SPATIAL DISTRIBUTION OF TWITTER DISCUSSION TOPICS REGARDING COVID-19 AND RELATED PUBLIC HEALTH POLICIES

A Thesis
by
HARRISON BROWN

Submitted to the School of Graduate Studies
at Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF ARTS

December 2022
Department of Geography and Planning

SPATIAL DISTRIBUTION OF TWITTER DISCUSSION TOPICS REGARDING COVID-19 AND RELATED PUBLIC HEALTH POLICIES

A Thesis
by
HARRISON BROWN
December 2022

APPROVED BY:

Dr. Kara Dempsey
Chairperson, Thesis Committee

Dr. Kathleen Schroeder
Member, Thesis Committee

Dr. Maggie Sugg
Member, Thesis Committee

Dr. Saskia van de Gevel
Chairperson, Department of Geography and Planning

Marie Hoepfl, Ed.D.
Interim Dean, Cratis D. Williams School of Graduate Studies

Abstract

SPATIAL DISTRIBUTION OF TWITTER DISCUSSION TOPICS REGARDING COVID-19 AND RELATED PUBLIC HEALTH POLICIES

Harrison Brown
B.A., Languages, Literatures, and Cultures, Appalachian State University
B.A., Geography and Planning, Appalachian State University

Department Chairperson: Dr. Saskia van de Gevel

Since the onset of the COVID-19 pandemic, general public opinion in the United States has shown a decrease in trust towards public health organizations and public health campaigns. Using social media as a lens for public opinion, this paper highlights the spatial distribution of the most widespread discussion topics among Twitter users regarding COVID-19, vaccinations, mask mandates, social distancing, quarantines, and shelter-in-place procedures. Using Topic Modeling, this study examined more than 5 million Tweets from January 1, 2020, to January 1, 2022. This study performs Topic Modeling, using Latent Dirichlet Allocation (*LDA*) as an intermediate step in analyzing the spatial distribution of topics across geographic scales. Analyzing the Twitter data provided 8 latent topics, from COVID-19 misinformation to vaccine requirements at school. Through Twitter geolocation information, each topic shows a unique spatial distribution across the contiguous United States. Analyzing and interpreting the prevalence of these topics allows policymakers to better understand sentiments regarding COVID-19 and related public health campaigns at the community and state level.

Acknowledgments

My sincere thanks to my graduate and thesis advisor Dr. Kara Dempsey, the members of my thesis committee Dr. Kathleen Schroeder and Dr. Maggie Sugg, Dr. Saskia van de Gevel, and the faculty and staff of the Department of Geography and Planning at Appalachian State University. My thanks to the Twitter Developer team for approval of Academic Research Access, and to the authors of the various R and Python packages which made this analysis possible. Additional thanks to my friends and family who have supported me throughout my career as a graduate student and during my time writing this thesis.

Table of Contents

Abstract iv

Acknowledgments v

List of Tables vii

List of Figures viii

Foreword x

1 Introduction 1

2 Background 3

3 Research Questions 10

4 Methodology 13

5 Results 21

6 Discussion 28

7 Conclusion 31

References 32

Figures 36

Vita 48

List of Tables

Table 1: Brief descriptions of common themes within each topic.. 22

Table 2: A table of ‘sample’ Tweets with three Tweets from each topic. 23

List of Figures

Figure 1: Map of unique Place IDs within the dataset across the continental United States. Place IDs for US States were replaced with polygon geometries, rather than rectangular boundary boxes. 19

Figure 2: Count of unique Place IDs which overlap each county. This distribution highlights the bias of populated regions in the dataset in terms of prevalence of Place IDs. 20

Figure 3: General spatial distribution of all 8 topics. The values within each legend provide the percentage of the total Tweets of each topic. See Table 1 for the description of each Topic, and Figure 4 through Figure 11 for more detailed maps of the distributions. 24

Figure 4: Overview of Local Moran's I clusters. Figure 13 through Figure 20 provide more detailed visualizations of the clustering of each topic. 26

Figure 5: Spatial distribution of Topic 0 (Coronavirus Misinformation, Hoaxes, "China Virus") 36

Figure 6: Spatial distribution of Topic 1 (Quarantine, Shelter-in-Place, Missing out on Social Events) 37

Figure 7: Spatial distribution of Topic 2 (Vaccines, Vaccine Development, Vaccine Aversion, Vaccine Mandates) 38

Figure 8: Spatial Distribution of Topic 3 (COVID-19 Testing, COVID-19 Dismissal) 39

Figure 9: Spatial distribution of Topic 4 (Lockdown, Shelter-in-Place, Mask Mandates) 40

Figure 10: Spatial distribution of Topic 5 (Public Health, Healthcare, Hospital Workers) 41

Figure 11: Spatial distribution of Topic 6 (COVID-19 Deaths, Hospitalizations) 42

Figure 12: Spatial distribution of Topic 7 (School Closures, School Mask & Vaccine Mandates) 43

Figure 13: Results of Local Moran's I for Topic 0 44

Figure 14: Results of Local Moran's I for Topic 1 44

Figure 15: Results of Local Moran's I for Topic 2 45

Figure 16: Results of Local Moran's I for Topic 3 45

Figure 17: Results of Local Moran's I for Topic 4 46

Figure 18: Results of Local Moran's I for Topic 5 46

Figure 19: Results of Local Moran's I for Topic 6 47

Figure 20: Results of Local Moran's I for Topic 7 47

Foreword

The content of this thesis will be submitted to *Geohumanities*, a peer-reviewed journal part of the *American Association of Geographers*, published by *Taylor and Francis*. Portions of this thesis may be formatted or modified in accordance with the style guide of *Geohumanities*..

1 Introduction

Social media is a powerful source of qualitative information for the analysis of public opinion and sentiments of issues worldwide. Many government and private organizations have a strong presence on social media platforms; user engagement in the form of comments, replies, etc., allows for unique insight into the dialog between citizens and the organizations that promote and develop public health policy (Murphy et al. 2014; Klašnja et al. 2017; Neubaum and Krämer 2017; McGregor 2019). However, proper utilization of social media data requires careful qualitative and quantitative analysis. Tools such as Topic Modeling are capable of extracting latent (hidden) topics of discussion without the expenses involved in manual content coding (Blei, Ng, and Jordan 2003; Alghamdi and Alfalqi 2015; Garcia and Berton 2021). Additionally, Isoaho et al. (2021) argue that Topic Modeling is a useful tool in performing qualitative analysis on large-scale data when used for policymaking; the authors argue, however, that researchers need a rigorous understanding of the statistics and computational methodologies in order to produce meaningful results.

Whereas traditional qualitative analysis requires manual interpretation of relatively small numbers of documents (e.g. manually coding interview transcripts, survey responses, etc.), text mining allows for vast amounts of data – often in the range of hundreds of thousands to millions of data points – to be collected through Application Programming Interfaces (APIs) and subsequently processed and analyzed automatically through programming languages such as Python or R (Hong and Davison 2010; Pak and Paroubek 2010). Moreover, the potential of text mining social media information has drastically increased in recent years, due to an everincreasing amount of content published online on platforms such as Facebook, Twitter, or Reddit.

This study argues that properly understanding the topics within the social media discourse, especially in the context of public health and health geography, allows government and private organizations to better implement public policy and science communication campaigns. Within the United States, a widespread mistrust of organizations has unfolded, especially since the onset of the COVID-19 pandemic; both citizens and prominent political figures have publicly undermined the Centers for Disease Control and Prevention (CDC), as well as other public health agencies (Dyer 2020; Jaffe 2020; Piller 2020). This skepticism and distrust are potentially dangerous, as they cast doubt upon legitimate public health campaigns and encourage the spread of infectious diseases, namely COVID-19 (Bella 2020; Bernheim et al. 2020; Dave et al. 2020; Gonsalves and Yamey 2020; Casola et al. 2021). Public safety measures regarding COVID-19 require support not only from individuals, but from communities, governments, and private organizations as well.

This study examines public perception within social media data taken from Twitter.. Twitter data was collected through the Twitter API for Developers with Academic Access, which allows for full-archive retrieval of data and includes more metadata than standard access, which limits data collection to seven days before the date of query (Twitter API Documentation 2021). The metadata collected by the Twitter API includes public, user-reported geolocation data in the form of a Twitter Place Object (see Section 4.4), which allows for analysis of the spatial distribution of the semantic topics within the discourse. Section 4.1 describes the use of the Twitter API in further detail.

While many studies performed Topic Modeling on Twitter Data regarding COVID-19 (Boon-Itt and Skunkan 2020; Dubey 2020; Manguri, Ramadhan, and Amin 2020; Garcia and Berton 2021), few have analyzed the spatial distribution of topics from a geographic perspective (Mei et al. 2008; Ghosh and Guha 2013). This study intends to bridge the gap between text mining and spatial analysis methods; by analyzing the spatial distribution of latent discussion topics generated with *LDA* (i.e., in which geographic regions are certain discussion topics more or less prevalent), it is possible to gain a more comprehensive understanding of social media narratives regarding COVID-19 than through Topic Modeling alone.

2 Background

Topic Modeling and other text mining methods are useful when analyzing large sets of data that would be impractical to perform manually due to both the sheer number of data points and the increasing complexity of the data model required. By programmatically analyzing the latent topics within a corpus, patterns and processes within the data can be uncovered and interpreted quantitatively.

The following section describes the literature and methodology behind Topic Modeling, as well as the applications of social media data and the perspective of geography on matters of public health. Inviting a geographical perspective into this study not only allows for an understanding of the basic spatial distribution of the data, but also allows the lenses of human and political geography to help in explaining the issues.

2.1 Short Text

A key limitation to not only Topic Modeling, but text analysis in general, is the length of documents within a corpus. Longer documents, such as novels, news articles, websites, etc., contain more text-based content, and are therefore a richer source of information (Litosseliti 2010). Short text, on the other hand, which includes formats such as social media posts, comments, reviews, SMS messages, chat messages, and other media which is short in length, offers little word co-occurrence (words are often not repeated within a document), so care must be taken when performing text mining (Rashid, Shah, and Irtaza 2019; Short Text Classification n.d.). Additionally, text mining methods like sentiment analysis and Topic Modeling do not perform as well on short text data; due to the size of the corpus in this study (approximately 5 million Tweets), the shortcomings of short text can be mitigated through sheer number of input data points (Litosseliti 2010; Rashid, Shah, and Irtaza 2019).

2.2 Sentiment Analysis

Several studies regarding public opinion and Social Media-based text mining involve the use of sentiment analysis, a dictionary-based approach wherein each word in a document is assigned a “sentiment score”, either a single value ranging from “positive” to “negative”, or a vector of values, which denote the abundance of a set of “sentiments”, e.g. “fearful”, “angry”, “joyful”, etc. (Mejova 2009; Liu 2012).

Several studies within the literature note that sentiment analysis performs poorly on short text data – e.g. Tweets, comments, or other short social media posts – often failing to pick up the highly contextual nature of this type of data (Liu 2012; Clavel and Callejas 2015; Debortoli et al. 2016; Boon-Itt and Skunkan 2020). Additionally, sentiment analysis often fails to properly classify complex sentence structures due to its word-for-word approach. For example, the sentences “the painting is good” and “the painting isn’t bad” have subtly different but similar semantic meanings (indicating positive sentiment towards a painting); however, sentiment analysis may classify the former as a positive phrase and the latter as a negative phrase, due to the presence of the word “bad”.

Initially, this study explored performing Sentiment Analysis on the corpus of Twitter data. The classification results, however, were first, inaccurate due to the short text nature of Twitter data and second, unable to convey meaningful information about the semantically complex topics of discussion within the dataset. Sentiment analysis is worth mentioning in the context of this study, as several studies have used SA as a means of quantifying user opinion regarding COVID-19 (Boon-Itt and Skunkan 2020; Dubey 2020; Manguri, Ramadhan, and Amin 2020; Garcia and Berton 2021). While sentiment analysis is a useful text mining methodology, its use may not always be appropriate on short text data; this study, therefore, uses Topic Modeling, which can provide more meaningful qualitative analysis on short text social media data.

2.3 Topic Modeling

Topic Modeling describes a framework of tools used within text mining to classify words and documents into semantic topics. By analyzing the distributions of words within a corpus of documents, Topic Modeling generates a set of latent, or “hidden” topics; documents that share common terms likely have overlapping themes, genres, and topics of discussion.

Different implementations of Topic Modeling algorithms exist; in this study, Latent Dirichlet Allocation (LDA) is used, as it is a widely supported Topic Modeling method that provides reasonable performance for large corpora (Blei, Ng, and Jordan 2003). The greatest utility of Topic Modeling is in analyzing large volumes of text; in the case of this study, more than 5 million individual Tweets were examined, containing an even larger number of connections *within* and *between* these Tweets.

This study uses several terms from the linguistics and text mining literature in order to maintain consistent terminology. The primary terms used in this study include:

- *Token*, the basic unit of text data, which includes single words (unigrams), and pairs or sequences of words analyzed together (bigrams and *n*-grams). Within this study, the term "word" is used interchangeably with *token* for the purpose of legibility, even if that token is composed of more than one grammatical word; for example, `COVID 19` is analyzed as a single token, despite the whitespace between `COVID` and `19`, and may therefore be considered a single token or "word".
- *Document*, an individual collection of tokens; in this study, a *document* refers to a single Tweet.
- *Corpus*, the collection of all documents used within the study. For this study, ≈ 5 million Tweets were collected (see Section 4.1).

Additional terminology is defined and explained as necessary, including parameters of models and other terms frequently encountered in the literature.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (*LDA*) is a generative, probabilistic model used for collections of discrete data, usually large corpora of text. Each document is defined as a mixture of topics, which are in turn composed of a mixture of tokens.

This study uses the *LDA* implementation from the `sklearn` Python package (Pedregosa et al. 2011).

LDA Input Parameters

The parameters outlined below control the *LDA* model, such as defining the number of topics and the shape of the distribution of topics within a document. An important step in the *LDA* process is assigning the ideal values for these parameters. The *short text* nature of Twitter data leads to additional challenges when applying parameters derived from the literature, as much of the foundational work was performed on "traditional" media containing longer passages of text (Blei, Ng, and Jordan 2003; Wallach 2006).

Number of Topics, k

The parameter k defines the number of topics for classification and usually ranges from ≈ 3 to ≈ 100. The "ideal" value of k is the number of "true," unknown topics within the corpus. As this true number is not known beforehand, the ideal value of k must be estimated by iterating different *LDA* models with differing values of k. As outlined in Blei, Ng, and Jordan (2003), the perplexity of the model (i.e., how well the probability model predicts the topics) dramatically increases with larger values of k, such that large values of k poorly estimate the distribution of topics. Therefore, the ideal value of k is that which is large enough to encapsulate the "true number" of topics within the corpus, but small enough to accurately model the distribution of topics across documents. Because of the size of the corpus and the approximate topics within (those defined by the Twitter API query; e.g., 'COVID', 'quarantine', etc.), it is not strictly necessary to optimize for the value of k; therefore, in this study, a value of $k = 8$ is used. In `sklearn`, k is the `n_components` parameter of the `LatentDirichletAllocation` class.

Document-Topic Prior

The `doc_topic_prior` parameter of `LatentDirichletAllocation` controls the prior of the document-topic distribution theta (θ). Within the literature, this parameter is often called alpha (α) (Hoffman 2010). In this study, a value of $\alpha = 0.125$ was used. The parameter α models the distribution θ of topics within each document, controlling the "shape" of the Dirichlet distribution used to assign topics within documents:

- At values of α close to 0, documents are likely to consist of a single topic
- Values of $\alpha \approx 1$ give a random distribution of topics within documents
- As $\alpha \to \infty$, topics become equally likely to occur

Topic-Word Prior

The `topic_word_prior` of `LatentDirichletAllocation` controls the topic-word distribution β; this topic-word prior is also known as η in the literature (Hoffman 2010). With a value of $\eta < 1$, topics are composed of words, while values of $\eta > 1$ generate topics with larger numbers of words. For this study, a value of $\eta = 0.05$ was used, as this provided topics with as little overlapping terms as possible, while still being descriptive enough to have sufficient meaning for interpretation. Due to the short text nature of Twitter data, and that by definition, all Tweets were queried by a select number of related keywords (e.g., `vaccine` and `quarantine`), this corpus will always have some overlap across topics, so optimizing the η parameter is less crucial than the other parameters like α and k.

2.4 Social Media Data

Social Media, as defined in Hill, Dean, and Murphy (2013), is the "collection of websites and web-based systems that allow for mass interaction, conversation, and sharing among members of a network". By using social media data, researchers are able to extract vastly greater amounts of data than from traditional formats (Murphy et al. 2014).

Performing qualitative analysis on social media data requires a different set of considerations than for traditional survey data. For example, it is important to consider if data from social media is truly representative of the general population; the demographic makeup of social media users, especially among different sites e.g., Twitter vs. Facebook vs. Reddit, is commonly considered different than the demographics of the general population – e.g., social media users are, on average, a younger subset of the population (Madden et al. 2013; Murphy et al. 2014).

Careful consideration must be made when performing analyses using social media data, especially on whether the data is truly representative of the general population. By obtaining a sufficiently large sample size, this study hopes to mitigate as much of the potential sampling bias as possible.

2.5 Geographic Perspective of Public Health Policy

In addition to the spatial distribution of Twitter discussion topics, this study aims to contribute to the literature by highlighting a methodology with potential applications in policymaking, for example in public health, science communication campaigns, and epidemiology. Understanding the issue from the perspective of *Geodemographics* is useful as a technique in offering 'valuable demographic context' to a number of public applications to better suit the needs of different communities (Harris, Sleight, and Webber 2005; Singleton and Longley 2009; Petersen et al. 2011).

Additionally, this study intends to investigate the literature regarding health geography, especially in the context of public policy and the decision-making process. This study can contribute to the health geography discipline by providing a methodology and results that allow for a more modern, social media driven approach to public perception of health policy (Cutchin 2007; Kearns and Collins 2010; Rosenberg 2017).

3 Research Questions

This study intends to answer research questions relating to Topic Modeling, spatial analysis, and health geography, to meaningfully contribute to the literature. In answering these questions, this study introduces an easily implemented methodology for analyzing the patterns of semantic topics within a corpus of data, and subsequently modeling the spatial patterns and distributions of these topics from a geographical perspective.

3.1 Topic Modeling

With the use of Topic Modeling, what latent (hidden) discussion topics exist within the Twitter data, and what do they highlight about public sentiment regarding COVID-19 and related issues? For example, are social media users primarily concerned with the epidemiological aspects of COVID-19 (e.g., health issues, transmissibility), social issues (e.g., lack of socialization and interaction from social distancing), political issues (e.g., "government overreach"), or other topics? Performing Topic Modeling reduces much of the potential sampling bias in qualitative analysis; as the topics developed in *LDA* are *latent*, information assumed about the topics is not known beforehand. Topics, therefore, which must be "read between the lines" are more easily modeled than in traditional qualitative analysis (Blei, Ng, and Jordan 2003).

3.2 Geospatial Analysis

From the lens of geospatial information systems (*GIS*), what is the spatial distribution across the United States of the latent topics uncovered through Topic Modeling (see Section 5). Are there regions that show a significant prevalence or absence of a particular topic, and does that pattern exist across spatial scales e.g. regional, state, county? Additionally, this study aims to codify the significance of these topic distributions using spatial autocorrelation metrics such as Local Moran's I. In doing so, statistically significant clusters and outliers can be uncovered across the United States.

3.3 Public Health

Topic Modeling has been used in many studies regarding the COVID-19 pandemic (Boon-Itt and Skunkan 2020; Garcia and Berton 2021), but the results of Topic Modeling are often seen as the "last step" in the analysis. This study intends to highlight the utility of Topic Modeling as a bridge between qualitative and quantitative data, which can be used in subsequent spatial analysis.

Using the geospatial results from Topic Modeling can facilitate public health policy and communication efforts to better tailor messaging and practices to the needs of the community. For example, a community that is concerned about the negative effects of large gatherings may have an increase in the number of Tweets about said gatherings, which could be analyzed through Topic Modeling (Bernheim et al. 2020; Dave et al. 2020). Doing so may lead to a greater understanding of the necessary planning and policy required to successfully communicate public health messaging to said community.

3.4 Study Specifications

The corpus used in this study is composed of approximately 5 million geolocated Tweets regarding COVID-19 and related public health policies. A large spatial study (e.g., the entire United States) requires a larger corpus to adequately sample as much of the study area as possible. As the majority of Tweets in general are not geolocated, 5 million geolocated Tweets represents a significant corpus of data.

With a geographical lens, this study expands upon the existing literature by analyzing the issue from a geospatial perspective by using user-reported geolocation information to quantify the distribution of topics across space. As of the time of writing (October 2022), few studies within the literature have integrated Topic Modeling with GIS (Mei et al. 2008; Ghosh and Guha 2013) – none are known to the author to study Topic Modeling of COVID-19 using GIS.

The results of Topic Modeling can be used in a meaningful manner by public health officials and decision-makers. This study intends to outline a comprehensive, easy-to-understand framework for Topic Modeling Twitter data and mapping the resulting spatial distribution of topics, such that public health policy can reflect the state of the general public's concerns and sentiments specific to disparate regions. For example, by Topic Modeling the outbreak of a novel virus on Twitter, health officials can better tailor messaging and communication efforts to more effectively encourage proper health procedures, such as social distancing and mask wearing. In addition, regions, communities, and demographic groups across the United States may have different sentiments towards "top-down" public health campaigns, so local governmental institutions must consider the concerns of their communities when planning for potential future outbreaks.

4 Methodology

4.1 Data Collection

Twitter API

The data for this study were collected from the Twitter API with Academic Access (Twitter API Documentation 2021), using the `academictwitteR` package for `R` (Barrie and Ho 2021). The Tweets were acquired with the following methods; R functions in this section are from the `academictwitteR` package unless otherwise stated:

1. After applying for Academic access – which allows for full-archive access to the Twitter API – the bearer token for the Twitter Developer account used was linked to the `R` environment through `set_bearer()`. This sets the necessary authentication credentials for full-archive collection.

2. A query for the data was constructed for data collection using `build_query()`, which outlines:

   - "Keywords" to be indexed; keywords are case insensitive, and include alternate grammatical forms (plural, past tense, etc.) not listed here for brevity . These keywords include COVID-19 related terms ('COVID', 'coronavirus', 'corona', 'SARS-COV2'), quarantine, social distancing, masks, mask mandates, and vaccines & vaccinations.
   - Language filter for Tweets; `lang = 'en'`
   - Whether user-reported geolocation exists; `has_geo = TRUE`
   - Additional filters, such as whether Tweets are "retweets" (`is_retweet = FALSE`) or "quotes" (`is_quote = FALSE`)

3. Start and end dates for indexing were set to `2020-01-01` and `2022-01-01`, respectively.

4. An optional but important step involves obtaining a count of potential Tweets using `count_all_tweets()`, which provides information about how many Tweets would be obtained with the current query. This step is useful, as the Twitter API has a limit on the number of Tweets they can pull per month (Twitter API Documentation 2021). For the Academic Access level used in this study, the monthly cap is 10,000,000 Tweets per month. This step prevents queries which are too broad from unintentionally "maxing-out" the monthly cap.

5. Using `get_geo_tweets()`, a corpus for the chosen query was collected in the form of paginated `.json` files, which contain up to 500 Tweets per file. This data contains information for each Tweet including, but not limited to:

   - `source`, the platform or application, e.g., "Twitter for Android"
   - `tweet_id`, a unique identifier
   - `text`, the text content
   - `user_username` and `author_id`, two identifiers for the Tweet's author
   - `created_at`, the timestamp of when the Tweet was created
   - "Metric" fields, such as `like_count`, `quote_count`, `retweet_count`, etc.
   - `place_id`, if `has_geo = TRUE` is set in `build_query()`, which contains a Twitter place object's place ID (Geo objects n.d.)

6. After the Tweets were collected, the resulting `.json` files were parsed with the `json` Python package and concatenated into tabular DataFrames for ease of subsequent manipulation.

In this study, between 300,000 and 500,000 Tweets per hour were collected, with a total of approximately 5 million Tweets, after removing duplicate Tweets that contained more than one keyword.

4.2 Text processing

Stopwords

Stopwords – terms like “the”, “at”, “and”, etc. which offer little significant semantic information – were filtered from documents. The `NLTK` Python package provides methods for removing stopwords in the `nltk.corpus.stopwords` module (Bird, Klein, and Loper 2009). In addition to the standard list of stopwords outlined in Lewis et al. (2004), Twitter-specific stopwords such as “AMP”, “RT”, etc. were removed. Wallach (2006) notes that, when word order is important, stopwords should not be removed; *LDA*, however, is a “bag-of-words” approach, in which word order is not considered (Blei, Ng, and Jordan 2003). Additionally, tokens such as URLs, emoji, emoticons, and superfluous punctuation were removed.

Duplicate Removal

As several queries were performed, duplicate Tweets existed within the corpus (e.g., a Tweet that mentions both “COVID-19” and “Vaccine”). These duplicate Tweets were removed by filtering their `tweet_id` field, leaving only a single instance of each Tweet. Additionally, this study removes duplicate Tweets from the same `user_id`, so a user posting repeat information does not artificially inflate the occurrence of certain words.

Word stemming

Word stemming is the process of removing grammatical declinations and conjugations, such that the base form of a word is returned. For example, “run,” “running,” and “runs” reduce to the form “run.” Stemming, also known as lemmatization, is done to preserve the semantic meaning of terms, while maintaining a manageable list of tokens across the corpus. Word stemming was performed with the Natural Language Tool Kit (*NLTK*) Python package (Bird, Klein, and Loper 2009).

Tokenization

Tokenization refers to splitting a document (i.e., a Tweet) into a series of tokens, or words. Tokenization can be done to split a document into individual words, or into *bigrams* or *n-grams*, which are sequences of two or more concurrent words. Latent Dirichlet Allocation is a "bag-of-words" approach to text mining, wherein word order does not factor into word-topic distribution. By creating *bigrams* and *n-grams*, word order is somewhat maintained, as phrases such as "do not" remain linked, rather than be split into the single term *unigrams* "do" and "not." Especially for *short text* data, this process is crucial, as the length of the document (i.e., 280 characters) limits the amount of information contained within each document. The literature suggests the potential benefits of using *trigrams*, but notes that the co-occurrence of *trigrams* is sparse in *short text*, leading to low term co-occurrence when using *trigrams* – the distribution of *trigrams* across documents is much lower than that of *unigrams* or *bigrams* (Bao et al. 2009; Hong and Davison 2010; Ostrowski 2015; Debortoli et al. 2016).

4.3 Topic Modeling

This study performed Topic Modeling through the use of the `sklearn` Python package; the following methodology section describes the workflow for creating the LDA model and applying it to the text.

The following sections describe the steps taken to create a topic model of the data using LDA:

1. A subset of approximately 15% of the overall corpus was used to create initial training data. This was done not only to reduce the computational resources required, but to reduce model over-fitting.

2. The training data was cast into a document-term matrix, representing the count of each token within each document. The document-term matrices for the training and testing data were created using the `sklearn` class `CountVectorizer`. Both DTMs selected unigrams and bigrams from the corpus, and filtered out tokens found in fewer than 10 documents to reduce the memory processing time required to fit the DTM to the LDA model.

3. The training document-term matrix was fit to a Latent Dirichlet Allocation model using `sklearn`. Latent Dirichlet Allocation was performed using the parameters outlined in Section 2.3, with the `LatentDirichletAllocation` class from `sklearn`. As stated in Section 2.3 the *LDA* parameters of $\alpha = 0.125$ and $\beta = 0.05$ were used.

4. The remaining testing data was processed in chunks into document-term matrices, and each matrix was fit to the training LDA model to apply the topics generated to the data.

5. The resulting data were concatenated into a Data Frame, with columns for the Tweet ID and for the proportion of each topic within the document (found in θ).

4.4 Geolocation

Geolocation is collected from the Twitter API in the form of a `place id`, a unique string of characters that reference a Place Object. These objects were designed by Twitter to encode a unique location, such as *Manhattan*, *New Jersey*, or *France* (Geo objects n.d.). The following attributes are included in each Place Object:

- `id`, or the `place id`
- `url`, which is used to acquire a `.json` file of place information
- `place_type`, such as `city`, `state`, etc.
- `name`, a short, human-readable representation of the place name, e.g. “Manhattan”

- full_name, the full, human-readable name, e.g. “Manhattan, N.Y.”
- country_code and country, e.g. “US” and “United States”, respectively
- bounding_box, a bounding box of the coordinates which enclose the place

The .json files generated by the academictwitteR package also include both a place_id for each Tweet, and the information outlined above for each place_id. Therefore, using this methodology, it is possible to link a topic distribution to each Tweet, each Tweet to a place_id, and each place_id to a georeferenced location on Earth, allowing for an analysis of the distribution of topics across a geographical area.

Each .json file was parsed to retrieve Place Objects, resulting in a Data Frame of each Object’s place_id, name (full_name), and bounding_box. Geolocations were selected to those within the Continental United States (states excluding Alaska and Hawaii), in order to maintain contiguity. The results of Topic Modeling (a Data Frame of tweet_id and each topic, indexed 0 through 7), were joined in a one-to-many relationship with the Place Objects table, such that each Tweet contained geolocation information.

The place_id bounding boxes for the contiguous 48 states were each replaced with a shapefile geometry of that state; using the original bounding boxes would have led to a high level of overlap between states. For example, the bounding box of California almost entirely overlaps the bounding box of Nevada, and Tweets geotagged as “California” would have also counted towards the distribution of Tweets in Nevada.

Topics were then aggregated to each unique place_id to reduce computational resources when performing spatial analysis. The resulting Data Frame consisted of each unique place_id, its bounding_box (converted to a GeoSeries using the geopandas Python package), the sum of each of the eight topics, and the sum of the number of Tweets within each place_id. Each aggregated topic column therefore represented the approximate number of Tweets of the given topic within the `place_id` and, when divided by the total number of Tweets, normalized the data to the percentage of Tweets within each location.

Overlapping Geolocations

The geolocations within the data were at a range of spatial scales, from state-wide to individual buildings. As various `place_id` bounding boxes overlapped one another (see Figure 1), it was necessary to summarize the percentage of topics across the study area into areal units.

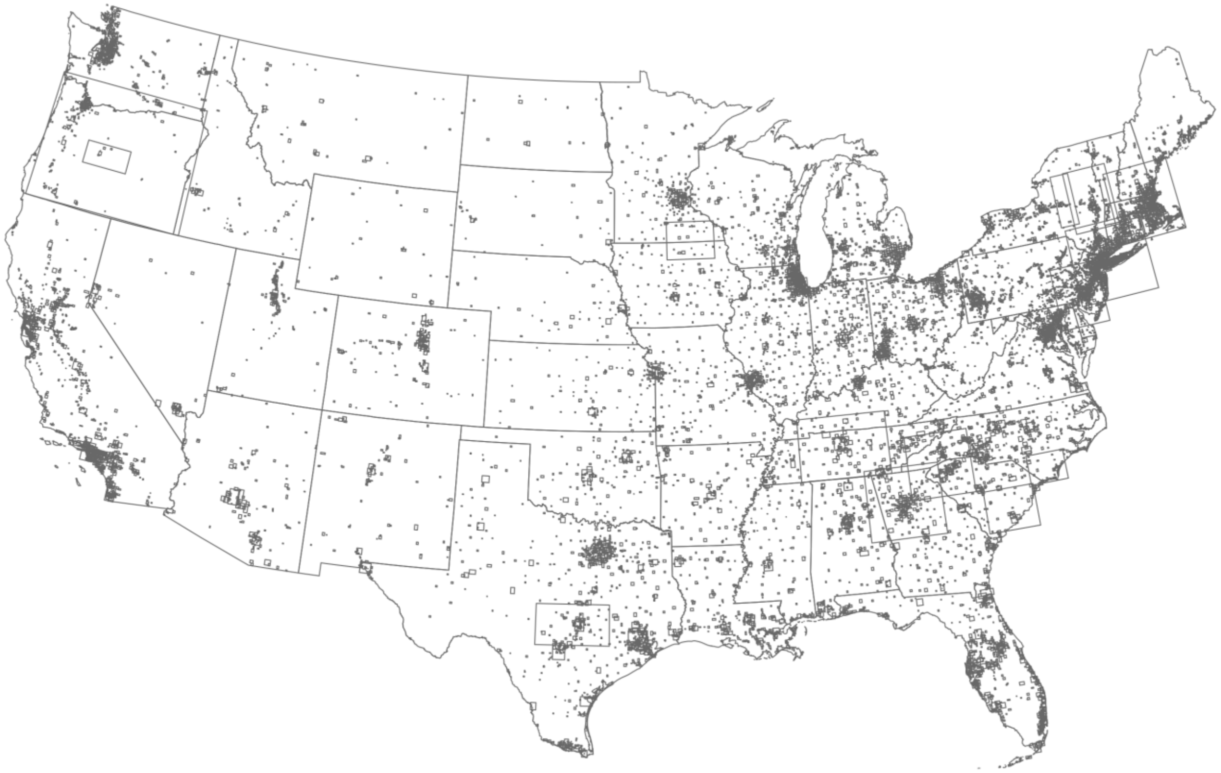

Figure 1: Map of unique Place IDs within the dataset across the continental United States. Place IDs for US States were replaced with polygon geometries, rather than rectangular boundary boxes.

The prevalence of topics across the contiguous United States was calculated using the Summarize Within tool from ArcGIS Pro 2.9.3. Topics within each `place_id` were summarized to the county level, resulting in the percentage of each topic within each county. The Summarize Within tool calculated the weighted average of topic percentages, with the weight proportional to the area of the `place_id` within the given county.

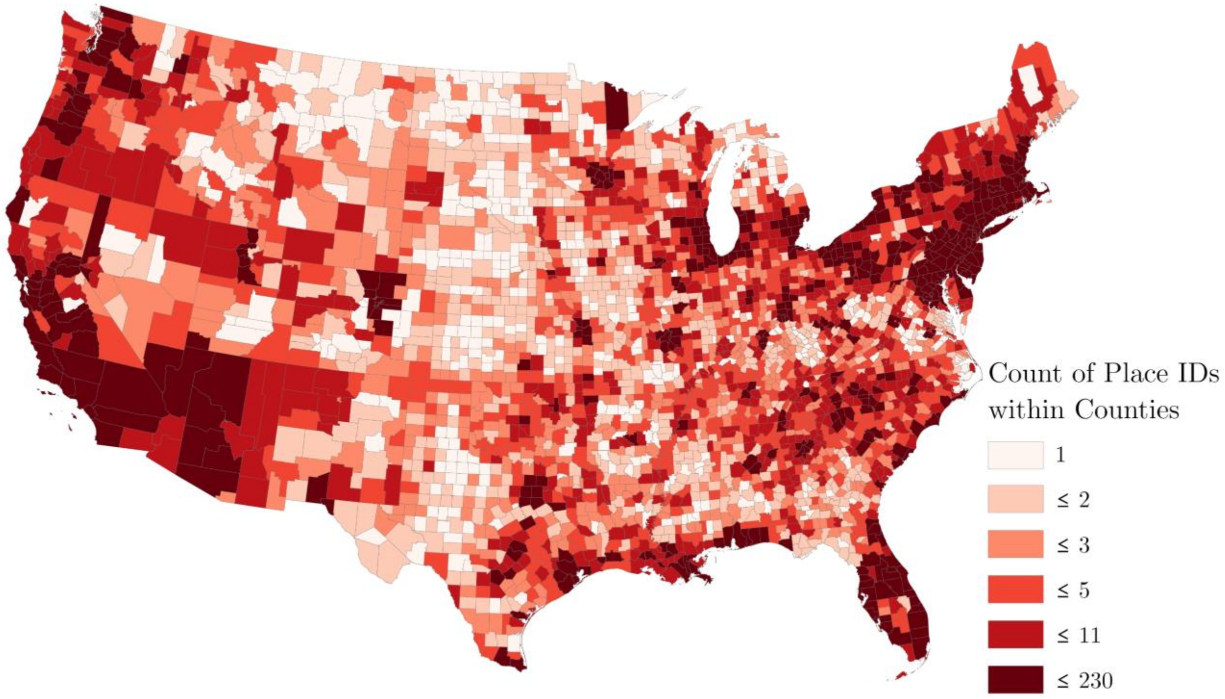

Figure 2: Count of unique Place IDs which overlap each county. This distribution highlights the bias of populated regions in the dataset in terms of prevalence of Place IDs.

4.5 Local Moran's I

Following the summarization of topic percentages within each county, this study analyzed local indicators of spatial autocorrelation (*LISA*) using ArcGIS Pro, in order to provide more statistically significant measures of regional hotspots and clustering. Figure 13 through Figure 20 (see Figures, pp. 36) provide the results for *LISA* analysis at a 95% confidence level.

5 Results

The results of this study are a percentage of Tweets of each of the eight latent topics within each county-level area. By analyzing the prevalence of each topic within counties, this study highlights general trends in the spatial distribution of the given topic. As the data were acquired at a variety of spatial scales – e.g. a Tweet located in Texas versus a Tweet located at a specific building within Houston, TX – careful consideration of the aggregated results is necessary. As a result, the maps generated as a result of this study are intended to show broad strokes in the distribution of topics.

The results of the Topic Modeling process (the terms used within each topic, and a summary of said topic), as well as the spatial distribution of these topics are outlined below.

5.1 Topic Modeling

This study performed topic modeling with $k = 8$ Topics. The short text nature of the data, as well as intrinsic overlap in the dataset caused by collecting topics with a keyword search limited the ability to estimate the ideal number of topics using metrics such as perplexity or topic coherence. Therefore, this number of topics was chosen to provide legibility in figures in tables, as well as to limit the necessary computational resources required to perform further analysis on all topics.

While automated methods exist for generating descriptive names of topics, e.g. the `LabelTopics()` function from `textmineR` (Jones 2019), a major challenge in Topic Modeling is providing adequate, descriptive labels to topics without introducing user bias. The following table is a listing of commonly occurring elements within each topic, and was determined by manually interpreting roughly 100 randomly selected Tweets for each topic.

Table 1: Brief descriptions of common themes within each topic.

LATENT TOPIC	DESCRIPTION
Topic 0	Coronavirus misinformation, hoaxes, “China Virus”
Topic 1	Quarantine, shelter-in-place, missing out on social events
Topic 2	Vaccines, vaccine mandates, vaccine development
Topic 3	COVID-19 testing, COVID-19 dismissal, “’rona”
Topic 4	Lockdown, Social Distancing, mask mandates
Topic 5	COVID-19 treatment, public health and healthcare, hospitalizations
Topic 6	COVID-19 deaths, symptoms, workplace vaccine requirements
Topic 7	Shelter-in-place, school closures, school-place mask requirements

Table 1 provides descriptions of each topic, with common elements seen in Tweets of each topic. Due to the overlap in common terms across topics, it is important to interpret the Tweet as a whole, rather than each word, in order to gain an understandable description of each topic. Although individual “elements” between topics may overlap, it is important to understand that the categorization of each topic incorporates all elements together. For example, *Topic 1* and *Topic 7* both include discussion of “shelter-in-place”, but are distinct due to the other elements in these topics, like “quarantine” for *Topic 1* and *“school closures”* for *Topic 7*.

An important takeaway regarding the use of Topic Modeling is that several latent topics, notably coronavirus misinformation, hospitalizations, school closures, and workplace vaccine requirements, were not explicitly queried as keywords; they represent the *LDA* model “reading between the lines” and successfully uncovering hidden topics within the corpus. As the primary limiter for a large-scale Twitter study is the rate limit on monthly queried Tweets, an exhaustive keyword search regarding every possible discussion topic is infeasible; therefore, Topic Modeling allows for a more detailed description of social media discussion topics without drastically increasing the size of the Twitter query.

Table 2: A table of 'sample' Tweets with three Tweets from each topic.

TOPIC	TEXT
Topic 0	fine print back biden presid yard sign n…
Topic 0	heard news flash donald trump go limit i…
Topic 0	march 13 ask take respons crisi happen c…
Topic 1	prepar everyth success i'm sure fair pos…
Topic 1	littl bit done semest especi hard sat ch…
Topic 1	ummm use quarantin time pleas watch daws…
Topic 2	knowledg base establish knowledg quest v…
Topic 2	flu vaccin decreas symptom period regard…
Topic 2	real vaccin prevent diseas covid jab pre…
Topic 3	could quarantin lot specul around deja v…
Topic 3	everi movi show watch charact touch anyt…
Topic 3	send prayer well wish hope alright feel …
Topic 4	must practic social distanc must cautiou…
Topic 4	absolut superspread event glad see least…
Topic 4	social distanc individu scare stay home …
Topic 5	new branson missouri open busi time covi…
Topic 5	busi liabl ill famili member worker cont…
Topic 5	wuhan show world end lockdown begin covi…
Topic 6	7 day averag texa covid death 296 sept 1…
Topic 6	mayor announc houston health depart repo…
Topic 6	new case berkeley alameda counti case tr…
Topic 7	new covid vaccin appoint avail 97035 saf…
Topic 7	2:30 call closest cv first covid shot to…
Topic 7	vaccin drive come februari 8th come get …

Table 2 provides three sample Tweets for each topic. The text for these Tweets is in the form of the processed Tweet text (outlined in Section 4.2); out of caution regarding the Twitter Developer terms of service, the full, unprocessed text, list of Place IDs, or any other potentially sensitive or identifiable information is not presented in this study.

5.2 Spatial Distribution

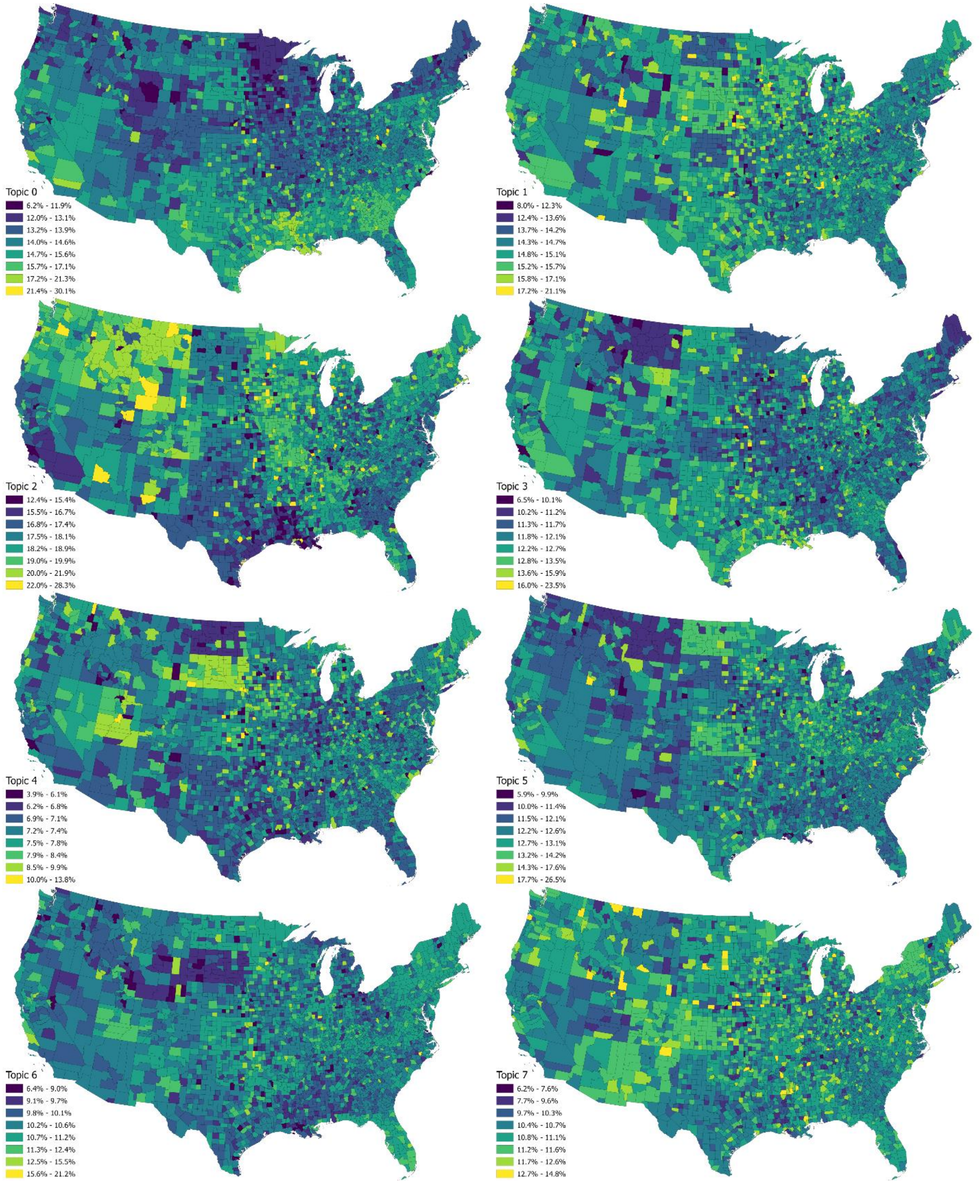

Figure 3: General spatial distribution of all 8 topics. The values within each legend provide the percentage of the total Tweets of each topic. See Table 1 for the description of each Topic, and Figure 5 through Figure 12 for more detailed maps of the distributions.

Figure 3 highlights the spatial distribution of each topic as a percentage of the total topics within each county. Each topic has a pattern of distribution across states and within states. The values within this map are intended to show the general trend in the pattern, rather than provide rigorous statistics that can be compared to e.g. demographic variables. Important to note for this study is the disparity between the statistical distributions of each topic. Topic 4, for example, shows a narrower range of relatively low values, while Topic 2 has a wider range of relatively high values.

5.3 Topic Distributions

Each of the eight latent topics shows a unique spatial distribution across the contiguous United States. Though caution must be exercised due to the aggregated nature of the data, interpreting each of the distributions allows for insights into the regions and communities that discuss a given topic more or less frequently. Below is a brief discussion of the distribution of a selection of topics, based on the results of Local Moran's I analysis.

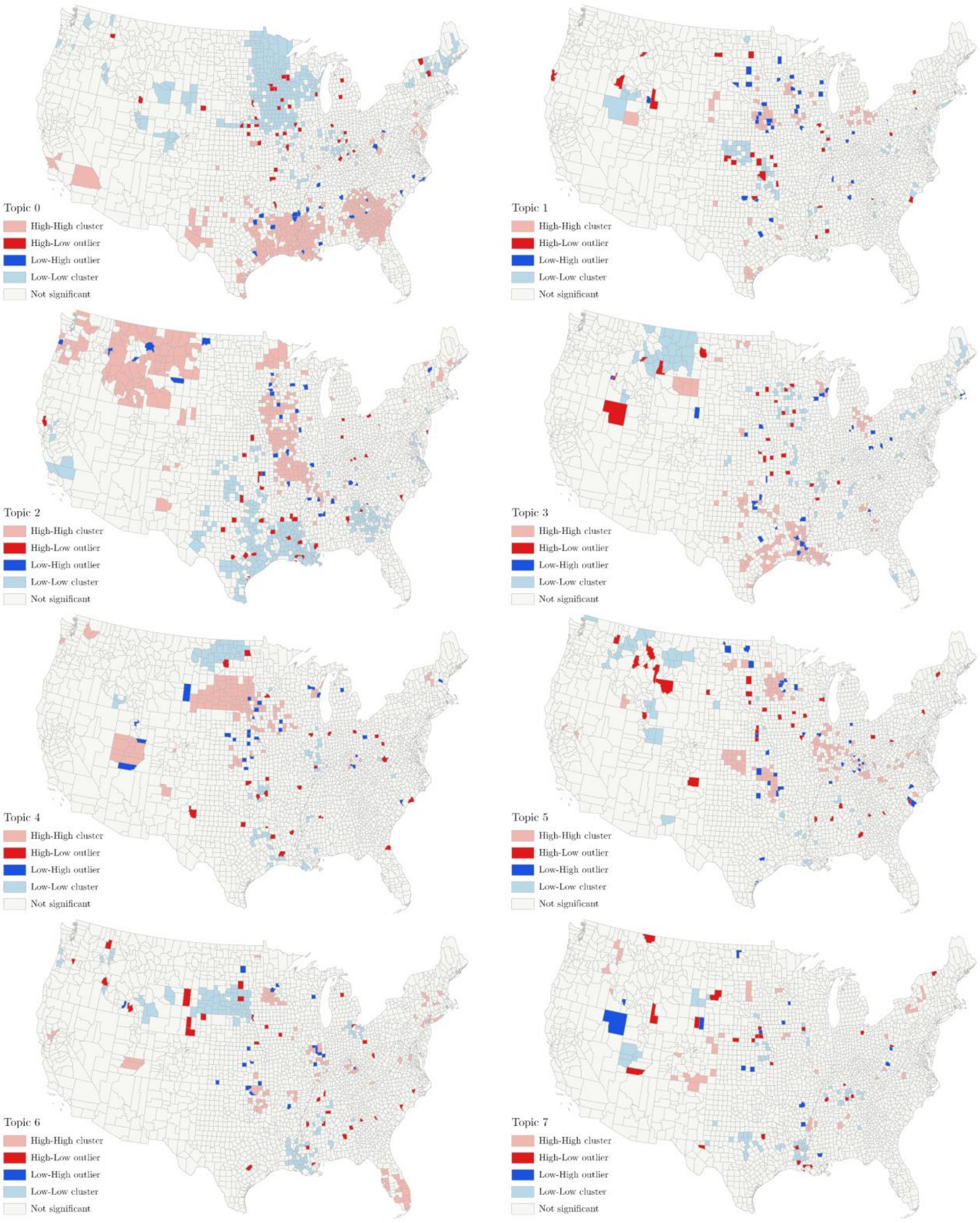

Figure 4: Overview of Local Moran's I clusters. Figure 13 through Figure 20 provide more detailed visualizations of the clustering of each topic.

Topic 0:

The general distribution of this topic shows a significant high-high cluster in Louisiana, Georgia, Mississippi, and eastern Texas. Isolated Low-High outliers exist in these regions, but the overall pattern is consistent within these states. Additionally, Low-Low clusters are seen in Minnesota, Iowa, Wisconsin, and New England, with isolated High-Low outliers in several counties.

Topic 1:

The spatial distribution of this topic appears less distinct and clustered than Topic 0; however, significant High-High and Low-Low clusters exist primarily in midwestern states e.g. Kansas and Nebraska.

Topic 2:

This topic shows an approximately reversed pattern compared to Topic 0; significant High-High clusters are seen in Missouri, Iowa, Montana, and Idaho, while significant Low-Low clusters appear in Georgia, Louisiana, and Texas.

Topic 3:

The distribution of this topic matches the pattern seen in Topic 0 in Louisiana and Texas. A significant Low-Low cluster is seen in Montana, with minor LL clusters in the Midwest.

Topic 4:

This topic shows a significant High-High cluster in South Dakota, Iowa, and Utah, with Low-Low hotspots in North Dakota, Illinois, Oklahoma, and Texas. Statistically significant outliers of both kinds exist scattered throughout the Great Plains and Midwest.

Topic 5:

The distribution of this topic shows High-High clustering in Minnesota, Illinois, Kentucky, Oklahoma, and parts of Virginia and West Virginia. Low-Low clusters are seen in Montana and Utah.

Topic 6:

The largest High-High clusters for this topic are found in Florida, Minnesota, Oklahoma, Illinois, and southern New England. Low-Low clusters encompass much of South Dakota, Louisiana, and northwestern Ohio and southeastern Michigan.

Topic 7:

Clusters for this topic are not as statistically significant as in other topics, but minor Low-Low hotspots exist in Louisiana, Tennessee, and portions of Texas and Utah.

In general, the topics with the largest and most continuous clusters are Topic 0 and Topic 2, which relate to COVID-19 misinformation and vaccine development, respectively. This likely reflects the highly controversial nature of the terms observed within these topics; much of social media discourse revolves around vaccine development and the COVID-19 pandemic (and related misinformation campaigns).

6 Discussion

6.1 Public Health

A key point of discussion regarding the results of this study is how to better plan for public health messaging campaigns. As shown in Figure 3 and Table 1, discussions regarding vaccinations, healthcare, and questioning the legitimacy of COVID-19 have unique spatial distributions. Organizations within communities should plan messaging campaigns around the topics discussed frequently (and those discussed infrequently), in order to better tailor these messaging campaigns around the needs of the communities they serve. For example, a community that is highly concerned with COVID-19 symptoms and school-place vaccine requirements (e.g., Topics 6 and 7) might consider implementing hybrid in-person/online school sessions.

Topics 0 and 3, notably, highlight the widespread public misunderstandings regarding COVID-19 and the lack of confidence in national public health messaging campaigns. Due to the aggregated nature of this data, caution must be exercised when performing correlation tests with e.g. demographic or voting pattern variables. Additionally, spatial autocorrelation may play a significant part in the patterns observed on the individual county level. However, the overall trends in the data, especially with Topics 0, 2, and 3, *generally* mimic both conservative/liberal and rural/urban patterns. Future studies may wish to analyze local spatial autocorrelation patterns to determine statistically significant hotspots of chosen topics.

6.2 Limitations

Due to rate limits from the Twitter API for Developers, a major time constraint in any full-archive Twitter search is the time needed to process an API query. Using the `academictwitteR` R package (or alternatives such as `twarc` for Python), it is possible to save the results of the query as it is processed. Approximately 300,000 Tweets per hour were collected, leading to a total query time of roughly 17 hours. As the query is rate-limited by the Twitter API based on the researcher's API key, it is not possible to circumvent these rate limits (and is a violation of the Twitter API terms of service); therefore, adequate planning is necessary to ensure enough time to process each query.

One limitation of this study is the bias towards urban population centers; areas with a higher population naturally have a larger density of Tweets, so the data must be normalized as a percentage of the total Tweets in an area. The prevalence of geolocated Tweets themselves is also an important consideration. As Twitter Place IDs appear more frequently in higher-population areas (as there are simply more buildings and locations in these areas), Tweets which are geolocated to a specific location appear less frequently in sparsely populated areas – i.e., a Tweet in a rural community may only be geolocated to the state level, as a Twitter-generated Place ID may not exist for all locations.

The Modifiable Areal Unit Problem (MAUP) factors into this study, as Twitter Place Objects exist at a range of scales (Fotheringham and Wong 1991). Geolocation information was aggregated to the county level as it was able to highlight general urban-rural patterns. An important methodological limitation in this study is the generalization of state-level Tweets to the county level; this study assumes that a state-level Tweet (e.g. a Tweet geolocated only in Texas) is representative of all counties within the state evenly, regardless of population. When summarizing Tweets to the county level, the mean percentage of the state-level Tweet is taken, weighted by the overlap between the state and each county – counties with a larger area receive a larger weight of the state-level Tweet average. State-level Tweets in this study show a uniform distribution across the counties within each state, and are primarily useful in highlighting patterns *across* states.

Social media data – in this case, geolocated Tweets – may not be represented of general public opinion due to several factors. As mentioned, the demographic of Twitter users differs compared to the average demographics of the general population (age, race, income, etc.) Additionally, the content of geolocated Tweets may differ from that of non-geolocated Tweets

— users may be less likely to report certain kinds of information, or author content that may be potentially identifiable, as there may be a higher perceived risk of self-identification when using geolocation information. Therefore, this study only intends to highlight the spatial distribution of geolocated Tweets, rather than the distribution of public perception in general.

A major constraint of this study is the ability to optimize Latent Dirichlet Allocation parameters. Given the size of the corpus, optimizing the number of topics, the document-topic prior α, the topic-word prior η, as well as other parameters within the LDA implementation in `sklearn`, time constraints and limitations on computational resources prevented ideal conditions for the different LDA parameters.

7 Conclusion

The results of this study offer insights into broad trends in the distribution of topics across the continental United States. The specific spatial distribution of each topic is unique ranges in statistical significance; some topics show more distinct clustering at larger scales, while others are less confined to a single region and instead appear roughly equally across the study area. Subsequently analyzing the Local Moran's I at each county provides statistically significant clusters and outliers at the 95% confidence level.

Performing Topic Modeling allows researchers to describe and interpret latent discussion topics within the corpus of social media data. Rather than manually coding each topic, this process allows for both efficient classification of discussion topics, as well as uncovering "hidden" topics within the discourse, which may not be accounted for with a simple keyword search. Additionally, the results of Topic Modeling (specifically, those from Latent Dirichlet Allocation) provide quantitative measurements of the distribution of topics within each Tweet;

these results, therefore, can be used to quantitatively model the spatial distribution of topics among geolocated Tweets.

By analyzing the spatial distribution of the resulting topics, general insights into the opinions within counties and communities can be determined. As Tweets may be geolocated at a variety of spatial scales, careful consideration must be taken to ensure any analysis is statistically rigorous and does not disproportionately bias one spatial scale over another.

References

Alghamdi, R., and K. Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6 (1).

Bao, S., S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*, 699–704. IEEE.

Barrie, C., and J. Ho. 2021. *academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint*. https://joss.theoj.org/papers/10.21105/joss.03272 (last accessed 1 September 2021).

Bella, T. 2020. “It affects virtually nobody”: Trump incorrectly claims COVID-19 isn’t a risk for young people. *The Washington Post*.

Bernheim, B. D., N. Buchmann, Z. Freitas-Groff, and S. Otero. 2020. The effects of large group meetings on the spread of COVID-19: the case of Trump rallies. In *Nina and Freitas-Groff, Zach and Otero, Sebastián, The Effects of Large Group Meetings on the Spread of COVID-19: The Case of Trump Rallies (October 30, 2020)*.

Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3:993–1022.

Boon-Itt, S., and Y. Skunkan. 2020. Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance* 6 (4):e21978.

Casola, A. R., B. Kunes, A. Cunningham, and R. J. Motley. 2021. Mask use during COVID-19: a social-ecological analysis. *Health Promotion Practice* 22 (2):152–155.

Clavel, C., and Z. Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing* 7 (1):74–93.

Cutchin, M. P. 2007. The need for the "new health geography" in epidemiologic studies of environment and health. *Health & place* 13 (3):725–742.

Dave, D. M., A. I. Friedson, K. Matsuzawa, D. McNichols, C. Redpath, and J. J. Sabia. 2020. *Risk aversion, offsetting community effects, and COVID-19: Evidence from an indoor political rally*. National Bureau of Economic Research.

Debortoli, S., O. Müller, I. Junglas, and J. vom Brocke. 2016. Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems* 39 (1):7.

Dubey, A. D. 2020. Twitter Sentiment Analysis during COVID-19 Outbreak. *SSRN*.

Dyer, O. 2020. Trump appointees tamper with renowned CDC publication, claiming that scientists are trying to "hurt the president." *BMJ* :m3589.

Fotheringham, A. S., and D. W. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A* 23 (7):1025–1044.

Garcia, K., and L. Berton. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing* 101:107057.

Geo objects. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo (last accessed 6 April 2022).

Ghosh, D., and R. Guha. 2013. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and geographic information science* 40 (2):90–102.

Gonsalves, G., and G. Yamey. 2020. Political interference in public health science during COVID-19. *bmj* 371.

Harris, R., P. Sleight, and R. Webber. 2005. *Geodemographics, GIS and neighbourhood targeting*. John Wiley and Sons.

Hill, C. A., E. Dean, and J. Murphy. 2013. *Social media, sociality, and survey research*. John Wiley & Sons.

Hoffman, M. 2010. Online variational Bayes for latent Dirichlet allocation (LDA). https://github.com/blei-lab/onlineldavb (last accessed 12 October 2022).

Hong, L., and B. D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, 80–88.

Isoaho, K., D. Gritsenko, and E. Mäkelä. 2021. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal* 49 (1):300–324.

Jaffe, S. 2020. Media reports reveal political interference at the US CDC. *The Lancet* 396 (10255):875.

Jones, T. 2019. *textmineR: Functions for Text Mining and Topic Modeling*. https://www.rtextminer.com.

Kearns, R., and D. Collins. 2010. Health geography. *A companion to health and medical geography* :15–32.

Klašnja, M., P. Barberá, N. Beauchamp, J. Nagler, and J. Tucker. 2017. Measuring public opinion with social media data. In *The Oxford handbook of polling and survey methods*. New York: Oxford University Press.

Litosseliti, L. 2010. *Research methods in linguistics*. London; New York: Continuum. http://www.myilibrary.com?id=851442 (last accessed 6 April 2022).

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5 (1):1–167.

Madden, M., A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. 2013. Teens, social media, and privacy. *Pew Research Center* 21 (1055):2–86.

Manguri, K. H., R. N. Ramadhan, and P. R. M. Amin. 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research* :54–65.

McGregor, S. C. 2019. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism* 20 (8):1070–1086.

Mei, Q., D. Cai, D. Zhang, and C. Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, 101–110.

Mejova, Y. 2009. Sentiment analysis: An overview. https://www.researchgate.net/publication/264840229_Sentiment_Analysis_An_Overview.

Murphy, J., M. W. Link, J. H. Childs, C. L. Tesfaye, E. Dean, M. Stern, J. Pasek, J. Cohen, M. Callegaro, and P. Harwood. 2014. Social media in public opinion research: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly* 78 (4):788–794.

Neubaum, G., and N. C. Krämer. 2017. Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media psychology* 20 (3):502–531.

Ostrowski, D. A. 2015. Using latent Dirichlet allocation for topic modelling in twitter. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 493–497. IEEE.

Pak, A., and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, 1320–1326.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Petersen, J., M. Gibin, P. Longley, P. Mateos, P. Atkinson, and D. Ashby. 2011. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems* 13 (2):173–192.

Piller, C. 2020. Undermining CDC. *Science* 370 (6515):394–399.

Rashid, J., S. M. A. Shah, and A. Irtaza. 2019. Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management* 56 (6):102060.

Rosenberg, M. 2017. Health geography III: Old ideas, new ideas or new determinisms? *Progress in Human Geography* 41 (6):832–842.

Short Text Classification. *MonkeyLearn*. https://monkeylearn.com/short-text-classification/ (last accessed 11 April 2022).

Singleton, A. D., and P. A. Longley. 2009. Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science* 88 (3):643–666.

Twitter API Documentation. 2021. *Twitter Developer Platform*. https://developer.twitter.com/en/docs/twitter-api (last accessed 19 November 2021).

Wallach, H. M. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 977–984. Pittsburgh, Pennsylvania: ACM Press http://portal.acm.org/citation.cfm?doid=1143844.1143967 (last accessed 2 December 2021).

Figures

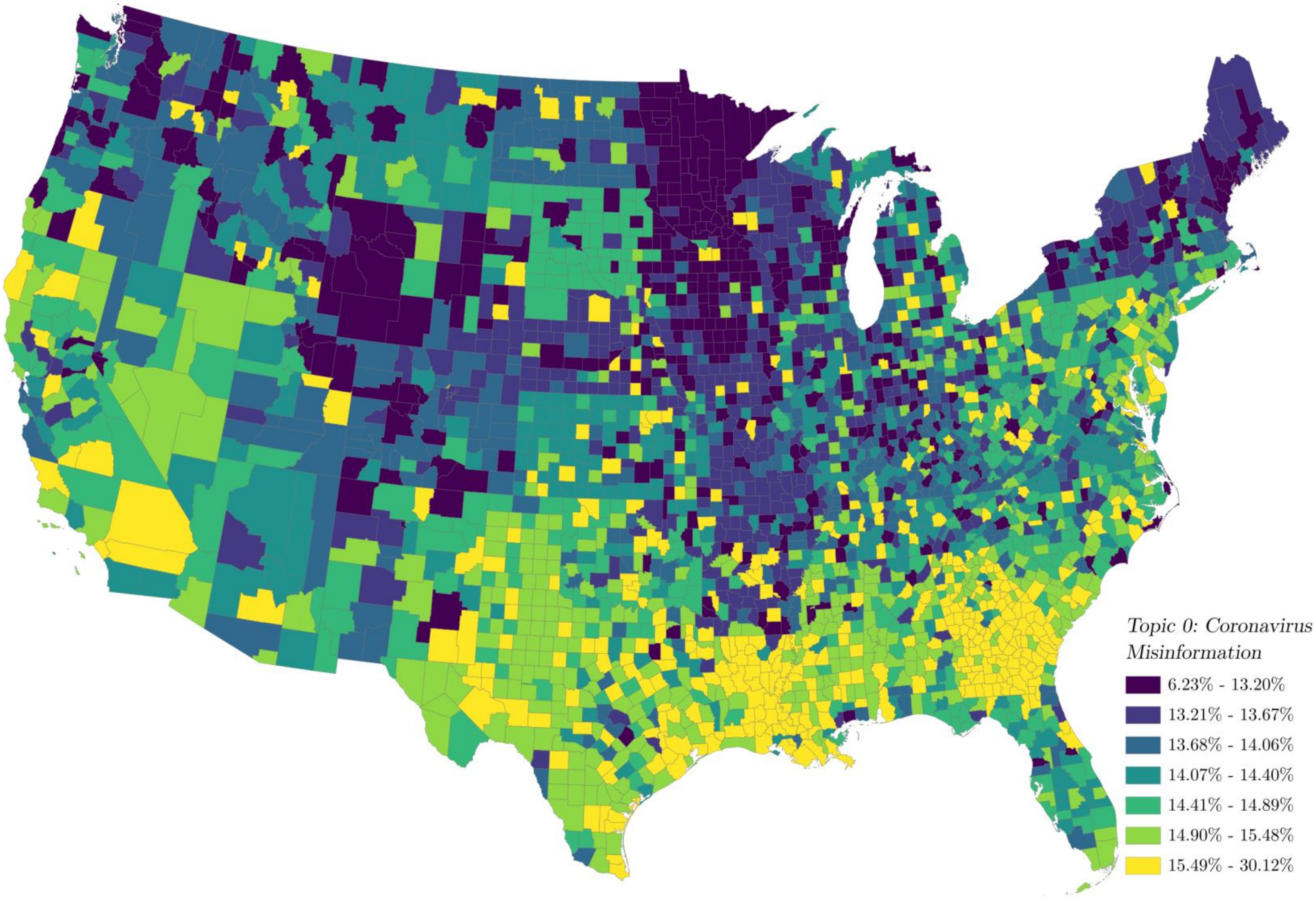

Figure 5: Spatial distribution of Topic 0 (Coronavirus Misinformation, Hoaxes, "China Virus")

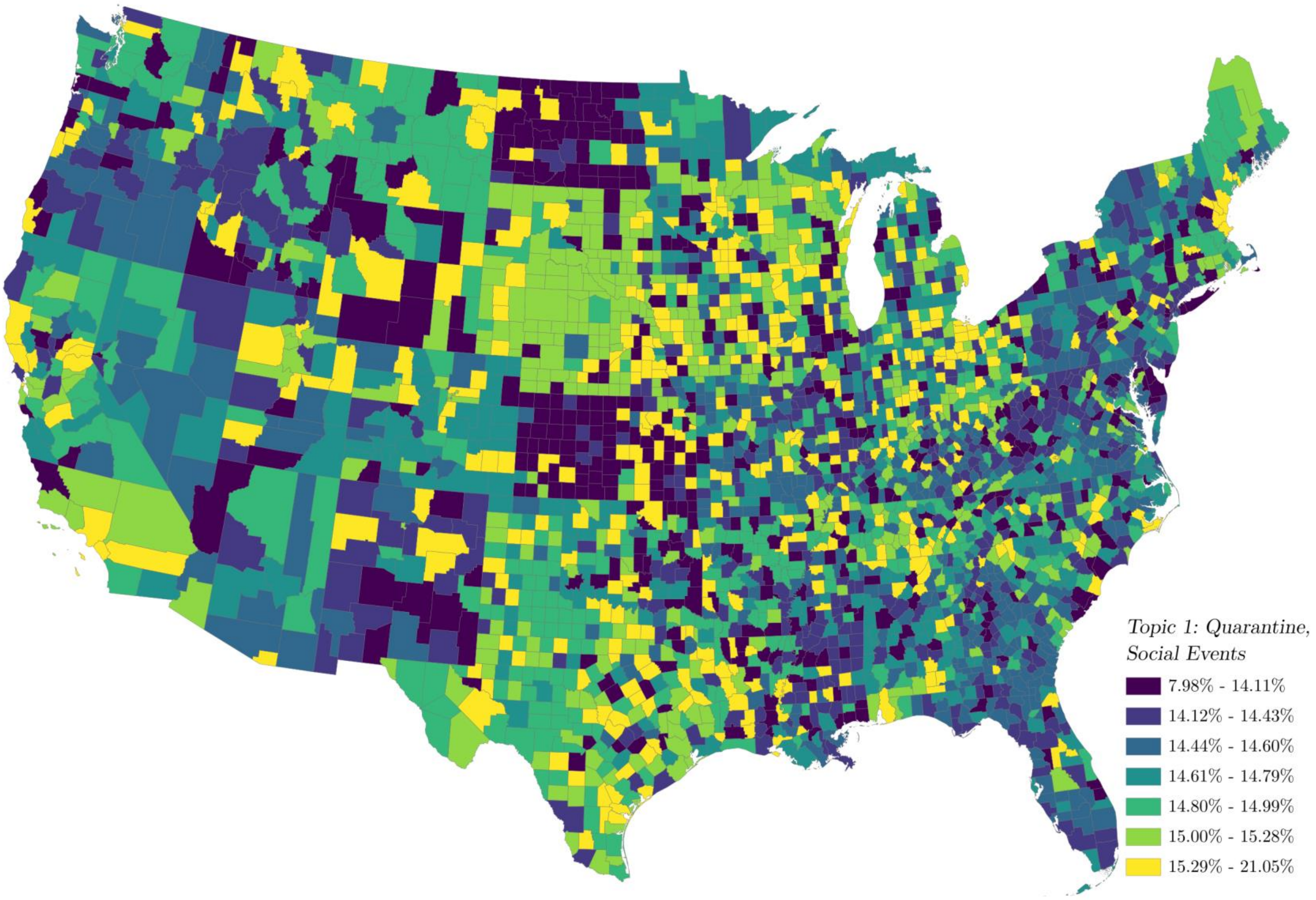

Figure 6: Spatial distribution of Topic 1 (Quarantine, Shelter-in-Place, Missing out on Social Events)

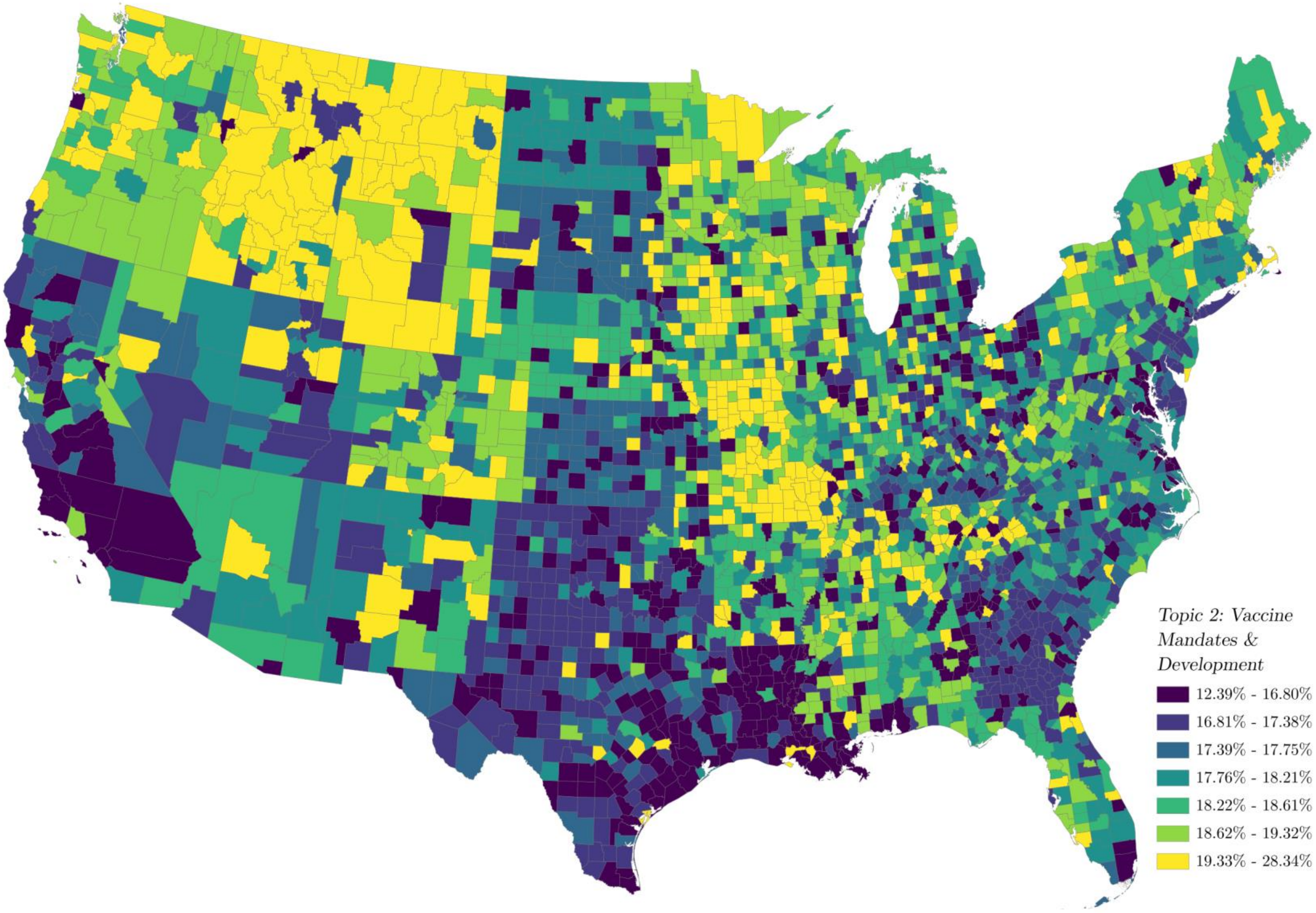

Figure 7: Spatial distribution of Topic 2 (Vaccines, Vaccine Development, Vaccine Aversion, Vaccine Mandates)

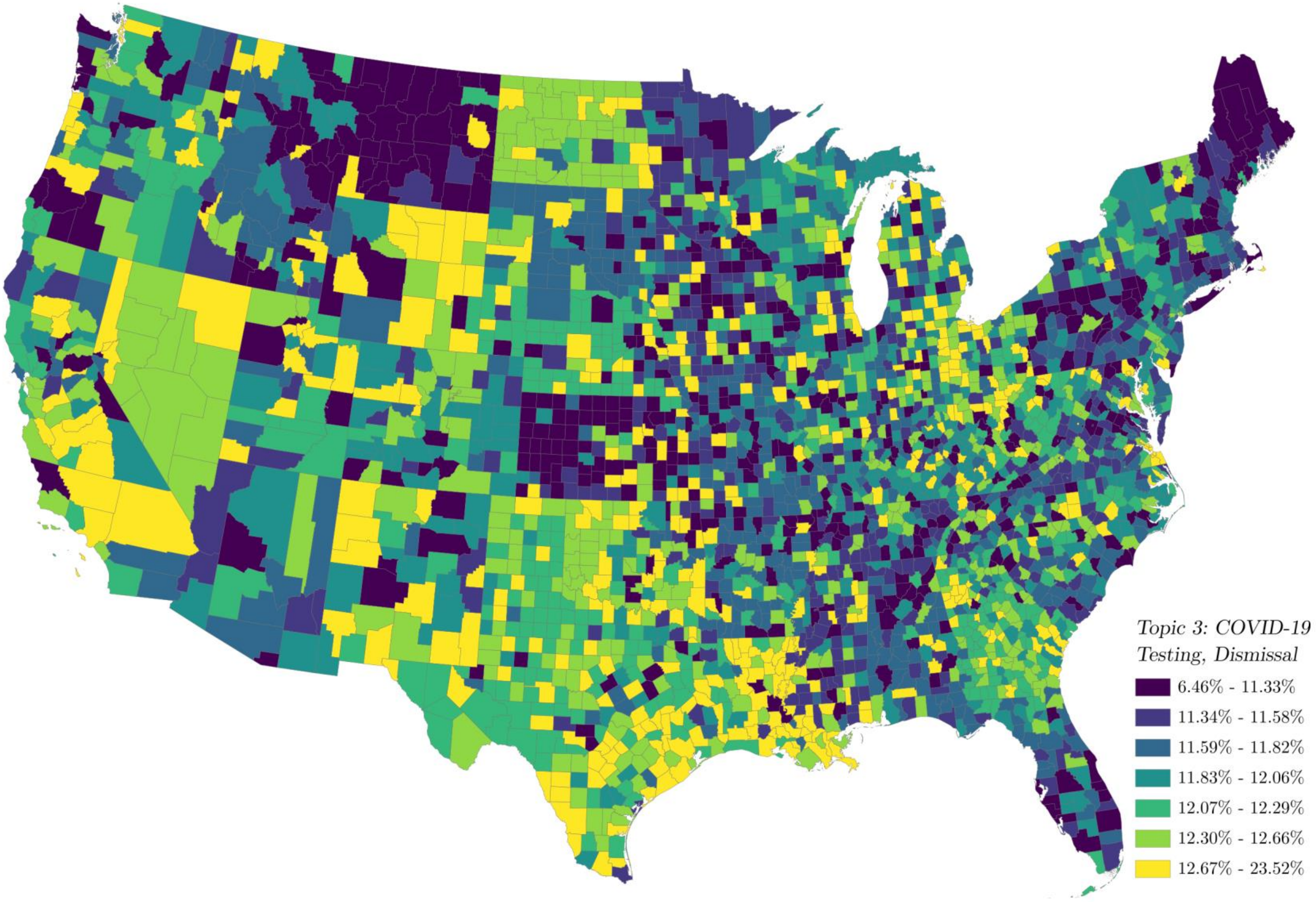

Figure 8: Spatial Distribution of Topic 3 (COVID-19 Testing, COVID-19 Dismissal)

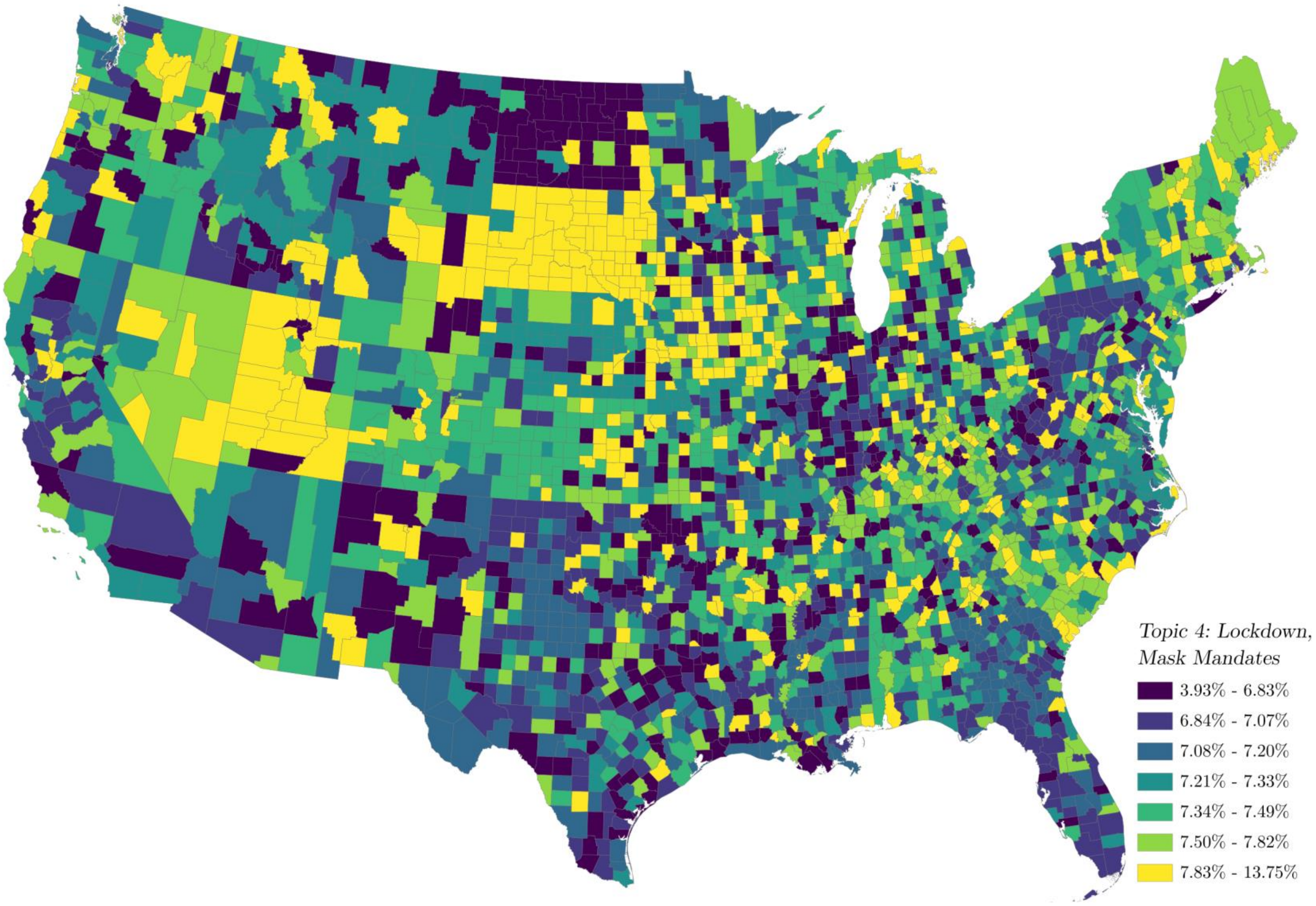

Figure 9: Spatial distribution of Topic 4 (Lockdown, Shelter-in-Place, Mask Mandates)

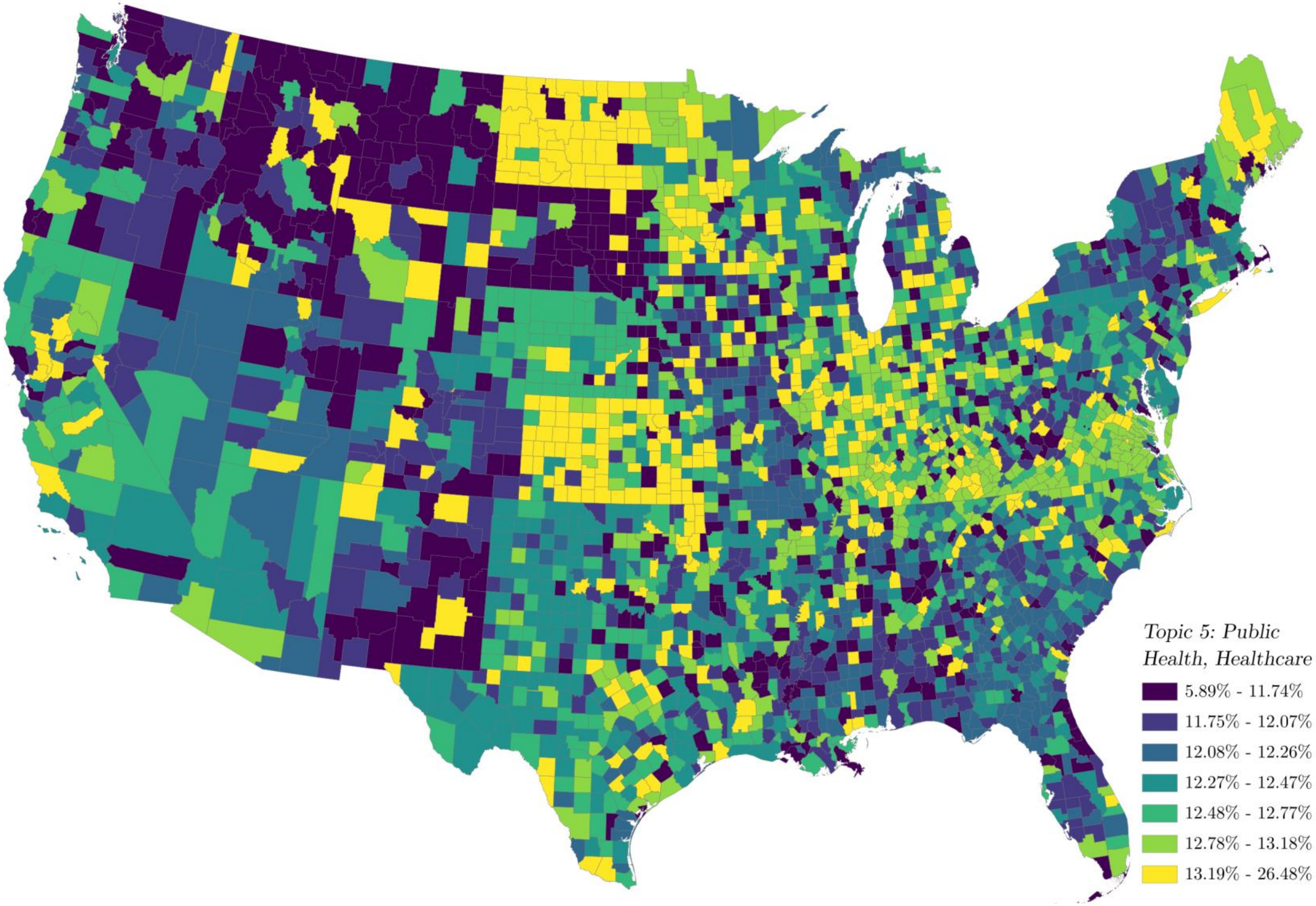

Figure 10: Spatial distribution of Topic 5 (Public Health, Healthcare, Hospital Workers)

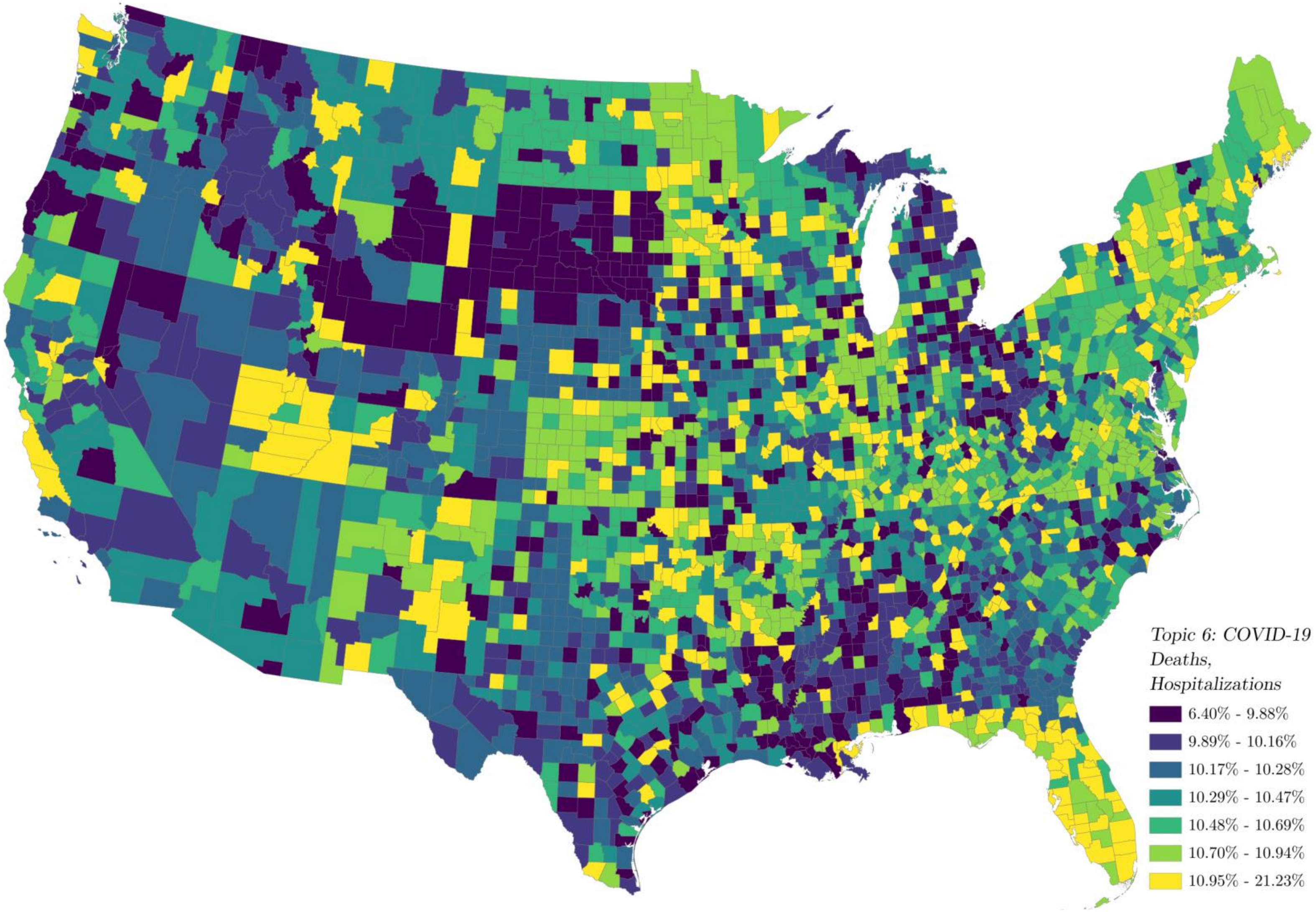

Figure 11: Spatial distribution of Topic 6 (COVID-19 Deaths, Hospitalizations)

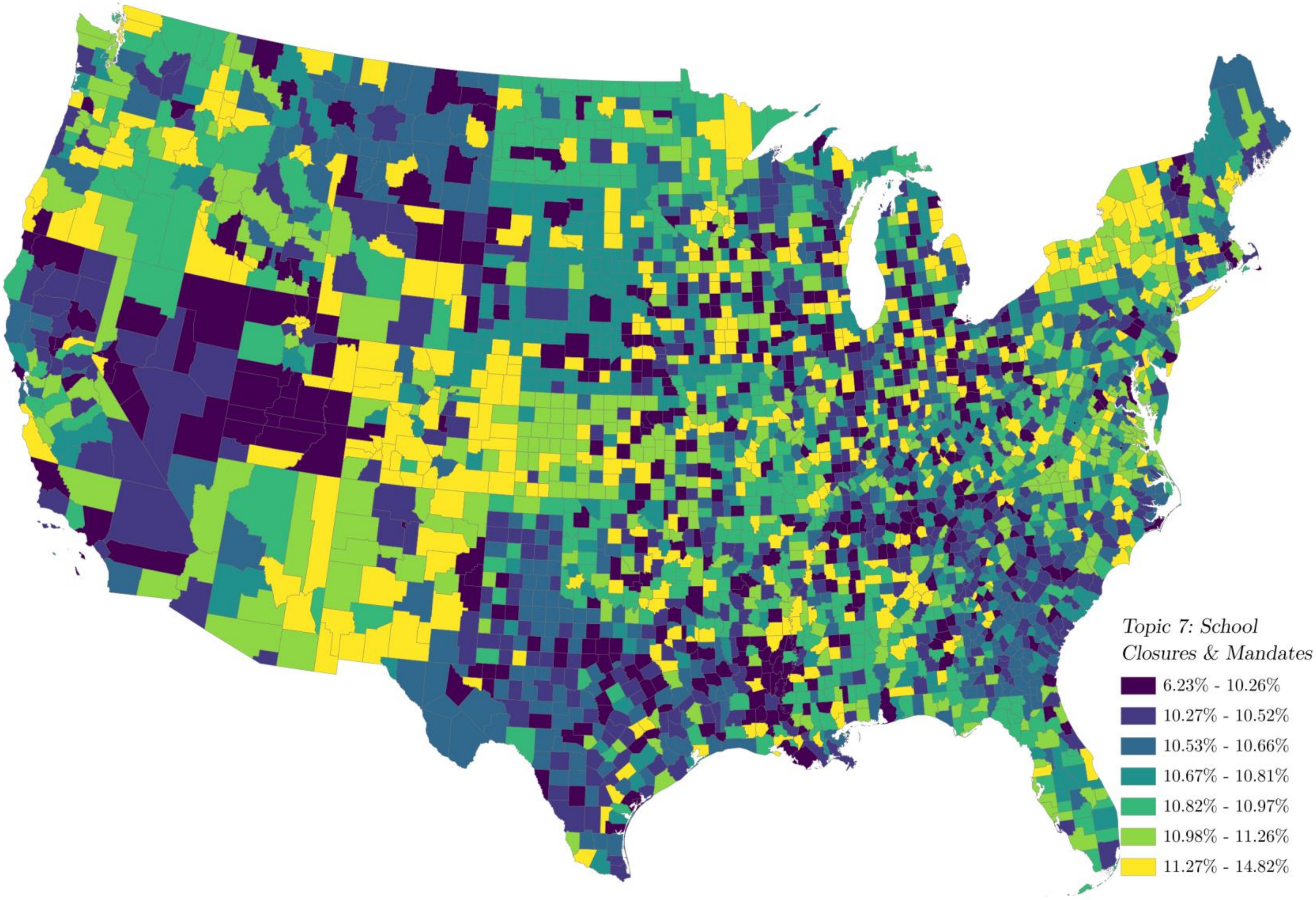

Figure 12: Spatial distribution of Topic 7 (School Closures, School Mask & Vaccine Mandates)

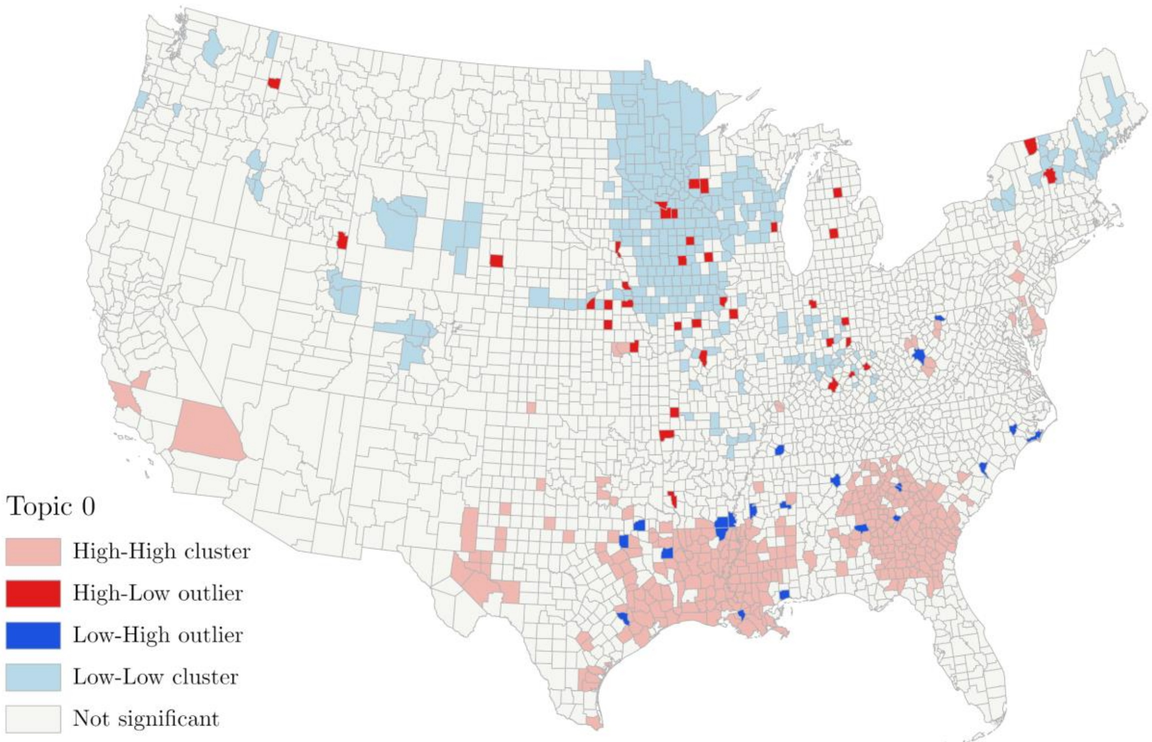

Figure 13: Results of Local Moran's I for Topic 0

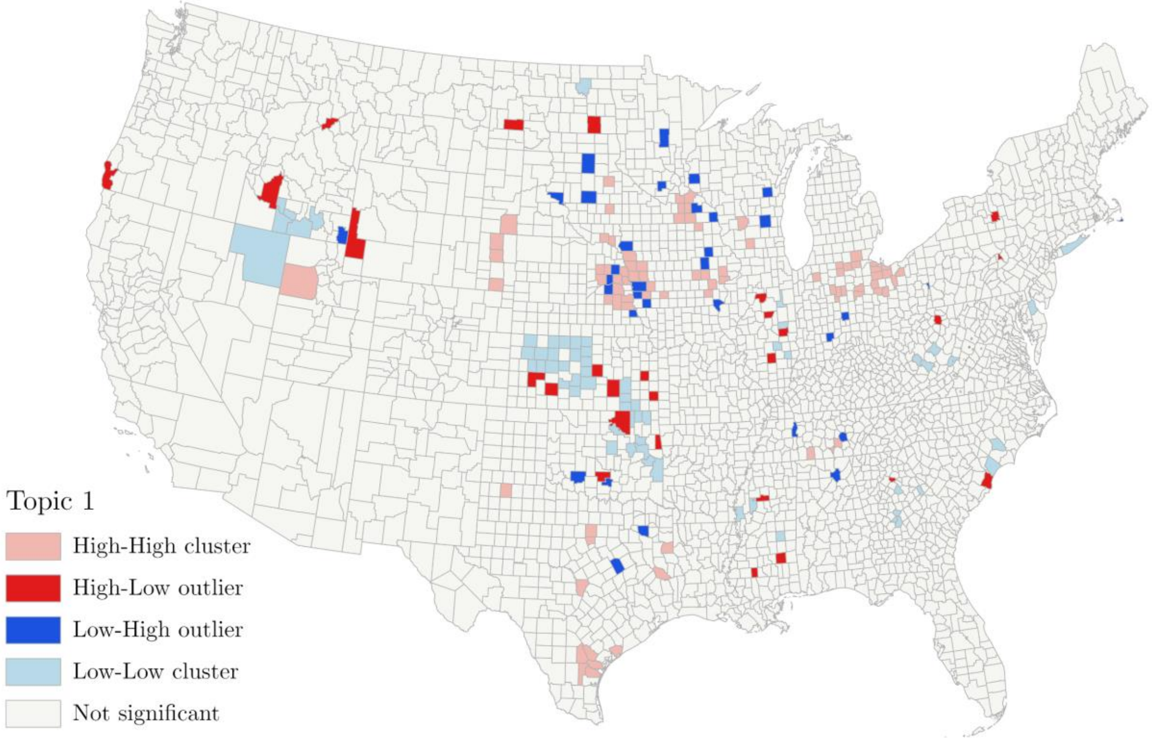

Figure 14: Results of Local Moran's I for Topic 1

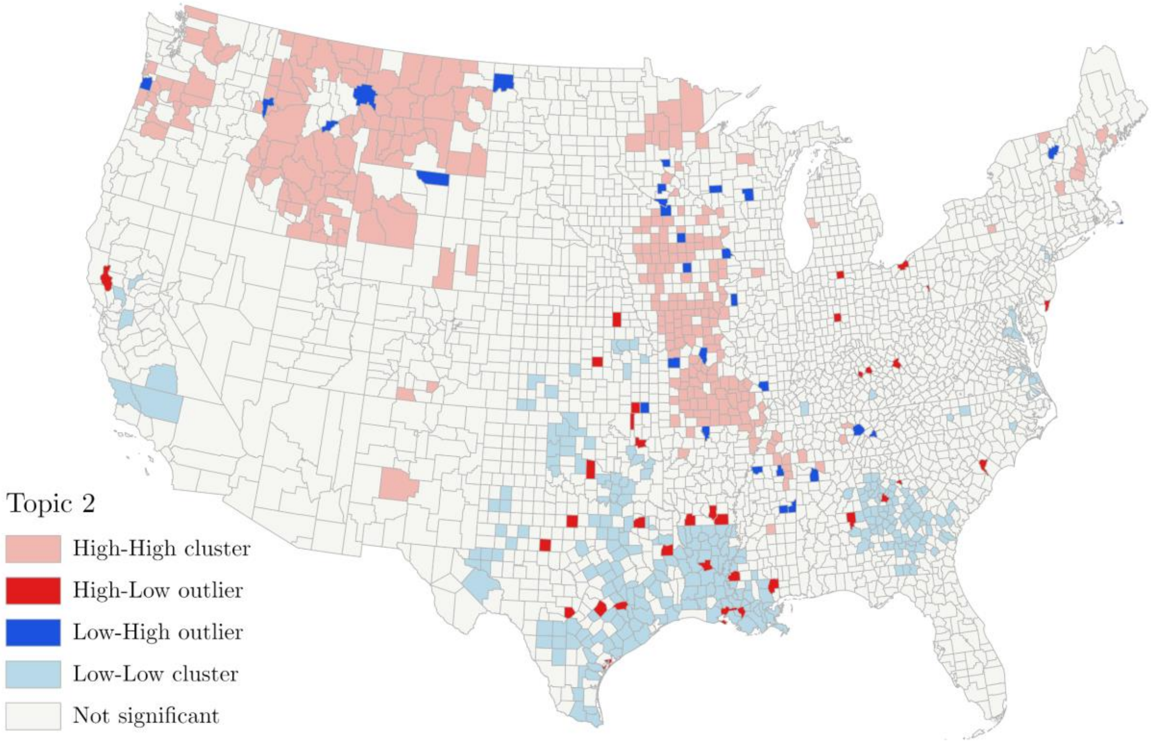

Figure 15: Results of Local Moran's I for Topic 2

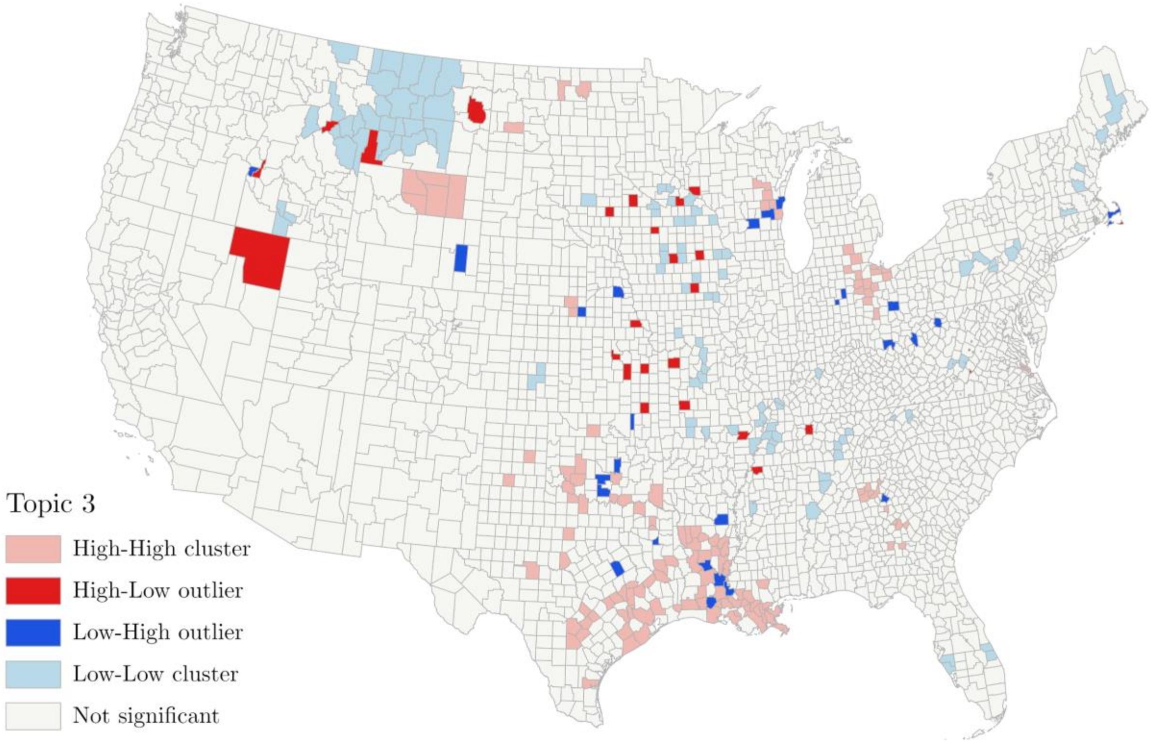

Figure 16: Results of Local Moran's I for Topic 3

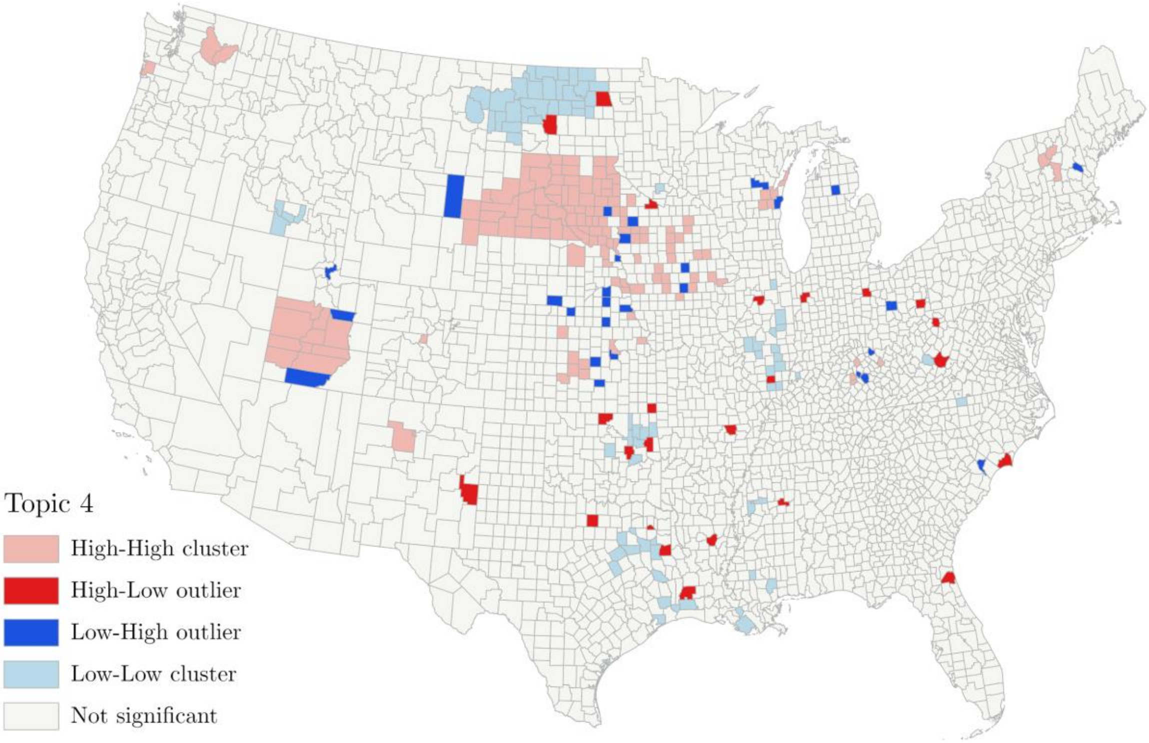

Figure 17: Results of Local Moran's I for Topic 4

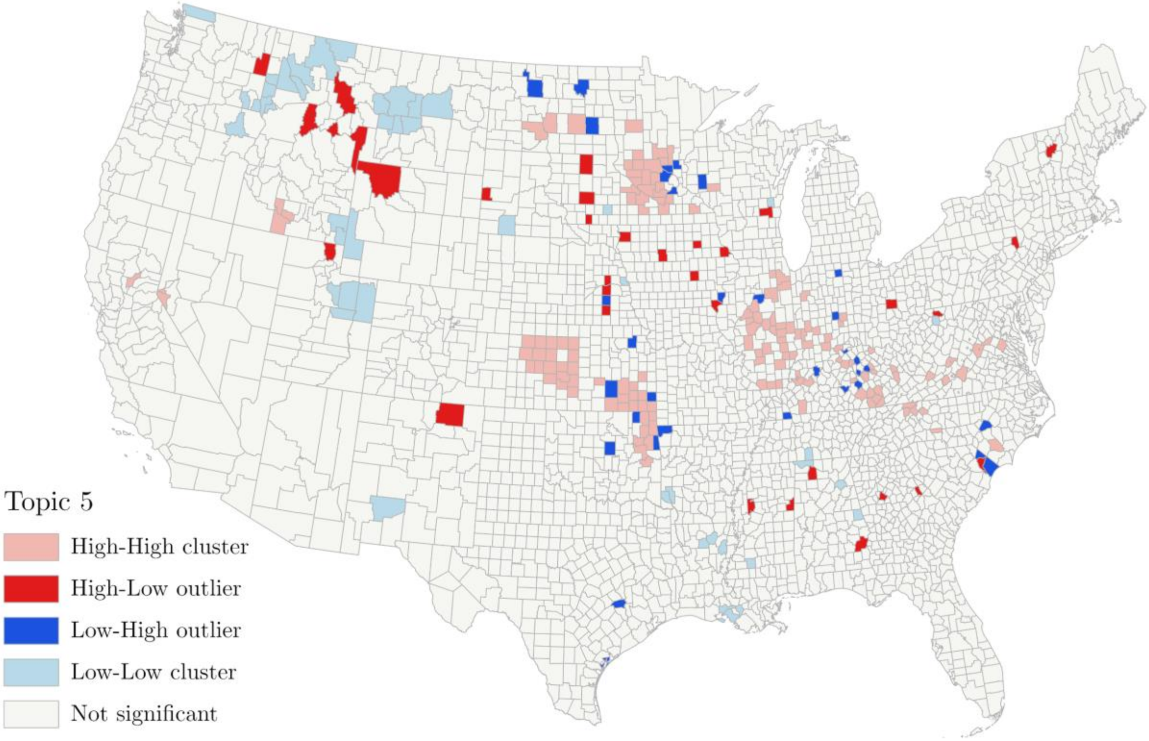

Figure 18: Results of Local Moran's I for Topic 5

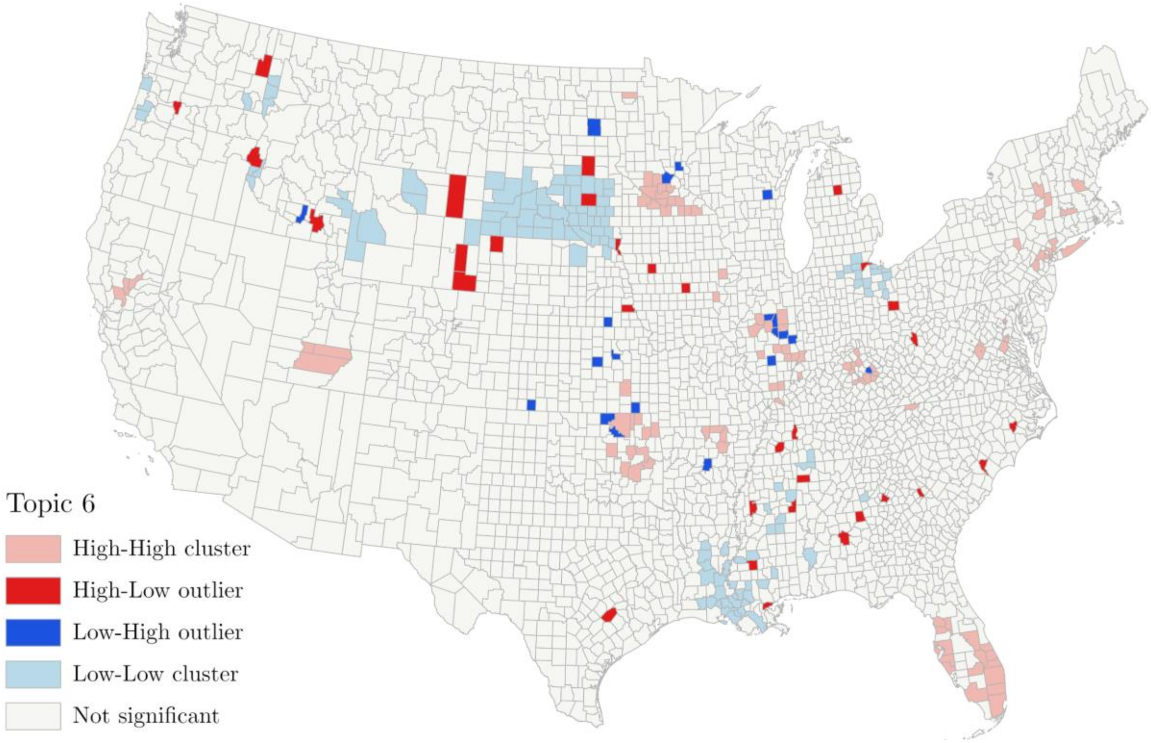

Figure 19: Results of Local Moran's I for Topic 6

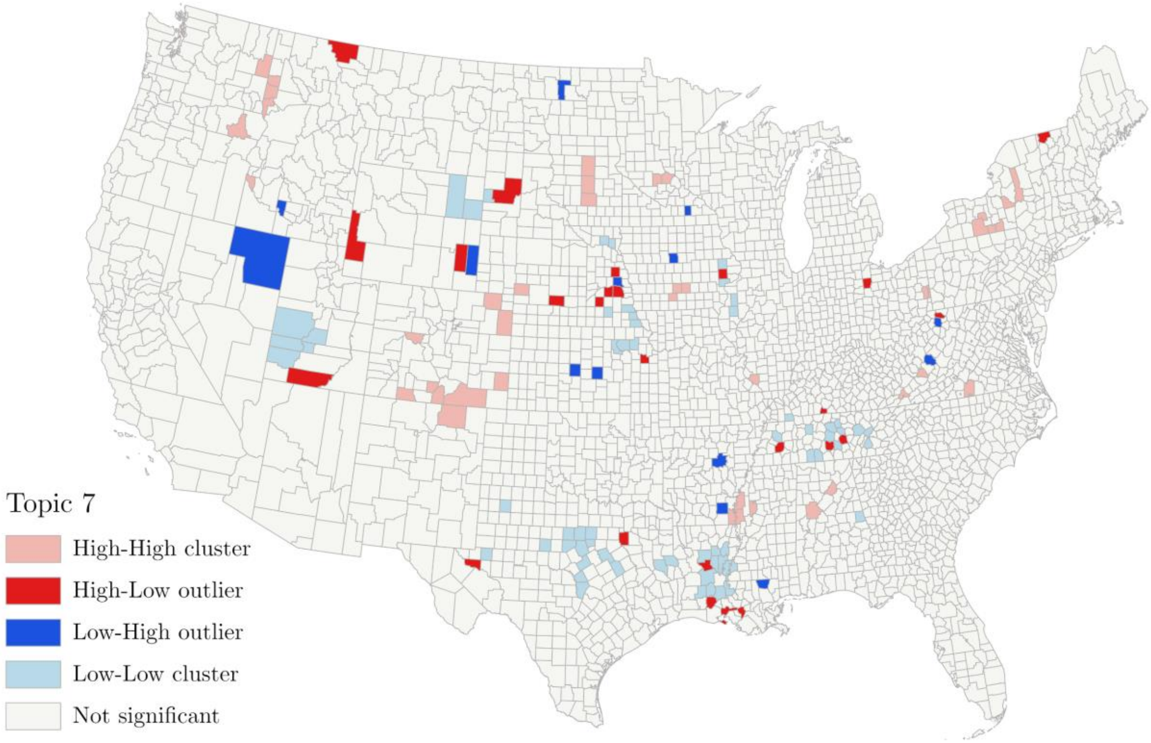

Figure 20: Results of Local Moran's I for Topic 7

Vita

Harrison Brown was born in Beaufort, South Carolina and moved to Boone, North Carolina at an early age. He attended Appalachian State University, where he graduated May 2020 with a Bachelor of Arts in Geography and Planning, and a Bachelor of Arts in German, from the Department of Languages, Literatures, and Cultures. In Spring, 2021, he was accepted into the master’s program at Appalachian State University, Department of Geography and Planning. His current research focuses on the intersection between social media data and human/political geography.

During his career as a master’s student, Harrison fostered skills in the R and Python programming languages, where he developed the necessary tools to perform the analysis portion of his thesis. In addition, he has worked as a graduate assistant for many GIS- and Remote Sensing-focused courses within the department, allowing him to gain skills in both GIS analysis, course instruction, and public speaking.

In December 2022, Harrison received his Master of Arts in Geography and Planning from Appalachian State University.