Adaptive Interpolation Algorithms for Temporal-Oriented Datasets

Jun Gao
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588, USA
jgao@cse.unl.edu

Abstract

Spatiotemporal datasets can be classified into two categories: temporal-oriented and spatial-oriented datasets depending on whether missing spatiotemporal values are closer to the values of its temporal or spatial neighbors. We present an adaptive spatiotemporal interpolation model that can estimate the missing values in both categories of spatiotemporal datasets. The key parameters of the adaptive spatiotemporal interpolation model can be adjusted based on experience.

1 Introduction

The fact of missing data can result in errors in many applications. For example, in the area of information visualization, missing data usually causes visualization failure or provides misleading interpretations of data [8]. The Standardized Precipitation Index (SPI) is a common and simple measure of the intensity and duration of drought at certain measured point locations [17, 20]. When there are missing data (e.g., a couple weeks gap), the SPI can not be calculated for any interval that includes the data gap [25].

The missing data can be estimated by interpolation methods based on the sampled values. There are many interpolation methods available for estimation. Some common methods are inverse distance weighting (IDW) [22], kriging [7], regression model [2], shape functions [16], splines [11], and trend surface analysis [26].

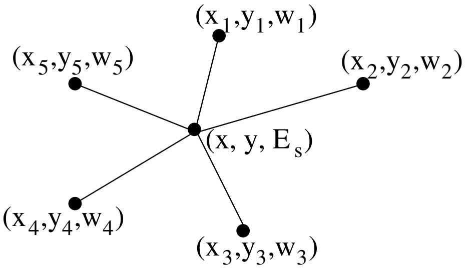

Figure 1. Target point and its spatial neighbors

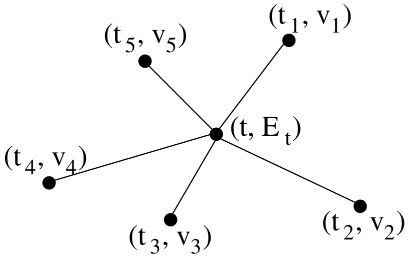

Figure 2. Target point and its temporal neighbors

Spatiotemporal datasets contain the information in both space and time. Let us look at a simple example. Suppose we need to interpolate the missing value of an attribute at the target point (x, y) at time t. The target point has five neighbors, $(x_1, y_1), \ldots, (x_5, y_5)$, and their values of that attribute at time t are known as $w_1, \ldots, w_5$. The target point also has some known values of the same attribute before or after time t, $(t_1, v_1), \ldots, (t_5, v_5)$. Figure 1 illustrates the spatial relation between the target point (x, y, E_s) (E_s is the estimate by spatial method) and its five surrounding neighbors (x_i, y_i, w_i) and Figure 2 illustrates the temporal relation between the target point (t, E_t) (E_t is the estimate by temporal method) at time t and its temporal neighbors (t_i, v_i).

Spatiotemporal datasets can be classified into two categories: temporal-oriented dataset and spatial-oriented dataset. In the temporal-oriented dataset the temporal relationship between the data values is stronger than the spatial relationship. For example, in the United States people who vote for Democrat will more likely vote for Democrat again in the next election. Hence, in the USA presidential election dataset, the outcomes in one state may be same for many years, while the outcomes of two neighboring states

Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)
1530-1311/06 $20.00 © 2006 IEEE

may be significantly different. Another example is the air ticket dataset. The air ticket price is more likely higher in a peak season for tourism and this trend may continue for many years. In the spatial-oriented dataset the spatial relationship between the data values is stronger than the temporal relationship. For example, in the climate dataset, the temperature sampled in one weather station may be very similar to that in a neighboring weather station, but may be very different from the temperature sampled one day ago. Another example is the dataset on heavy metal pollutants in floodplain soils. It is known that the heavy metal pollutants depend on several factors, and one of the most important is the distance to the river [2].

The spatiotemporal interpolation model developed by Gao and Revesz is a general interpolation model for spatiotemporal datasets [10]. Let E_s be the estimated value using spatial method, E_t the estimated value using temporal method, α the weight of E_s, and β the weight of E_t. Then the overall estimation E can be calculated as follows:

$$E = \alpha \times E_s + \beta \times E_t \qquad (1)$$

where $\alpha + \beta = 1$ and $0 \leq \alpha,\ \beta \leq 1$.

In order to apply the above interpolation model into a specific application, the following three questions need to be answered:

(1) What spatial interpolation method is used to determine E_s?

(2) What temporal interpolation method is used to determine E_t?

(3) What is the relationship between α and β?

There are many answers for the first two questions since there are numerous interpolation methods in the world. In this study we adopt IDW as the basic method, modify and improve it to estimate E_s and E_t. We choose IDW due to the following reasons. First, IDW is a popular method and used in problems as diverse as predicting of rainfall and temperature, mapping of crop spraying, monitoring extent of contaminated groundwater plumes or quantitatively assessing the extent of contamination in aquatic sediments [24]. Second, IDW is easy to use and has low computation charge [6]. Compared with other methods, most notably kriging, the IDW method is simpler to program and does not require pre-modelling or subjective assumptions in selecting a semi-variogram model [24]. Third, IDW provides a measure of uncertainty of the estimates that is directly related to the values being estimated, in contrast to kriging standard deviation which is based on the modelled semi-variogram [1].

A step function is a natural and simple solution to the third question. In a step function, a parameter σ and a threshold θ are needed. If $\sigma < \theta$, then we set $\alpha = 1$ (or $\alpha = 0$) and $\beta = 0$ (or $\beta = 1$). Another simple and natural solution is a linear function which is based on the linear combination of α and β.

The rest of the paper is structured as follows. Section 2 illustrates the representation of IDW-based spatiotemporal interpolation in constraint databases. Section 3 describes the application of adaptive spatiotemporal interpolation model in forecasting presidential election. Section 4 describes the experimental methods and results. Section 5 gives a short introduction on other spatiotemporal interpolations methods. Finally, Section 6 presents some ideas for future work.

2 IDW-based Spatiotemporal Interpolation in Constraint Databases

2.1 Inverse Distance weighting

IDW assumes that the value at an unsampled location is a distance-weighted average of values at sampled points within a defined neighborhood surrounding the unsampled point [2].

The relative importance of the known values is reflected by the weights assigned by the IDW method to them. In the IDW method the sum of the weights is equal to 1, and the weights are assigned proportionally to the inverse of the distance between the known and unknown locations.

Let λ_i be the weight for the individual location, and y_i the variable observed in the sampled location. IDW interpolations are of the form:

$$y = \sum_{i=1}^{N} \lambda_i \cdot y_i \qquad \lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})^p}$$

For simplicity in this study we choose $p = 1$.

2.2 Spatiotemporal Interpolation in Constraint Databases

Constraint databases generalize relational databases by finitely representable infinite relations [13, 18]. A constraint relation is a finite set of constraint tuples, where each constraint tuple is a conjunction of atomic constraints using the same set of attribute variables [13, 18]. A constraint database is a finite set of constraint relations [13, 18]. We describe how to represent the spatiotemporal interpolation method in a constraint database through the example in Figures 1 & 2.

Let $Neigh(x, y, x', y', i)$ be a relation that stores the five closest neighbors (x', y') of point (x, y), where $1 \leq i \leq 5$.

Then we can get the relation that stores the five distances between point (x, y) and its five closest neighbors. We denote this relation as $Dist(x, y, d, i)$, where d is a Euclidean distance between point (x, y) and one of its five closest neighbors. It can be expressed in Datalog as follows:

$$\begin{aligned} Dist(x, y, d, i) \quad :- \quad & d = \sqrt{(x - x')^2 + (y - y')^2}\ , \\ & Neigh(x, y, x', y', i). \end{aligned}$$

Using the IDW equation and the relation $Dist(x, y, d, i)$ we can get the relation $Weights(x, y, \lambda, i)$ which stores the weights of the five neighbors of point (x, y) as follows:

$$\begin{aligned} Weights(x, y, \lambda, i) \quad :- \quad & \lambda = \frac{d^{-1}}{d_1^{-1} + d_2^{-1} + d_3^{-1} + d_4^{-1} + d_5^{-1}}\ , \\ & Dist(x, y, d, i), \\ & 1 \leq i \leq 5. \end{aligned}$$

Where d_1 means the distance d in the relation $Dist(x, y, d, i)$ when $i = 1$. d_2 means the distance d when $i = 2$, and so on.

Now the spatial estimation value E_s of an attribute at time t at the target point (x, y) can be expressed by the constraint tuple as follows:

$$\begin{aligned} R_{space}(x, y, E_s) \quad :- \quad & E_s = w_1\lambda_1 + w_3\lambda_2 + w_3\lambda_3 \\ & \qquad + w_4\lambda_4 + w_5\lambda_5\ , \\ & Weights(x, y, \lambda, i), \\ & 1 \leq i \leq 5. \end{aligned}$$

Where λ_1 means the weight λ in the relation $Weights(x, y, \lambda, i)$ when $i = 1$ and so on.

Similarly, let $Neigh_{time}(x, y, t, t', i)$ be a relation that stores the five time shots close to time t at point (x, y) when the values of the attribute of interest are known. Then we can get the relation that stores the five temporal distances between time t and its five temporal neighbors. We denote this relation as $Dist_{time}(x, y, t, d_t, i)$, where d_t is the temporal distance between time t and one of its five temporal neighbors. It can be expressed in Datalog as follows:

$$\begin{aligned} Dist_{time}(x, y, t, d_t, i) \quad :- \quad & d_t = |t - t'|\ , \\ & Neigh_{time}(x, y, t, t', i). \end{aligned}$$

Using the IDW equation and the relation $Dist_{time}(x, y, t, d_t, i)$ we can get the relation $Weights_{time}(x, y, t, \lambda_t, i)$ which stores the weights of the five temporal neighbors of time t as follows:

$$\begin{aligned} Weights_{time}(x, y, t, \lambda_t, i) \quad :- & \\ & \lambda_t = \frac{d_t^{-1}}{d_{t1}^{-1} + d_{t2}^{-1} + d_{t3}^{-1} + d_{t4}^{-1} + d_{t5}^{-1}}, \\ & Dist_{time}(x, y, t, d_t, i), \\ & 1 \leq i \leq 5. \end{aligned}$$

Where d_{t1} means the distance d_t in the relation $Dist_{time}(x, y, t, d_t, i)$ when $i = 1$. d_{t2} means the distance d_t when $i = 2$, and so on.

Now the temporal estimation value E_t of an attribute at time t at the target point (x, y) can be expressed by the constraint tuple as follows:

$$\begin{aligned} R_{time}(x, y, t, E_t) \quad :- \quad & E_t = v_1\lambda_{t1} + v_2\lambda_{t2} + v_3\lambda_{t3} \\ & \qquad + v_4\lambda_{t4} + v_5\lambda_{t5}\ , \\ & Weights_{time}(x, y, t, \lambda_t, i), \\ & 1 \leq i \leq 5. \end{aligned}$$

Where λ_{t1} means the weight λ_t in the relation $Weights_{time}(x, y, t, \lambda_t, i)$ when $i = 1$ and so on.

Based on the above relations, the total estimation value E of point (x, y) at time t can be expressed by the following constraint tuple:

$$\begin{aligned} R(x, y, t, E) \quad :- \quad & E = \alpha \times E_s + \beta \times E_t\ , \\ & R_{space}(x, y, E_s), \\ & R_{time}(x, y, t, E_t). \end{aligned}$$

3 Application: Presidential Election Prediction

Presidential election prediction is an application where estimation is widely used. Many presidential election forecasting models were introduced and the accuracy of some models is admirable.

3.1 Presidential election forecasting models

The modern age of election forecasting began in the late 1970s. Among the earliest presidential forecasting models were [9, 23, 21, 14]. Most of these models have been amended, updated and are still used. Nearly all the presidential election forecasting models use *multi-variate ordinary least squares regression*, a common statistical method in the social sciences [12]. This approach enables the forecaster to identify factors that have influenced past election outcomes and determine how much weight should be given to each factor. The appropriate data for the present election are then inserted into the model to produce a forecast.

All these models are frequently cited for their use in forecasting and the accuracy is admirable, however, most of them share limitations. For example, the choice of factors to include in the model adds to the uncertainty. The decision to include one set of variables, such as presidential popularity and growth in GNP, rather than another, such as the rate of inflation and unemployment, changes the prediction outcome [12]. Most models are limited by the lack of historical information on the relationship between political and economic fundamentals and elections [12]. Hence we turn the direction into the historical election data itself and use it as the basis of spatiotemporal interpolations without a set of variables.

3.2 Spatiotemporal interpolation in election prediction

First let us look at an example. Suppose we are back in November 2004 and want to predict the percentage vote for John Kerry in the 2004 USA presidential election in Alachua county, Florida. The available data can be divided into two parts. One part is the previous election results in Alachua, namely the presidential election votes in 2000, 1996, 1992 and etc. The other part is the 2004 presidential election votes in the neighboring counties of Alachua. Before we apply the interpolation method to do the prediction, we need to clarify one thing. In general, interpolation is used for estimating unknown values at unsampled points, based on a set of measured values at sample points. However, we predict the presidential election results at county level or state level, which are regional data. We bridge the gap in this way. Each county can be treated as a point since it provides the measured values like total votes in that county, votes for individual candidates and the distance between two counties could be calculated as the distance between the centroids of two counties. Regarding the three questions raised in Section 1, we solve them as follows.

To calculate the temporal estimation E_t we use a variant of the IDW methods that measures "distance" in terms of time difference instead of spatial difference. We call this method *inverse linear temporal method.* we also introduce another method, *inverse exponential temporal method*, that assigns weights that decrease exponentially with the time difference, i.e., if we look back in time n years and have one data in each of the past n years, then the weight of the data i years back in time will be $\frac{1}{2^i}$ for $1 \leq i \leq (n-1)$ and $\frac{1}{2^{n-1}}$ for n years back. Note that the last two weights will be the same and with this rule the sum of the weights is still 1.

When we use the IDW method to calculate the spatial estimation E_s, a problem arises. It is not reasonable to use the actual votes in the neighboring counties, because those votes are not known yet. A possible solution is to use the estimated data in the neighboring counties, which can be created by many methods such as our inverse linear or inverse exponential temporal methods.

We first tried the IDW with uniform distances and it works as follows. Suppose we want to predict the votes for county C, which has the following neighboring counties, $N_1, N_2, \ldots, N_k$. We assume all the distances between counties C and $N_i, 1 \leq i \leq k$, are the same. Hence each neighbor N_i has exactly the same weight $\lambda_i = \frac{1}{k}$, $1 \leq i \leq k$. Instead of the uniform distances we also tried the real distances. We find that the differences between the two versions are extremely small in our case. Therefore, the much more complicated standard IDW method using exact distances can be simplified by IDW method with uniform distances using topological neighbors without any significant change in the accuracy of the result.

In order to use the step function we need to find the parameters σ and θ appropriate for election prediction application. We believe voting consistency over time in a county is a reasonable parameter. Hence we choose σ as the changes in the vote percentages of all pairs of subsequent presidential elections in a county. A smaller σ means that the values in a county are more consistent over time, thus we can rely more on the temporal interpolation method. We choose θ as a constant, say 1%, 2% and so on. In addition to the step functions, we also experimented with linear functions of the form $\alpha = c\,\sigma + d$ with different values for the constants c and d. However, the linear functions did not work as well as the step functions. One likely explanation is that the temporal and IDW methods give similar variations for most counties, that is, when the temporal estimation value is higher (or lower) than the original data, then the IDW estimation value is also higher (or lower). That makes it difficult to find a good linear function.

4 Evaluation

In order to test our idea, we estimated the votes for the 2004 democratic candidate for USA president (John Kerry) in states of California, Florida, and Ohio using our spatiotemporal interpolation method and compared them with the actual votes.

4.1 Methodology

Several measures are suitable for experimentally comparing the accuracy of interpolation methods. We use mean absolute error (MAE) and root mean square error (RMSE).

$$MAE = \frac{\sum_{i=1}^{N} |F_i - A_i|}{N} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{N} (F_i - A_i)^2}{N}}$$

F_i: Prediction value, A_i: Actual measurement, N: Number of data.

Let $VPstate_e$ be the estimated statewide vote percentage for a given party. Similarly, let $VPstate_a$ be the actual statewide vote percentage for a given party.

$$VPstate_e = \frac{\sum E_i \times V_i}{\sum V_i}$$

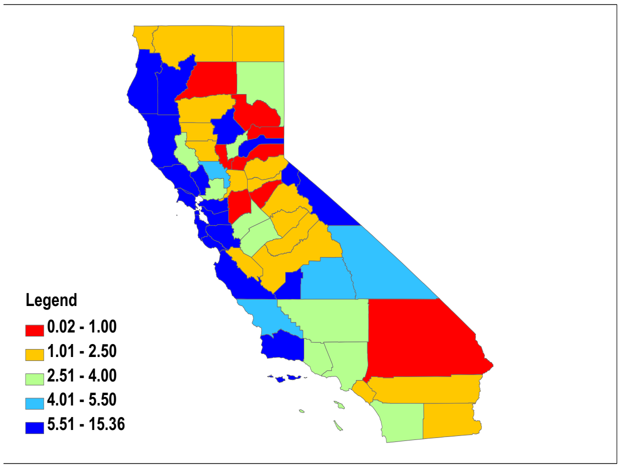

Figure 3. Differences between the estimated vote percentages using step functions and the actual vote percentages at the county level in California, USA

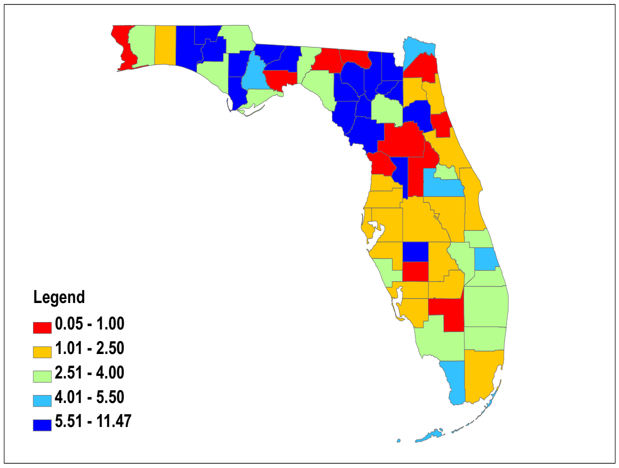

Figure 4. Differences between the estimated vote percentages using step functions and the actual vote percentages at the county level in Florida, USA

E_i: Estimated vote percentage for a given party in county i. V_i: The number of all voters in county i.

Then we can calculate the error of statewide total vote percentage (TE), which is a more interesting measure in the voting prediction area.

$$TE = |VPstate_e - VPstate_a|$$

4.2 Results

Figures 3-5 shows the voting prediction results on the 2004 USA presidential election at the county level in the states of California, Florida, and Ohio. The numbers of the legend indicate the differences between the estimated vote percentages using step functions and the actual vote percentages. We can see that for all the three states, the differences are less than 1% in some counties and less than 4% in most counties.

Table 1 records our experimental results at the state level. We can see that the performance of spatiotemporal step functions and inverse exponential temporal methods is the best, getting comparatively precise predictions, especially in predicting the 2004 USA presidential election in Florida. Spatiotemporal step functions (with $\theta = 7\%$) predict for the 2004 USA presidential election, the democratic candidate (John Kerry) will win 46.00% votes in Florida, and the actual result is 47.09%, hence the discrepancy (TE) is only 1.09%. This contrasts favorably with a CNN poll which predicted only 42% for John Kerry shortly before the election, i.e., it had a TE of more than 5%.

The experimental results for California and Ohio are also impressive. Inverse exponential temporal method shows slightly better performance, TE is 3.46 and 3.18 in California and Ohio, respectively. For all three states, MAE and RMSE are reasonably low, between 2.39 and 6.83.

5 Related Works

Alternative spatiotemporal interpolation methods are given in [15, 16, 20]. Chomicki et al. [4, 5] give an alternative classification of spatiotemporal data based on their algebraic closure properties. The pioneering paper for the constraint database representation of various types of spatiotemporal databases is [3].

6 Conclusion and Future Work

Election data is a typical temporal-oriented dataset. The experimental results show that our spatiotemporal interpolation method can be a basis for an effective voting prediction system. Of course, any real voting prediction system would need to be fine-tuned by considering many additional variables. We plan to exploit it by factoring more variables. Other possible extensions are to design other interesting applications and experiment with corresponding datasets. This paper did not consider periodic spatiotemporal objects, which are considered for example in [19]. It remains an interesting open question how to interpolate periodic spatiotemporal data.

Table 1. Comparison of step function, temporal and IDW methods

	California 2004			Florida 2004			Ohio 2004		
	TE	MAE	RMSE	TE	MAE	RMSE	TE	MAE	RMSE
Spatial IDW	8.65	11.60	9.67	4.88	7.98	9.05	8.75	11.31	7.60
Spatiotemporal Step Function ($\theta = 7\%$)	3.49	4.51	6.26	1.09	2.40	5.18	3.57	4.37	3.57
Spatiotemporal Step Function ($\theta = 8\%$)	3.55	4.77	6.38	1.10	2.40	4.72	3.89	4.66	3.88
Spatiotemporal Step Function ($\theta = 9\%$)	3.49	4.51	6.26	1.10	2.39	4.61	3.27	4.05	3.14
Temporal Inverse Linear	5.46	6.66	7.25	2.68	3.81	5.12	4.10	5.09	3.74
Temporal Inverse Exponential	3.46	4.48	6.01	1.10	2.39	4.59	3.18	3.99	3.10

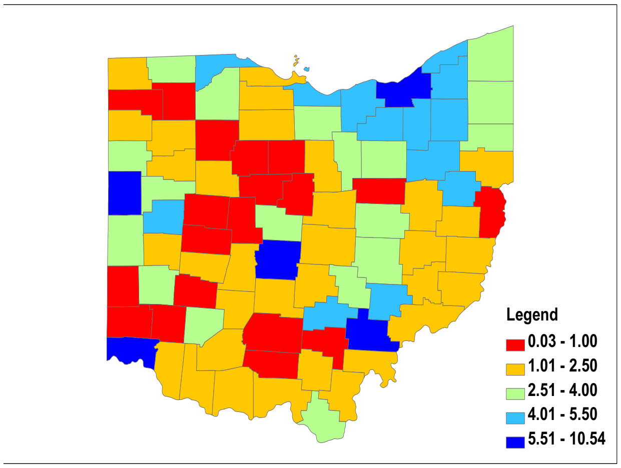

Figure 5. Differences between the estimated vote percentages using step functions and the actual vote percentages at the county level in Ohio, USA

7 Acknowledgements

The author was supported by an Othmer Fellowship of the University of Nebraska-Lincoln in the USA. Additional funding was also given by the USA National Science Foundation under grant EIA-0091530.

References

[1] G. S. Adisoma and M. G. Hester. Grade estimation and its precision in mineral resources: the Jackknife approach. *Mining Engineering*, 48(2):84-88, 1996.

[2] P. A. Burrough and R. A. McDonnell. *Principles of Geographical Information Systems*, Oxford, 1998.

[3] J. Chomicki and P. Revesz. Constraint-based interoperability of spatiotemporal databases. *Geoinformatica*, 3(3):211-243, 1999.

[4] J. Chomicki and P. Revesz. A geometric framework for specifying spatiotemporal objects. In *Proc. of 6th International Symposium on Temporal Representation and Reasoning*, IEEE Press, pages 41-46, May 1999.

[5] J. Chomicki, S. Haesevoets, B. Kuijpers, and P. Revesz. Classes of spatiotemporal objects and their closure properties. *Annals of Mathematics and Artificial Intelligence*, 39(4):431-461, 2003.

[6] F. Collins and P. Bolstad. A comparison of spatial interpolation techniques in temperature estimation. In *Proc. of the 3rd International Conference/Workshop on Integrating GIS and Environmental Modelling*, 1996.

[7] C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford University Press, 1998.

[8] C. Eaton, C. Plaisant, and T. Drizd. The challenge of missing and uncertain Data. In *Proc. of IEEE InfoVis Poster Compendium*, pages 40-41, IEEE Computer Society Press, 2003.

[9] R. Fair. The effect of economic events on votes for president. *Review of Economics and Statistics*, 60:159-173, 1978.

[10] J. Gao and P. Revesz. Voting prediction using new spatiotemporal interpolation methods, In *Proc. of the 7h Annual International Conference on Digital Government Research*, San Diego, May 2006.

[11] J. E. Goodman and J. O'Rourke, eds. *Handbook of Discrete and Computational Geometry*, CRC Press, New York, 1997.

[12] J. P. Greene. Forecasting follies. *The American Prospect*, vol 4 no. 15, 1993.

[13] P. Kanellakis, G. Kuper, and P. Revesz. Constraint query languages. *Journal of Computer and System Sciences*, 51(1):26-52, 1995.

[14] M. Lewis-Beck and T. Rice. Forecasting presidential elections: A comparison of naive models. *Political Behavior*, 6:9-21, 1984.

[15] J. Li, R. Narayanan, and P. Revesz. A shape-based approach to change detection and information mining in remote sensing. In C. H. Chen, editor, *Frontiers of Remote Sensing Information Processing* WSP, pages 63-86, 2003.

[16] L. Li and P. Revesz. Interpolation Methods for Spatiotemporal Geographic Data. *Journal of Computers, Environment, and Urban Systems*, 28(3):201-227, 2004.

[17] T. B. McKee, N. J. Doesken, and J. Kleist. The Relationship of Drought Frequency and Duration to Time Scales. In *Proc. of the 8th Conference on Applied Climatology*, American Meteorological Society, pages 179-184, 1993.

[18] P. Revesz. *Introduction to Constraint Databases*, Springer, New York, 2002.

[19] P. Revesz and M. Cai, Efficient querying of periodic spatiotemporal objects. *Annals of Mathematics and Artificial Intelligence*, 36(4):437-457, 2002.

[20] P. Revesz and S. Wu. Spatiotemporal reasoning about epidemiological data. *Artificial Intelligence in Medicine*, 2006.

[21] S. Rosenstone. *Forecasting Presidential Elections*. Yale University Press, New Haven, 1983.

[22] D. A. Shepard. A two-dimensional interpolation function for irregularly spaced data. In *Proc. of the 23rd ACM National Conference*, pages 517-524, 1968.

[23] L. Sigelman. Presidential popularity and presidential elections. *Public Opinion Quarterly*, 43:532-534, 1979.

[24] M. Tomczak. Spatial interpolation and its uncertainty using automated anisotropic inverse distance weighting (IDW) - Cross-validation/Jackknife approach. *Journal of Geographic Information and Decision Analysis*, 2(2):18-30, 1998.

[25] N. Wells, S. Goddard, and M. J. Hayes. A self-calibrating palmer drought severity index. *Journal of Climate*, 17(12):2335-2351, 2004.

[26] E. G. Zurflueh. Applications of two-dimensional linear wavelength filtering. *Geophysics*, 32:1015-1035, 1967.