University of Louisville Journal of Respiratory Infections

ORIGINAL RESEARCH

The City of Louisville Encapsulates the United States Demographics

Stephen Furmanek[1], MS MPH; Connor Glick[1], BS; Thomas Chandler[1], MPH; Mahder Tella[1], MPH; William Mattingly[1], PhD; Julio A Ramirez[1], MD; Timothy L Wiemken[2], PhD*

[1]Division of Infectious Diseases, School of Medicine, University of Louisville, Louisville, KY, USA; [2]Saint Louis University Center for Health Outcomes Research, Saint Louis, MO, USA

*stephen.furmanek@louisville.edu

Abstract

Background: One weakness that applies to all population-based studies performed in the United States (US) is that investigators perform population-based extrapolations without providing objective statistical evidence to show how well a particular city is a suitable surrogate for the US. The objective of this study was to propose and utilize a novel computational metric to compare individual US cities with the US average.

Methods: This was a secondary data analysis of publicly available databases containing US sociodemographic, economic, and health-related data. In total, 58 demographic, housing, economic, health behavior, and health status variables for each US city with a residential population of at least 500,000 were obtained. All variables were recorded as proportions. Euclidean, Manhattan, and the average absolute difference metrics were used to compare the 58 variables to the average in the US.

Results: Oklahoma City, Oklahoma had the lowest distance from the United States, with Euclidean and Manhattan distances in proportion of 0.261 and 1.519, respectively. Louisville, Kentucky had the second lowest distance for both Euclidean distance and Manhattan distance, with distances of 0.286 and 1.545, respectively. The average absolute differences in proportion for Oklahoma City and Louisville to the US average were 0.026 and 0.027, respectively.

Conclusion: To our knowledge, this represents the first study evaluating a method for computing statistical comparisons of United States city sociodemographic, economic, and health-related data with the United States average. Our study shows that among cities with at least 500,000 residents, Oklahoma City is the closest to the United States, followed closely by Louisville. On average, these cities deviate from the US average on any variable studied by less than 3 percent.

Recommended Citation:
Arnold, Forest W.; Burns, Mark V.; Mahmood, Kamran; Aden, Darmaan; Furmanek, Stephen; Tella, Mahder; Glick, Connor; Raghuram, Anupama; Beavin, Leslie; Cavallazzi, Rodrigo; Balcom, Dawn; Wolf, Leslie; Palmer, Kenneth E.; Carrico, Ruth; Ramirez, Julio A. and the Center of Excellence for Research in Infectious Diseases (CERID) Coronavirus Study Group (2020). "Endemic Human Coronaviruses in Hospitalized Adults with Community-Acquired Pneumonia in the City of Louisville, Kentucky," *The University of Louisville Journal of Respiratory Infections*: Vol. 4, Iss. 2, Article 4.

Received Date: November 4, 2019

Accepted Date: May 6, 2020

Published Date: June 19, 2020

Funding Source: The author(s) received no specific funding for this work.

Conflict of Interest: All authors declared no conflict of interest in relation to the main objective of this work.

Introduction

Population-based studies are necessary to define the burden of a particular disease in a well-defined geography. Once data on the burden of disease, such as incidence, hospitalizations, or mortality, is obtained for a defined population, extrapolations to larger geographies may be possible. Our group recently conducted a population-based study to define burden of community-acquired pneumonia (CAP) in Louisville, Kentucky, and used these estimates to extrapolate the burden of CAP to the United States (US). [1] Other investigators have conducted similar studies evaluating the burden of CAP in Chicago, Illinois and Nashville, Tennessee, extrapolating their results to the US population. [2] One weakness that applies to all population-based studies performed in the US is that investigators perform the extrapolations without providing objective statistical evidence indicating the city used is a suitable surrogate for the United States. A validation of these extrapolations is warranted. Sociodemographic, economic, and health-related statistics in the US are publicly available from several government agencies including the US Census Bureau and the Centers for Disease Control and Prevention (CDC). Data are typically aggregated for a range of geographies, including the city, and are also available for the United States as a whole.

Computational methods for comparing city and country data exist, but require borrowing techniques frequently used in other fields of science and mathematics. Specifically, the fields of cluster analysis and machine learning have measures of similarity and dissimilarity that are used to classify data. [3,4] By treating a city's or country's set of statistics as a set of coordinates, we can employ techniques used to mathematically compute distance, such as Euclidean or Manhattan distance. In this manner, cities with the least computed distance from the United States can be seen as being the closest, and thus most generalizable to the United States. Such cities may be the best candidates for conducting population-based studies.

The objective of this study was to propose novel computational methods for comparing sociodemographic, economic, and health-related data of major US cities to the US average, and to evaluate which cities are closest to the general US demographics.

Methods

Study Design

This was a secondary data analysis of publicly available databases containing US sociodemographic and health-related data taken from nationally administered surveys. Two entities were used for analysis: cities that met inclusion criteria, and the United States as a whole. Cities were eligible for inclusion in the study if 1) they had over 500,000 population, as cities with less population than this may not have a sufficient number of cases per 100,000 population to adequately extrapolate to the US population, and 2) they were included in one of the metropolitan and micropolitan statistical area (MMSA) reports of the 2017 Behavioral Risk Factor Surveillance System (BRFSS) SMART database. [5,6] The BRFSS is a yearly national survey that interviews US participants concerning several health behavior and health status variables. We chose only the largest city for MMSAs that consisted of multiple cities.

Variables

In addition to health behavior and health status variables from the 2017 BRFSS, sociodemographic and economic variables were taken for these cities from the 2017 Census Quick Facts [7], a summary of many demographic, housing, and economic data for the selected cities and United States taken from the 2017 American Community Survey 5-year estimates. [8] The American Community Survey is another yearly national survey which measures sociodemographic, housing, and economic variables of individuals and households; five year totals are combined to reduce the margin of error (MOE) for variable estimates. In total, 58 demographic, housing, economic, health behavior, and health status variables for each entity were used for analysis. A listing of the variables and their source dataset is found in **Supplementary Table 1**. All variables were reported as proportions; variable data was not standardized as the resulting distance measure would be less interpretable.

Distance measures

For each entity, each set of variables represented a 58-element vector in real space. For any two entities U and V, distance measures of Euclidean distance (d_e) and Manhattan distances (d_m) were used was used to measure the distance between entities, by the following formulae:

Where and correspond to the 58-dimensional vectors for the entities, and *n* was equal to 58. In this manner, the Euclidean distance could be seen as the "straight line" distance (in proportion) from one entity to another, and Manhattan distance can be seen as the sum of the absolute difference (in proportion) from one entity to another. Additionally, the average absolute distance, , was calculated by the following formula:

Notably, the average absolute difference was the Manhattan distance divided by the number of variables—this represented the average deviation (in proportion) from one entity to another per variable.

Statistical Analysis

Euclidean distance, Manhattan distance, and the average absolute difference were calculated for each pair of entities. For each city, distances from the US statistics were reported. Additionally, hierarchical clustering with Euclidean distance using Ward's method was performed, and a cluster dendrogram was created. Clusters were identified using a cutoff height of half the total height of the dendrogram. To visualize similarity and dissimilarity, a heatmap was created to show the distance between each city and the United States using Euclidean distance. R version 3.5.1 [9] and R package factoextra [10] were used for analysis.

Results

From a total of 136 MMSAs, 28 cities met inclusion criteria of at least 500,000 residents. Data for the 58 sociodemographic, housing, economic, health behavior, and health status variables for each of the 28 cities are shown in **Supple-**

mentary Table 2.

Oklahoma City, Oklahoma had the lowest distance from the United States, with a Euclidean distance of 0.261 and a Manhattan distance of 1.519. Louisville, Kentucky had the second lowest distance for both Euclidean distance and Manhattan distance, with distances of 0.286 and 1.545, respectively. Euclidean distance and Manhattan distance from the United States for all cities meeting inclusion criteria are shown in **Table 1**.

Table 1. Euclidean distance, Manhattan distance and average absolute differences from cities to the United States. Cities appear in order based on Euclidean distance.

City	Euclidean Distance	Manhattan Distance	Average Absolute Difference
Oklahoma City, OK	0.261	1.519	0.026
Louisville, KY	0.286	1.545	0.027
Indianapolis, IN	0.315	1.549	0.027
Nashville, TN	0.329	1.631	0.028
Jacksonville, FL	0.354	1.783	0.031
Portland, OR	0.358	1.948	0.034
Denver, CO	0.369	1.994	0.034
Columbus, OH	0.399	1.869	0.032
Fort Worth, TX	0.401	1.952	0.034
Phoenix, AZ	0.415	1.795	0.031
Albuquerque, NM	0.429	1.679	0.029
Austin, TX	0.438	2.219	0.038
Charlotte, NC	0.472	1.954	0.034
Seattle, WA	0.520	2.743	0.047
Chicago, IL	0.544	2.259	0.039
Sacramento, CA	0.568	2.527	0.044
Boston, MA	0.571	2.765	0.048
Milwaukee, WI	0.592	2.411	0.042
Philadelphia, PA	0.603	2.283	0.039
Dallas, TX	0.635	2.979	0.051
San Antonio, TX	0.695	2.693	0.046
Washington, DC	0.696	2.903	0.050
Houston, TX	0.699	3.132	0.054
New York, NY	0.716	2.994	0.052
Los Angeles, CA	0.781	3.262	0.056
Baltimore, MD	0.815	2.665	0.046
Memphis, TN	0.884	3.470	0.060
El Paso, TX	0.995	3.731	0.064

Hierarchical clustering and distance heatmaps
Results from hierarchical clustering using Euclidean distance are shown in **Figure 1**. The height represents the Euclidean distance between either entities or clusters. The maximum height between clusters was found to be 1.85. With a cut-point of height of 0.925, four distinct clusters were identified, indicated with boxes around each cluster. Heat maps for the distances between cities and the United States are shown in **Figure 2**.

Discussion

To our knowledge, this represents the first study comparing the sociodemographic, economic, and health-related statistics of US cities with the US average. This study shows that among cities with at least 500,000 residents, Oklahoma City, is the most similar to the US, followed closely by Louisville. On average, Louisville deviates from any given US variable studied by less than 3 percent. Using two different techniques to evaluate the distance between each city and the US, Euclidean distance and Manhattan distance, Louisville remains as the second most closest city to the US.

We identified four demographic clusters of cities in the US. Cluster A (as labeled in the dendrogram) can be seen in the center of the heatmap and contains the US average as well as the values of 8 cities: Oklahoma City, Oklahoma; Louisville, Kentucky; Indianapolis, Indiana; Nashville, Tennessee; Jacksonville, Florida; Portland, Oregon; Denver, Colorado; and Columbus, Ohio. These cities are the best representation of the US population according to these calculations. The

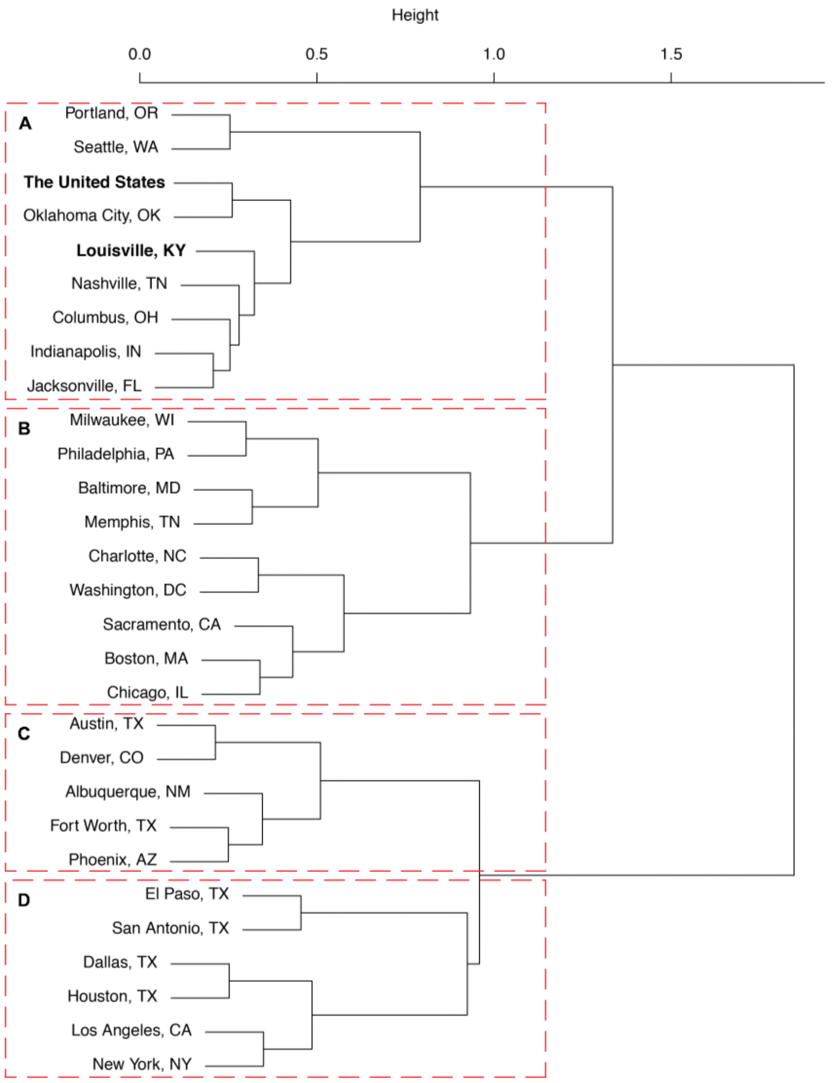

Figure 1. Hierarchical clustering dendrogram of cities and the United States by Euclidean distance.

best extrapolation generalizability will be obtained from cities within cluster A, followed by cities in cluster B. The weakest generalizability will originate from cities in clusters C and D.

One strength of this study is that we chose a great number and variety of sociodemographic, economic, health behavior, and health status variables. The importance of this is two-fold: 1) cities with outlying deviations for one or two variables will be less impacted by them, and 2) it improves the concept of generalizability by incorporating more than just simple demographics.

Our use of Euclidean distance and Manhattan distance means that we can measure the deviation in an interpretable way. Each distance represents the distance in proportion from the United States variables. Future research in this field can be performed using more sophisticated techniques such as cosine similarity or measures of correlated distance, such as Pearson or Spearman correlation distances.

Our study has several limitations. It is possible that smaller cities may be proximally closer to the United States' statistics, but there may not be a large enough population to adequately study the disease in question. For the purposes of conducting population-based studies, cities with at least 500,000 residents are likely necessary to have a sufficient population for studying disease. Additionally, while we were exhaustive in the breadth of variables selected, they were chosen, and it is possible that such a selection could unintentionally bias the distances produced. Future studies using

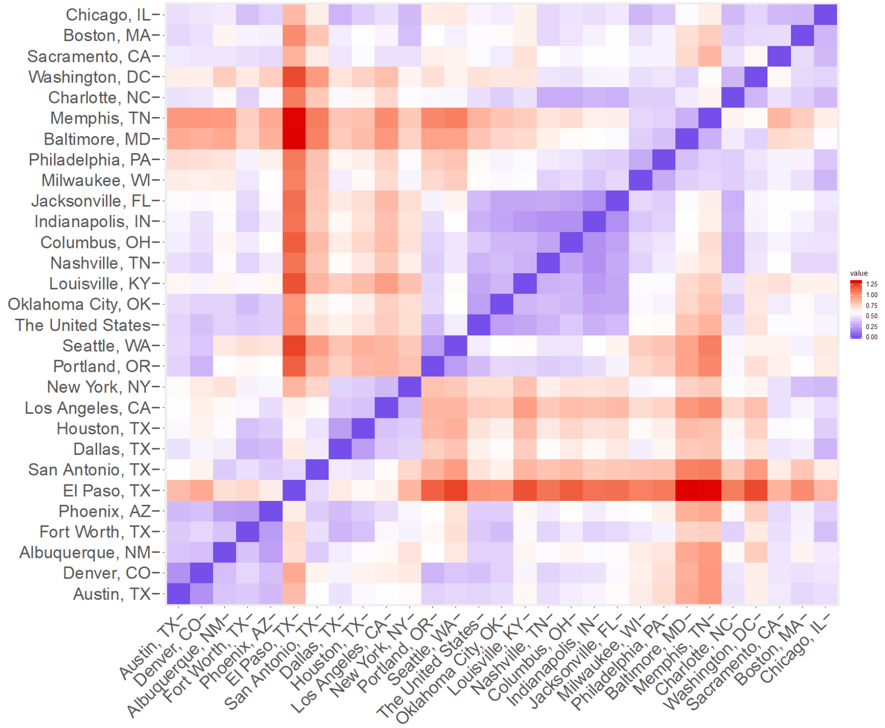

Figure 2. Euclidean distance heat map. Blue colors show similarity; red colors show dissimilarity by Euclidean distance.

this technique may prefer a random sampling of variables to the ones we selected. Additionally, for our hierarchical clustering, we chose a subjective cut point for clustering; a sensitivity analysis on this cut point was not performed.

It is important to note that the variables selected were estimates and had accompanying margins of error (MOEs), and a further limitation of our study is that we did not incorporate such margins of error into our calculations. Future studies should implement ways of handling MOEs to account for the underlying uncertainty. One method could be an inverse MOE weighted distance, akin to inverse variance weighted averages.

In conclusion, this study offers an objective approach that can be used to help define the validity of extrapolating US estimates from any particular population-based study performed in the US. Our data indicates that Oklahoma City and Louisville most closely represent the average US statistics. Data from population-based studies performed in the city of Louisville result in valid extrapolations to the US as a whole given its high level of similarity to the US compared to other large cities.

References

1. Ramirez JA, Wiemken TL, Peyrani P, Arnold FW, Kelley R, Mattingly WA, et al.; University of Louisville Pneumonia Study Group. Adults hospitalized with pneumonia in the United States: Incidence, epidemiology, and mortality. Clin Infect Dis. 2017 Nov;65(11):1806–12.
2. Jain S, Self WH, Wunderink RG, Fakhran S, Balk R, Bramley AM, et al.; CDC EPIC Study Team. Community-acquired pneumonia requiring hospitalization among U.S. adults. N Engl J Med. 2015 Jul;373(5):415–27.
3. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. 6th ed.: Pearson-Prentice Hall; 2007.
4. Kassambara A. Practical guide to cluster analysis in R: Unsupervised machine learning. STHDA; 2017 Aug 23.
5. Behavioral Risk Factor Surveillance System SMART. City and County Survey Data [Internet]U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2017.
6. Behavioral Risk Factor Surveillance System Survey Data. [Internet]U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2017.
7. U.S. Census Bureau. U.S. Census Bureau Quickfacts: United States. 2017. Available from: https://www.census.gov/quickfacts/fact/table/US/PST045218
8. US Census Bureau. ACS demographic and housing estimates, 2013–2017 American Community Survey 5-year estimates.

9. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna.

10. Kassambara A, Mundt F. Factoextra: extract and visualize the results of multivariate data analyses. R package version. 2017;1(4):2017.