USING SOCIAL MEDIA TO COLLECT FINE SCALE PUBLIC OPINION

A Thesis submitted to the faculty of
San Francisco State University
In partial fulfillment of
the requirements for
the Degree

Master of Science

In

Geographic Information Science

by

Benjamin Harrison Wheeler

San Francisco, California

May 2018

CERTIFICATION OF APPROVAL

I certify that I have read Using Social Media to Collect Fine Scale Public Opinion by Benjamin Harrison Wheeler, and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of the requirement for the degree Master of Science in Geographic Information Science at San Francisco State University.

XiaoHang Liu, Ph.D.
Professor of Geography

Jerry Davis, Ph.D.
Professor of Geography

USING SOCIAL MEDIA TO COLLECT FINE SCALE PUBLIC OPINION

Benjamin Harrison Wheeler
San Francisco, California
2018

Fine-scale public opinion is important to politicians, special interest groups, and social organizers; it, however, remains cost-prohibitive. Twitter provides a wealth of information about public opinion. Most studies of public opinion using Twitter focus at the state level while important campaign decisions are made at finer scales. This study evaluates the usefulness of Twitter as a source of detecting fine-scale public opinion by expanding to the county level through collecting tweets based on a broader set of candidate-related hashtags as search terms, and the utilization of user location, a free-form text field, for resolving a greater number of tweets to a location on the Twitterscape. To explore the utility of county-level Twitter data, I present a case study of the New York State Democratic and Republican Presidential Primary Elections. This study reveals an urban bias in Twitter data, correlations as high as 0.78 between percentage of tweets about a candidate (Twitter share) and vote share, a significant relationship between Twitter share the day before an election and vote share, and a higher rate of the use of coordinates for tweets favoring discussion about Democratic candidates. These results signal the usefulness of Twitter as a fine-scale sensor of public opinion. Additionally, due to the wide-scale use of geographic weighted regression (GWR), a method is formalized for moving from an OLS model to a GWR using Moran's I as a diagnostic on model residuals to identify the potential for non-stationarity in model relationships. This approach also emphasized the use of multiple hypothesis testing corrections, as these corrections have been overlooked by many users of GWR. Approaches for visualizing the outputs of such models are illustrated employing data visualization schemes that are easily repeatable on standard mapping software.

I certify that this Abstract is a correct representation of the content of this thesis.

Chair, Thesis Committee Date

ACKNOWLEDGEMENTS

I would like to thank my wife who saw me through this process with unending support. I must also acknowledge the instrumental assistance of Tina Mendez, Ph.D., without whose guidance, this paper would never have reached completion. I would also like to thank Uriel Garcia for his invaluable support in programming the correlation analyses. As well as the invaluable wit and editing of my brother David. Dany Mendez and Lina Mendez, Ph.D. for their advice, support, and encouragement to narrow the scope of this paper. I would also like to thank Prof. Jerry Davis and Prof. XiaoHang Liu for their patience as I struggled through this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures Page

LIST OF APPENDICES

CHAPTER ONE

1. Introduction

In fall 2015, Gallup Polls announced that they would not be polling what is regarded as the "horse race" for the 2016 Presidential nominations (Shephard 2015). The announcement followed a major rework of Gallup after a wildly incorrect prediction of President Barack Obama losing to Mitt Romney. A number of major pollsters decided to no longer follow the Democratic and Republican primary elections as closely as they had in the past, creating an interesting gap in information.

Public opinion is at the heart of the democratic process of the United States. Due to the complexities of the Electoral College for presidential elections, state-level opinions dictate the next President of the United States. Similarly, sub-state regional opinions dictate the rise and fall of senators and congresswomen. Despite its importance, fine-scale studies of public opinion remain elusive and limited in scope. The cost-prohibitive method of phone polling provides a significant barrier to finer scale polling. The proliferation of social media coupled with the advent of Big Data has led to an ever-increasing network for gauging political opinions, allowing for the analysis of millions of records.

Finer scale understanding of public opinion using social media requires an understanding of how geographic reality maps onto the landscape of the internet. This has been called the “web information landscape” by Tsou et al. (2013), and later termed “cyber geography” (Tsou and Leitner 2013). One medium for understanding both public opinion and cyber geography is Twitter. Twitter is a microblogging service which consists of small 140 character public messages called tweets.

Tweets are accessible to the public through Twitter’s public application programming interface (API)—which allows utilization of the software for the purpose of designing applications—and are accompanied by a wealth of metadata. Twitter has become a heavily studied medium for this reason. One form of metadata commonly associated with tweets that makes them particularly useful for finer scale analyses is geographic data. This data can come in the format of coordinates, places (polygons) and a free-form location field (Ajao, Hong, and Liu 2015). Like in other social media, Twitter relationships are found geographically local (Quercia et al. 2012), suggesting that our relationships on Twitter reflect geographic reality. This suggests a Twitter landscape or Twitterscape as it has been termed in this paper.

Twitter has been shown to be useful for identifying a number of physical phenomena including crime (Gerber 2014), disasters (Sakaki, Okazaki, and Matsuo 2010), relationships (Quercia, Capra, and Crowcroft 2012) and social movements (Hemsley and Eckert 2014). Additionally, Twitter has been shown to be useful in both spatial and aspatial capacities as an indicator for public opinion (Bermingham and Smeaton 2011; DiGrazia et al. 2013; O’Connor et al. 2010; Tsou et al. 2013). While this is the case, Twitter has also been shown to have a demographic and a spatial bias, exhibiting an overrepresentation of urban populations(Arthur and Williams 2017; Mislove et al. 2011; Mitchell et al. 2013). While this presents problems for making sweeping assumptions about the American public, it does not preclude the data from being useful. This is particularly true in politics where only a narrow scope of the American public is truly important in deciding the outcome of an election. In the United States, the use of the Electoral College emphasizes swing states and a narrow subset of the undecided population as the focal point of elections. As such, demographically or geographically biased indicators should not necessarily be rejected.

While a number of studies explore the geography of political opinions on the Twitterscape (Beauchamp 2015; Calvo and Escolar 2003), little work has been done to produce finer scale methods for investigating sub-state patterns of public opinion on the Twitterscape in the US.

In an effort to investigate sub-state patterns of public opinion on the Twitterscape, a case study of New York–a geographically and demographically diverse state–was performed to study the 2016 Presidential primaries for both the Democratic and Republican Parties.

The primary goal of this research is to build a model to predict public opinion at the county-level by incorporating data from the Twitterscape. To this end, I will examine several questions. First, in order to build the model, I look at how to collect a greater number of tweets, and how to use geographic data to extrapolate county-level opinions.

Then I investigate the characteristics of the locations of tweeters, whether tweets contribute to prediction accuracy or not, and how Twitter share correlates to election outcomes. Finally, I question how effective traditional models are that use demographics only to predict public opinion, and whether traditional models are most suitable considering the Twitterscape is geographically local.

2. Study Area

2.1. New York State

The State of New York is a legal administrative, judicial and legislative boundary of the United States of America. It resides on the East Coast of the US, one of the thirteen original colonies; it is closely connected to the surrounding states economically. It borders Vermont, Connecticut, Pennsylvania, New Jersey, Massachusetts and the country of Canada. New York is composed of several urban centers including Buffalo, Rochester, Syracuse, Albany, New York City, and Long Island.

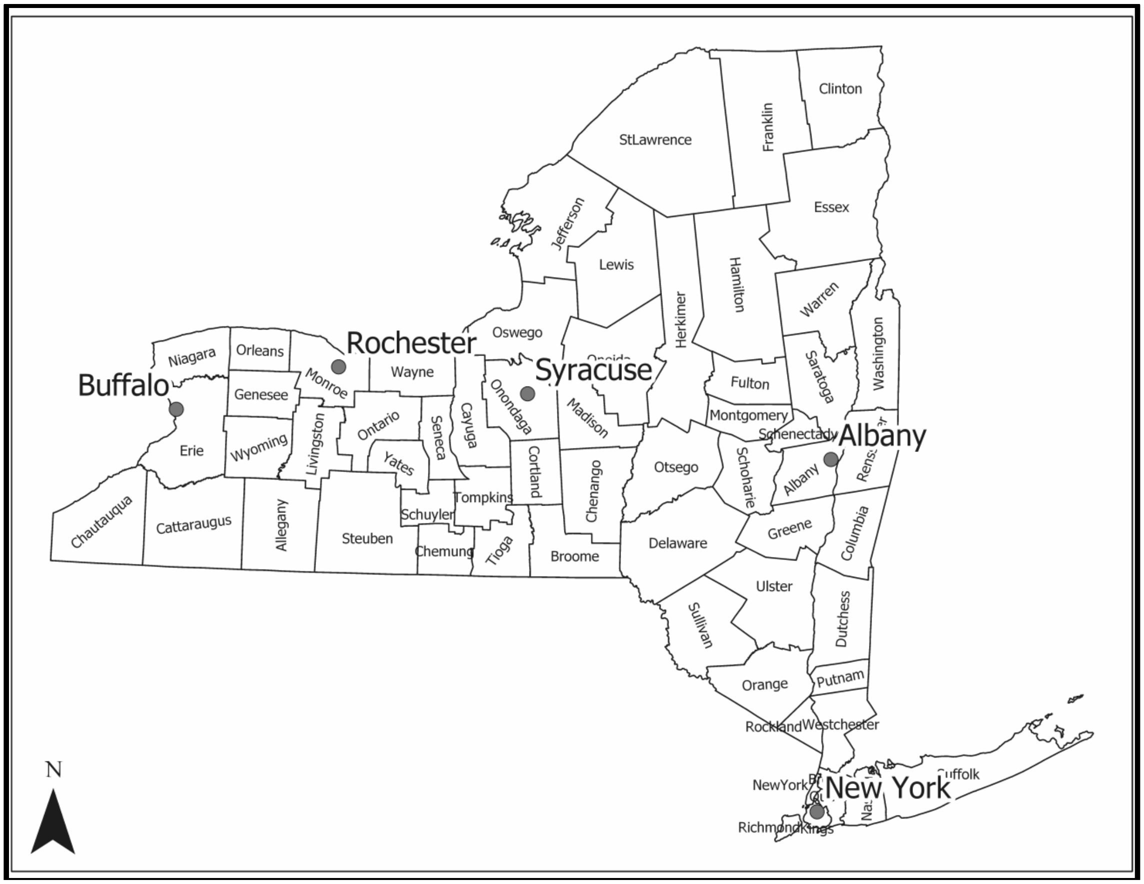

Figure 1. Map of the study area

3. Significance of Research

Three interconnected aims collectively support the overarching goal of producing fine-scale public opinion data. They are to (1) capture a larger cross-section of the political conversation on Twitter, (2) understand and evaluate the merit of county-level aggregation of the Twitterscape, and (3) produce a procedure for investigating localized political phenomena using geographic weighted regression.

3.1. Capture a Larger Cross-Section of the Political Conversation on Twitter

DiGrazia et al. (2013), Tsou et al. (2013) and Shi et al. (2012) utilize detection of relevant tweets using first and last name pairs of candidates (e.g., *Mitt Romney*). No studies of public opinion using hashtags and the candidate's name were found in the literature. Analyses of specific hashtags, and of their changes in meaning over time (Small 2011), discourse analyses related to specific hashtags (e.g., Hemsley and Eckert 2014), but no method or solution for sampling the Twitter stream to identify a broader scope of tweets related to public opinion could be found. It stands to reason that sampling of a population based on a first and last name configuration could produce sample biases, demographically or politically. In an effort to increase the population of tweets, relevant hashtags were also used to detect additional dialogues (e.g., *#FeelTheBern*, *#MakeAmericaGreatAgain*).

In addition to this, Tsou et al. (2013) studies phenomena at the state or regional level, and this study attempts to increase the richness of the political opinion landscape through the use of both place and the free-form location field ("user location") associated with Tweets by imitating Mislove et al. (2011) and their use of the user location field for aggregating tweets at a county level. By utilizing several sources of geographic information, more tweets can be resolved to a location on the map, hypothetically allowing for a more granular landscape by capturing a more comprehensive subset of the population.

3.2. Evaluate the Merit of County-Level Aggregations of the Twitterscape

Digrazia et al. (2013) produces an aspatial model with a high correlation between Twitter mentions and the actual share of votes a candidate garnishes on election day, while suggesting that this removes the necessity for geographic association of a candidate's support, their study population consists of regionalized political competitions that lack a national scale, as such this option is not applicable for primaries for Presidential races that draw international attention. Tsou et al. (2013) attempt to resolve public opinion to the state or metropolitan level using coordinates collected for metropolitan areas. The structure of their method, however, neglects rural counties. The structure of this study aims to aggregate and compare correlations for county-level data and vote share in an effort to provide a less urban-centric understanding of public opinion.

3.3. Produce a Procedure for Investigating Localized Political Models Using Geographic Weighted Regression

Political partisanship in the United States varies over space and has been shown to have an inherent geographic structure (e.g., Tam Cho, Gimpel, and Hui 2013; Mellow and Trubowitz 2005; McKee and Teigen 2009; Morrill, Knopp, and Brown 2007). This structure (in a simplistic example) biases towards a more liberal urban population, moderate/conservative suburban populations and rural populations that are more conservative. This phenomenon can be demonstrated at various scales including from the city-level to the state level (Morrill, Knopp, and Brown 2007). Phenomena that exhibit spatial heterogeneity may present themselves as clustering or dispersion of errors higher than would be expected at random when modeled using traditional linear models (Fotheringham, Brunsdon, and Charlton 2002). One method of exploring complex spatial relationships is geographic weighted regression(GWR) which allows relationships in the model to vary over space (Fotheringham, Brunsdon, and Charlton 2002). Despite a burgeoning body of literature surrounding this method, there are no succinct procedural papers for determining the need for a geographic weighted regression, outlining methodological decisions in building a model, and evaluating the outputs. To achieve this aim, a procedure is suggested later in this paper.

4. Overview

The remainder of this paper will be organized into three chapters, following this introduction (Chapter One). Chapter Two will outline the methods by which tweets are collected and processed for all candidates with an evaluation of their correlations with vote share (i.e., performance at the polling booth). This is to include additional findings and discussion of spatial and partisan biases on the political Twitterscape. Chapter Three will follow and introduce a formalized procedure for moving to geographic weighted regression, in addition to developing this procedure, a model for Donald Trump will be evaluated utilizing it, with an emphasis on Twitter-derived variables.

CHAPTER TWO: CAPTURING A LARGER CROSS-SECTION OF THE TWITTER CONVERSATION AND COUNTY LEVEL CORRELATION ANALYSIS

1. Data Collection and Processing

1.1. Tweet Data Collection

The Twitter API has two primary functions that pertain to collecting tweets, the search/REST API, and the Streaming API. The Streaming API initiates a connection to the Twitter server and provides tweets as they appear, while the Search API returns tweets from the last seven days within certain parameters. Morstatter et al. (2013) demonstrated that the Streaming API is not quite a random sample of the entire Twitter stream. They refer to a 1% limitation on the Streaming API, meaning that, the API will return a volume of tweets no greater than 1% of the entire volume of the Twitter stream. There is no official documentation on what the actual limit is. The Streaming API can be used with several different methods of filtering. These methods include filtering for keywords provided to it, as well as an option to set geographic or spatial filters (e.g., tweets within 10 miles of the center of New York City). While the second option sounds particularly useful for this study, Hemsley and Eckert (2014) assert that only 1-2% of all tweets are geospatially enabled, thus severely limiting the potential sample population. In order to capture more than a cross-section of that 1-2% of tweets, a set of hashtags relevant to the election was used in conjunction with the Streaming API to collect all tweets available through the API without a spatial filter. Hypothetically, this would enable the program to capture more users located in New York than solely those with geospatially enabled tweets.

1.2. Programming a tool to find and save tweets

Twitter uses JavaScript Object Notation (JSON)—a file format that is easily read and written in most programming languages—to deliver tweets (Crockford 2006; Twitter Inc. 2016). A Python program, *tiipwriter*, was written to store JSON tweets from the Twitter Streaming API in text files (B. Wheeler 2016). The program initiated streaming from Twitter using another Python package, Tweepy (Roesslein 2016), and a set of search terms, and subsequently stored the return of information from Twitter, JSON format tweets, in a text file. In order to prevent the creation of unmanageably large text files, as well as the possibility of the file becoming corrupted over time, every hour *tiipwriter* stopped streaming, closed the text file being written and opened a new text file and began streaming again (B. Wheeler 2016). In order to compensate for tweet surges beyond what the system could handle, if an error occurred for any reason that would crash streaming, the program would begin streaming again or attempt to do so until *n* number of times occurred, as specified by the user.

1.3. Qualitative Review of Hashtags

Relevant Twitter hashtags were collected by searching for each of the candidates in the election. The list of Twitter hashtags was assembled in a qualitative and iterative manner by searching Twitter for candidate names, then subsequently identifying related hashtags;

these hashtags were then recorded and included in the qualitative review. Additionally, review of the Twitter streams of users who identified themselves as clear supporters of a specific candidate was also used to find hashtags that might not occur concurrently with other hashtags. This produced a list of hashtags related to each candidate. In the Democratic Primary, there were only two candidates on the ballot: Bernie Sanders, and Hillary Clinton. In the Republican Primary, there were four names on the ballot: Ted Cruz, John Kasich, Donald Trump, and Ben Carson who pulled out of the race prior to the election and requested his votes be voided (MSNBC, 18 April 2016). Tweets related to Carson were not collected, because it is impossible to determine how many people cast their votes for him since he voided them. This paper uses the terms by which candidates were referred to most commonly in the election, principally the use of first names for Democratic candidates and last names for Republican candidates. Hereafter, Bernie Sanders and Hillary Clinton will be referred to as Bernie and Hillary, respectively and Ted Cruz, John Kasich and Donald Trump will be referred to as Cruz, Kasich and Trump, respectively

1.4. Terms and Hashtags Used for Collection

The list of hashtags used for collecting tweets can be found in Table 1. Data collection began on 4/17/2016 04:49:00 AM EST and finished 4/20/2016 09:42:00 PM EST.

Table 1. Terms and hashtags used for streaming

Candidates	Terms
Donald Trump	'#Trumpers', '#AlwaysTrump', 'Donald Trump', '#Trump', '#DumpTrump', '#TrumpTrain', , '#WeAreTrump', '#VoteTrump', '#TeamTrump', '#VoteTrump2016', '#TrumpWins', '#Trump2016', '#IStandWithTrump', '#TrumpIsAlwaysRight', '#DonaldTrump', '@RealDonaldTrump', '#NeverTrump', 'RealDonaldTrump', '#MakeAmericaGreatAgain'
John Kasich	'John Kasich', '#JohnKasich', '@JohnKasich', '#DropOutKasich', '#Kasich2016', '#KasichGroundGame'
Ted Cruz	'#CruzControl', '#CruzForPresident', '#Trusted', #OnlyCruz', '#UniteWithCruz', '#ChooseCruz','#NeverCruz', '#Cruz2016', '#TedCruz', '#CruzCrew', '@TedCruz', 'Ted Cruz'
Hillary Clinton	'Hillary Clinton', '@HillaryClinton', '#ReadyForHillary', '#HillYes', '#Hillary2016', '#ImWithHer', '#LoveTrumpsHate',
Bernie Sanders	'#StandWithBernie', '#Bernie2016', '#Sanders2016', '#VoteTheBern', '#FeelTheBern', '@SenSanders',

1.5. Election and Demographic Data Collection

County level primary election data was collected from the New York State Board of Elections (New York State Board of Elections 2016). The county geographic data, as well as the socioeconomic and demographic data used in building the prediction model, were retrieved from the US Census Bureau’s site (http://www.census.gov).

2. Post Processing

A Python program was written to parse out the following information from each JSON format tweet: User ID, the unique identifier for each user, Tweet ID, the unique identifier for each tweet, the text of the tweet itself, a timestamp, and geographic information. For simplicity, the output files consisted of simple text files of the User ID number, the tweet ID, and the relevant data field.

2.1. Geographic Information and Assigning Tweet Location

A variety of methods for determining the location of Twitter users can be found in the literature (for an extensive literature review see Ajao et al., 2015). There are three straightforward methods for determining the location of a user given the metadata included in a JSON tweet. These stem directly from three specific data fields that might be included in a tweet: coordinates, user location, and place (Twitter Inc. 2016). A fourth, ad-hoc source resulted from the identification of user locations that were the product of a specific application UberSocial. The individual process for each method is outlined below.

2.1.1. Coordinates

Coordinates (referred to as Coords in Twitter documentation) is an option that users opt-into for their tweets to be accompanied with the location from which they sent it (Twitter Inc. 2016). Since it is optional, only a very small percentage of tweets (1-2%) have coordinates (Hemsley and Eckert 2014). It should be noted, in regards to the purpose of this study that this geographic data reflects the location of the sender of the tweet, not necessarily the home location of the sender.

2.1.2. User Location

User location is an open text field, and is not restricted to verifiable location data; some users go so far as to put coordinate pairs into the field, while others use their state or city only. The field is also home to a number of internet-chic snarky comments like "behind you" or fictional locations like "Middle-Earth" as noted in Crampton et al. (2013). No effort to clean this data was made; the raw field was provided to the Google Maps API for geocoding in order to create a tabular output of location as latitude-longitude coordinate pairs. This process was modeled after Mislove et al. (2011). It is very similar to Quercia et al. (2012), who used a different geocoder, Yahoo!PlaceMaker API, to geocode the same information (Quercia, Capra, and Crowcroft 2012). Information that was not resolvable was removed from the analysis.

2.1.3. UberSocial User Locations

A subset of data in user location was discovered in post-processing for tweets produced using a popular mobile app UberSocial. The platform enables a user to quickly add a coordinate representing their current location to a tweet, which then updates the field "user location" on the fly, and a number of individual users were found to have multiple user locations over the course of several days due to their use of this application. UberSocial allows the user to place a coordinate of their current location in their user location field with a prefix of "ÜT:" or "US:" this allowed for easy identification of these coordinates. Due to the implicit geographic accuracy and granularity of these coordinates they were included in this analysis and considered to have similar accuracy to the source Coordinates.

2.1.4. Place

Places are added to tweets by a user selecting a name out of a drop-down or a search menu (e.g., San Francisco, CA). Places are represented as four bounding coordinate-pairs in Twitter metadata. The upper-left and lower-right coordinate-pairs were averaged and the centroid of the four coordinates was used to resolve the place to a single coordinate-pair.

2.1.5. Hierarchy of Tweet Location Resolution

Each of the aforementioned data sources were independently plotted and subsequently filtered based on whether or not they were found to be inside of New York State. As a tweet may contain all of the above data sources, a hierarchy assigning the "true" location of a tweet based on the data source with the highest assumed accuracy is employed. This method is simple from a logic and calculation stand-point. The location of a tweet in this analysis was determined using the following hierarchy from highest accuracy to lowest: (1) Coordinates, (2) UberSocial user locations (Uber-coordinates), (3) Place, and (4) user location.

For instance, if a tweet contained a coordinate, a place, and a user location within New York, the (assumed) most precise value, the coordinate, was used. The logic behind this decision is as follows: Coordinates reflect the true GPS location of the user at the time of posting, therefore it stands to reason that this would be a truer location than a user picked location (Place), or the user location field which, due to its geocoded nature, is likely to be the lowest quality. If the user had a user location inside of New York (e.g. "Buffalo, NY"), but Coordinates outside of it, the tweet would still be in our dataset since the process excluded tweets outside of New York at the geographic source level, not at the final output level. The individual tweets were then aggregated at the county level.

2.2. Partisanship, Hashtag and First-Last Detection

Partisanship was calculated by adapting a method used by Tsou et al. (2013) and Digrazia et al. (2013), in which a candidate's first and last name is used to designate a tweet as supporting a specific candidate (referred to as first-last detection hereafter). Tsou et al. (2013) and Digrazia et al. (2013) utilize this method because it reduces crosspollination from other Donalds or other Sanderses (i.e. Col. Sanders). The terms utilized for first-last detection in this paper were "Donald Trump", "Hillary Clinton", "Bernie Sanders", "John Kasich", and "Ted Cruz". Neither Tsou et al. (2013) nor Digrazia et al. (2013) stated their exact method for handling calculations when more than one candidate is mentioned in the body of a tweet. Due to the lack of a salient method in the literature a simple method for assigning partisanship was devised. In order to calculate partisanship for each tweet, a Python script searched each tweet for first and last name mentions for each candidate. The candidate with the most mentions was treated as the favored candidate, showing that candidate to be the primary topic of the tweet, regardless of the positive or negative sentiments expressed. The term "favored" was chosen due to the user favoring discussion of one candidate over and above another. The same process was repeated for all tweets using the keywords in Table 1(referred to as hashtag detection hereafter) in lieu of using first-last detection only. If a favored candidate could not be determined, it was counted as “unclear” and excluded from further analysis. Tweets with negative and antagonistic hashtags such as *#DumpTrump* were still included as contributing towards the favor of a candidate. This is supported by O’Connor et al. (2010) who note that sentiment is not necessarily reflective of public opinion, (i.e. all publicity may be good publicity).

Additional methods for review included tabulating tweets by party based on the affiliation of the favored candidate, as well as summing the frequency of individual mentions summed by party and then compared by party. To illustrate, if the count of topics associated with Trump, Kasich and Cruz was greater than the count of topics associated with Hillary and Bernie, it was counted as “Republican”, and vice-versa “Democrat”. If the numbers of associated topics mentioned were equal, partisanship was counted as “Unclear”.

2.3. Spatiotemporal Binning

Each tweet comes with a timestamp in UNIX time (also known as epoch-time). UNIX time is calculated in seconds since January 1st, 1970. The tweets gathered for this study were segmented into seven time blocks according to the following list: (1) April 17th, (2) April 18th, (3) April 19th (before the election was called 9PM), (4) All tweets collected prior to the 19th at 9pm EST, (5) April 19th (after election was called), (6) April 20th and (7) all tweets collected. Note, tweets collected before the election outcome was anounced (9 PM) was used to avoid possible inflation by news media tweets discussing winners or gloating tweets that might skew results.

The tweets were tabulated at the county level for partisanship determined by hashtag and first-last detection, as well as using several subsets of tweet best-available-location data (1) Coords only, (2) Coords and Uber-coordinates (3) Coords, Uber-coordinates and Place, and (4) Coords, Uber-coordinates, Place and user location.

3. Analysis of Tweet and Vote Data

3.1. Vote Share and Twitter Share

The following equations were modeled after Digrazia et al. (2013). Twitter share (tw_s) is the sum of candidate $i's$ total tweets divided by the sum of tweets for n candidates by political party. Vote share (v_s), similarly is the sum of candidate $i's$ total votes divided by the sum of tweets for n candidates by political party. Equation 1, below, was used to calculate Twitter share (tw_s) for each county for each aforementioned spatiotemporal bin. Equation 2 was used to calculate vote share (v_s) at the county level using the official election results.

$$tw_S(i) = \frac{tw_{Candidate(i)}}{\sum_{j=1}^{n} tw_{Candidate(j)}} \quad x\,100 \qquad (2.1)$$

$$v_S(i) = \frac{v_{Candidate(i)}}{\sum_{j=1}^{n} v_{Candidate(j)}} \; x\,100 \qquad (2.2)$$

3.2. Correlation Analysis between Twitter Share and Vote Share

This section of the study examines whether Twitter share related to a candidate, or a grouping of candidates, is correlated with vote share at the county level. Because the vote share data was not normally distributed (Zar 2005), Spearman's Rank Correlation Coefficient (Spearman's rho) is used. Tsou et al. (2013) used all candidates in a single correlation test (i.e. Trump, Kasich and Cruz all in the same category), in addition to that, this research will also examine the correlation for each candidate and their grouping for each temporal bin, by geographic data sources, and by method of detection (hashtag detection vs. first-last detection).

4. Results and Discussion

4.1. Tweets Collected and Summary Statistics

A total of 4.47 million tweets were collected during the collection period and of those, 2.37 million tweets were geospatially placed. Of the 2.37 million geospatially placed tweets, 220,593 tweets were geo-located within New York State. This required geocoding the total 600,000 unique user locations, in order to select the 41,376 users that were placed within New York State using their user location information. These 41,376 users produced 212,762 tweets (~96.4% of tweets collected within New York State). The locations of the remaining tweets were resolved through geotags.

4.2. Twitter Users and Tweets in New York

43,650 users were collected with a total of 220,593 tweets and a frequency of ~5.05 tweets per user (Figure 2 illustrates the geographic spread). The population of New York was estimated to be 19,745,289 in 2016 and with 220,593 tweets, 0.011 tweets per capita for the entire state of New York in 2016 (roughly 1 in 90). The rate of tweets per user for the entire state of New York is 5.05. The number of Twitter users in New York 43,650 collected in our study, breaks down to roughly 1 in 452 people (0.00221 users per capita).

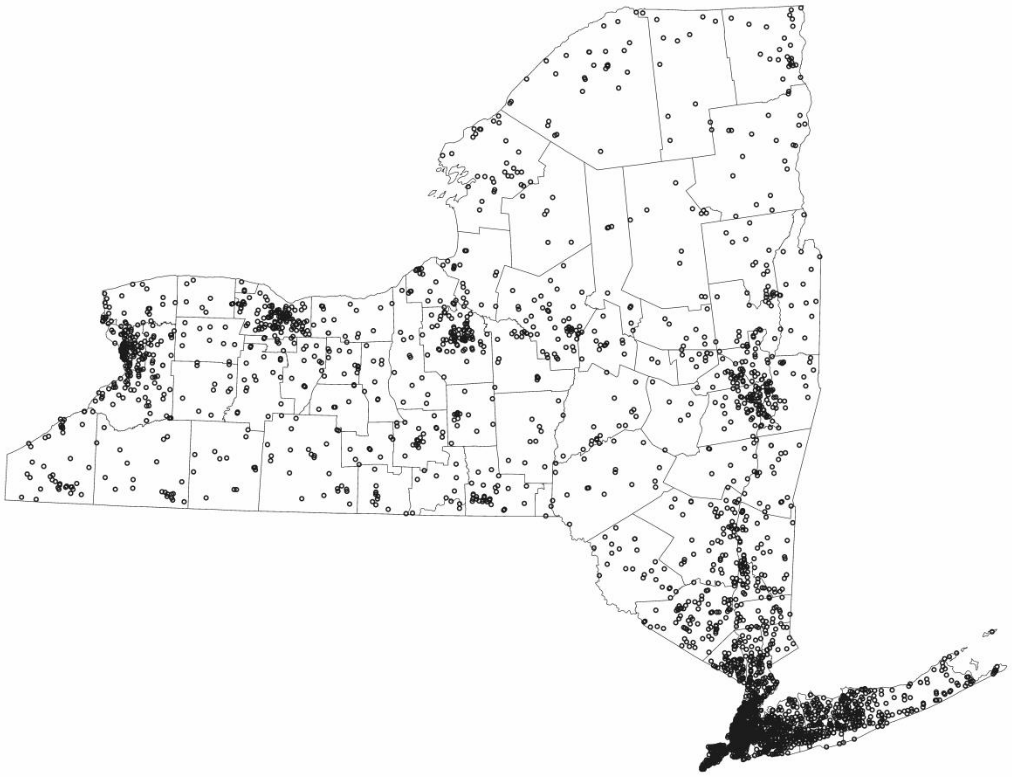

Figure 2. Plot of all tweets collected within study area

4.3. Patterns and Trends in the Location of Tweeters

4.3.1. Candidate Breakdown over Study Period

Trump had the most tweets over the study period, 92,796, while Hillary had the second largest count of tweets, 45,104, followed by Bernie with 35,906, Cruz with 23,230, tweets about an unclear candidate with 20,025, and lastly Kasich with 3,532 (Figure 3).

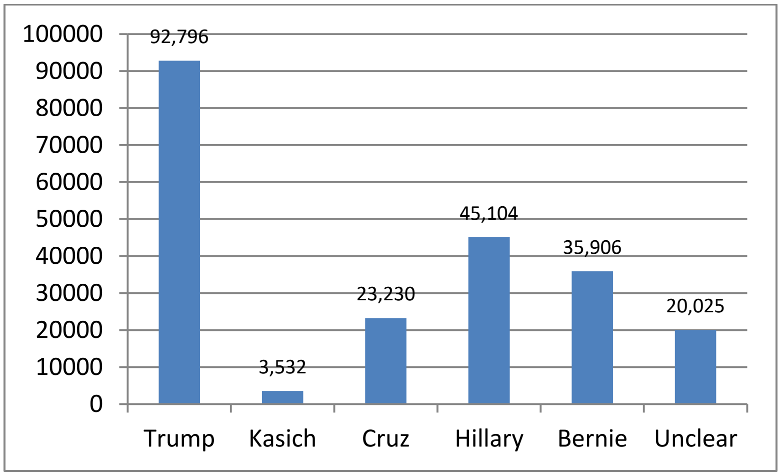

Figure 3. Tweets by candidate favored in the overall study period

For both methods of detection, all candidates had an upward trending number of tweets favoring them from the 17th to the 19th, followed by a steep decline the day after the election. Many of the favored tweet counts returned to levels similar to or below the number of tweets on the 17th. Note: the 17th does not include a 24hr day, as collection began at 4:49AM EST, and the 20th, does not include a 24hr day, as collection ended at 9:42PM.

The total number of tweets for Bernie between 12am and 9pm were greater on the day of the election than for Hillary during the same time period, but Hillary demonstrated a greater number of tweets during all other time periods. Trump, Cruz, and Kasich maintained their ordinal positions throughout the entire collection period for both first-last detection and hashtag detection (Table 2). Hillary demonstrated almost twice as many tweets as Bernie all time periods using first last detection. One anomaly of interest on the day of the election, using first-last detection the total numbers of tweets for Bernie were less than half of those for Hillary, while using hashtag detection Bernie had a greater number of tweets.

When using first-last detection, the total detection decreased from 220,593 to 57,630 total tweets. Specific candidates retained a higher number of tweets while other candidates had a much lower percentage of their tweets where they were mentioned using their first and last name (See Table 2). For instance, Bernie Sanders only had 12.32% of his original tweet count using hashtags and first and last names. This was likely due to the accidental exclusion of the term “Bernie Sanders” from the key terms (refer to Table 1), which were used for identifying tweets, this likely influenced the percentage of tweets in this category. Ted Cruz retained the highest number of tweets with ~40.28% of his total tweet count, followed by John Kasich who retained 40.01% of his tweet count. Despite varied percentages of tweet count retention, ordinal relationships within parties remained the same.

Table 2: Candidate tweets by day for hashtag (#) and first-last (FL) detection.

	Trump		Kasich		Cruz		Hillary		Bernie		Unclear		Total	
Period	#	FL	#	FL	#	FL	#	FL	#	FL	#	FL	#	FL
17th	17373	3716	508	192	4160	1408	7449	2149	6473	853	3927	1390	39890	9708
18th	26257	6950	1270	437	6235	2788	11272	3274	8522	1199	5023	1560	58579	16208
19th til 9PM	25477	6252	965	456	6226	2521	12186	2583	14275	1204	4411	1191	63540	14207
19th after 9PM	9769	3452	338	163	1957	816	8647	2620	3269	620	3016	1054	26996	8725
20th	13920	3283	451	165	4652	1825	5550	1816	3367	547	3648	1146	31588	8782
All Before 9pm	69107	16918	2743	1085	16621	6717	30907	8006	29270	3256	13361	4141	162009	40123
Totals	92796	23653	3532	1413	23230	9358	45104	12442	35906	4423	20025	6341	220593	57630
% Total	-	25.49%	-	40.01%	-	40.28%	-	27.59%	-	12.32%	-	31.67%	-	26.13%

4.3.2. Party Breakdown

Two methods were used to produce estimates of partisanship; The first is a summation of each of the identified candidates tabulated in the candidate breakdown into their respective party. It reflects a sum across rows of Table 2, by party of affiliation and method of detection (Table 3). The second method is candidate agnostic and shows a much smaller number of unclear tweets, Partisanship for Table 4 was calculated using the terms in Table 1 by party of affiliation, and the party with the most related mentions was chosen as the favored topic for the tweet.

Table 3: Tweet partisanship by favored candidate by day

Source	Democrat		Republican		Unclear	
	#	FL	#	FL	#	FL
17th	13,922	3,002	22,041	5,316	3,927	1,390
18th	19,794	4,473	33,762	10,175	5,023	1,560
19th til 9PM	26,461	3,787	32,668	9,229	4,411	1,191
All Before 9PM, 19th	60,177	11,262	88,471	24,720	13,361	4,141
19th after 9PM	11,916	3,240	12,064	4,431	3,016	1,054
20th	8,917	2,363	19,023	5,273	3,648	1,146
Count	81,010	16,865	119,558	34,424	20,025	6,341

Table 4: Tweet partisanship by favored topics by day

Source	Democrat		Republican		Unclear	
	#	FL	#	FL	#	FL
17th	15,007	3,169	23,925	6,033	958	506
18th	21,342	4,690	35,988	11,033	1,249	485
19th til 9PM	27,817	3,935	34,514	9,748	1,209	524
All Before 9PM, 19th	64,166	11,794	94,427	26,814	3,416	1,515
19th after 9PM	12,949	3,408	13,078	4,750	969	567
20th	9,907	2,479	20,816	5,908	865	395
Count	87,022	17,681	128,321	37,472	5,250	2,477

4.4. Characteristics of the location of tweeters

Tweets collected for the state of New York demonstrated a high clustering around urban populations, particularly, in the New York Tri-State Area. Looking at maps of overall placement, there also appears to be clustering around the urban centers of Buffalo, Rochester, Syracuse, and Albany tabulations of these values can be found in Appendix C. This is further elucidated in Table 6. New York's primary metropolitan area, referred to as the Tri-State Area, was the largest region with high volume of tweets. Long Island (Nassau and Suffolk Counties) and New York City contributed ~78% of the tweets.

4.4.1. Tweet Location Source and Partisanship

In general, tweets focused on Democratic candidates tended to have a higher quality spatial location source. This is illustrated in Table 5 by Coordinates having three times as many mentions of Democratic candidates as Republican candidates (with less than half the population of tweets), as well as the Uber-coordinates tweet volume for Democratic candidates being 1.7x greater than Republican candidates. Conversely, Place mentions were relatively similar with Republicans having slightly more than Democrats (51.80% to 45.06%). User location was the only location collected more frequently for tweets mentioning Republican candidates than for tweets mentioning Democratic candidates (58.57% to 39.07% respectively). There are far more tweets mentioning Republican candidates with greater frequency. Comparing this to Table 5 below, one can see that tweets favoring Democrats had an increased rate of higher quality location information, although this phenomenon is not reflected as clearly in the first-last detection method.

Table 5. Tweet sources and partisanship. This table displays the quantity of tweets from Table 4 with different source locations, as well as their party alignment, it shows the percentage of each spatial source (Coordinates, Uber-coordinates, Place, and user location).

Tweet Sources	Democrats				Republicans				Unclear			
	#	%Source	FL	%Source	#	%Source	FL	%Source	#	%Source	FL	%Source
Coordinates	694	72.67%	68	53.54%	244	25.55%	56	44.09%	17	1.78%	3	2.36%
Uber-coordinates	422	61.25%	83	42.56%	254	36.87%	102	52.31%	13	1.89%	10	5.13%
Place	2788	45.06%	286	29.67%	3205	51.80%	629	65.25%	194	3.14%	49	5.08%

User locations	83118	39.07%	17244	30.60%	124618	58.57%	36685	65.11%	5026	2.36%	2415	4.29%
Total	87022		17681		128321		37472		5250		2477	

4.4.2. City and Town Clusters

The majority of tweets occurred in and around cities and their surrounding suburbs, the city with the largest population of tweets being New York City. Almost 90% of all tweets originated in the top three metropolitan areas (See Appendix B). 653 of 994 cities in New York had at least one tweet geocoded or placed within its boundaries. Table 6 shows tweet counts for the top ten most tweeting cities and towns within New York, which also make up almost 90% of the tweets collected.

Table 6: Top ten cities and towns in New York State by tweet count

Rank	City or Town	Tweets	% Total
1	New York City*	160,919	72.95%
2	Buffalo	8,816	4.00%
3	Islip**	4,690	2.13%
4	Rochester	4,029	1.83%
5	Syracuse	3,176	1.44%
6	Albany	2,975	1.35%
7	NewBurgh*	2,206	1.00%
8	Brookhaven**	1,598	0.72%
9	Hempstead**	1,228	0.56%
10	Mount Pleasant*	1,102	0.50%

*Tri-State Area
**Long Island

4.5. Tweets by Source of Spatial Information

The highest volume source of location information when using hashtag-detection was by far user location, followed by Place, then Coordinates, and Uber-coordinates. Similarly, utilizing first-last detection (Table 7) reflected the same structure with the exception that Uber-coordinates contained almost 50% more tweets than coordinates using first-last detection. Figures 2 and 5 display the spatial distributions for all tweets collected.

Table 7. Tweet count of each source in New York State.

Source	User Location	Coordinates	Place	Uber-coordinates	Total
Count (#)	212,762	955	6,187	689	220,593
Count (FL)	56,344	127	964	195	57,630

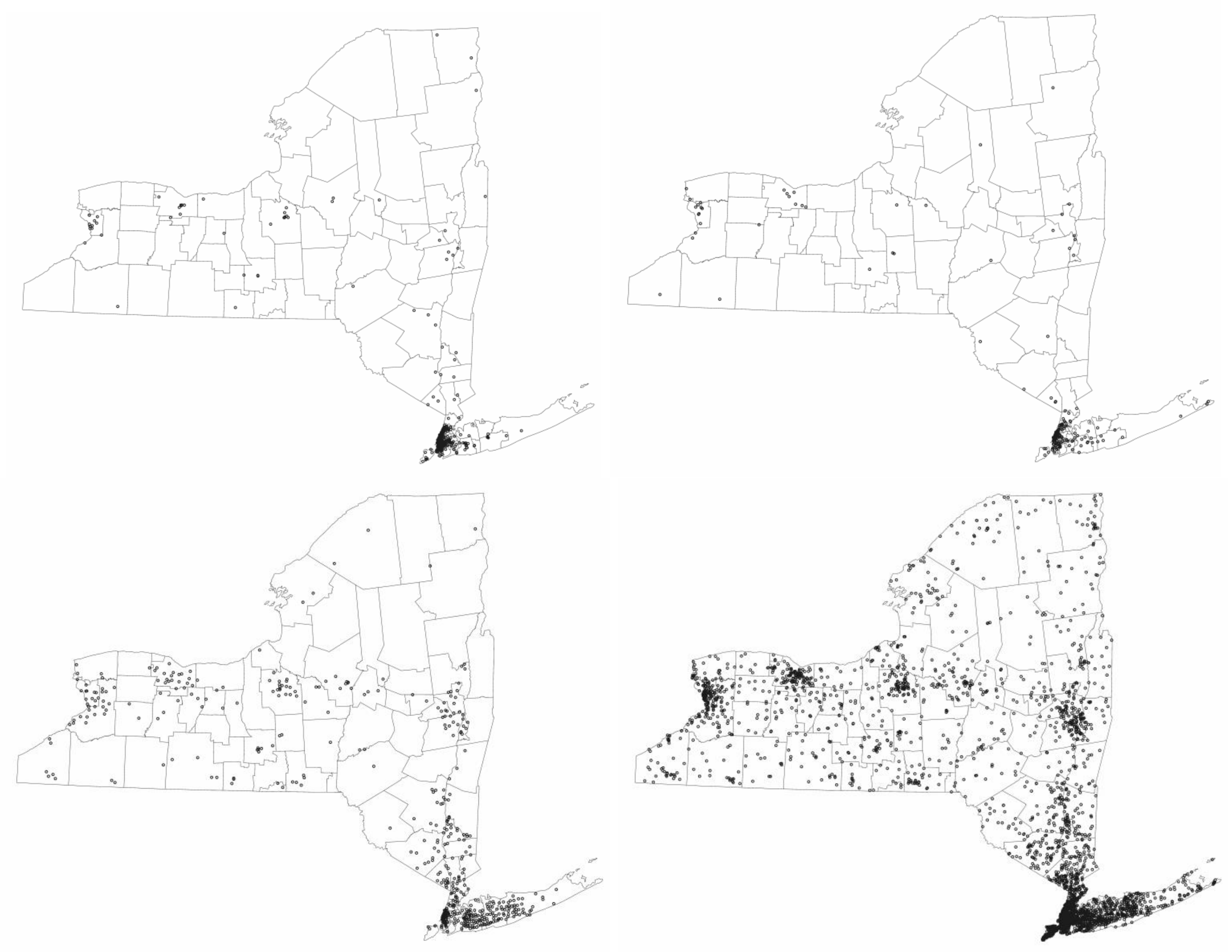

Figure 4: Various Spatial Data Sources. by best available data source (a) Coordinates (b) Uber-coordinates, (c) Place (d) User location

4.5.1. Tweet Location Source and Temporal Binning

The following table is the results of binning each source by day and time period (Table 8). While all sources increase over time reflecting the increase of tweets, higher quality sources increase at a greater rate than user location.

Table 8: Tweet count of each source by period, for hashtag (#) and first-last (FL) detection.

Source	User Location		Coordinates		Place		Uber-coordinates		Total	
	#	FL	#	FL	#	FL	#	FL	#	FL
17th	38628	9526	173	31	1000	130	89	21	39890	9708
18th	56525	15786	178	25	1692	333	184	64	58579	16208
19th til 9PM	61030	13904	449	34	1814	219	247	50	63540	14207
19th after 9PM	25808	8503	114	26	970	159	104	37	26996	8725
20th	30771	8625	41	11	711	123	65	23	31588	8782
All Before 9pm, Day of Election	156183	39216	800	90	4506	682	520	135	162009	40123
Count	212762	56344	955	127	6187	964	689	195	220593	57630

4.6. Correlation between Tweets and Demonstrated Opinion

The below table shows a similar analysis to Tsou et al. (2013) and Digrazia et al. (2013) for all spatial sources, which includes user location, Uber-coordinates, Place and Coordinates. It has been broken down by Republican and Democrat, and individual candidates. Twitter share here is calculated using most-mentioned candidate for hashtag-detection and first-last detection (Table 9). Worth noting from Table 9, Kasich and Cruz have a negative correlation with Twitter share, while the groupings of Trump and Kasich (T&K) and Trump and Cruz (T&C) both showed a substantial positive correlations , 0.78 and 0.61 respectively for the day of the election. These positive correlations were higher than Republicans as a group, which still resulted in a correlation of .513 for the day before the election; the diminished correlation is contributed to by the overall spread within the race at the bottom (see Figure 5).

Hillary and Bernie did not have the same structure as the Republican candidates, but both displayed a low correlation between Twitter and their overall performance in the election. After the election however, there is a correlation based on an increase in Hillary discussion over and above Bernie Sanders following her win in the election. There was a noticeable lift/ correlation on the day of the election when looking at Democrats as a group. However, evaluating them individually you see that Hillary displays a higher correlation with Twitter while Bernie's values stay below 5% correlation with the exception of the day after the election.

Table 9. Vote share vs. Twitter share, by time period for all spatial sources, and hashtag detection(#), and first-last detection (FL)

	TS17		TS18		TS19b4		TS19aft		TS20		TS_b4		*TS*	
Candidates	#	FL	#	FL	#	FL	#	FL	#	FL	#	FL	#	FL
Republicans	0.415	0.371	0.513	0.426	0.504	0.449	0.499	0.476	0.360	0.310	0.511	0.443	0.493	0.470
Democrats	-0.062	-0.173	-0.066	-0.165	0.199	-0.109	-0.024	-0.118	0.224	0.011	-0.013	-0.226	0.061	-0.239
All	0.455	0.256	0.529	0.319	0.639	0.383	0.533	0.373	0.437	0.229	0.632	0.417	0.631	0.454
Trump (T)	-0.025	-0.034	0.110	0.004	0.050	0.150	0.002	-0.053	0.076	0.030	0.106	0.171	0.110	0.116
Kasich (K)	0.069	0.141	0.229	0.159	0.244	0.069	0.133	0.007	0.014	-0.079	0.224	0.193	0.249	0.195
Cruz (C)	-0.193	-0.164	-0.184	-0.140	-0.168	-0.155	-0.091	-0.084	0.029	-0.024	-0.187	-0.078	-0.172	-0.064
T&K	0.665	0.600	0.763	0.712	0.783	0.768	0.726	0.705	0.719	0.700	0.777	0.743	0.784	0.768
T&C	0.540	0.461	0.651	0.565	0.670	0.607	0.640	0.608	0.486	0.405	0.684	0.612	0.681	0.653
K&C	-0.417	-0.334	-0.330	-0.427	-0.424	-0.467	-0.347	-0.390	-0.469	-0.506	-0.415	-0.429	-0.468	-0.452
Hillary	0.108	0.062	0.155	0.205	0.061	0.187	0.217	0.197	0.352	0.390	0.037	0.135	0.150	0.199
Bernie	0.003	-0.187	0.035	-0.028	0.061	-0.101	0.046	-0.210	0.291	-0.217	0.037	0.094	0.150	0.158

4.6.1. Correlations for Other Subsets of Spatial Data

Tables of Spearman's rho correlations were calculated at each spatiotemporal break for each method of detection, for each of the following subsets of spatial data: Coordinates, Coordinates and Uber-coordinates, Coordinates, Uber-coordinates and Place. Due to the sparseness of data in those subsets, only Table 9 has been included in results, the remaining tables have been included in Appendix A for reference.

4.6.2. Contextual Observations

Contextual observations related to the correlations of Trump, Kasich and Cruz can be found in Figure 5. Figure 5 is an example of the clustering of data points that appears in the data for the Republican candidates. Each data point represents a county in NY, and each county is represented three times (once for each candidate). Figure 5 shows Trump's values primarily clustered in the high-high region of the map. Although Kasich has much higher vote share than Cruz, his Twitter share is quite low when compared to the other candidates.

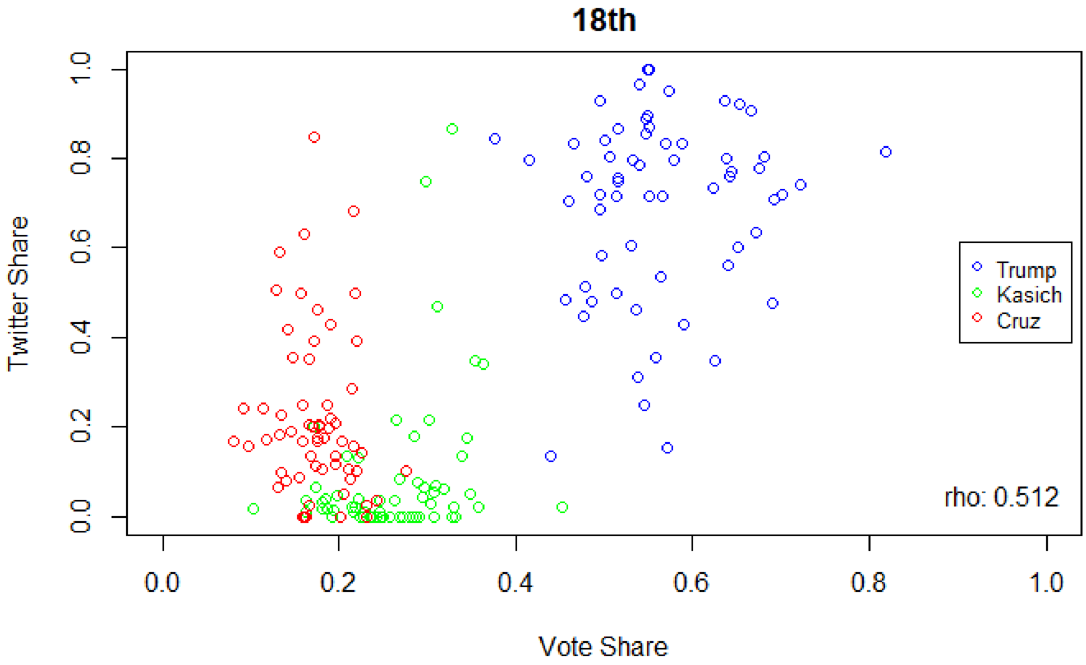

Figure 5. Twitter share (TS18) vs. vote share by county for each Republican candidate. Each county is represented three times, once per candidate.

5. Discussion of Chapter Two

5.1. Discussion of Capturing a Larger Cross-Section of the Political Conversation on Twitter

An enormous number of tweets were captured in 4 days of collection, 4.47 million, roughly 50% of those tweets were resolved spatially (2.27 million tweets) which reflects a rate that is lower than the 67% noted in Barbera & Rivero (2015). This may be partially influenced by the desire for individuals to distance themselves from their political opinions. One of the aims of this paper was to capture a greater number of tweets to widen the breadth of the discourse being analyzed. The method described in Tsou et al. (2013) for collecting relevant tweets (circa 2012) using a spatial filter for multiple locations in the US only led to 82,000 tweets being collected in 3 months. Conversely, with my method, 220,593 relevant tweets were collected in four days using candidate relevant hashtags, only 26% of which included a first-last name combination of a candidate. This suggests that a greater number of tweets can be collected using more than first-last detection coupled with user-location.

An interesting trend when looking at first-last detection is the percentage of the overall sample that contained the candidates' full name (see Table 2). Using first-last detection, both Trump and Hillary had a lower rate of tweets using their first and last name than either Kasich or Cruz. Kasich and Cruz both had roughly 40% of their original tweet count, while Trump and Hillary had 25.49% and 27.59% respectively. Qualitatively, this reflects the common use of *Hillary, Trump,* and *Bernie* as more common than using their full names in the election. It is possible that these candidates were sampled with a greater number of hashtags leading to an under-representation of their first-last name proportions. It is also likely that due to the emphasis on Twitter as an ideological battleground for these three candidates, their popularity led to a greater number of hashtags than for non-trending candidates. This phenomenon is reflected in viral movements such as #occupy, that lead to the massive creation of hashtags surrounding their topic (Hemsley and Eckert 2014). It is important to note that Bernie Sanders displays a distinctly low percentage of first-last tweets in comparison to Hillary Clinton (see Table 2); this is likely due to an error in the collection period where the term *Bernie Sanders* was inadvertently excluded from the collection terms. This error renders the democratic side of tweet collection practically useless for drawing overarching conclusions about the population's tweeting behavior.

One of the key fundamental limitations of this study and this method is the difficulty of determining what percentage of the true tweet population was sampled. This manifests itself in several ways: (1) 1% limitation on total twitter stream sampling; (2) addition of new trending hashtags or non-searchable tweets that are clear to humans but not to Twitter crawlers (see Tufekci, 2014) and (3) the two aforementioned issues make it impossible to know if oversampling is occurring on one population over another.

The 1% limitation on total Twitter stream sampling reduces the chance that collecting all terms via a single streaming connection will capture a complete population. All terms collected on a single streaming instance muddles the ability to evaluate first-last detection against detection using hashtags. A more appropriate experimental design to determine differences in collection terms should collect each candidate separate from one another, as well as their hashtags separate from first-last name detection. This design would allow a truer understanding of the population captured through each method.

The second issue is the addition of new-trending hashtags during the period of collection. These are impossible to anticipate; as such it can be assumed that one would not be able to capture all relevant twitter conversation on a single topic. Additionally, Tufekci (2014) describes the use of an “invisible” method of interaction through the use of screenshots (which are functionally invisible to our methods). This kind of tweet coupled with newly introduced trending hashtags could render a large proportion of the true population of tweets invisible to crawling.

5.2. Discussion of Correlation between Twitter Share and Vote Share

Results of the Spearman’s rho analyses reveal some interesting trends. For the Republican side, the highest correlations were found for groupings of candidates, specifically those between Trump (the dominant candidate) and an opponent. Correlations with the highest values were produced using all spatial sources (i.e., Coordinates, user location, Uber-coordinates, and Place). The highest correlation found was for the grouping of Trump and Kasich (T&K for Twitter share of total tweets collected (TS) with a positive correlation of 0.784, followed by 0.783 for Twitter share share of tweets on the day of election before 9PM (TS19b4). The third highest and perhaps most relevant were the correlations found using all the Republican candidates and all spatial sources. It is clear that when all candidates are compared together a higher rho value is achieved. This study achieves very similar correlations to the findings of Tsou et al. (2013). Tsou found a correlation of 0.56 for one day before the election, with another high correlation value for a narrower subset of the candidates returned a correlation of 0.75 for the day of the election. Similarly, this study achieved a correlation of 0.513 for Twitter share for Republicans the day before the election and 0.783 for a subset(i.e., Trump and Kasich) on the day of the election. These minimally differing results suggest that Twitter performance correlates with the performance of a candidate in real life at the polling booth and that the county level results in this study utilizing hashtag detection and user location as a source reflect similar values to those achieved by Tsou et al. (2013). This study also noted higher correlations when using hashtag detection rather than just using first-last detection.

One challenge is that this trend does not continue for the individual candidates. The highest correlation found in the individual candidates was with Kasich The correlation between vote share and Twitter share for Kasich varied from 0.24 (TS19b4) to 0.014 (TS20), with correlation values between 0.069 and 0.24 leading up to the election (TS17, TS18, TS19b4), followed by a dip after the election (TS19_aft, TS20) (see Table 9). These dips may be related to a flurry of news media and opportunistic twitter-bots that would add noise to the data, leading to a decrease in correlation between Twitter share and vote share, but more research would be needed to investigate if there are motivators for the decrease in correlation or if they are just the product of noise. It is likely that over-segmentation of data has led to noise in the data. There were stronger correlations for variables representing many days (TS_B4) than many other time periods, suggesting that longer time increments may be necessary and more consistent in capturing relationships at finer-scales.

It is evident that there is a relationship between Twitter share and political performance, and that there are several factors that are at play in this election that prevent a straightforward test of public opinion via Twitter. Each of the Republican candidates in New York were in themselves very distinct in their voting performance as well as Twitter performance. Ted Cruz and John Kasich both display a wide range of correlations with Twitter performance. Figure 5 shows clear regions in a scatter plot; Kasich is associated with strong political performance (relative to Cruz) with a low Twitter share. The low values for Twitter share returned for Kasich reflect the overall "silent" majority of the voting population, the older, less tech-savvy crowd with a lower propensity to tweet. He performed much better than Ted Cruz in New York; Ted Cruz's strong presence on the New York Twitterscape failed to translate in the real world at the polling booths. This may have been in part due to the small number of white Evangelicals in New York, whom had been Ted Cruz' main supporters (CBS NEWS, April 19th). From a qualitative stand-point, it was clear that John Kasich had little traction on Twitter from a cursory investigation of his Twitter page and his supporters. This is also reflected in the relatively few number of search terms that were usable for identifying Kasich and the corresponding political conversations. Trump had an overall strong response in New York. This is heavily impacted by New York being a home state for Donald Trump and so he had an advantage that is well reflected in the numbers. Trumps supporters were very active on Twitter, and this coupled with his strong performance across the state in the polls is reflected in Figure 5.

In looking at the correlations on the Democratic side, Bernie's reflected Twitter share is likely lower than should be expected based on the percentages found of first-last detection (see Table 4), meaning Hillary tweets are likely over-represented. Such poor sampling would lead to a misrepresentation of any true correlations that may lie in the data. While a positive correlation as high as 0.199 was found the day of the election for Democrats, it is impossible to know if the correlation is truly relevant due to the underlying sampling error.

The primary goal of testing correlations between Twitter share and vote share was to see if detecting public opinion at the county level was feasible and would reveal similar correlations to those found in the literature for other scales (e.g., state level). An additional goal was to see if specific geographic sources (e.g., coordinates) would produce a similar or superior result to those found in the literature. This was not the case in my data; it is clear based on these findings that county-level detection is not feasible for a more limited set of source data over a period of three days. It may be feasible to produce similar results using other methods over the course of weeks or months. However, that is beyond the scope of this study. One limitation to this study was the minimal geographic scope and lack of political performance heterogeneity on the Republican side. Trump dominated his opponents across the board, on Twitter and in the polling booth. A study focusing on a region or multiple regions where a candidate experiences both low and high performance would have been more appropriate than the case study chosen and would have allowed for a more refined investigation into the geography of the Twitterscape.

CHAPTER THREE: FINE-SCALE PUBLIC OPINION AND GEOGRAPHIC WEIGHTED REGRESSION

1. Using County Demographics and Tweets to Predict Election Outcomes

1.1. OLS

To limit the scope of this study, vote share was modeled for Donald Trump only. Typical election models predict election outcomes (vote share of the candidate) by utilizing county-level demographics such as; percentage of the population that is white, percentage of the population with a Bachelor's degree and above, percentage of the population that is female, as well as median household income, and median age. Our study additionally includes three Twitter derived variables: the candidate's Twitter share two days prior to the election(TS17), Twitter share one day prior (TS18), and Twitter share on the day of the election before the election was called (TS19b4). Digrazia et al. (2013) also included district partisanship, incumbency, and CNN coverage as variables. These variables were not used in this research due to either non-applicability or being outside of the scope of this thesis. Two OLS models are implemented (1) an OLS of typical demographic data without Twitter variables, and (2) an OLS model with the three additional Twitter derived variables.

1.2. Formalizing a Procedure for Geographically Weighted Regression

New York is a state with high spatial heterogeneity in landscape, urbanity (or contrarily rurality), and demographics. This makes it challenging to build an effective model to predict its election outcomes. Certain landscape-scale demographics contribute to OLS model performance; however, such a model may miss certain non-stationary attributes. This is well described in Calvo & Escolar (2003). In fact, if spatial non-stationarity exists in a model, it violates the assumption of an OLS that the samples are independent, and identically distributed (i.i.d.). No clearly laid out procedure for iterating through this process was found in the literature, and as such the following procedure is proposed here and expanded upon in the following sections:

First, produce a properly specified OLS regression. Then compute Moran's I for the residuals of the OLS regression (visualizing as necessary). If residuals are significantly spatially autocorrelated (and the model is properly specified), then initiate a Geographic Weighted Regression. Next correct p-values for the Multiple-Hypothesis Testing Problem, and display the variables, coefficients, t-values (significance) visually to evaluate the findings of the GWR. Finally, compute Moran's I for residuals of the GWR regression.

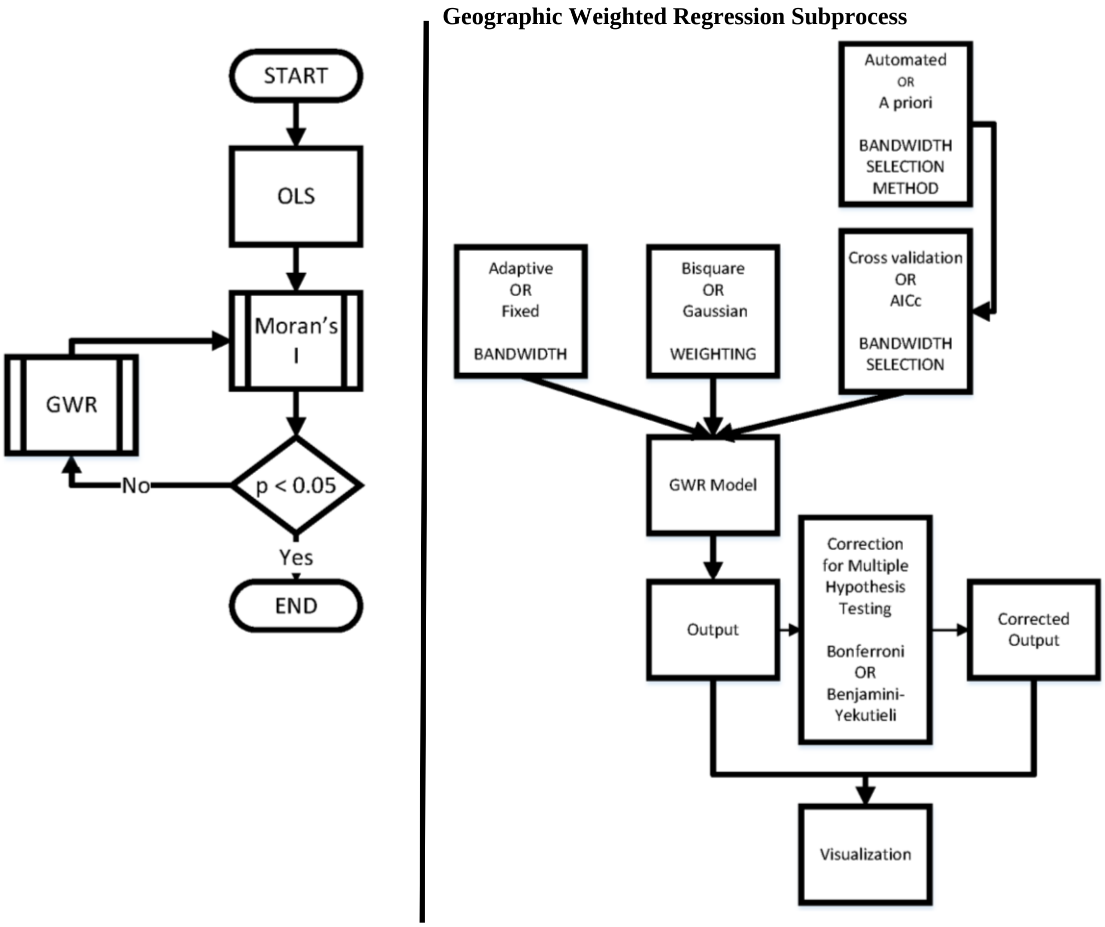
Geographic Weighted Regression Subprocess
START
OLS
Moran's I
GWR
p < 0.05
No
Yes
END
Automated OR A priori BANDWIDTH SELECTION METHOD
Adaptive OR Fixed BANDWIDTH
Bisquare OR Gaussian WEIGHTING
Cross validation OR AICc BANDWIDTH SELECTION
GWR Model
Output
Correction for Multiple Hypothesis Testing Bonferroni OR Benjamini-Yekutieli
Corrected Output
Visualization

Figure 6. Flow chart for evaluating an OLS for non-stationary behavior.

One measure for understanding if errors within a model are spatially dependent or not is Moran's I, which looks for spatial autocorrelation (Moran 1950). Moran's I values range from -1 to +1. A value of -1 suggests negative spatial autocorrelation (similar or high values dispersed greater than can be expected at random), and a value of +1 suggests positive spatial autocorrelation (similar or high values clustered greater than can be expected at random). A value of zero reflects no autocorrelation. Spatial autocorrelation in residuals is a key indicator that specific underlying geographic trends may be missing from a model.

Anselin et al. (2006) emphasize the natural process of checking Moran's I when developing a model prior to moving to a form of spatially-aware regression, although not explicitly mentioning GWR (Anselin, Syabri, and Kho 2006). Kupfer & Farris (2007) note that one method for assessing the improvement of a model is to determine if there is a reduction in the spatial autocorrelation of the residuals as shown by Zhang et al. (2005). Zhang et al. (2005) use the reduction of spatial autocorrelation over an OLS model as an indicator of an improved model. A geographically weighted regression (GWR) allows model coefficients to vary over space and by doing so allows for non-stationary relationships with explanatory variables. This allowance better accommodates for underlying spatial structures that may cause spatial autocorrelation in residuals when evaluate using global methods (Zhang, Gove, and Heath 2005). Spatial autocorrelation itself can be tested using a Monte Carlo permutation test as described by Fotheringham et al. (2002). This method involves producing *n* number of random variations of the source-data and locations, and re-computing Moran's I; If Moran's I from the test is significantly different from the mean when ($p < .05$), then it is likely that the data is clustered or dispersed. For this study 999 permutations were conducted. Queen's method was used for computing neighbors (allowing edge-connected and corner-connected neighbors).

A geographic weighted regression should use the equation from a properly specified OLS model in a hope to understand additional information about any nonstationary relationships that may exist. GWR is a relatively new statistical model that has become increasingly more popular in the field of geography. The mathematics behind a GWR is well described in Fotheringham et al. (2002). The basic underpinnings of a GWR model rely on a set of coordinates, (u_i ,v_i), that represent each data point in space (Fotheringham, Brundson, and Charlton 2002). For each point (u_i ,v_i) a separate regression is calculated, such that for n data points, there will be n regression equations calculated with its own set of coefficients ($\boldsymbol{\beta}$), R^2, residuals and $\hat{y}$.

A global OLS would calculate a regression using the following equation: (Fotheringham, Brundson, and Charlton 2002)

$$y_i = \beta_0 + \sum_{\mathrm{k}} \beta_k \mathrm{x}_{\mathrm{ik}} + \varepsilon_{\mathrm{i}} \quad (3.1)$$

While a geographic weighted regression would use Equation 3.2: (Fotheringham, Brundson, and Charlton 2002)

$$y_i = \beta_0(u_i, v_i) + \sum_{\mathrm{k}} \beta_k(u_i, v_i)\mathrm{x}_{\mathrm{ik}} + \varepsilon_{\mathrm{i}} \quad (3.2)$$

This is subject to a weighting scheme for each record as such:

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = (\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{y} \quad (3.3)$$

The above formula (Eq. 3.2) is an adaptation of the basic OLS formula for finding a beta matrix, where $\mathbf{X}$ is the design matrix $(n \times p)$ and $\boldsymbol{W}(u_i, v_i)$ is a weighted matrix calculated at each point (u_i, v_i), essentially n times (Although this would only hold true if predictions were calculated only at known locations). $\boldsymbol{y}$ is a vector of the dependent variable; in this case, it is vote share (D. C. Wheeler and Páez 2010). $\mathbf{X}$ is the design matrix of the explanatory variables. $\widehat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$ matrix, and $\boldsymbol{W}(u_i, v_i)$ is a matrix representing the weights at each point. The primary reason for demonstrating this is in order to discuss the issue of distribution of weighting, and its related issues. For each point (u_i ,v_i), a matrix of weights is calculated. This matrix of weights can be calculated in a number of different schemes, however, kernel methods using bi-square or Gaussian functions are most common (Fotheringham, Brundson, and Charlton 2002). A Gaussian kernel produces smoother model predictions as the model is still affected by features outside of the bandwidth (nominally), conversely a bi-square distribution quickly drops to weighted values of 0 after the bandwidth has been reached (Guo, Ma, and Zhang 2008). For the purposes of this study, a bi-square kernel will be used. There are two kinds of kernels: fixed and adaptive. Adaptive kernels use a specific number of nearest neighbors of (u_i ,v_i), while a fixed kernel would be calculated using a specific distance from (u_i ,v_i). An adaptive kernel is considered most useful for observations that are irregularly spaced (D. C. Wheeler and Páez 2010). It has been suggested as distinctly useful for human geography, as the density of observations may differ between cities and rural areas (Brundson, Fotheringham & Charlton (1999). Similarly, Mennis (2006) chooses an adaptive bandwidth to account for the varying sizes of administrative boundaries (e.g. census, tracts). This research also uses an adaptive kernel to generate the weight matrix for GWR.

Wheeler and Paèz (2010) posit that kernel selection "is less critical than selection of the kernel bandwidth parameter for estimation results." The reason bandwidth is so critical to the results of a GWR is its smoothing effect. A small kernel bandwidth promotes more localized estimations, with a higher heterogeneity, while a larger kernel bandwidth promotes a smoother model with lower heterogeneity (Brunsdon, Fotheringham, and Charlton 1999).

The selection of bandwidth has been identified as a challenge for geographic weighted regression, and the method by which bandwidth is selected does not have clear decision-making process. There are several methods by which one can obtain a bandwidth. (1) A user-selected bandwidth, (2) automated bandwidth selection, (3) a value selected from the literature. Automated bandwidth selection is used for this study in lieu of a priori values. Automated bandwidth selection can be done using cross-validation which is regarded as a key choice, also known as leave-one-out cross-validation (LOOCV) (e.g., Wheeler and Paéz 2010, Calvo and Escolar 2003). However, there are a number of other methods as outlined in Wheeler and Páez (2010) such as AIC and corrected AIC (AICc) (Fotheringham, Brunsdon, and Charlton 2002). Numerous adaptations and corrections for these methods have been suggested or utilized (e.g., Farber & Paez, 2007; Mcmillen, 2010). Farber and Paez (2007) reference the choice of Nakaya et al. (2005) to use AICc due to how AICc penalizes model complexity. One advantage of AICc is that it also allows the comparison of models to each other in order to test their relative performance against one another. However, Farber and Paez (2007) state that there is no support for the assertion that either cross-validation or AICc is superior to the other. For the purposes of this study, cross-validation will be used to determine the bandwidth. Because the purpose of this study is exploratory, regardless of the model outcomes, it is likely the choice of cross-validation over AICc is immaterial.

A geographic weighted regression analysis was implemented using the R Package GWmodel (Gollini et al. 2015), using a bi-square kernel and an adaptive bandwidth calculated using cross-validation (n = 25). The variables used were the same as the aforementioned OLS model.

1.3. Method for Reviewing the GWR Outputs

There are several elements in the GWR output to examine and they are outlined in Fotheringham et al. (2002). For each data point in the analysis, a set of "local" statistics exists, including model residuals, regression coefficients (estimates of effect of independent variables on the dependent variable), t-values that represent the significance of the corresponding local coefficient, R^2 values to show what percentage of the data is explained by the model, predicted values based on the existing ones($\hat{y}$ or "y hat").

Although methods to examine these elements have been improved upon over time, the basics of the outline remain:

1. Corrections to p-values for multiple hypothesis testing
2. Cartographically/visually examine coefficients and t-values (significance)

1.3.1. Corrections for Multiple Hypothesis Testing

The multiple hypothesis testing problem, also known as the multiplicity problem, arises from the increased probability of committing a Type I error(false positive) when a hypothesis is tested on subsets of the same data (Shaffer 1995). In a GWR model using n data points, each data point represents a regression equation subject to the same probability of a type I error, however, this results in testing the same hypothesis n times for each parameter p, which results in np hypotheses to be tested (Charlton, Byrne, & Fotheringham, 2002). As the number of hypothesis tests goes up, the chance of Type I error increases, often sharply, according to Shaffer (1995). To accommodate for this, correction procedures have been devised to adjust the significance level (α) to a new threshold, such that the p-value necessary to be considered significant must be much lower (e.g., $p < .0001$ instead of $p < 0.05$). Two kinds of approaches are common: methods that minimize family-wise error rate (FWER) and methods that minimize false discovery rate (FDR).

Methods that minimize FWER seek to reduce the chance of even one Type I error, one such method, the Bonferroni correction for m tests at α significance level would use an adjusted α calculated as α/m . This is a straightforward, but very conservative control of Type I errors and is suggested by Fotheringham et al. (2002).

A less conservative alternative to the Bonferroni Correction is the Benjamini-Yekutieli Correction Procedure which specifically controls the false discovery rate (FDR), and so maintains its ability to detect true relationships (statistical power) better than the Bonferroni correction which is meant to minimize the chance of even one false-positive (FWER) (Benjamini and Yekutieli 2001; Charlton, Byrne, and Fotheringham 2002).

Both of these methods are easily implemented in R using the function *p.adjust* from the *stats* package (R Core Team 2016). The package GWmodel and its included function *gwr.t.adjust* will return these values for variables in a GWR (Gollini et al. 2015). The function was adapted from its original state to prevent rounding of the values returned by GWmodel. Due to the exploratory nature of this study, the less conservative Benjamini-Yekutieli correction procedure was used.

1.3.2. Cartographic Methods for GWRs

In this method, the coefficients and t-values that are resultant of each analysis are mapped for visualization. Visualization is important for both understanding and interpreting the results of a GWR (Mennis 2006). Fotheringham et al. (2002) introduce the concept of displaying t-values and coefficient values alongside one another. However, this requires an individual to visually map values over to one another (Mennis 2006). Several methods are outlined by Mennis (2006) primarily introducing the concept of obscuring insignificant results. This highlights areas that have a significant relationship (i.e., enabling easier interpretation of maps). While Matthews and Yang (2015) suggest an improvement on this methodology by placing t-value isolines over a map of the coefficients, utilizing t-value isolines placed on top of the coefficients leads to map-clutter. I disagree that this is an improvement on the methodology. However, the concept of determining differences between t-values remains a valid point.

In small-scale thematic maps, complexity has been found to reduce an individual's ability to recall choropleth maps (MacEachren 1991, citing MacEachren 1982). While Mennis (2006) and Matthews and Yang (2015) both suggest either adding map-clutter (isolines) or added complexity in color-schemes (e.g., multiple color schemes based on significance). In an effort to reduce th**e** complexity of maps and simply and effectively communicate the significance and values of the coefficients, three maps were made: a map of the t-values, a map of the coefficients illustrating the values for only the counties that are significant before a Benjamini-Yekutieli correction, and a third map (if applicable) showing counties that are significant after a Benjamini-Yekutieli correction.

2. Results and Discussion

2.1. Ordinary Least Squares Multivariate-Regression

Two OLS regression models to predict Trump's vote share were built using the following variables in common: Percentage of the population with a bachelor's degree and higher (x_{bach}), percentage of the population female (x_{fem}), median age of the population (x_{age}), percentage of the population that is white (x_{white}), median household income in 2015 (x_{income}). The second OLS model, also used Twitter share for the Trump on the 17th (x_{TS17}), Twitter share for Trump on the 18th (x_{TS18}), and Twitter share for Trump on the 19th before 9pm (x_{TS19b4}).

These result in the following equations:

$$\boldsymbol{v_S} =$$

$$.6186 - 0.008774x_{bach} + 0.1757x_{fem} + .002668x_{age} + -.3682x_{white} + .000005390x_{income} \quad (3)$$

$$\boldsymbol{v_S = 0.490461 - 0.04768x_{TS17} + 0.107328x_{TS18} - 0.02518x_{TS19b4} - 0.00941x_{bach} + 0.417739x_{fem} + 0.001818x_{age} - 0.35486x_{white} + 0.00000577x_{income}} \quad (4)$$

Percentage of the population with a bachelor's degree and higher (x_{bach}), percentage of the population that is white (x_{white}), and median household income in 2015 (x_{income}) were all found to be significant for both models, Twitter share for Trump on the 18th (x_{TS18}) was significant as well for the second model. The first model had an adjusted R^2 of 0.5615. The model including Twitter data resulted in an adjusted R^2 of 0.5909 showing a slight improvement with the addition of Twitter variables. The residuals from the model including Twitter variables are shown in Figure 7 displaying a slight right-skew.

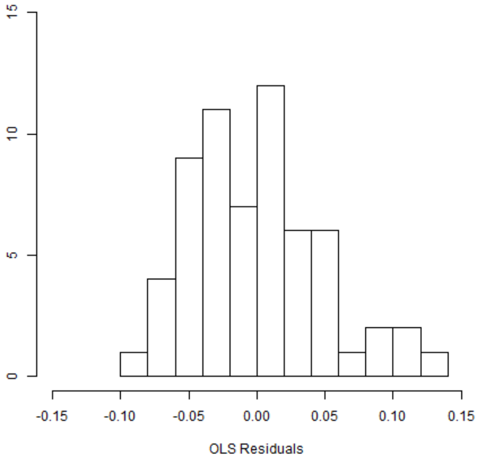

Figure 7. Histogram of residuals from OLS for Trump vote share including Twitter variables

2.2. Moran’s I on OLS Residuals

To determine whether the residuals are randomly distributed, a Moran’s I test was undertaken using queen adjacency. This resulted in a Moran’s I of 0.3719 which was found to be highly significant statistically, suggesting that the residuals from OLS are clustered. Figure 7 displays a red line for Moran’s I that displays Moran’s I compared against 999 permutations of a Monte Carlo simulation of Moran’s I based on a random resampling of the source data.

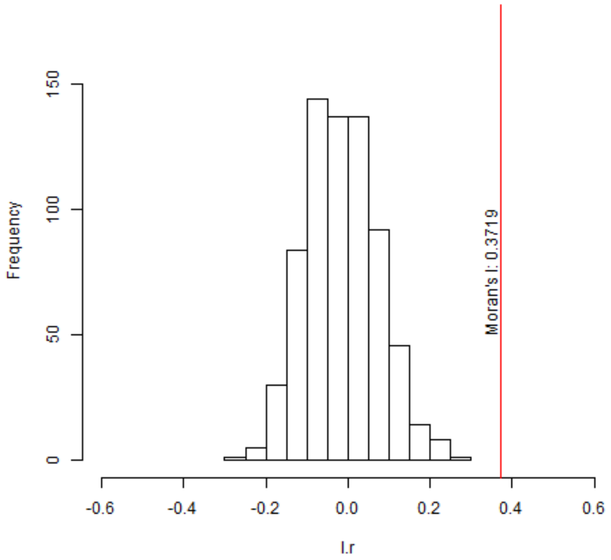

Figure 7. Monte Carlo simulation of Moran's I.

2.3. Outputs of Geographic Weighted Regression

The GWR model demonstrated an adjusted R^2 increase from 0.5909 for the OLS to 0.7480. Figure 7, demonstrates bands of higher R^2 in the center of the state in the Finger-Lakes region and Rochester as well as on Long Island and the New York boroughs (Figure 9). Additionally, the geographic weighted regression shows an overall reduction in the magnitude of residuals when compared to the OLS, particularly in the urban centers of New York, Buffalo, and Rochester. (See Figure 11d and Figure 11f)

The outcome includes results significant at an α of 0.10 and less using both an unadjusted p-value and an adjusted p-value using a Benjamini-Yekutieli correction. All variables had regions of significance before the correction (Table 10). After the Benjamini-Yekutieli

(BY) correction only three variables retained significance: median household income, percentage white, and percentage of population with a Bachelor’s degree or higher (Table 11). Maps of all non-twitter variables can be found in Appendix B.

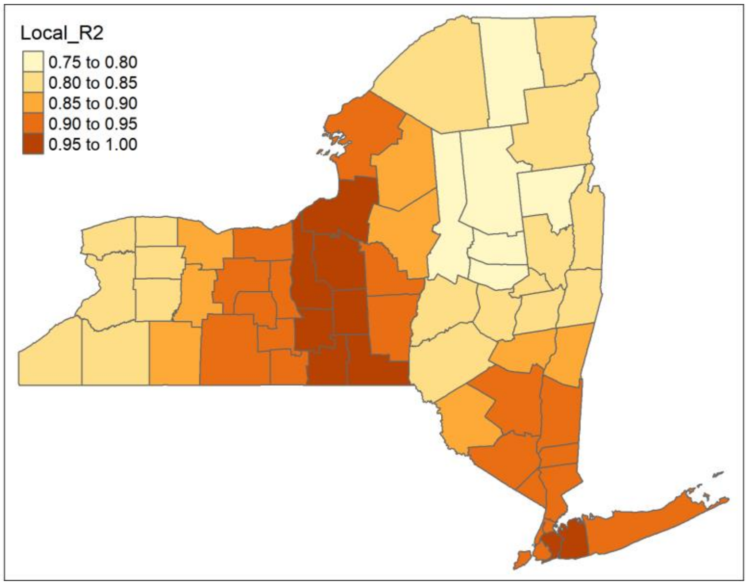

Figure 9. Local R^2 values for GWR model of Trump vote share.

Table 10. GWR coefficients and significance. This table represents basic information about coefficients significant at varying levels of significance (***α***), accompanied by the number (n) of counties significant at each ***α*.**

Column	*α*	*n*	Coefficient		
			min	max	median
Intercept	0.1	55	0.7808	2.1382	1.1651
	0.05	47	0.8773	2.1382	1.2708
	0.01	6	1.6604	2.1382	1.9135
Twitter share 18th	0.1	24	0.0926	0.1576	0.1316
	0.05	13	0.1131	0.1576	0.1441
	0.01	0	-	-	-
Twitter share 19th (before 9PM)	0.1	7	-0.0980	0.1097	0.1045
	0.05	0	-	-	-
	0.01	0	-	-	-
Twitter share 17th	0.1	12	-0.0964	-0.0767	-0.0887
	0.05	7	-0.0964	-0.0881	-0.0914
	0.01	0	-	-	-
Percentage Bachelor’s Degree or above	0.1	45	-0.0119	-0.0048	-0.0089
	0.05	40	-0.0119	-0.0048	-0.0091
	0.01	21	-0.0119	-0.0082	-0.0111
Percentage Female	0.1	5	-2.3188	-1.5837	-2.0749
	0.05	1	-2.0749	-2.0749	-2.0749
	0.01	0	-	-	-
Median Age	0.1	12	-0.0202	0.0102	-0.0163
	0.05	7	-0.0202	-0.0157	-0.0180
	0.01	0	-	-	-
Percentage White	0.1	31	-0.8458	-0.2125	-0.6237
	0.05	25	-0.8458	-0.3534	-0.6915
	0.01	19	-0.8458	-0.3534	-0.7412
Household Median Income (Estimate for 2015)	0.1	23	-7.8322E-06	6.3789E-06	6.0041E-06
	0.05	22	4.3580E-06	6.3789E-06	6.0265E-06
	0.01	20	4.8024E-06	6.3789E-06	6.0540E-06

Table 11. GWR Coefficients and Significance after Benjamini-Yekutieli correction. This table represents coefficients significant after a correction for multiple hypothesis testing.

Column	*α*	*n*	Coefficient		
			min	max	median
Percentage with Bachelor's Degree or above	0.1	16	-0.0119	-0.0092	-0.0114
	0.05	16	-0.0119	-0.0092	-0.0114
	0.01	16	-0.0119	-0.0092	-0.0114
Percentage White	0.1	14	-0.8458	-0.3534	-0.7818
	0.05	3	-0.8138	-0.5095	-0.5697
	0.01	3	-0.8138	-0.5095	-0.5697
Household Median Income (Estimate for 2015)	0.1	18	5.6231E-06	6.3789E-06	6.06E-06
	0.05	18	5.6231E-06	6.3789E-06	6.06E-06
	0.01	18	5.6231E-06	6.3789E-06	6.06E-06

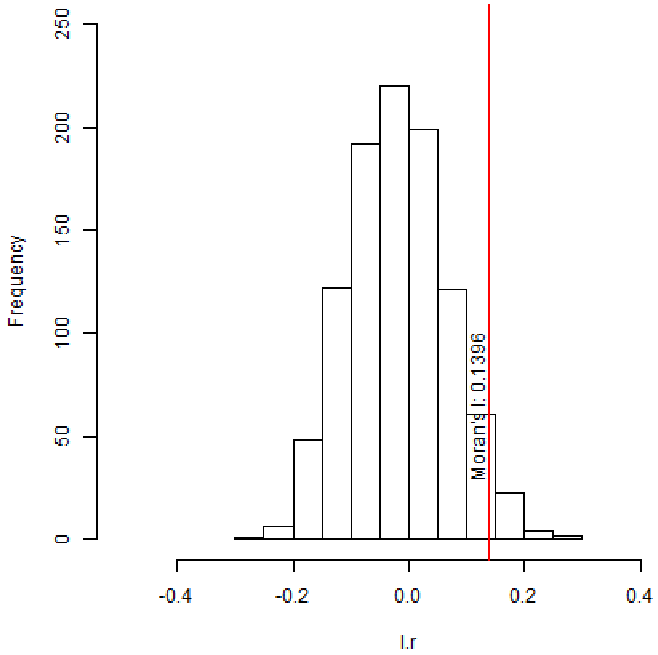

Figure 10. Moran's I of residuals from geographic weighted regression.

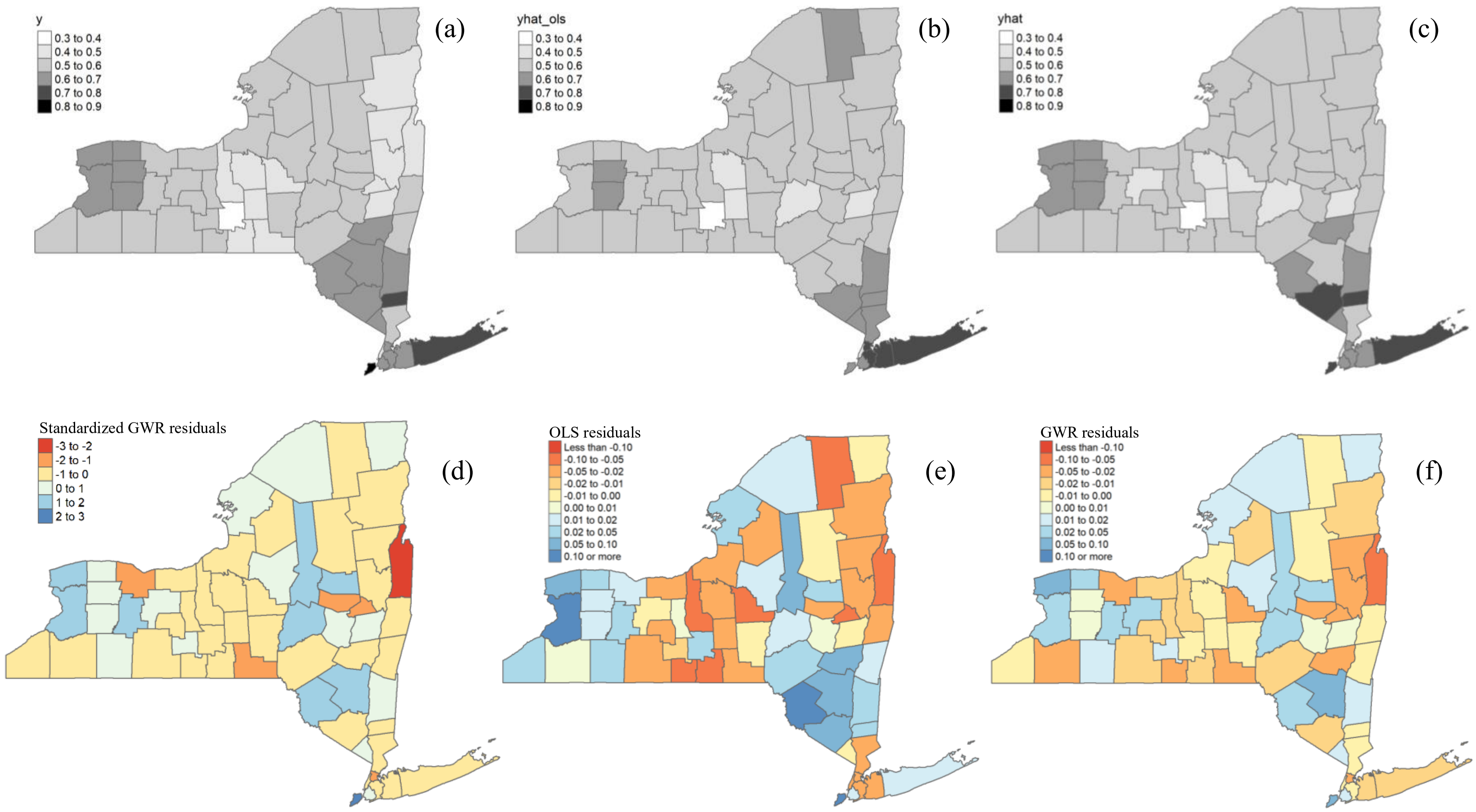

Figure 11. (a) Observed vote share for Trump (Y), (b) predicted vote share for Trump using geographic weighted regression ($\widehat{Y}$), (c) Predicted vote share for Trump using OLS ($\widehat{Y}$), (d) Standardized GWR residuals, (e) residuals from OLS predictions [b-a] (f) residuals from GWR predictions [c – a].

2.3.1. Candidate Twitter Share April 17th (TS17)

12 counties displayed a significant negative coefficient with an α of 0.10 for Twitter share on April 17th (Figure 12, see Table 10). These values were relatively small and correspond to a decrease in 0.1% of vote share per percentage point increase in Twitter share. There is a clear clustering around Finger Lakes and a portion of Southern Tier to the south. Moving out from the band of significant counties, Oswego to the north and Schoharie were both significant as well. After a correction using the Benjamini-Yekutieli procedure, no counties were significant at an α of 0.10 or less.

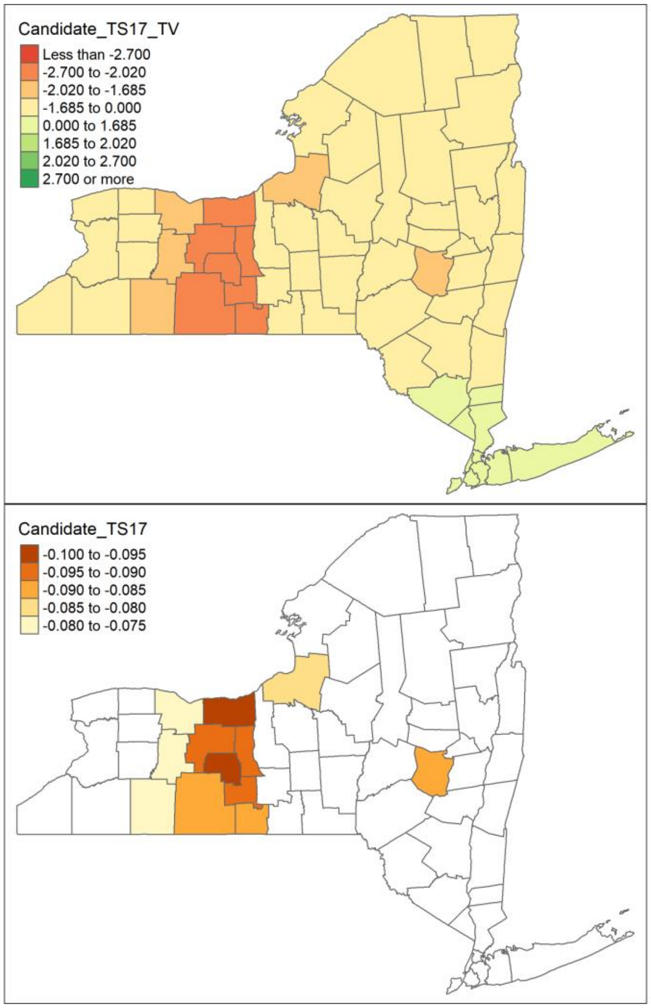

Figure 12. Candidate Twitter Share April 17th (A) local t-values (B) local coefficient values where the t-value indicates significance at an α of 0.10.

2.3.2. Candidate Twitter Share April 18th (TS18)

Candidate Twitter share on April 18th (TS18) had 24 counties that displayed a significant positive relationship with vote share (Figure 13, see Table 10).. These were significant at 90% or greater significance. The significant values were clustered in the western portion of the state as well as some significant counties in the Tri-State area along the New Jersey and Pennsylvania state-lines and Rensselaer County. The coefficients for Twitter share ranged from as much as 0.1576 and as small of a value as 0.0926. These values were relatively small and correspond to an increase of between 0.16% and 0.09% of a percentage point of vote share per percentage point increase in Twitter share. After a correction using the BY procedure, no counties were significant at an α of 0.10 or less.

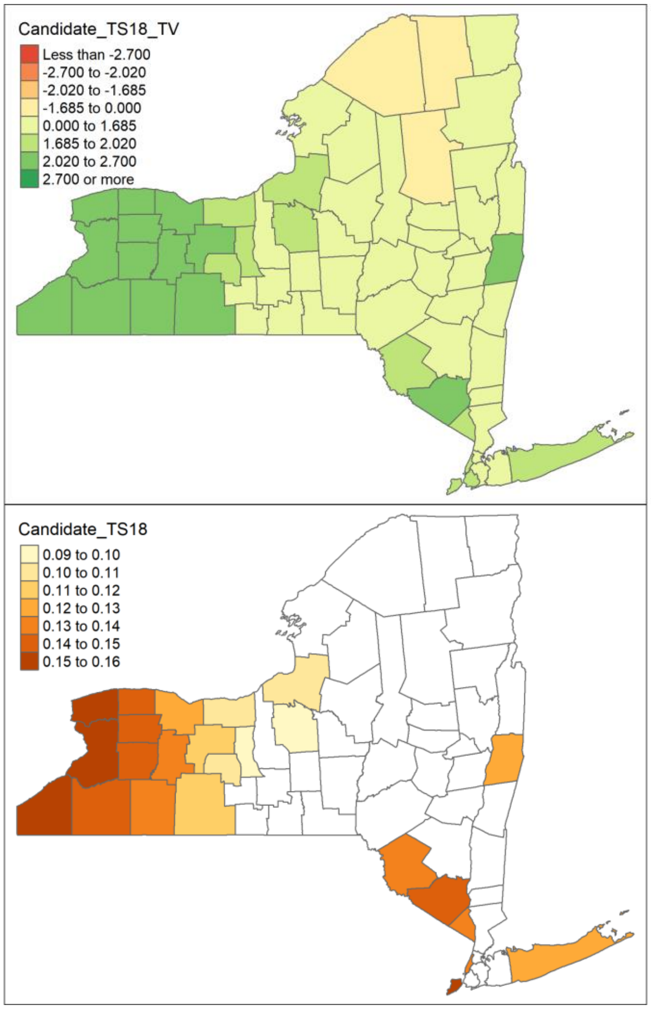

Figure 13. Candidate Twitter Share April 18th (A) local t-values (B) local coefficient values where the t-value indicates significance at an α of 0.10.

2.3.3. Candidate Twitter Share April 19th Before the Election was Called (TS19b4)

Candidate Twitter share on April 19th before the election was called (TS19b4) had seven counties that were significant at an α of 0.10 (Figure 14, see Table 10). The associated model included a maximum coefficient of 0.1097 and a minimum coefficient of -0.098 with a median value of 0.1045. No counties were significant at an α of 0.05 or less. After a correction using the Benjamini-Yekutieli procedure, no counties were significant at an α of 0.10 or less. Positive correlations are found in the boroughs of New York City and Nassau County, while one county (Rensselaer) has negative correlation coefficient. TS19b4 was the only layer to display significant positive and negative coefficients.

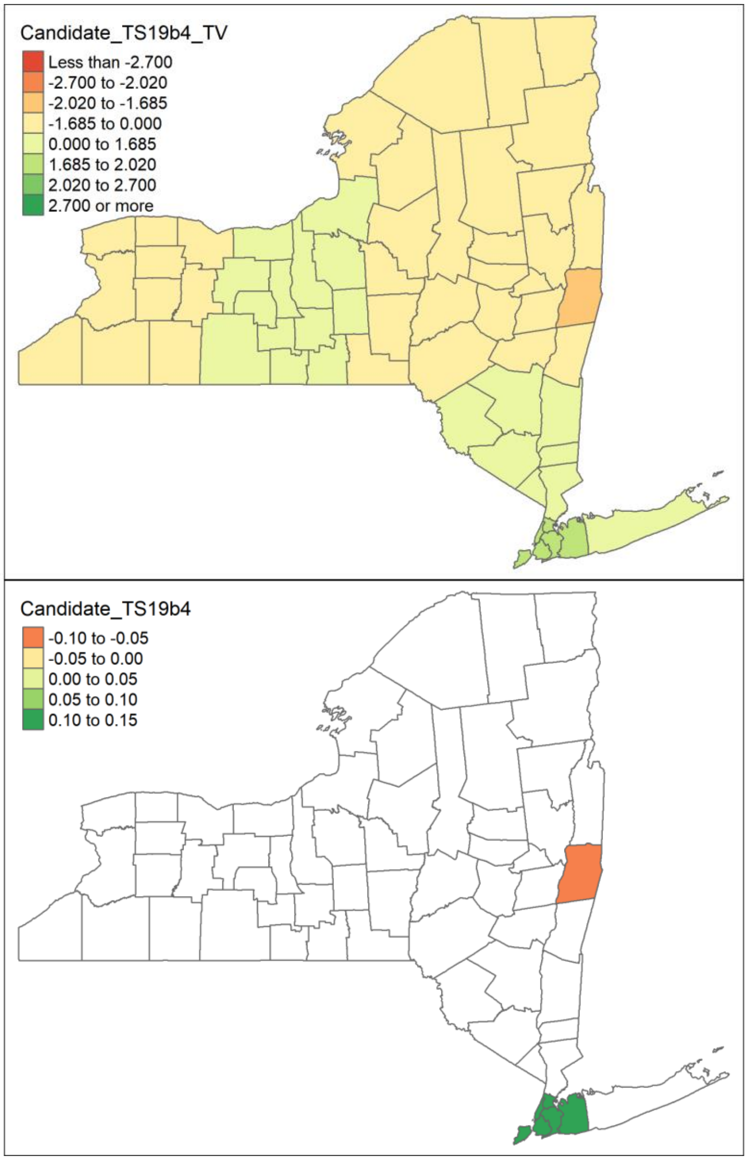

Figure 14. Candidate Twitter Share April 19th Before the Election was Called. (A) local t-values (B) local coefficient values where the t-value indicates significance at an α of 0.10.

3. Discussion of Effectiveness of Using County Demographics and Tweets to Predict Election Outcomes

The OLS revealed a significant positive relationship between median household income and the Twitter share the day prior to the election (TS18). Both percentage of the population that has attained a Bachelor's degree and higher and percentage of the population that is white showed a negative relationship with vote share in the model. The significant value for TS18 suggests an overall weak positive correlation between Trump's performance and the mentions the day before the race. Tsou et al. (2013) finds a relationship with demonstrated opinion and Twitter, showing strong correlations for the day-of and the day before. The significance of these results at the county level suggests that this is not just a state-level phenomenon but also a more localized one. The correlation for TS18 is 0.11 (Table 9), suggesting a low positive correlation at the candidate-specific level, which is reflected as significant in both the GWR and the OLS model, despite the significance disappearing after a Benjamini-Yekutieli correction.

In the GWR, all variables had at least one county where a significant result was found. Notably, TS17 and TS19b4 had regions of significance in addition to TS18. TS17 showed a weak negative relationship with Twitter share two days prior, focused around Rochester New York and the surrounding counties in the Finger Lakes region (Figure 12). Twitter Share on the day of the election (TS19b4) shows a weak positive relationship with vote share. However, this relationship was focused almost exclusively in the boroughs of New York City. None of these variables retained significance following a

Benjamini-Yekutieli correction. Additional study should be devoted to the investigation of what underlying relationships the outcome of the geographic weighted regression may be indicators or the product of.

The focus of this study was to establish a relationship between Twitter and public opinion. Twitter demonstrated a significant albeit minimal positive relationship with Donald Trump's performance at the individual scale for the day before an election, conversely Tweets two days prior to the election (TS17) displayed a weak negative relationship in the Finger Lakes Regions.

One of the aims of this study was to produce a framework for investigating finer scale relationships using geographic weighted regression. The geographic weighted regression increased the adjusted R^2 value from 0.5909 for the OLS to 0.7480. This suggests an increase in its capability to accurately model the outcome of the race. With GWR, spatial autocorrelation also decreased in the residuals, as can be seen by comparing Figure 7 to Figure 10. While the residual is less spatially autocorrelated, there was still significant auto-correlation in the residuals of the final model. This suggests that there are other underlying spatial processes not demonstrated in the data. These may be reflected in some measure of partisanship. Further study and another iteration of models would be important to understand the underlying geography.

One possible flaw in choosing New York as a case study is that its metropolitan areas adjoin many other metropolitan areas which may mute "local geography" in favor of arbitrary legislative or federal geography. While a study of the whole country would have been more meaningful in understanding trends, a state with a more distinct geography like California might have provided a better case study with geographic barriers between other states and stark contrasts between urban and rural counties endemic to the state. This election lacked a region under which Trump performed poorly. Investigating the relationship at an individual candidate level would require the identification of low-performing counties by expanding the study area geographically or through a re-framing of the methods to look at performance above and below mean performance.

It should be noted that due to the nature of this data, an innumerable number of divisions and variations of the data could be tested against, investigated for nuances, and discussed. This is well-verbalized by Cournot (1843), "...it is clear that nothing limits...the number of features according to which one can distribute [natural events or social facts] into several groups or distinct categories." (Shaffer, 1995 citing Cournot). Twitter similarly has a seemingly endless potential for spatiotemporal segmentation, and while research and investigation of these data can produce meaningful results, this paper has discussed a minuscule subset of the potential for cross-sections and tabulations of this dataset.

CHAPTER FOUR:

The results of this case study provide insight and direction for future studies of Twitter and fine-scale detection of public opinion. The primary goal of this research was to build a model to predict public opinion at the county-level by incorporating data from the Twitterscape. Moreover, there were three foundational aims of this paper; (1) capture a larger cross-section of the political conversation on Twitter, (2) understand and evaluate the merit of county-level aggregation of the Twitterscape, and (3) produce a procedure for investigating localized political phenomena using geographic weighted regression. These aims ultimately supported the primary goal. Additionally, several unexpected findings were discovered in the process of evaluating the aforementioned aims.

In order to capture a larger cross-section of the political conversation on Twitter, the decision was made to stream all tweets rather than limiting the study to tweets using coordinates. Additionally, hashtag detection was included, utilizing hashtags related to candidates to pick up implicit mentions over and above first-last detection (utilizing just the first and last names of the candidates for detection). Hashtag detection was found to capture different subsets of the data and four times as many tweets. Furthermore, the use of geocoded user-locations provided a foundational ~213,000 of the ~221,000 tweets placed within New York thereby providing a richer spatial dataset that allowed for the aggregation of data at the county level.

With the intention of understanding and evaluating the merit of county-level aggregation of the Twitterscape, correlation analyses were implemented that found substantial correlations between Twitter share and vote share. The identification of non-trivial correlations (0.51, 0.78, etc) supported similar results to those found at the state level by other researchers. This suggests that the relationship between Twitter share and vote share demonstrates a positive correlation at the county level similar to those demonstrated at less granular scales. Future studies seeking to increase the granularity of this data in regards to public opinion should be undertaken to confirm these findings. Investigation also revealed exceptions to the usefulness of social media in detecting the performance of certain parties or demographics, most notably the overall strong performance of John Kasich in New York State despite a poor Twitter presence. A look at the underlying spatial data sources also suggests inherent biases based on one's method of determining the location of a tweet. For example, based on the data collected in this study, tweets containing Coordinates have a strong Democratic bias, as such processes developed around using only Coordinates might oversample Democratic opinions. Another consideration for sampling is the inherent urban bias that was found in the dataset.

Additionally, a procedure for investigating localized political phenomena using geographic weighted regression is formalized in this paper. Specifically, a procedure for progressing from an OLS to a geographic weighted regression to investigate non-stationarity. This method utilizes Moran's I to look for spatial autocorrelation in residuals as an indicator of non-stationarity with the ultimate goal of identifying the potential for non-stationary behaviors that may be the product of underlying spatial processes. Notably, the proposed method emphasizes using a correction for multiple hypothesis testing. With the expansive use of GWR over the last decade, many studies lack any mention of correcting for multiple hypothesis testing; this paper hopefully provides a succinct framework for better studies in the future. The implementation of this model resulted in an overall reduction in spatial autocorrelation from the OLS to the GWR, although it was still significantly clustered. Additionally, a higher R^2 was achieved in the GWR. The OLS model resulted in a significant Twitter share variable, the day before the election (TS18); this variable was found to be significant in the GWR, however after a Benjamini-Yekutieli correction, only demographic variables retained significance. While this does not entirely rule out the usefulness of Twitter share variables for the prediction of vote share, it illustrates the value of a correction operation in interpreting a model.

The hope is that this paper will raise important questions and criticisms of Big Data and hopefully promote incremental movement towards more responsible fine-scale models. Fine-scale detection of public opinion could be useful for special interest groups as well as candidates to target potential voters with greater accuracy and fewer resources.

REFERENCES

Ajao, Oluwaseun, Jun Hong, and Weiru Liu. 2015. “A Survey of Location Inference Techniques on Twitter.” *Journal of Information Science* 1: pp 1-10.

Anselin, Luc, Ibnu Syabri, and Youngihn Kho. 2006. “GeoDa: An Introduction to Spatial Data Analysis.” *Geographical Analysis* 38(1): 5–22.

Arthur, Rudy, and Hywel Williams. 2017. “Scaling Laws in Geo-Located Twitter Data.” arXiv preprint arXiv:1711.09700

Barbera, Pablo, and Gonzalo Rivero. 2015. “Understanding the Political Representativeness of Twitter Users.” *Social Science Computer Review* 33(6): 712–29.

Beauchamp, Nick. 2015. “Predicting and Interpolating State-Level Polling Using Twitter Textual Data.” *APSA Annual Meeting* 2015: 1–14.

Benjamini, Yoav, and Daniel Yekutieli. 2001. “The Control of the False Discovery Rate in Multiple Testing under Dependency.” *Annals of Statistics* 29(4): 1165–88.

Bermingham, Adam, and Alan F Smeaton. 2011. “On Using Twitter to Monitor Political Sentiment and Predict Election Results.” *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*: 2–10.

Brunsdon, Chris, A Stewart Fotheringham, and Martin Charlton. 1999. “Some Notes On Parametric Significance Tests for Geographic Weighted Regression.” *Journal of Regional Science* 39(3): 497–524.

Calvo, Ernesto, and Marcelo Escolar. 2003. “The Local Voter: A Geographically Weighted Approach to Ecological Inference.” *American Journal of Political Science* 47(1): 189–204.

Charlton, Martin, Graeme Byrne, and Stewart Fotheringham. 2002. “Multiple Dependent Hypothesis Tests in Geographically Weighted Regression.” *Measurement*: 2–6.

Cournot, AA. 1843. “Exposition de La Theorie Des Chances et Des Probabilites.” In *Reprinted 1984 as Vol. 1 of Cournot’s Oeuvres Completes*, ed. B Bru. Paris: Vrin.

Crampton, Jeremy W. et al. 2013. “Beyond the Geotag: Situating ‘big Data’ and Leveraging the Potential of the Geoweb.” *Cartography and Geographic Information Science* 40(2): 130–39.

Crockford, Douglas. 2006. “The Application/json Media Type for Javascript Object Notation (Json).” https://buildbot.tools.ietf.org/html/rfc4627.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. “More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior.” *PLoS ONE* 8(11): 1–5.

Farber, Steven, and Antonio Páez. 2007. “A Systematic Investigation of Cross-Validation in GWR Model Estimation: Empirical Analysis and Monte Carlo Simulations.” *Journal of Geographical Systems* 9(4): 371–96.

Fotheringham, A S, C Brundson, and M Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*.

Gerber, Matthew S. 2014. “Predicting Crime Using Twitter and Kernel Density Estimation.” *Decision Support Systems* 61(1): 115–25.

Gollini, Isabella et al. 2015. “GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weithed Models.” *Journal Of Statistical Software* 63: 1–50.

Guo, Luo, Zhihai Ma, and Lianjun Zhang. 2008. “Comparison of Bandwidth Selection in Application of Geographically Weighted Regression: A Case Study.” *Canadian Journal of Forest Research* 38(9): 2526–34.

Hemsley, Jeff, and Josef Eckert. 2014. “Examining the Role of ‘Place’ in Twitter Networks through the Lens of Contentious Politics.” *Proceedings of the Annual Hawaii International Conference on System Sciences*: 1844–53.

Kupfer, John A., and Calvin A. Farris. 2007. “Incorporating Spatial Non-Stationarity of Regression Coefficients into Predictive Vegetation Models.” *Landscape Ecology* 22(6): 837–52.

MacEachren, Alan M. 1982. “The Role of Complexity and Symbolization Method in Thematic Map Effectiveness.” *Annals of the Association of American Geographers* 72(4): 495–513.

———. 1991. “The Role of Maps in Spatial Knowledge Acquisition.” *Cartographic Journal, The* 28(2): 152–62.

Matthews, Stephen A., and Tse Chuan Yang. 2012. “Mapping the Results of Local Statistics: Using Geographically Weighted Regression.” *Demographic Research* 26: 151–66.

McMillen, Daniel P. 2010. “Issues in Spatial Data Analysis.” *Journal of Regional Science* 50(1): 119–41.

Mennis, Jeremy. 2006. “Mapping the Results of Geographically Weighted Regression.” *The Cartographic Journal* 43(2): 171–79.

Mislove, Alan et al. 2011. “Understanding the Demographics of Twitter Users.” *Artificial Intelligence*: 554–57.

Mitchell, Lewis et al. 2013. “The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place.” *PLoS ONE* 8(5).

Moran, P. A. P. 1950. “Notes on Continuous Stochastic Phenomena.” *Biometrika* 37: 17–23.

Morstatter, Fred, J Pfeffer, H Liu, and KM Carley. 2013. “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose.” *Proceedings of ICWSM*: 400–408.

Nakaya T, A. S. Fotheringham, C. Brunsdon and M. Charlton. 2005. “Geographically Weighted Poisson Regression for Disease Association Mapping.” *Statistics in medicine* 24(17): 2695–2717.

New York State Board of Elections. 2016. “2016 Election Results.” https://www.elections.ny.gov/2016ElectionResults.html (August 1, 2017).

O’Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge, and Noah a Smith. 2010. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” (May): 122–29.

Quercia, Daniele, Licia Capra, and Jon Crowcroft. 2012. “The Social World of Twitter: Topics, Geography, and Emotions.” In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, , 298–305.

R Core Team. 2016. “R: A Language and Environment for Statistical Computing.”

Roesslein, Joshua. 2016. “Tweepy Documentation.” 3.6.0. http://tweepy.readthedocs.org/en/v3.5.0/.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. 2010. “Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors.” *Proceedings of the 19th International Conference on World Wide Web*: 851–60.

Shaffer, Juliet Popper. 1995. “Multiple Hypothesis Testing.” *Annual Review of Psychology* 46: 561–84.

Shephard, Steven. 2015. “Gallup Gives up the Horse Race.” *POLITICO*. https://www.politico.com/story/2015/10/gallup-poll-2016-pollsters-214493.

Small, Tamara A. 2011. “What the Hashtag?” *Information, Communication & Society* 14(6): 872–95.

Tsou, Ming-Hsiang et al. 2013. “Mapping Social Activities and Concepts with Social Media (Twitter) and Web Search Engines (Yahoo and Bing): A Case Study in 2012 US Presidential Election.” *Cartography and Geographic Information Science* 40(4): 337–48.

Tsou, Ming-Hsiang, and Michael Leitner. 2013. “Visualization of Social Media: Seeing a Mirage or a Message?” *Cartography and Geographic Information Science* 40(August 2015): 55–60.

Tufekci, Zeynep. 2014. “Big Questions for Social Media Big Data : Representativeness , Validity and Other Methodological Pitfalls.” *ICWSM* 14: 505–14.

Twitter Inc. 2016. “Tweet Objects.” https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json (August 2, 2016).

Wheeler, Benjamin. 2016. “Tiipwriter.” https://github.com/benjaminwheel3r/tiipwriter.

Wheeler, David C, and Antonio Páez. 2010. “Geographically Weighted Regression.” In *Handbook of Applied Spatial Analysis*, , 461–86.

Zar, Jerrold H. 2005. “Spearman Rank Correlation.” In *Encyclopedia of Biostatistics*, Chichester, UK: John Wiley & Sons, Ltd.

Zhang, Lianjun, Jeffrey H. Gove, and Linda S. Heath. 2005. “Spatial Residual Analysis of Six Modeling Techniques.” *Ecological Modelling* 186(2): 154–77.

APPENDIX A

Correlation Tables for Twitter Share Variables by Spatial Source

The below correlation tables have very low values of N (number of counties tested for correlation), as such the Spearman Rank Correlation Coefficient provides exceptionally high correlations. These correlations are misleading and exclude a majority of the study area. They have been included for transparency, and they provide the potential for future study and review. The challenges of extrapolating these values to the rest of the study area accentuate a primary aim of this study, which was to provide a higher-volume detection to capture a sufficient sample in a shorter period.

The following tables of Spearman’s rho were calculated at each spatiotemporal break for each method of detection, for each of the following subsets of spatial data: Coordinates (Table A1 and A2), Coordinates and Uber-Coordinates (Tables A3 and A4), Coordinates, Uber-Coordinates and Place (Tables A5 and A6).

Table A1: Vote Share vs. Twitter Share by Period for Coords, and Hashtag Detection.

	TS17	TS18	TS19b4	TS19aft	TS20	TS_b4	TS
R_vote share	0.802	0.607	0.379	0.643	0.694	0.683	0.657
D_vote share	-0.116	-0.416	0.076	0.752	-0.017	0.059	0.126
All_vote share	0.616	0.440	0.379	0.729	0.449	0.541	0.569
Trump	0.099	0.777	-0.185	0.592	0.866	0.178	0.284
Kasich	0.040	0.593	NA	NA	NA	0.249	0.214
Cruz	NA	-0.160	0.041	0.436	0.577	0.082	0.238
TK	0.773	0.763	0.603	0.817	0.875	0.751	0.754
TC	0.832	0.663	0.339	0.776	0.938	0.766	0.747
KC	0.326	0.170	-0.222	-0.064	-0.174	0.079	0.088
H	0.155	-0.238	0.386	0.493	0.089	0.234	0.271
B	0.155	-0.238	0.386	0.493	0.089	0.234	0.271

Table A2: Vote Share vs.Twitter Share by Period for Coords, and First-Last Detection.

	TS17_FL	TS18_FL	TS19b4_FL	TS19aft_FL	TS20_FL	TS_b4_FL	TS_FL
R_vote share	0.922	0.596	0.461	-0.046	0.365	0.642	0.548
D_vote share	0.513	0.265	0.522	0.600	0.683	0.568	0.449
All_vote share	0.730	0.519	0.517	0.215	0.504	0.643	0.546
Trump	0.632	0.738	0.447	-0.500	0.000	0.329	0.489
Kasich	0.949	0.738	NA	NA	NA	0.619	0.450
Cruz	NA	-0.258	0.224	0.500	0.866	0.121	0.442
TK	0.946	0.551	0.700	0.270	0.414	0.699	0.666
TC	0.894	0.651	0.575	0.088	0.488	0.654	0.606
KC	0.764	0.440	-0.208	-0.541	0.131	0.366	0.267
H	0.135	0.500	0.000	-1.000	0.866	0.335	0.149
B	0.135	0.500	0.000	-1.000	0.866	0.335	0.149

Table A3. Vote Share vs. Twitter Share by Period for Coords and UberCoordinates, and Hashtag Detection.

	TS17	TS18	TS19b4	TS19aft	TS20	TS_b4	TS
R_vote share	0.580	0.650	0.526	0.592	0.639	0.617	0.564
D_vote share	-0.119	0.061	0.065	0.483	0.546	-0.028	0.104
All_voteshare	0.471	0.562	0.459	0.617	0.687	0.564	0.566
Trump	0.190	0.402	0.050	-0.172	0.021	0.087	-0.005
Kasich	0.047	0.051	0.054	-0.309	NA	0.053	-0.032
Cruz	0.003	-0.055	0.072	0.336	-0.041	-0.140	-0.044
TK	0.627	0.746	0.718	0.691	0.786	0.725	0.681
TC	0.620	0.770	0.592	0.667	0.771	0.727	0.669
KC	0.096	-0.074	-0.161	-0.114	-0.324	-0.143	-0.152
H	-0.231	-0.186	0.314	0.309	0.362	0.016	0.076
B	-0.231	-0.186	0.314	0.309	0.362	0.016	0.076

Table A4. Vote Share vs. Twitter Share by Period for Coords and UberCoordinates, and First-Last Detection.

	TS17_FL	TS18_FL	TS19b4_FL	TS19aft_FL	TS20_FL	TS_b4_FL	TS_FL
R_vote share	0.538	0.525	0.397	0.232	0.323	0.482	0.400
D_vote share	0.457	0.418	0.148	0.588	0.821	0.253	0.221
All_vote share	0.521	0.548	0.422	0.357	0.489	0.501	0.429
Trump	0.511	0.449	0.198	-0.335	-0.247	0.190	-0.040
Kasich	0.518	0.017	0.227	-0.274	NA	0.127	0.009
Cruz	-0.156	0.046	0.039	-0.010	-0.165	-0.286	-0.130
TK	0.624	0.640	0.610	0.313	0.558	0.635	0.480
TC	0.545	0.654	0.452	0.229	0.277	0.552	0.480
KC	0.274	-0.058	-0.106	-0.228	-0.290	-0.129	-0.148
H	-0.047	-0.157	-0.498	-0.393	0.655	-0.494	-0.362
B	-0.047	-0.157	-0.498	-0.393	0.655	-0.494	-0.362

Table A5: Vote Share vs. Twitter Share by Period for Coords and UberCoordinates, and Place, and Hashtag Detection.

	TS17	TS18	TS19b4	TS19aft	TS20	TS_b4	TS
R_vote share	0.425	0.340	0.415	0.535	0.496	0.452	0.503
D_vote share	-0.046	0.017	0.242	0.019	0.183	0.098	0.067
All_vote share	0.455	0.401	0.524	0.517	0.503	0.553	0.567
Trump	-0.100	-0.058	-0.150	-0.216	-0.170	-0.071	-0.035
Kasich	-0.192	0.159	-0.041	0.026	-0.084	0.078	0.163
Cruz	-0.159	-0.399	-0.369	-0.253	-0.326	-0.324	-0.208
TK	0.667	0.581	0.664	0.700	0.746	0.688	0.725
TC	0.520	0.364	0.498	0.613	0.598	0.524	0.600
KC	-0.417	-0.321	-0.427	-0.279	-0.487	-0.345	-0.293
H	-0.108	0.005	0.291	0.010	0.140	0.135	0.110
B	-0.108	0.005	0.291	0.010	0.140	0.135	0.110

Table A6: Vote Share vs. Twitter Share by Period for Coords and UberCoordinates, and Place, and First-Last Detection.

	TS17_FL	TS18_FL	TS19b4_FL	TS19aft_FL	TS20_FL	TS_b4_FL	TS_FL
R_vote share	0.256	0.287	0.422	0.406	0.435	0.396	0.424
D_vote share	0.289	0.272	0.451	0.155	0.657	0.204	0.023
All_vote share	0.328	0.378	0.516	0.429	0.533	0.449	0.416
Trump	0.353	-0.150	-0.326	-0.198	-0.139	-0.102	-0.152
Kasich	-0.024	0.314	0.136	-0.432	0.023	0.157	0.096
Cruz	0.185	-0.423	-0.388	-0.177	-0.035	-0.257	-0.222
TK	0.575	0.609	0.684	0.646	0.703	0.665	0.693
TC	0.353	0.249	0.428	0.467	0.479	0.427	0.472
KC	-0.282	-0.362	-0.357	-0.469	-0.344	-0.307	-0.340
H	-0.179	-0.157	-0.139	-0.018	0.284	-0.428	-0.244
B	-0.179	-0.157	-0.139	-0.018	0.284	-0.428	-0.244

APPENDIX B

Tabulations of Micropolitan, Metropolitan and Rural tweets

Rural

99.04% of tweets were assigned to a micropolitan or metropolitan area, which leaves 0.96 % of tweets that have not been assigned to a metropolitan or micropolitan area. The remaining 2,120 tweets that are not included in any micropolitan or metropolitan regions are rural (Table 10).

Table B1: Comparison of metropolitan, micropolitan, and rural tweet counts.

Location	Tweets Collected	% of Total (235,767)	FL Tweets Collected	% of Total (57,630)
Metropolitan	215,135	97.15%	56,296	97.69%
Micropolitan	3,338	1.51%	875	1.52%
Rural	2,120	0.96%	459	0.80%
Total	220,593	-	57,630	-

Metropolitan and Micropolitan Communities

183,629 Tweets (83.24%) were found to be in the NY portion of the Tri-State metropolitan area. Of the tweets that were included using first-last detection, 48,061 (83.40%) were found to be in the Tri-State metropolitan area (Table B1). Overall, 1.5% of tweets were placed within Micropolitan communities. (Table B2)

Table B2: Tweets in Metropolitan Statistical Areas.

Metropolitan Statistical Area (M1)	**Tweets Collected (#)**	**% of Total (220,593)**	**Tweets Collected (FL)**	**% of Total (57,630)**
New York-Newark-Jersey City, NY-NJ-PA	182,431	82.70%	47,730	82.82%
Buffalo-Cheektowaga-Niagara Falls, NY	11,684	5.30%	3,547	6.15%
Rochester, NY	6,136	2.78%	1,581	2.74%
Albany-Schenectady-Troy, NY	4,918	2.23%	1,173	2.04%
Syracuse, NY	4,298	1.95%	1,021	1.77%
Glens Falls, NY	1,228	0.56%	139	0.24%
Kingston, NY	1,198	0.54%	331	0.57%
Utica-Rome, NY	1,142	0.52%	282	0.49%
Ithaca, NY	867	0.39%	175	0.30%
Binghamton, NY	718	0.33%	191	0.33%
Watertown-Fort Drum, NY	390	0.18%	75	0.13%
Elmira, NY	125	0.06%	51	0.09%
Total:	**215,135**	**97.53%**	**56,296**	**97.69%**

Table B3: Tweets in Micropolitan Statistical Areas.

Micropolitan Statistical Area (M2)	**Tweets Collected**	**% of Total (220,593)**	**FL Tweets Collected**	**% of Total (57,630)**
Jamestown-Dunkirk-Fredonia, NY	1,306	0.59%	317	0.55%
Amsterdam, NY	324	0.15%	56	0.10%
Cortland, NY	240	0.11%	45	0.08%
Ogdensburg-Massena, NY	203	0.09%	78	0.14%
Corning, NY	200	0.09%	51	0.09%
Plattsburgh, NY	198	0.09%	48	0.08%
Hudson, NY	160	0.07%	47	0.08%
Oneonta, NY	153	0.07%	38	0.07%
Olean, NY	147	0.07%	58	0.10%
Gloversville, NY	125	0.06%	33	0.06%
Auburn, NY	94	0.04%	29	0.05%
Seneca Falls, NY	88	0.04%	37	0.06%
Malone, NY	58	0.03%	20	0.03%
Batavia, NY	42	0.02%	18	0.03%
Total:	**3,338**	**1.51%**	**875**	**1.52%**

Table B4: Tweets in Metropolitan Combined Statistical Areas.

Combined Statistical Area (CSA)	Tweets Collected	% of Total (220,593)	FL Tweets Collected	% of Total (57,630)
New York-Newark, NY-NJ-CT-PA	183,629	83.24%	48,061	83.40%
Buffalo-Cheektowaga, NY	11,831	5.36%	3,605	6.26%
Albany-Schenectady, NY	6,755	3.06%	1,448	2.51%
Rochester-Batavia-Seneca Falls, NY	6,266	2.84%	1,636	2.84%
Syracuse-Auburn, NY	4,392	1.99%	1,050	1.82%
Ithaca-Cortland, N**Y**	1,107	0.50%	220	0.38%
Elmira-Corning, NY	325	0.15%	102	0.18%
Total	214,305	97.15%	56,122	97.38%

APPENDIX C

GWR Supplemental Figures of Demographic Characteristics

The following appendix graphically displays the demographic characteristics included in the GWR model for Donald Trump. Descriptive results are included, although not discussed.

Median Age (age)

Median Age of the population had 12 counties that were significant at an α of 0.10. This included a maximum coefficient of 0.0102 and a minimum coefficient of -0.0202 with a median value of -0.0163. Of those counties 7 were significant at an α of 0.05 . This interval had a maximum coefficient of -0.0157 and a minimum coefficient of -0.0202 with a median value of -0.018. No counties were significant at an α of 0.01. After a correction using the Benjamini-Yekutieli procedure no counties were significant at an α of 0.10 or less. Negative correlations are shown in the Tri-state area (the boroughs of New York, Nassau, Westchester, Rockland and Orange counties), and positive correlations are shown for Washington, Rensslaer, and Chenango counties.

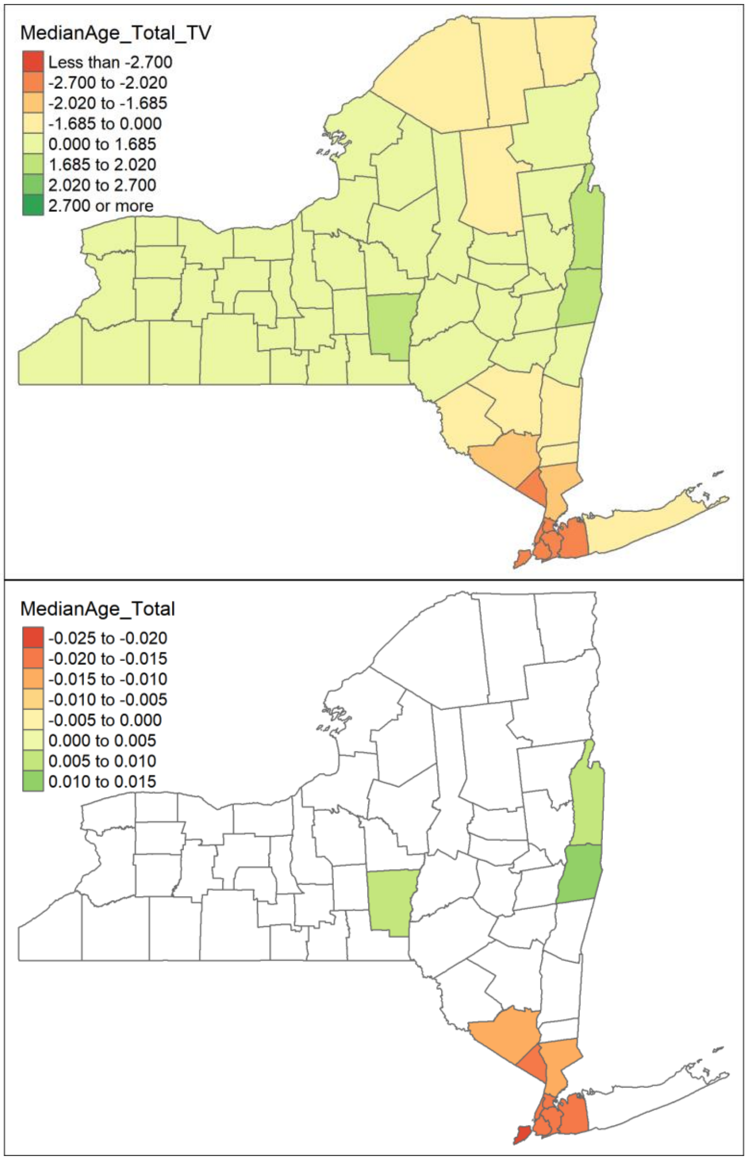

Figure B1. Median Age of Population. (a) local t-values (b) local coefficient values where the t-value indicates significance at an α of 0.10.

Percentage of Population with Bachelor's Degree or higher (bach)

Percentage of Population with Bachelor's Degree or higher had 45 counties that were significant at an α of 0.10. This included a maximum coefficient of -0.0048 and a minimum coefficient of -0.0119 with a median value of -0.0089. Of those counties 40 were significant at an α of 0.05 . This interval had a maximum coefficient of -0.0048 and a minimum coefficient of -0.0119 with a median value of -0.0091. Of those counties 21 were significant at an α of 0.01. This interval had a maximum coefficient of -0.0082 and a minimum coefficient of -0.0119 with a median value of -0.0111. After a correction using the Benjamini-Yekutieli procedure, 16 counties were significant at an α of 0.10 or less. Of those all 16 counties were significant at an α of 0.01. This interval had a maximum coefficient of -0.0092 and a minimum coefficient of -0.0119 with a median value of -0.0114.

Uncorrected p-values for Percentage Bachelor's Degree and higher was significant for most counties in the state, however, after the correction the New York portion of the tri-state area and Long Island is still significant while the other counties have dropped off. The OLS coefficient had a similar coefficient -0.00941 which is within the minimum and maximum of the GWR's results.

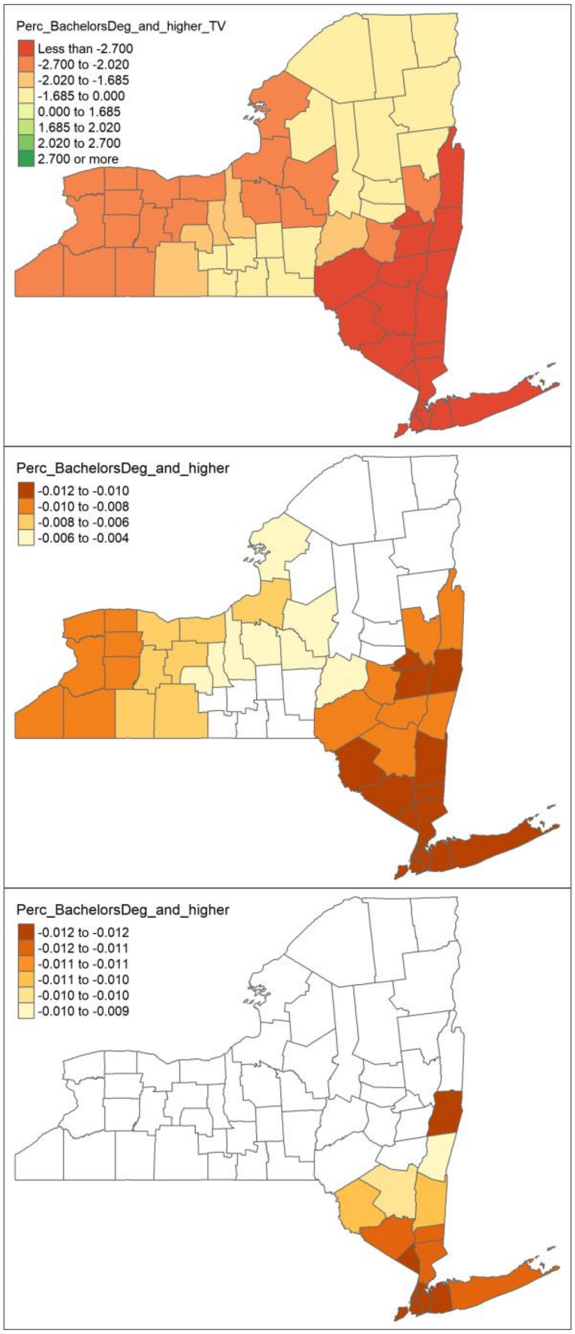

Figure B2. Percentage of Population with Bachelor's Degrees and Higher. (a) local t-values (b) local coefficient values where the t-value indicates significance at an α of 0.10. (c) local coefficient values where the t-value indicates significance at an α of 0.10 after a Benjamini-Yekutieli correction.

Percentage of Population, Female (fem)

Percentage of Population Female (fem) had 5 counties that were significant at an α of 0.10. This included a maximum coefficient of -1.5837 and a minimum coefficient of -2.3188 with a median value of -2.0749.Of those counties 1 were significant at an α of 0.05 . This interval had a maximum coefficient of -2.0749 and a minimum coefficient of -2.0749 with a median value of -2.0749. No counties were significant at an α of 0.01. After a correction using the Benjamini-Yekutieli procedure no counties were significant at an α of 0.10 or less.

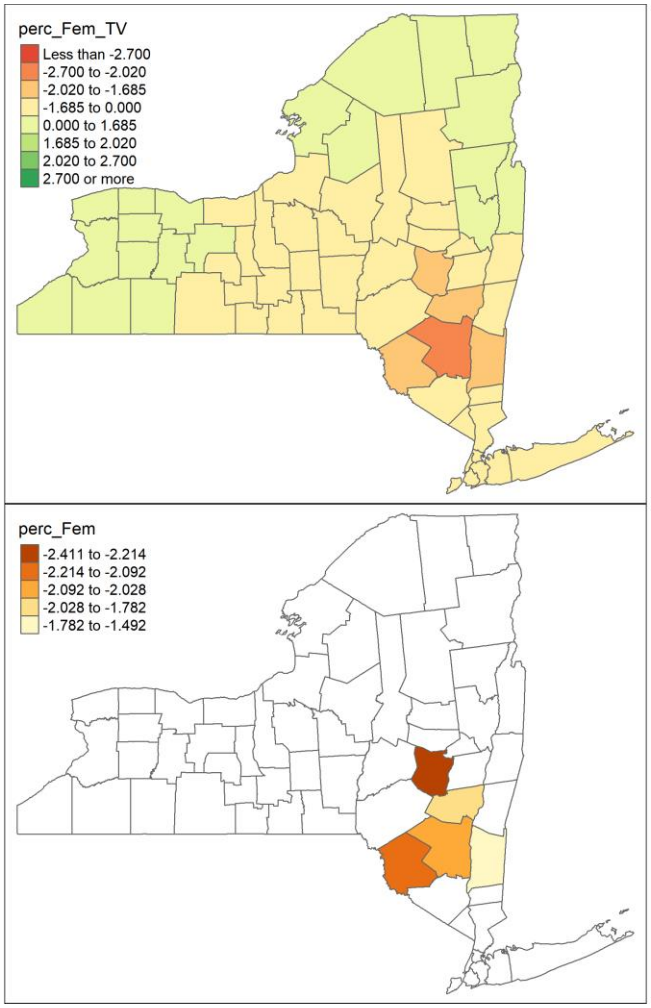

Figure B3. Percentage of Population Female. (A) local t-values (B) local coefficient values where the t-value indicates significance at an α of 0.10.

Percentage of Population that is White (white)

Percentage of Population that is White had 31 counties that were significant at an α of 0.10. This included a maximum coefficient of -0.2125 and a minimum coefficient of -0.8458 with a median value of -0.6237. Of those counties 25 were significant at an α of 0.05 . This interval had a maximum coefficient of -0.3534 and a minimum coefficient of -

0.8458 with a median value of -0.6915. Of those counties 19 were significant at an α of 0.01. This interval had a maximum coefficient of -0.3534 and a minimum coefficient of -0.8458 with a median value of -0.7412. After a correction using the Benjamini-Yekutieli procedure, 14 counties were significant at an α of 0.10 or less. This interval had a maximum coefficient of -0.3534 and a minimum coefficient of -0.8458 with a median value of -0.7818. Of those 3 counties were significant at an α of 0.01. This interval had a maximum coefficient of -0.5095 and a minimum coefficient of -0.8138 with a median value of -0.5697. Geographically, the percentage white is significant in Western New York and Finger Lakes Area and just outside of the Tri-State area, reaching up into the Capital District and Mohawk Valley. It also includes Suffolk County on Long Island. After the correction, the expansiveness reduces but a number of counties in the Capital District and Western New York retain their significance.

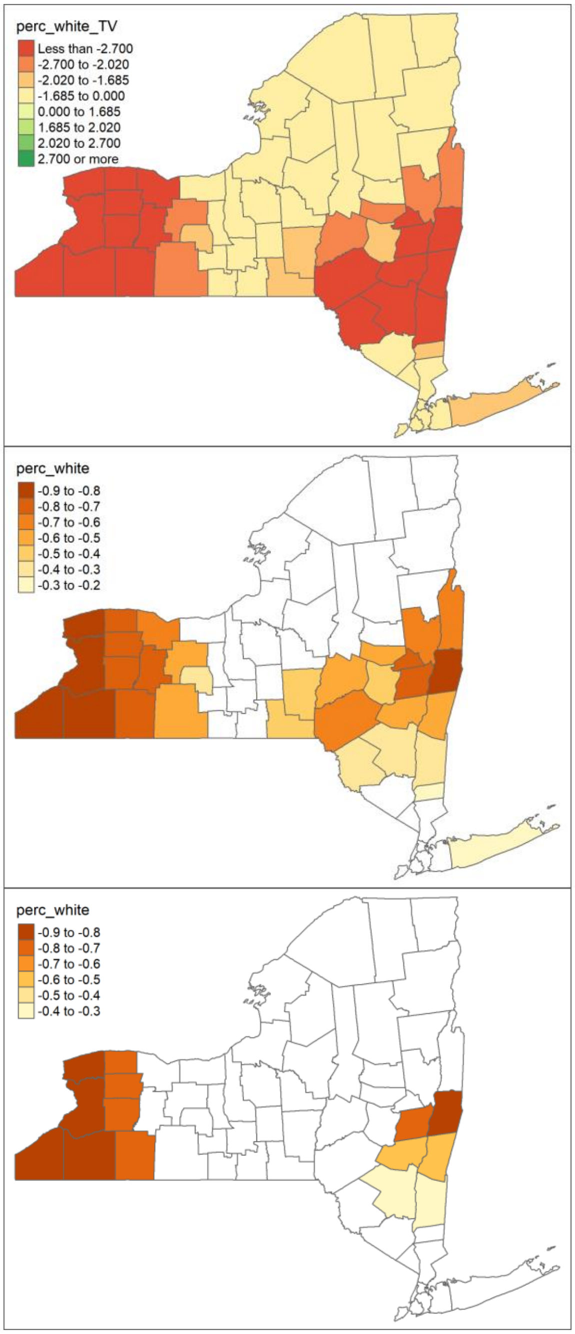

Figure B4. Percentage of population white (a) local t-values (b) local coefficient values where the t-value indicates significance at an α of 0.10. (c) local coefficient values where the t-value indicates significance at an α of 0.10 after a Benjamini-Yekutieli correction.

Median Household Income

Median Household Income had 23 counties that were significant at an α of 0.10. This included a maximum coefficient of 0.00000638 and a minimum coefficient of -0.00000783 with a median value of 0.000006. Of those counties 22 were significant at an α of 0.05. This interval had a maximum coefficient of 0.00000638 and a minimum coefficient of 0.00000436 with a median value of 0.00000603. Of those counties 20 were significant at an α of 0.01. This interval had a maximum coefficient of 0.00000638 and a minimum coefficient of 0.0000048 with a median value of 0.00000605. After a correction using the Benjamini-Yekutieli procedure, 18 counties were significant at an α of 0.10 or less. All of those 18 counties were significant at an α of 0.01. This interval had a maximum coefficient of 0.00000638 and a minimum coefficient of 0.00000562 with a median value of 0.00000606.

For a majority of these counties every 10,000 dollars increase in Median Income would relate to a 5-6 percentage points increase in vote share. One county, Jefferson had a negative correlation with Household Median Income with a potential drop in 7 percentage points per $10,000. After a Benjamini-Yekutieli correction the counties farthest from New York did not retain their significance: Schenectady, Washington, Delaware and Scoharie.

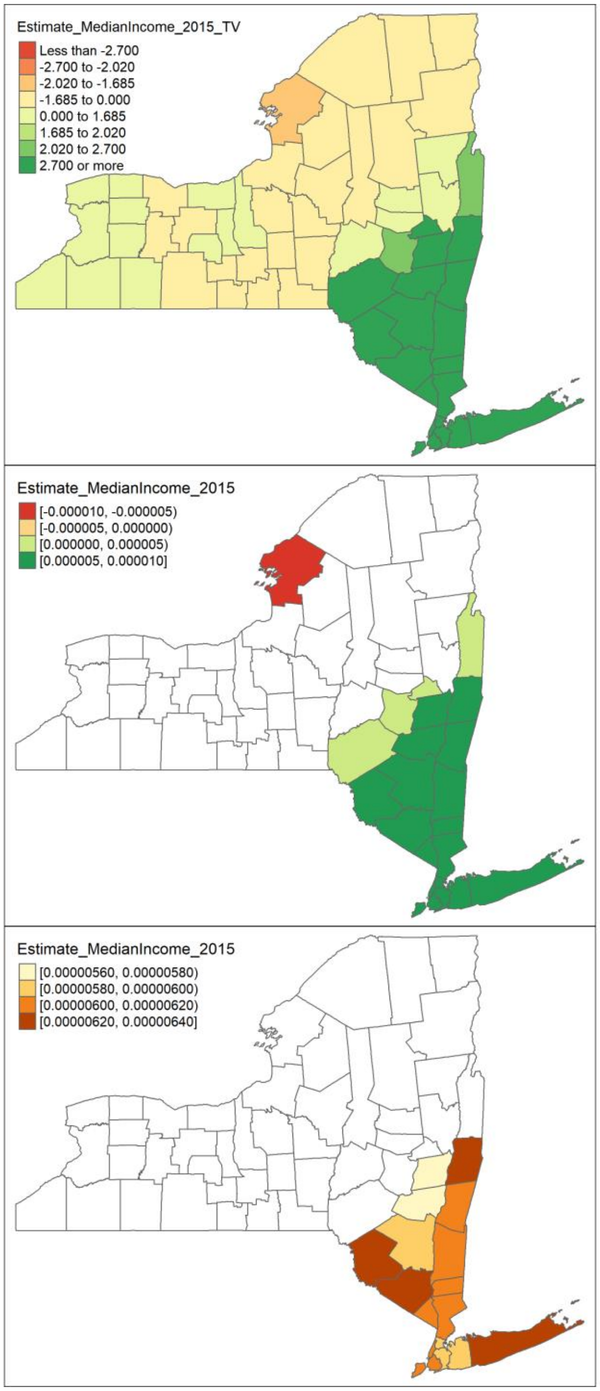

Figure B5. Median Household Income (a) local t-values (b) local coefficient values where the t-value indicates significance at an α of 0.10. (c) local coefficient values where the t-value indicates significance at an α of 0.10 after a Benjamini-Yekutieli correction.

Model Intercept

Intercept had 55 counties that were significant at an α of 0.10. This included a maximum coefficient of 2.1382 and a minimum coefficient of 0.7808 with a median value of 1.1651.Of those counties 47 were significant at an α of 0.05. This interval had a maximum coefficient of 2.1382 and a minimum coefficient of 0.8773 with a median value of 1.2708. Of those counties 6 were significant at an α of 0.01. This interval had a maximum coefficient of 2.1382 and a minimum coefficient of 1.6604 with a median value of 1.9135. After a correction using the Benjamini-Yekutieli procedure no counties were significant at an α of 0.10 or less.

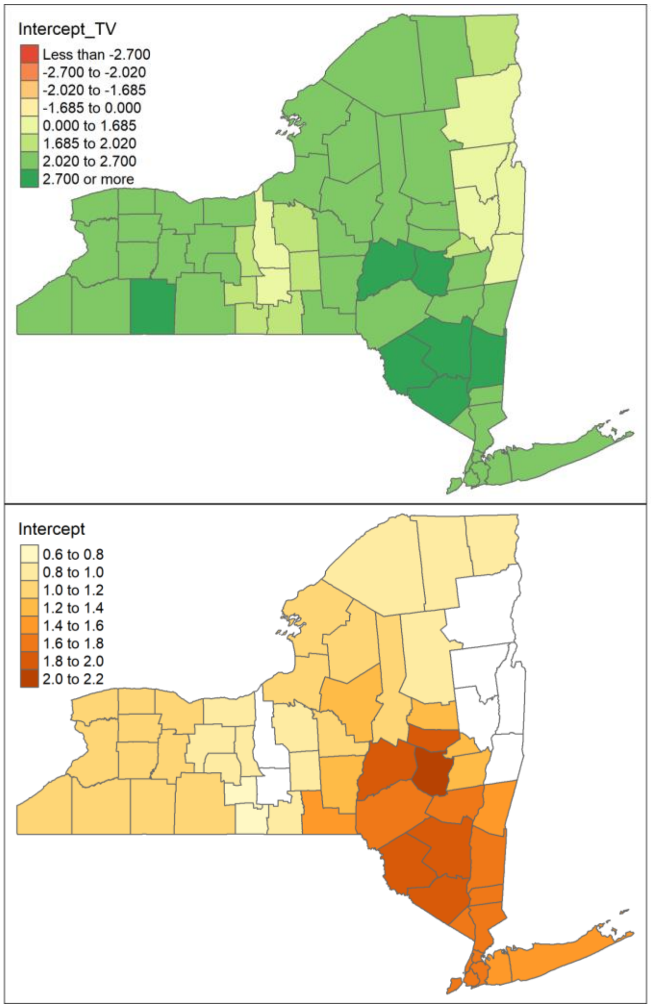

Figure B6. Model Intercept (A) local t-values (B) local coefficient values where the t-value indicates significance at an α of 0.10.