Visualization of Temporal-Oriented Datasets

Jun Gao
University of Nebraska-Lincoln
Lincoln, NE 68588, USA

Peter Revesz
Max Planck Institute for Computer Science
66123 Saarbrucken, Germany

Abstract

Visualization and interpolation naturally go together and strengthen each other. The paper presents an interpolation method as a pre-processing tool to generate the missing data to be visualized. The method is applied in election prediction, a typical temporal-oriented spatiotemporal dataset. The experimental results demonstrate the ability of the methods to accurately predict the presidential election.

1 Introduction

In many applications there is a need to represent numeric data in a form which has more visual impact [9]. Visualization is a powerful way to facilitate data analysis [14]. For example, visualization tools can help to unveil hidden patterns and relationships among variables, present abstract statistical concepts and complicated data structures in a concrete manner [23].

While visualization can be highly effective in the recognition of patterns and trends, pool handling of missing data might lead to misleading data interpretation [5]. There are numerous sources for missing data, such as broken instruments, data-entry errors and data-processing mistakes. Given the intrinsic collection and presentation influenced reasons behind missing data, avoiding missing values is nearly impossible, and the amount of missing data is likely to increase proportionally with the size of the set [5].

Missing data can be estimated by interpolation methods based on the sampled values. Interpolation methods have an increasing presence in advanced scientific databases and are closely related to visualization techniques [15]. Visualization and interpolation strengthen each other. If a good interpolation technique suggests itself naturally, then by applying it first, we can usually get a better visualization. However, in some cases it is hard to find a good interpolation function or such a function would be too complex to compute efficiently. In those cases, the visualization can itself serve as a useful interpolation method, because the human eye can then see patterns that would be too complex to capture mathematically. Occasionally, an interpolation technique may also be detrimental and hide a more naturally emerging patterns. Therefore, one may also try to generate visualizations both with and without the use of a pre-processing interpolation and then see whether the emerging pattern can be clearer observed in one than in the other. If the merging pattern is clearer without the interpolation technique, then that could be an indication that the interpolation technique may not be appropriate to the current data set.

The spatiotemporal interpolation model developed by Gao and Revesz is a general interpolation model for spatiotemporal datasets [7]. It could be used as a basic preprocessing tool to generate the missing data to be visualized. The model works as below. Suppose we need to estimate a missing value in a spatiotemporal dataset. 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 $0 \leq \alpha,\ \beta \leq 1$.

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, from common sense we know that 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 may be significantly different. 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. Since election data is a typical temporal-oriented dataset and what people are most interested in is who will win in the coming election, instead of doing an interpolation we apply the interpolation algorithm in predicting the outcome of 2004 USA presidential election and use it as a pre-processing tool to generate the missing data to be visualized.

The rest of the paper is structured as follows. Section 2 discusses the current presidential election forecasting models. Section 3 describes the spatiotemporal interpolation as the pre-processing tool for visualization in predicting presidential election. Section 4 gives some visualization results. Finally, Section 5 presents some ideas for future work.

2 Presidential election forecasting models

The modern age of election forecasting began in the late 1970s. Among the earliest presidential forecasting models were [6, 21, 20, 10]. Most of these models have been amended, updated and are still used. The core of Fair's model [6] is economic conditions and incumbency. It consists of seven variables, three economic (two measures of per capita GDP growth and one of inflation) and four political (incumbency, terms in office, party, and war). Sigelman's model [21] analyzes the connection between presidential approval ratings and subsequent election results. Rosenstone's model [20] modifies the usual vote by conditions that prevail in a given election such as the economy, war, incumbency, region, and trends over time. Lewis-Bech and Rice's model [10] is a adaptation of Edward Tufte's approval rating and economic performance model to forecast both congressional and presidential elections. Aramowitz [1] amended this model by appending a "time for a change" variable (i.e., a penalty if the president's party has been in office two or more terms) to it. Forecast produced by Aramowitz's model have been consistently accurate. Campbell and Wink [2] built a model using only two indicators, the trial-heat poll and second quarter GDP growth in the year of the elections. This model is noteworthy for its simplicity and accuracy. Chappell [3] developed a model that predicts the election result in each state rather than for a whole country. His methods is based on growth in the national economy, nationwide Gallup Poll results during the campaign, and each states voting record in the previous presidential election. Lewis-Beck and Tien's model [11] is based on economic growth in the first half of the election year, July presidential approval ratings, and a survey indicator of the publics outlook for peace and prosperity. Lichtman [13] devised a systems based on patterns evident in elections since 1860. He identified 13 keys to the presidential election and predicted the winning presidential candidate based on the number of keys favoring each party's candidate. This approach is more analytical and less number-oriented than the other models.

With the exception of Lichtman's, nearly all of the previous discussed models use *multi-variate ordinary least squares regression*, a common statistical method in the social sciences [8]. 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 [8]. Most models are limited by the lack of historical information on the relationship between political and economic fundamentals and elections [8]. Hence we consider if we can turn the direction into the historical election data itself and use it as the basis of spatiotemporal interpolations without a set of variables.

3 Interpolation as a pre-processing tool for visualization

As stated in Section 1 we do the prediction instead of interpolation in the application of presidential election, since people is more interested in knowing who will win in the coming election. Another point is that we need to look back in order to know how well our interpolation method is working. We can not say that we have a good forecasting method for the 2008 presidential elections, because that election is still two years away. We can certainly do a prediction even now, but we would have to wait two more years until we know how good our predictions are. In order to use the interpolation model (i.e., Equation 1) in this specific application, we need to choose the spatial and temporal interpolation methods and decide the relationship between the spatial and temporal weights.

In this study we adopt inverse distance weighting (IDW) as the basic method, modify and improve it to estimate E_s and E_t. We choose IDW because of its ease of use and low computation charge [4]. And furthermore IDW is a popular method used in diverse problems such as predicting of rainfall and temperature and mapping of crop spraying [22]. The main assumption of IDW is that values of locations closer to the unsampled location are more similar to the value to be estimated than values of locations further away. IDW interpolations are of the form:

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

where λ_i is the weight for the individual location, and y_i is the variable observed in the sampled location.

In this particular study for each county y is the vote percentage for John Kerry in the 2004 USA presidential election, N is the number of neighboring counties, and p is chosen to be one for simplicity. The left problem is how to determine d for each county. We apply two versions of calculation of d in both the spatial interpolation and temporal interpolation.

When we use the IDW method to calculate the spatial estimation E_s, we tried to calculate d as both the uniform distance and real distance. In the version of uniform distance the distance between a county and any of its neighbors is one, that is, $d = 1$. In the version of real distance d is calculated by the real distance between the centroids of a county and any of its neighbors. The experimental results show that the differences between the two versions are extremely small in our case. Therefore, we adopt the IDW method with uniform distances.

To calculate the temporal estimation E_t we measure d in terms of time difference instead of spatial difference. In the version of *inverse linear temporal method* the weights are assigned proportional to the inverse of the time difference, while in the version of *inverse exponential temporal method* the weights are assigned to 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.

There is a problem in this prediction case. When we calculate E_s 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.

At this point it is time to determine the relationship between α and β. A natural choice is the step function, where a parameter σ and a threshold θ are needed. If $\sigma < \theta$, then we set $\alpha = 1$ (or $\alpha = 0$) and $\beta = 0$ (or $\beta = 1$). σ is chosen 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. The experimental results demonstrate that this simple and natural solution generate exciting performance. We also tried a little more complicated relationship, linear function which is based on the linear combination of α and β. 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 Visualization

In this section, we show the visualization of interpolation data and give some analysis of the quality of interpolation.

Figures 1-6 illustrate the voting prediction results on the 2004 USA presidential election in the states of California, Florida, and Ohio at the county level. Figures 1, 3, and 5 indicate the results in terms of the differences between the actual vote percentages and the estimated vote percentages using our spatiotemporal interpolation model based on step functions. 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. In Figures 2, 4, and 6 the dashed line shows the actual vote percentage in each county and the solid line describes the estimated vote percentages using our spatiotemporal interpolation model in each county. We can see that in the three states for most counties the discrepancy is low and it almost disappears for some counties.

In Figures 7 and 8 red counties vote for republic candidate and blue counties vote for democratic candidate in Florida for 2004 USA presidential election. Figure 7 is based on the actual results while Figure 8 is our interpolated results. We can see that only two out of 67 counties are different in two figures.

In order to analyze the quality of interpolation we conduct the experiments based on three measures comparing the accuracy of interpolation methods, mean absolute error(MAE), root mean square error (RMSE), and error of statewide total vote percentage (TE), which is a more interesting measure in the voting prediction area. TE is calculated as the difference between the actual statewide vote percentages and the estimated statewide vote percentages.

Table 1 illustrates the quality of prediction at the state level in terms of TE, MAE, and RMSE. 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 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%. 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 also reasonably low.

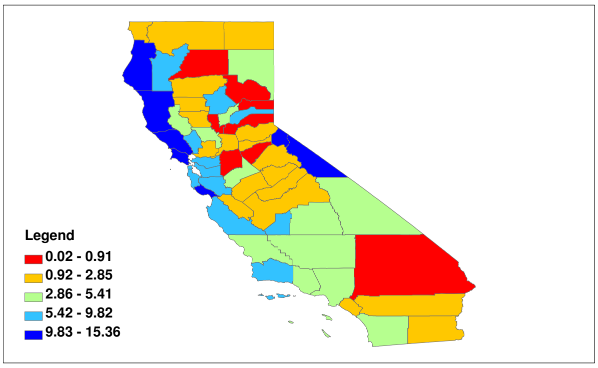

Figure 1. Prediction accuracy in California, USA

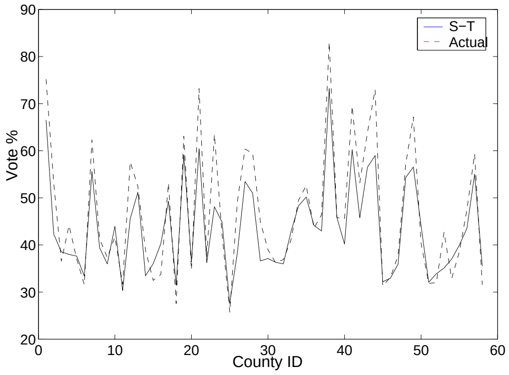

Figure 2. Predicted and actual voting in California

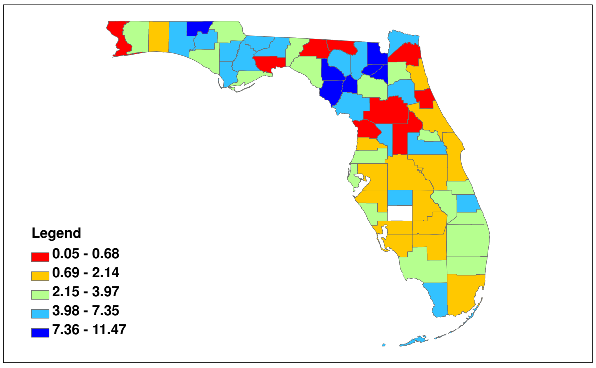

Figure 3. Prediction accuracy in Florida, USA

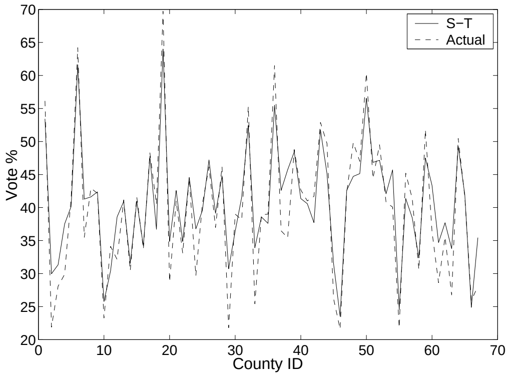

Figure 4. Predicted and actual voting in Florida

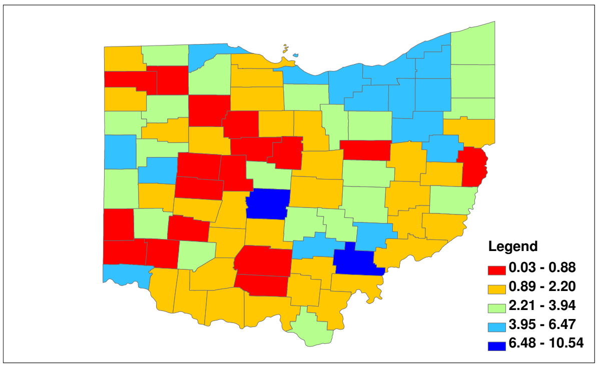

Figure 5. Prediction accuracy in Ohio, USA

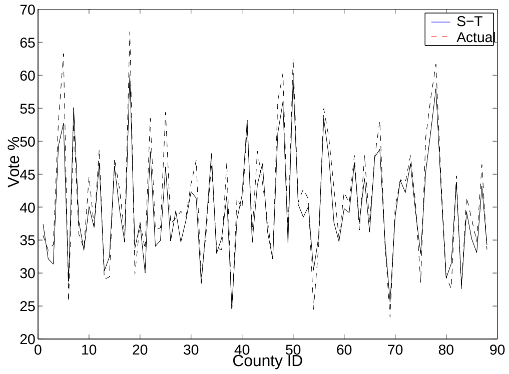

Figure 6. Predicted and actual voting in Ohio

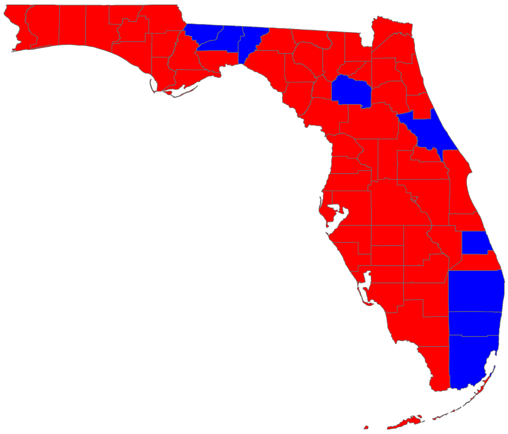

Figure 7. Actual results: red and blue counties in Florida, USA

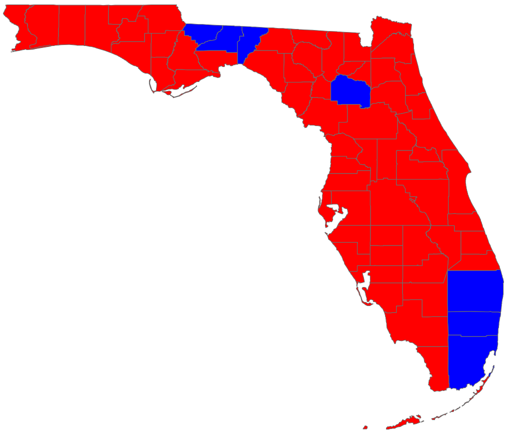

Figure 8. Interpolated results: red and blue counties in Florida, USA

5 Conclusion and Future Work

This paper discusses a pre-processing tool to generate the data to be visualized for the temporal-oriented spatiotemporal datasets. For spatiotemporal applications the querying and visualization become easy in constraint databases [16, 12]. For example, the visualization of recursively defined concepts can not be handled in an easy way by some relational databases and knowledge-based systems, however, it is easy to maintain in constraint databases [18, 19]. Constraint databases integrate database technology with constraint solving methods to visualize complex spatiotemporal problems [17]. We are currently extending the usage of the pre-processing tool in the constraint databases. We are also branching out to other applications that require the interpolation and visualization.

Table 1. TE, MAE, and RMSE results

		Spatial IDW	S-T step function	Temporal inverse linear	Temporal inverse expo
CA	TE	8.65	3.49	5.46	3.46
	MAE	11.60	4.51	6.66	4.48
	RMSE	9.67	6.26	7.25	6.01
FL	TE	4.88	1.09	2.68	1.10
	MAE	7.98	2.40	3.81	2.39
	RMSE	9.05	5.18	5.12	4.59
OH	TE	8.75	3.57	4.10	3.18
	MAE	11.31	4.37	5.09	3.99
	RMSE	7.60	3.57	3.74	3.10

6 Acknowledgement

The first author was supported by an Othmer Fellowship of the University of Nebraska-Lincoln in the USA, and the second author was supported by a Humboldt Research Fellowship of the Alexander von Humboldt Foundation in Germany. Additional funding was also given by the USA National Science Foundation under grant EIA-0091530.

References

[1] A. Abramowitz. Bill and Als Excellent Adventure: Forecasting the 1996 Presidential Election. In J. Campbell and J. Garand, eds, *Before the Vote: Forecasting American National Elections*, pages 47-56, Thousand Oaks, CA: Sage Publications, 2000.

[2] J. Campbell and K. Wink. Trial-heat forecasts of the presidential vote. *American Politics Quarterly*, 18:251-269, 1990.

[3] H. Chappell. Forecasting Presidential Elections in the United States. Entry Prepared for the Encyclopedia of Public Choice, Charels Prowley and Friedrich Schneider, eds., Springer, 2004.

[4] F. Collins and P. Bolstad. A Comparison of Spatial Interpolation Techniques in Temperature Estimation. In *Proc. of the Third International Conference/Workshop on Integrating GIS and Environmental Modelling*, 1996.

[5] C. Eaton, C. Plaisant, and T. Drizd. The challenge of missing and uncertain Data. In *Proc. of IEEE Info-*

Vis Poster Compendium, pages 40-41, IEEE Computer Society Press, 2003.

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

[7] J. Gao and P. Revesz. Adaptive spatio-temporal interpolation methods. In *Proc. of the 8th Joint Conference on Information Sciences, 1st International Conference on Geometric Modeling, Visualization & Graphics*, pages 1622-1625, 2005.

[8] J. Greene. Forecasting Follies. *The American Prospect*, vol 4 no. 15, 1993.

[9] F. Hussain and M. Sarfraz. On Visualisation of Statistical Data. In *Proc. of International Conference on Information Visualization*, pages 343-346, 1997.

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

[11] M. Lewis-Beck and C. Tien. The Future in Forecasting: Prospective Presidential Models. In J. Campbell and J. Garand, eds, *Before the Vote: Forecasting American National Elections*, pages 83-102, Thousand Oaks, CA: Sage Publications, 2000.

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

[13] A. Lichtman. *The Keys to the White House: A Surefire Guide to Predicting the Next President*. Lanham, MD: Madison Books, 1996.

[14] C. Olston and J. D.Mackinlay. Visualizing Data with Bounded Uncertainty. In *Proc. of the IEEE Symposium on Information Visualization*, pages 37-40, 2002.

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

[16] P. Revesz and L. Li. Constraint-Based Visualization of Spatial Interpolation Data. In *Proc. of 6th International Conference on Information Visualization*, pages 563-569, IEEE Press, 2002.

[17] P. Revesz and L. Li. Constraint-Based Visualization of Spatiotemporal Databases. In M. Sarfraz, editor, *Advances in Geometric Modeling*, pages 263-276. John-Wiley Inc., 2003.

[18] P. Revesz and S. Wu. Visualization of Recursively Defined Concepts. In *Proc. of the 8th International Conference on Information Visualization (IV)*, pages 613-621, IEEE Press, 2004.

[19] P. Revesz and S. Wu. Spatiotemporal Reasoning about Epidemiological Data. *Artificial Intelligence in Medicine*, 2006, to appear.

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

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

[22] 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.

[23] C. H. Yu and S. Stockford. Evaluating spatial- and temporal-oriented multi-dimensional visualization techniques. *Practical Assessment, Research & Evaluation*, 8(17), 2003.