Imputation of Life Expectancy at Birth for Utah HPI

# **BACKGROUND**

To generate HPI scores for Utah census tracts, life expectancy at birth (LEB) is used to 1) screen individual indictors for a positive association with life expectancy, and 2) calculate domain (policy action area) weights using a weighted quantile sums algorithm.[1](#_ENREF_1), [2](#_ENREF_2) In 2018, life expectancy calculations were made available through a collaborative project of the National Association for Public Health Statistics and Information Systems, CDC's National Center for Health Statistics (NCHS), and the Robert Wood Johnson Foundation (RWJF). The project is known as the United States Small-Area Life Expectancy Estimates Project (USALEEP) and uses pooled data from 2010 to 2015 for United States census tracts. Exclusion criteria for USALEEP differs from our Utah HPI eligibility criteria. There are 42 HPI-eligible census tracts in Utah that do not have LEB data reported by USALEEP. Imputation of LEB not only reduces information loss but potentially reduces biases from missing information. Imputation is particularly suited to this situation because the amount of missing data is small (~5%).

There are many methods for imputing missing data.3 Some take into account geographical proximity and others use non-missing covariates to build parametric or non-parametric models. However, because HPI is geographically based, a key question guiding an appropriate imputation method is the degree to which the distribution of census tracts with missing LEB data are geospatially clustered or random. Our VCU partners suggested one of two approaches depending on the distribution of census tracts with missing LEB ("missing tracts"). If missing tracts were randomly distributed, the kth nearest neighbor (KNN) algorithm is the preferred imputation method. Despite its name, KNN matches a missing data point with its closest k neighbors in a multi-dimensional space, which, in our case, is constructed from the non-missing values of 25 HPI indicators rather than x-y coordinates of geographically proximate census tracts. Alternatively, if the distribution of missing tracts were geographically clustered, then non-missing LEB values of the most geographically proximate census tracts would be preferable for imputation.

This report describes the assessment of the geospatial distribution of missing LEB data to inform a preferred imputation method, which is then applied to impute missing LEB data for the 42 missing tracts.

# **METHODS**

## Data Sources

* Population (B01001\_001E) and group quarters composition (B26001\_001E) for Utah census tracts from the American Community Survey, 2015-2019
  + HPI-eligible were those with a population of 1500 or greater and group quarters percentage less than 50%
* Life expectancy of Utah census tracts from USALEEP – downloaded from the website of the National Center for Health Statistics (<https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html#data>).

## Geographic Distribution of Missing Census Tracts

The Monte-Carlo simulation of join-count statistics was used to assess the geographic distribution of census tracts with missing data.5 The join-count statistic is a method of measuring the degree of clustering or dispersion of binary nominal data (i.e. yes/no) among a set of spatially adjacent polygons. The method calculates the observed number of adjacent neighbor pairs with similar attributes compared to the expected number of adjacent neighbor pairs with similar attributes. In this case, it is measuring the number of neighboring census tract pairs both missing census data compared to how many missing-missing neighbor pairs would be expected if missingness was randomly distributed. If missing census tracts were clustered in our sample, significantly more missing-missing neighbor pairs would be present than if missingness was randomly distributed. Monte-Carlo simulation of join-count statistics with 1,000 simulations was used to provide most reliable results. Monte-Carlo simulation test for join-count statistics is based on random permutations of the dataset and does not rely on the assumption that the test statistic is normally distributed.

Adjacent polygons for the join-count statistic were defined using the Rook criteria. This means that two polygons were considered adjacent neighbors if they share a common boundary. This differs from the Queen criteria which only requires that adjacent neighbors share a common vertex. A one-sided alternative hypothesis was selected to test the alternative hypothesis that the number of like similar neighbor pairs is more than expected by random chance. Because many metrics of spatial clustering or dispersion may be sensitive to geographic scale, the Monte-Carlo simulation of join-count statistics was conducted for all HPI-eligible census tracts in Utah, and also for a subset of HPI-eligible census tracts only in Salt Lake, Utah, and Davis Counties.

## Imputation Methodology

Because join-count statistics indicated that the spatial distribution of missing tracts was not random, missing LEB data was imputed from geographically proximate census tracts with USALEEP LEB data. Geographic adjacent neighbors were defined using the Rook criteria. All Utah census tracts missing LEB data had at least two adjacent neighbors. The imputed LEB for missing census tracts was computed using the arithmetic mean of the LEB values of the identified adjacent census tracts.

All imputation analysis was conducted with R version 4.1.2.

# **RESULTS**

Forty-two of the 525 HPI-eligible census tracts in Utah, or 7.3%, are missing life expectancy at birth data. The majority of census tracts with missing data are in Salt Lake, Utah, and Davis counties, the three most populous counties in Utah.

**Figure 1. A) Distribution of census tracts with missing LEB data in Utah. B) Distribution of census tracts with missing LEB data in the Salt Lake City metro area.**

A B

Map

Description automatically generated Map

Description automatically generated

The Monte Carlo simulation of join-count statistic indicated that the distribution of census tracts with missing data is clustered (Table 1), both when testing join-count statistics using all HPI-eligible census tracts in Utah, and when using a subset of HPI-eligible census tracts in Salt Lake, Utah, and Davis counties.

**Table 1. Monte-Carlo Simulation of Join-Count Statistics for Census Tracts Missing LEB Data in Selected Utah Geographies**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Geography | Join-Count Statistic (Observed missing-missing pairs) | Mean of Simulation (expected missing-missing pairs) | p Value | Distribution |
| Salt Lake, Utah and Davis Counties | 12 | 6.68 | 0.02 | Clustered |
| Utah State | 16 | 7.85 | 0.002 | Clustered |

Because there is evidence of census tracts with missing data being clustered, LEB was computed from geographically proximate census tracts with USALEEP data for LEB. All census tracts had at least two adjacent neighbors.

# **DISCUSSION**

Using join-count statistics, we found evidence that census tracts in Utah missing LEB data are clustered. Therefore, we imputed LEB for the 42 census tracts missing data using geographic neighbors, rather than covariate near neighbors. This methodology to analyze the geographic distribution of census tracts with missing data differed from our methodology when imputing LEB data for the California HPI. We previously used the Average Nearest Neighbor Index (ANNI) to assess clustering or dispersion of missing data. ANNI is a method for spatial point data and is calculated using distance between census tract centroids. It is therefore very sensitive to differences in census tract shape and size. Because the join-count statistic is a measure of clustering or dispersion for areal data, we determined it may be a more appropriate choice.

# **REFERENCES**

1. Gennings C, Sabo R, Carney E. Identifying subsets of complex mixtures most associated with complex diseases: Polychlorinated biphenyls and endometriosis as a case study. *Epidemiology* 2010;21:S77–S84.

2. Maizlish N, Delaney T, Dowling H, Chapman D, Sabo R, Woolf S, et al. California Healthy Places Index: Frames Matter. *Public Health Rep.* 2019;134:354-362. https://journals.sagepub.com/doi/10.1177/0033354919849882

4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York.: John Wiley & Sons; 1987.

5. Oyana, Tonny J., and Florence Margai. *Spatial Analysis : Statistics, Visualization, and Computational Methods*, Taylor & Francis Group, 2015.*ProQuest Ebook Central*, https://ebookcentral-proquest-com.proxy.lib.ohio-state.edu/lib/ohiostate-ebooks/detail.action?docID=2127158.