Clustering Time Series Data Considering Both Trend Filtering and Subject Level Closeness

Zhiren Wang

{zhirenwang@uchicago.edu}

Collegiate Division-Mathematics, The University of Chicago, IL, USA

Abstract. This paper presents a novel clustering methodology that integrates both longitudinal trends and multi-dimensional proximity of lung cancer rates across 50 U.S. states and Washington D.C. from 1969 to 2019. By synthesizing temporal dynamics and local similarities, the proposed hybrid algorithm refines traditional clustering techniques to offer a comprehensive understanding of state-level lung cancer trends. The methodology combines K-means clustering with a medoid-based optimization approach, capturing the evolving patterns in lung cancer rates while mitigating the impact of outliers. Results reveal distinct clusters that reflect shared historical trajectories in health outcomes, outperforming traditional methods such as K-means, Radial Basis Function (RBF) networks, and Support Vector Machines (SVM) in sensitivity to temporal variations. The findings provide insights into regional health behaviors and interventions, highlighting the significance of integrating time-series analysis with clustering frameworks in public health research.

Keywords: Clustering Methodology, Time Series Analysis, Medoid-Based Optimization, Lung Cancer.

1 Introduction

Cluster analysis is a powerful tool widely used across various fields, including quantitative marketing, biology, and epidemiology, to uncover patterns in complex datasets. This paper applies cluster analysis to explore lung cancer rate trends across all 50 U.S. states and Washington D.C. over a 51-year period from 1969 to 2019. By examining both general and regional trends, we aim to identify commonalities and variations in lung cancer incidence. While traditional time series analyses provide insight into temporal trends, they often fail to account for geographic, socioeconomic, and political differences that may influence cancer rates across states. Conversely, static cross-sectional analyses offer little insight into how trends evolve over time. This paper addresses these limitations by introducing a novel clustering methodology that combines both time series trends and cross-state similarities, offering a more holistic understanding of lung cancer dynamics in the U.S. from 1969 to 2019.

The first section focuses on identifying overarching temporal patterns that characterize lung cancer rate changes across the U.S. This analysis investigates whether the states exhibit a consistent trend, with a shared peak period between 1990 and 2000, and examines potential nationwide factors such as public health policies and medical advancements that may have contributed to this pattern. The second part of this section delves into regional differences,

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354637

classifying states based on the year they reached peak lung cancer rates and exploring the socioeconomic and political factors influencing these variations.

The next section introduces various clustering methods, such as K-means, Radial Basis Function (RBF) networks, and support vector machines (SVM), to categorize the states into distinct groups. This section applies dimensionality reduction techniques, such as Principal Component Analysis (PCA), to better visualize and interpret the clustering results. The clustering process aims to uncover similarities between states that reflect the complex interplay of lung cancer trends over time. Next, we apply autoregressive models to evaluate the temporal dependence of lung cancer rates within each state, shedding light on how historical data influences future trends.

Finally, by combining time series analysis with clustering techniques, we propose a novel framework that simultaneously considers the longitudinal trends of lung cancer rates and the multi-dimensional proximity of states. This integrated approach provides a more comprehensive understanding of state-level clustering and reveals key regional differences, with implications for broader research areas involving time- and space-driven data.

2 Related works

Clustering time series data is a crucial technique for uncovering patterns, especially in fields dealing with large datasets like finance, biology, and medicine. The challenge lies in addressing both the underlying trends in time series and the closeness between subjects. Various methods have been proposed to capture these dimensions, each contributing unique approaches to handling high-dimensional, noisy, or incomplete data.

Wenig et al. [1] proposed the JET algorithm, which combines coarse-grained pre-clustering with efficient hierarchical clustering. This method addresses challenges in variable-length time series, such as those encountered in jet engine testing. JET demonstrates superior accuracy and computational efficiency compared to existing techniques. In a complementary approach, Fokianos and Promponas [2] explored clustering using spectral density functions, applying frequency domain representations like the periodogram to define similarity measures, particularly for biological datasets such as gene expression data.

Recent advancements in deep learning have further influenced time series clustering. Alqahtani et al. [3] reviewed deep time-series clustering (DTSC) methods, demonstrating that techniques like modified Deep Convolutional Autoencoders can capture spatial and temporal dependencies more effectively than traditional models, especially for complex movement data. Additionally, Zakaria et al. introduced unsupervised shapelets for clustering, focusing on local patterns rather than relying on distance-based methods [4]. Similarly, Paparrizos and Gravano developed the k-Shape and k-MultiShapes algorithms, using a shape-based distance (SBD) measure to enhance clustering accuracy without complex parameter tuning [5]. The robustness of k-Shape was further validated by Paparrizos and Gravano, who demonstrated its effectiveness across diverse datasets [6].

Holder et al. [7] conducted an extensive evaluation of elastic distance measures, including Dynamic Time Warping (DTW) and Move-Split-Merge (MSM). Their findings revealed that MSM and Time Warp Edit (TWE) distances, combined with k-medoids clustering, outperform traditional DTW-based approaches in terms of accuracy and interpretability. SOMTimeS, introduced by Javed et al. [8], is a self-organizing map-based algorithm that incorporates pruning strategies to reduce unnecessary DTW computations, achieving faster runtimes without sacrificing accuracy. SOMTimeS demonstrated its utility in both benchmark datasets and healthcare applications, showcasing its scalability and effectiveness. Similarly, Ma et al. [9] introduced Deep Temporal Clustering Representation (DTCR), a seq2seq-based model that integrates temporal reconstruction and K-means objectives to improve cluster structures. They further examined the role of pre-trained models in clustering tasks, showcasing the advantages of Transformer-based architectures and transfer learning for clustering tasks involving large-scale datasets [10].

Anomaly detection is another critical area, as highlighted by Izakian and Pedrycz [11-12], who employed fuzzy C-means clustering to identify structural changes in time series data by combining original and autocorrelation representations. They further refined this work using Dynamic Time Warping distance for shape-based clustering, which handles time misalignments paired with fuzzy clustering techniques [13]. Meanwhile, Khaleghi et al. extended the clustering framework to handle both offline and online settings [14]. They proposed asymptotically consistent algorithms that cluster time series based on their generating distributions. A broader review by Aghabozorgi et al. emphasized the growing significance of unsupervised learning in time-series clustering amid the rise of big data, underscoring the need for scalable, domain-independent methods [15]. These recent advances reflect a transition from traditional distance-based techniques to advanced approaches utilizing shape analysis, feature extraction, deep learning, and fuzzy logic, enhancing accuracy and interpretability across various applications.

3 Lung cancer rate trends

This section analyzes the trends in lung cancer incidence across all 50 U.S. states and Washington D.C. over a 51-year period, spanning from 1969 to 2019. The analysis is guided by two primary objectives.

The first objective is to assess whether a generalizable temporal pattern exists that characterizes the overall trend in lung cancer rates across all states. Specifically, this entails examining whether the 51-year evolution of lung cancer rates exhibits consistent features across the entire dataset, despite the diversity of each state.

The second objective is to explore whether meaningful regional variations exist in the lung cancer rate trends. This aspect seeks to uncover whether certain states deviate from the general trend and whether distinct temporal patterns can be divided into groups. Identifying regional differences could reveal important insights about the role of localized factors and might help tailor future public health interventions to address specific needs in different parts of the country.

3.1 General trend analysis

To address the first objective, we construct a time-series plot that represent the annual lung cancer rate for all 50 U.S. states and Washington D.C. from 1969 to 2019.

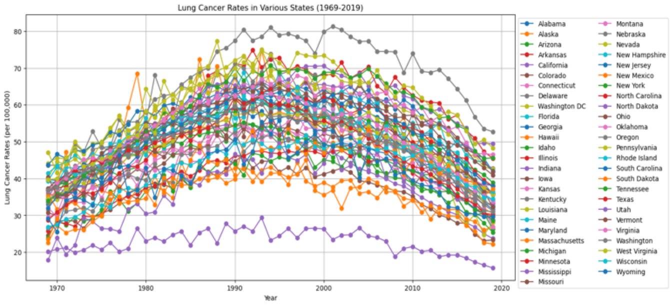

Fig. 1. Lung Cancer Rates in Various States.

In nearly every state, the time-series data exhibits a characteristic inverse bell-shaped curve, with lung cancer rates rising to a peak before subsequently declining. This peak consistently occurs within a relatively narrow timeframe of 1990 to 2000, suggesting the presence of common underlying drivers influencing lung cancer rates across the country.

Several plausible explanations warrant further investigation to explain this synchronized pattern. One potential factor could be changes in public health policies during this period, particularly those related to smoking cessation efforts, which intensified during the 1980s and 1990s. This era saw the implementation of stricter tobacco regulations, increased taxation on cigarettes, and widespread anti-smoking campaigns. Additionally, shifts in smoking behavior itself, such as a reduction in smoking prevalence following the dissemination of information on the dangers of tobacco, could account for the observed decline in rates after the peak.

Furthermore, advances in medical treatments, particularly in the early detection and management of lung cancer, may have played a role in the downward trend observed after the peak. Innovations in diagnostic tools, such as the increased use of imaging technologies, may have facilitated earlier detection, contributing to improved survival rates and a reduction in incidence over time. Finally, socioeconomic factors, including changing patterns in healthcare access and education, could have influenced both the initial rise in lung cancer rates and the subsequent decline.

While the time-series analysis provided clear evidence of a generalizable trend, the specific mechanisms driving the rise and fall of lung cancer rates during this period remain to be fully understood.

3.2 Individual trend analysis

To address the second objective—understanding regional variations in lung cancer trends—I conduct a more granular analysis by grouping states according to the year in which their lung cancer rates reached the peak. The peak year is a critical point in the temporal trajectory of lung cancer incidence, reflecting the culmination of various contributing factors before the onset of a decline. By categorizing states based on their peak year, I aim to uncover temporal and regional disparities in lung cancer dynamics that have been masked by the broader nationwide trend.

The states are divided into three categories based on the year of their peak rates: (1) states that reached their peak before 1992, (2) between 1993 and 1995, and (3) after 1996. This stratification is chosen to capture potential shifts in lung cancer trends over time while maintaining meaningful group sizes for comparison. The resulting distribution is as follows: 12 states reached their peak before 1992, 25 states peaked during 1993 to 1995, and 14 states peaked after 1996.

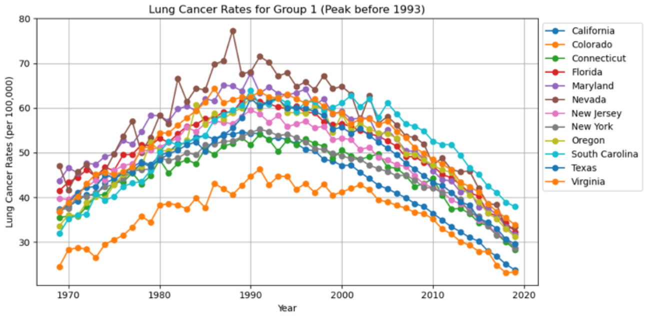

Fig. 2. Lung Cancer Rates for Group 1 (Peak before 1993).

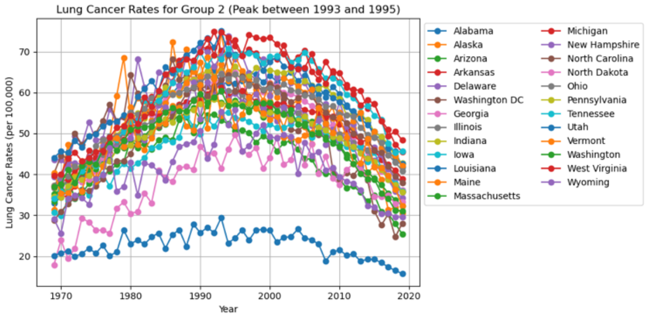

Fig. 3. Lung Cancer Rates for Group 2 (Peak between 1993 and 1995).

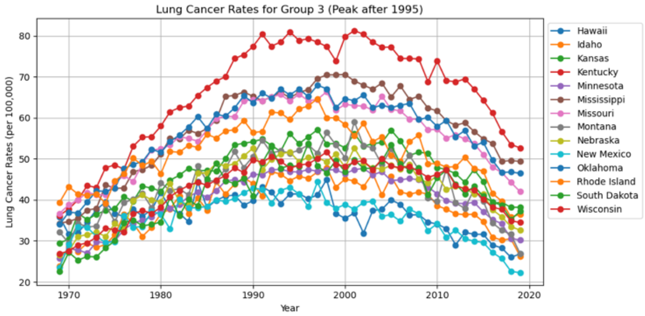

Fig. 4. Lung Cancer Rates for Group 4 (Peak after 1995).

The concentration of peak years within the 1993-1995 window, encompassing nearly half of the states, raises questions about the timing and nature of the factors influencing lung cancer incidence. The fact that 25 states shared this brief peak window suggests the presence of a strong nationwide influence, such as the introduction and enforcement of stricter tobacco control policies, which gained momentum in the early 1990s. This period saw the passage of laws mandating smoke-free environments, increased cigarette taxes, and a surge in anti-smoking media campaigns.

In contrast, the states peaking before 1992 and after 1996 present more varied trajectories. To further explore the distinct lung cancer trends observed across the different groups, we select two representative states from each group for a more detailed, state-specific analysis. The selection of states from each group was made with the intention of capturing both geographic diversity and variability in socioeconomic, political, and geographical contexts. Here are the time-series plots and analysis for the representative states:

California lung cancer trend

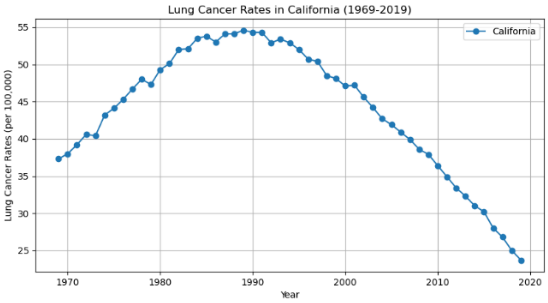

Fig. 5. Lung Cancer Rates in California (1969-2019).

California exhibits a relatively smooth lung cancer rate change, characterized by a steady and consistent increase from 1969 to 1989, followed by a marked and sustained decline. This pattern reflects the state's early and aggressive anti-smoking policies.

New York lung cancer trend

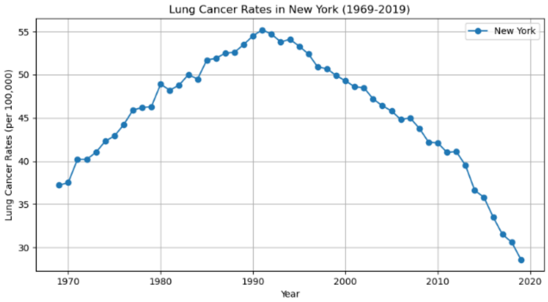

Fig. 6. Lung Cancer Rates in New York (1969-2019).

The trend in New York mirrors that of many northeastern states, showing a gradual increase in lung cancer rates from 1969 through 1991, followed by a pronounced decline. This decline reflects the effectiveness of public health policies, including higher taxes on cigarettes, restrictions on tobacco advertising, and the establishment of smoke-free environments in public places.

Alabama lung cancer trend

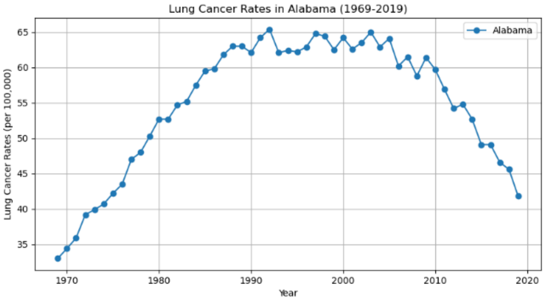

Fig. 7. Lung Cancer Rates in Alabama (1969-2019).

The lung cancer rate in Alabama shows a clear upward trend from 1969 through the mid-1980s, reflecting the growing prevalence of smoking and limited public health interventions during that period. Rather than an immediate decline, the trend is characterized by a fluctuating plateau in 1990s to 2000s. This plateau suggests a period of stagnation in either public health efforts or in behavioral changes related to smoking.

Illinois lung cancer trend

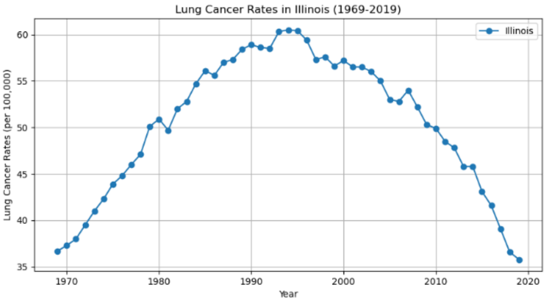

Fig. 8. Lung Cancer Rates in Illinois (1969-2019).

The lung cancer rate in Illinois shows a steady increase to a peak in 1994, reflecting the growing prevalence of smoking and limited public health interventions during this time. Following this peak, the rates plateaued in the late 1990s and early 2000s, suggesting stagnation in public health efforts or behavioral changes among smokers.

Kentucky lung cancer trend

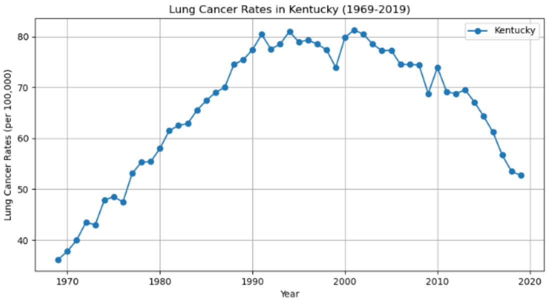

Fig. 9. Lung Cancer Rates in Kentucky (1969-2019).

Kentucky's lung cancer trend reflects the state's longstanding challenges with high smoking rates and limited public health interventions. Unlike states that saw rapid declines post-peak, Kentucky's rates show slow decline through the early 2000s, indicating the persistence of high smoking rates and slower adoption of anti-smoking measures. The subsequent decline, while evident, is uneven and far less pronounced than in states with stronger public health infrastructure.

Mississippi lung cancer trend

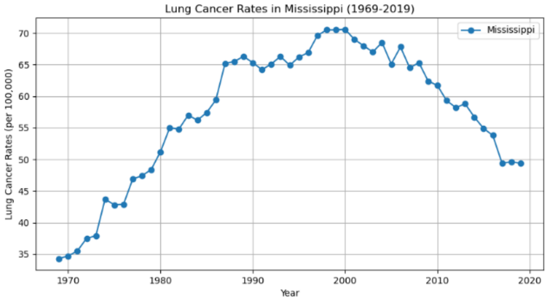

Fig. 10. Lung Cancer Rates in Mississippi (1969-2019).

Mississippi's lung cancer trend is characterized by a steady increase until 2000, at which point the rate begins to decline, though slower and more irregular compared to other states. The data suggests that while lung cancer rates in Mississippi peaked later than in many other states, the decline that followed was not as pronounced. This gradual and inconsistent decline reflects deeper public health challenges in the region, including economic disparities and limited access to healthcare.

3.3 Interpretation

By dividing states into 3 groups based on the peak year of lung cancer rates, we can investigate common characteristics of states in each group and formulate hypotheses about why their peaks occurred when they did:

Group 1: Peak Before 1992: States in this group, such as California, New York, and Florida, experienced their peak lung cancer rates before 1993. These are often states with large populations and early public health interventions, such as stringent anti-smoking campaigns and clean air regulations. Many of these states are in the Northeast and West, regions known for early adoption of anti-smoking legislation and higher public awareness of the risks associated with smoking. This early peak aligns with their proactive public health policies, industrial regulation, and higher socioeconomic status.

Group 2: Peak Between 1993 and 1995: The majority of states fall into this category, indicating a broader national trend during this period. Many states in this group are part of the Midwest and the South, where the tobacco industry historically played a significant economic role. The timing of the peak suggests these states began to see the effects of anti-smoking campaigns and policy changes slightly after the first group, likely due to later implementation or slower adoption of public health measures.

Group 3: Peak After 1996: States like Kentucky, Missouri, and Mississippi, which experienced their peak lung cancer rates after 1995, tend to be more rural, with less urbanization and possibly delayed implementation of smoking restrictions. Many of these states are located in regions where smoking was culturally more prevalent, and public health campaigns may have faced more resistance or slower uptake. Furthermore, these states might have had less access to advanced medical technologies and screening programs, which could delay the identification of lung cancer trends and peaks.

4 Cluster analyses

In this section, we aim to employ various clustering methodologies to categorize all 50 U.S. states and Washington D.C. into distinct clusters based on their characteristics. Each state is represented as a 51-dimensional data point within a multi-dimensional space, where the dimensions correspond to years 1969 to 2019. By using different clustering methods, we assess the similarities between states and identify which method most effectively captures the inherent structure of the data.

The clustering methods we will apply include K-means clustering, Generalized Additive Models (GAM), Radial Basis Function (RBF) networks, and support vector machines with hyperplane classification. Following the clustering analysis, we will apply Principal Component Analysis (PCA) to reduce the dimensionality of our data from 51 dimensions to two principal components. It transforms the original variables into a new set of uncorrelated principal components, which capture the maximum variance present in the data. This reduction facilitates a clearer visualization and interpretation of the clustering results.

4.1 Three clustering methods

The subsequent graphs will illustrate the individual clustering outcomes obtained from the three distinct methods, providing insights into the similarities and differences between states. Through comparative analysis, we determine which clustering approach yields the most coherent and interpretable groupings of the states.

Radial Basis Function (RBF)

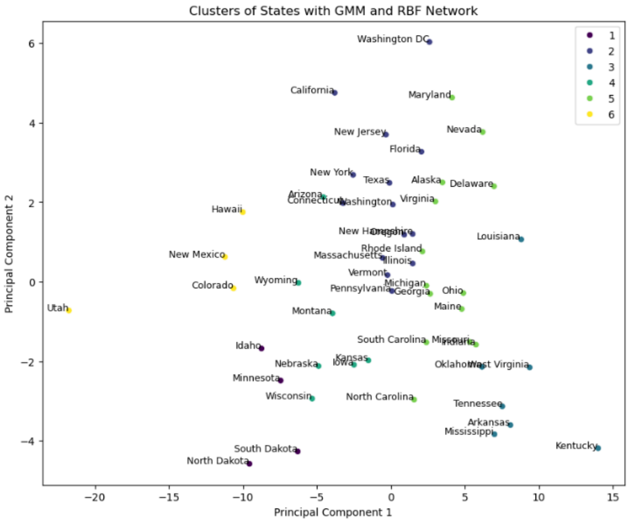

Fig. 11. Clusters of States with GMM and RBF Network.

RBF clustering employs a radial basis function to determine the proximity of data points to centroids in multi-dimensional space, mathematically expressed as:

$$K(x_i, c_j) = exp(-\gamma\ ||x_i - c_j||^2)$$

where x_i is a data point, c_j is a centroid, and γ controls the kernel width. This method groups nearby data points, resulting in compact clusters that capture local structures effectively. However, RBF clustering may overlook broader trends and relationships over time, as it prioritizes proximity and local compactness, which can fragment larger, meaningful patterns.

K-Means

K-Means is an unsupervised clustering technique that partitions data into K clusters by minimizing the variance within each cluster. It uses Euclidean distance to assign points to the nearest centroid. K-Means is sensitive to initial centroid placement and can converge to local minima, leading to clusters that may not reflect the true structure of the data.

Support Vector Machine (SVM)

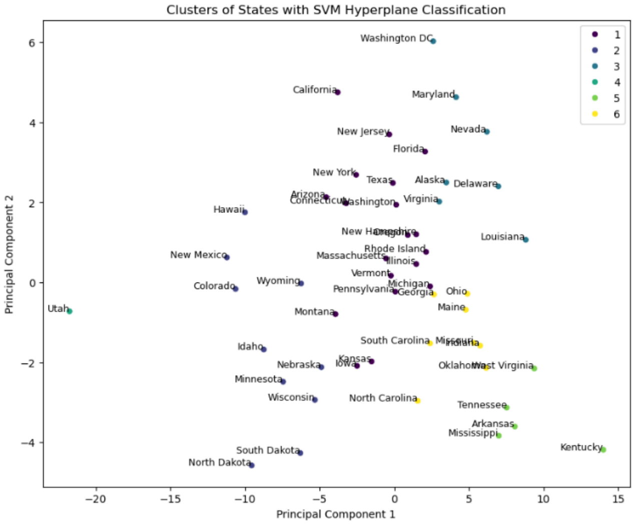

Fig. 12. Clusters of States with SVM Hyperplane Classification.

SVM is primarily a supervised learning algorithm used for classification and regression tasks. However, it can also be adapted for clustering purposes by using the concept of support vectors to identify the most important data points that define the boundaries between different clusters. This can lead to a focus on extreme values and outliers, resulting in clusters that may be overly sensitive to specific states or lung cancer rates.

4.2 E. Analysis and comparison

RBF Clustering

The RBF clusters show both local and broader regional groupings: Cluster 1 is consistent with lower lung cancer rates, emphasizing regional similarities; cluster 2 captures high-population states with generally higher rates, similar to the SVM result, showing significant healthcare influences; cluster 3 again highlights the Southern U.S., emphasizing long-term health trends and shared risk factors; cluster 4 shows some overlap with K-Means, indicating common regional health dynamics; cluster 5 demonstrates RBF's capacity to group based on healthcare and socioeconomic access, capturing states with shared characteristics affecting lung cancer trends; cluster 6 mirrors the K-Means grouping for the less populated states.

RBF clustering's focus on proximity allows it to emphasize local health patterns effectively. However, it may struggle to integrate broader trends across the states, leading to fragmentation of states that share significant historical health patterns.

K-Means Clustering

The K-Means clusters yield distinct geographic and demographic groupings: Cluster 1 features states with lower lung cancer rates, likely reflecting shared demographic factors and healthcare practices; cluster 2 groups East Coast states with generally high healthcare accessibility and similar lung cancer statistics; cluster 3 captures a broader array of states, potentially revealing shared health risks related to socioeconomic factors prevalent in the South; cluster 4 retains a similar grouping to the SVM method, indicating that proximity impacts cluster assignment; cluster 5 consolidates states with generally high populations and diverse socioeconomic contexts, emphasizing factors like healthcare access and prevention efforts; cluster 6 clusters states often seen as having similar environmental conditions and lung cancer risk profiles.

K-Means efficiently captures spatial and demographic similarities but may simplify relationships by not considering how temporal trends influence clustering. As a result, states with differing trajectories over time could end up in the same cluster, potentially masking underlying health patterns.

Hyperplane SVM Clustering

The SVM clusters show a mix of states with shared characteristics: Cluster 1 contains states with relatively high lung cancer rates, including populous states like California and New York, which often exhibit significant variations in healthcare access and environmental factors affecting lung health; cluster 2 groups states with lower population densities and similar healthcare profiles, indicating a focus on less populated areas that may share specific regional health trends; cluster 3 captures states in the Southern U.S., highlighting a region with historically high lung cancer rates linked to factors such as smoking prevalence and socioeconomic conditions; cluster 4 is singular, representing Utah, possibly indicating unique state policies or demographics affecting lung cancer rates; cluster 5 shows a diverse group of states, indicating shared socioeconomic factors influencing lung cancer risk; cluster 6 clusters states from the Southeast and Midwest, linking them through shared risk factors like socioeconomic status and healthcare access.

The SVM method's focus on boundaries allows it to identify significant groupings based on extreme values and regional variations. However, it may overlook nuanced, continuous trends between neighboring states, particularly those sharing similar health outcomes over time.

5 Autoregression analysis

For each state's lung cancer rate data from 1969 to 2019, we apply an autoregressive model of order 1 (AR(1)) to capture the temporal dependence of lung cancer rates on their immediate past values. The data is first pre-processed by removing any missing values to ensure accuracy in the subsequent analysis. After cleaning, an AR(1) model is fitted to the data, where the lung cancer rate at time t is modeled as a function of the rate at time t-1, along with an error term. The general form of the AR(1) model can be expressed as:

$$Yt = \alpha + \beta Yt - 1 + \epsilon t$$

The autoregressive model is fitted for each state, and several key parameters are recorded: the intercept α, the slopeβ, and two model evaluation criteria: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC provide a trade-off between model fit and complexity, with lower values indicating better models. The following table presents the results of the autoregressive (AR(1)) analysis for each U.S. state:

Table 1. Autoregression Results.

State	Intercept	Slope	AIC	BIC
Alabama	3.5212	0.9394	210.5280	216.2641
Alaska	10.1519	0.8095	330.7516	336.4877
Arizona	0.0070	0.9961	227.8955	233.6316
Arkansas	3.9427	0.9370	247.0776	252.8136
California	-2.2721	1.0451	159.4804	165.2164
Colorado	1.4350	0.9599	212.4235	218.1596
Connecticut	-0.2934	1.0033	210.6976	216.4336
Delaware	5.5170	0.9012	301.0493	306.7854
Washington DC	2.1660	0.9537	295.0836	300.8197
Florida	-1.7387	1.0302	183.2487	188.9848
Georgia	1.5287	0.9715	212.9110	218.6471
Hawaii	7.4512	0.7926	255.0782	260.8089
Idaho	4.7927	0.8774	250.0369	255.7730
Illinois	0.6472	0.9869	178.6853	184.4198
Indiana	3.6869	0.9371	214.2314	219.9675
Iowa	3.7214	0.9216	221.6777	227.3538
Kansas	4.7383	0.9034	226.7045	232.5113
Kentucky	4.7027	0.9339	242.095	247.8265
Louisiana	1.5299	0.9738	230.2132	235.9462
Maine	5.5939	0.8989	266.549	272.2851
Maryland	-1.3835	1.0122	223.2260	228.9620
Massachusetts	0.3416	0.9917	195.2212	200.9573
Michigan	1.7989	0.9661	180.3249	186.0609
Minnesota	1.9431	0.9539	176.1174	181.8534
Mississippi	3.9319	0.9370	216.2350	221.9711
Missouri	2.9491	0.9488	213.3013	219.0373
Montana	4.9697	0.8860	284.3537	290.0898
Nebraska	4.5650	0.8974	240.4758	246.2119
Nevada	1.4231	0.9700	283.7662	289.5022
New Hampshire	5.7451	0.8849	227.6677	283.4038
New Jersey	-1.9023	1.0347	183.0718	188.8079
New Mexico	4.6773	0.8679	246.6184	252.3544
New York	-1.9085	1.0378	163.5096	169.2457
North Carolina	2.8295	0.9483	205.5377	211.2737
North Dakota	6.6914	0.8309	237.1184	278.8545
Ohio	2.3630	0.9583	192.9451	198.6812
Oklahoma	4.8366	0.9188	226.7553	231.5094

Oregon	1.4059	0.9711	237.9064	243.6425
Pennsylvania	1.3088	0.9738	179.5990	185.3351
Rhode Island	3.2105	0.9366	250.8772	256.6133
South Carolina	2.8713	0.9472	217.5529	223.2889
South Dakota	6.3076	0.8555	264.6215	270.3576
Tennessee	3.5232	0.9436	219.2108	224.9468
Texas	-0.7839	1.0132	197.6945	203.4305
Utah	5.6360	0.7503	225.2002	230.9362
Vermont	9.5055	0.8020	301.1435	306.8796
Virginia	0.2377	0.9943	201.2063	206.9424
Washington	-0.0855	0.9993	212.8687	218.6048
West Virginia	3.8374	0.9394	243.4652	249.2013
Wisconsin	3.0304	0.9327	194.1780	199.9141
Wyoming	7.6282	0.8168	291.1577	296.8938

From the results, we observe that the average slope across all states is 0.9404. This mean slope value provides insight into the general strength of the autoregressive effect across the entire dataset, representing the average degree to which the lung cancer rates of one year are influenced by those of the previous year for all U.S. states. Specifically, a slope of 0.9404 implies that, on average, approximately 94.04% of the lung cancer rate from the prior year is carried forward into the next year. This points to a high level of temporal autocorrelation, meaning that the progression of lung cancer rates is highly consistent over time, with current rates being heavily dependent on historical values.

The fact that the mean slope is slightly below 1 indicates that while the autoregressive effect is strong, it is not perfect. The gap between the slope and 1 reflects some degree of variability or noise in the data, suggesting that factors outside the scope of the autoregressive model might influence the yearly changes in lung cancer rates. This could be due to external time-dependent effects not captured by the AR model.

To gain a deeper understanding of the data, this mean slope can be compared with the slopes of individual states. States with slopes that deviate significantly from the mean may stand out as having unique temporal behaviors. For example, a state with a slope significantly greater than 0.9404 would exhibit an even stronger autoregressive effect, where current lung cancer rates are more rigidly tied to the past, suggesting less variability over time. Conversely, a state with a slope much lower than the mean might demonstrate weaker autocorrelation, implying greater year-to-year fluctuations in lung cancer rates.

6 Trend filtering and Subject-Level closeness

The objective of this section is to develop a comprehensive clustering methodology that simultaneously accounts for both the longitudinal trends of lung cancer rates from 1969 to 2019 and the multi-dimensional proximity of the all 50 U.S. states and Washington D.C.. By synthesizing these two aspects, we create a novel clustering framework that offers a more holistic understanding of the data. This method not only facilitates deeper insights into the clustering of states but also has the potential for broader application across various research domains.

6.1 Methodological framework

This section presents a hybrid clustering algorithm designed to capture both temporal trends and local similarity within high-dimensional time series data. Our method refines initial cluster assignments iteratively, using a combination of K-means clustering and medoid-based optimization. The algorithm is applied to U.S. state-level lung cancer data spanning 51 years, aiming to discover meaningful clusters that reflect both aligned time trends and local similarity.

K-means initialization

K-means minimizes the sum of squared Euclidean distances between each data point and the centroid of its cluster. This algorithm assigns each state to the cluster whose centroid is closest in terms of Euclidean distance. However, the initial clusters require refinement to align with both time-series trends and local similarity.

Medoid-based Mean Path Calculation

Instead of using the arithmetic mean to calculate the "mean path" of each cluster, we employ a medoid-based approach. A medoid is the element within the cluster that minimizes the total distance to all other elements, ensuring robustness against outliers. For a cluster C_k, the medoid is defined as:

$$x_{medoid} = \arg min_{x_j \epsilon C_k} \sum_{x_i \epsilon C_k} ||x_i - x_j||_2$$

This ensures that the representative path of each cluster reflects the most typical behavior without being skewed by extreme values.

Reassignment of States to Clusters

After computing the medoid path for each cluster, we reassign each time series to the cluster with the closest medoid. The new assignment is based on minimizing the Euclidean distance between the time series and the medoid path:

$$Cluster(i) = \arg min_k ||x_i - x_{medoid,k}||_2$$

This step ensures that each state aligns with the cluster most similar to its temporal pattern.

Iterative Refinement and Convergence

The algorithm iterates between medoid calculation and state reassignment until convergence. At iteration t, the cluster assignments C^((t)) are compared with those from the previous iteration $C^{(t-1)}$. The process stops when the assignments stabilize:

$$C^{(t)} = C^{(t+1)}.$$

This iterative approach guarantees monotonic convergence, meaning the sum of distances within clusters cannot increase. While convergence to a global minimum is not guaranteed, the medoid-based approach minimizes sensitivity to outliers and ensures stable clustering results.

Mathematical interpretation

The medoid calculation can be interpreted as a LASSO regression with only the intercept, where the goal is to minimize the total error across clusters by choosing representative paths. Each iteration reduces within-cluster variability, improving alignment between clusters and underlying temporal trends.

6.2 Results and visualization

The final clusters are visualized by plotting the mean time series paths for each cluster over the 51 years. Additionally, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset and project the data into two components for visualization.

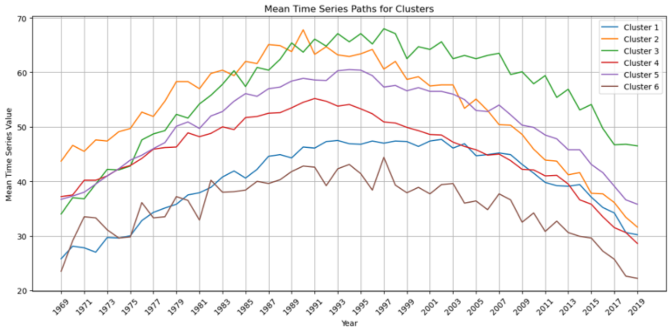

Fig. 13. Mean Time Series Paths for Clusters.

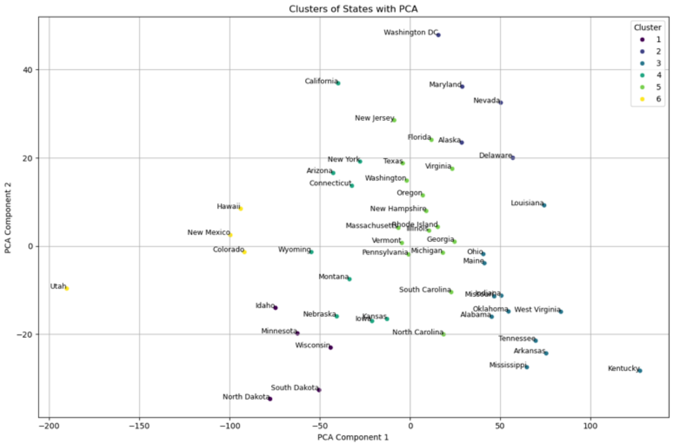

Fig. 14. Clusters of States with PCA.

The clustering achieved by the proposed method reveals a greater sensitivity to temporal dynamics and nuanced regional behavior in the lung cancer trends. Specifically:

Cluster 1 (e.g., Idaho, Minnesota, North Dakota) contains Midwestern states that generally exhibit slow-changing, stable health behaviors, reflecting modest but consistent improvements in lung cancer rates.

Cluster 2 (e.g., Alaska, Delaware, Washington DC) captures regions with divergent social structures (such as urban vs. rural contexts) but which share a unique temporal trajectory in disease rates, likely driven by transient population patterns and unique healthcare interventions.

Cluster 3 (e.g., Alabama, Mississippi, West Virginia) comprises Southern states with historically higher smoking rates and delayed peaks in lung cancer mortality, which aligns with socioeconomic challenges that impacted the decline of tobacco use in these regions.

Cluster 5 (e.g., Florida, Georgia, Illinois, New York) includes a mix of urbanized, economically developed states, which display early peaks in cancer rates followed by more rapid declines, in line with early public health interventions.

Cluster 6 (e.g., Colorado, Hawaii, Utah) groups states with outlier behaviors—regions characterized by relatively low smoking prevalence and unique geographic or policy factors that reduced lung cancer rates earlier than in other states.

By capturing both regional similarities and temporal trends, the proposed method allows for clusters that reflect not only spatial proximity but also shared historical trajectories in health outcomes, which traditional methods struggle to capture.

6.3 Comparison with traditional methods

This section provides a comparison of the proposed clustering method with traditional approaches, including K-Means clustering, Radial Basis Function (RBF) Networks, and Support Vector Machines (SVM).

The method developed in this paper focuses on clustering time-series data by integrating both temporal trends and local similarity. Instead of relying solely on the mean or centroid, it aims to identify a representative time-series path that minimizes the aggregated Euclidean distance to all series within the same cluster.

While K-Means is computationally efficient and works well for spherical clusters in Euclidean space, it assumes that the mean is a suitable representative of the cluster. However, this assumption can be limiting for time-series data, where temporal dependencies and trends must be considered. In contrast, the proposed method u is more robust to outliers and better suited for capturing complex temporal patterns.

The clustering process of RBF involves finding prototypes that minimize the radial distances from data points. These methods are particularly useful for nonlinear mappings, but the underlying assumption is that all data points are embedded within a fixed feature space defined by the RBF kernel. This limits the method’s ability to adapt to dynamic patterns over time. In comparison, the proposed clustering method explicitly incorporates temporal trends into the clustering process.

SVM-based clustering seeks to find maximum-margin hyperplanes that separate clusters, which may not be optimal for capturing time-series trends. Unlike SVMs, which focus on finding hyperplanes to separate data points, the proposed method aligns the clustering process with temporal dynamics by treating each time series as a multidimensional path and iteratively refining clusters based on both trend and local similarity.

The key conceptual distinction between the proposed approach and the traditional methods lies in the treatment of time series as evolving paths rather than static points in a feature space. The proposed method emphasizes the temporal coherence within clusters, which allows it to identify clusters with shared dynamic trends over time. Moreover, the iterative medoid-based approach ensures robustness to outliers, as the medoid path minimizes aggregated distances instead of relying on mean values.

Comparison with RBF Networks

The RBF-based clusters show that this method is able to capture non-linear relationships, but it still assumes the data points exist in a fixed feature space defined by the RBF kernel. As a result, RBF fails to capture longitudinal variations that occur within clusters over time. For example, RBF places both California and Maryland in the same cluster (Cluster 2) even though these states may have experienced different peak years for lung cancer rates.

The proposed method identifies distinct temporal trajectories—for instance, California's trend aligns more closely with other early-intervention states like New York and Connecticut, while Maryland follows a different trajectory grouped with urban areas like Washington DC.

Comparison with K-Means Clustering

The K-Means results show clear spatial groupings, suggesting that K-Means primarily clusters states based on static regional proximity. However, K-Means is limited by its reliance on centroids as averages, which overlooks temporal dynamics. For instance, K-Means places California and New York in the same cluster (Cluster 4), but these states have different historical trajectories in tobacco regulation. The proposed method, however, reflects these nuanced differences by placing New York and California in separate clusters. This illustrates the advantage of using medoid-based clustering for time-series data, where each series is represented by its most representative path rather than a mean trajectory.

Comparison with SVM-Based Clustering

The SVM-based results show that this method prioritizes maximizing the margin between clusters but lacks sensitivity to evolving trends within each cluster. SVM clusters are often defined by sharp boundaries, which are suitable for separating distinct classes but may overlook gradual transitions or shared historical patterns. For instance, SVM places Iowa, Kansas, and New York in the same cluster, despite significant differences in their cancer rate trajectories over time.

The proposed method’s iterative approach, in contrast, ensures that clusters evolve based on the progression of trends, not just separability by margins. For example, states with late peaks (e.g., Kentucky and Alabama) are grouped separately from states with early declines (e.g., California and New York), ensuring greater alignment between clusters and time-based health patterns.

7 Conclusion

The proposed clustering framework marks a significant advancement in the analysis of longitudinal data by addressing the limitations of traditional clustering methods. By incorporating both temporal trends and local similarities, this methodology allows for a more nuanced understanding of lung cancer rates across U.S. states, revealing clusters that align with shared historical trajectories and regional behaviors. The iterative refinement process, centered on medoid calculations, enhances robustness against outliers and emphasizes the dynamic nature of health outcomes. The results underscore the importance of considering temporal dynamics in clustering analyses, particularly in fields such as public health, where interventions and health policies must adapt to changing trends. Future research may extend this framework to explore other health outcomes or geographical contexts, fostering a deeper understanding of how health trends evolve over time and contribute to regional disparities. This approach not only enriches the literature on clustering methodologies but also serves as a foundation for more effective public health strategies that consider the complexity of longitudinal data.

References

[1] Wenig, P., Höfgen, M., & Papenbrock, T. (2024). JET: Fast estimation of hierarchical time series clustering. Engineering Proceedings, 2(1), 15–25.

[2] Fokianos, K., & Promponas, V. J. (2012). Biological applications of time series frequency domain clustering. Journal of Time Series Analysis, 33(6), 744–756.

[3] Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep time-series clustering: A review. Electronics, 10(23), 3001.

[4] Zakaria, J., Mueen, A., & Keogh, E. (2012). Clustering time series using unsupervised shapelets. In 2012 IEEE 12th International Conference on Data Mining (pp. 785–792). IEEE.

[5] Paparrizos, J., & Gravano, L. (2015, May). k-shape: Efficient and accurate clustering of time series. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 1855–1870).

[6] Paparrizos, J., & Gravano, L. (2017). Fast and accurate time-series clustering. ACM Transactions on Database Systems, 42(2), 1–49.

[7] Holder, C., Middlehurst, M., & Bagnall, A. (2023). A review and evaluation of elastic distance functions for time series clustering. Knowledge and Information Systems, 66, 765–809.

[8] Javed, A., Awan, M. I., Ahmad, H. F., Qadir, J., Yasin, M., & Rehman, A. (2024). SOMTIMES: Self-organizing maps for time series clustering and its application to serious illness conversations. Data Mining and Knowledge Discovery, 38(3), 813–839.

[9] Ma, Q., Zheng, J., Li, S., & Cottrell, G. W. (2019). Learning representations for time series clustering. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada.

[10] Ma, Q., Liu, Z., Zheng, Z., Huang, Z., Zhu, S., Yu, Z., & Kwok, J. T. (2024). A survey on time-series pre-trained models. IEEE Transactions on Knowledge and Data Engineering, 36(12), 7536–7554.

[11] Izakian, H., & Pedrycz, W. (2013). Anomaly detection in time series data using a fuzzy c-means clustering. In 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS) (pp. 1513–1518).

[12] Izakian, H., & Pedrycz, W. (2014). Anomaly detection and characterization in spatial time series data: A cluster-centric approach. IEEE Transactions on Fuzzy Systems, 22(6), 1612–1624.

[13] Izakian, H., Pedrycz, W., & Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence, 39, 235–244. DOI: 10.1016/j.engappai.2014.11.005.

[14] Khaleghi, A., Ryabko, D., Mary, J., & Preux, P. (2016). Consistent algorithms for clustering time series. Journal of Machine Learning Research, 17(3), 1–32.

[15] Aghabozorgi, S. R., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering—a decade review. Information Systems, 53, 16–38.