Spatial Analysis of News Sources

Andrew Mehler, Yunfan Bao, Xin Li, Yue Wang, and Steven Skiena

Abstract— People in different places talk about different things. This interest distribution is reflected by the newspaper articles circulated in a particular area. We use data from our large-scale newspaper analysis system (*Lydia*) to make entity datamaps, a spatial visualization of the interest in a given named entity. Our goal is to identify entities which display regional biases. We develop a model of estimating the frequency of reference of an entity in any given city from the reference frequency centered in surrounding cities, and techniques for evaluating the spatial significance of this distribution.

Index Terms—GIS, Geographic Visualization, Text and Document Visualization, Information analytics, WWW data visualization, Spidering, Newspapers

1 Introduction

Periodical publications represent a rich and recurrent source of knowledge on both current and historical events. The *Lydia* project seeks to build a relational model of people, places, and other entities through natural language processing of news sources and the statistical analysis of entity frequencies and co-locations. We encourage the reader to visit our website (*http://www.textmap.com*) to see our analysis of recent news obtained from over 500 daily online news sources.

Lydia tracks the occurrences of hundreds of thousands different entities arising in these news sources. An exciting consequence of this is that we can establish regional biases in the news, by analyzing the relative frequency that entities are mentioned in different news sources. We can report the results of our analysis through data maps reflecting the interest in a given entity as a function of location.

Typical datamaps of interest are presented in Figures 1-3. The datamap for New York governor *George Pataki* focuses on his home state but also exhibits a secondary concentration in Iowa. This is explainable by Pataki's presidential ambitions and the significance of the Iowa Caucuses, the first test of the U.S. presidential primary season. The map for Phoenix Sun's basketball star *Steve Nash* reflects home town fan interest in both his current and previous (Dallas Mavericks) teams. Datamaps of geographical locations also show interesting biases. News interest in *Mexico* is significantly heavier around the U.S. Mexico border, particularly in southern Texas. *Washington, DC* reflects national interest in its capitol city, with stronger concentrations centered in the District of Columbia (reflecting local interest) and the State of Washington (reflecting natural language processing artifacts in resolving city references from state references). National figures such as Vice President Dick Cheney show little regional bias, while former international movie star Arnold Schwarzenegger is today primarily a California state political figure.

It is the geographical bias among primary news sources which permits us to construct maps of relative interest in particular entities. These biases are illustrated in Table 1, which present significantly over-represented entities in each of three major American newspapers, as scored by the number of standard deviations above expectation. These over-represented entities include local political and business figures (e.g. *Brad Fitzpatrick* of LiveJournal and television star *Oprah Winfree*, based in San Francisco and Chicago, respectively), local sports figures/teams (e.g. *Steve McMichael* and *Dwayne Wade*), and even local dialect (e.g. *Estados Unidos* in heavily Cuban Miami).

San Francisco Chronicle		Chicago Tribune		Miami Herald	
Entity	Score	Entity	Score	Entity	Score
Wuksachi Lodge	24.39	Steve McMichael	24.38	Estados Unidos	16.49
Brad Fitzpatrick	24.39	Chicago Tribune	23.20	Broward	14.65
Golden Gate Park	15.29	Richard J. Daley	15.89	Dwyane Wade	13.60
Bay Area	12.03	White Sox	13.86	Miami-Dade County	12.48
San Francisco, CA	10.20	Ozzie Guillen	10.42	Marlins	11.55
Giants	4.66	Oprah Winfrey	10.39	Adam Kidan	11.08

Table 1. Overrepresented Entities in Three Major U.S. Newspapers

Note that these source biases reflect interest in the primary location the given paper. However, the set of U.S. cities with spiderable online daily newspapers is surprisingly small, so sophisticated modeling and analysis is needed to interpolate this data throughout the United States. Our contributions in this paper are:

- *News Source and Coverage Analysis* – We discuss the basic mechanics of large-scale news acquisition and analysis, including spidering and duplicate document identification. We visualization techniques to demonstrate how news sources are distributed around the country. We do *not* discuss the details of our entity extraction / NLP analysis, which has been previously presented in [8, 9, 10].
- *Source-Influence Modeling for Entity Analysis* – Interpolating entity distributions from roughly 500 different newspapers to reflect relative interest over the entire United States requires some sophisticated modeling. In this paper, we present the details of our news source-influence model. This model is based on computing an appropriate sphere of influence for each given newspaper, as a function of its circulation, location, and the population distribution of the United States. We also describe our model for allocating the relative contribution of all news sources influencing each location.
- *Visualization Techniques for Data Maps* – The engineering of our spatial news analysis system required a variety of decisions concerning visualization techniques, which may be of independent interest.
- *Identifying Interesting Datamaps* – Our system is capable of constructing datamaps for thousands of different entities on a daily basis – far more than can be exhaustively viewed by any human observer. Many datamaps are uninteresting in that they show no

- *Andrew Mehler is with the Department of Computer Science, Stony Brook University, E-mail: mehler@cs.sunysb.edu*
- *Yunfan Bao is with the Department of Computer Science, Stony Brook University, E-mail: ybao@cs.sunysb.edu*
- *Xin Li is with the Department of Computer Science, Stony Brook University, E-mail: xinli@cs.sunysb.edu*
- *Yue Wang is with the Department of Computer Science, Stony Brook University, E-mail: yuewang@cs.sunysb.edu*
- *Steven Skiena is with the Department of Computer Science, Stony Brook University, E-mail: skiena@cs.sunysb.edu*

Manuscript received 31 March 2006; accepted 1 August 2006; posted online 6 November 2006.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

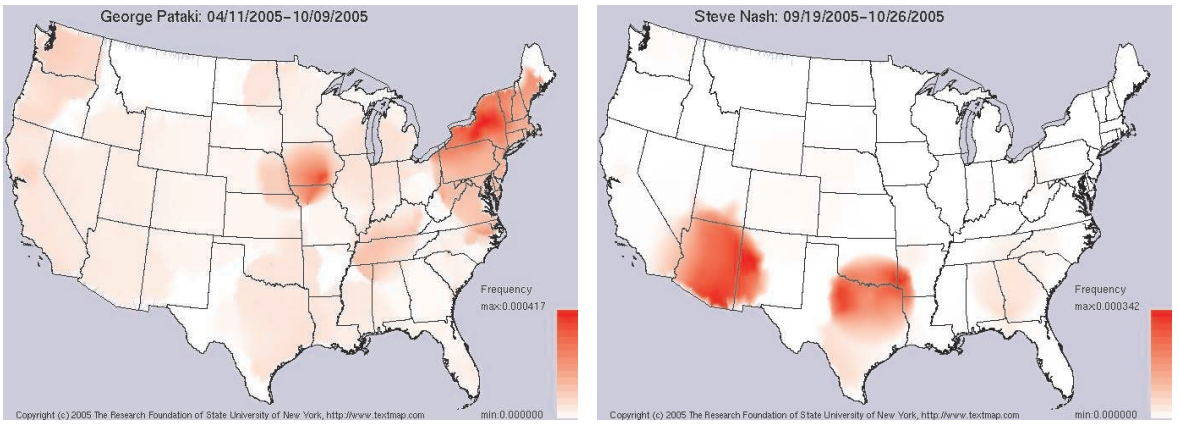

Fig. 1. Entity datamaps for New York Governor *George Pataki* (l) and Dallas/Phoenix basketball star *Steve Nash* (r).

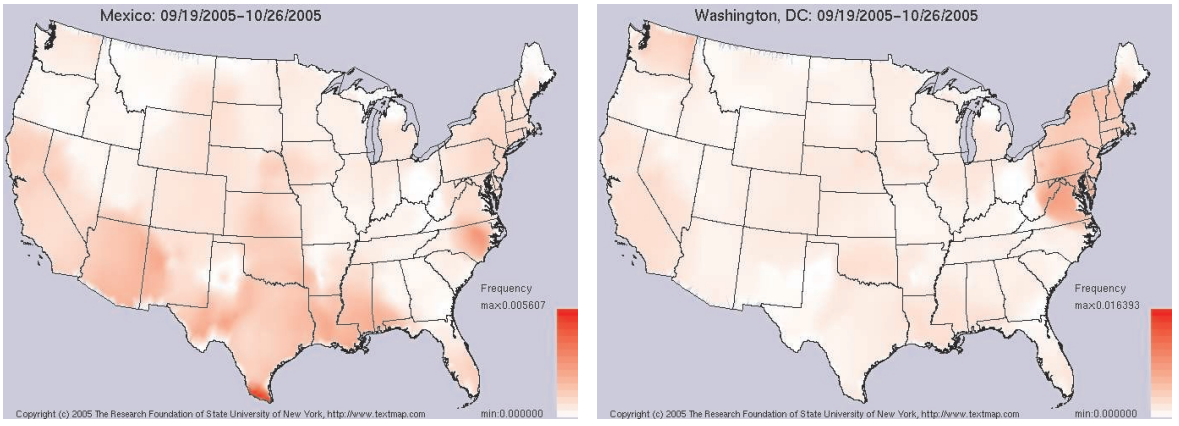

Fig. 2. Datamaps for two geographic entities, namely *Mexico* and *Washington, DC*.

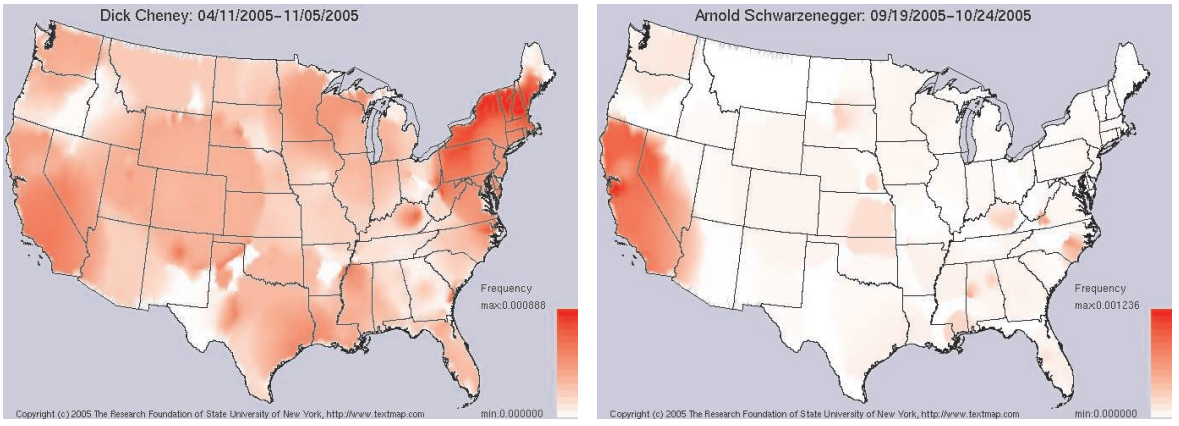

Fig. 3. Datamaps for Vice President Dick Cheney (l) and California Governor Arnold Schwarzenegger (r).

regional bias. Identifying the most interesting datamaps for visual inspection requires the development of statistical methods for evaluating geographical bias.

We propose a total of five different discrimination functions, each a variant of two basic methods, namely variance analysis and connected-component histograms. We present the results of computational experiments that demonstrate that, while all methods can successfully distinguish spatially-interesting entity maps from those of unbiased entities and random distributions, the best in practice appear to be the weighted variance and maximum gap discriminators.

1.1 Previous Work

Our work lies at the intersection of modeling and visualization. We start with gross geographic information about news content centered around select newspaper locations, and attempt to model and predict the spatial level of interest throughout create the exact data. This differs from many data map applications e.g. [13], where one seeks to smooth or interpolate data to make it the information easier to visualize. News-specific modeling lie at the foundation of our work.

There is an enormous literature on fitting and analysis of geographic data. This work is not directly applicable to us because (1) our focus is on larger-scale aspects of news processing and modeling, and (2) relatively simple approaches work well enough for us now.

Still, we give a brief survey of this literature. Tufte [18] gives an overview of the principles of visualizing data. Statistical geographic maps are studied in depth in [16], including spatial statistical methods such as *spatial auto correlation*. We expect both geographically-biased and unbiased maps both to have high spatial autocorrelation, a result of our modeling and the nature of news. Comprehensive overviews of map layout, color, and all aspects of map creation are covered in Brewer [2], who states that sequential coloring schemes should rely on lightness as the primary method of contrast.

Fuentes [4] discusses methods for interpolating spatial processes, and models for spatial correlations. This methods utilize the *variogram*, a model which assumes the variance is determined only by the distance between two points. A nearest neighbor method for estimating the variogram is given in [19]. Kriging is another process used for interpolating spatial data [17]. Mitchell [12] gives an overview of spatial analysis, describing statistics for measuring continuous geographic data. We believe the phenomena we are studying to be continuous, or at least piecewise continuous. Significance statistics for spatial processes compare an observed data set to random data, and determine if a given process is random. Methods such as the local G-Statistic exist for clustering continuous spatial data.

Fotheringham, et. al. discuss Quantitative Geography in [3]. They discuss spatial regression techniques, statistical spatial inference, and spatial modeling.

Miller and Han [11] give frameworks and algorithms for data mining tasks on spatial and geographical databases. They focus on representation, rule mining, clustering, outlier detection, and other data mining tasks. Considerable work has been done on smoothing algorithms to help visualize data, including head-banging (a median-based smoother) [13, 5], splines, surface modeling, and wavelet transforms. Smoothing algorithms start with precise data points (for instance, precise soil measurements made at particular locations) and then seek to interpolate where no measurements were taken and filter out noise.

Lydia is the front-end analysis system we employ in this project to do entity extraction on our newspaper sources. An in-depth discussion on the architecture of the *Lydia* natural language processing (NLP) pipeline can be found in [10]. *Lydia* has been adapted to work on a variety of other text sources, including as blogs [9], and served as the basis for a question answering system discussed in [8].

2 Text Acquisition and Analysis

The data for our analysis come from U.S. newspaper websites. In this section, we describe the mechanics of acquiring representative news text through spidering and duplicate article detection before reporting an analysis of the coverage breadth of our news sampling.

2.1 Text Acquisition / Spidering

Our text is acquired from online newspaper sources by spidering the websites. A *spider* is a program that attempts to crawl an entire web domain, and download all the web-pages. It would clearly be infeasible for us to build custom spiders for each of the roughly 800 daily newspapers in the United States and approximately 300 daily English language newspapers overseas. Instead, we developed a universal spider that downloads all the pages from a newspaper website, extracts all new articles, and normalizes them to remove source-specific formatting and artifacts.

Our spiders are built around the popular program *wget* [6] with the correct parameters; regulating the recursion depth (two levels suffices for most newspapers), user identification (via cookies), and wait time (for politeness, we never hit a website more than once per second). The news sources are divided by time zone, with many (at least 30) newspapers spidered in parallel across a given zone. Each download starts at 12:30AM local time. Each newspaper takes about 20-80 minutes to download, with a raw download size of 40-130MB.

2.2 Duplicate Article Identification

An interesting issue we faced concerned identifying duplicate and near-duplicate news articles. Repeated instances of given news articles can skew the significance of our spatial trends analysis, so we need eliminate duplicate articles before subsequent processing. Duplicate articles appear both as the result of syndication and the fact that old articles are often left on a website and get repeatedly spidered.

Schleimer et al.[14] describe a clever solution to this problem in the context of plagiarism detection. By comparing hash codes on all overlapping windows of length w appearing in the documents, we can identify whenever two documents share a common sequence of length w, although at the cost of an index at least the size of the documents themselves. However, the index size can be substantially reduced by a factor of p with little loss of detection accuracy by only keeping the codes which are congruent to 0 mod p. This will result in a different number of codes for different documents, however. We discovered little loss of detection will happen if we select the c *smallest* codes congruent to 0 mod p for each article. The Karp-Rabin string matching algorithm [7] proposes an incremental hash code such that all codes can be computed in linear time.

Through experimentation, we discovered that taking the 10 smallest hashes of windows of size 150 characters that are congruent to 0 mod 100 gives (1) a good sub-sampling of the possible hashes in a document, (2) a reasonable probability that if two articles are near duplicates, then they will collide on at least two of these hashes and (3) a reasonable probability that if two articles are unique, then they will not collide on more than one of these hashes. Our experimental set of $3,583$ newspaper days resulted in a total of $253,523$ unique articles with $185,398$ exact duplicates and $8,874$ near duplicates.

2.3 News Coverage Analysis

Every local newspaper has a readership centered around the city in which it is located. The size of the *sphere of influence* around a given paper is a function of (1) its readership (naturally measured by circulation or a proxy such as web hits) and (2) geographic population density. Details of our influence analysis will be presented in Section 3, but here we discuss its consequences on sampling density and potential for multi-source integration.

We have attempted to spiders all of the roughly 800 daily U.S. newspapers we are aware of through authoritative web sources. Many newspaper websites are no longer active or are highly seasonal (particularly school newspapers which cease activity for summer and other break periods). Others (primarily small limited circulation papers) employ robot.txt files or even block IP address to prevent us from spidering them. The upshot is that we get occasional spidering data from roughly 600 online newspapers, however on any given day only about 500 sources are active.

We can use our model of influence to visualize the coverage of the sources we spider. Figure 4(l) presents a datamap where a city's heat is a function of the number of newspapers it is influenced by. We get

significant coverage throughout the entire country, excepting isolated border locations around Maine, Minnesota, and Texas. The regions covered by more than ten sources are emphasized in the binary map of Figure 4(r). Many of these areas are surprisingly scattered around the country, reflecting intense competition among small papers in many local markets.

Figure 5 measures the relative raw volume of text that influence residents of each location. The national media centers of California

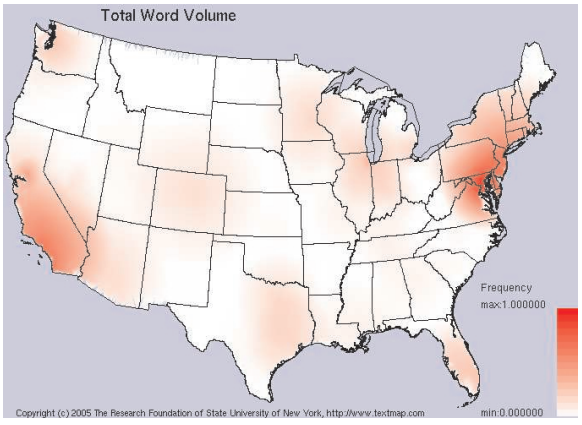

Fig. 5. Media exposure by location as measured by volume of total words.

and New York are significantly over-represented by this measure.

3 Source-Influence Modeling

The *heat* any given entity e generates in a given location s is a function of its frequency of reference of e in each of the newspapers that have influence over that city. The relative frequency of entity e in the source s_i is given by

$$\text{heat}(e,s) = \frac{\text{references}(e,s)}{\sum \text{references}(e_i,s)}. \quad (1)$$

Thus each heat value is from 0 to 1, giving the frequency of reference to the particular entity. A heat value of 0.05 implies that the entity is referenced once for every 20 entity references over the universe of all entities. The heat of an entity is a relative measure. Areas of high news volume may boast more absolute references to a given entity, but the popularity (heat) is determined by its *frequency* of reference.

3.1 Estimating Source Influence

Any analysis involving both the *New York Times* and *Ithaca Times* must capture the *influence* relation between a news source and a location. This influence relation must take into account the distance from the source location, the circulation of the source, and other relevant features.

Constructing this influence function forms the basis of our datamap model. The radius of influence of a news source depends on the circulation of the newspaper, and the population of nearby cities. Cities inside of this circle will be influenced an amount depending on their distance from the center; with maximum influence at the newspaper's location. This is captured in the equation

$$influence(s,c) = \begin{cases} 0 & \text{if } distance(s,c) > \text{radius of influence of s} \\ f(distance(s,c)) \times \text{max-influence}(s) & \text{else} \end{cases}$$

where f is some decay function. When f is linear, the influence of a newspaper can be thought of as a cone centered at the newspaper, with height the maximum-influence, and base the circle of influence. Of course other decay functions can also be justified. Cities lying outside this circle will receive zero influence from the given source. Circles associated with coastal cities will have a larger radius because there is no population where there is water. The population covered is still what the model predicts.

The maximum influence of a newspaper source is a combination of various circulation and ranking statistics of the source. In particular, we use a weighted combination of Alexa's reach per million web traffic analysis (www.alexa.org) and the weekday average circulation of the newspaper. Together they estimate the number of readers (both online and paper) of the newspaper. We estimate the online readership by multiplying the Alexa reach per million by the population of the U.S. in millions. We estimate the radius of influence supported by a given printed circulation through an estimate of the frequency with which people subscribe to newspapers. We model that 10% of the population covered by the radius of influence should equal the readership estimate. This parameter was anecdotally chosen, and seems reasonable.

3.2 Integrating Multiple Sources

Once the influence function of each source has been defined, the heat at a location can be calculated by a weighted average of the reference frequency of the entity in the source, weighted by the sources influence on the location.

For our model, we discretize space into roughly 25,000 U.S. cities/towns of all sizes. For a given C of cities, S of news sources, and E of entities, we get for the heat of an entity e at city c as per:

$$\text{heat}(e,c) = \frac{\sum_{s \in S} \text{heat}(e,s) \times \text{influence}(s,c)}{\sum_{s \in S} \text{influence}(s,c)}. \quad (2)$$

4 Visualization Issues

To render our datamaps, we use mesa/openGl graphics libraries. We know how to calculate the heat at cities, and openGl will interpolate heat between cities if given a polygon mesh. To get a polygon mesh from our set of cities, we used Shewchuk's C program *triangle* [1, 15]. This creates a Delauney triangulation of the cities.

To get the coloring for the datamaps, we set the scale such that the maximum heat value gets the highest red value, and the other values are scaled linearly. This makes two datamaps incomparable, since they are on different scales. However, using absolute scales makes most heat effects unobservable since very low values are imperceivable. Our method ensures we will always have maximum contrast between the highest and lowest heat values.

5 Identifying Significant Datamaps

Once we have the ability to calculate a large number of entity datamaps, we are faced with the problem of automatically screening which of these are interesting to look at, i.e. suggest significant geographic bias vs. entities of uniform national interest.

In this section we present methods for quantifying the geographic disparity of a datamap. Geographic disparity does not have a precise definition, and proves difficult to quantify for a variety of reasons. Because the population density is highly non-uniform, the relative sizes of significant intensity proves misleading. The area covered by "red" and "blue" states on the electoral map overstates the degree of red/Republican dominance due to their strength in the sparsely populated West. The news distributions resulting from our model proves to have large numbers of local optima. Do these represent distinct regional sources of elevated interest or are they modeling artifacts?

We consider two distinct classes of methods, based on variance analysis and connected component analysis respectively.

5.1 Variance Analysis

Statistical variance measures the deviation of data values from their mean. We would expect that datamaps showing higher statistical variance of intensity values more likely reflects regional biases, although similar variance measures can be derived from simple checkerboard patterns.

We define two scores reflecting this measure:

- *Variance* – Our datamap construction gives us heat values for each of 25,374 cities. For this measure, we compute the variance of these 25,374 heat values.

- *Weighted Variance* – The variance measure will be biased by absolute heat. If we scale every value on a datamap by some constant, the variance will also be scaled (by the square of the constant), although the underlying distribution is essentially the same. For this measure we divide the variance by the mean of the 25,374 values.

5.2 Connected Component Analysis

Consider a datamap with very high geographic disparity. If we look at the ten cities with the highest heat values, we would expect them to be clustered close together. If the datamap has no disparity, then the top ten cities will probably be scattered all about the country. Suppose we continue to look at the top 20, 30, 40 cities and so on. Datamaps with high disparity should have clusters of cities, while datamaps with no disparity should remain scattered. This motivates the idea of the *connected component profile* for a datamap.

Consider the graph G on the set of cities c induced by adding an edge between every pair of nearby cities, That is $(c_1, c_2) \in E$ if $distance(c_1, c_2) < d$. We can define a profile for a datamap by counting the number of connected components in G only including cities above a certain heat value. The profiles for regional figure *Wayne Gretzky* and national term *America* are shown in Figure 6. We see the regional term has a long period of a small number of connected components, corresponding to the high concentration area.

We propose three different features of the profiles for scoring methods:

- *Largest Gap* – A large gap between a connected component change would suggest that the entity is drawn from two separate (national and local) probability distributions.

- *Weighted Gap* – This method again uses the largest gap, but divides this number by the maximum heat. Thus terms that are generally popular wont be more heavily favored as they are in the first method.

- *Percentage Gap* – Finally we score based on the largest *percentage* heat change between a component change.

5.3 Results

To quantitatively evaluate our geographical bias measures, and avoid personal bias judgments as to the relative geographic disparity of heat maps, we conducted a large-scale experiment assessing how well they distinguish maps of regionally-biased entities from (1) maps of entities with presumed uniform national interest and (2) random maps generated under two different models. In the first model, the frequency of our imaginary entity in each news source is chosen from a uniform distribution, while in the second model it is chosen from a binomial distribution. Datamaps generated under these models are shown in Figure 7. The uniform model has greater local variance, while the binomial distribution is more globally smooth (fewer extreme values) because the values are centered around the mean in a binomial distribution.

We made two sets of real entity datamaps for our experiments. Entities likely to be geographically-biased include United States cities and local sports teams. Entities likely to have little bias include foreign cities, country names, national political figures, and entertainment terms. In total, we constructed 128 unbiased datamaps, and 400 biased datamaps. We also generated sets of random-valued datamaps, a total of 200 datamaps each for the uniform and binomial distributions.

The results of our scoring method are presented in Tables 2 and 3. For all five methods we calculate the mean, median, max, and min on each set for the raw score, and for the ranks. From these experiments, we can conclude that the *weighted gap method* gives the best results, since in general it scored biased maps higher than unbiased maps, and both higher then random maps. However, even though the other four scoring methods scored random maps above our biased datamaps, they score biased maps higher then the unbiased maps. Thus for real maps, each of the scoring methods has some significance.

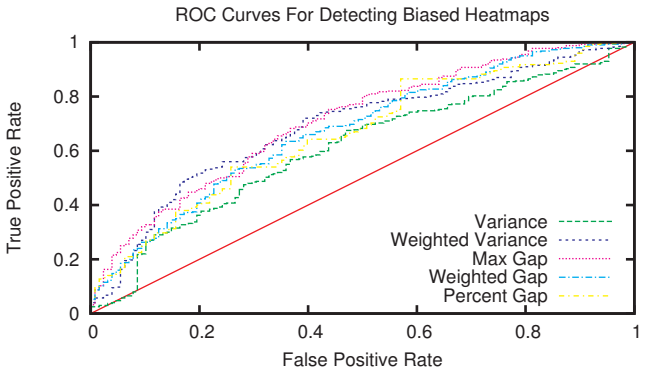

Fig. 8. ROC Curves For Datamap Classification

Weighted Gap	mean	min	median	max
biased	0.519	0.080	0.494	0.996
unbiased	0.367	0.053	0.323	0.947
uniform	0.080	0.035	0.070	0.208
binomial	0.098	0.037	0.088	0.254
Percent Gap	mean	min	median	max
biased	6.23	0.300	2.00	243.3
unbiased	2.08	0.273	1.25	18.0
uniform	7.55	0.294	1.68	967.6
binomial	2.53	0.411	2.49	4.55
Max Gap	mean	min	median	max
biased	1.66e-3	7.00e-6	5.12e-4	2.80e-2
unbiased	5.10e-4	6.00e-6	1.36e-4	9.34e-3
uniform	7.47e-2	6.55e-2	3.33e-2	2.05e-1
binomial	1.57e-3	5.19e4	1.33e-3	5.00e-3
Weighted Variance	mean	min	median	max
biased	6.60e-4	7.53e-6	2.30e-4	1.31e-2
unbiased	2.61e-4	3.50e-6	7.57e-5	3.23e-3
uniform	9.70e-2	5.58e-2	9.37e-2	1.58e-1
binomial	1.18e-3	7.56e-4	1.16e-3	1.94e-3
Variance	mean	min	median	max
biased	1.76e-7	1.80e-11	9.79e-9	9.14e-6
unbiased	9.35e-8	1.32e-12	3.36e-9	1.92e-6
uniform	3.36e-2	1.59e-2	3.22e-2	5.50e-2
binomial	8.13e-6	4.94e-6	8.00e-6	1.46e-5

Table 2. Performance of our 5 different scoring methods.

We are interested in distinguishing geographic bias among real world maps. Figure 8 shows the *Receiver Operating Curve* (ROC) for classifying the real world maps. We see that each scoring method is substantially above the 45-degree line. Thus all methods do better than random guessing. This, coupled with the fact that the methods are not perfectly correlated with each other (see Table 4) lead us to believe that a fusion method (a method that is a combination of all the methods) should do even better. A fusion method cannot improve when underlying different scoring methods are highly correlated, thus effectively all saying the same thing.

6 Conclusions

In this paper, we explored methods of spatially representing the popularity of news entities, by interpolating over reference frequencies in newspapers. We have presented a model for generating such data maps. Our visualizations show the geographic popularity of an entity, and possible geographic biases. To help identify maps that display in-

Weighted Gap	mean	min	median	max
biased	246	0	232	719
unbiased	354	21	352	867
uniform	747	450	770	927
binomial	685	408	690	925
Percent Gap	mean	min	median	max
biased	437	1	459	922
unbiased	589	25	660	927
uniform	506	0	553	925
binomial	395	134	367	911
Max Gap	mean	min	median	max
biased	576	200	632	926
unbiased	728	213	784	927
uniform	100	0	100	199
binomial	432	238	439	630
Weighted Variance	mean	min	median	max
biased	609	200	623	926
unbiased	728	216	769	927
uniform	100	0	100	199
binomial	367	238	367	498
Variance	mean	min	median	max
biased	649	245	641	926
unbiased	707	410	742	927
uniform	100	0	100	199
binomial	301	200	302	402

Table 3. Rank Performance of five different scoring methods.

	data	Wei. Gap	Max Gap	% Gap	Variance	Wei. Var.
Weighted Gap	bias	1	0.708	0.431	0.376	0.651
	unbias	1	0.722	0.856	0.621	0.738
Max Gap	bias	0.708	1	0.883	0.865	0.985
	unbias	0.722	1	0.966	0.910	0.931
Percent Gap	bias	0.431	0.883	1	0.984	0.928
	unbias	0.856	0.966	1	0.877	0.934
Variance	bias	0.376	0.865	0.984	1	0.919
	unbias	0.621	0.910	0.877	1	0.980
Weighted Var.	bias	0.651	0.984	0.928	0.919	1
	unbias	0.738	0.931	0.934	0.980	1

Table 4. Pearson Correlation Coefficients for five scoring methods.

teresting geographic bias, we have developed methods of scoring the maps using a spatial 'connected components' method. Our model has been implemented in the *Lydia* system (www.textmap.com) and currently generates thousands of different maps each day.

In future work, we hope to add new features and evaluation techniques to our visualizations. We would like to investigate the use of different decay functions and parameters to use in our model, and perhaps seek better theoretical or empirical justification for them. We also seek to compare our methods of evaluation to other recognized spatial statistics, such as spatial autocorrelation. Other directions we are interested in are more associated with technical aspects of our *Lydia* system, including (1) dynamic maps that animate the changes in news sentiment over a period of time and (2) sentiment maps which measure spatial biases in how entities are liked or disliked, instead of just how often they are mentioned.

ACKNOWLEDGEMENTS

We thank the referees for suggestions which significantly improved the presentation of this paper.

This research was partially supported by NSF grants DBI-0444815 and EIA-0325123.

REFERENCES

[1] H. Bao, J. Bielak, O. Ghattas, L. F. Kallivokas, D. R. O'Hallaron, J. R. Shewchuk, and J. Xu. Large-scale simulation of elastic wave propagation in heterogeneous media on parallel computers. *Computer Methods in Applied Mechanics and Engineering*, 1998.

[2] C. Brewer. *Designing Better Maps*. ESRI Press, 2005.

[3] A. S. Frotheringham, C. Brunsdon, and M. Charlton. *Quantitative Geography*. Sage Publications, 2000.

[4] M. Fuentes. Spatial interpolation of environmental processes. *Spatial Statistics Through Applications*, 2002.

[5] K.M. Hansen. Head-banging: robust smoothing in the plane. *IEEE Transactions on Geoscience and Remote Sensing*, 29:369–378, 1991.

[6] K. Hemenway and T. Calishain. *Spidering Hacks*. O'Reilly and Associates, Sebastopol CA, 2003.

[7] R. Karp and M. Rabin. Efficient randomized pattern-matching algorithms. *IBM J. Research and Development*, 31:249–260, 1987.

[8] J.H. Kil, L. Lloyd, and S. Skiena. Question answering with lydia. *The Fourteenth Text Retrieval Conference Proceedings (TREC 2005)*, November 15-18 2005.

[9] L. Lloyd, P. Kaulgud, and S. Skiena. News vs. blogs: Who gets the scoop. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, March 27-29 2006.

[10] L. Lloyd, D. Kechiagas, and S. Skiena. Lydia: A system for large-scale news analysis. *12th International Conference, SPIRE 2005, Lecture Notes in Computer Science*, 377:161–166, 2005.

[11] H. Miller and J. Han. *Geographic Data Mining & Knowledge Discovery*. CRC, 2001.

[12] A. Mitchell. *GIS Analysis Volume 2: Spatial Measurements & Statistics*. ESRI Press, 2005.

[13] M. Munigole, L. Pickle, and K. Simonson. Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine*, 18:3201–3209, 1999.

[14] S. Schleimer, D. Wilkerson, and A. Aiken. Winnowing: Local algorithms for document fingerprinting. In *22th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems*, pages 76–85, San Diego, California, USA, 2003.

[15] J.R. Shewchuk. Delaunay refinement mesh generation, 1997.

[16] T.A. Slocum, R.B. McMaster, F.C. Kessler, and H. Howard. *Thematic Cartography and Geographic Visualization*. Pearson Prentice Hall, 1999.

[17] M. L. Stein. *Interpolation of Spatial Data*. Springer, 1999.

[18] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[19] K. Yu and J. Mateu. *Nonparametric nearest-neighbor variogram estimation*. Springer, 1999.

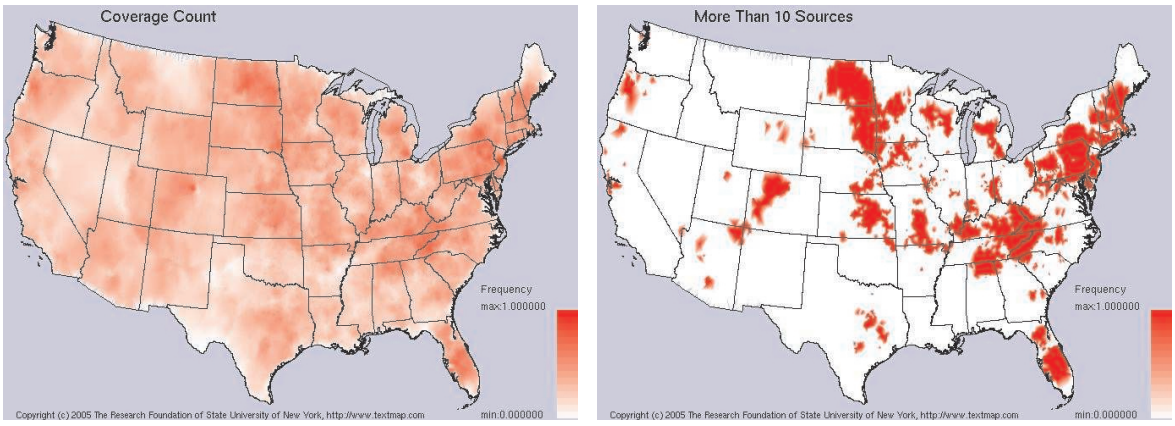

Fig. 4. The number of different news sources influencing each U.S. city (l), and the number of cities influenced by more than ten sources (r).

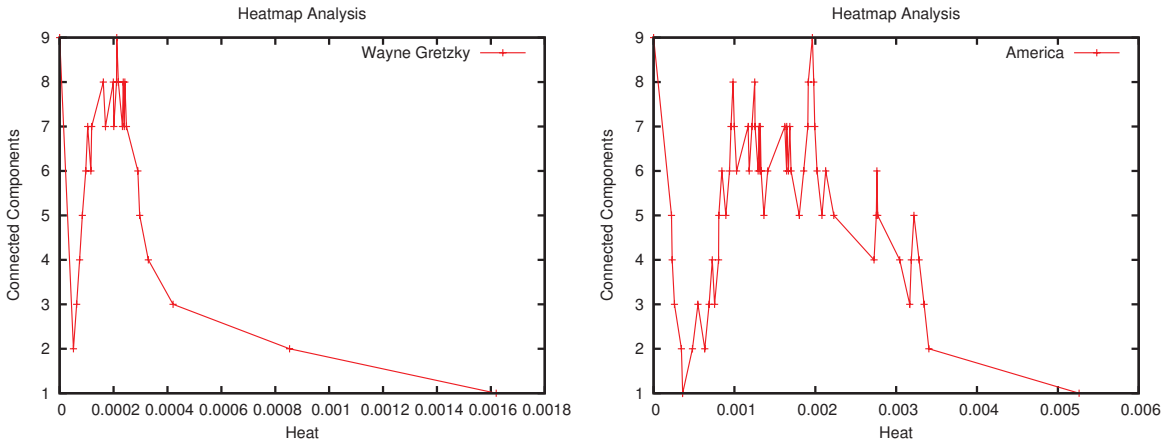

Fig. 6. Connected Component Profiles for Wayne Gretzky and America.

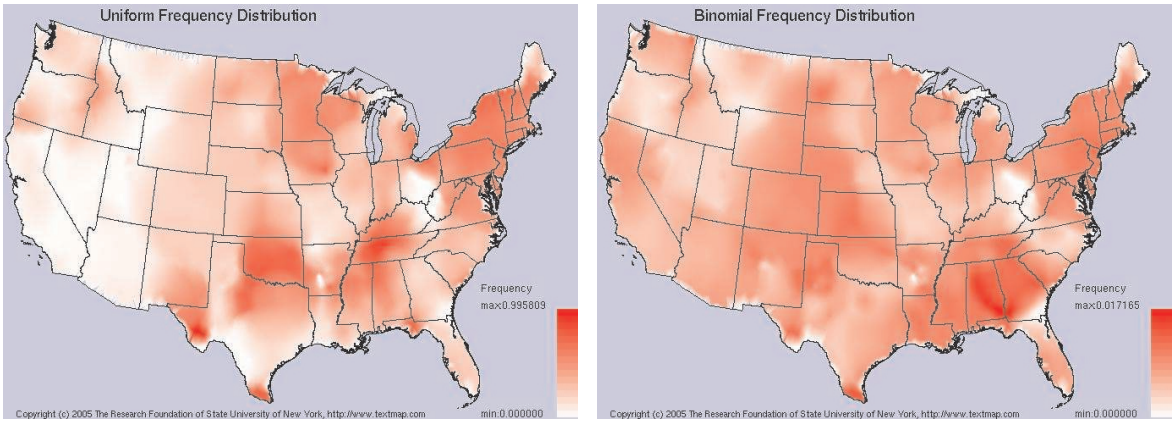

Fig. 7. Random Datamaps. The frequency of this imaginary entity in each source is given by a random uniform distribution, or binomial distribution