with USEPA, we concluded that IEc would focus on adapting the results of the USALEEP study, due to its national coverage in the United States and its publicly available results at the census tract level in the form of abridged life tables.

# USALEEP Study and Derivation of Mortality Incidence Rates

The following provides an overview of the USALEEP study, followed by the methods used to derive mortality rates from the USALEEP supplementary material and results and limitations of the approach.

## Overview of USALEEP Study

USALEEP[[3]](#footnote-4) develops and applies a methodology for estimating a set of abridged period life tables for U.S. census tracts for the 2010-2015 period. This project provides the first set of abridged life tables at the census tract level for all states and the District of Columbia. The National Center for Health Statistics (NCHS), in collaboration with the U.S. Department of Housing and Urban Development (HUD), geocoded death records considered to be of high quality to census tracts reflecting place of residence. These geocoded death records were combined with population and socioeconomic information from the 2010 decennial census and the American Community Survey (ACS) for 2011-2015 to develop life tables in three phases:

1. Determine age-specific death rates from census tracts with sufficient death counts for all age groups, population sizes of 5,000 or more, and which display a regular or typical age pattern, or schedule, of mortality.
2. Fit a zero-truncated Poisson and negative binomial model to the Phase 1 census tracts to model death counts based on tracts’ socioeconomic and demographic characteristics. Fill in missing age-specific death counts for other tracts using model predictions.
3. Calculate abridged life tables for all census tracts with population sizes over 5,000.

In Phase 1, 4,639 census tracts were identified as suitable for the Phase 2 parametric model out of a total of 74,001 census tracts identified by both the 2010 census and ACS population datasets over the period. The 69,362 census tracts were identified as unsuitable for the following reasons:

1. 4,481 tracts have a 6-year total population size of zero (in both the 2010 Census and ACS datasets);
2. 2,560 tracts are in Maine and Wisconsin, excluded because of insufficient death records over the period;[[4]](#footnote-5)
3. 60,685 tracts have at least one age group with zero deaths;
4. 222 tracts with a population less than 5,000; and
5. 1,102 tracts had irregular age-specific mortality patterns (i.e., inconsistent with a standard log-scale mortality schedule).

The fifth criterion is based on a mortality pattern that is “universally observed in human populations,” which has three characteristics: (1) mortality at birth is higher than subsequent death rates up until middle adulthood; (2) mortality is lowest in middle childhood, around ages 5-14; and (3) after the lowest point, mortality accelerates with age although it may decelerate in very old ages. An example of this pattern for 2013 is shown in Exhibit 1 (also Figure 1 in the USALEEP Report). Any census tracts that did not meet these three criteria were excluded from fitting the model.

Exhibit 1. Age pattern (schedule) of mortality: United States, 2013

A green line graph with numbers

Description automatically generated

The remaining 4,639 tracts were used in the statistical model in Phase 2. The mortality rates from these “model” tracts were trained using socio-economic and demographic characteristics of the tracts. These include “quartiles of median household income, population density, and the proportions of the population that are non-Hispanic black, Hispanic, and had a 4-year college degree or higher in the census tract; and a binary variable indicating whether the census tract belonged to a Purchased/Referred Care Service Delivery Area (PRCSDA).”[[5]](#footnote-6) In Phase 3, abridged life tables were developed for 11 age groups: 0, 1-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and 85+ using a series of equations and assumptions.

The report authors note multiple limitations to their derivation of mortality. First, the authors rely on population data from ACS, which is based sampling from the U.S. population, as opposed to a complete survey from the 2010 decennial census. Second, the authors note that the regression model is trained using census tracts with higher death counts. For example, 51 percent of tracts are in the southeastern United States where death rates are elevated.

USALEEP outputs include abridged life tables for all census tracts in the dataset. However, the authors do not provide census tract mortality rates. In the following section, we outline our methods for estimating incidence rates from the published life tables.

## Deriving Mortality Rates from USALEEP

We estimate tract- and age-specific mortality rates (M) from the published life tables by dividing the number of people dying within each age group (ndx) by the person-years lived within each age group (nLx).[[6]](#footnote-7) An example of an abridged life table and the mortality calculation is shown in Exhibit 2, using values from a census tract in Alabama.

Exhibit 2. Example of an abridged life table and IEc mortality incidence rate (M) calculations

|  |  |  |  |
| --- | --- | --- | --- |
| Age Group | nd(x) | nL(x) | M |
| Under 1 | 1,974 | 98,224 | 0.0201 |
| 1-4 | 871 | 390,364 | 0.0022 |
| 5-14 | 600 | 968,553 | 0.0006 |
| 15-24 | 832 | 961,393 | 0.0009 |
| 25-34 | 2,290 | 945,785 | 0.0024 |
| 35-44 | 1,898 | 924,846 | 0.0021 |
| 45-54 | 1,637 | 907,169 | 0.0018 |
| 55-64 | 7,618 | 860,889 | 0.0088 |
| 65-74 | 18,660 | 729,499 | 0.0256 |
| 75-84 | 32,148 | 475,459 | 0.0676 |
| 85 and older | 31,472 | 43,836 | 0.7179 |
| Note: Columns in light blue are directly from the USALEEP Supplementary Material. Column in dark blue is based on IEc calculations. | | | |

While USALEEP provides abridged life tables for most of the census tracts in the United States, some are excluded primarily due to insufficient underlying data, as described above. Missing tracts are mapped in red in Exhibit 3. Many of these tracts represent low population areas, though some missing tracts also occur in high population areas including Maricopa County and Los Angeles County. For missing tracts, we impute rates by using county-level mortality rates from the BenMAP-CE database to fill in missing values.

A map of the united states

Description automatically generatedExhibit 3. Census tracts missing from USALEEP abridged life tables

# Validation of USALEEP-derived rates

To assess the quality of the tract-level mortality incidence estimates derived from USALEEP, EPA and IEc compared these rates against a collection of mortality datasets at comparable spatial scales, listed in Exhibit. In addition, we used available data in BenMAP-CE for select U.S. counties that are each comprised of a single Census tract.

**Exhibit 4. Historical Administrative Data Sources**

|  |  |  |  |
| --- | --- | --- | --- |
| **Location** | **Source** | **Timescale** | **Spatial Scale** |
| Alameda County, CA | Public Health Department | 2010-2015 | Census Tract |
| New York City, NY | NYC Department of Health and Mental Hygiene | 2010-2014 | Census Tract |
| Washington D.C. | Department of Health Planning | 2000-2015 | Neighborhooda |
| Small U.S. Counties (counties with only one tract) | CDC WONDER | 2013-2017 | Census Tract |
| Colorado | Department of Public Health | 2010-2020 | Census Tract |
| a D.C. data was provided at the neighborhood level. We aggregate the USALEEP tract-level modeled rates to the neighborhood level (population weighted average) for this comparison. | | | |

We compared the USALEEP tract-level incidence rates to mortality incidence rates within this subset of tracts using numerous tests, including calculating correlation coefficients, summary statistics, relative differences, and Root Mean Squared Errors (RMSEs) between USALEEP rates and historical rates, and plotting scatterplots and density plots. Exhibit 5 shows the calculated correlation coefficients for each region, including age-specific results for older adults (age 65 and older) for Washington D.C. Exhibit 6 shows scatterplot and density plot comparisons for Alameda County as an example.

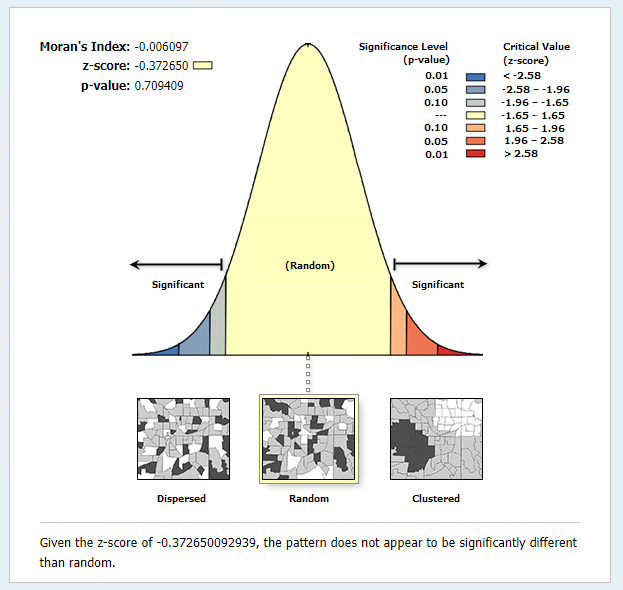
**Exhibit 5. Correlation coefficients comparing USALEEP rates and observed rates**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Region** | **Age Range** | **Number of Tracts** | **Pearson Correlation Coefficient** | **Spearman Correlation Coefficient** | **Lin Correlation Coefficient** |
| Alameda County, CA | 0-99 | 356 | 0.90 | 0.88 | 0.88 |
| New York City, NY | 0-99 | 2,030 | 0.86 | 0.74 | 0.80 |
| Washington D.C. | 0-99 | 48a | 0.98 | 0.93 | 0.97 |
| Washington D.C. | 65-99 | 48a | 0.71 | 0.76 | 0.69 |
| Small U.S. Counties (counties with only one tract) | 0-99 | 220 | 0.51 | 0.50 | 0.47 |
| Coloradob | 0-99 |  | 0.94 |  |  |
| a 48 neighborhoods in D.C.  b Colorado data is maintained and analyzed by Priyanka DeSouza at UC-Denver. A smaller subset of comparisons was conducted. | | | | | |

**Exhibit 6. Scatterplot and density plot comparing USALEEP rates and observed rates for Alameda County, CA**

|  |  |
| --- | --- |
| A graph with a red line  Description automatically generated | A close-up of a sign  Description automatically generatedChart, histogram  Description automatically generated |

We also conducted spatial validation tests using the Global Moran’s I and Local Moran’s I, two spatial autocorrelation tools that evaluate if relative differences between historical data and USALEEP data display spatial patterns. These tests were used to understand if differences between the datasets are random or if they displayed any sort of systemic pattern. Results of the Global Moran’s I test, shown for Alameda County as an example, are displayed in Exhibit 7. The Global Moran’s I shows that the relative differences between historical rates and USALEEP rates in Alameda county are randomly dispersed, suggesting that there is no spatial pattern to the differences between rates.

**Exhibit 7. Global Moran’s I test for Alameda County, CA**

|  |  |
| --- | --- |
|  |  |

Exhibit 8 shows the results of the Global Moran’s I for all locations analyzed. The relative differences between historical and modeled rates in the NYC dataset are spatially clustered, suggesting a systematic over or underestimate of mortality by the USALEEP rates compared to the administrative data in some areas.

**Exhibit 8. Global Moran’s I results for three regions**

|  |  |
| --- | --- |
| **Region** | **Global Moran’s I Pattern** |
| Alameda County, CA | Random |
| New York City, NY | Clustered |
| Washington D.C. | Random |

Exhibit 9 shows Local Moran’s I results for the three regions analyzed. In all three regions, the Local Moran’s I shows most Census tracts in white, signifying insignificant clustering of relative differences between historical and modeled mortality incidence rates. The light pink and light blue areas indicate tracts with clustering, while the red and dark blue areas indicate tracts with spatial outliers. We see the greatest incidence of clustering in NYC, consistent with the Global Moran’s I result, and suggesting some neighborhood specific differences not captured in the USALEEP results.

**Exhibit 9. Local Moran’s I results for three regions**

|  |  |
| --- | --- |
| Alameda County, CA | New York City |
| A map of the alameda county  Description automatically generated |  |
| Washington D.C. | |
|  | |

Finally, we conducted mortality analyses in BenMAP-CE using the tract-level USALEEP incidence rates and county-level incidence rates to compare results using these different incidence sources. The analysis using the USALEEP incidence rates yielded slightly fewer total deaths (5,650 compared to 5,692 deaths across the contiguous US), with similar spatial patterns at the county level. Results of this comparison are shown in Exhibit 10.

**Exhibit 10. Mortality impacts using USALEEP incidence rates (top) and county incidence rates from CDC WONDER (bottom)**

A map of the united states

Description automatically generated

# Conclusion and Discussion

Based on our validation tests, we find that the tract-level mortality rates derived from USALEEP appear suitable for analyses conducted at sub-county geographic resolutions. For large-scale air pollution health benefit analyses, the rates result in aggregate counts comparable to existing county-scale estimates while facilitating estimation of health effects that may vary across census tracts due to differences in underlying mortality and morbidity rates in the baseline. Our validation tests generally showed a strong correlation (>0.8) between modeled (i.e., derived from USALEEP) and observed incidence rates from other sources at the tract level. In some instances, we did observe spatial clustering of tract-level differences between USALEEP-derived rates and observed values, suggesting that local factors may not be fully reflected in the USALEEP predictions.

While the USALEEP-derived mortality incidence rates allow for researchers and regulatory analysts to conduct health benefits analyses at finer spatial scales, we emphasize some limitations of these data and recommend areas for future research.

First, the USALEEP study is accompanied by several uncertainties and limitations. While some tracts (4,639) reflect observed mortality rates, most tracts (63,586) are predicted using a statistical model trained on the observed set, and as noted by the study’s authors, the training set includes tracts with higher mortality rates. In addition, the resulting predictions are not available for all tracts. For eight percent of census tracts, we impute mortality incidence estimates using county-level rates; other imputation procedures could be explored.

Second, some of the resulting incidence rates may require modification to be consistent with the values in EPA’s BenMAP-CE program. EPA has conventionally omitted neonatal mortality to better align with available epidemiological evidence; however, these deaths are currently counted in the USALEEP abridged life tables (and the resulting incidence rates). Given that rates reflect all-cause incidence rates during the period from 2010 to 2015, additional analysis may be warranted to project these rates into future decades and to reflect cause-specific rates (e.g., respiratory mortality).

Given these limitations, we recommend that analysts leverage existing mortality incidence rates (e.g., county rates in BenMAP-CE) in tandem with any application of the USALEEP-derived mortality incidence rates to explore the sensitivity of results to data source and to highlight any potentially problematic estimates. We also note that further work could be pursued to characterize uncertaintyin the new tract-level mortality incidence rates. USALEEP abridged life tables provide both standard error of the probability of dying within each age group and standard error of life expectancy, which can be used to estimate standard error around mortality rates.

# References

Arias E, Escobedo LA, Kennedy J, Fu C, and Cisewski J. (2018). U.S. Small-area life expectancy estimates project: Methodology and results summary. National Center for Health Statistics. Vital Health Stat 2(181).

Bennett JE, Li G, Foreman K, et al. (2015) The future of life expectancy and life expectancy inequalities in England and Wales: Bayesian spatiotemporal forecasting. *Lancet 2015*; 386: 163–70.

Dwyer-Lindgren L, Stubbs RW, Bertozzi-Villa A, et al. (2017). Variation in life expectancy and mortality by cause among neighbourhoods in King County, WA, USA, 1990–2014: a census tract-level analysis for the Global Burden of Disease Study 2015. *Lancet Public Health 2017*; 2: 400–10

Kostaki A and Panousis V. (2001). Expanding an Abridged Life Table. *Demographic Research* 5(1): 1–22.

1. See the BenMAP-CE User Manual, Appendix D for more information on BenMAP-CE’s mortality incidence datasets: <https://www.epa.gov/benmap/benmap-ce-manual-and-appendices> [↑](#footnote-ref-2)
2. For example, CDC WONDER may report 250 deaths for a given age group, cause of death, and county. If this county is composed of 100 Census tracts, most tracts are likely to have fewer than 20 deaths, necessitating suppression of these results and limiting analysts’ ability to derive incidence rates based on observed data. [↑](#footnote-ref-3)
3. Arias E, Escobedo LA, Kennedy J, Fu C, Cisewski J. U.S. small-area life expectancy estimates project: Methodology and results summary. National Center for Health Statistics. Vital Health Stat 2(181). 2018. [↑](#footnote-ref-4)
4. Tracts in Maine and Wisconsin have been included in the model in a more recent iteration, though the USALEEP documentation has not been updated to account for this. [↑](#footnote-ref-5)
5. The PRCDA indicates the geographic area within which PRC services will be made available by the Indian Health Service (IHS) to members of an identified Indian community who reside in the area. [↑](#footnote-ref-6)
6. USALEEP abridged period life tables are available at: <https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html#life-expectancy>. [↑](#footnote-ref-7)