Paper RW03

Understanding the spread of COVID-19 in the United States using topology and machine learning for time series

Sergey Glushakov, Intego Group, Maitland, FL, USA
Iryna Kotenko, Intego Group, Kharkiv, UKRAINE
Kostiantyn Drach, Aix-Marseille University, Marseille, FRANCE /
Intego Group, Kharkiv, UKRAINE

ABSTRACT

Real-world data can be essential for our understanding of clinical data, especially with the emergence of phenomena such as the COVID-19 outbreak. In this paper, we analyze how the spread has advanced across the U.S. using novel graph-based machine-learning techniques. A topological model was extracted from several publicly available datasets and represented as a graph. In this graph, every node corresponds to a single county (over 3000 nodes), whereby two counties are connected with an edge only if they have similar patterns in the advance of the pandemic spread over specific timeframe. This geometric, data-driven approach reveals insights that would otherwise have been difficult to identify through the implementation of standard statistical methods alone. The focus on topological properties helps to identify the underlying geometry of the datasets and to discover a set of unrelated features that may be causing the similarity in the spread of COVID-19 across the U.S.

INTRODUCTION

In this paper, the authors present novel machine-learning techniques that rely on graphs as the fundamental approach to structuring and analyzing complex real-world data. The authors analyze how the spread of COVID-19 has advanced in every county of the USA and look for similarities in a specific timeframe. Focusing on the geometric properties of datasets, we unveil a set of unrelated features that could have caused similarities in the pandemic spread.

Researchers worldwide have constructed multiple data models related to the spread of COVID-19; in the current experiment, however, the analysis focuses on the early stage of the pandemic. The objective is to identify why in some regions of the USA the pandemic spread faster after the first case was confirmed than in others which showed much slower patterns. Social, economic, demographic, political, and other factors which could influence similarities in pandemic spread are addressed.

The analysis was performed on a county-by-county basis, yielding a topological data model in graph form in which each of 3,142 nodes corresponds to one county, and two nodes are connected if they share similarities. The graph was built using Topological Data Analysis (TDA) – a novel approach to building a visual representation of a complex dataset. The dataset incorporated a variety of outcomes corresponding to the number of confirmed cases and deaths in each county over a specified time interval.

The key focus of the experiment is to investigate the spread of the pandemic since its outset. Thus, the first confirmed and recorded case of COVID-19 in each county was taken as the starting point of the observation interval. Given the speed with which the pandemic has advanced, it was critical not only to select an appropriate starting point but also to limit the analysis by carefully selecting the endpoint of the time interval to make it relevant without overloading the model. The jump in the number of cases was significant, going, for example, from 451 to 330,384 between March 8 and April 5; therefore, the observed time interval was set at 39 days.

After the graph was built, real-world data were used to integrate into the model any predictors which might be responsible for similarities in the early-stage spread of the pandemic. Over 250 predictors from different publicly available sources were used during the course of the experiment. Further, we performed a statistical analysis of discovered patterns to explain similarities in the spread of the pandemic based on the predictors integrated into the model. At any time, additional predictors can be added into the model to expand the search of unrelated features that might be responsible for similarities in pandemic spread.

1. DESIGN OF THE EXPERIMENT

This research relied on graphs as the fundamental approach to structuring and analyzing complex data. It explored the advance of COVID-19 in every county of the United States, looking for similarities which have occurred since the beginning of the pandemic. Focusing on the geometric properties of datasets, the experiment aimed to unveil a set of unrelated features that could have caused similarities in the spread of the pandemic.

Our experimental model is based on real-world data derived from a variety of publicly available sources (see Figure 1). These data correspond with the outcomes of interest, which are further analyzed to generate meaningful insights. Taking a much broader perspective, novel machine learning (ML) algorithms coupled with advanced analytics are used to extract meaningful patterns of information that will help describe the similarities in how the COVID-19 pandemic has spread across the United States since the first confirmed cases. Multiple modern approaches can be applied to real-world data with a focus on finding previously unseen or unknown patterns. ML and other advanced methods allow the data themselves to reveal patterns and relationships that provide new insights.

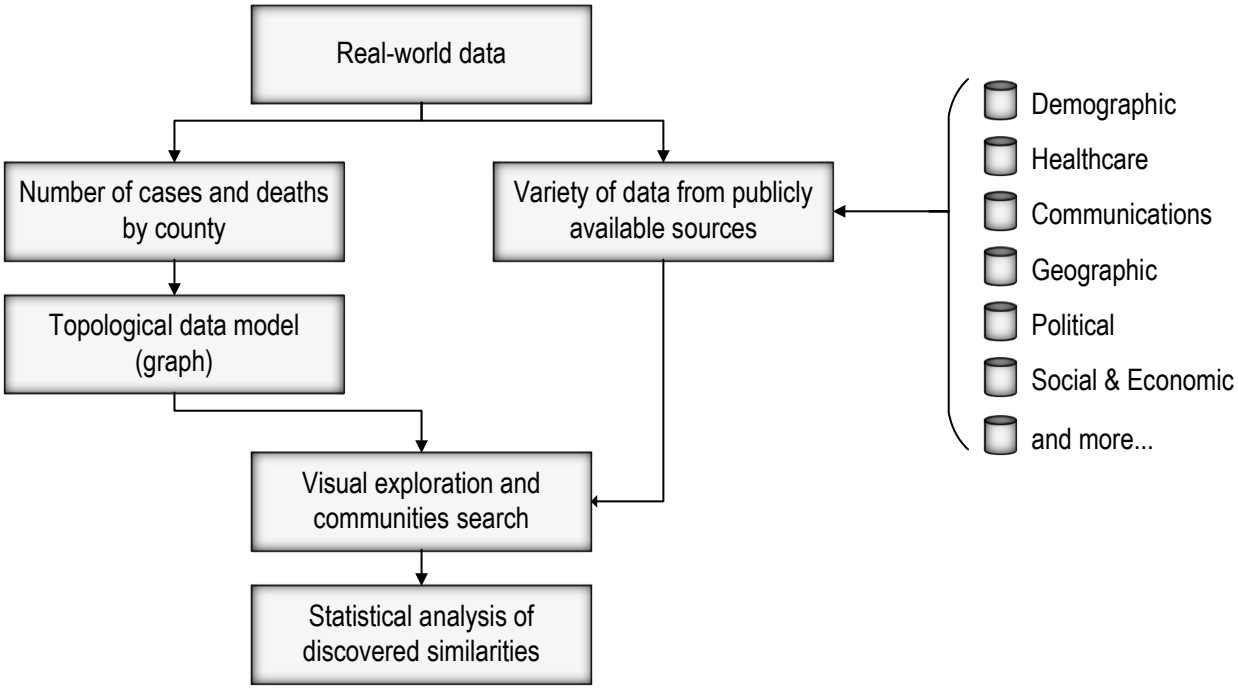


Figure 1. Real-world data used in the experiment

Years of research led by our group of data scientists have enhanced our expertise in Topological Data Analysis (TDA), the modern approach on which we rely to discover hidden patterns in large and complex datasets. Besides traditional analytics, such as geometric, statistical, and data-driven algorithms, TDA employs a wide range of ML techniques, both supervised and unsupervised by humans. These algorithms are called unsupervised ML because, unlike supervised learning, there is no correct answer. Algorithms are left to discover and present hidden insights in the data by themselves. Utilizing unsupervised ML allows researchers to extract comprehensive topological data maps represented by graphs without first having to develop a hypothesis. This is extremely important for real-world analytics when expected results may not be obvious, considering the volume and variety of data that need to be processed.

1.1. EXPERIMENT WORKFLOW

The goal of the experiment is to find similarities in the spread of COVID-19 since the beginning of the pandemic or, in other words, since the first confirmed cases were reported in the United States. The analysis is performed on a county-by-county basis, which will yield a topological data model in graph form in which every node corresponds to one county (total 3,142 nodes), and two nodes are connected if they share similarities. The graph will be built using TDA based on the number of **outcomes** corresponding to the number of confirmed cases and deaths in each county over a specified time interval.

After the graph is built, real-world data will be used to integrate into the model any **predictors** which may be responsible for similarities in the early-stage spread of the pandemic. Over 250 predictors from different publicly available sources are used during the course of the experiment.

The key element of the analysis is the visual exploration of the graph. During this step, the researcher investigates the geometric properties of the graph, looking for anomalies such as a high concentration of nodes or interesting shapes that may indicate similarities of patterns in data. Every subgroup of nodes discovered in this way is then further statistically analyzed to find the possible reasons for the geometric anomalies.

Researchers worldwide have constructed multiple data models related to the spread of COVID-19; in the current experiment, however, the analysis focuses on the early stage of the pandemic. The objective is to identify why in some regions of the United States the pandemic spread faster after the first case was confirmed than in others, which showed much slower patterns. Social, economic, demographic, political, and other factors which could influence similarities in pandemic spread will be addressed. Table 1 illustrates the order of steps in the experimental set-up.

Table 1. Phases in the experimental set-up

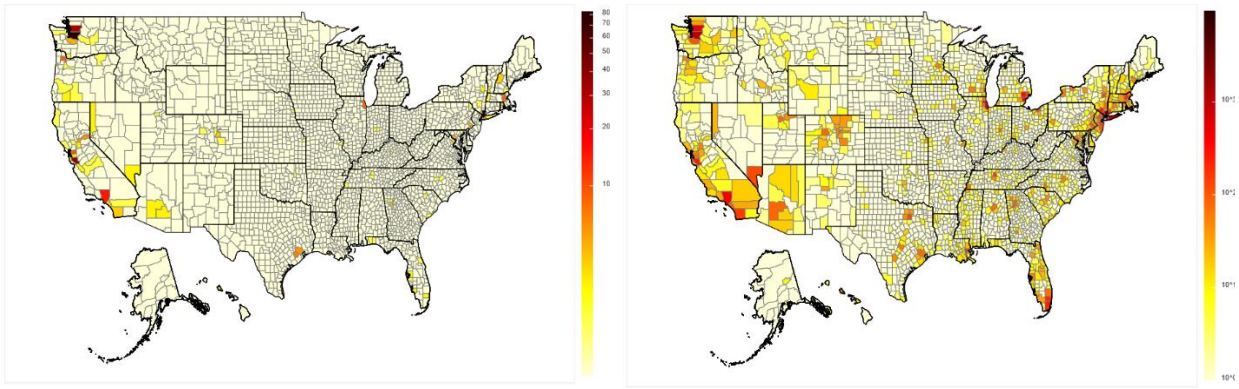
1) Define **outcomes**	Integrate the number of confirmed cases and deaths in every county of the United States over a specified timeframe since the beginning of the pandemic
2) Build the graph	Based on defined outcomes and using TDA, build a graph revealing the geometric properties of the dataset and hidden interdependencies
3) Integrate **predictors**	Integrate real-world data (from a variety of publicly-available sources) which may influence similarities in the spread of the pandemic. (The number of predictors can be expanded at any time after the model is built)
4) Perform a visual exploration	Perform a visual exploration of the graph and automatic search for communities, looking for anomalies such as a high concentration of nodes or interesting shapes
5) Interpretation of findings	Perform a statistical analysis of discovered patterns to explain similarities in the spread of the pandemic based on the predictors integrated into the model

1.2. SOURCES OF DATA USED TO DEFINE THE OUTCOMES

Since the outbreak of the COVID-19 pandemic, multiple organizations and authorities across the globe have continuously collected statistical data related to the occurrence and spread of the disease and, as a result, accumulated a large volume of complex real-world data. In this paper, we use data obtained from an open-source data repository available at Github and named *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* [1]. Collected data include aggregated data sources such as the World Health Organization, European Centre for Disease Prevention and Control, US Centers for Disease Control and Prevention, etc. In this dataset, data related to US statistics are represented at state or county/city level, including data from public health authorities all over the US states and counties, while non-US data sources are aggregated at country/region or province level. The current research focuses on data which correspond to the location and number of confirmed COVID-19 cases and deaths in all affected counties of the US.

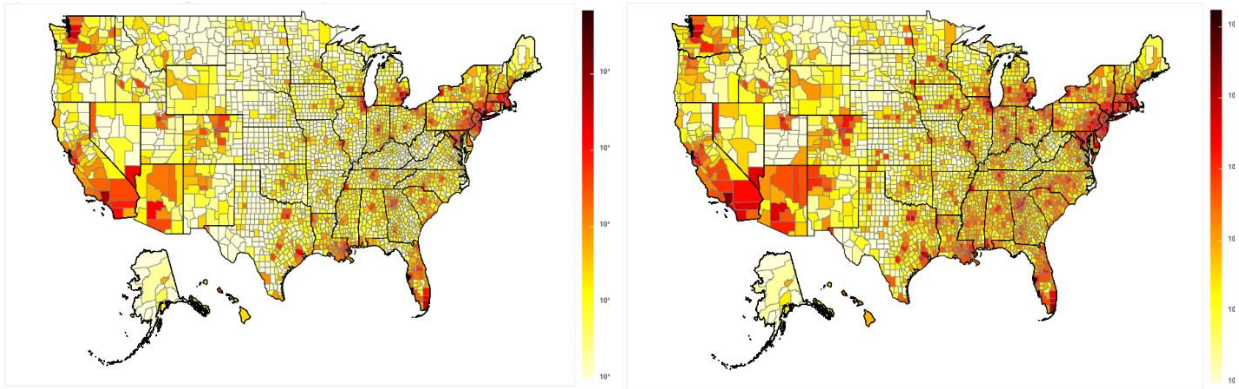
The data used to set up the outcomes for the model were collected at the level of counties (administrative or political subdivisions of a state in the United States) and county equivalents (other functionally equivalent subdivisions under US jurisdiction). In total, the resulting dataset includes data related to 3,142 counties and county equivalents. The selected dataset includes the number of confirmed COVID-19 cases and the number of deaths reported by each county.

Data collection started on January 22, 2020, when the first confirmed case of COVID-19 was reported in King County, Washington State, and has since been continuously undertaken. Figure 2 illustrates the rapidity of the spread of the pandemic across the United States during a short period. There were only 451 confirmed cases on the 47th day (March 8) in 92 counties. Just 14 days later, on March 22, there were 32,899 cases, with 1,136 counties affected. Snapshots c) and d) illustrate how fast the pandemic advanced thereafter, reaching 330,384 and 1,151,933 cases on April 5 and May 3, respectively.



a)
The number of confirmed cases
on March 8, 2020: 451 cases in 92 counties

b)
The number of confirmed cases
on March 22, 2020: 32,899 cases in 1,136 counties



c)
The number of confirmed cases
on April 5, 2020: 330,384 cases in 2,406 counties

d)
The number of confirmed cases
on May 3, 2020: 1,151,933 cases in 2,831 counties

Figure 2. The spread of the pandemic across counties

1.3. SELECTION OF OBSERVATION TIME INTERVAL

The key focus of the experiment was to target the spread of the pandemic since its outset. Thus, the first confirmed and recorded case of COVID-19 in each county was taken as the starting point of the observation interval. Many counties recorded no new cases for several days after the first case was reported, so the day the second confirmed case was recorded was taken as the start day to reduce the gap in the number of days for which no new data were collected.

In order to implement ML algorithms for time series, it was necessary to select a time interval of the same fixed length for every county, as this would allow the construction of a graph using TDA based on an n-dimensional vector

representing each county over a specific time interval. Given the speed with which the pandemic has advanced, it was critical not only to select an appropriate starting point but also to limit the analysis by carefully selecting the endpoint of the time interval to make it relevant without overloading the model.

Governmental and public health authorities across the United States implemented a stay-at-home policy to prevent the spread of COVID-19. During the stay-at-home period, typical measures included shuttering non-essential business operations and requiring people to stay home unless performing essential activities or attending medical facilities. The stay-at-home restriction was adopted in most states. The experiment aimed to analyze how COVID-19 spread in every county at the beginning of the outbreak and during the stay-at-home period when contacts among people were limited in most counties and conditions for spreading the disease in different counties were as similar as possible.

For our model, the first day after the stay-at-home restrictions were lifted for each county was set as a maximum limit of our observation time interval. In other words, for each county, the observation time interval was determined to lie within **Start_day** – the second confirmed case in the county, and **End_day** – the day of release from the stay-at-home order based on statewide orders [2]. The jump in the number of cases was significant, going, for example, from 451 to 330,384 between March 8 and April 5, with an additional 821,549 cases between April 5 and May 3 (see Figure 2); therefore, our time interval was set at 39 days so as not to overload the model. Figure 3 below illustrates the selection of the start day and end day of the observation interval, whereby the start day was selected as the day of the second confirmed case and the end day was selected as the release day of the stay-at-home period.

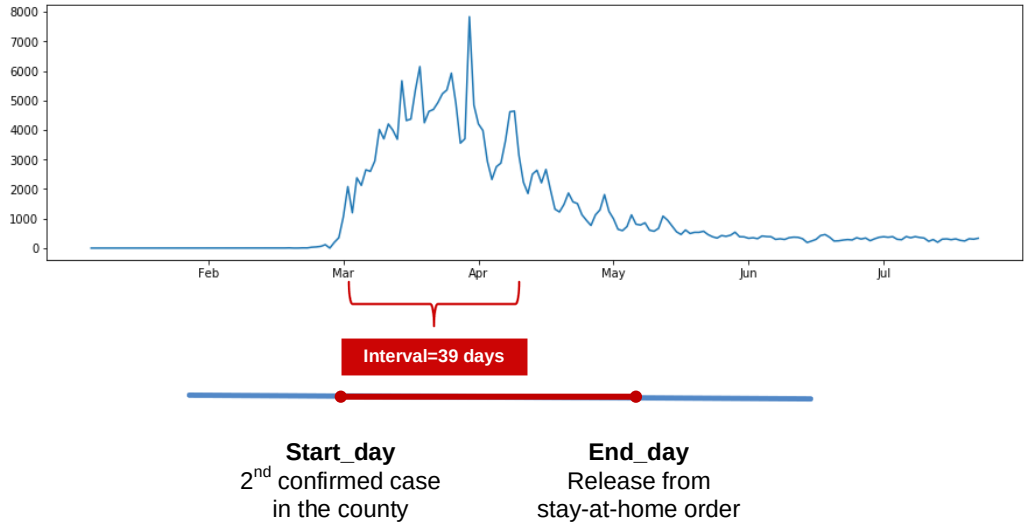


Figure 3. Selection of time series observation interval

In the scenario when more than 39 days elapse between start_day and end_day in a specific county, the observation interval was limited to the 39th day after the second reported case. In the opposite scenario when there were not enough data between start_day and end_day to cover a 39-day interval, the empty days were filled with zeros while the observation interval was expanded by an additional 5 days. It was assumed that social activities in the first five days after the release day from stay-at-home order did not affect the number of confirmed cases because of the disease incubation period and the delay in the availability of COVID-19 testing results. Figure 4 illustrates the selection of the start_day and end_day of the observation interval. The first diagram shows the observation interval within the 39-day limit, while the second demonstrates the scenario in which not enough data points are available to cover the 39-day time series.

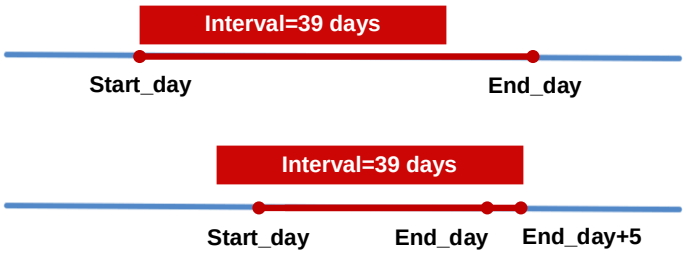


Figure 4. Selection of the start_day and end_day of the observation interval

This 39-day observation interval was used to structure a table of outcomes for the model. The n-dimensional vector built for every county combined 13 data points of confirmed cases and 13 data points of deaths, with data aggregated every three days. This vector, with the additional transformation described in Section 2, was further processed by the computational platform to build a graph using TDA.

1.4. ANALYSIS OF EPIDEMIC CURVES

Epidemic curves are an important component of public health, especially during a pandemic outbreak, allowing epidemiologists to represent the onset of disease cases over time visually. An epidemic curve is a histogram or plot illustrating the onset and progression of an outbreak of infectious disease in a particular population over a specific time [3]. The time interval is displayed on the X-axis and case numbers are shown on the Y-axis. This visual representation provides useful information on the size, pattern of spread, time trend, and exposure period of the outbreak.

For the purpose of the current experiment, epidemic curves were built for every county based on the number of confirmed cases. The resulting epidemic curves are different in terms of a shape (e.g., the peak of confirmed cases differs across counties, falling at the beginning, middle, or end of the curve) and of total number of confirmed cases of the disease reported during the observation period. To incorporate all counties into the model during the pre-processing stage, the data were first normalized using the standard score (also known as the z-score). The normalized epidemic curves were used to compare different counties in terms of the varying number of confirmed cases within the observation interval.

After normalization, the epidemic curves were grouped according to the similarity of shape of the epidemic curves. The k-means clusterization method was selected as a tool for grouping the epidemic curves into clusters. The k-means method starts with the selection of the number of groups (i.e., clusters) into which the epidemic curves should be grouped. In the current experiment, the number of groups was set to five types of epidemic curves. When performing the clusterization, the k-means method enables calculation of the center of a cluster, which is then used to determine the type of shape of the epidemic curves within the cluster. Next, all counties were grouped according to similarity of epidemic curves and assigned to one of five types to illustrate similarity in terms of the pattern of disease spread (see Figure 5).

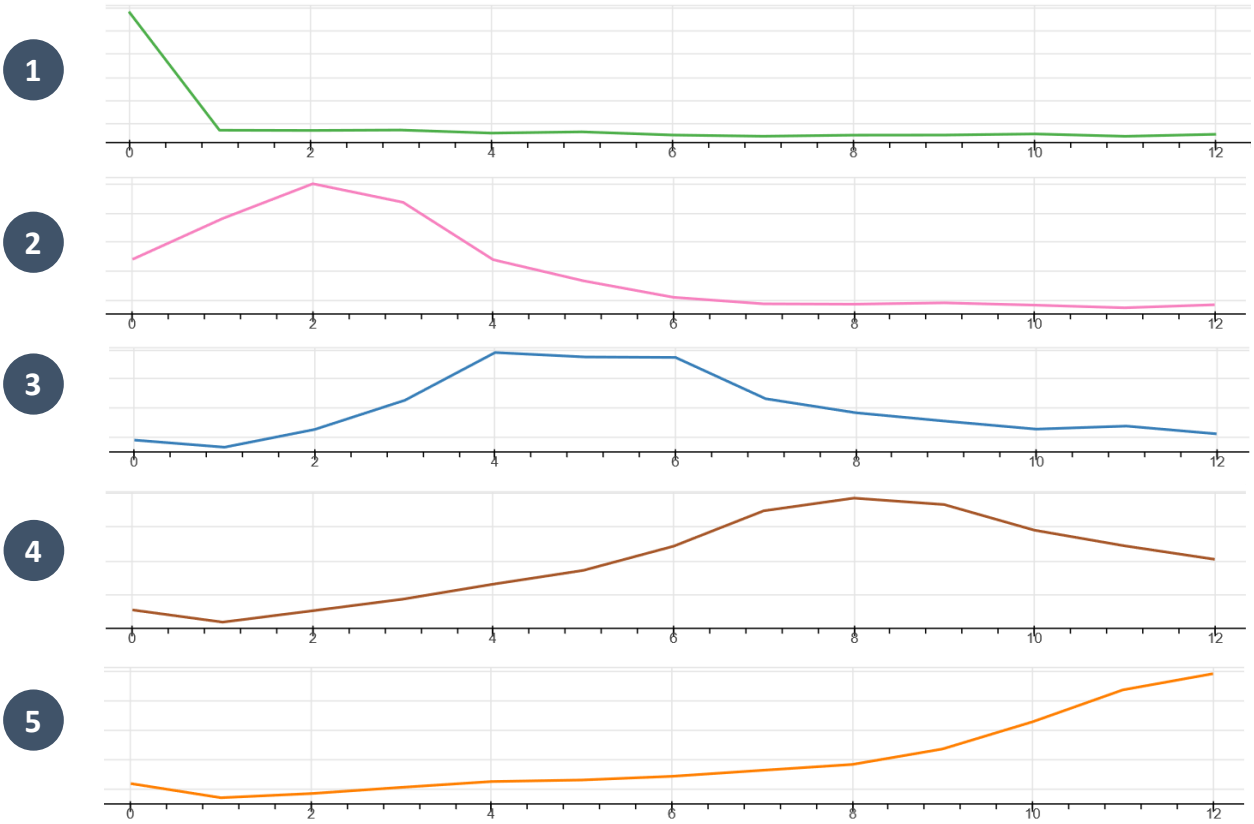


Figure 5. Five groups of epidemic curves built according to the number of confirmed cases

For illustrative purposes, we assigned a specific color to the epidemic curve of each county, using five determined types. By combining administrative and geographical maps of the United States (see Figure 6), it was possible to visually identify the similarity of pandemic spread based on geographical features. For example, counties located in mountainous areas, which probably indicates a lower population density, have a similar pattern: a peak at the beginning and a flattening of the curve thereafter (type 2 epidemic curve; colored pink). This visual approach, combined with the additional analysis of a variety of predictors (social, demographic, political, etc.), provides a much broader picture of the reasons behind the spread of the pandemic. The same approach was used to analyze the number of deaths integrated into the model for each county.

2. BUILDING THE MODEL USING TOPOLOGICAL DATA ANALYSIS

TDA is a novel approach to building a visual representation of a complex dataset. This analysis allows the extraction of comprehensive graphs from a dataset to provide a compressed graphical representation of a multidimensional set of interrelated outcomes. When applied to clinical data, this graph consists of nodes corresponding to patients participating in a clinical study and edges connecting those who share similarities. The same approach was used for the current experiment: TDA was applied to build a graph in which every node corresponds to a unique county while two counties with a similar pandemic spread pattern are connected by an edge. This section focuses on the concept of the geometric properties of a dataset to understand how graphs can be extracted from complex datasets and further modern ML algorithms and visual exploration techniques can be used to detect subgroups of counties that share similarities.

2.1. TOPOLOGY AND DATA MINING

Topology is a field of mathematics that deals with the properties of objects that remain invariant under continuous deformation. Imagine a surface that is made of a very thin and elastic material. The surface can be bent, stretched, or crumpled in any way; however, it cannot be torn and its parts cannot be glued together. As the surface is deformed, it changes in many ways, but some properties remain the same. The idea underpinning topology is that some geometric properties depend not on the exact shape of an object but, rather, on how its parts are combined.

As a simple example, consider geometric figures on the plane representing the numerical digits 0, 1, 2, … 9. For a topologist, various representations of the digit 0 are equivalent since they can all be continuously transformed into each other without cutting or gluing (see Figure 7 a-d). It is possible to change the size, thickness, or slope of the digit 0 through continuous deformation; however, one property remains invariant: the object separates the plane into two regions, namely the interior and the exterior. At the same time, 0 is not topologically equivalent to 1 or 8: 1 does not encircle a region and 8 contains two holes (see Figure 7e). The topological classification of the digits 0, 1, 2, … 9 results in the following five classes:

$$\{0\}, \{1, 2, 3, 5, 7\}, \{4\}, \{6, 9\}, \{8\}.$$

The digits in any of the classes are topologically identical, but no two digits taken from distinct classes are equivalent from the topological point of view.

The number of holes in a geometric object is a basic topological property. Another significant property is connectedness. Intuitively, an object is connected if it consists of a single piece. For example, the curve representing 0 is connected; if any two points are removed from it, it will become disconnected. Pieces of a disconnected object that are themselves connected are referred to as connected components. In the mathematical study of topology, all of these intuitive concepts are examined on a rigorous basis and generalized to higher dimensions.

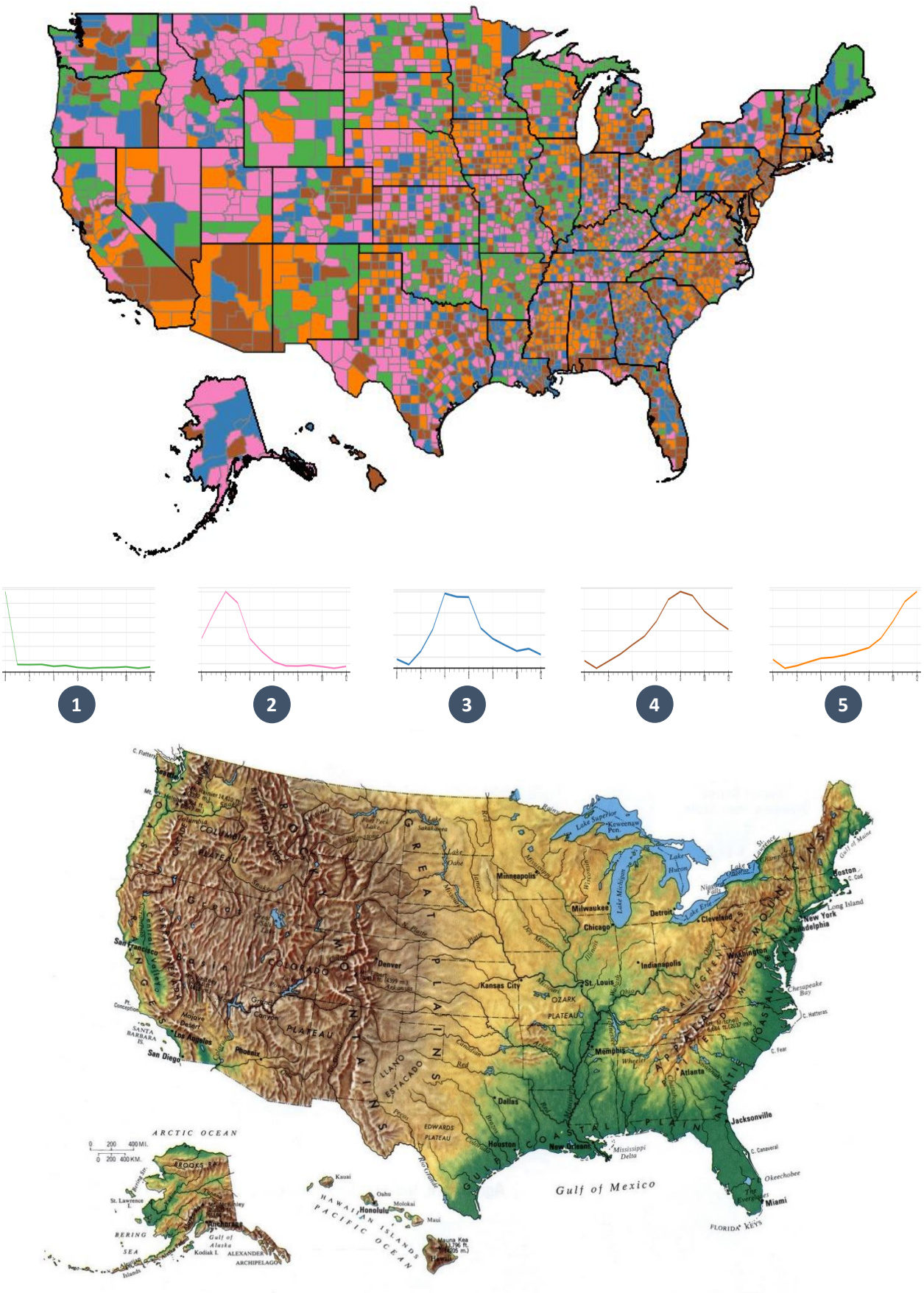


Figure 6. Five types of epidemic curves on administrative and geographical maps

a) b) c) d) e)

Figure 7. Different representations of the digit 0 (a-d) are topologically equivalent. All share a common topological property: they divide the plane into an interior region and an exterior region. The digit e) is not equivalent to 0 since it encloses two internal regions

Topology deals with abstract mathematical entities, such as curves and surfaces, that consist of an infinite number of points. In practice, however, all datasets are necessarily finite. Recently, a new field has emerged at the crossroads of topology and data science. TDA aims to extract topological data, that is, qualitative information, from finite sets of data points. It involves exploring datasets (viewed as finite clouds of points in multidimensional space) at multiple scales or resolutions, from fine- to coarse-grained. Given a complex dataset, TDA can be used to extrapolate the underlying topology and build a compressed yet comprehensive topological summary of the dataset. TDA exploits various methods and algorithms stemming from computational topology and geometry, statistics, and data mining. For detailed expositions of the mathematical theories that underpin TDA and certain applications in biology, see [4-6] and the references therein.

2.2. UNDERSTANDING CLINICAL DATA USING TOPOLOGY

Topology was originally developed to distinguish between the qualitative properties of geometric objects. It can be used in conjunction with the usual data-analytic tools for the following tasks:

1) **Characterization and classification.** Topological features succinctly express qualitative characteristics. In particular, the number of connected components of an object is of importance for classification.
2) **Integration and simplification.** Topology is focused on global properties. From the topological perspective, a straight line and a circle are locally indistinguishable; however, they are not equivalent if they are considered as a whole. Topology offers a toolbox to integrate local information about an object into a global summary. Thus, topology can provide the researcher with a natural "big-picture" view of complex, multidimensional data.
3) **Features extraction.** Topological properties are stable. The number of components or holes is likely to persist under small perturbations or measurement errors. This is essential in data mining applications because real data are always noisy.

In the context of clinical research, the dataset under study is typically a table of outcomes in a particular clinical trial. The table rows correspond to individual participants in the clinical trial, and the columns contain information on specific outcome measures of interest, such as lab tests, vitals, questionnaires, etc. Given a table of clinical outcomes, two types of parameters are required to generate a graph using TDA. The first of these is a projection, a function that is used to stratify patients into subpopulations (bins). The second is a distance function, which measures the proximity between patients. The distance function makes it possible to connect the related patients with similar outcomes within each population.

To be considered for further analysis, a graph extracted from the dataset using TDA algorithms should meet certain requirements. Namely, it should:

- accurately represent the original dataset;
- eliminate the features of the dataset that are not relevant to the purpose of the study;
- reduce the complexity of the features that are shown on the data map; and
- be insensitive to small noise, such as errors of measurement.

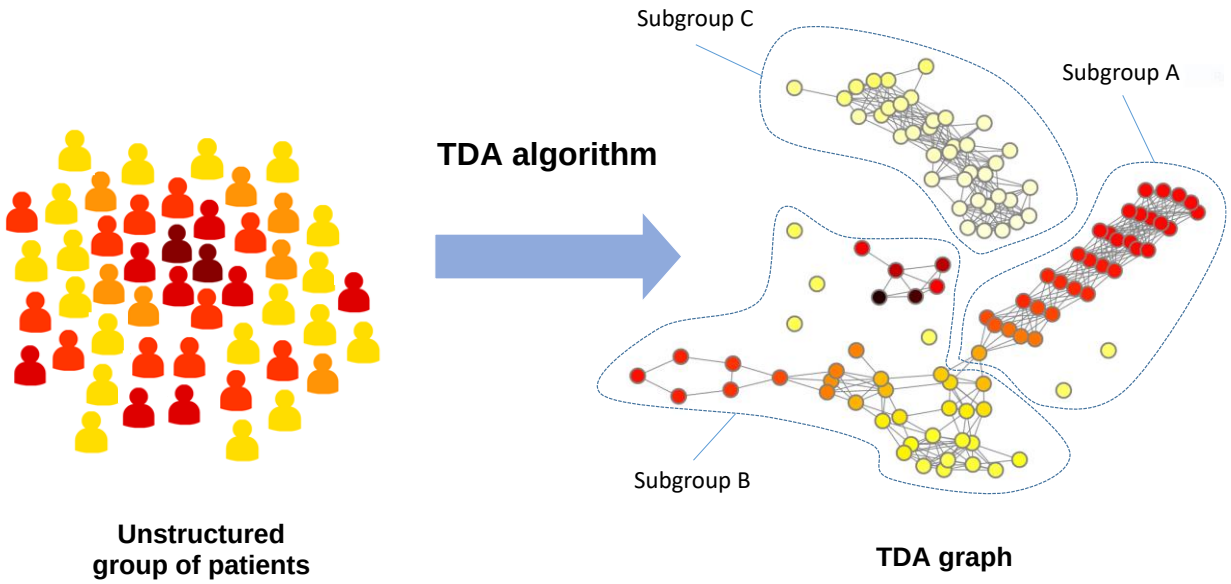


Figure 8. Discovery of multivariate patterns in clinical trial outcomes
A graph represents groups of patients structured according to the similarity of clinical outcomes

The core idea of clinical data mining using TDA relies on the visual discovery of subgroups of related patients in a graph (see Figure 8) that presents the relevant information about the dataset in a compact and efficient manner. For the clinical dataset, the following criteria have to be met to perform the analysis:

- **Each node represents a patient** – a graph extracted from a clinical dataset is actually a graphical representation of the dataset in which each node represents an individual (trial subject).
- **Similar nodes are connected** – two nodes representing similar patients (in terms of a predefined set of clinical outcomes) are connected with an edge.
- **Coloring focused on specific outcomes** – the color of the nodes helps to highlight emerging patterns in the data and identify subgroups of patients related to the distribution of a variable of interest.
- **Visual discovery of subgroups** – clusters or "communities" of nodes on a graph reflect a segmentation of patients that may indicate robust patterns within the data.

The same approach was used in the current experiment, for which the dataset was built using real-world data obtained from publicly available sources. Each row of the resulting dataset corresponds to one out of 3,142 counties. At the same time, columns represent the outcomes related to the numbers of confirmed cases and deaths within the observation interval (39 days). The core idea of the experiment is to build a graph that will highlight possible similarities in pandemic spread patterns using the geometric properties of data. Further, by integrating various predictors from different areas, such as demographic, social, geographic, etc., the revealed similarities can be described.

After a graph is constructed according to selected outcomes, the researcher then visually explores it to discover interesting subgroups within the data. For example, the isolated components of a graph or highly interlinked groups of nodes that form communities may indicate meaningful relationships within the dataset.

2.3. OUTCOMES SELECTED FOR THE MODEL

In the present experiment, TDA was used to build a graph that visually represents the geometric properties of the dataset. In other words, TDA allowed a comprehensive graph to be extracted from the original dataset to provide a compressed graphical representation of a multidimensional set of interrelated outcomes. The graph consists of 3,142 nodes corresponding to each county and edges connecting the counties that are similar in terms of outcomes.

As described above, the analysis was performed with respect to a 39-day observation interval. Thus, the outcomes selected in this study include 13 (39 / 3) normalized numbers of confirmed cases, the logarithm of the total number of confirmed cases, 13 (39 / 3) normalized numbers of cases of death, and the logarithms of the total number of deaths calculated for each county. Therefore, the total number of outcomes for each county is 13 + 1 + 13 + 1 and equals 28 (see Figure 9).

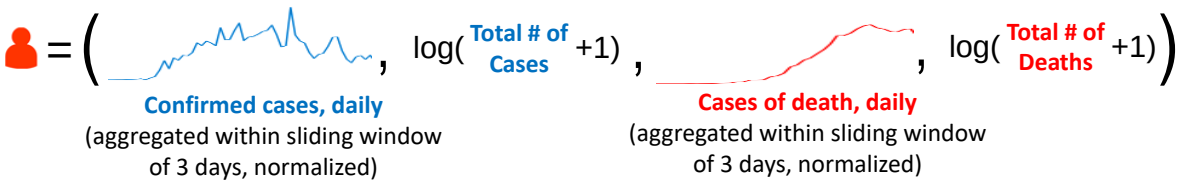


Figure 9. Outcomes selected for each county

Therefore, the relevant information about the spread of COVID-19 in each county is encoded by a 28-dimensional vector, as illustrated in Figure 10:

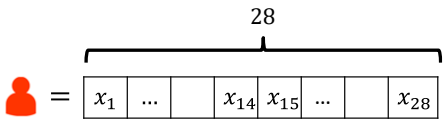
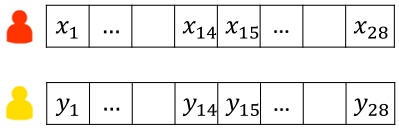


Figure 10. 28-dimensional vector constructed for each county

Encoding this information into this vector allowed investigation of the dataset based on four characteristics of the spread of COVID-19: the shape of epidemic curves built on confirmed cases, the total number of confirmed cases, the shape of the epidemic curve built on the number of deaths, and the total number of deaths.

To determine similarities between two counties while building a graph, it is necessary to calculate the distance between two 28-dimensional vectors. For that purpose, a distance function based on modified Euclidean distance was used to measure the similarity in the spread of the pandemic between the counties (see Figure 11).



$$d(\,,\,) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_{14} - y_{14})^2} + \sqrt{(x_{15} - y_{15})^2 + \ldots + (x_{28} - y_{28})^2}$$

Figure 11. The distance function measuring similarities between two counties

Thereafter, using the two-dimensional projection based on density and centrality of nodes with respect to selected outcomes, the resulting TDA graph was built.

2.4. DEFINE THE MOST REPRESENTATIVE GRAPH USING THE KOLMOGOROV COMPLEXITY

When applying the TDA algorithm to a complex dataset based on various parameters, the algorithm generates a large volume of graphs, including those that capture nonrelevant noise. The last step is to determine the most representative graph based on available parameters of the topological model (the number and placement of stratifying bins, parameters in density and centrality estimates, etc.). For this purpose, we used Kolmogorov complexity.

Kolmogorov complexity originated in theoretical computer science and is a measure of information contained in a string or an arbitrary array of letters and digits. It measures how well an object (array, graph, text, etc.) can be compressed.

For a given string (array), the Kolmogorov complexity is determined as the smallest length of the compressed description of the string (array). For example, the string "11111111111111111111" can have the following compressed description: "1 × 20 times". However, in the string "10011011110101011110" the digits seem to be spread randomly and it is difficult to give a shorter description of the string than a direct list of the digits within it. Therefore, the Kolmogorov complexity of the second string is higher than that of the first string.

The measure of information contained in graphs can be similarly estimated by using Kolmogorov complexity. Specifically, when the nodes of a graph are numbered, it can be described by a graph adjacency matrix, in which "1" is placed at the intersection of row i and column j if the nodes numbered i and j are connected, and "0" is placed at the intersection of row i and column j if the nodes numbered i and j are not connected.

Graphs with a regular structure of edges have low Kolmogorov complexity, while graphs with a random structure of edges have high Kolmogorov complexity. Figure 12 shows a complete regular graph with 10 nodes that has a Kolmogorov complexity of 103.79, while Figure 13 shows a random graph with 10 nodes that has a Kolmogorov complexity of 437.61:

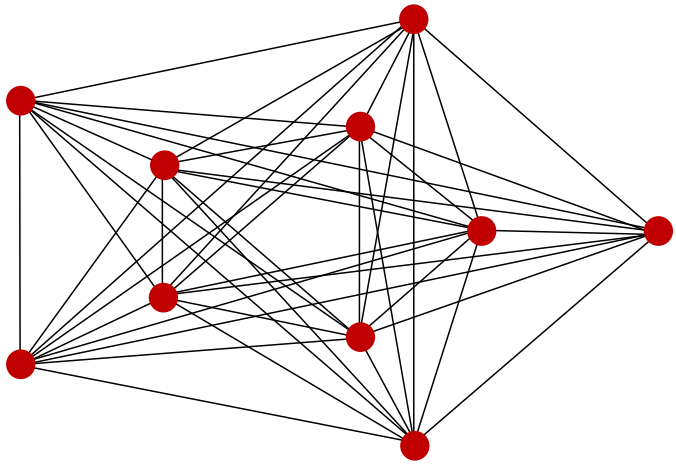


Figure 12. A regular graph having low Kolmogorov complexity (K.C. = 103.79)

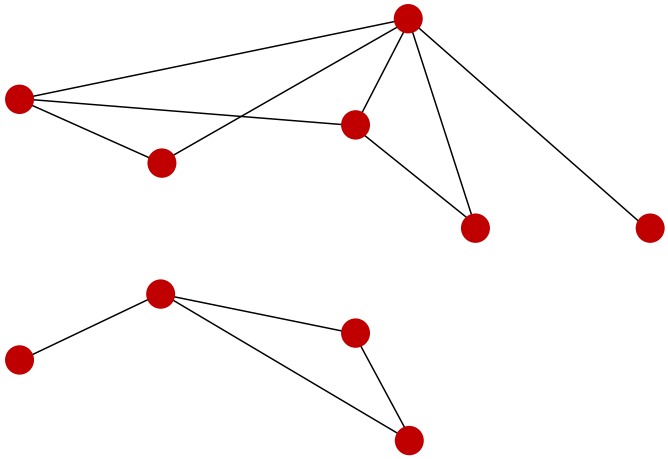


Figure 13. A random graph having high Kolmogorov complexity (K.C. = 437.61)

Although it has been proven that the Kolmogorov complexity cannot be calculated in a general case, there are various methods for estimating it. An example is the block decomposition method, which includes the decomposition of a string (array) into blocks having a limited length, estimation of the Kolmogorov complexity of each block, and summing up the estimates according to the information theory rules [7].

In the current experiment, when the Kolmogorov complexity of the graph was estimated, consideration was given to partitioning the nodes of the graph into stratifying bins and the number of nodes in each stratifying bin, as well as to partitioning the graph into separately connected subgraphs and the number of connected subgraphs. The Kolmogorov complexity was measured within each stratifying bin separately. It was assumed that high Kolmogorov complexity implies a random distribution of edges in subgraphs of the graph, which in turn makes the influence of the node-clustering algorithm less significant than the influence of the projection, the distance function, and the number and placement of stratifying bins.

Experiments have shown that in graphs with low Kolmogorov complexity, nodes within the same stratifying bin tend to combine into clusters with a high density of edges. Conversely, graphs with high Kolmogorov complexity have a more uniform distribution of edges; that is, the selected parameters of the projection in such graphs distribute the nodes over the stratifying bins more evenly. Thus, in graphs having high Kolmogorov complexity, the nodes within the stratifying bins tend to group into clusters of approximately equal size.

Upon measuring the Kolmogorov complexity within each stratifying bin, we determined graphs with high Kolmogorov complexity, that is, graphs in which subgraphs corresponding to the stratifying bins have high Kolmogorov complexity. Finally, the most representative graph was selected from this group (see Figure 14).

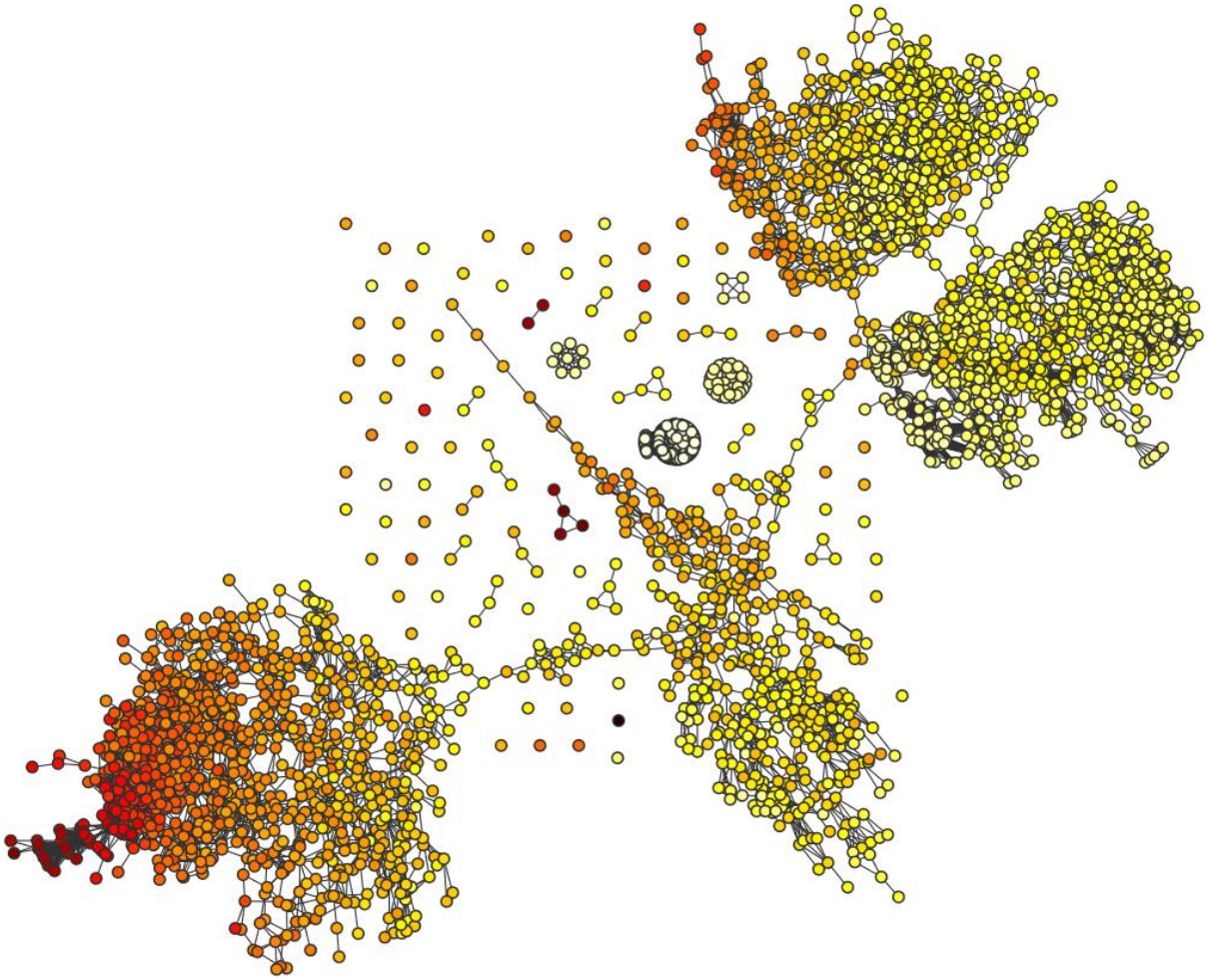


Figure 14. The topological model of COVID-19 spread over the United States

The most representative graph capturing key features of our dataset with selected outcomes will be further analyzed using ML algorithms, visual exploration, and statistical analysis of predictors. To summarize, here are a few essential points that describe this graph:

- Each node on the graph corresponds to one US county (total 3,142 nodes).
- Two nodes are connected with an edge if they are similar in terms of predefined outcomes.
- Outcomes integrate the number of confirmed cases and deaths within a 39-day observation interval from the start of the pandemic.
- The similarity in pandemic spread is measured by a modified Euclidean distance function.
- The resulting graph is selected based on estimates for its Kolmogorov complexity.

2.5. AUTOMATIC COMMUNITIES DETECTION

The key feature of a graph is a community structure, which relates to the way the nodes are organized in communities. Specifically, many edges connect nodes within the same community (or cluster), while comparably few edges connect nodes between different communities [8]. These clusters or communities can be considered to represent independent structures within the graph, and the detection of those independent communities is one of the key goals in the analysis of large graphs that represent complex relationships within datasets.

In graphs that represent real-world systems, the distribution of edges over subgroups of nodes is usually non-uniform. This reflects the possible presence of some hidden structures and patterns in the graph, and hence in the real-world data from which the graph was created. Specifically, some groups of nodes may have high concentrations of edges, while the concentrations of edges between these groups of nodes may be low. This structure takes the form of an intermediate-scale graph structure known as a community structure, or a cluster structure, where a group of densely connected nodes is referred to as a community. Figure 15 illustrates an example of a community structure within a graph that contains three clusters of nodes with dense internal connections and comparably fewer connections between clusters.

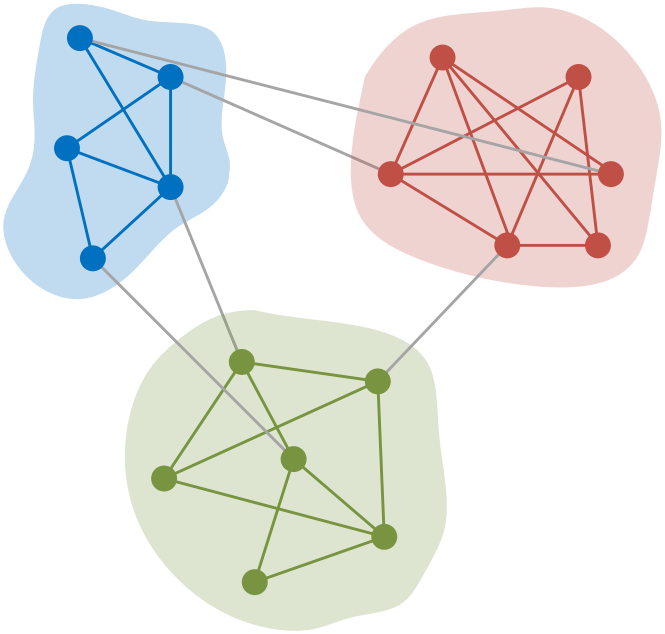


Figure 15. A schematic representation of a simple graph that has a community structure
The graph contains three communities of densely connected nodes that have a much lower density of connections (gray edges) between them

Communities, or clusters, are groups of nodes within a graph that are likely to share common properties and/or play similar roles. In view of this, where possible, the aim of community detection is to identify communities within the graph and their hierarchical organization by using the information that is contained within the graph topology alone. Identifying communities according to the topological properties of the graph only allows the classification of nodes according to their structural position on the graph. Thus, nodes with a central position in their communities share the largest number of edges with the other nodes within the community, which may indicate the important role they play in the stability of the community. On the other hand, nodes that are located at the boundaries between communities may play an important role as mediators in the relationships and exchange between different communities.

The problem of graph clustering, intuitive at first sight, is actually not well defined. Though numerous attempts have been made to analyze real-world systems based on the community structure in multiple disciplines and through practical applications, graph theory does not define the problem of graph clustering, and no universally accepted definitions of a community or partitioning into communities have emerged. Therefore, the concepts of a community and partitioning into communities are to some extent arbitrary and must be determined by researchers according to the specific problem under consideration.

Detecting communities within a graph (especially large ones) can be computationally difficult if the number of communities within the graph is unknown and the size and density of the communities are unequal. However, several ML algorithms have been developed and used for community search with varying degrees of success in recent years. The current research uses various community detection algorithms to unveil robust connections and hidden relationships in the dataset, based on the geometric properties of the graph.

2.6. PREDICTORS USED FOR STATISTICAL ANALYSIS

After a topological data model is extracted, the researcher visually explores the graph to discover interesting subgroups within the data. These subgroups can be further studied by utilizing standard statistical methods to determine the predictors that may be responsible for the similarity of the pandemic spread observed within the identified subgroup of counties.

A very common situation in statistics occurs when the distribution of an outcome (or response variable) is related to one or several predictors (or explanatory variables). A standard approach used by researchers to study the relationship between a predictor and an outcome is the application of a suitable statistical model. The model selection depends on the data types of the predictor and outcome (quantitative, binary, categorical, etc.) and often involves additional assumptions concerning the distribution of the outcome. To describe similarities in pandemic spread, we integrated over 250 predictors into our model, available from a variety of public sources (see Figure 16). At any time, additional predictors can be easily integrated into the analysis without any changes in the topological model.

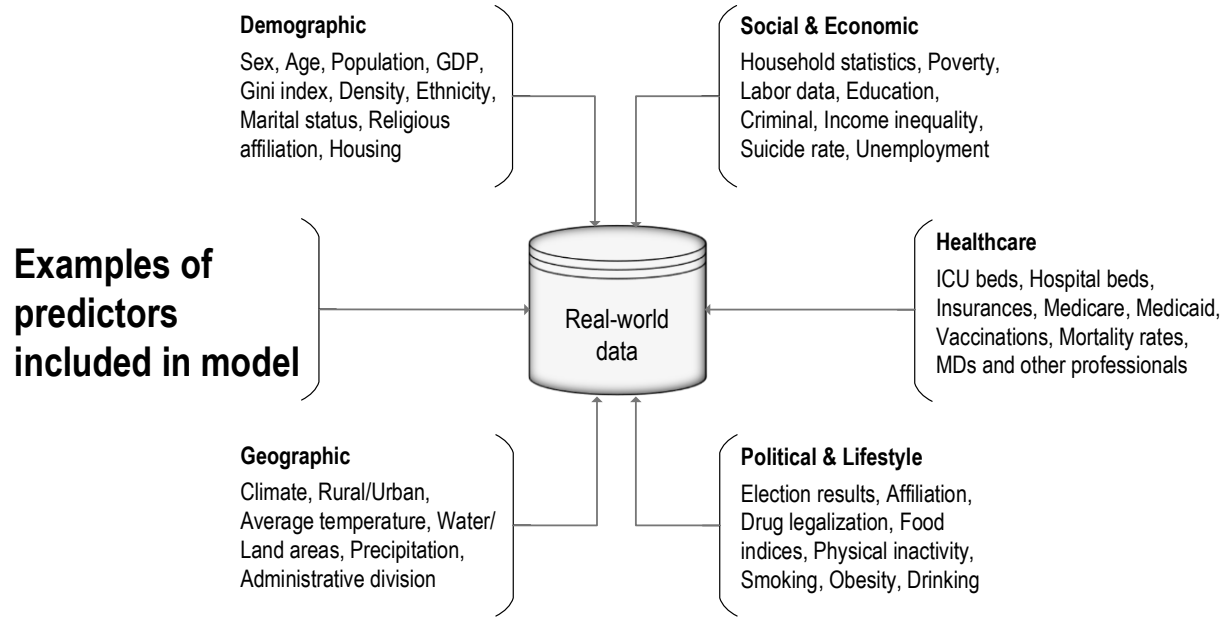


Figure 16. Group of predictors incorporated into the model.

The interactive visualization provides researchers with an opportunity to manually perform a visual inspection of a graph to identify regions of interest. For example, the nodes that form fork-like structures or loops might be of interest for further research. In addition, isolated components or highly concentrated groups of nodes that form communities may indicate meaningful relationships in the outcomes dataset. While performing a visual inspection, the researcher can also re-color the graph in accordance with the median value of an outcome or predictor selected from the corresponding datasets. The use of color codes may highlight how a subgroup of nodes represented by a given region of the graph might be different from the rest of the nodes.

The researcher can select any region of the data map that exhibits interesting geometric properties to perform further statistical analysis. After running statistical tests, a table of predictors with the corresponding p-values can be calculated to determine whether the distribution of the predictors for the selected subgroup of nodes is different from that of the rest of the nodes. If the desired significance level of any predictor is found to be statistically significant, the researcher can construct a histogram that represents normalized frequency distributions of the predictor for both the nodes in the selected region of the graph and the rest of the nodes. The same can be done when comparing two different selected regions of nodes with each other.

For the purpose of the statistical analyses undertaken for the present experiment, continuous, mixed, binary, and categorical (non-binary) univariate predictors were differentiated according to a variable type. Continuous predictors were examined using the standard two-sample Kolmogorov-Smirnov test. This method verifies whether two data samples are obtained from the same distribution. The test assumes that the underlying distributions are continuous (no ties in the variables' values are allowed). To examine the statistical association between two samples within the categorical data, Fisher's exact test and the χ^2 test were used for the binary and non-binary categorical variables, respectively. Figure 17 describes the variable type, data source, and other features of the predictors that were integrated into the model.

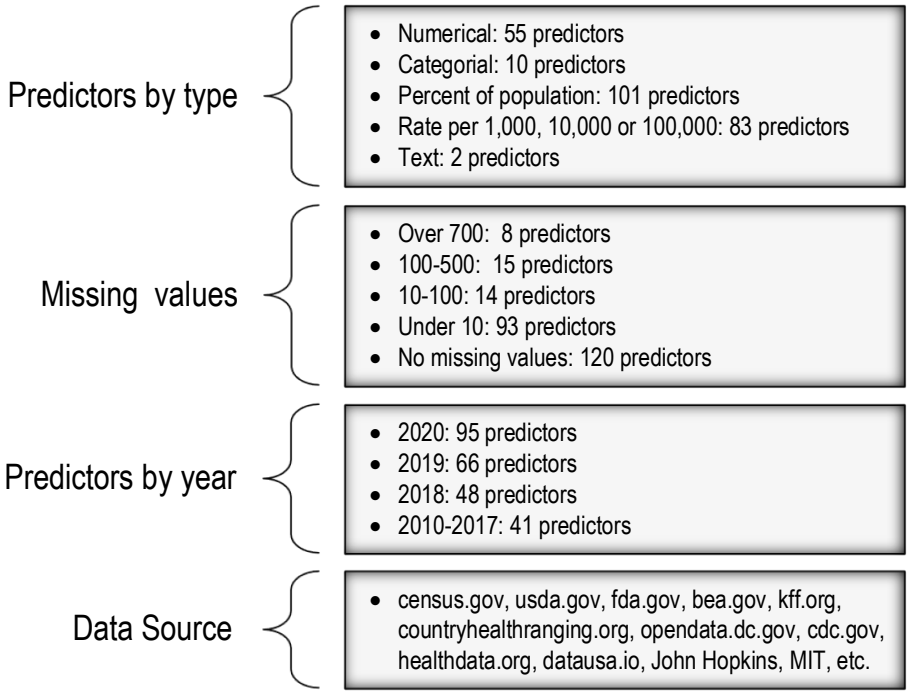


Figure 17. The characteristics of predictors integrated into the model

3. VISUAL EXPLORATION AND INTERPRETATION OF FINDINGS

This section investigates the geometric properties of the resulting graph, looking for anomalies such as a high concentration of nodes or interesting shapes that may indicate similarities of patterns in data. Every community of nodes discovered is then statistically analyzed to find possible reasons for the geometric anomalies.

3.1. FOUR MAJOR COMMUNITIES

Looking at the geometric properties of the graph extracted from the dataset that embeds the number of confirmed cases and deaths within a 39-day interval (see Figure 14), we can clearly distinguish four main regions. In other words, based on TDA algorithms and following the predefined parameters, the geometry of the dataset unveils four communities of counties that have more similarities (number of edges) with each other than with the rest of the counties. Figure 18 illustrates the four communities to which we refer.

The total number of counties within these four communities is 2,743 (out of 3,142), with the largest, the purple community located on the lower left, corresponding to 1,036 counties. Let's first consider the map of the USA and the distribution of these four communities. We colored the counties on the US administrative map according to the colors of the nodes corresponding to the counties in the detected communities (see Figure 19). An analysis of the map clearly indicates that counties colored red and green lie in regions with a lower population density. In contrast, purple counties are located in densely-populated areas.

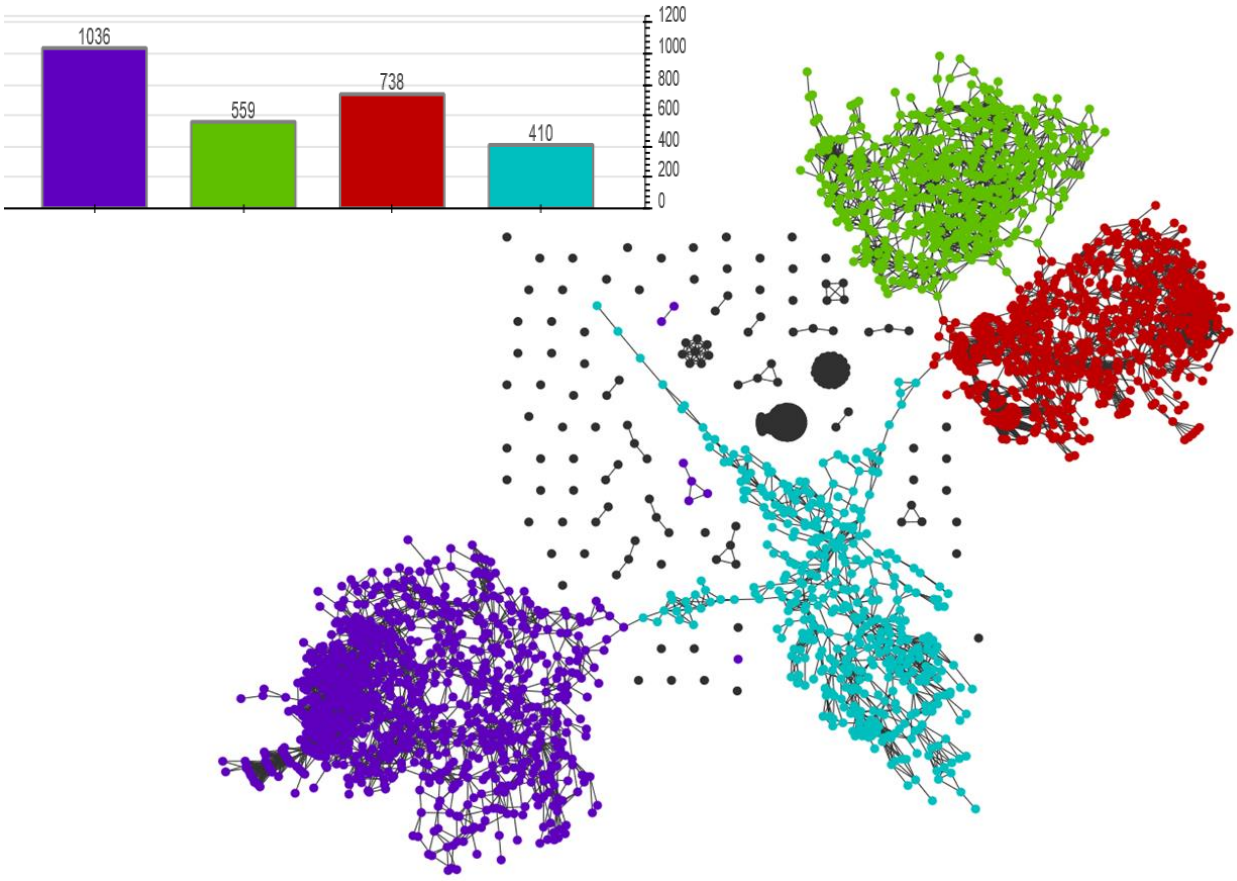


Figure 18. The four major communities detected on the graph

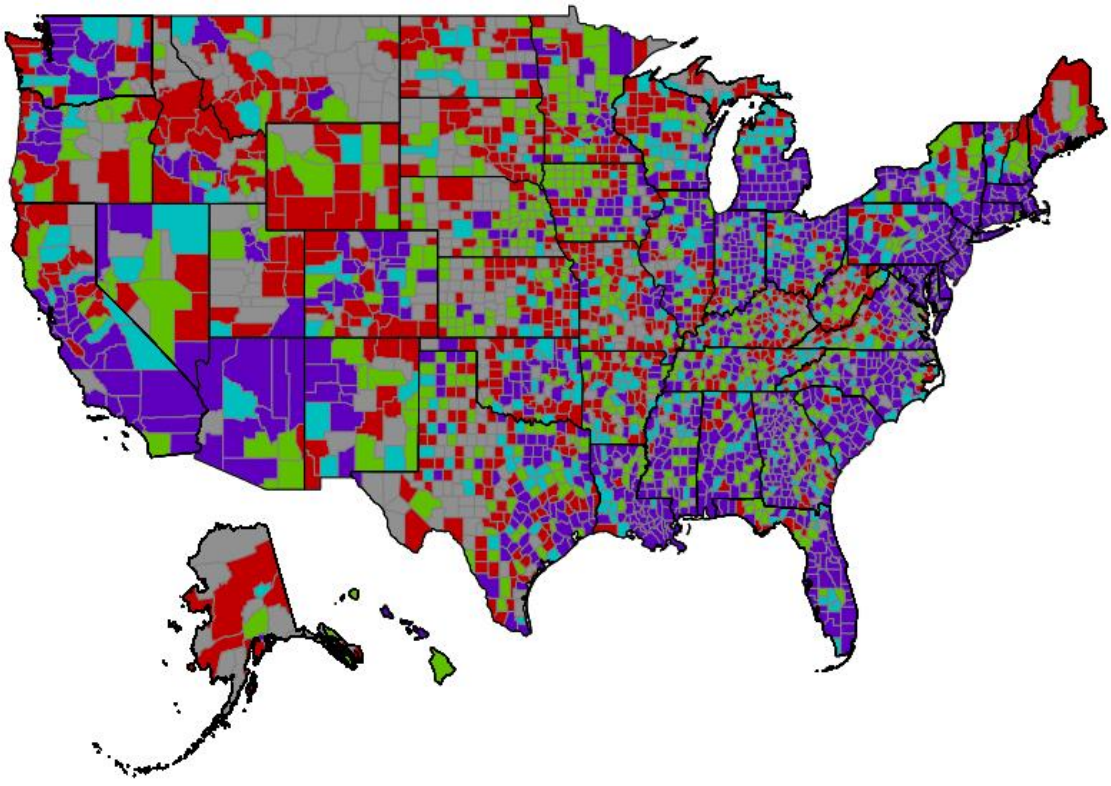


Figure 19. The US administrative map with the four communities detected on the graph

Let's next analyze how the four major communities differ in terms of pandemic spread patterns. To determine the significance of the type of onset and progression of an outbreak of the disease, the graph was re-colored according to the epidemic curves described earlier (see Figure 5). As we can see in Figure 20, the lower-left and upper-right communities are similar in terms of epi-curves. In these two communities, the nodes are mostly colored blue (3), brown (4), and orange (5), which corresponds to the delayed outbreak of the disease. The lower-right and central communities are mostly colored green (1) and pink (2), which corresponds to the disease outbreak at the beginning of the observation interval, with a further significant reduction of confirmed disease cases.

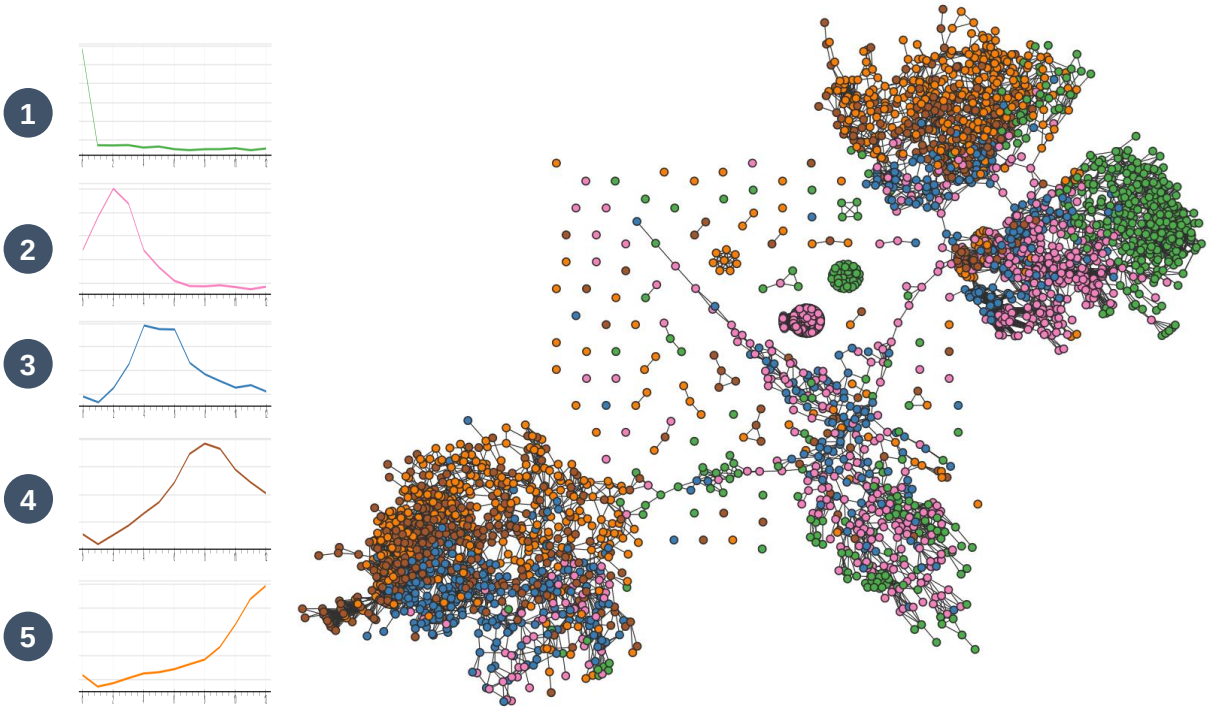


Figure 20. The graph colored according to the **epidemic curves** within the observation interval

To further determine specific features of these four communities in terms of the predefined outcomes, the graph was colored according to the number of confirmed cases within the 39-day observation interval (see Figure 21) and number of deaths (see Figure 22). The dark color indicates a higher number of cases, while the yellowish color indicates the opposite.

Looking at these two colorings, we can clearly see that a lower-left (circled in purple) and upper-right community (circled in green) have similar patterns in regard to the number of confirmed cases as compared with the other two communities (cyan in the center and red on the lower right). On the other hand, referring to Figure 22, these two communities, purple and green, have different patterns in the number of deaths: the green community, on the upper right, has far fewer deaths. In other words, when the pandemic started, a similar pattern of spread was seen in both purple and green communities. At the same time, the number of deaths indicates that counties located in the green community handled the virus more effectively, which substantially reduced the number of fatalities. Even though there are almost twice as many counties in the purple than in the green community (1,036 versus 559), this is still important for further analysis.

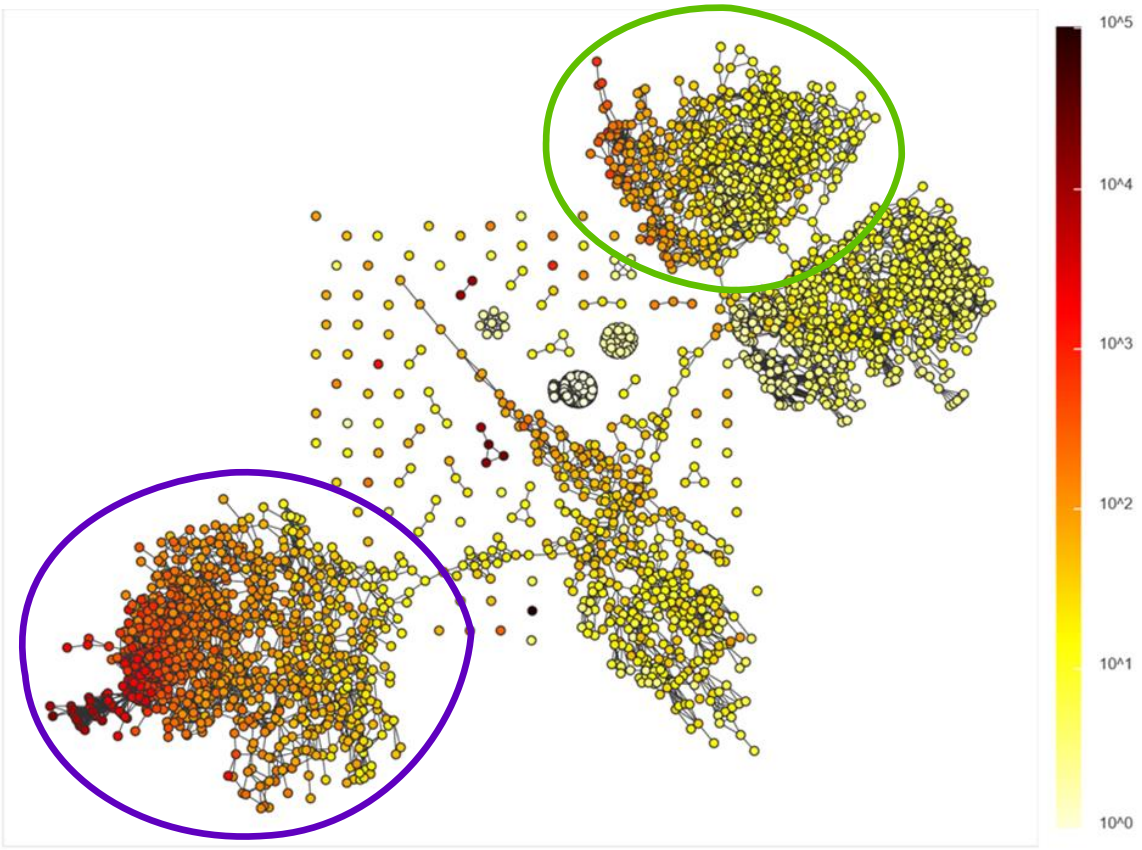


Figure 21. The graph colored according to the **number of confirmed cases** within the observation interval

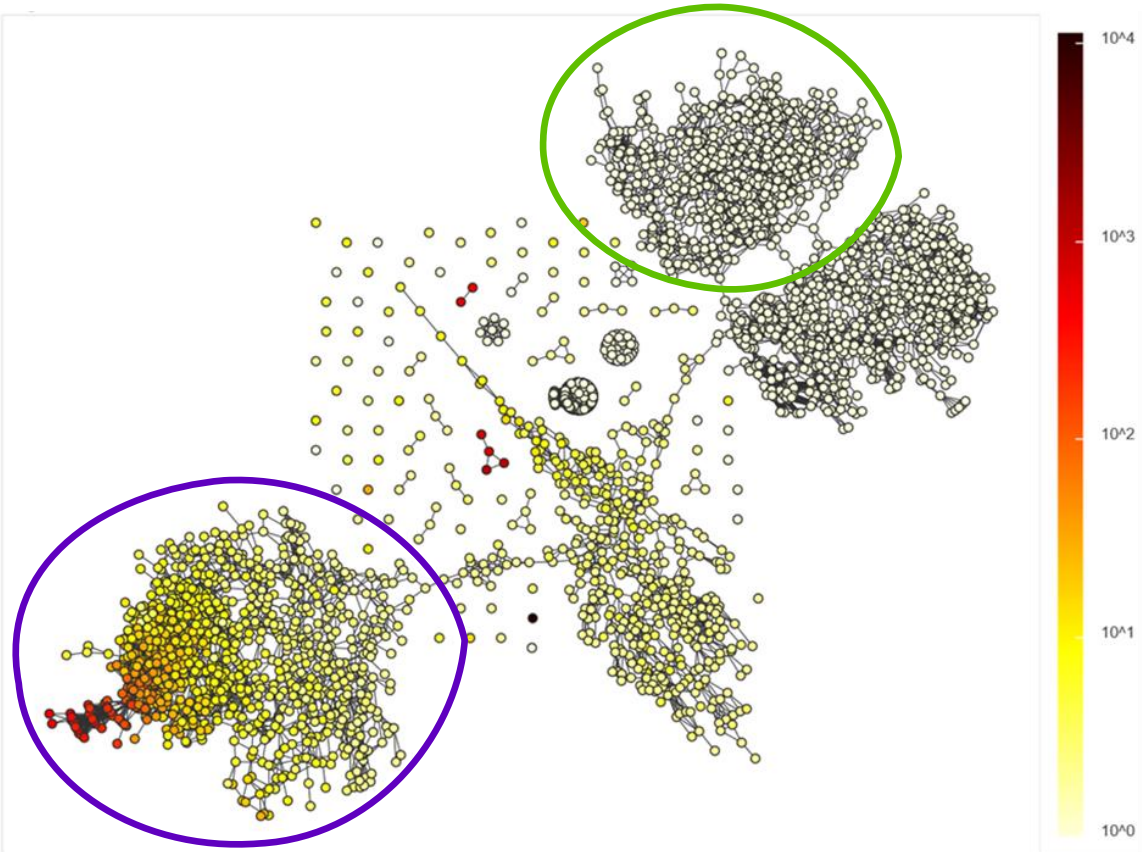


Figure 22. The graph colored according to the **number of deaths** within the observation interval

These two communities, purple and green, will become of interest for further analysis to determine **why the number of deaths was significantly different** while at the same time the spread of the disease showed a clear similarity, based on the shapes of epi-curves and the number of confirmed cases. For this purpose, we perform a statistical analysis of discovered patterns to explain similarities in the spread of the pandemic based on the predictors integrated into the model.

After running statistical tests in these two communities, a table of predictors with the corresponding p-values was calculated to determine whether the distribution of the predictors for the selected purple subgroup of nodes was different from that of the selected green nodes subgroup. If the desired significance level of any predictor was found to be statistically significant, we were able to construct a histogram representing normalized frequency distributions of the predictor for both the nodes in the selected purple region of the graph and those from the green region.

For the purpose of the statistical analyses undertaken for the present experiment, continuous, mixed, binary, and categorical (non-binary) univariate predictors were differentiated according to variable type. Continuous predictors were examined using the standard two-sample Mann–Whitney–Wilcoxon test. To examine the statistical association between two samples within the categorical data, Fisher's exact test and the χ^2 test were used for the binary and non-binary categorical variables, respectively.

In the next step, we analyzed various predictors with statistically significant p-values ≤ 0.05 to describe the differences between the purple and green communities. Thus, Figure 23 shows a histogram that compares the population in both. The purple community appeared to have more densely-populated counties, including cities and big urban centers. In contrast, the green community turned out to have moderately-populated counties, mostly located in rural areas (see Figure 19 for more detail).

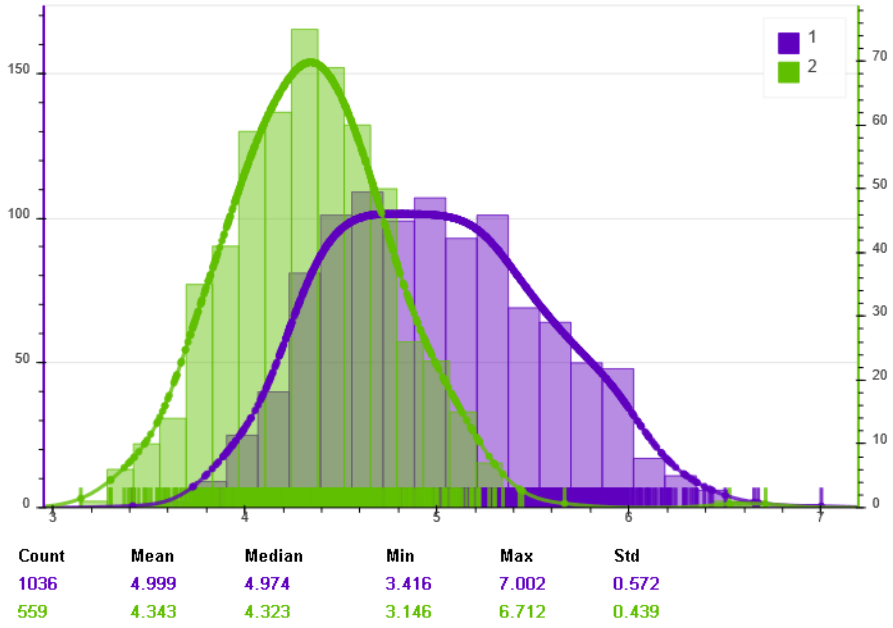


Figure 23. The histogram showing the **logarithm of population** in the green and purple communities

The next statistically significant predictor is the urban influence code, also called an “urban/rural score.” In the histogram illustrated in Figure 24, the value “1” on the scale corresponds to big urban centers while “12” corresponds to villages. The purple community mostly encompassed big urban centers, while the green community turned out to include counties in non-urban areas.

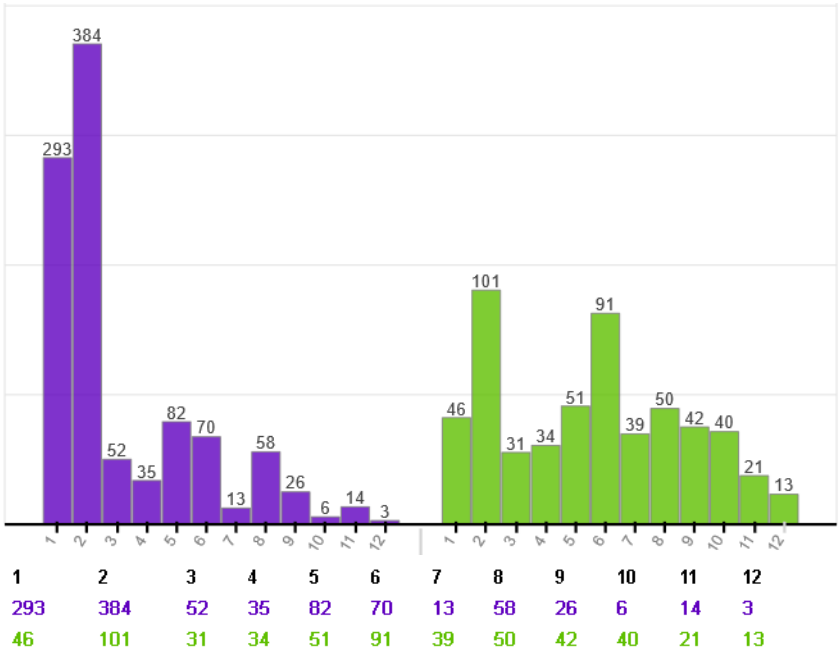


Figure 24. The histogram showing the **urban influence code** in the green and purple communities

Another predictor found to be significant is the public transportation system, which might be responsible for the difference in the number of deaths between the green and purple communities. We compared these communities based on the public transportation score (see Figure 25). The public transportation score indicates how well a location is served by public transit. The purple community appeared to have a much higher public transportation score than the green one.

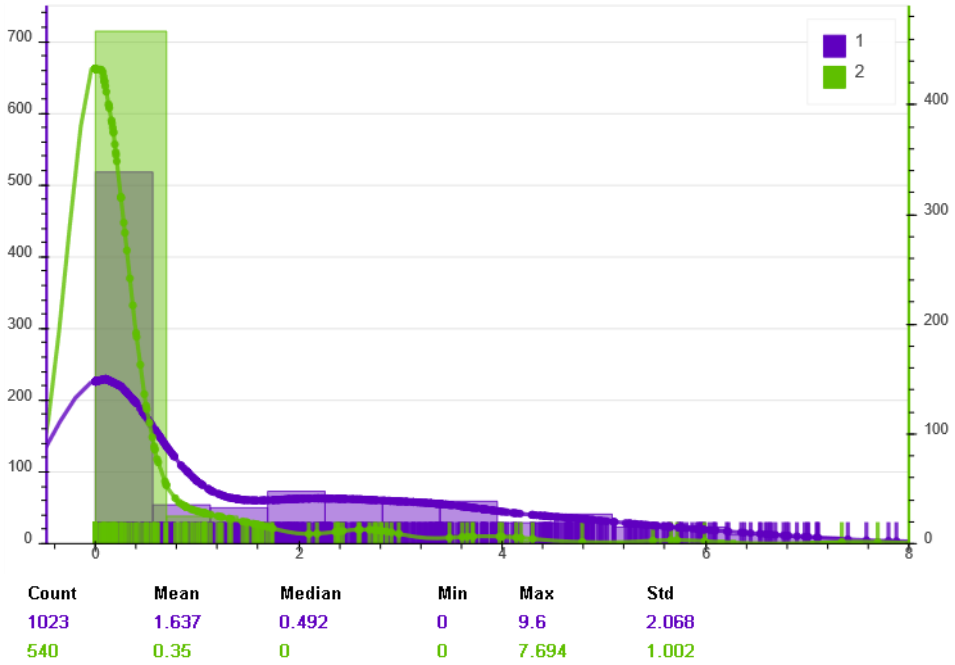
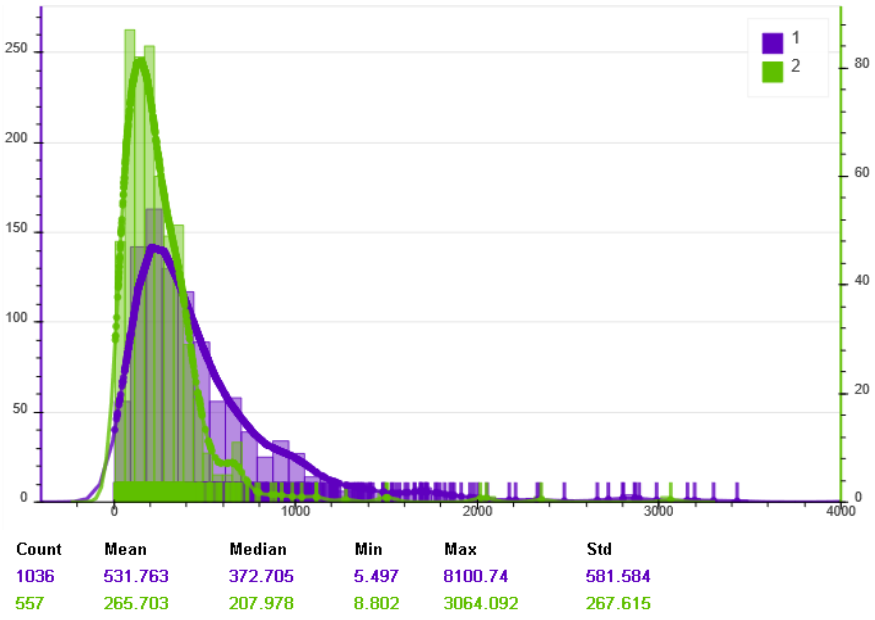


Figure 25. The histogram showing the **public transportation score** in the green and purple communities

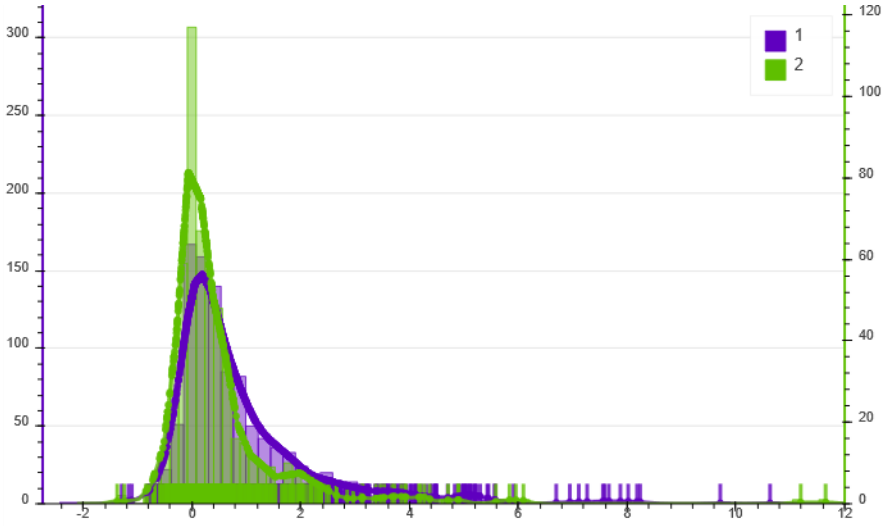
Analyzing more predictors in the transportation segment, Figure 26 shows a histogram comparing the purple and green communities based on highway length per 1,000 square kilometers. This predictor was found to be statistically significant with a p-value ≤ 0.05. The purple community appeared to have a much greater highway length than the green one.



Count	Mean	Median	Min	Max	Std
1036	531.763	372.705	5.497	8100.74	581.584
557	265.703	207.978	8.802	3064.092	267.615

Figure 26. The histogram showing the **highway density** in the green and purple communities

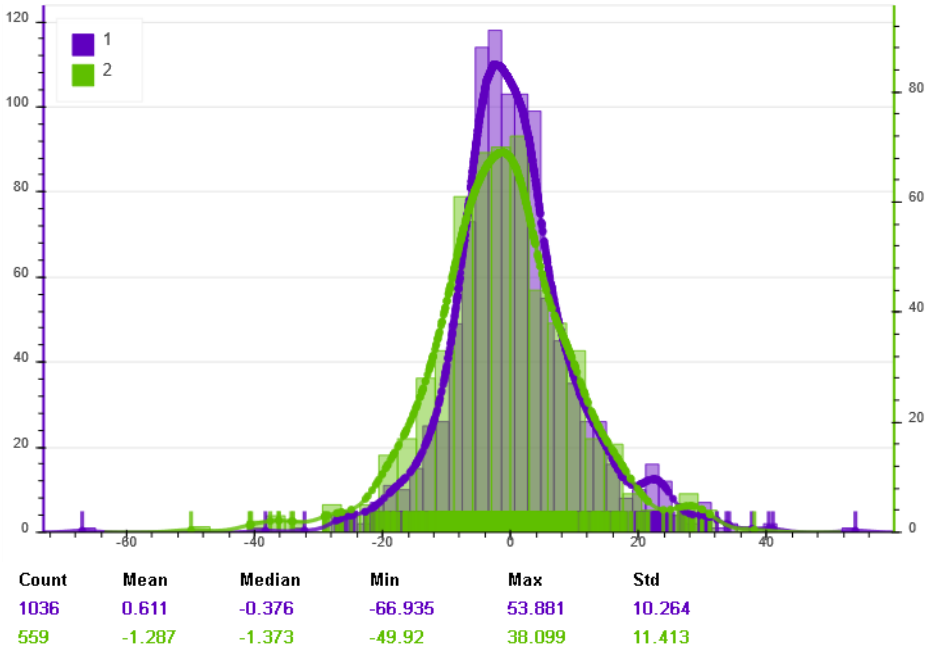
The international migration rate within the purple and green communities was also compared. This rate illustrates the difference between the number of people coming to the county from outside of the USA and the number of people leaving the county for international destinations throughout the year. Figure 27 shows this histogram. The purple community appeared to have approximately twice the international migration rate of the green community.



Count	Mean	Median	Min	Max	Std
1036	0.953	0.482	-1.292	13.998	1.47
559	0.584	0.192	-1.374	26.042	1.617

Figure 27. The histogram showing the **international migration rate** in the green and purple communities

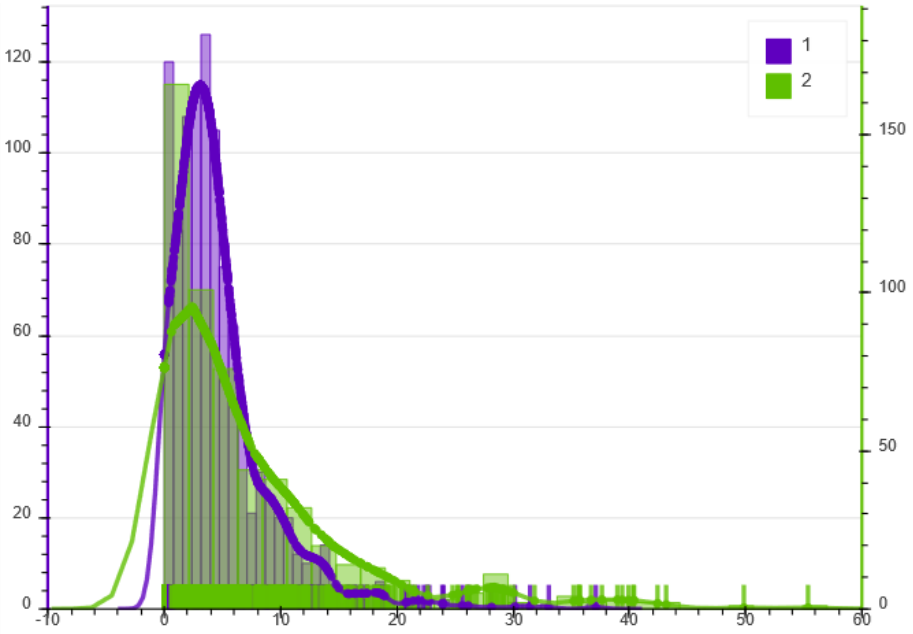
Likewise, we analyzed the domestic migration rate, which, in contrast with the international migration rate, refers to the difference in the number of people coming into and leaving the county from within the USA (not internationally). Figure 28 shows a histogram comparing the purple and green communities according to domestic migration rate. The purple community appeared to have a positive domestic migration rate, which means a surplus in arrivals into the counties of this community. The green community had a negative domestic migration rate, which means more people depart this community's counties than arrive.



Count	Mean	Median	Min	Max	Std
1036	0.611	-0.376	-66.935	53.881	10.264
559	-1.287	-1.373	-49.92	38.099	11.413

Figure 28. The histogram showing the **domestic migration rate** in the green and purple communities

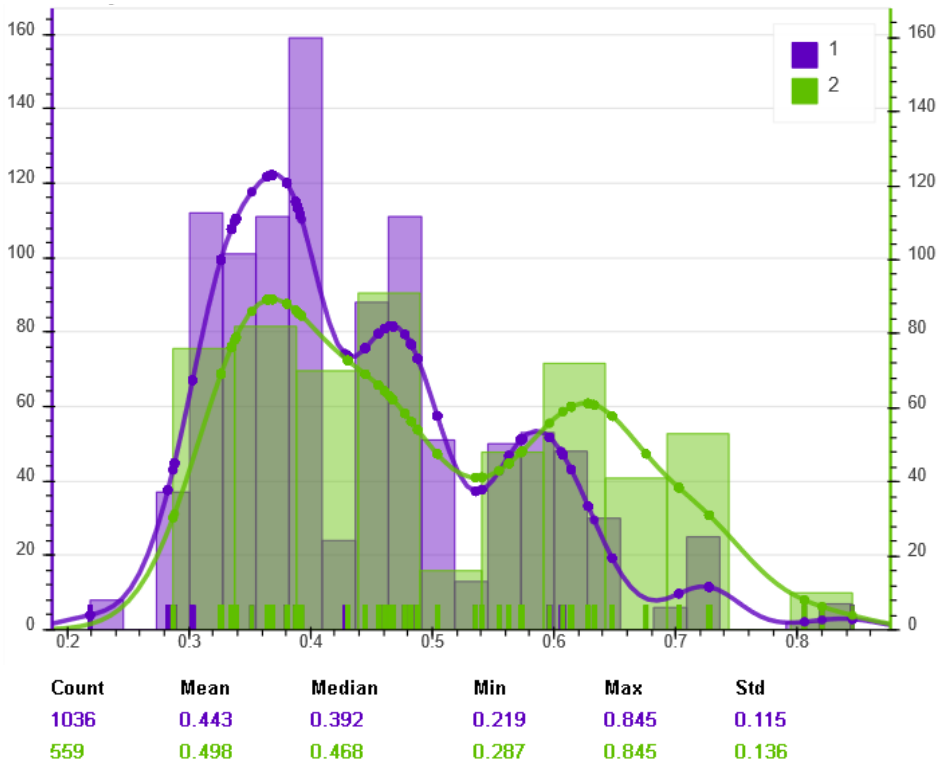
Another group of predictors that might be responsible for the difference in the number of deaths across the purple and green communities relates to the healthcare system. In the next step, the graph was re-colored according to the number of hospitals per 100,000 people. Figure 29 shows a histogram comparing the purple and green communities. The green community appeared to have twice the number of hospitals per 100,000 people than the purple one. This could be one reason that counties in the green community handled the pandemic outbreak more effectively even though they experienced a similar pattern in the number of cases as counties from the purple community.



Count	Mean	Median	Min	Max	Std
1036	4.847	3.762	0	37.116	4.524
559	7.9	4.569	0	200.602	12.563

Figure 29. The histogram showing the **number of hospitals** in the green and purple communities

Analyzing the healthcare system further, the purple and green communities were also compared according to the number of general practice physicians per 1,000 people. Figure 30 shows a histogram comparing the purple and green communities. The green community appeared to have a slightly higher number of general practice physicians per 1,000 people than the purple one.



Count	Mean	Median	Min	Max	Std
1036	0.443	0.392	0.219	0.845	0.115
559	0.498	0.468	0.287	0.845	0.136

Figure 30. The histogram showing the **number of general practice physicians** in the green and purple communities

Multiple other predictors integrated into the model were found to be statistically significant with a p-value ≤ 0.05 and could be responsible for the discrepancy between the green and purple communities in terms of the number of deaths. At any time, new predictors can be added to the model to find additional features that might be responsible either for similarities or dissimilarities in the patterns within selected communities.

3.2. MACHINE LEARNING-DETECTED COMMUNITIES

By constructing a graph representing the original dataset with the number of confirmed cases and deaths, we were able to undertake a visual exploration to discover sub-populations within the data. For example, isolated components of the graph or highly interlinked groups of nodes may indicate meaningful relationships in the dataset. As datasets can span a large number of nodes and edges, visual inspection and further discovery of sub-populations within the graph can be challenging or even misleading. Thus, we rely on some known machine-learning algorithms applied to the automatic detection of sub-populations using a graph-based community search (see Section 2.5).

Community detection relies on revealing subgroups of densely connected nodes, with many edges connecting nodes of the same community and comparatively few edges connecting nodes of different communities. Such communities can be considered to represent relatively independent areas of a graph and help identify and exploit relevant relationships in the dataset.

The machine-learning algorithm was applied to our graph to detect communities that might not be obviously identifiable when only visual exploration techniques are used. Figure 31 illustrates the result of the automatic community search algorithm. We focus our **analysis on the comparison of blue and yellow communities, since they showed similar patterns** in the number of confirmed cases and number of deaths (see colors of nodes in Figure 21 and Figure 22). Also, concentrating our analysis on differences between these two communities may enable us to discover reasons why the TDA algorithm has split green and yellow communities, which were among the four major communities described earlier (see Figure 18).

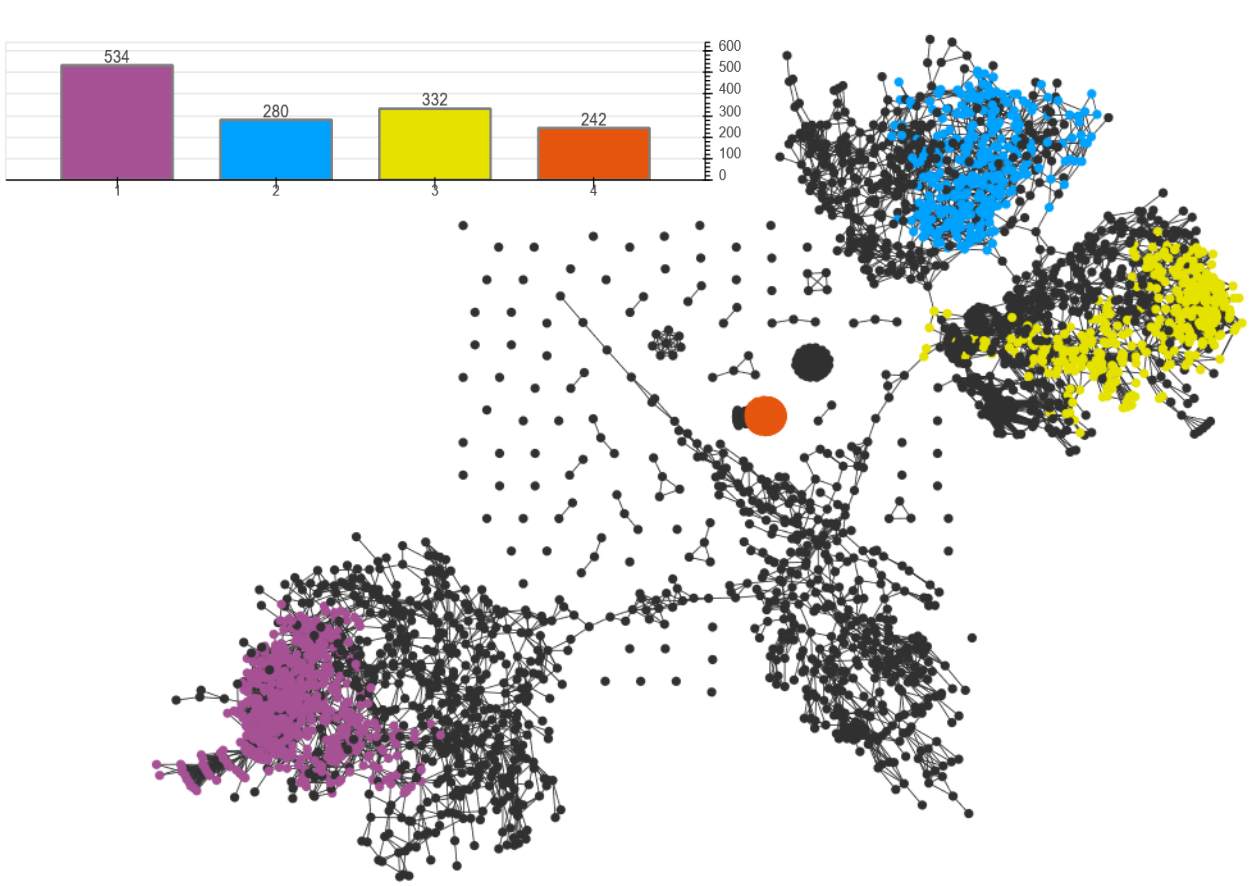


Figure 31. Communities detected on the graph using machine-learning algorithms

The first statistically significant predictor for the machine learning-detected communities appeared to be the number of days which passed between the date on which the stay-at-home restriction was adopted and the start day of the observation interval, i.e., the date of detection of the second confirmed case in the county.

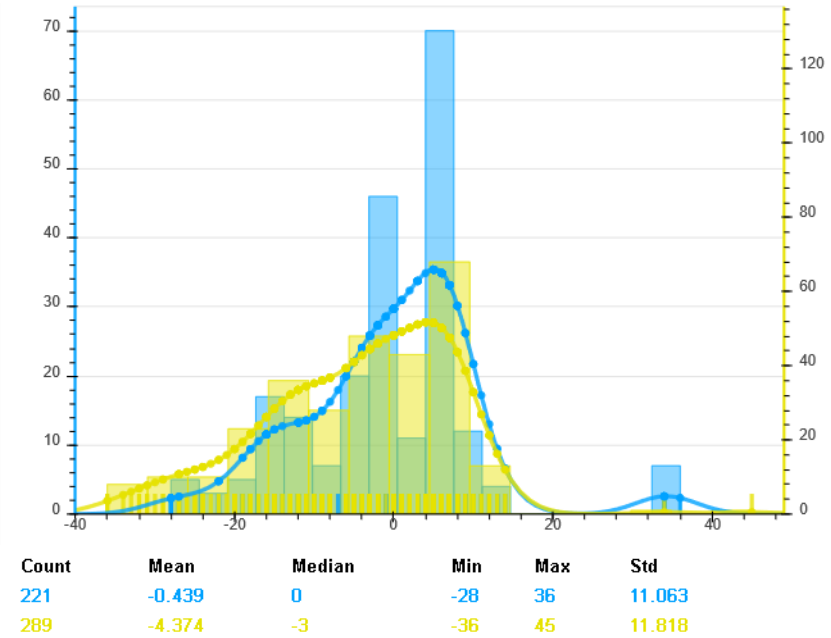
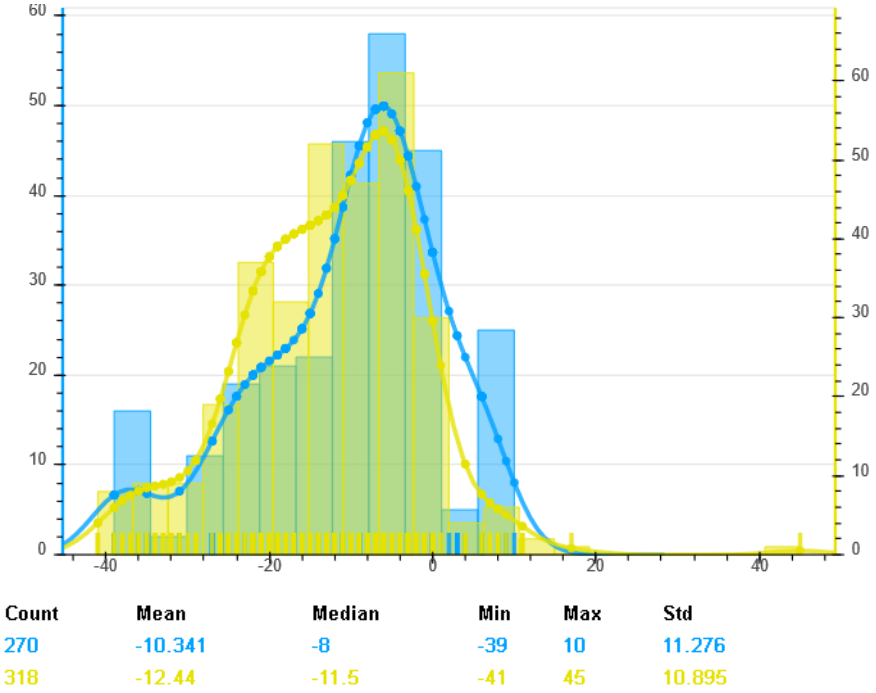


Figure 32. The histogram comparing the yellow and blue communities according to the number of days which passed between the date on which the stay-at-home restriction was adopted and the start date of detection of confirmed cases in each county

Figure 32 shows a histogram comparing the blue and yellow communities according to the number of days which passed between the date on which the stay-at-home restriction was adopted and the start date of detecting confirmed cases in each county. In the yellow community (with the smaller number of confirmed cases of the disease), the delta (number of days) between the stay-at-home restriction date and the start day of the observation interval appeared to be negative in most counties. This means that the stay-at-home restriction was adopted before the second confirmed case of the disease was detected in these counties (see Figure 32). It was determined that in the blue community counties, the stay-at-home restriction was adopted after the start day.

In the next step, the graph was re-colored based on the delta between the date on which restrictions were introduced for bars and restaurants and the start day of the observation interval (see a histogram comparing the yellow and blue communities in Figure 33). In the yellow community, the bars and restaurants were closed earlier with respect to the start day than they were in the blue community.



Count	Mean	Median	Min	Max	Std
270	-10.341	-8	-39	10	11.276
318	-12.44	-11.5	-41	45	10.895

Figure 33. The histogram comparing the yellow and blue communities according to the number of days which passed between the date on which restrictions were introduced for bars and restaurants and the start date of detection of confirmed cases in each county

3.3. SIMILARITY IN EPIDEMIC CURVES

As discussed in Section 1.4., an epidemic curve is a histogram illustrating the onset and progression of an infectious disease outbreak in a particular population over a specific time. The time interval is displayed on the X-axis, and case numbers are shown on the Y-axis. This visual representation provides useful information on the size, pattern of spread, time trend, and exposure period of the outbreak. For the current experiment, epidemic curves were built for every county based on the number of confirmed cases.

The resulting epidemic curves are different in terms of shape (e.g., the peak of confirmed cases differs across counties, falling at the beginning, middle, or end of the curve) and total number of confirmed cases of the disease reported during the observation period (see Figure 5). Previously, all counties were grouped according to the similarity of epidemic curves and assigned to one of five types to illustrate similarity in terms of disease spread pattern. We used a machine-learning algorithm to distribute all communities between the two groups to speed up the experiment. Counties with epidemic curves 1, 2, and 3 (green, pink, and blue), with a peak at the beginning or the middle and a flattening of the curve thereafter, as shown in Figure 5, were combined into a first community (see the violet community in Figure 34). Counties with epidemic curves 4 and 5 (brown and orange), with a peak at the end of the curve or a continuously rising curve, as shown in Figure 5, were combined into a second community (see the light brown community in Figure 34). In other words, **violet counties were able to take control of the pandemic** within an observational time interval and flatten the curve. At the same time, the **light brown communities experienced a rise in the pandemic** with no sign that control was taken.

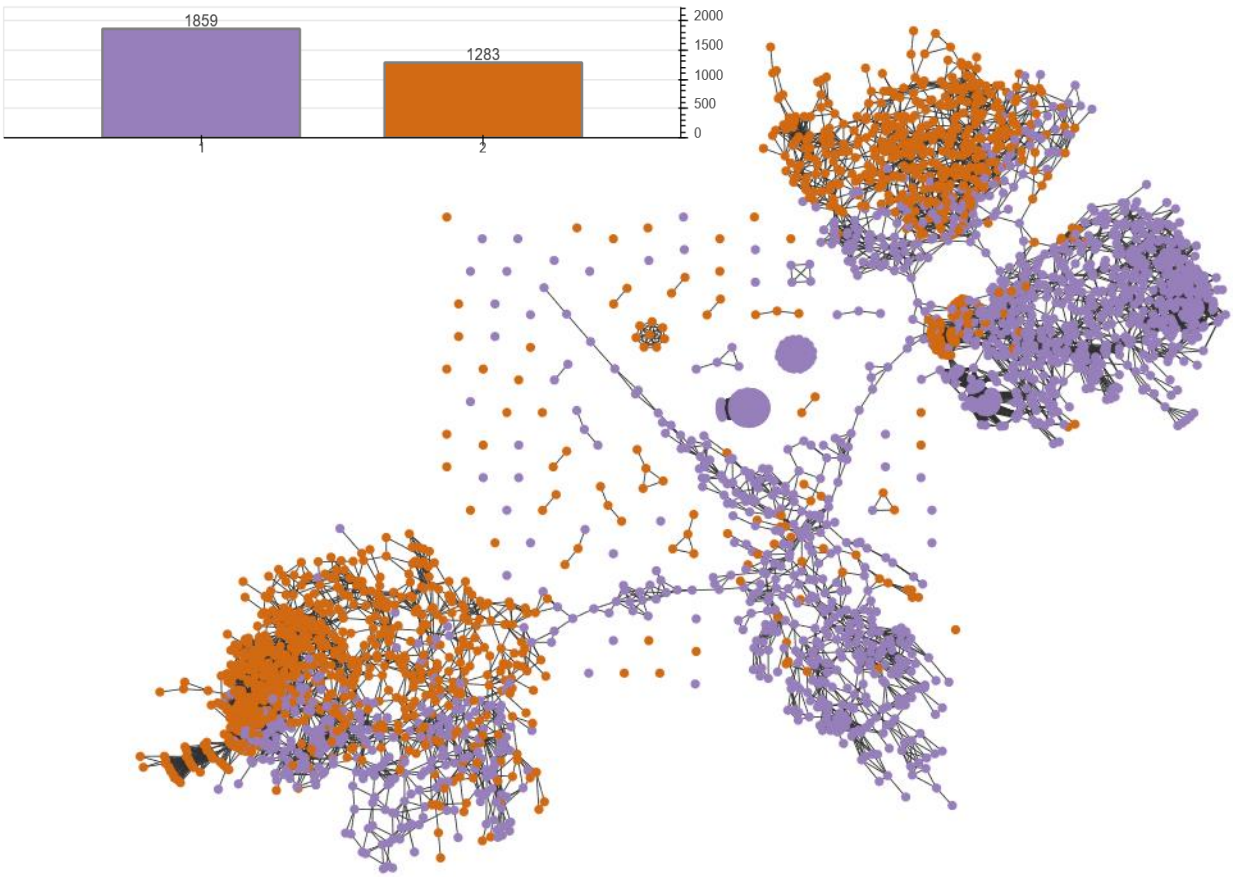


Figure 34. The color of nodes indicates the distribution of counties according to the epidemic curves (violet counties were able to flatten the curve while light brown counties experienced a rise in the number of confirmed cases)

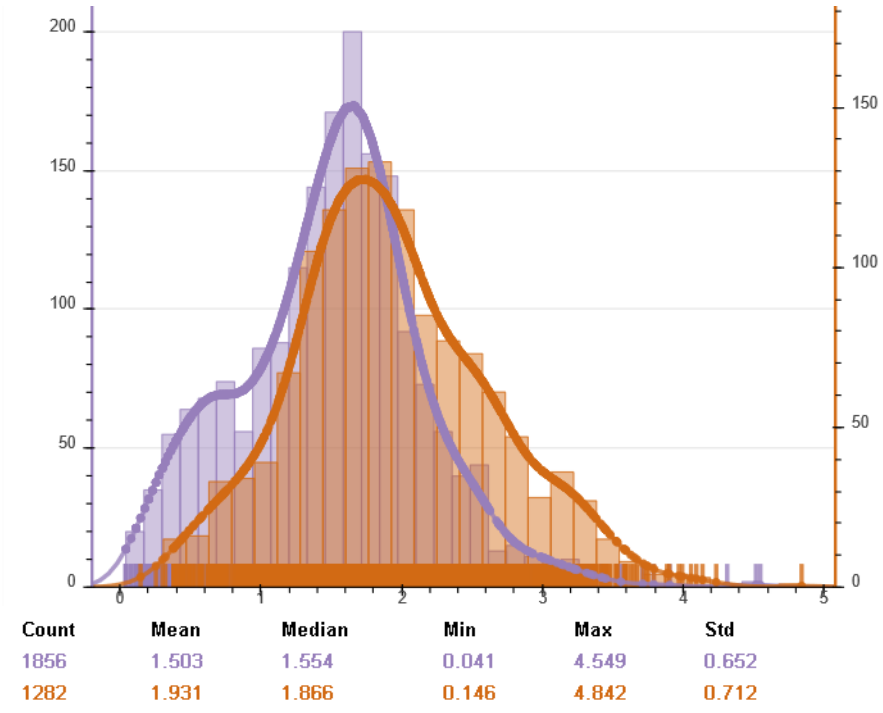


Figure 35. The histogram comparing the light brown and violet communities based on population density

We further compared these two communities using the statistical methods described above to analyze various predictors integrated into the model. Figure 35 shows a histogram comparing the light brown and violet communities based on the population density predictor, which was statistically significant with a p-value ≤ 0.05. The light brown community (which included counties with epidemic curves showing a controlled spread of the disease) appeared to have a population density about three times that of the violet community (where low or no control of the disease spread was observed).

The next statistically significant predictor for the light brown and violet communities appeared to be the delta between the stay-at-home restriction date and the start day of the observation interval (see Figure 36). In the violet community, the stay-at-home restriction date and start of observation appeared to be approximately the same day. In the light brown community, the stay-at-home restriction date fell about six days later than the observation interval's start day.

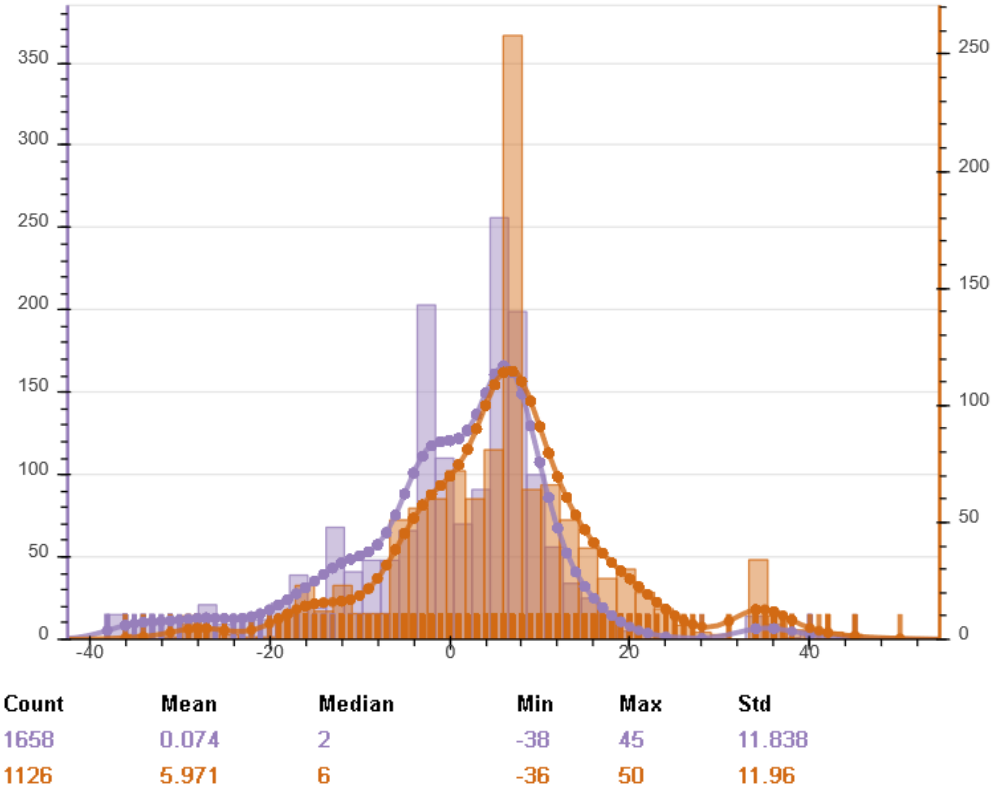
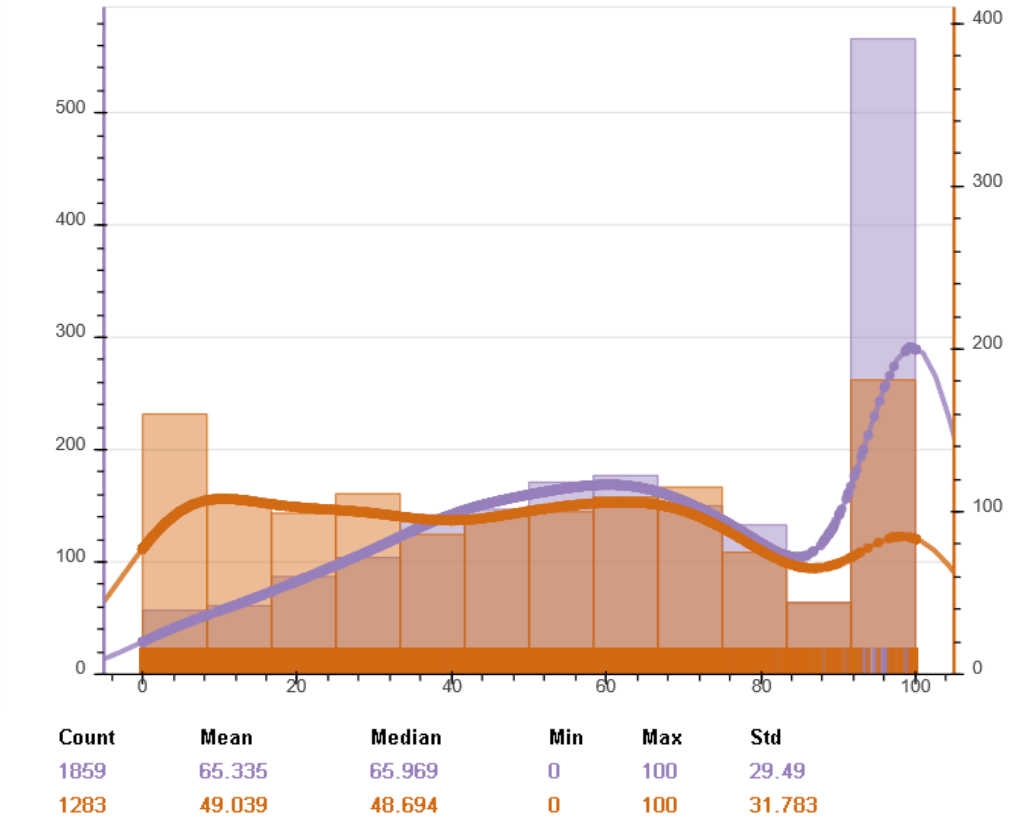


Figure 36. The histogram comparing the light brown and violet communities according to the number of days between the stay-at-home restriction date and the start date of detection of confirmed cases

Next, we analyzed communities based on the percentage of the population living in a rural area. Figure 37 shows a histogram with this analysis. The violet community appeared to have a higher percentage of population living in a rural area than the light brown community.



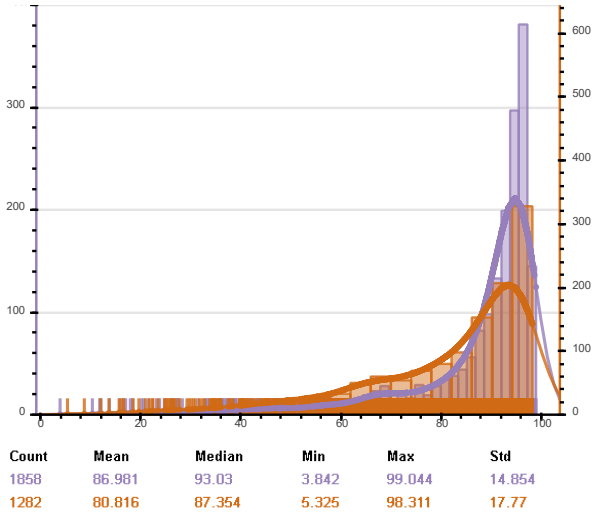
Count	Mean	Median	Min	Max	Std
1859	65.335	65.969	0	100	29.49
1283	49.039	48.694	0	100	31.783

Figure 37. The histogram comparing the light brown and violet communities according to the percentage of the population living in a rural area

Further, several predictors related to the race of the population are statistically significant. Figure 38 (a-e) shows a histogram comparing the light brown and violet communities according to the percentage of different races. The violet community appeared to have a higher percentage of Caucasian and Native American populations than the light brown community. The light brown community had a higher percentage of African American, Hispanic, and Asian populations than the violet community.

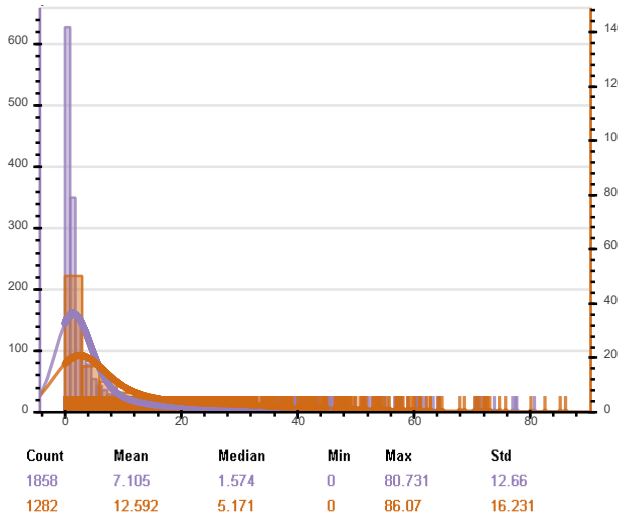
The next statistically significant predictor for the light brown and violet communities was found to be a climate region. In the violet community, the number of counties located in central, west, north central, northwest, and south climate regions and Alaska was higher than in the light brown community. In contrast, in the light brown community, the number of counties belonging to the northeast and southeast climate regions was higher than in the violet community. Figure 39 shows the USA's climate regions, and Figure 40 shows histograms comparing the light brown and violet communities according to the climate region.

The data mining experiments indicated various predictors that could influence the spread of COVID-19 in the USA. These include geographic conditions, population density, urban influence, public transportation, highway length, migration, number of hospitals, number of general practice physicians, number of days passed between introducing the stay-at-home restriction and the start date of detection of confirmed cases of the disease, and many others. At any time, additional predictors can be easily integrated into the model to expand the search of parameters that might be responsible for similarities in pandemic spread discovered by using TDA, machine learning algorithms, and visual exploration techniques.



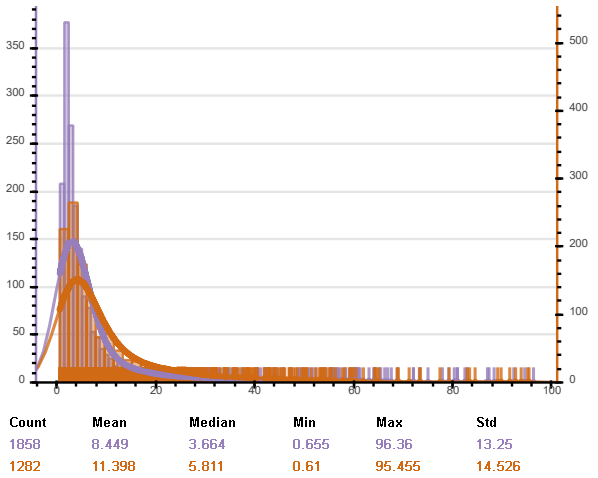
Count	Mean	Median	Min	Max	Std
1858	86.981	93.03	3.842	99.044	14.854
1282	80.816	87.354	5.325	98.311	17.77

a) Percentage of Caucasian



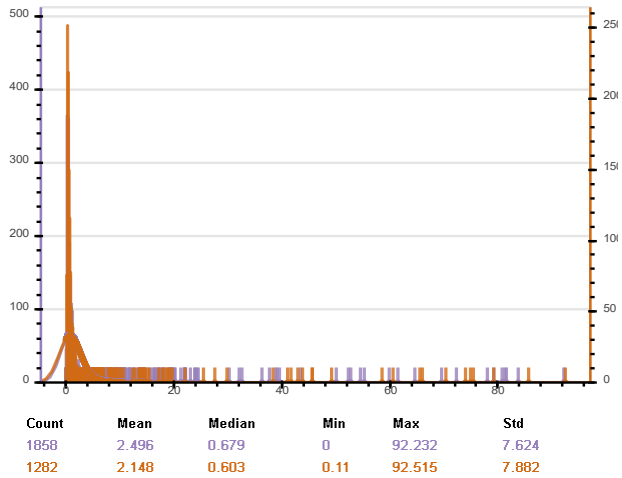
Count	Mean	Median	Min	Max	Std
1858	7.105	1.574	0	80.731	12.66
1282	12.592	5.171	0	86.07	16.231

b) Percentage of African American



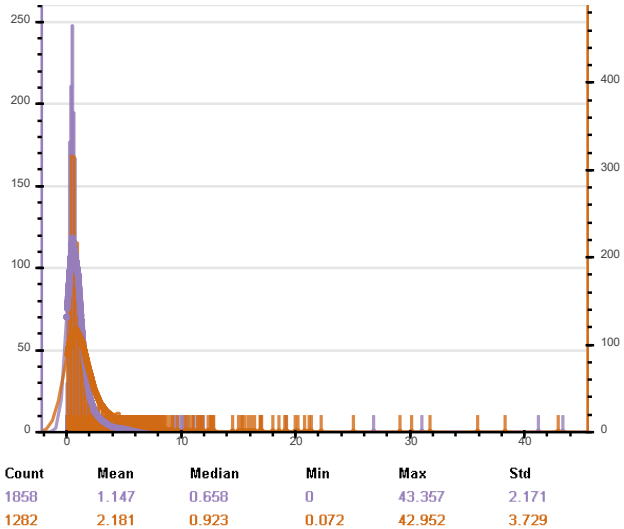
Count	Mean	Median	Min	Max	Std
1858	8.449	3.664	0.655	96.36	13.25
1282	11.398	5.811	0.61	95.455	14.526

c) Percentage of Hispanic



Count	Mean	Median	Min	Max	Std
1858	2.496	0.679	0	92.232	7.624
1282	2.148	0.603	0.11	92.515	7.882

d) Percentage of Native American



Count	Mean	Median	Min	Max	Std
1858	1.147	0.658	0	43.357	2.171
1282	2.181	0.923	0.072	42.952	3.729

e) Percentage of Asian

Figure 38. The histogram comparing the light brown and violet communities based on the percentage of different races within the communities

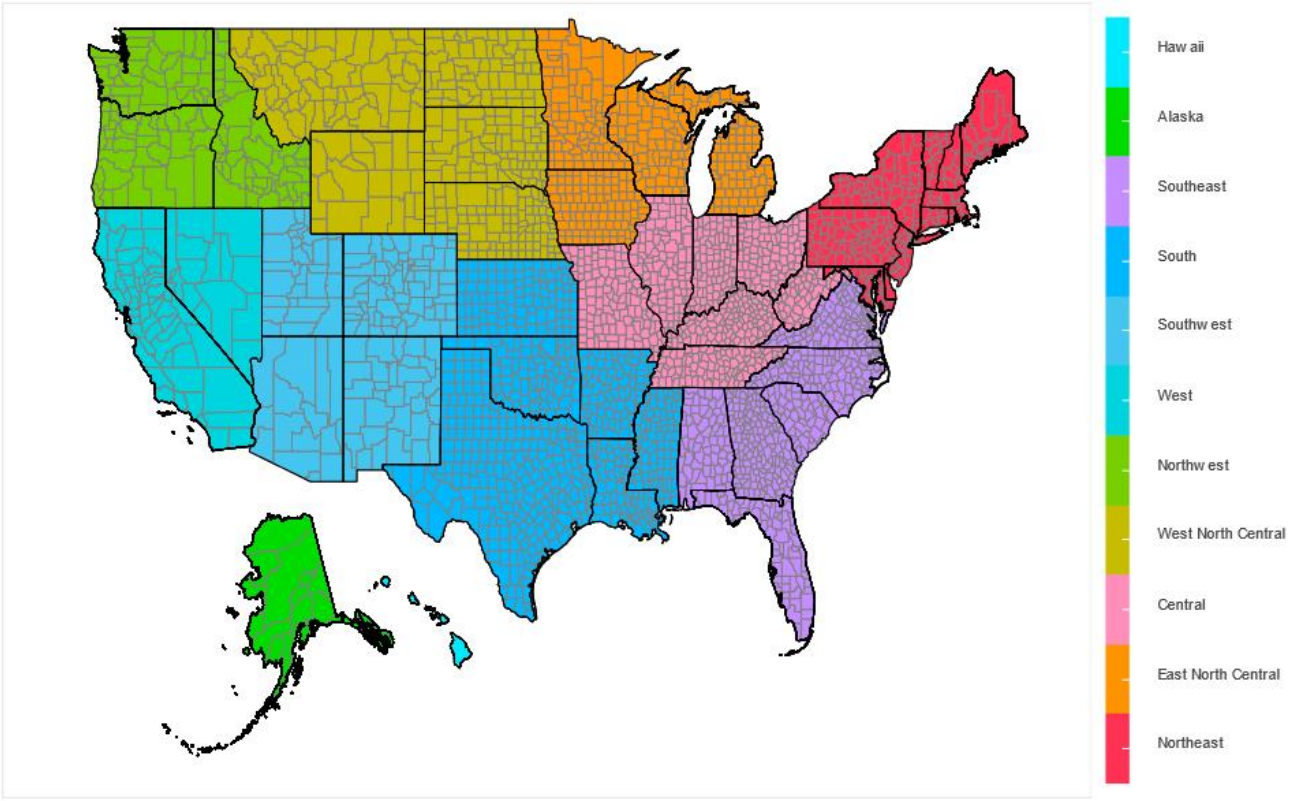
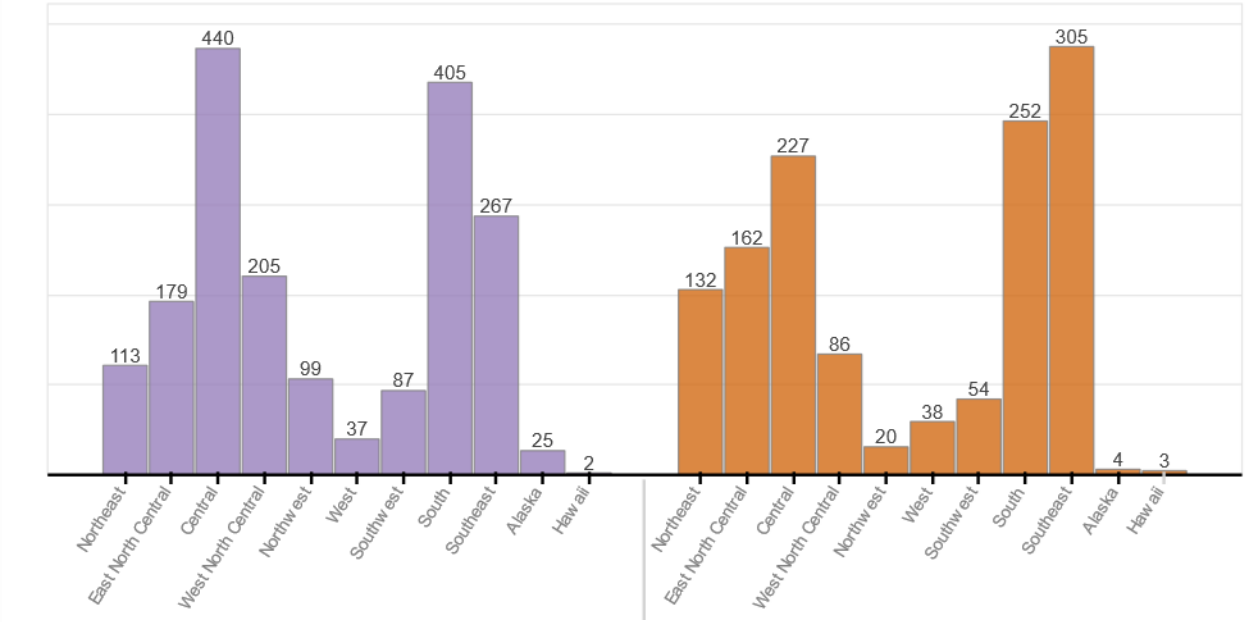


Figure 39. The climate regions of the USA



Northeast	East North Central	Central	West North Central	Northwest	West	Southwest	South	Southeast	Alaska	Hawaii
113	179	440	205	99	37	87	405	267	25	2
132	162	227	86	20	38	54	252	305	4	3

Figure 40. The histograms comparing the light brown and violet communities according to the climate region of the counties

4. CONCLUSION

In this paper, the authors present novel machine-learning techniques that rely on graphs as the fundamental approach to structuring and analyzing complex real-world data. The authors analyze how the spread of COVID-19 has advanced over the USA, looking for similarities in a specific timeframe. The analysis was performed on a county-by-county basis, by extracting from the dataset a topological data model represented as graph in which each of 3,142 nodes corresponds to one county, and two nodes are connected if they share similarities. After the graph was built, real-world data were used to integrate into the model any predictors which might be responsible for similarities in the early-stage spread of the pandemic. Over 250 predictors from different publicly available sources were used during the course of the experiment. The data mining experiments indicated various predictors that could influence the spread of COVID-19 in the USA. These include geographic conditions, population density, urban influence, public transportation, highway length, migration, number of hospitals, number of general practice physicians, number of days passed between introducing the stay-at-home restriction and the start date of detection of confirmed cases of the disease, and many others. At any time, additional predictors can be integrated into the model to expand the search of parameters that might be responsible for similarities in pandemic spread discovered by using topological data analysis, machine learning algorithms, and visual exploration techniques.

REFERENCES

[1] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (can be accessed at https://github.com/CSSEGISandData/COVID-19)
[2] A List of Statewide "Stay-at-home" Orders (https://www.littler.com/publication-press/publication/stay-top-stay-home-list-statewide)
[3] Torok, Michelle (2003). "Focus on Field Epidemiology" (https://nciph.sph.unc.edu/focus/vol1/issue5/1-5EpiCurves_issue.pdf). North Carolina Center for Public Health Preparedness.
[4] Edelsbrunner, Herbert; Harer, John (2010). Computational Topology: An Introduction. American Mathematical Soc. ISBN 9780821849255.
[5] Zomorodian, Afra (2005). Topology for Computing. Cambridge University Press. ISBN: 9780511546945.
[6] Carlsson, Gunnar (2009). "Topology and data". Bulletin of the American Mathematical Society. 46(2): 255-308.
[7] Zenil, H.; Hernández-Orozco, S.; Kiani, N.A.; Soler-Toscano, F.; Rueda-Toicen, A.; Tegnér, J. A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity. Entropy 2018, 20, 605.
[8] Fortunato, Santo. (2009). Community Detection in Graphs. Physics Reports. 486.

ACKNOWLEDGMENTS

We would like to acknowledge **Victoriia Shevtsova** (Intego Group, Ukraine) and **Lyudmyla Polyakova** (Kharkiv National University, Ukraine) for handling the computational experiment and being core members of the research team. We would also like to acknowledge Illia Skliar, Anna Petrenko and Natalia Averianova, students on the Kharkiv National University's Biostatistical Programming training program, who helped in gathering data and preparing the database of predictors used during the course of the experiment. Without you, this research and the experiment would not have been possible.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the group of authors at:

Contact: Sergey Glushakov
Company: Intego Group
Address: 555 Winderley Place, Ste. 129, Maitland, FL 32751
Work Phone: 407.641.4730
Email: sergey.glushakov@intego-group.com
Web: www.intego-group.com