A Data Mining Framework for Analysing Geospatial-Temporal Data

A Thesis

Presented to

The Faculty of the College of Graduate Studies

Lamar University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Raunak Sarbajna

December 2019

A Data Mining Framework for Analysing Geospatial-Temporal Data

RAUNAK SARBAJNA

Approved:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name

Sujing Wang

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name

Stefan Andrei

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name

Committee Member

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Stefan Andrei

Chair, Department of Computer Science

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Lynn M. Maurer,

Dean, College of Arts and Sciences

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

William E. Harn

Dean, College of Graduate Studies

©2019 Raunak Sarbajna

No part of this work may be reproduced without permission except as indicated by the “Fair Use” clause of the copyright law. Passages, image, or ideas taken from this work must be properly credited in any written or published materials.

ABSTRACT

A Data Mining Framework for Analysing Geospatial-Temporal Data

by

Raunak Sarbajna

Change analysis and automated storytelling are essential techniques in understanding patterns and trends in multifaceted, time-series geospatial-temporal data. In this thesis, we introduce a novel data mining framework for detecting and analysing how the patterns of features change over time and space in spatio-temporal data and automatically generate interesting stories based on a variety of unique factors. We discuss and develop techniques to generate appropriate spatial polygons from raw georeferenced point data while preserving its meta data and shape. Our framework utilizes appropriate change predicates that are data agnostic and can capture the widest possible polygon movements. We then develop and use new measures of interestingness to aid in generating automated stories based on the change analysis results. Our change monitoring framework then creates a change graph that captures the changes in the spatial clusters and our change summarization framework creates specific change summaries based on the change graph types. We evaluate the effectiveness of our framework through case studies involving 2016 Twitter data with sentiment scores attached and 2017 drought data from North America. The experimental results show that our framework can discover interesting change patterns and useful information from spatial-temporal data.

ACKNOWLEDGEMENTS

This research is supported by Lamar University Research Enhancement Grant.

Table of Contents

List of Tables v

List of Figures vi

Chapter Page

1 Introduction 1

2 Literature Review 5

3 Methodology 11

Change Analysis Predicates 11

Polygon Generation from Point Data Sources 12

Change Analysis for Polygonal Map Data Sources 14

Storytelling Techniques for Geospatial Data 17

4 Case Study 19

Data Sets 19

Dataset Preprocessing 22

Drought Data Sources 27

Emotion Spatial Clusters from Twitter 32

Analysis 34

5 Conclusion and Future Work 36

References

Appendices

Appendix A

Appendix B

List of Tables

Table Page

Table 1. Layer Specification for Drought Data 21

Table 2. Layer Specification for Twitter Data 21

List of Figures

Figure Page

Figure 1. Change Analysis framework Architecture 16

Figure 2. Spatial reference for Twitter Data 19

Figure 3. Spatial reference for Drought Data 20

Figure 4. United States Drought Monitor Data 20

Figure 5. Twitter shapefile data inserted into Postgres db 22

Figure 6. Emotion polygons created from June 5 Twitter data 25

Figure 7. Emotion polygons created from June 6 Twitter data 25

Figure 8. Emotion polygons created from June 7 Twitter data 26

Figure 9. Emotion polygons created from June 1 Twitter data

Using Convex Hulls 26

Figure 10. Continental USA original monthly drought polygon

distribution for 2017 27

Figure 11. Increase in Drought Affected Regions in Texas

following Hurricane Harvey 29

Figure 12. Decrease in Drought Affected Regions in Texas

following Hurricane Harvey 30

Figure 13. Shifting areas of drought affected regions in California

following the 2017 wildfires 32

Figure 14. Change in Twitter emotion polygons from June 5 to June 6 33

Figure 15. Change in Twitter emotion polygons from June 7 to June 7 34

Chapter 1

Introduction

Analyzing change in spatial-temporal data is critical for many applications including developing early warning systems that monitor environmental conditions, detecting political unrest and crime monitoring. Change analysis models are essential in understanding larger patterns and trends in multifaceted, time-series geographic data. The purpose of this study is to detect spatio-temporal changes within sequential (time-series) geospatial-temporal data.

Analytical methods of spatial and spatio-temoral have developed considerable recently. Most of the newer techniques capture only spatial correlation or covariance structures over polygonal lattice data, and there are few methods involving polygon, point and line spatial data. For most remotely sensed imagery, the phenomena are represented as vector points, lines, polygonal models or raster surfaces. Polygonal models classify objects, while rasters classify fields. However, many situations cannot be classified and mapped using such categories, because they contain elements of both, such as, wildfires, hurricanes, droughts or ecological hotspots. Methods for characterizing space-time change in such phenomena are currently poorly developed.

Spatial relationships between multiple polygons are not very well defined. The most widely used models of spatial relations is the point set topology based 9-intersection model by Egenhofer and Herring. The model has been further refined by Chen et al. using Voronoi diagrams of spatial objects. The modern approach combines the 9-intersection model and metric parameters, as proposed by Liu and Deng. However, the complexity of spatial relationships between polygons makes characterizing them using topological relations defined by point-set topology a very challenging analytical and computational task.

Detecting changes in land-use/land-cover is one of the most fundamental and common uses of remote sensing image analysis. One of the most rudimentary forms of change detection is the visual comparison of two images by a trained interpreter. With an effective display system large enough to display both images simultaneously and to explore and digitize with a cursor tracking to the same location in both images, this is a quick method that can be used to locally collect valuable GIS compatible data while streaming the images themselves over a relatively low-bandwidth Internet connection.

Modern technology digitizes wide sources of information constantly, with hour after hour of data being stored and most of it being unprocessed raw data. This is especially true with remote sensing networks and resultant geo-spatial imagery. Much of this data has meta-information associated with it which can reveal patterns or trend if properly classified and analyzed. Performing basic change analysis on a subset of this data gives us a lot of insights that would otherwise go unrecognized.

The basic metrics of spatial relations between moving polygons are useful to describe changes in edges or centroids. Over a larger time period, changes in distance (related to velocity) and direction combine to describe the underlying phenomena with some degree of certainty. However, an issue arises in quantifying distance and directional relationships where polygons both shift and change in size. This is where polygons as objects and rasters as fields are combined to get finer metrics for both. In this case, most real world polygons are attached metadata that serve the same purpose.

Digital algorithms also exist for change detection. Unclassified images can be compared on a pixel-by-pixel or patch-by-patch basis; classified images can be compared with the results indicating changes in specific classes over time. Visually comparing co-registered images from two dates is always the first place to start, even if the goal is to use an automated algorithm for classification or change detection Most image processing packages include tools to swipe one image over the other, flicker between images, and view images side-by-side. In some cases, heads-up digitizing may be used to identify and classify change; in other cases, visual inspection is used to help select the most appropriate automated change detection technique.

Generating actionable intelligence from geospatial data requires change detection capabilities that are beyond human grasp. Whether we are looking to detect changes in the number of parked cars at retail outlets, changes to buildings, or changes in the position of a satellite dish one is often looking for changes that are practically impossible for the human eye to detect. Even when detected changes can be observed by the human eye; given the immense volume of geospatial data we simply cannot hire enough people to stare at video, LiDAR, and satellite imagery monitoring changes. Smart and efficient Change Detection software is crucial.

A common understanding is that most big data available today is either archival, media or web scrapes [25]. However, a large source of that data is from Geographic Information Systems (GIS), and the tools available to interpret them easily are lacking [26]. The multiplicity of APIs has standardized access and structuring, but they limit much of the meta-data associated with them. Most publicly available (i.e. non-governmental) 'big data' sources with spatial components revolve around data scraped from mobile software platforms, including twitter, Instagram, Snapchat and reviews on mapping apps [26]. The recent API changes by Google Maps, for instance, aptly show that data content, and thus its meaning, is subject to regulation that is outside the control of researchers. The goal of this research project is to detect and analyze how the patterns of features change over time and space in spatio-temporal data and then automatically generate stories based on the sequential spatio-temporal data.

Our approach provides a change monitoring framework which creates a change graph that captures the changes in spatial land uses clusters and a change summarization framework that creates specific change summaries based on the change graph based on the change story types.

Our research contributions are summarized as follows:

1. A novel change analysis framework that works on vectored polygonal datasets
2. New change predicates that are data agnostic and can work on a large spectrum of data
3. A new measure of interestingness to aid in generating automated storytelling based the change analysis results

The rest if the thesis is structured as follows. Chapter 2 reviews previous literature on the subject and discusses related work. Chapter 3 introduces our data mining framework and lays out the methodology in detail. Chapter 4 evaluates the framework with case studies on drought datasets pre and post Harvey, pre and post California wildfires and twitter emotion polygons. Chapter 5 provides a conclusion and discusses potential future expansions to the framework.

Chapter 2

Literature Review

Liu and Deng have created the model for modern metric based spatial relationships between polygons. They updated the original 9-intersection model by Egenhofer and Herring, elimination limitations. Liu and Deng’s metric spaces avoid problems with the immensity of the spatial object complement operation, restricts sensitivity of boundary data errors and makes representing dynamic topological relations easier. They introduce geometrical metric parameters of point sets, such as their distance, area and the circumference. This is divided into segmentation parameters and nearness parameters. The finals steps are (1) to find and locate spatial objects, (2) to make a approximate classification for spatial relations between a selected reference object and a target one using the 9-intersection model, and (3) to identify the spatial relations in more detail by applying metric parameters.

Robertson et al. coined the term moving polygons for spatio temporal polygon data (STAMP). They developed the original work by Sadahiro and Umemura and improved upon the original event based framework. They first grouped polygons based on factors such as distance thresholds and intersection points. These groups are then evaluated according to three event types and in four hierarchical levels. The typology of events to describe these geometric changes include Generation, Disappearance, Expansion and Contraction. These are modified as needed according to the dataset, with movement events such as Displacement, Convergence, Fragmentation and Divergence.

A survey of the classical change detection algorithms can be found in the Lu et al. [3] paper and tells us that the integrated GIS and remote sensing approaches yield the best results. However, they are very sensitive to registration accuracies between images. Thus, images must be properly orthorectified and georeferenced, especially because the changes in the emotion polygons are so subtle. This assumes the emotions are to be treated as just another feature in the map, like any other category.

The ability to develop spatially distributed models of topographic change is presenting new capabilities in geomorphic research, as seen in James et al. High resolution maps of elevation change indicate locations, processes, and rates of geomorphic change, and provide a means of calibrating temporal simulation models. Methods of geomorphic change detection (GCD), based on gridded models, may be applied to a wide range of time periods by utilizing cartometric, remote sensing, or ground-based topographic survey data to measure volumetric change. Advantages and limitations of historical DEM reconstruction methods are reviewed with a focus on coupling them with subsequent DEMs to construct DEMs of difference (DoD), which can be created by subtracting one elevation model from another, to map erosion, deposition, and volumetric change.

Since our data is primarily in an urban environment, with all the grid like rigidity that entails, it is a good idea to look at change detection algorithms optimized for urban environments. One of the hardest aspects to measure is to distinguish between change and no-change, as well as different kinds of change. Comparing image differencing, image regression, tasselled-cap transformation and chi square transformation, Ridd and Liu [3] find image differencing to be the most consistent, with a sustained overall accuracy of >80%.

It is useful to have a programming-oriented study comparing several of the change detection algorithms using MATLAB, rather than pure application-oriented comparison, in order to have a benchmark. Minu and Shetty [5] analyzed image differencing, image ratioing, change vector analysis, tasseled cap transformation and principal component analysis for efficiency and effectiveness. Although their area of study was not urban but a variety of land use/ land cover, change vector analysis gave the best overall accuracy.

We also studied two novel methods that are recent developments and are showing promising results: Neighborhood Correlation Image and Comprehensive Change Detection Method, both of which are optimized for remote sensing imagery but can be adapted to vectorized maps without loss of generality.

The change detection model using Neighborhood Correlation Image (NCI) logic works because of the obvious fact that the same geographic area (e.g., a 3x3 pixel window) on two dates of imagery will tend to be highly correlated if little change has occurred, and uncorrelated when change occurs [1]. Computing the piecewise correlation between two data sets demonstrates that NCIs contain change information and that NCIs may be powerful tools for change detection.

A high-performance remote sensing method called Comprehensive Change Detection Method (CCDM) integrates spectral-based change detection algorithms and a novel change model called Zone, which extracts change information from two Landsat image pairs [2]. This can be easily modified to work on the Twitter-based emotional grading maps. This method is simple, easy to operate, widely applicable, and capable of capturing anthropogenic changes like our area of interest.

Storytelling techniques are an effective summarization method to succinctly organize extensive information. Traditional storytelling has been mostly successful on news articles, blogs, as well as structured databases. However, traditional storytelling techniques tend to perform poorly on social media content, such as Twitter, where text lacks proper form and function [11]. Moreover, the ability to support dynamic storylines as they evolve is critical to modelling fast moving social media streams such as Twitter. Dos Santos et al. [21] introduced a set of methods to automatically derive stories over linked entities in tweets. They model a story as a graph of entities propagating through spatial regions in a temporal sequence, and controls search space complexity by suggesting regions of exploration. They developed algorithms to conduct storytelling to model tweets over space and time, reasoning over spatio-temporal features, and devise spatio-temporal storylines based on connectivity strength.

Kumar et al. [14] proposed an efficient storytelling implementation that embeds the CARTwheels [15] redescription mining algorithm which utilizes induced classification trees to model redescriptions in an A\* search procedure, using the CARTwheels to supply next move operators on search branches to the A\* search procedure. Vocht et al. [15] proposed the implementation of an optimized algorithm controlling the pathfinding process to obtain more homogeneous search domain and retrieve more links between adjacent hops in each path to improve the semantic relatedness of concepts mentioned in a story by increasing the relevance of links between nodes through additional domain delineation and refinement steps. Chen et al. [20] proposed a multimodal imitation learning via generative adversarial networks (MIL-GAN) method to directly model users' interests as reflected by various data by imitating users' demonstrated storylines. MIL-GAN model is designed to learn the reward patterns given user-provided storylines and then applies the learned policy to unseen data. Santos et al. [21] combined storytelling and Spatio-logical Inference (SLI) to generate rules of interaction among entities and measure how well they forecast a real-world event.

Hossain et al [13] introduced Google Fusion Tables(GFT) that offers collaborative data management in the cloud for data scientists to enable the integration of increasingly complex geospatial data to support storytelling. The paper focused on introduction of overview of map processing in GFT, the architecture overview of GFT, and how to scale to large datasets, massive and complex polygon datasets. GFT provides a useful tool for storytelling through interactive maps.

Kumar et al. [14] formulated storytelling as a generalization of redescription mining. Stories are defined as chains of redescriptions. They proposed an efficient storytelling algorithm as A\* search around the outputs of a CARTwhells redescription mining algorithm. The efficiency and scalability of the proposed algorithm were evaluated by three application case studies: word overlaps in large English dictionaries, exploring connections between gene sets in a bioinformatics data set, and relating publications in the PubMed index of abstracts.

Hossain et al. [19] proposed an approach to automatically construct stories between entities in large document collections that can help from directed chains of relationships, with support for co-referencing, evidence marshaling, and imposing syntactic constrains on the story generation process. A new optimization techniques based on concept lattice mining is used to rapidly construct stories on massive datasets.

Chen et al. [20] introduced an approach, multimodal imitation learning via generative adversarial networks (MIL-GAN) for generating storyline on unseen data. It can directly model users’ interests as reflected by various data. This approach is used to learn the reward patterns given user-provided storylines and then applies the learned policy to unseen data.

Santos et al. [21] introduced three methods of association analysis, Distance-based Byesian Inference, Spatial Association Index, and Spatio-logical inference, to capture relatedness among real-world events in high data volumes, and to model similar events that are described disparately under high data variability. It takes as input a set of geotemporally-encoded text streams about violent events called “storylines”. This study demonstrated that spatio-temporal storytelling is able to capture important associations among violent events reported in social media and traditional datasets.

Chapter 3

Methodology

Change Analysis Predicates

We begin by creating our change predicates [6]. There are three change functions that are necessary for complete change predicates, we define them as follows:

Here, is a spatial cluster polygon from the batch timed and is another spatial cluster polygon from a batch timed , with indicating a time increase. We denote cluster intersection as and union as . denotes the area covered by the polygon , upto a desired level of accuracy.

The change function is a measure of similarity. measures whether individual polygons coincide. To measure concurrence between one polygon and a set of polygons, we use the functions. Here are spatial clusters from a batch , and is a cluster from a different batch . Other basic functions include , which indicates the Centroid of a polygon as defined as:

…(1)

… (2)

where and A = the area of the polygon. We also use a simple Manhattan distance function which is tuned to the required georeferenced accuracy. We use all these functions to define these seven change predicates:

Implementations of these change predicates can be found in the Chapter 4.

Polygon Generation from Point Data Sources

We improve upon the change analysis framework called Aconcagua [6]. The system expects an input of emotion polygons annotated with emotion assessment scores, with +1 representing a very high positive emotion and -1 representing a very high negative emotion. While this method does lead to inconsistencies in locations due to georeferencing inaccuracies, we find that the 8 ~ 10m precision [8] works well within city limits.

There are several ways of using the numerous point data we have obtained from this step into an actual polygonal map:

1. Creating closed contour lines for contour lines that lie on the boundary of the observation area.
2. Creating a convex hull from points with similar scores.
3. Creating alpha-shapes or general concave hulls by selecting the minimal Delauney triangulation of the points.

We now elaborate on creating these polygons.

For contour lines, we start by selecting all point objects belonging to the same emotion group. We pick a random point within a group and connect all neighboring spots by straight lines, where ‘neighbor’ is defined somewhat arbitrarily according to close distances. We then move on to another point and restart the process. The results after all points are connected is a triangulated irregular network (TIN). All lines are then marked using an equal interval classification scheme, where the contour interval is chosen depending on the size of the polygon under consideration. Then, selecting an arbitraty point in the TIN, contour lines are threaded through ticks of equal value.

A convex hull can be though of as the smallest convex geometry that encloses all the points considered. If the group of points forms a line or is only a single then the convex hull is considered a degenerate and is ignored. Otherwise convex hulls are very fast to calculate using the Graham Scan algorithm, as the points are already sorted by latitude and longitude. The algorithm stars at the South Western most point and moves Eastward. For every coordinate triplet, the procedure checks whether or not it’s a concave corner. If it is, then the middle point cannot lie on the convex hull. This cross product calculation is also very fast. Convex hulls are widely used and are implemented in most common GIS systems. The shapes formed by a convex hull are however necessarily blocky and do not reflect natural distributions.

For a generalization of the convex hull, the concave hull is used. The most common method of concave hulls is Alpha Shapes. Alpha shapes usually look like generalized and more optimal versions of contour interpolations. Alpha shapes are very strong at formalizing the shape of a spatial point data set, unlike convex hulls. They area subgraph of the Delauney triangulation of the entire geometry. Given a set of points, the Delauney triangulation can generate a large family of shapes. We add the parameter ‘alpha’ to control the level of detail. The alpha shape degenerates to the point set for alpha tending to zero, and a large value of alpha begins to approximate the convex hull.

Change Analysis for Polygonal Map Data Sources

Our approach uses three primary set operations: union, intersection and erase. We calculate the area of each individual polygon within each map layer. We then execute a union operation and calculate area. The union layer now contains the original areas of both layers and the areas of the overlapping polygons - we now need to query them properly to prepare for calculating the change percentage and tabulating intersection.

To outline the polygon, we examine several different methods:

1. We find features common to either of the layers but not both, essentially performing a symmetrical difference
2. We erase the larger of the polygons from the smaller, thus retaining only the growth, and do vice-verse for shrinking
3. We perform simple intersection and then invert selection to get changed regions.

We then use these techniques:

1. **Data Pre-Processing**: This involves curation of datasets with obvious georeferencing errors. This would preferable be done by minimizing the root mean square error.
2. **Parametrization of polygons**: Calculate shape and area parameters for each individual polygon with each map layer.
3. **Analysis through Symmetrical Difference**: Extract features common to either of the layers.
4. **Polygon Union Computation**: Union sequential layers to contain the original areas of both layers and the areas of overlapping polygons.
5. **Polygon Erase Operation**: Erase larger polygons from smaller (or vice-versa) for detection of growth/shrinking.
6. **Polygon Intersection/Invert**: Selecting and then labelling the changed regions.

The framework can be found in Figure 1.

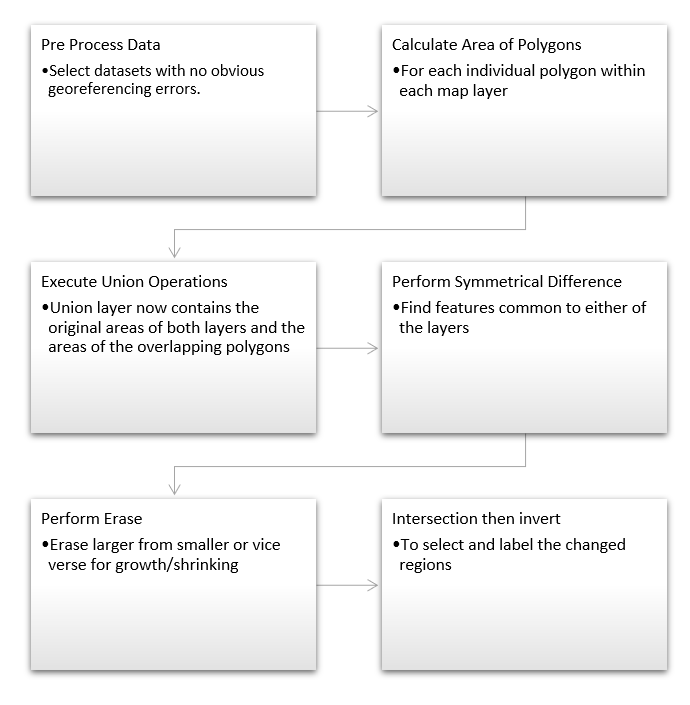


Figure 1: Change Analysis framework Architecture

Storytelling Techniques for Geospatial Data

In order to tell a coherent spatio-temporal data story from the change analysis output, we propose using an interestingness function for this task.[1]

We choose an area of interest which is a rectangle bounded by coordinates and . The length of the diagonal is . We choose a set of polygons SCP, whose centroid lies within .

The Thinness ratio measure takes a maximum value of 1 for a circle. Objects of regular shape have a higher thinness ratio than similar irregular ones. We define the thinness ratio of a polygon as:

… (3)

It not only contains those polygons but also their associated characteristics. For example, a SCP could contain a set of spatial clusters represented by polygons, their average drought score, total area of each polygon, centroid coordinates of each polygon and other summaries for each spatial cluster (polygon). We define the function as:

… (4)

We define a threshold $\omega$, which ensures that a narrative will only be generated an object such that . The parameters for are:

The parameters and are special cases. is to be used for those SCP whose primary change was growth/shrinking, henceforth referred to as . is to be used for those SCP whose primary change was shifting, henceforth referred to as . The threshold parameters need to be finely tuned so as to not exclude those polygons who fall through exceptions. Once we have a suitable selection of polygons and chose a threshold value, we can create a summary narrative based on that.

Thus, our Interestingness function is hence calculated as:

… (5)

… (6)

Chapter 4

Case Study

Data Sets

We use two datasets for our case study:

1. Drought dataset from the North American Drought Portal [28].
2. Twitter posts from the United states [9].

The Twitter data contains the timestamp, latitude, longitude and the text of each tweet, which are then processed and tokenized. The spatial reference can be seen in Figure 2 and the layer specification in Table 2.

The layer specification for the Drought shapefiles is summarized in Table 1. the spatial reference for the Drought shapefiles can be seen in Figure 3. The original structure of the drought monitor data can be seen in figure 4.

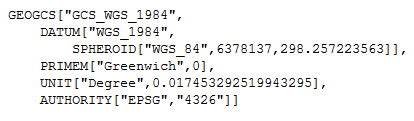


Figure 2: Spatial reference for Twitter Data

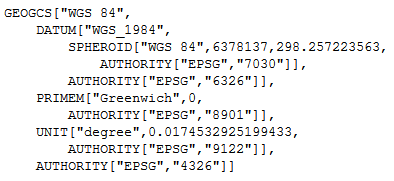


Figure 3: Spatial reference for Drought Data

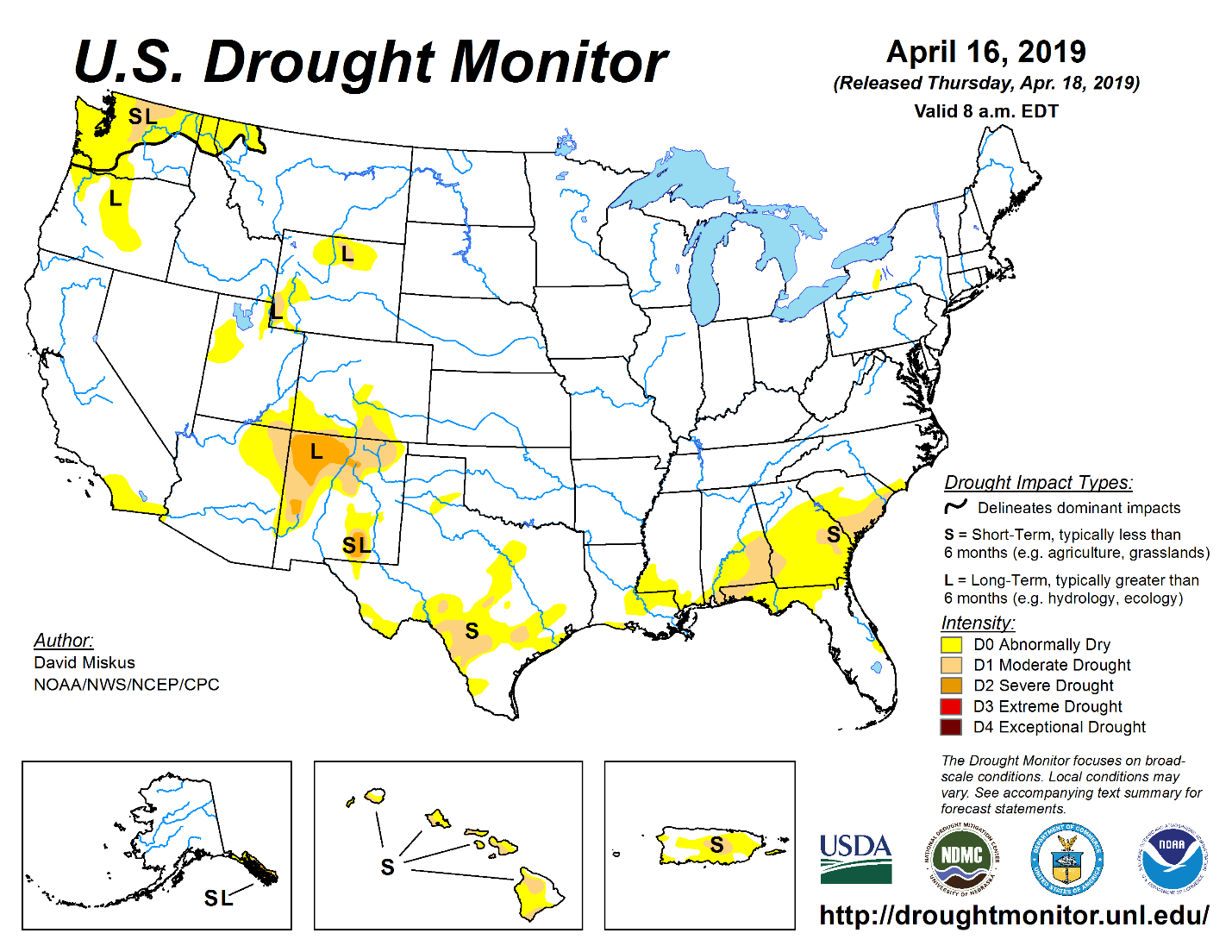


Figure 4: United States Drought Monitor Data (reproduced with permission)

Table 1: Layer Specification for Drought Data

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Width | Precision |
| Long | Real | 24 | 15 |
| Lat | Real | 24 | 15 |
| Id | Integer | 9 | 0 |
| dnstyTh | Real | 24 | 15 |
| avgScor | Real | 24 | 15 |
| numTwts | Integer | 9 | 0 |
| stdDev | Real | 24 | 15 |
| batchNm | Integer | 9 | 0 |
| geodata | String | 80 | 0 |

Table 2: Layer Specification for Twitter Data

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Width | Precision |
| FIPS\_ADMIN | String | 4 | 0 |
| GMI\_ADMIN | String | 7 | 0 |
| ADMIN\_NAME | String | 42 | 0 |
| FIPS\_CNTRY | String | 2 | 0 |
| GMI\_CNTRY | String | 3 | 0 |
| CNTRY\_NAME | String | 40 | 0 |
| POP\_ADMIN | Integer | 9 | 0 |
| TYPE\_ENG | String | 26 | 0 |
| TYPE\_LOC | String | 50 | 0 |
| SQKM | Real | 16 | 2 |
| SQMI | Real | 16 | 2 |
| COLOR\_MAP | String | 2 | 0 |

Dataset Preprocessing

Our initial approach to this problem was to store all shapefiles of the Twitter dataset in a postgres database with a GIS addon and perform operations in python. We used psycopg2 and osgeo libraries to import, process and visualize maps. However, this lead to many problems with interconversions between georeferencing schemes, while converting from WKT geometry to PostGIS geography.

We start with basic point data, which contains latitude/longitude, along with metadata identifying value of interest, whether that is drought level or emotion value. We insert the contents of the shapefile into a PostGreSQL database using the shp2pgsql toolkit that comes along with the PostGIS extension. The results can be seen in Figure 5.

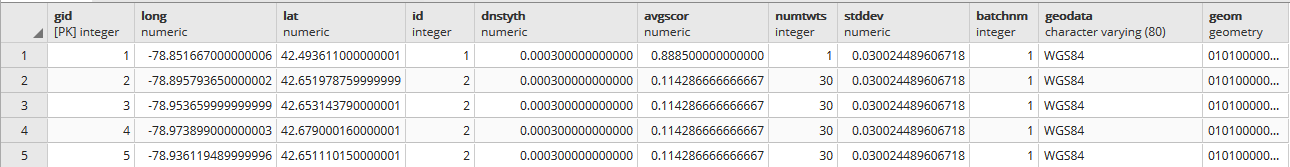


Figure 5: Twitter shapefile data inserted into Postgres db

As our framework relies on using polygonal data, we cannot use this. So we begin by creating a convex hull and alpha shapes of points and inserting it into a new table, while preserving related meta-data. Due to engine limitations, we need to be careful and insert only those convex hulls or alpha shapes that are polygons specifically, not points or lines.

Our query to generate concave hulls were:

INSERT INTO public.june1poly (avgscor, numtwts, geodata, id, batchnm, geom)

(SELECT d.avgscor, d.numtwts, d.geodata, d.id, d.batchnm, ST\_ConvexHull (ST\_Collect(d.geom))

FROM public."2014-06-01 " AS d

GROUP BY (d.id, d.avgscor, d.numtwts,d.geodata, d.batchnm)

HAVING ST\_GeometryType(ST\_ConvexHull (ST\_Collect(d.geom))) = 'ST\_Polygon')

For concave hulls, we needed to select a shrink factor target percentage between 0.5 and 0.99 depending upon the geometry in question. We do not allow for polygons with holes due to problematic edge cases. Our query for this is as follows:

INSERT INTO public.june1poly (avgscor, numtwts, geodata, id, batchnm, geom)

(SELECT d.avgscor, d.numtwts, d.geodata, d.id, d.batchnm, ST\_ConcaveHull (ST\_Collect(d.geom), 0.75)

FROM public."2014-06-01 " AS d

GROUP BY (d.id, d.avgscor, d.numtwts,d.geodata, d.batchnm)

HAVING ST\_GeometryType(ST\_ConcaveHull (ST\_Collect(d.geom), 0.75)) = 'ST\_Polygon')

Following this, we need to add some meta information into the shapefiles, such as centroid, perimeter lengths and area. The query for each of them are as follows:

1. Centroid:

UPDATE public.june1poly

SET centroid=ST\_Centroid(geom)

1. Area:

ALTER TABLE public.june1poly ADD COLUMN IF NOT EXISTS area double precision

UPDATE public.june1poly SET area=ROUND((ST\_Area(geom::geography))::numeric,2)

1. Perimeter

SELECT ST\_Perimeter(j2s.geom, true) FROM public.june1poly j2s

We repeat variations of this process for every shapefile. For the exact code to run each query, see Appendix A. Figures 6, 7 and 8 show the results of our polygon generation by using concave shapes from what was previously point data. Figure 7, especially, shows interesting patterns in how much more naturalistic the polygon is, as compared to Figure 9, which was generated without using alpha shapes.

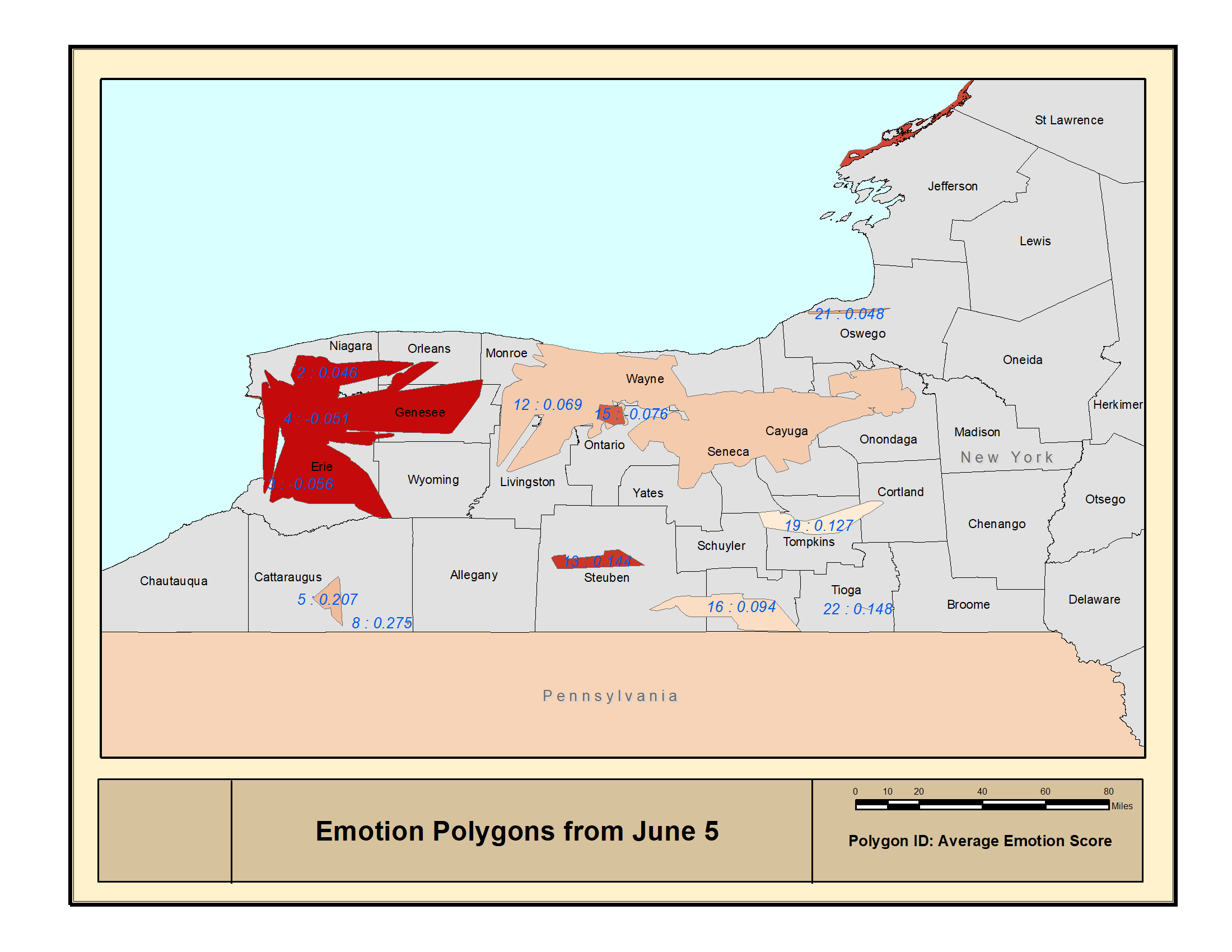


Figure 6: Emotion polygons created from June 5 Twitter data

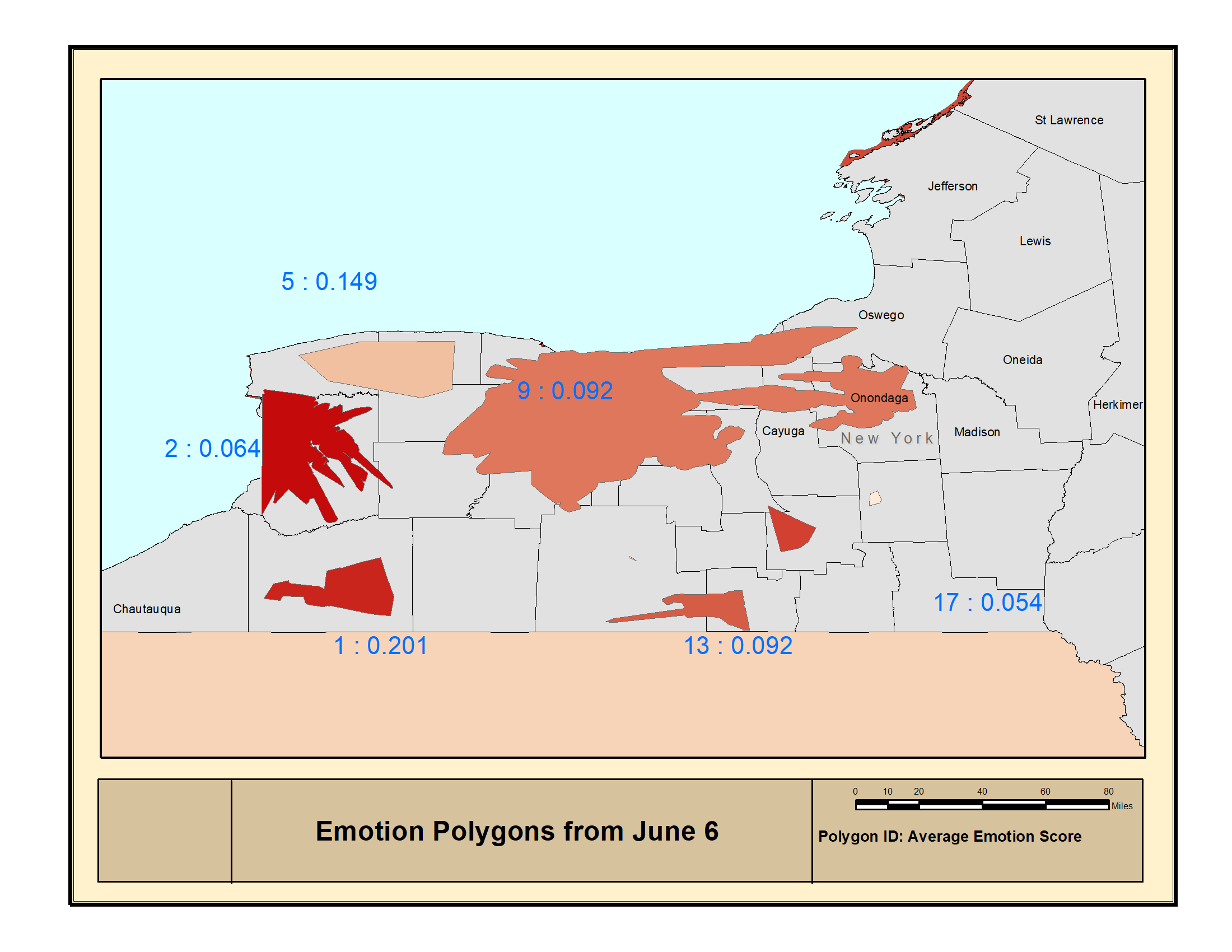


Figure 7: Emotion polygons created from June 6 Twitter data

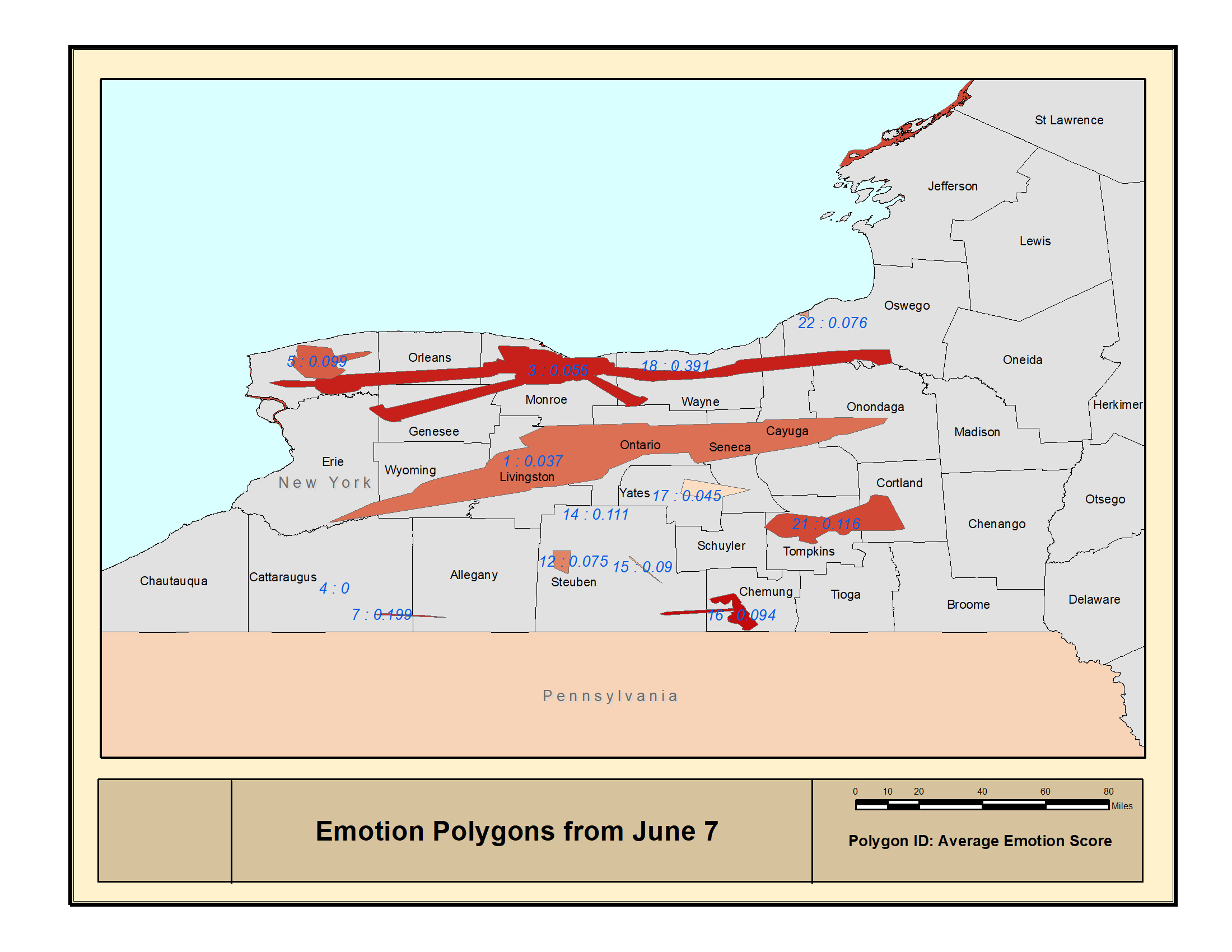


Figure 8: Emotion polygons created from June 7 Twitter data

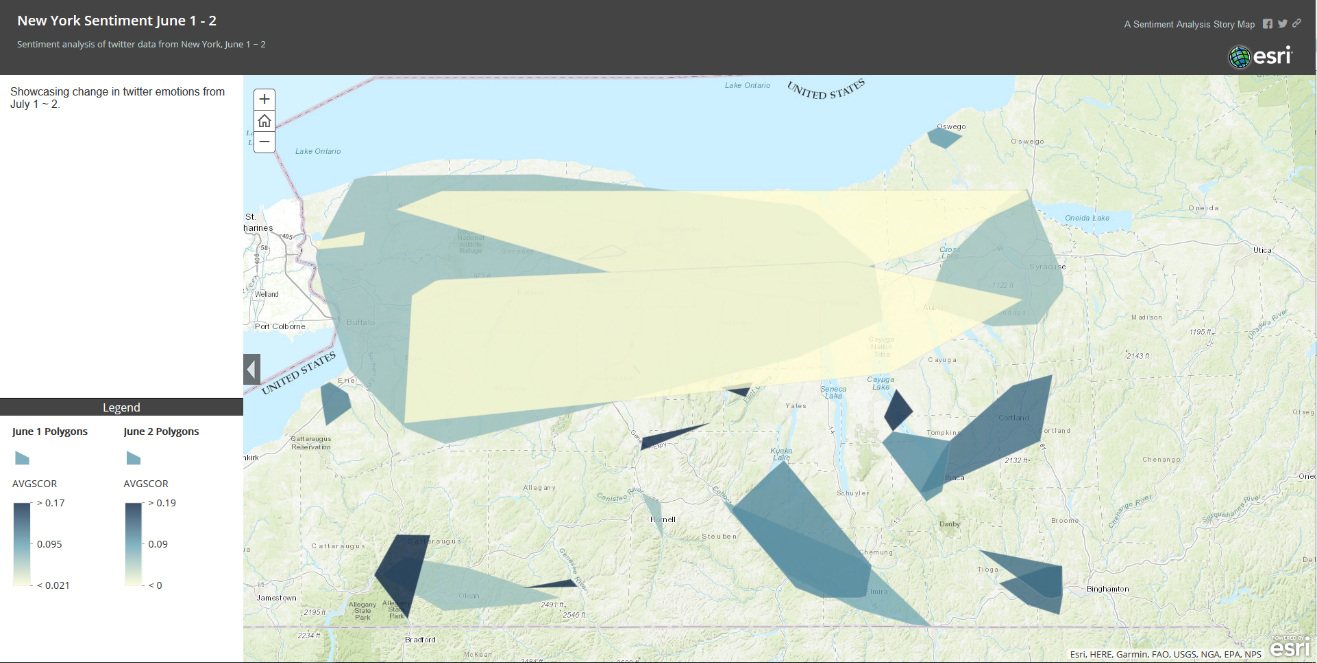


Figure 9: Emotion polygons created from June 1 Twitter data Using Convex Hulls

Drought Data Sources

The original polygons for the drought datasets, before our analysis is in Figure 10.

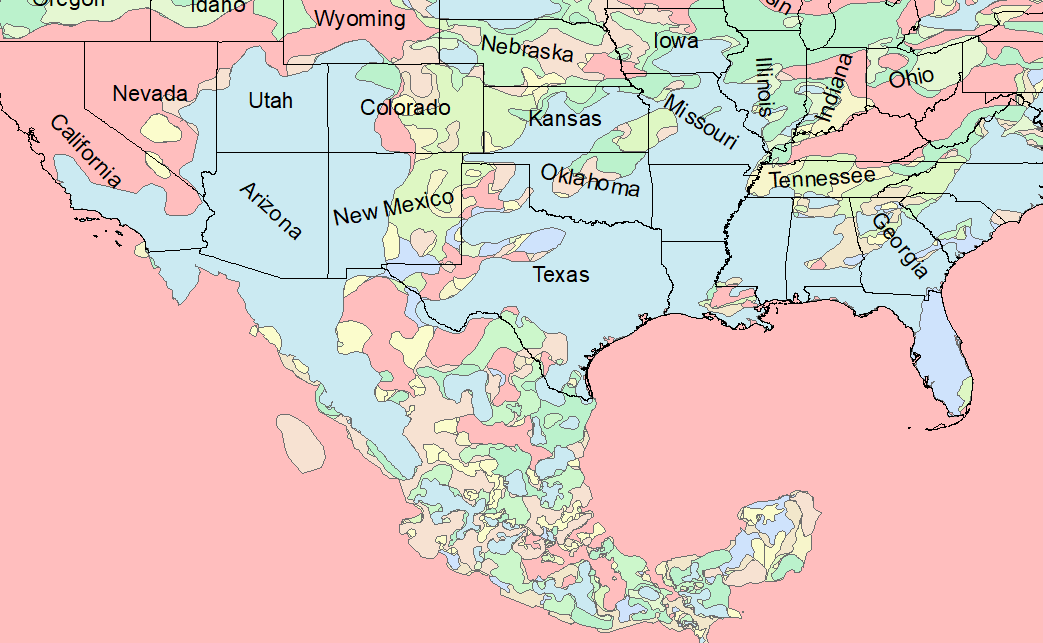


Figure 10: Continental USA original monthly drought polygon distribution for 2017

Now, we demonstrate the three change predicates as discussed in Chapter 3.

1. To detect polygons that are increasing in size, we check for similar IDs, intersection and then the rate of overlap. We initially check whether the polygons intersect at all before querying for amount of overlap. This leads to faster processing as it discards the many combinations where the polygons don't touch each other. Our query is structured as:

SELECT DISTINCT j2.\*

FROM public.june1poly j1, public.june2poly j2

WHERE ST\_INTERSECTS(j1.geom, j2.geom)

AND

(ST\_AREA(ST\_INTERSECTION(j2.geom, j1)) /st\_area(j2.geom)) > .85

A result of this operation performed on the drought dataset on September ~ October, 2017 can be seen in Figure 11.

1. To detect polygons that are shrinking in size, we check for similar IDs, and lower rates of overlap. This can be modified based on need.

SELECT DISTINCT j2.\*

FROM public.june1poly j1, public.june2poly j2

WHERE ST\_INTERSECTS(j1.geom, j2.geom)

AND

(ST\_AREA(ST\_INTERSECTION(j2.geom, j1)) /st\_area(j2.geom)) < .25

A demonstration of the output of this operation performed on the drought dataset on September ~ October, 2017 is found in Figure 12.

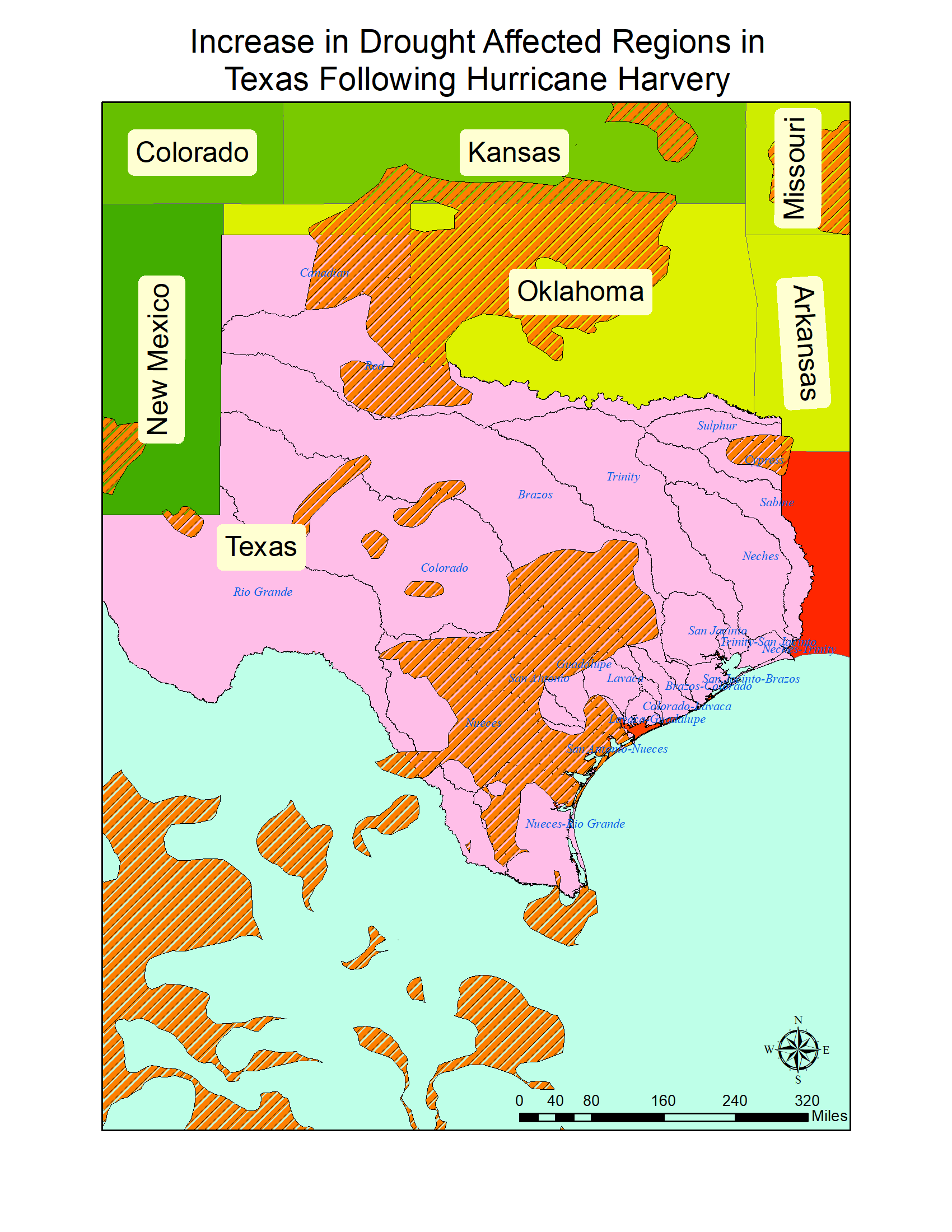


Figure 11: Increase in Drought Affected Regions in Texas following Hurricane Harvey

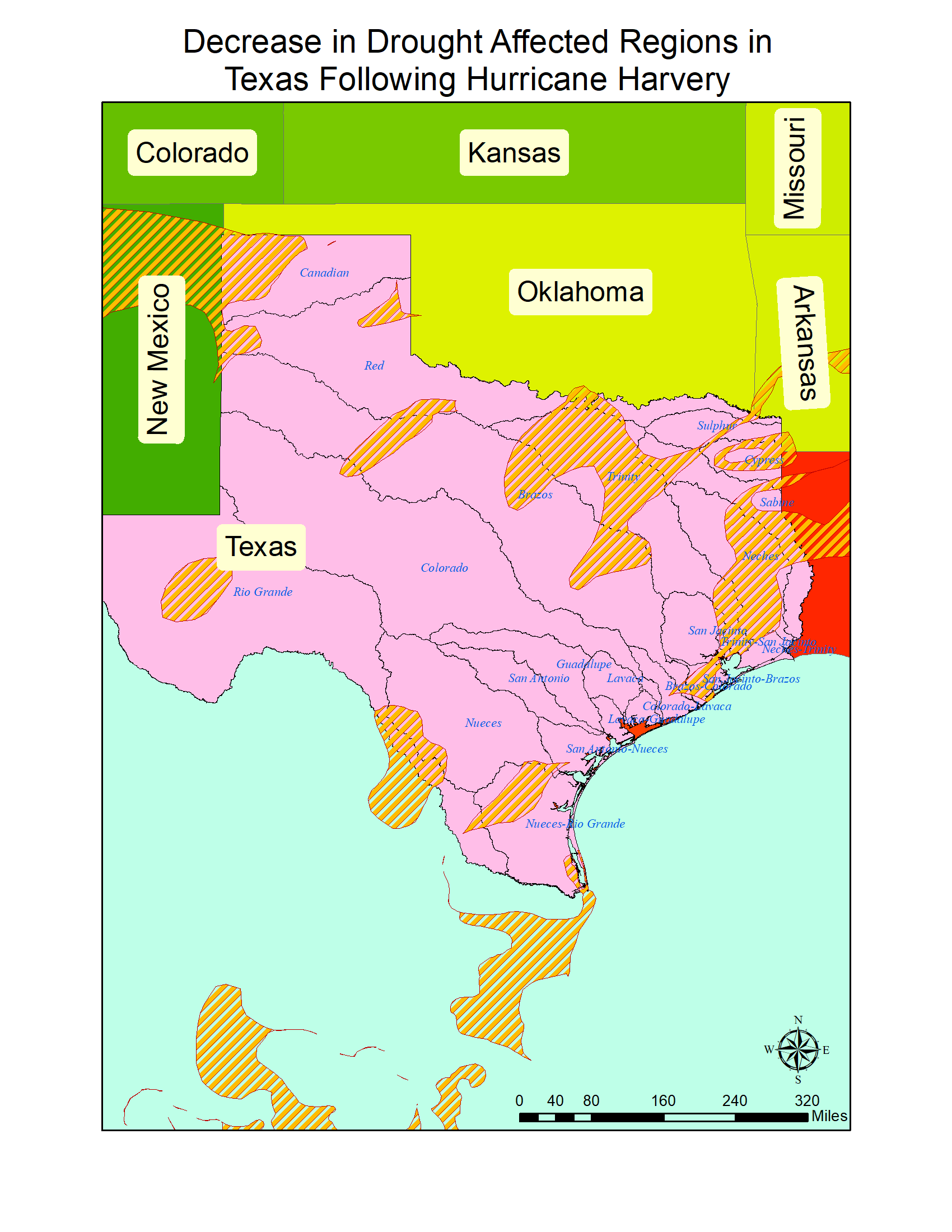


Figure 12: Decrease in Drought Affected Regions in Texas following Hurricane Harvey

1. To detect polygons that have shifted we compare their centroids and check if they have moved over 75km. Remember these polygons are created through a convex hull of points, which cannot ensure the centroid will lie within the polygon itself. Which is why we are taking a sufficiently large bounding value for the polygon.

SELECT ST\_Distance\_Spheroid

(j1.centroid, j2.centroid, 'SPHEROID["WGS 84",6378137,298.25]'), j1.id

FROM public.june1poly j1, public.june2poly j2

WHERE j1.id = j2.id AND ST\_Distance\_Spheroid(j1.centroid, j2.centroid,

'SPHEROID["WGS 84",6378137,298.25]')>75000;

An example of polygons shifting around can be found in Figure 13, highlighting changes in the California area following the 2017 wildfires.

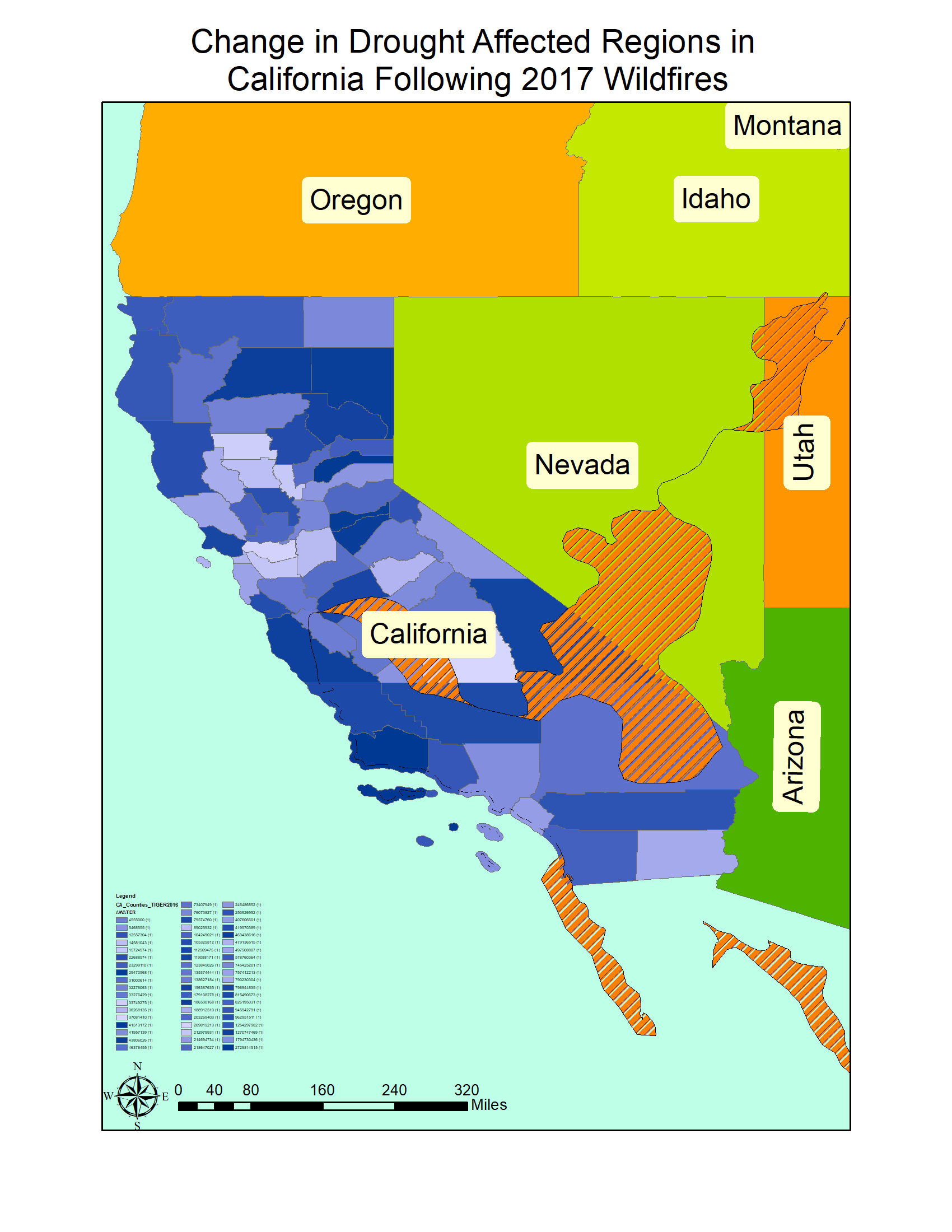


Figure 13: Shifting areas of drought affected regions in California following the 2017 wildfires

Emotion Spatial Clusters from Twitter

The tool used for creating the emotion clusters was the K2 framework [7].The spatial reference for the twitter data can be seen in figure 2.

According to K2, the best package for the emotion score was the Valence Aware Dictionary and Sentiment Reasoner (VADER) system [10], whose analyser parses the tokenized text and checks within a lexicon for words with strong sentiment. The final score is created from the weighted average of all sentimental words and lies within the range [-1, 1], as is required for our Aconcagua implementation.

The results of running our framework on June 5, 6 and 7 can be seen in Figure 15 and Figure 15.

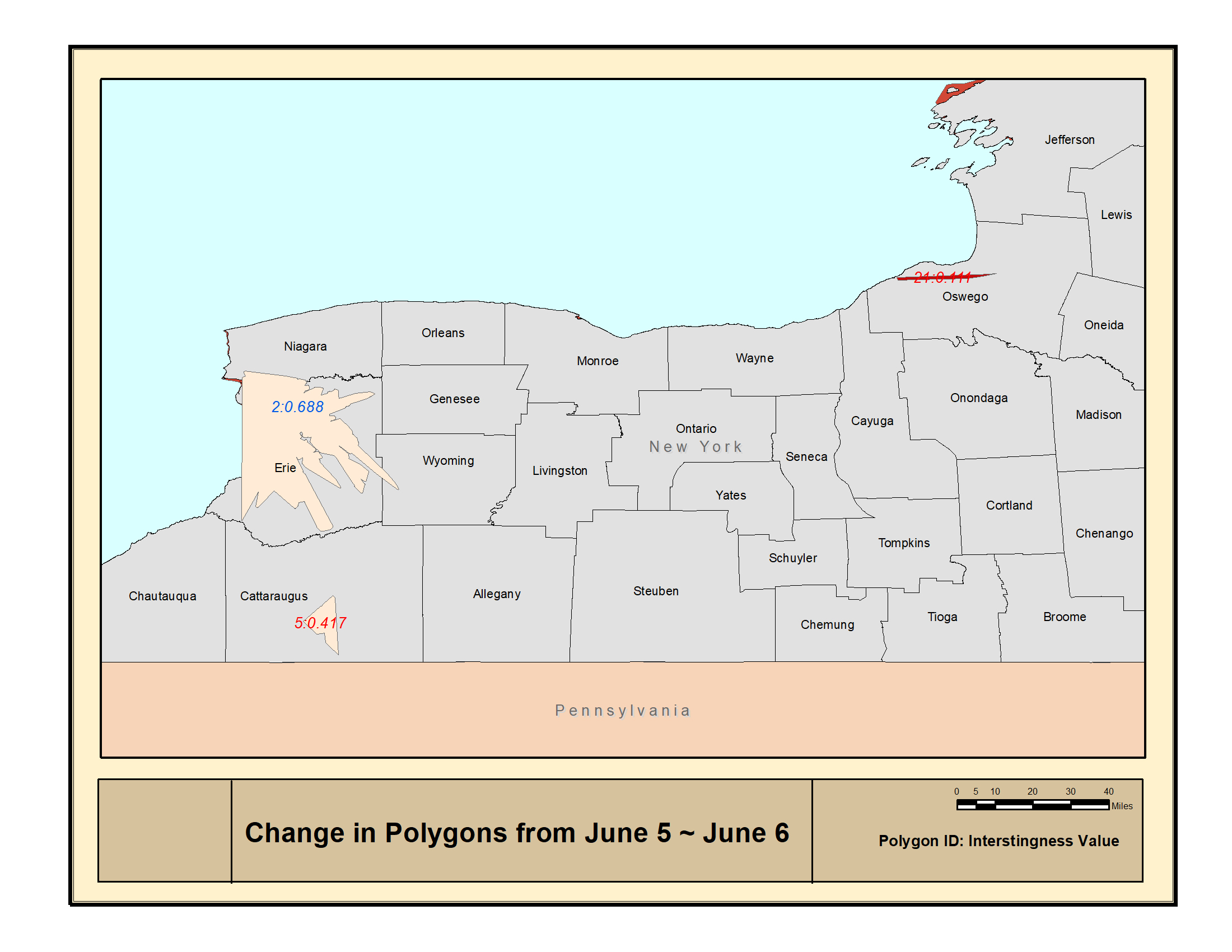


Figure 14: Change in Twitter emotion polygons from June 5 to June 6

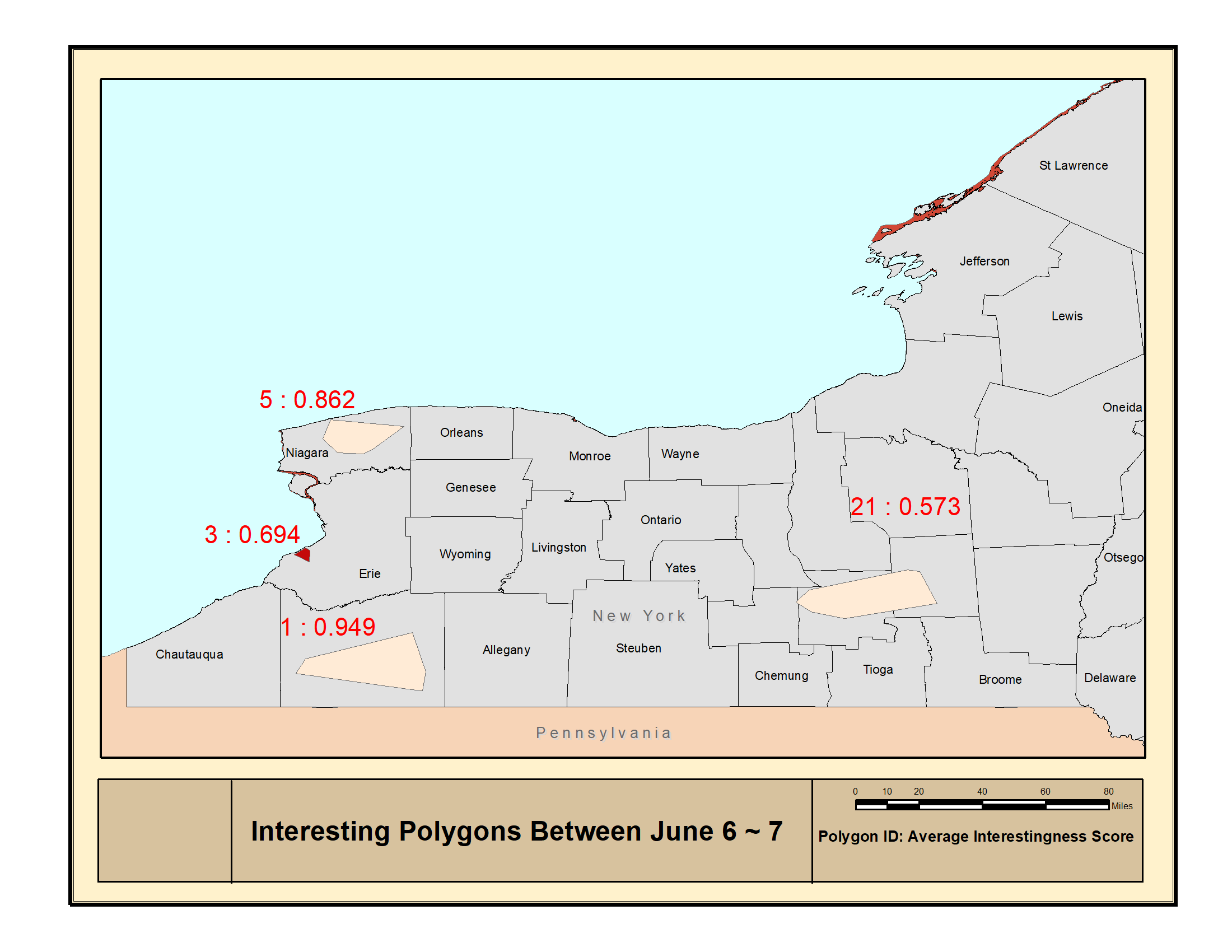


Figure 15: Change in Twitter emotion polygons from June 6 to June 7

Analysis

We focus on two specific regions for this case study. First, Figure 8 shows the regions where areas of drought grew in California after the wildfires in 2017. We noticed patterns of increasing drought surrounding the regions that were burnt down, with especially large period upstream from rivers. Our procedure works well dealing with the large number of small polygons in the region that occur due to isolated wildfires. However, we had difficult dealing with convex polygons that intersected with each other multiple times.

Next, we inspect the region in Texas after Hurricane Harvey. Figure 9 shows the regions where drought affected regions increased following the disaster and Figure 10} shows the regions decreased.

We observe that drought affected regions decrease at a high rate around the South Eastern Texas and Louisiana region. However, there is no clear relation between drought prone regions and river basins close to the coast. We believe that high amount industrial regions create a micro climate that affects the water content entering the soil.

According to our results, the drought prone regions increased substantially around the West Texas region both during and after Hurricane Harvey, with a larger increase as winter came around. However, this could be an artefact of the fact that the original dataset had very few polygons in the Texas area, which leads to broader conclusions. Our technique is still severely dependent on the resolution of the input imagery.

The results of our interestingness calculation, as seen in Figure 14 tell us that the smaller polygons with high emotion score and large amounts of variability are the most interesting. This tells us that regions with most outspoken or passionate tweets are the ones we should focus on when we are generating automated stories, which is obvious. Comparing Figure 12 with Figure 7 and Figure 8, we find that the large central polygon with very low emotion scores do not have much variance, telling us that that large group of people in a rural do not change their opinions dramatically day by day.

Chapter 5

Conclusion and Future Work

Our experimental studies show that our change detection and analysis framework can successfully detect changes in spatiotemporal datasets. Our change predicates can work on a variety of data sources, including both polygonal and point datasets. Our framework is able to handle both geographically large and small map sources without any significant georeferencing distortions being introduced.

There is a lot work left in dealing with irregularly shaped convex polygons that appear in poorly georeferenced real-world data. We believe this is because the sensitivity of the geographic operations in PostGIS or ArcGIS. We are currently working on extending our framework to deal with similar cases.

References

[1] J. Im and J. Jensen. 2005. A change detection model based on neighborhood correlation image analysis and decision tree classification. Remote Sensing of Environment 99, 3 (2005), 326–340. DOI:http://dx.doi.org/10.1016/j.rse.2005.09.008

[2] Suming Jin, Limin Yang, Patrick Danielson, Collin Homer, Joyce Fry, and George Xian. 2013. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. Remote Sensing of Environment 132 (May 2013), 159–175. DOI:http://dx.doi.org/10.1016/j.rse.2013.01.012

[3] D. Lu, P. Mausel, E. Brondízio, and E. Moran. 2004. Change detection techniques. International Journal of Remote Sensing 25, 12 (June 2004), 2365–2401. DOI:http://dx.doi.org/10.1080/0143116031000139863

[4] Merrill K. Ridd and Jiajun Liu. 1998. A Comparison of Four Algorithms for Change Detection in an Urban Environment. Remote Sensing of Environment 63, 2 (1998), 95–100. DOI:http://dx.doi.org/10.1016/s0034-4257(97)00112-0

[5] S. Minu and Amba Shetty. 2015. A Comparative Study of Image Change Detection Algorithms in MATLAB. Aquatic Procedia 4 (March 2015), 1366–1373. DOI:http://dx.doi.org/10.1016/j.aqpro.2015.02.177

[6] K. Elgarroussi, S. Wang, R. Banerjee, and C. F. Eick, “Aconcagua: A Novel Spatiotemporal Emotion Change Analysis Framework,” in Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, New York, NY, USA, 2018, pp. 54–61.

[7] R. Banerjee, K. Elgarroussi, S. Wang, A. Talari, Y. Zhang, and C. F. Eick, “K2: A Novel Data Analysis Framework to Understand US Emotions in Space and Time,” Int. J. Semantic Computing, vol. 13, no. 01, pp. 111–133, Mar. 2019.

[8] Geomenke, “How Accurate is the GPS on my Smart Phone? (Part 2),” Community Health Maps, 07-Jul-2014. .

[9] W. Z.-M.-L. I. for the S. S. Person, “Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness,” Jan. 2016.

[10] C. J. Hutto, VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social .. 2019.

[11] X. Huang, A. An, and N. Cercone, “Comparison of Interestingness Functions for Learning Web Usage Patterns,” in Proceedings of the Eleventh International Conference on Information and Knowledge Management, New York, NY, USA, 2002, pp. 617–620.

[] W. Wang, Lu Chen, K. Thirunarayan, and Amit P. Sheth, Harnessing Twitter "Big Data" for Automatic Emotion Identifcation, in Proc. of the IEEE 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing.

[] R. F. Dos Santos, S. Shah, F. Chen, A. Boedi-hardjo, P. Butler, C.T. Lu and N. Ramakrishnan, "Spatiotemporal storytelling on twitter", Virginia Tech Computer Science Technical Report, 2015, [online] Available: http://vtechworks.lib.vt.edu/handle/10919/24701.

[20] D. Kumar, N. Ramakrishnan, R. F. Helm and M. Potts, Algorithm for storytelling, in IEEE Transactions on Knowledge and Data Engineering, 2008, Volume 20 Issue

6, pp. 736-751.

[] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts and R. R. Helm, Turning CARTwheels: an alternating algorithm for mining redescriptions, in Knowledge Discovery and Data Mining, 2004, pp. 266-275.

[] L. D. Vocht, C. Beecks, R. Verborgh, T. Seidl, E. Mannens and Rik Van De Walle, Improving Semantic Relatedness in Paths for Storytelling with Linked Data on the Web, in Revised Selected Papers of the ESWC 2015 Satellite Events on The Semantic Web: ESWC 2015 Satellite Events, May 31-June 04, 2015.

[] Z. Chen, X. Zhang, A. P. Boedihardjo, J. Dai, Chang-Tien Lu, Multimodal storytelling via generative adversarial imitation learning, in Proceedings of the 26th International Joint Conference on Artifcial Intelligence, August 19-25, 2017, Melbourne, Australia.

[] R. D. Santos, S. Shah, F. Chen, A. Boedihardjo, Chang-Tien Lu, N. Ramakrishnan, Forecasting location-based events with spatio-temporal storytelling, in Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pp. 13-22, November 04-04, 2014, Dallas/Fort Worth, Texas.

[] Jayant Madhavan, Sreeram Balakrishnan, Kathryn Brisbin, Hector Gonzalez, Nitin Gupta,Alon Halevy, Karen Jacqmin-Adams, Heidi Lam, Anno Langen, Hongrae Lee,Rod McChesney, Rebecca Shapley, Warren Shen, “Big Data Storytelling through Interactive Maps”

[] M. Shahriar Hossain1, Patrick Butler1, Arnold P. Boedihardjo2, Naren Ramakrishnan, Storytelling in Entity Networks to Support Intelligence Analysts, KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1375-1383 ,Beijing, China , August 12 - 16, 2012

[] Zhiqian Chen , Xuchao Zhang , Arnold P. Boedihardjo , Jing Dai , Chang-Tien Lu, Multimodal storytelling via generative adversarial imitation learning, Proceedings of the 26th International Joint Conference on Artificial Intelligence, August 19-25, 2017, Melbourne, Australia

[] Raimundo F. Dos Santos, Jr. , Arnold Boedihardjo , Sumit Shah , Feng Chen , Chang-Tien Lu , Naren Ramakrishnan, The big data of violent events: algorithms for association analysis using spatio-temporal storytelling, Geoinformatica, v.20 n.4, p.879-921, October 2016

[] Deept Kumar, Naren Ramakrishnan, Richard F. Helm, and Malcolm Potts. 2008. Algorithms for Storytelling. IEEE TKDE20, 6 (2 2008), 736–751.DOI:http://dx.doi.org/10.1145/1188913.1188915

[] “GRACE Tellus Data,” GRACE Tellus. [Online]. Available: https://grace.jpl.nasa.gov/data/get-data. [Accessed: 26-Jul-2019].

[] “SARAL - eoPortal Directory - Satellite Missions.” [Online]. Available: https://directory.eoportal.org/web/eoportal/satellite-missions/s/saral. [Accessed: 26-Jul-2019].

[] “Top 5 sources of big data | Artificial Intelligence | Data Science |,” 26-Nov-2017. [Online]. Available: https://www.allerin.com/blog/top-5-sources-of-big-data. [Accessed: 26-Jul-2019].

[] J. Thatcher, “Big Data, Big Questions| Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data,” International Journal of Communication, vol. 8, no. 0, p. 19, Jun. 2014.

[] “Livehoods.” [Online]. Available: http://livehoods.org/research. [Accessed: 27-Jul-2019].

[] “Shapefiles of drought intensity and impact for the North American Drought Portal.” [Online]. Available: https://www1.ncdc.noaa.gov/pub/data/nidis/shapefiles/. [Accessed: 01-Aug-2019].

Appendices