ADAPTIVE INTERPOLATION METHODS FOR SPATIOTEMPORAL DATA

by

Jun Gao

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Peter Revesz

Lincoln, Nebraska

December, 2006

ADAPTIVE INTERPOLATION METHODS FOR SPATIOTEMPORAL DATA

Jun Gao, Ph.D.

University of Nebraska, 2006

Advisor: Peter Revesz

Interpolation methods are needed for spatiotemporal data sets to fill in missing data and predict the future. Spatiotemporal interpolation is more challenging than pure spatial or temporal interpolations, and currently there are only a few known spatiotemporal interpolation methods.

We analyze spatiotemporal data sets by introducing spatial and temporal relationship strength measures for them. Based on the relative strengths of the spatial and the temporal relationships in the data sets, we classify them as being spatial-dominated or temporal-dominated.

This analysis of spatiotemporal data sets allows us to introduce a class of adaptive spatiotemporal interpolation methods. An adaptive spatiotemporal interpolation method combines a spatial interpolation method with a temporal interpolation method in such a way that the degree of reliance on the two components is proportional to the measured spatial and temporal relationship strengths. Hence a spatial-dominated spatiotemporal data set would be interpolated more like a spatial data set. Similarly, a temporal-dominated spatiotemporal data set would be interpolated more like a temporal data set. Adaptive spatiotemporal interpolation reduces in a flexible way the spatiotemporal interpolation problem to the problem of pure spatial and temporal interpolation.

Adaptive reduction works in principle for both point-based and region-based spatiotemporal data sets. Although there are many point-based spatial interpolation

methods, there is a lack of region-based spatial interpolation methods. As many spatiotemporal data sets are region-based, as a practical matter, we also propose two region-based variations on the well-known inverse distance weighting interpolation method. The first variation assigns uniform weights between neighbors and the second variation assigns weights proportional to the centroid distances. We also propose a new temporal interpolation method, called the exponential decay temporal interpolation method. Finally, we test the adaptive spatiotemporal interpolation methods on a spatial-dominated climate data set and on a temporal-dominated election data set.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor Professor Peter Revesz. Without his brilliant ideas and guidance I would not have been able to complete this dissertation. He has helped me to expand the breadth and depth of my knowledge and research by providing many insights into my research problems. I am greatly indebted to him.

I would like to thank Professor Ashok Samal, Professor Jun Wang, and Professor Keng Siau for serving on my thesis committee. I sincerely appreciate their help and valuable comments on my dissertation.

I am grateful to the Othmer Fellowship at the University of Nebraska-Lincoln, the Department of Computer Science & Engineering, and NSF EIA-0091530 and EPSCoR programs for their financial support of my courses, research, and travel expenses when presenting this work.

I would like to thank my parents who have provided great support and encouragement throughout my education. Finally, I would like to thank my husband Zhirong for his understanding, love, and support throughout all these year.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Missing data is a problem that is regularly encountered in databases. Given any large dataset, it is likely that there will be missing values scattered in it. Missing data could be a potential threat to the validity of a research study. For example, the Standardized Precipitation Index (SPI) is a common and simple measure of the intensity and duration of drought at certain measured point locations [47, 52]. 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 [64]. As another example, the climate community and many federal agencies use climate data sets to model natural resources which will be used by many agencies to make decisions [16]. However, those climate data sets usually have some missing values.

Interpolation algorithms, which fill in the missing values, are important in many areas, such as Geographic Information Systems (GIS), environmental studies, image processing, and remote sensing. For example, interpolation can offer insights into significant geological structure and behavior, which may not be otherwise apparent [63].

Interpolating animal tracking data solves the problem of uneven sampling [61]. Interpolation is important for geographically distributed statistics for agricultural productions, disease prevalence, pollution levels, soil types, precipitation, and temperatures. Interpolation can also help accounting for topographic effects in spatial climate analysis [26].

Spatiotemporal data, which have both spatial and temporal dimensions [34], are used in many applications. For example, in mobile computing, mobile device users can move in space and register their location at different time instances. In GIS tracking animals and weather conditions will create spatiotemporal data by storing locations of observed objects over time. The interpolation of spatiotemporal data is considerably less well-understood than the interpolation of pure spatial data. This dissertation presents some new ideas on this topic.

1.2 Overview of Contributions of This Dissertation

After a literature review about interpolation methods in Chapter 2, we present the following main contributions.

Spatiotemporal data classification: In Chapter 3, which is partly based on [22], we introduce a new classification of spatiotemporal data. We describe the spatial and temporal *relationship strength measures* that are applicable for any spatiotemporal data. We apply these measures to show that particular spatiotemporal data are classifiable as *spatial-dominated* or *temporal-dominated*. We call a spatiotemporal data spatial-dominated, if the spatial relationship is stronger than the temporal relationship, else we call it temporal-dominated.

An adaptive spatiotemporal interpolation method: In Chapter 4, which is based on [19], we propose a novel adaptive spatiotemporal interpolation method. The adaptive interpolation method can be used for both spatial-dominated and temporal-dominated data. The basic idea behind the adaptive method is to combine a pure spatial interpolation method with a pure temporal interpolation method. The combination is flexible to allow leaning more on the former in case of spatial-dominated data and more on the latter in case of temporal-dominated data. We propose three combination types: *step function-based combination*, *linear function-based combination*, and *variance comparison-based combination*.

Spatial interpolation of region-based spatial data: Although there are many point-based spatial interpolation methods, there is a lack of region-based spatial interpolation methods. As many spatial and spatiotemporal data are region-based, we need to modify the point-based spatial interpolation methods to be applicable to region-based data. In Chapter 5, which is based on [20], we propose two region-based variations on the well-known inverse distance weighting interpolation method: *IDW with uniform weights (IDWU)* and *IDW with centroid distance weights (IDWC)*.

Exponential decay temporal interpolation: In Chapter 6, which is based on [21], we propose the *exponential decay temporal interpolation method*, which we compare with an IDW-based temporal interpolation method. A special case of IDW-based interpolation is the *inverse linear temporal method*.

Climate data interpolation: In Chapter 7, which is based on [18], we apply our new spatiotemporal interpolation methods to climate data. Climate data is spatial-dominated and point-based. The experimental results show that our adaptive method is better than several pure spatial interpolation or pure tem-

poral interpolation methods.

Election prediction: In Chapter 8, which is partly based on [20, 22], we apply our new spatiotemporal interpolation methods to the USA presidential election data. Election data is temporal-dominated and region-based. The experimental results show that our adaptive method is better than several pure spatial interpolation or pure temporal interpolation methods.

After these main contributions, we also give in Chapter 9 conclusions and suggestions for future work.

Chapter 2

Literature Review

In Section 2.1 we describe a classification of point-based spatiotemporal data and region-based spatiotemporal data. In Section 2.2 we introduce spatial interpolation methods for point-based spatiotemporal data. In particular we give in Section 2.2.1 a detailed introduction to the inverse distance weighting (IDW) interpolation method which we use in our research. In Section 2.3 we describe several spatiotemporal interpolation methods. In Section 2.4 we introduce prediction as a special case of interpolation and describe several commonly used presidential election forecasting models.

2.1 Point-Based and Region-Based Spatiotemporal Data

In [41] Li and Revesz categorize the spatiotemporal databases into three groups according to whether the data representation is based on points, regions, or constraints. Point-based spatiotemporal relations have the scheme of $(x, y, t, w_1, w_2, \ldots, w_m)$, where (x, y) is the point location and t is the time instance, w_i $(1 \leq i \leq m)$ indicate the

features at point (x, y) and time t. Region-based spatiotemporal databases have both spatial and temporal parts. The spatial part has the scheme of (*RegionID, boundary*), where *RegionID* is a unique identifier for each polygonal-shaped region and *boundary* is a sequence of corner vertices. The temporal part has the scheme of $(RegionID, t, w_1, w_2, \ldots, w_m)$, where t is the time instance and w_i $(1 \leq i \leq m)$ record the features in that region at time t.

We conduct the experiments on two different data sets, *climate data* and *USA presidential election data*. The climate data records the climate observations in weather stations in Nebraska and Colorado. The general form is $(StationID,$ *latitude, longitude*$, t, w_1, w_2, \ldots, w_m)$, where *StationID* is the unique identifier of a weather station, (*latitude, longitude*) is its location, t specifies the time instance, and w_i $(1 \leq i \leq m)$ are fields that record the values of minimum temperature, maximum temperature, amount of precipitation and etc. The USA presidential election dataset records the presidential election results in different states. In each state the voting results are recorded by counties. The general form is $(CountyName, t, w_1, w_2, \ldots, w_m)$, where t specifies the election year, and w_i $(1 \leq i \leq m)$ represent the total votes and the votes for different candidates. According to [41] the climate data can be identified as a point-based spatiotemporal data and the USA presidential election data as a region-based spatiotemporal data.

2.2 Spatial Interpolation for Point-Based Data

There are several characteristics of spatial interpolation methods, including point-based versus region-based, global versus local, exact versus approximate, stochastic versus deterministic, and gradual versus abrupt [15]. Point-based interpolation methods estimate values at specific points in space based on the values and locations of

other sample points in space. For example, a point-based interpolation method could predict carbon monoxide concentration at a specific latitude and longitude based on the measurements from a monitoring network. Region-based interpolation methods estimate values for entire areas based on data available for a different set of areas. For example, estimate the population in a county based on the population in its neighboring counties.

2.2.1 Inverse Distance Weighting

Inverse distance weighting (IDW) [55] is an example of a point-based, local, exact, deterministic, and gradual interpolator in which points closer to the measured data points receive more weight in the averaging formula. Since IDW is a point-based interpolator, for point-based spatiotemporal data like climate data, we can just use the standard IDW as the spatial interpolator to do the spatial estimation part.

IDW has been used to interpolate spatial data by many authors, for example, by Legates and Willmont [35] and Stallings et al. [59]. The main assumption of IDW is that if A, B and C are three different locations, such that A is closer to B than to C, then the value we are interested in (temperature, precipitation, percentage of voters preferring a particular candidate, etc.) is also closer between A and B than between A and C. Hence, if the value at location A is unknown, while the values at locations B and C are known, then the value at B should be more important than the value at C in estimating the value at A.

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 some power of the distance* between the known and unknown locations.

IDW interpolations are of the form [29]:

$$w(x,y) = \sum_{i=1}^{N} \lambda_i \cdot w_i \tag{2.1}$$

$$\lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})^p} \tag{2.2}$$

where $w(x,y)$ is the interpolated value at point (x,y), λ_i is the weight for the individual sampled neighbor of point (x,y), w_i is the variable observed in the sampled neighbor, N is the number of closest sampled neighbors, d_i is the Euclidean distance between each (x_i, y_i) and (x,y), and p is the exponent, which influences the weighting of w_i on w.

For simplicity in the following we assume that $p = 1$. Therefore,

$$\lambda_i = \frac{\frac{1}{d_i}}{\sum_{k=1}^{N} \frac{1}{d_k}} \tag{2.3}$$

Example 2.2.1 In Figure 2.1, point (x,y) is the target point where the value E_s of an attribute at time t needs to be interpolated. The target point has five spatial neighbors, $(x_1, y_1), \ldots, (x_5, y_5)$, and their values of that attribute at time t are known as $w_1, \ldots, w_5$.

In order to follow the IDW interpolation, we first calculate the distance d_i between the target point (x,y) and its ith neighbor (x_i, y_i) as follows.

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2}$$

We then calculate the λ_i for the ith neighbor using Equation 2.3 as follows.

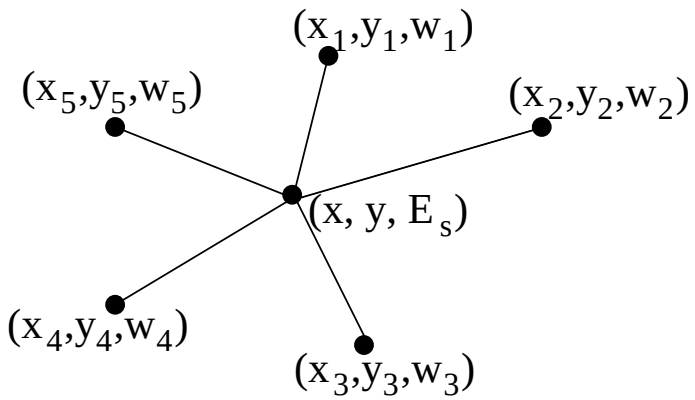

Figure 2.1: Target point and its spatial neighbors

$$\lambda_i = \frac{\frac{1}{d_i}}{\sum_{k=1}^{5} \frac{1}{d_k}}$$

Now we can calculate E_s using Equation 2.1 as follows.

$$E_s = \sum_{i=1}^{5} \lambda_i \cdot w_i$$

Example 2.2.2 Assume that $A = (5, 0)$, $B = (0, 0)$, and $C = (20, 0)$ and the value at A is unknown but the values at B and C are 100 and 200, respectively. Then, the number of known points is $N = 2$. We use the subscripts B and C instead of numbers in this simple example. We can calculate that:

$$\lambda_B = \frac{\frac{1}{5}}{\frac{1}{5} + \frac{1}{15}} = \frac{3}{4}$$

$$\lambda_C = \frac{\frac{1}{15}}{\frac{1}{5} + \frac{1}{15}} = \frac{1}{4}$$

Hence the value of A will be interpolated based on B and C to be:

$$\begin{aligned} w_A &= \lambda_B \times w_B + \lambda_C \times w_C \\ &= \frac{3}{4} \times 100 + \frac{1}{4} \times 200 \\ &= 125 \end{aligned}$$

Note that since point C is three times more distant than B is from point A, the weight λ_C is only a third of the weight λ_B. Hence w_A is much closer to w_B than to w_C.

2.2.2 Other Spatial Interpolation Methods

Other spatial interpolation methods for point-based data are kriging [13], regression model [4], shape functions [43], splines [24], and trend surface analysis [67]. A relatively new type of interpolation methods employ artificial neural networks (ANN) [58, 45, 39, 57].

There is an extensive literature on the comparison of the various interpolation techniques. Early reviews of interpolation techniques include [33, 53]. Burrough and McDonnell [4] provide a solid introduction and detailed overview to different types of interpolation, and discuss how the assumptions influence the final result.

2.3 Spatiotemporal Interpolation Methods

Spatiotemporal interpolations are frequently found in applications such as digital image processing and human vision [62, 46, 12]. In those applications spatiotemporal

interpolations are used to reconstruct the images from spaced samples or estimate the motion of moving objects.

Spatiotemporal interpolation in GIS started from the incorporation of three-dimensional data structures into existing GIS and the development of spatiotemporal representations, i.e., four-dimensional representations [3]. Miller describes an approach to interpolation in four dimensions using Kriging in [48].

Alternative spatiotemporal interpolation methods are given in [40, 43, 52]. Revesz and Wu [52] give a general method to model a class of recursively defined spatiotemporal concepts and apply the method to represent the epidemiological definitions and predictions about the spread of infectious diseases. In [43] Li and Revesz design a spatiotemporal method which reduces the spatiotemporal interpolation problem to a regular spatial interpolation case. The method based on IDW works in two steps. First, it interpolates using a piecewise linear function the measured value over time at each sample point. Second, it substitutes the desired time instant into the regular IDW interpolation functions. For example, the first step works as follows.

Assume the value at the location i at time t_{i1} is w_{i1}, and at time t_{i2} the value is w_{i2}. The value at the location i at any time between t_{i1} and t_{i2} can be approximated using a piecewise linear function in the following way:

$$w_i(t) = \frac{t_{i2} - t}{t_{i2} - t_{i1}} w_{i1} + \frac{t - t_{i1}}{t_{i2} - t_{i1}} w_{i2}$$

Then the value of the unsampled point at location (x, y) at time t can be interpolated as follows.

$$w(x, y, t) = \sum_{i=1}^{N} \lambda_i w_i(t) \quad , \qquad \lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})^p}$$

Chomicki et al. [7, 9] 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 [8].

2.4 Prediction as a Special Case of Interpolation

Usually interpolation is used to estimate the data which are missing or unknown in an arbitrary time. Prediction is used to predict the unknown value in the future. Therefore, generally speaking, prediction is a special case of interpolation.

Election prediction is an interesting issue in the prediction area, and the presidential election prediction may be the most difficult and exciting among them. We pick it as a special example showing how to use interpolation methods to do the prediction.

The modern age of election forecasting began in the late 1970s. Among the earliest presidential forecasting models were [17, 56, 54, 37]. Most of these models have been amended, updated and are still used. The core of Fair's model [17] 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 [56] analyzes the connection between presidential approval ratings and subsequent election results. Rosenstone's model [54] 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 [37] 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 [5] 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 [6] 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 [38] 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 [44] 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. Table 2.1 gives a simple summary of several above 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 [25]. 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 [25]. Most models are limited by the lack of historical information on the relationship between political and economic fundamentals and elections [25]. 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.

Table 2.1: Presidential forecasting models

Model author(s)	Indicators	Authors' area
Aramowitz	economic growth in the first half of the election year, presidential approval rating in June, term	political scientist
Campbell and Wink	trial-heat poll and second quarter GDP growth in the election year	political scientist
Chappell	growth in the national economy, nationwide Gallup Poll results during the campaign, and each states voting record in the previous presidential election	economist
Lewis-Beck and Tien	economic growth in the first half of the election year, presidential approval rating in July, and survey on peace-prosperity	political scientist
Lichtman	13 keys: party mandate, contest, incumbency, third, short-term economy, long-term economy, policy change, social unrest, scandal, foreign/military failure, foreign/military success, incumbent charisma, challenger charisma	historian
Fair	two measures of per capita GDP growth and one of inflation, incumbency, terms in office, party, and war	economist
Rosenstone	party, key issues, the economy, war, incumbency, region, and trends over time	political scientist

Chapter 3

Classification of Spatiotemporal Data Sets

We introduce a new classification of spatiotemporal data sets based on the spatial and temporal relationship strengths among them. Section 3.1 describes the spatial and temporal relationship strength measures that are applicable for any spatiotemporal data. Section 3.2 gives the definition about how to classify the spatiotemporal data as being spatial-dominated or temporal-dominated.

3.1 Relationship Strength Measures

Spatiotemporal data contains information in both space and time. We can measure the spatial and temporal relationship strength for them. For a spatiotemporal data set, the spatial relationship strength can be quantified as a parameter, denoted as S_σ; and the temporal relationship strength can be quantified as a parameter, denoted as T_σ. Although the calculations of S_σ and T_σ largely depend on the particular data set, we can still give some general ideas about how to calculate these two parameters.

Table 3.1: Temperatures on same day for ten consecutive years of weather stations in Nebraska

	Station 1 Sep 8 1986-1995	Station 2 Mar 7 1989-1998	Station 3 Apr 1 1993-2002	Station 4 Jun 6 1994-2003	Station 5 Feb 9 1994-2003	T_σ
	38	12	27	58	13	
	46	31	16	58	21	
	51	12	16	55	31	
	60	31	22	55	10	
	50	31	31	43	23	
	67	23	22	52	28	
	37	10	27	50	28	
	42	9	19	53	1	
	55	14	17	59	24	
	38	21	28	47	10	
Mean	48.4	19.4	22.5	53	18.9	
Std Deviation	10.15	9.16	5.48	5.16	9.85	7.96

From a spatiotemporal data set, we take a set of values which vary in space but not in time. For example, the set $(p_1, p_2, \ldots, p_M)$ indicates the values taken at the neighboring points at the same time, then this set may show the relations in space among the values of the neighboring points. Similarly, assume we have a set of values which vary in time but not in space. For example, the set $(q_1, q_2, \ldots, q_N)$ indicates the values taken on the same day in different years at the same location, then this set of values may show the relations among values in time. If a data set is normally distributed, then the standard deviations of $(p_1, p_2, \ldots, p_M)$ and $(q_1, q_2, \ldots, q_N)$ can be used to measure S_σ and T_σ, respectively. That is, S_σ can be calculated as follows:

Table 3.2: Temperatures in neighboring weather stations in Nebraska

	Station 1 9/8/1991	Station 2 3/7/1994	Station 3 4/1/1998	Station 4 6/6/1999	Station 5 2/9/2002	S_σ
Self	67	23	22	52	24	
Neighbor 1	68	22	27	49	22	
Neighbor 2	65	23	26	48	26	
Neighbor 3	63	25	21	55	26	
Neighbor 4	69	25	25	58	21	
Neighbor 5	71	25	25	53	28	
Mean	67.17	23.83	24.33	52.5	24.5	
Std Deviation	2.86	1.33	2.34	3.73	2.66	2.58

$$S_\sigma = \sqrt{\frac{\sum_{k=1}^{M}(p_i - \overline{p_i})^2}{M}} \tag{3.1}$$

$$\overline{p_i} = \frac{1}{M}\sum_{k=1}^{M} p_i \tag{3.2}$$

Similarly, T_σ can be calculated as follows:

$$T_\sigma = \sqrt{\frac{\sum_{k=1}^{N}(q_i - \overline{q_i})^2}{N}} \tag{3.3}$$

$$\overline{q_i} = \frac{1}{N} \sum_{k=1}^{N} q_i \tag{3.4}$$

Let us look at an example showing how to calculate T_σ and S_σ for a spatiotemporal data set.

Example 3.1.1 Suppose we have a climate data set recording the daily minimum temperatures in 50 years for weather stations in Nebraska. Tables 3.1 and 3.2 demonstrate how to calculate T_σ and S_σ for this climate data. The five sets of values in Table 3.1 are the daily minimum temperatures on the same day for ten years at five weather stations. The stations, days and years are selected randomly. The set for station 1 records the minimum temperature on September 8th from 1986 to 1995, station 2 March 7th from 1989 to 1998, station 3 April 1st from 1993 to 2002, station 4 June 6th from 1994 to 2003, and station 5 February 9th from 1994 to 2003. We calculate the standard deviation for each set of values and choose the mean of the five standard deviation values as T_σ.

Table 3.2 records the temperatures in neighboring stations for the same five target stations in Table 3.1. We only choose the five closest neighbors for each target station. The days for each station are station 1 September 8th in 1991, station 2 March 7th in 1994, station 3 April 1st in 1998, station 4 June 6th in 1999, and station 5 February 9th in 2002, respectively. We calculate the standard deviation for each set of values and choose the mean of the five standard deviation values as S_σ. We can see that for this case of climate data, S_σ (2.58) is much smaller than T_σ (7.96).

3.2 Spatial-Dominated VS Temporal-Dominated

Now we introduce a new classification of spatiotemporal data set based on the relative strengths of the spatial and the temporal relationships in the data set.

Definition 3.2.1 *For a spatiotemporal data set, if $S_\sigma < T_\sigma$, then we call it spatial-dominated, else temporal-dominated.*

Consider again the climate data set in Example 3.1.1. Since S_σ (2.58) is much smaller than T_σ (7.96), according to Definition 3.2.1 the climate data is spatial-dominated.

In the spatial-dominated data set the spatial relationship between the data values is stronger than the temporal relationship. For example, in a climate data set, 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 a data set 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 [4].

In the temporal-dominated data set the temporal relationship between the data values is stronger than the spatial relationship. For example, in the United States people who vote for Democratic presidential candidate will more likely vote for Democratic candidate again in the next election. Hence, in the USA presidential election data set, the outcomes in one state may remain the same for many years, while the outcomes of two neighboring states may be significantly different. Another example is an air ticket data set. The air ticket price is more likely higher in a peak season year after year.

Now let us look at a specific example of election data showing why the election data set is temporal-dominated.

Example 3.2.1 Suppose we have a USA presidential elections data set recording the vote percentages for democratic candidates in the past 50 years for all counties in Florida. Tables 3.3 and 3.4 show how to calculate T_σ and S_σ for this election data

Table 3.3: Vote percentages for democratic candidate in six USA presidential elections in Florida

	Alachua	Clay	Jefferson	Marion	St. Johns	T_σ
1980	52.27	31.63	57.16	37.94	36.64	
1984	46.43	20.28	47.82	30.02	28.73	
1988	48.82	22.97	46.73	33.09	29.29	
1992	49.61	23.33	48.55	35.44	30.74	
1996	53.9	28.16	52.9	41.08	34.43	
2000	55.25	25.48	53.89	40.39	32.1	
Mean	50.21	25.27	52.18	36.33	31.99	
Std Deviation	3.34	4.07	4.1	4.31	3.07	3.78

set. The five counties are chosen randomly. Table 3.3 records the vote percentages for democratic candidate in six USA presidential elections for five counties in Florida. The five sets of values in Table 3.3 are the vote percentages for democratic candidates in six consecutive elections starting from 1984, that is, elections in 1984, 1988, 1992, 1996, and 2000, for the five counties, Alachua, Clay, Jefferson, Marion, and St. Johns, respectively. We calculate the standard deviation for each set of values and choose the mean of the five standard deviation values as T_σ.

Table 3.4 records the vote percentages for democratic candidates in neighboring counties of the above chosen five counties. As examples, the column of Alachua contains the election votes in 1984 in its neighboring counties, Clay in 1988, Jefferson in 1992, Marion in 1996, and St. Johns in 2000. We calculate the standard deviation for each set of values and choose the mean of the five standard deviation values as

Table 3.4: Vote percentages for democratic candidate in USA presidential elections in neighboring counties in Florida

	Alachua (1984)	Clay (1988)	Jefferson (1992)	Marion (1996)	St. Johns (2000)	S_σ
Self	46.43	22.97	48.55	41.08	32.1	
Neighbor 1	36.18	48.82	49.12	53.9	25.48	
Neighbor 2	20.28	28.34	45.36	44.44	40.74	
Neighbor 3	32.59	35.96	35.6	40.25	51.25	
Neighbor 4	33.82	36.72	34.55	44.63	46.14	
Neighbor 5	35.81	42.23		47.75		
Neighbor 6	30.02	29.29		45.56		
Neighbor 7	40.61			49.28		
Neighbor 8	29.65					
Mean	33.93	34.9	42.64	45.86	39.14	
Std Deviation	7.34	8.84	7.06	4.44	10.42	7.62

S_σ. We can see that for this case of election data, T_σ (3.78) is much smaller than S_σ (7.62). Hence we think this is a temporal-dominated data set.

Chapter 4

Adaptive Spatiotemporal Interpolation

We describe a new adaptive spatiotemporal interpolation method which is a combination of spatial interpolation and temporal interpolation. Section 4.1 describes the adaptive method and its features. Section 4.2 discusses the general idea and issues in applying the adaptive method. Section 4.3 introduces step function-based, linear function-based and variance comparison-based combination relationships to solve the combination issue mentioned in Section 4.2.

4.1 Adaptive Method of Spatiotemporal Interpolation

The adaptive interpolation method developed by Gao and Revesz [19] is a general method to interpolate spatiotemporal data. The basic idea behind the adaptive method is to combine a pure spatial interpolation method with a pure temporal interpolation method. The combination is flexible to allow leaning more on the former in

case of spatial-dominated data and more on the latter in case of temporal-dominated data. Let E_s be the interpolated value using the spatial interpolation method, E_t the interpolated value using the temporal interpolation method, α the weight of E_s, and β the weight of E_t. Then the overall interpolated value E can be calculated as follows:

$$E = \alpha \times E_s + \beta \times E_t \tag{4.1}$$

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

The adaptive method offers the following features:

1. It captures the essence of interpolation of spatiotemporal data at a high level. It could be used as a starting point to develop a formal interpolation methodology for a spatiotemporal data set.

2. It is *Open* and *flexible*, which means the method is domain and application independent. We can explain it as follows. From the point view of interpolation methods, we can choose any spatial interpolation method or temporal interpolation method appropriate for a specific application. From the point view of combination, we can define a particular relationship between the spatial estimate and temporal estimate for an application.

3. A main strength of this method is its ability to borrow information across both space and time. Borrowing information across both space and time is certainly not a new idea. In climate prediction and many other fields, people have borrowed information across time, usually from the past, through dynamical models. However, our method can be viewed as a generalization from spatial or temporal interpolation methods to a method interpolating in both spatial and

temporal directions.

One thing to note about is that our general method is not strictly *"general"*. We have pre-selected the linear combination relationships among final estimate, spatial estimate and temporal estimate, and α and β. Linear relationships are often the first approximation used to describe any relationship. When there is very little information to determine what the relationship is, assuming a linear relationship is simplest and thus is a reasonable starting point. For example, round-trip delay time (RTT) is significant in systems that require two-way interactive communication. Transmission Control Protocol (TCP) implementations attempt to predict future round-trip times using a linear function as below:

$$t_{RTT} = \beta \cdot t_{RTTsample} + (1 - \beta) \cdot t_{RTT}$$

where β is a constant between 0 and 1 that controls how rapidly the t_{RTT} adapts to changes.

It should be noticed that how to determine the weights of spatial and temporal estimates is closely related to the specific application. In order to find the right relationship we need to develop computation models based on history data. For example, in the application of climate data interpolation, we choose the standard deviation of elevations of weather stations, spatial estimates and temporal estimates to build the computation model, calculate and verify the interpolation result with history data.

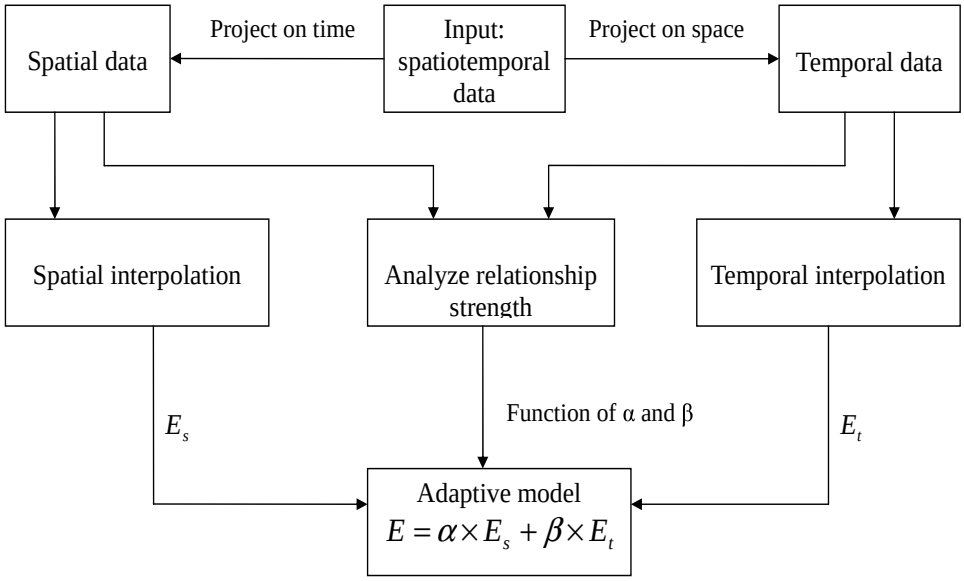

Figure 4.1: Flow chart describing general idea of applying adaptive methods

4.2 Applying the Adaptive Method

4.2.1 General Idea of Applying the Adaptive Method

Figure 4.1 illustrates a flow chart which describes a general idea of applying our adaptive spatiotemporal interpolation method. Spatiotemporal data inputs usually contain both spatial information and temporal information. For example, data from surrounding climate weather stations are spatial information, and local historical weather records are temporal information. Therefore, we project the data into the time dimension and get the spatial data, and project into space dimension and get the temporal data. We then choose a type of spatial interpolation method to do the estimation on the spatial data input. Similarly, we choose a type of temporal interpolation method to do the estimation on the temporal data input. We also need to analyze the relationship strengths and determine the values of alpha and beta. After that we can calculate the final estimate according to the adaptive model.

4.2.2 Issues in Applying the Adaptive Method

In order to follow the flow chart described in Section 4.2.1 to apply the adaptive spatiotemporal interpolation method to a specific application, we need to answer the following three questions:

(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. We adopt IDW as the basic method, modify and improve it into appropriate spatial interpolation and temporal interpolation methods for calculating 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 [60]. Second, IDW is easy to use and has low computation charge [10]. 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 [60]. 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 [2]. We have given a detailed introduction of IDW in Section 2.2.1.

To solve the third question we propose two kinds of combination relationships: *step function-based*, *linear function-based*, and *combination variance comparison-based combination* which are described in Section 4.3.

4.3 Combination Types

4.3.1 Step Function-Based Combination

Considering the relationship between α and β, a natural combination function is a step function as shown in Figure 4.2. First let us explain why step function-based combination is a natural choice. Let $I_{i,t}^M$ be the interpolated value for some location i at time t using method M and $O_{i,t}$ be the observed value for the location i at time t. In particular, we have $I_{i,t}^S$ for $I_{i,t}^M$ using the spatial interpolation method and $I_{i,t}^T$ for $I_{i,t}^M$ using the temporal interpolation method.

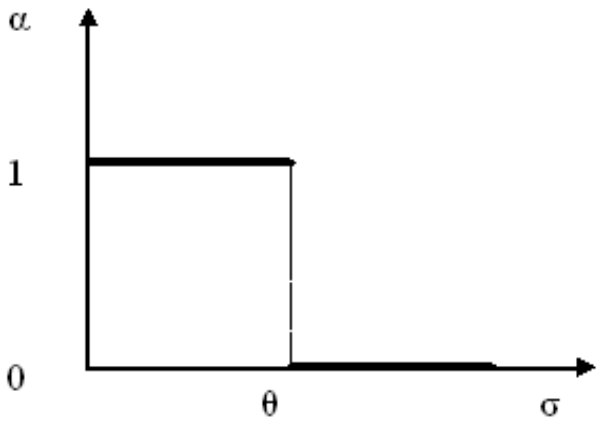

Figure 4.2: Step function-based combination

Let $D_{i,t}^M$ be the absolute difference between the interpolated value using method M and the observed value at location i at time t, that is,

$$D_{i,t}^M = |I_{i,t}^M - O_{i,t}| \tag{4.2}$$

In a step function, we find some parameter σ_i and fix some threshold value θ. If for most locations with $\sigma_i < \theta$, we have $D_{i,t}^S < D_{i,t}^T$ and for most locations with

$\sigma_i \geq \theta$, we have $D^S_{i,t} \geq D^T_{i,t}$, then it means for locations with $\sigma_i < \theta$ the spatial interpolation method is more accurate and for locations with $\sigma_i \geq \theta$ the temporal interpolate method is more accurate. Therefore, for locations with $\sigma_i < \theta$, we set $\alpha = 1$ and $\beta = 0$, which enforces that we use the pure spatial interpolation method; and for locations with $\sigma_i \geq \theta$, we set $\alpha = 0$ and $\beta = 1$, which enforces that we use the pure temporal interpolation method. In summary,

$$\begin{cases} \alpha = 1,\ \beta = 0 & \text{if } \sigma_i < \theta \\ \alpha = 0,\ \beta = 1 & \text{if } \sigma_i \geq \theta \end{cases} \tag{4.3}$$

The selection of the parameters σ_i and θ in a step function is according to the specific application.

The data set shown in Figure 4.3 gives an example showing how step function works. It is an ideal situation, because for all locations on the left side of the vertical dashed line $\sigma_i = 6$, we have $D^S_{i,t} < D^T_{i,t}$ and for all locations on the right side of the vertical dashed line $\sigma_i = 6$, we have $D^S_{i,t} \geq D^T_{i,t}$.

4.3.2 Linear Function-Based Combination

Another simple and natural combination is a linear function-based combination. For example, we have a data set shown in Figure 4.4. In this data set, when $\sigma_i < 11$, $D^S_{i,t}$ increases with the increase of σ_i; when $\sigma_i \geq 11$ the increasing trend continues, however, the degree becomes less. It means that the accuracy of spatial interpolation is decreasing with the increase of σ_i, and after some threshold θ (i.e., $\theta = 11$ in Figure 4.4), the deceasing degree becomes less.

We can simplify this situation into a linear function shown in Figure 4.5. When $\sigma_i < \theta$ the weight of spatial interpolated values varies inversely with σ_i, else keep the weight some constant r. In summary,

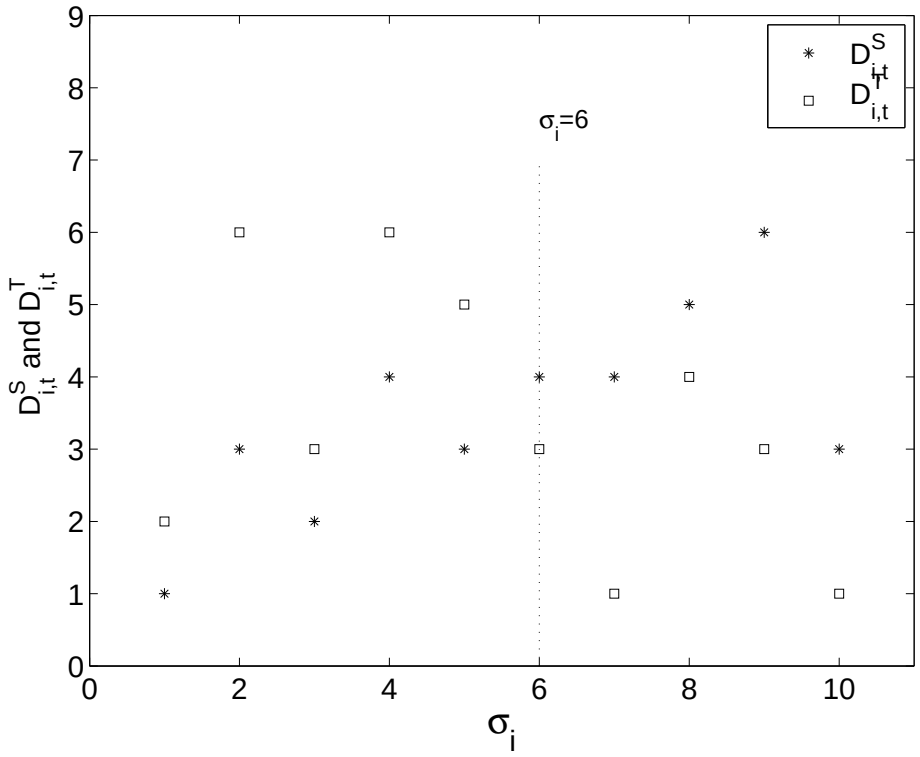

Figure 4.3: Each location appears twice as a star and a box within the vertical line $\sigma_i = \theta$. We see that for $\sigma_i < 6$, the stars are lower than the boxes, and for $\sigma_i \geq 6$, the boxes are lower than the stars.

$$\alpha = \begin{cases} 1 - \sigma \times \frac{1-r}{\theta} & \text{if } \sigma_i < \theta \\ r & \text{if } \sigma_i \geq \theta \end{cases} \tag{4.4}$$

where r is a rate constant between 0 and 1.

Similarly to the step function-based combination, the parameters σ_i, θ, and r are selected according to the specific application.

4.3.3 Variance Comparison-Based Combination

The third type of combination is called variance comparison-based combination. In this combination, we compare the two relationship strength measures. If spatial re-

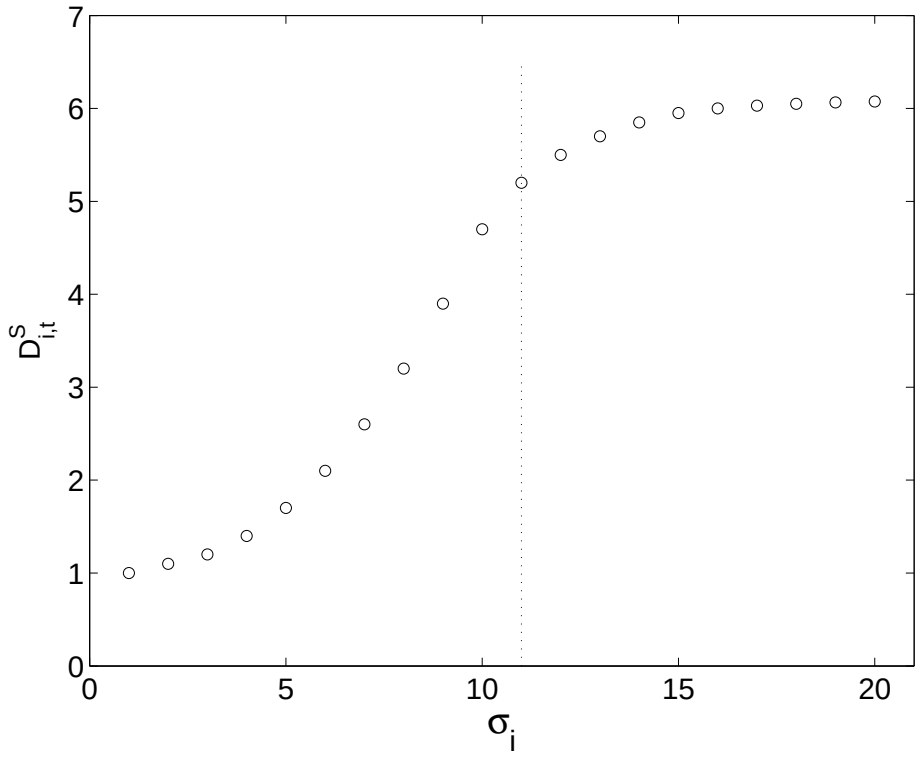

Figure 4.4: Variation of $D^S_{i,t}$ with σ_i

lationship is stronger than temporal relationship, we use spatial interpolation only; if temporal relationship is stronger than spatial interpolation, we use temporal interpolation only. In summary,

$$\begin{cases} \alpha = 1,\ \beta = 0 & \text{if } S_\sigma < T_\sigma \\ \alpha = 0,\ \beta = 1 & \text{if } S_\sigma \geq T_\sigma \end{cases} \tag{4.5}$$

One advantage of the variance comparison-based combination is that it is easier to use and there is no need for additional parameters, for example, the parameter θ in the step function-based combination.

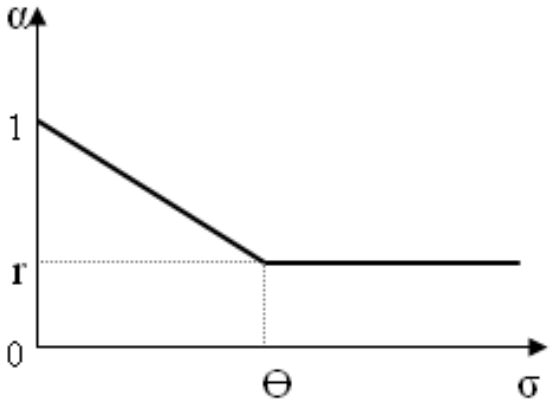

Figure 4.5: Linear function-based combination

4.3.4 Methods to Determine α and β

Now we give a method to determine α and β which is applicable to all spatiotemporal data sets.

Give a spatiotemporal data set, we first apply the measures described in Section 3.1 to find the spatial and temporal relationship strengths.

Based on the spatial and temporal relationship strengths, we classify the data set to spatial-dominated or temporal-dominated.

We choose a spatial interpolation method and a temporal interpolation method for this data set. Then we can calculate the $D^S_{i,t}$ or $D^T_{i,t}$ as described before.

We decide how to choose the parameters σ_i and constant threshold θ. For a particular data set, we may need other parameters.

Finally, we analyze the relations between $D^S_{i,t}$ (or $D^T_{i,t}$) and σ_i. We then can decide whether step function-based or linear function-based combination is more appropriate for this spatiotemporal data set.

Chapter 5

Interpolation of Region-Based Spatial Data

Although there are many point-based spatial interpolation methods, there is a lack of region-based spatial interpolation methods. As many spatial and spatiotemporal data are region-based, we need to modify the point-based spatial interpolation methods to be applicable to region-based data. In Section 5.1 and Section 5.2 we propose two region-based variations on the well-known IDW interpolation method: *IDW with uniform weights* (IDWU) and *IDW with centroid distance weights* (IDWC), respectively.

To spatially interpolate the region-based data, we find the *centroid* of the region. In geometry, the centroid of an object X in n-dimensional space is the intersection of all hyperplanes that divide X into two parts of equal moment about the hyperplane. Informally, it is the "average" of all points of X. After finding the centroid of a region, we assign total value of data in that region to the centroid, and treat it as point interpolation. For example, in the USA presidential election data set, the data of votes are counted in county level. That is, the data are region-based (or

Table 5.1: Latitude and longitude of centroid of 67 counties of Florida, USA

County name	Latitude	Longitude
Alachua	29.676436	-82.379953
Baker	30.287517	-82.236268
Bay	30.219170	-85.638788
...		
Wakulla	30.144620	-84.366174
Walton	30.637995	-86.155962
Washington	30.630591	-85.638396

county-based). In order to interpolate such data, we could find the centroid of each county and assign to its centroid the total votes or the votes for a party candidate for that county. Table 5.1 lists the centroid of 67 counties in Florida, which is obtained from the official web site *www.census.gov*. By using the latitude and longitude of its centroid as the simulator of locations for each county, we could treat this region-based data as point-based data.

In order to apply IDW in this region-based data interpolation, one problem arises. How can we calculate the distance between two regions? One natural solution is to calculate the distance between the centroid of two counties. However, we realize that since all surrounding counties are neighbors of a target county, the distances between neighbors should be equal in this sense. Hence it is also reasonable to use uniform distances in this case. Therefore, to use IDW to spatially interpolate region-based spatiotemporal data, we propose two variations of the standard IDW, that is, *IDW with uniform weights (IDWU)* and *IDW with centroid distance weights (IDWC)*.

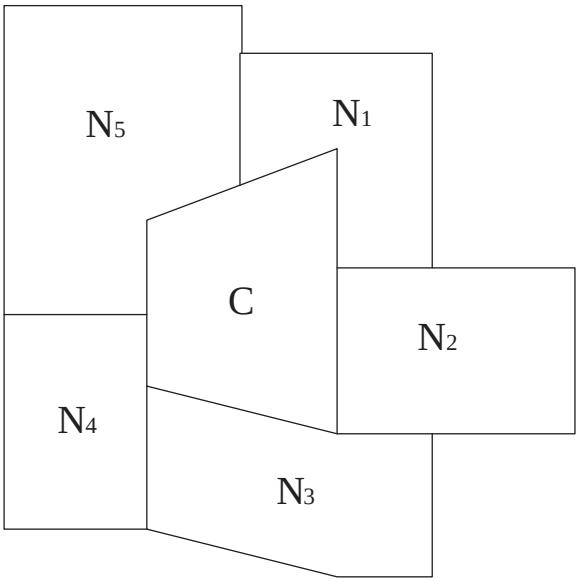

Figure 5.1: County C and its neighbors

5.1 IDW with Uniform Weights

In this variation of IDW, all the distances between the target point (centroid of target region) and one of its neighboring points (centroid of its neighboring regions) are the same. 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 by Equation 2.3 each neighbor N_i has exactly the same weight $\lambda_i = \frac{1}{k}$, $1 \leq i \leq k$. For example, in Figure 5.1, county C has five neighbors and each neighbor has the same weight $\frac{1}{5}$.

Example 5.1.1 Assume we need to interpolate the value in Region R. Let $y_1 = 100$, $y_2 = 200$, $y_3 = 300$, and $y_4 = 400$ be the values measured at the surrounding regions of Region R, that is, R_1, R_2, R_3, and R_4, respectively. By *IDWU* the distances between Region R and Region R_i $(1 \leq i \leq 4)$ are the same. Using Equation 2.3 we can calculate that:

$$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \frac{1}{4}$$

Hence, using Equation 2.1 the value of Region R will be interpolated based on Regions R_1, R_2, R_3, and R_4 to be:

$$\begin{aligned} y &= \sum_{i=1}^{4} \lambda_i \cdot y_i \\ &= \frac{1}{4} \times 100 + \frac{1}{4} \times 200 + \frac{1}{4} \times 300 + \frac{1}{4} \times 400 \\ &= 250 \end{aligned}$$

5.2 IDW with Centroid Distance Weights

In this case the distance between the target region and its surrounding region is calculated as the distance between the two corresponding centroid. Consider again the example of neighboring counties. Now the distances between counties C and $N_i, 1 \leq i \leq k$ should be the distances between the centroids of those counties. Consider the example shown in Figure 5.1, if we use IDWC, the distances d_1, d_2, d_3, d_4, and d_5 should be like in Figure 5.2. Because of the near-spherical shape of the Earth, calculating an accurate distance between two points requires the use of the following spherical geometry formulae introduced by Weisstein [65].

$$distance = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2} \tag{5.1}$$

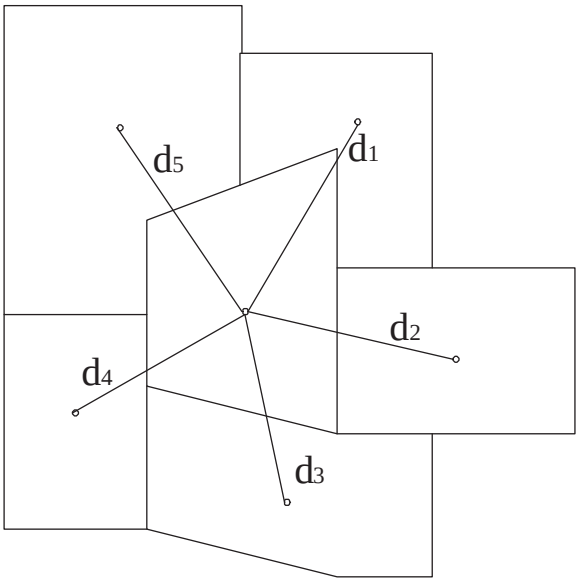

Figure 5.2: Centroids and distances

where for $0 \leq i \leq 1$

$$\begin{aligned} x_i &= R \times \cos(long_i) \times sin(90^\circ - lat_i) \\ y_i &= R \times \sin(long_i) \times sin(90^\circ - lat_i) \\ z_i &= R \times \cos(90^\circ - lat_i) \\ R &= 6368KM \end{aligned}$$

Example 5.2.1 As in Example 5.1.1 we need to interpolate the value in Region R. At this time we assume the distances between the centroid of Region R and the centroid of Region R_i $(1 \leq i \leq 4)$ are $d_1 = 5$, $d_2 = 10$, $d_3 = 15$, and $d_4 = 20$, respectively. By $IDWC$ and using Equation 2.3 we can calculate that:

$$\lambda_1 = \frac{\frac{1}{5}}{\frac{1}{5}+\frac{1}{10}+\frac{1}{15}+\frac{1}{20}} = \frac{12}{25}$$

$$\lambda_2 = \frac{\frac{1}{10}}{\frac{1}{5}+\frac{1}{10}+\frac{1}{15}+\frac{1}{20}} = \frac{6}{25}$$

$$\lambda_3 = \frac{\frac{1}{15}}{\frac{1}{5}+\frac{1}{10}+\frac{1}{15}+\frac{1}{20}} = \frac{4}{25}$$

$$\lambda_4 = \frac{\frac{1}{20}}{\frac{1}{5}+\frac{1}{10}+\frac{1}{15}+\frac{1}{20}} = \frac{3}{25}$$

Hence, using Equation 2.1 the value of Region R will be interpolated based on Regions R_1, R_2, R_3, and R_4 to be:

$$\begin{aligned} y &= \sum_{i=1}^{4} \lambda_i \cdot y_i \\ &= \frac{12}{25} \times 100 + \frac{6}{25} \times 200 + \frac{4}{25} \times 300 + \frac{3}{25} \times 400 \\ &= 192 \end{aligned}$$

Chapter 6

Exponential Decay Temporal Interpolation

We present two approaches to temporal interpolation. In one approach, which we describe in Section 6.1, we apply the spatial IDW method to the temporal interpolation and measure the "distance" as time difference instead of spatial difference. A more interesting approach is the *exponential decay temporal interpolation method*, which is described in Section 6.2.

6.1 IDW Applied to Temporal Data

This approach considers distance as a regular time difference. For example, if we want to estimate the minimum temperature of May 1st in 2002, then the distance between May 1st in 2002 and May 1st in 2003 is 365 days, and the distance between May 1st in 2002 and May 1st in 2000 is 730 days. Then follow the IDW method and use Equation 2.3, the weights are assigned proportional to the inverse of the time distance. Therefore, we also call this method *inverse linear temporal method*.

Example 6.1.1 In Figure 6.1 assume we need to interpolate an unknown value of an attribute at a target point at time t. The target point has some known values of the same attribute before or after time t, that is, $(t_1, v_1), \ldots, (t_5, v_5)$.

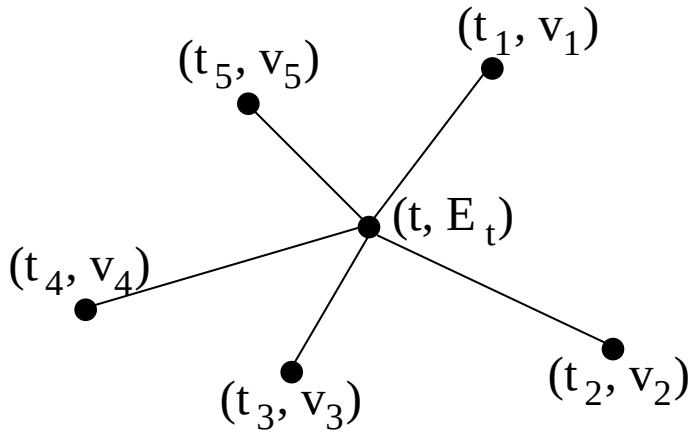

Figure 6.1: Target point and its temporal neighbors

In order to follow the inverse linear temporal method, we first calculate the distance d_i between the time t and its ith temporal neighbors t_i as follows.

$$d_i = |t_i - t|$$

We then calculate the λ_i for the ith neighbor using Equation 2.3 as follows.

$$\lambda_i = \frac{\frac{1}{d_i}}{\sum_{k=1}^{5} \frac{1}{d_k}}$$

Now we can calculate E_t using Equation 2.1 as follows.

$$E_t = \sum_{i=1}^{5} \lambda_i \cdot v_i$$

Example 6.1.2 Assume we need to interpolate the value at time t for some location. Let $y_1 = 100$, $y_2 = 300$, $y_3 = 500$, and $y_4 = 700$ be the values measured at different time instances t_i $(1 \leq i \leq 4)$ at that location. Assume that the regular time distances between time t and t_i $(1 \leq i \leq 4)$ are $d_1 = 1$, $d_2 = 2$, $d_3 = 3$, and $d_4 = 4$, respectively. Using Equation 2.3 we can calculate that:

$$\lambda_1 = \frac{\frac{1}{1}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = \frac{12}{25}$$

$$\lambda_2 = \frac{\frac{1}{2}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = \frac{6}{25}$$

$$\lambda_3 = \frac{\frac{1}{3}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = \frac{4}{25}$$

$$\lambda_4 = \frac{\frac{1}{4}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}} = \frac{3}{25}$$

Hence the interpolated value at time t will be the following:

$$\begin{aligned} y &= \sum_{i=1}^{4} \lambda_i \cdot y_i \\ &= \frac{12}{25} \times 100 + \frac{6}{25} \times 300 + \frac{4}{25} \times 500 + \frac{3}{25} \times 700 \\ &= 284 \end{aligned}$$

6.2 Exponential Decay Temporal Interpolation

Exponential decay or growth is observed in scientific fields and experiments such as population growth, compound interest, radioactive decay, and value depreciation. This gives the idea to consider the influence of any event to decay exponentially with time. The general definition of exponential function is as follows.

Definition 6.2.1 *Let a be a positive real number. The exponential function with base a is the function:*

$$f(x) = ka^{cx}$$

We apply the exponential function in assigning the weights and have the *exponential decay temporal interpolation method*. For simplicity we choose the base $a = 2$. In this approach the weights decrease exponentially with the time difference.

For example, 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.

Example 6.2.1 We consider again the data from Example 6.1.2. Now we use the *exponential decay temporal interpolation method*. Then the weights are assigned as follows:

$$\lambda_1 = \frac{1}{2}\,, \quad \lambda_2 = \frac{1}{4}\,, \quad \lambda_3 = \frac{1}{8}\,, \quad \lambda_4 = \frac{1}{8}$$

Hence the interpolated value at time t will be the following:

$$
\begin{aligned}
y &= \sum_{i=1}^{4} \lambda_i \cdot y_i \\
&= \frac{1}{2} \times 100 + \frac{1}{4} \times 300 + \frac{1}{8} \times 500 + \frac{1}{8} \times 700 \\
&= 275
\end{aligned}
$$

As another example, we look at how to use these two approaches in the USA presidential election prediction.

Example 6.2.2 Given the six USA presidential election results in 2000, 1996, 1992, 1988, 1984, and 1980, and predict the outcome of the USA presidential election in 2004. In this case when we apply the inverse linear temporal interpolation method, we consider the time distance between 2000 and 2004 is one (even though it means four years), the time distance between 1996 and 2004 is two and so on. Actually this specification is still consistent with the idea of regular time distances, but simpler and easier to understand. We use the subscripts in year instead of numbers in this example. Using Equation 2.3 we can calculate that:

$$\lambda_{2000} = \frac{\frac{1}{1}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = \frac{60}{147}$$

$$\lambda_{1996} = \frac{\frac{1}{2}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = \frac{30}{147}$$

$$\lambda_{1992} = \frac{\frac{1}{3}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = \frac{20}{147}$$

$$\lambda_{1988} = \frac{\frac{1}{4}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = \frac{15}{147}$$

$$\lambda_{1984} = \frac{\frac{1}{5}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = \frac{12}{147}$$

$$\lambda_{1980} = \frac{\frac{1}{6}}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}} = \frac{10}{147}$$

When we consider the *exponential decay temporal interpolation method*, the weights are as follows:

$$\begin{aligned}
\lambda_{2000} &= \frac{1}{2} \\
\lambda_{1996} &= \frac{1}{4} \\
\lambda_{1992} &= \frac{1}{8} \\
\lambda_{1988} &= \frac{1}{16} \\
\lambda_{1984} &= \frac{1}{32} \\
\lambda_{1980} &= \frac{1}{32}
\end{aligned}$$

The weight $\frac{1}{32}$ occurs twice to make the sum of the weights 1.

Chapter 7

Spatiotemporal Interpolation of Climate Data

The interpolation of climate data has been a focus of research for a long time. Because the availability of climatic measurements varies spatially and temporally, sets of climate data are usually incomplete. The sources of missing data may be that the instrument was broken and the data was never recorded, that there was a break in the data transmission, or that there was a mistake during data-entry and data processing. Techniques are necessary to fill gaps in data sequences before accurate valuation analysis can be performed.

One motivation of this dissertation is that climatology researchers mainly use spatial interpolation methods to do the estimation. However, in some situations, spatial interpolation methods are not accurate enough. Suppose the IDW interpolation method is used and consider the following scenarios.

(1) In mountainous regions, the assumptions used by the IDW method do not hold (see Section 2.2.1 for the assumptions of IDW).

(2) Some weather stations may not have enough nearby stations for estimation,

while the assumption of IDW is based on certain number of surrounding stations.

(3) Several nearby stations have data for the same time instance, and spatial methods can be used for the estimation, but the estimation accuracy is poor. For example, if we define "nearby" as within 50 miles, but all the nearby stations are between 45 to 50 miles, then the accuracy will be poor.

Therefore we recognize that the temporal methods can be useful in combination with spatial methods in the regions where spatial methods can not work well in themselves.

In this chapter we discuss how to apply our adaptive spatiotemporal interpolation method to interpolate climate data. Section 7.1 to 7.3 describe how to solve the three issues in applying our adaptive spatiotemporal interpolation method to climate data. Section 7.4 gives the evaluation. We compare the step function based and linear function based spatiotemporal interpolation methods with pure spatial interpolation and pure temporal interpolation methods. We also compare the results in mountainous regions with those in plain areas.

7.1 Determination of E_s

In Section 2.1 we identify the climate data as point-based data and since IDW is a point-based spatial interpolator, we adopt the standard IDW as the spatial interpolation method to spatially interpolate climate data. Assume that some climate data are missing at one target weather station. Following the IDW method described in Section 2.2.1, we need to find the closest neighboring stations. In this study we look for the five closest neighboring stations of a target station, that is, $N = 5$ in Equation 2.3. Table 7.1 gives an example of the five closest neighboring weather stations of a given

target station. After selecting the neighboring stations we can calculate the distances between the target station and its neighboring stations using Equation 5.1. Based on the distances we can calculate the weights for each neighbor using Equation 2.3. Now we have the weights and observations for individual neighbor, it is ready to calculate the spatial estimate using Equation 2.1.

Table 7.1: Neighboring stations of a target weather station

Station ID	Latitude	Longitude
1	40.36889	-105.51083
2	40.26694	-105.83222
3	40.18532	-105.86667
4	39.99194	-105.26667
5	39.89333	-105.86285
Target	40.22889	-105.51833

7.2 Determination of E_t

Two things need to be noticed when we adopt the temporal interpolation. One thing is that adopting temporal interpolation actually means that in reconstructing missing data, using the local historical weather records could be an alternative to using data from neighboring stations. The other thing is that the reconstruction of missing climate data and weather forecasting are different because for the former the data collected both before or after the missing gap can be used.

Assume that in a station some climate data on a particular day of a particular year is missing and the historical records are available. We apply the *inverse linear temporal method* described in Section 6.1 to calculate the temporal estimate. We first

calculate the time distances as described in Section 6.1. Based on the time distances we calculate the weights for each temporal neighbor using Equation 2.3. Then we can calculate the temporal estimate using Equation 2.1 since we have the weights and observations for individual neighbor.

7.3 Determination of α and β

We adopt the step function-based combination and the linear function-based combination which are proposed in Section 4.3 to determine α and β. Therefore, in this application of climate data interpolation, we have the following methods: our adaptive spatiotemporal interpolation method with step function-based combination (ASTS), adaptive spatiotemporal interpolation method with linear function-based combination (ASTL), the spatial IDW interpolation method (IDW) and the inverse linear temporal interpolation method (LT).

According to the definitions of $I_{i,t}^{M}$ and $O_{i,t}$ in Section 4.3, in this application $I_{i,t}^{M}$ is the interpolated value for station i at time t using method M and $O_{i,t}$ is the observed value for station i at time t. In particular, we have $I_{i,t}^{IDW}$, $I_{i,t}^{ASTS}$, $I_{i,t}^{ASTL}$, and $I_{i,t}^{LT}$. Let σ_i be the standard deviation of the elevations among the target station i and its closet N neighbors, that is,

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N}(S_{i,j} - T_i)^2}{N}} \tag{7.1}$$

where T_i is the elevation of the target station i, and $S_{i,j}$ is the elevation of the jth neighboring station of the target station i.

According the definition of $D_{i,t}^{M}$ in Section 4.3, in this application $D_{i,t}^{M}$ is the absolute difference between the interpolated value using method M and the observed

value at station i at time t. In particular, we have $D_{i,t}^{IDW}$ and $D_{i,t}^{LT}$.

The constant threshold θ can be fixed as $100, 200$ and so on. For some θ, if most stations with $\sigma_i < \theta$ have smaller $D_{i,t}^{IDW}$, while most stations with $\sigma_i \geq \theta$ have smaller $D_{i,t}^{LT}$, then it means spatial interpolation is more accurate for stations with $\sigma_i < \theta$ and linear temporal interpolation is more accurate for stations with $\sigma_i \geq \theta$. Hence a step function-based combination as show in Figure 4.2 would be a good choice.

Now we consider the linear function-based combination. When σ_i increases, the neighboring stations are less likely to be on the same elevation with the target station, therefore we should decrease the weights of spatial estimate and increase the weights of temporal estimate. Hence, the linear function shown in Figure 4.5 satisfies this intuition well.

7.4 Evaluation

7.4.1 Design of the Study

We randomly select weather stations in Colorado and Nebraska and use the minimum daily temperature data of the time period from 1993 to 2003. Figures 7.1 and 7.2 show the weather stations we have used. We assume that the minimum daily temperature on a particular day (e.g., May 1st) of a particular year (e.g., 2002) is unknown and has to be reconstructed from the data from the neighboring weather stations and its local weather records. That is, the missing data can be interpolated based on the minimum daily temperature data on May 1st, 2002 from neighboring weather stations and its own data on May 1st, 1993 until 2001, and 2003. We interpolate the minimum daily temperature for the May-August period in 2002 for 50 weather stations in both Colorado and Nebraska.

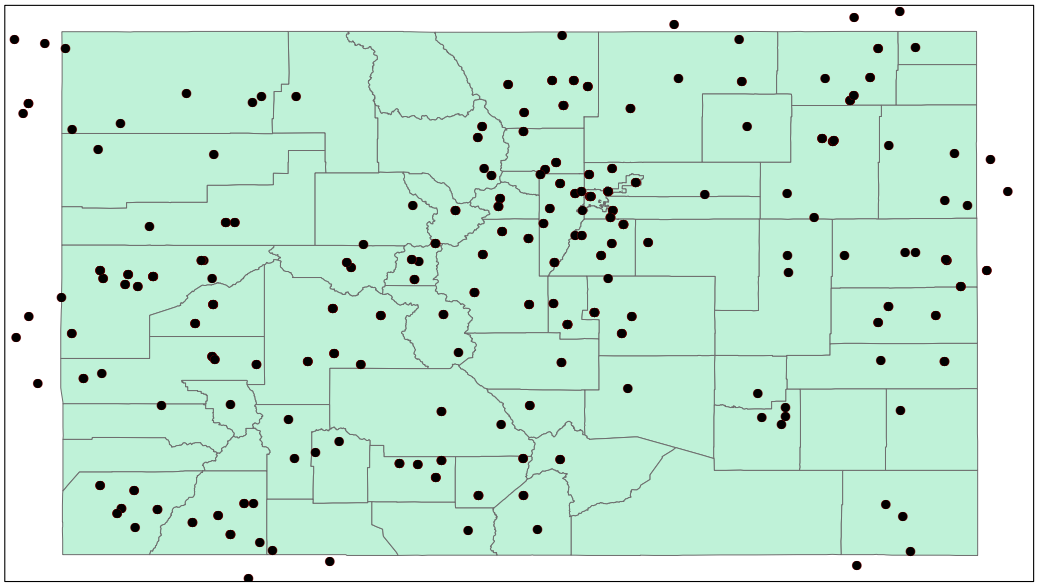

Figure 7.1: Weather stations in Colorado

7.4.2 Evaluation Methods

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

$$MAE = \frac{\sum_{i=1}^{N} |I_i - O_i|}{N} \tag{7.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (I_i - O_i)^2}{N}} \tag{7.3}$$

where

I_i: Interpolated value.

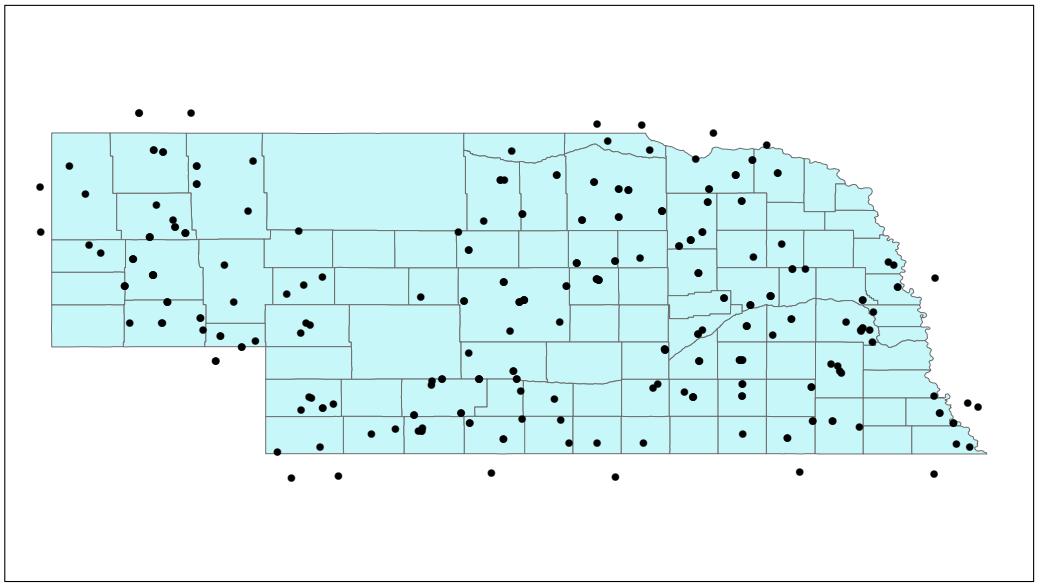

Figure 7.2: Weather stations in Nebraska

O_i: Observation value.

N: Number of data.

7.4.3 Comparison of Four Methods

Figure 7.3 gives some idea about how step function-based combination works. The x-axis is σ_i and the y-axis is $D_{i,t}^{IDW}$ and $D_{i,t}^{LT}$ (See Section 7.3 for the definitions of σ_i, $D_{i,t}^{IDW}$ and $D_{i,t}^{LT}$). Each station has a star and a box on the same vertical line, where the star is $D_{i,t}^{IDW}$ and the box is $D_{i,t}^{LT}$. We add a vertical line on the σ_i value equal to 500. On the left of the line are weather stations with $\sigma_i < 500$ and on the right of the line are weather stations with $\sigma_i \geq 500$. We can find that most stations on the left have smaller $D_{i,t}^{IDW}$ and most stations on the right have smaller $D_{i,t}^{LT}$. Therefore, 500 seems a reasonable threshold value for this data set.

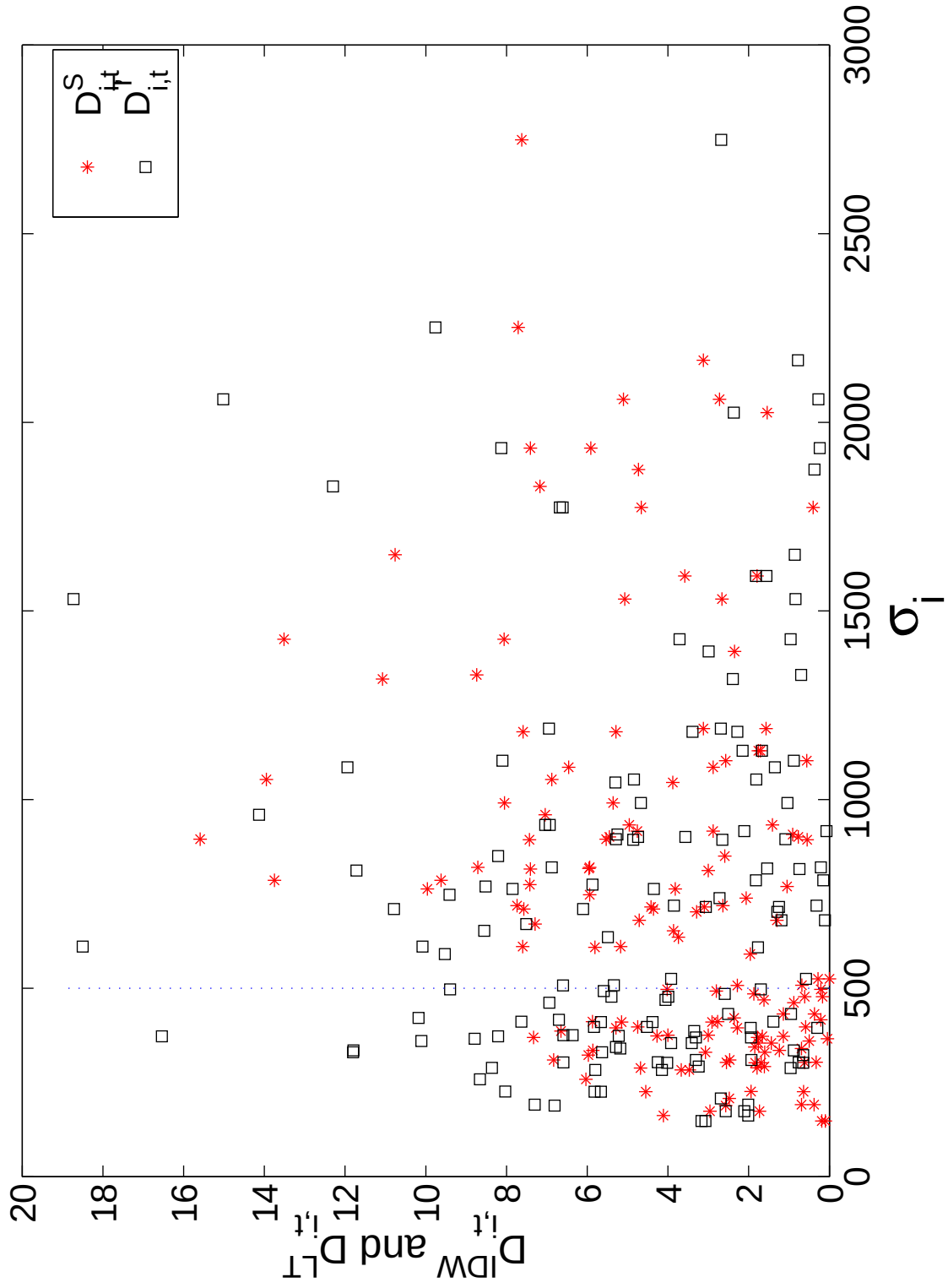

Figure 7.3: Each station appears twice as a star and a box within the vertical line $\sigma_i = \theta$. We see that for $\sigma_i < 500$, the stars are lower than the boxes in general, and for $\sigma_i \geq 500$, the boxes are lower than the stars. Both the stations and dates are randomly chosen.

To test the best performance of the ASTS method, we have tried the threshold values (100, 150, ..., 1450, 1500). we compared their MAE and RMSE with those of the IDW and the LT methods. In Table 7.2, the MAE and RMSE columns summarize the results for the various methods. We can see that the ASTS method yields better performance than either the IDW or the LT method. The IDW method has 9% and the LT method has 25% less accurate MAE than the best the ASTS method. Similarly, the IDW method has 4% and the LT method has 28% less accurate RMSE than the best ASTS method has.

Table 7.2: Comparison of ASTS, IDW and LT

Method	Best parameters	MAE	$\frac{Method's MAE}{BestMethod's MAE}$	RMSE	$\frac{Method's RMSE}{BestMethod's RMSE}$
ASTS	$\theta = 950$	3.8452	1.00	4.6988	1.00
IDW	$N = 5, p = 1$	4.1958	1.09	4.8912	1.04
LT		4.8114	1.25	6.0262	1.28

In order to test the performance of the ASTL method, we did experiments on 40 threshold values (100, 200, 300, ..., 4000) and 10 rates (0.0, 0.1, ..., 0.9), and recorded the best combination of those parameters and results in Table 7.3. We can see from Table 7.3 that the ASTL method yields much better performance than either the IDW or the LT method. The IDW method has 21% less accurate MAE than the best the ASTL method. The LT method has 39% less accurate MAE than the best the ASTL method. Similarly, the IDW method has 15% and the LT method has 42% less accurate RMSE than the best ASTL method has.

Figures 7.4 and 7.5 show the MAE and RMSE of 50 weather stations in Colorado, respectively. Both figures record the result of the best ASTL method, the IDW method, and the LT method. From these two figures we can see that the ASTL

Table 7.3: Comparison of ASTL, IDW and LT

Method	Best parameters	MAE	$\frac{Method's MAE}{Best Method's MAE}$	RMSE	$\frac{Method's RMSE}{Best Method's RMSE}$
ASTL	$r = 0.3, \theta = 1400$	3.4598	1.00	4.2586	1.00
IDW	$N = 5, p = 1$	4.1958	1.21	4.8912	1.15
LT		4.8114	1.39	6.0262	1.42

method has better performance than the other two. While Figure 7.4 demonstrates the trend of MAE, Figures 7.6 to 7.9 show the MAE value of individual weather station for the IDW, ASTS, ASTL, and LT methods. Similarly, while Figure 7.5 demonstrates the trend of RMSE, Figures 7.10 to 7.13 show the RMSE value of individual weather stations for the IDW, ASTS, ASTL, and LT methods.

Figures 7.6 to 7.8 show the MAE of 50 Colorado weather stations using the IDW, ASTS, and ASTL methods, respectively. For the legends, the color blue means the range of MAE is [0.00, 2.00], light blue [2.01-3.00], green [3.01-4.00], yellow [4.01-6.00], and red [6.01-10.00]. We can see that in general both spatiotemporal methods yield better results than IDW, and the ASTL method is the best. Figure 7.9 shows the MAE of 50 Colorado weather stations using the LT method. We did not use the same legend as in Figures 7.6 to 7.8, because the minimum MAE by the LT method is 3.68, more than 2.00. Obviously the LT method is the worst among the four methods.

Table 7.4 records the minimum, average, and maximum values of MAE using the four methods in Colorado. We can see that the IDW, ASTS and ASTL methods share the same minimum value 1.29, while the LT method has a much higher minimum value 3.68. The ASTL method has the lowest maximum value of MAE, followed by the LT, ASTS and IDW methods.

Figures 7.10 to 7.12 show the RMSE of 50 Colorado weather stations using the

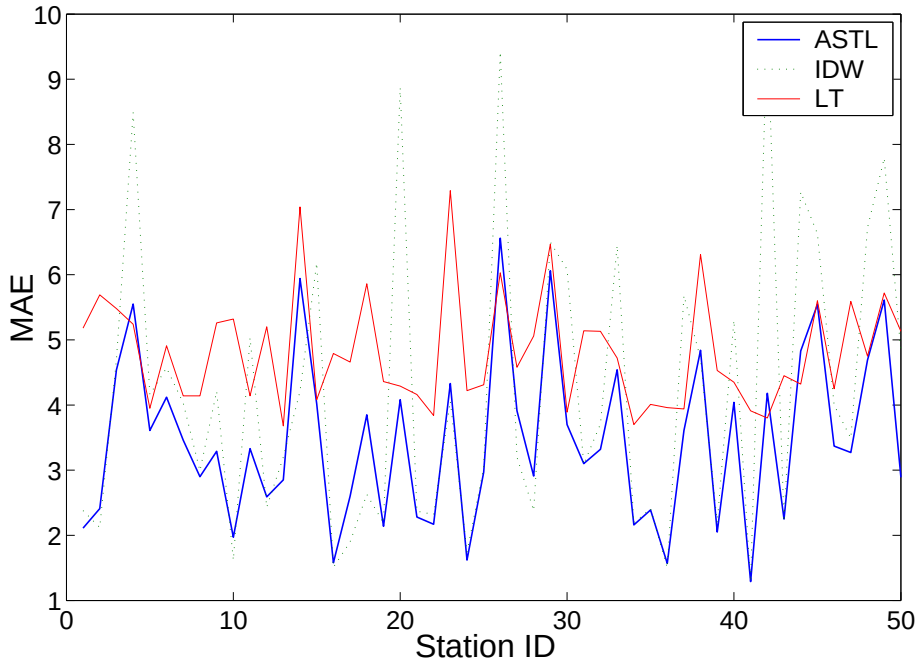

Figure 7.4: MAE of 50 Colorado stations over May to August 2002

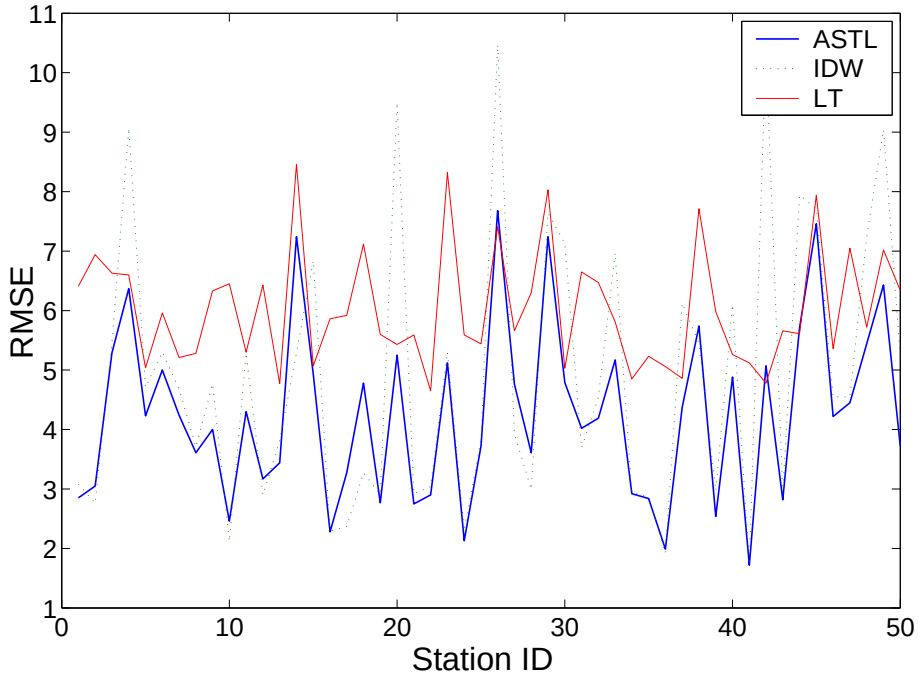

Figure 7.5: RMSE of 50 Colorado stations over May to August 2002

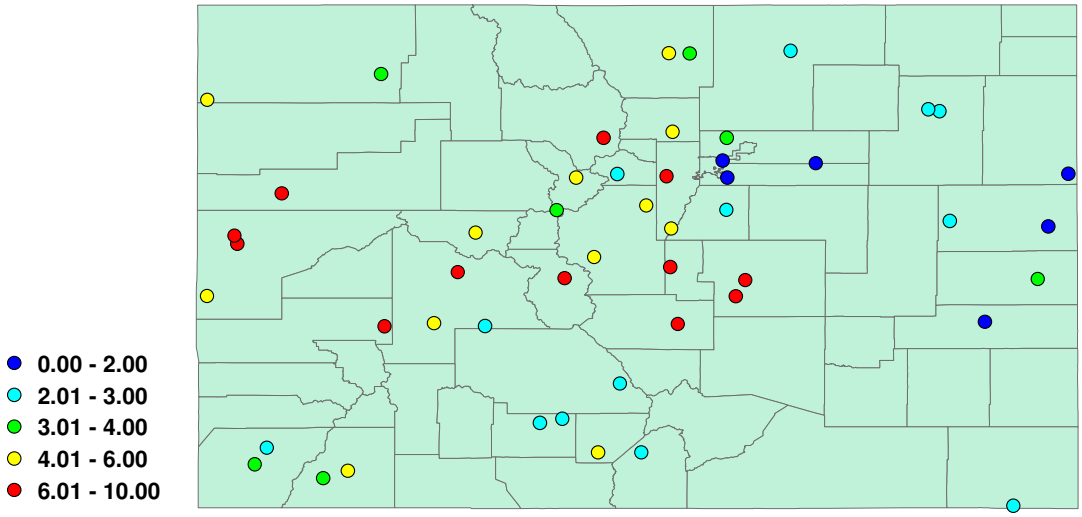

Figure 7.6: MAE of weather stations in Colorado using IDW over May to August 2002

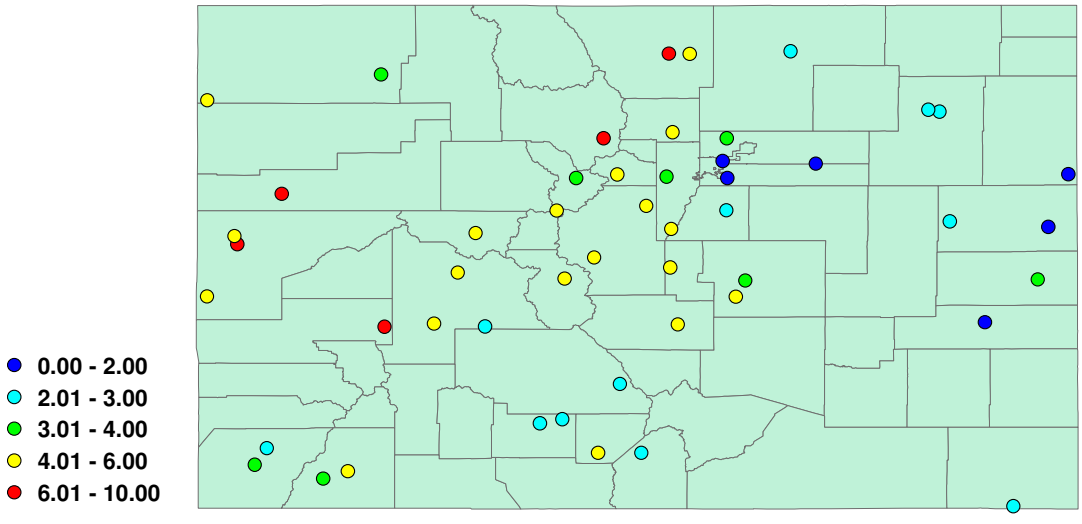

Figure 7.7: MAE of weather stations in Colorado using ASTS over May to August 2002

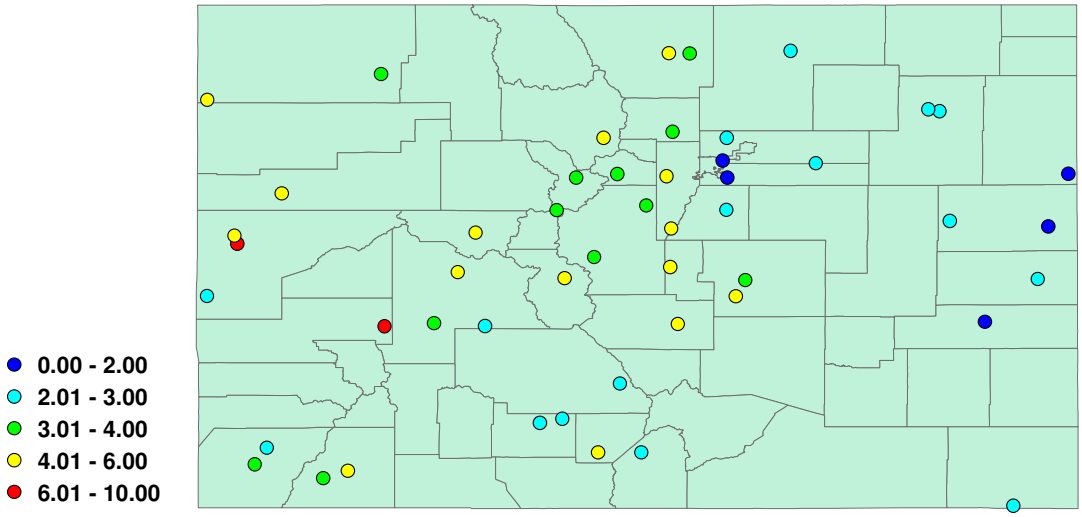

Figure 7.8: MAE of weather stations in Colorado using ASTL over May to August 2002

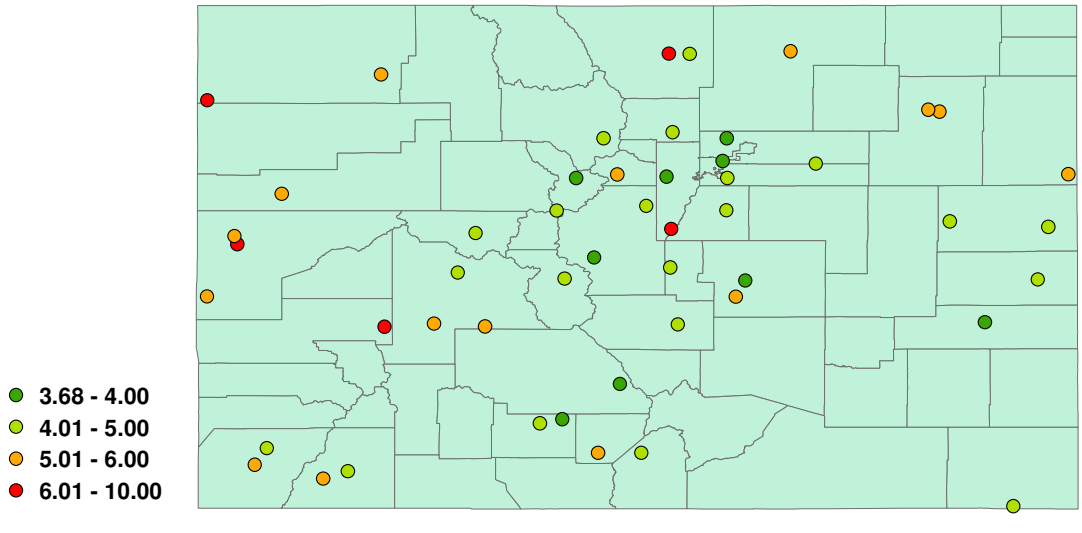

Figure 7.9: MAE of weather stations in Colorado using LT over May to August 2002

Table 7.4: Minimum, average, and maximum values of MAE and RMSE using four methods in Colorado

	MAE			RMSE		
	Min	Avg	Max	Min	Avg	Max
IDW	1.29	4.20	9.82	1.73	4.89	10.47
ASTS	1.29	3.85	9.42	1.73	4.70	10.47
ASTL	1.29	3.46	6.56	1.72	4.26	7.68
LT	3.68	4.81	7.29	4.65	6.03	8.46

IDW, ASTS, and ASTL methods, respectively. For the legends, the color blue means the range of RMSE is [0.00, 2.50], light blue [2.51-3.50], green [3.51-4.50], yellow [4.51-6.50], and red [6.51-11.00]. We can see that the results of RMSE and MAE are similar, that is, the two spatiotemporal methods still yield better results than IDW, and the ASTL is still the best. Figure 7.13 shows the RMSE of 50 Colorado weather stations using the LT method. The figure for the LT method also uses a different legend, because it yields much different results compared with those of the other three methods. Similarly to that of MAE, the LT method is the worst among the four methods.

Table 7.4 records the minimum, average, and maximum values of RMSE using the four methods in Colorado. It shows a similar trend to that of MAE. The first three methods still have smaller minimum value (1.72 or 1.73) than the LT method which has 4.65. For the maximum value of RMSE, the ASTL method still has the lowest RMSE, followed by the LT , ASTS and the IDW methods.

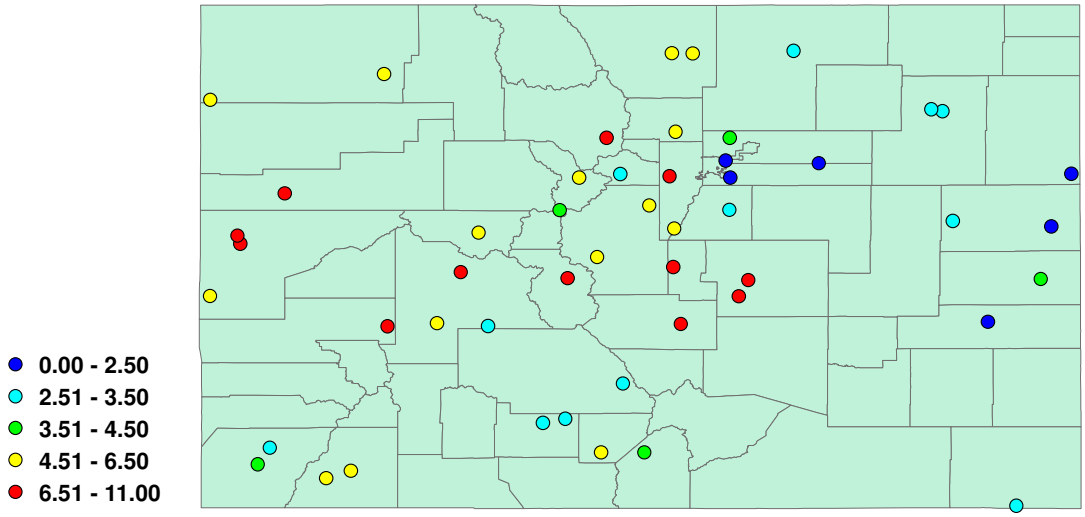

Figure 7.10: RMSE of weather stations in Colorado using IDW over May to August 2002

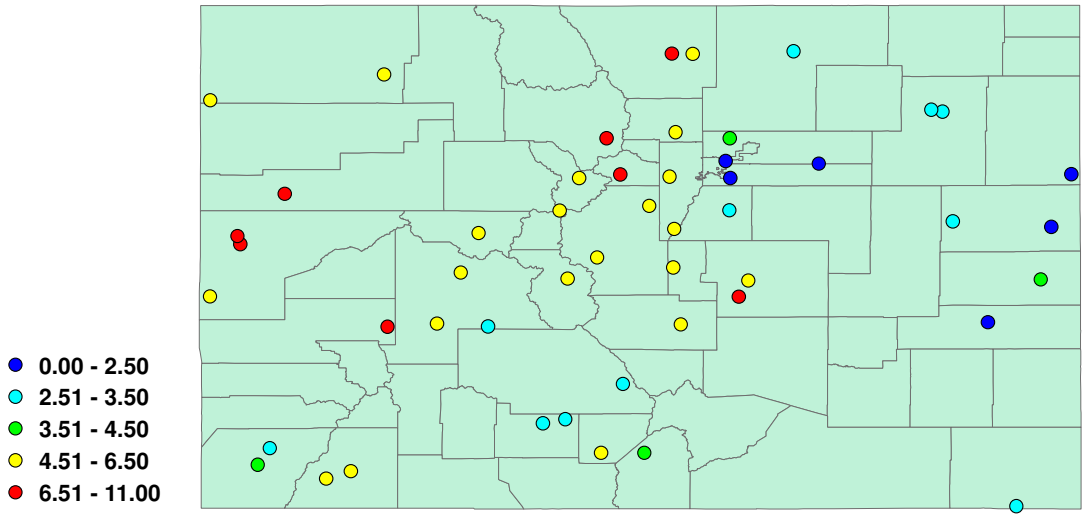

Figure 7.11: RMSE of weather stations in Colorado using ASTS over May to August 2002

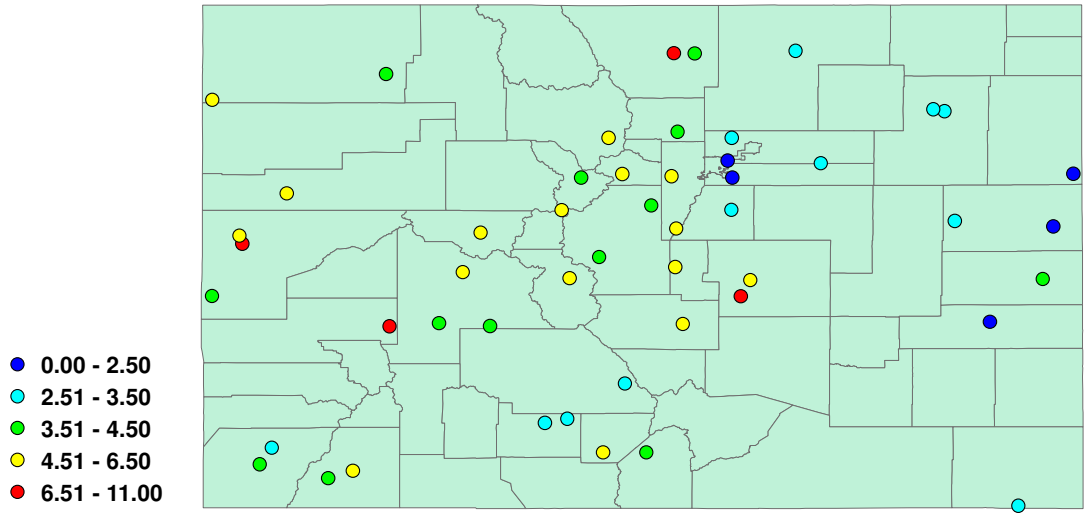

Figure 7.12: RMSE of weather stations in Colorado using ASTL over May to August 2002

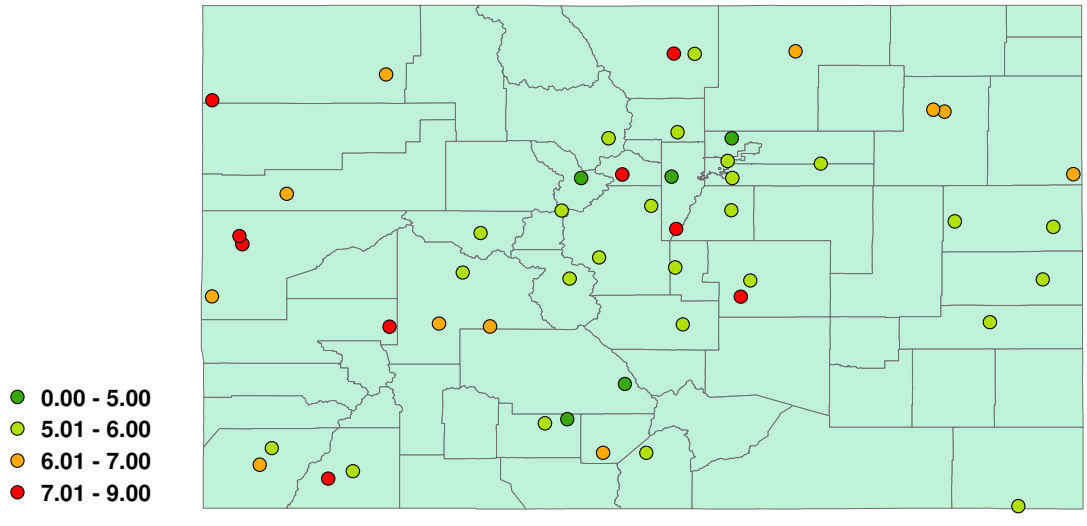

Figure 7.13: RMSE of weather stations in Colorado using LT over May to August 2002

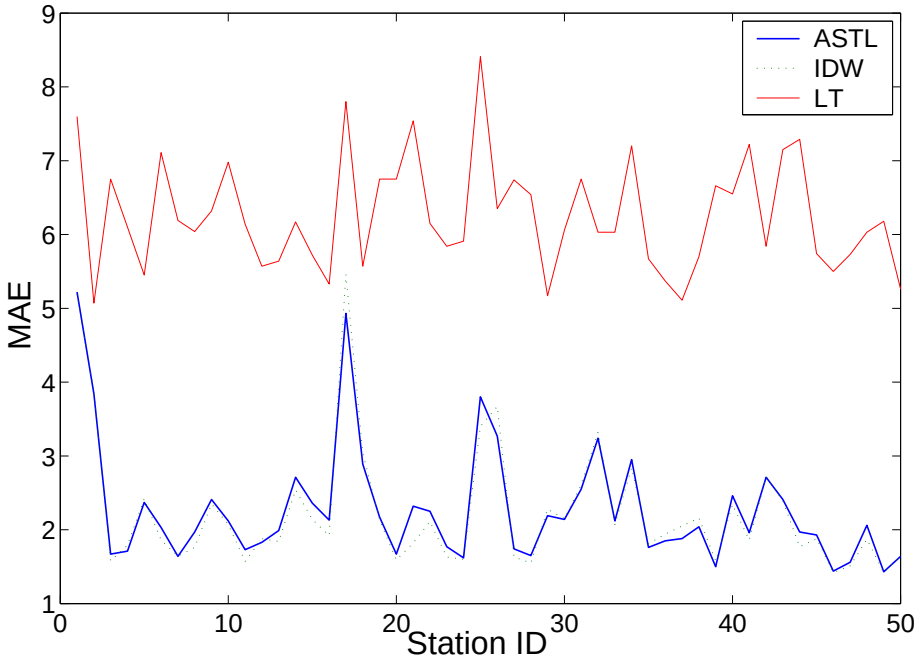

Figure 7.14: MAE of 50 Nebraska stations over May to August 2002

7.4.4 Comparison of Mountainous Regions and Plain Areas

We did experiments on Nebraska weather stations too. In this case the IDW method yields the best performance among the tested methods. Figure 7.14 shows the MAE of 50 weather stations in Nebraska and records the result of the best ASTL method, the IDW method, and the LT method. We can see that the IDW method yields slightly better performance than the ASTL method, and the LT method is the worst case. While Figure 7.14 demonstrates the trend of MAE, Figures 7.15 to 7.18 show the MAE value of individual weather station for the IDW, ASTS, ASTL, and LT methods.

Figures 7.15 to 7.17 show the MAE of 50 Nebraska weather stations using the IDW, ASTS, and ASTL methods, respectively. For the legends, the color blue means the range of MAE is [0.00, 2.00], light blue [2.01-3.00], green [3.01-4.00], yellow [4.01-5.00], and red [5.01-6.00]. Compared with the results in Colorado, the value of MAE

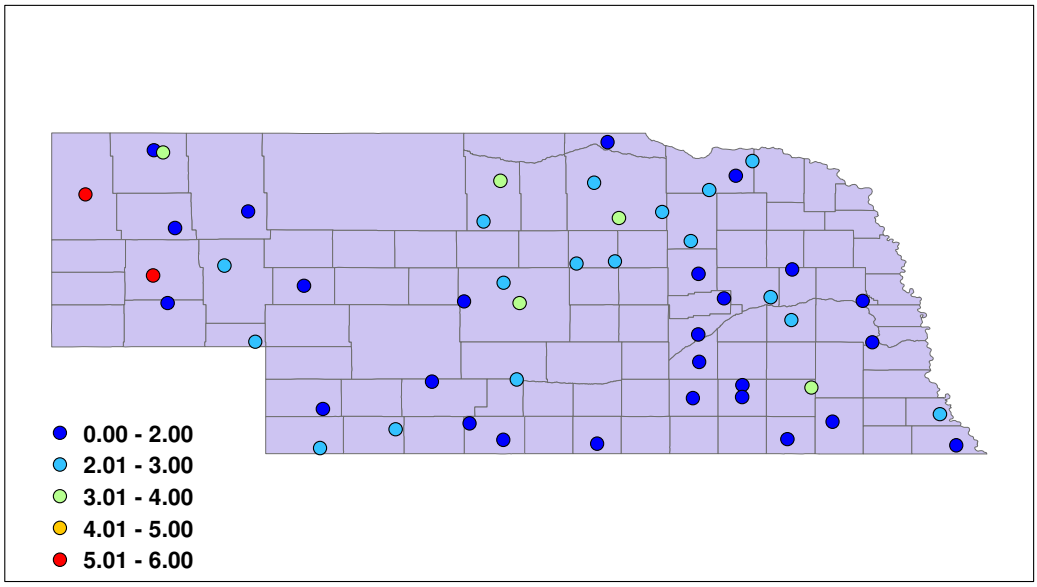

Figure 7.15: MAE of weather stations in Nebraska using IDW over May to August 2002

is smaller. The maximum MAE is less than 6.00. We can see that among the three methods, the IDW method is the best, with the IDW method slightly better than the ASTL method, and the ASTL method slightly better and the ASTS method.

This result can be explained as follows. Since Nebraska is a plain area, the weather stations in Nebraska have a better chance of having a close neighbor with similar heights than weather stations have in Colorado which is a mountains area.

Similarly to the results in Colorado, the LT method still yields the worst performance, and the difference between the LT method and the other three methods becomes more obvious in Nebraska. Figure 7.18 shows the MAE of 50 Nebraska weather stations using the LT methods. We did not use the same legend due to the same reason in Colorado.

Table 7.5 records the minimum, average, and maximum values of MAE using the four methods in Nebraska. We can see that the IDW, ASTS and ASTL methods have

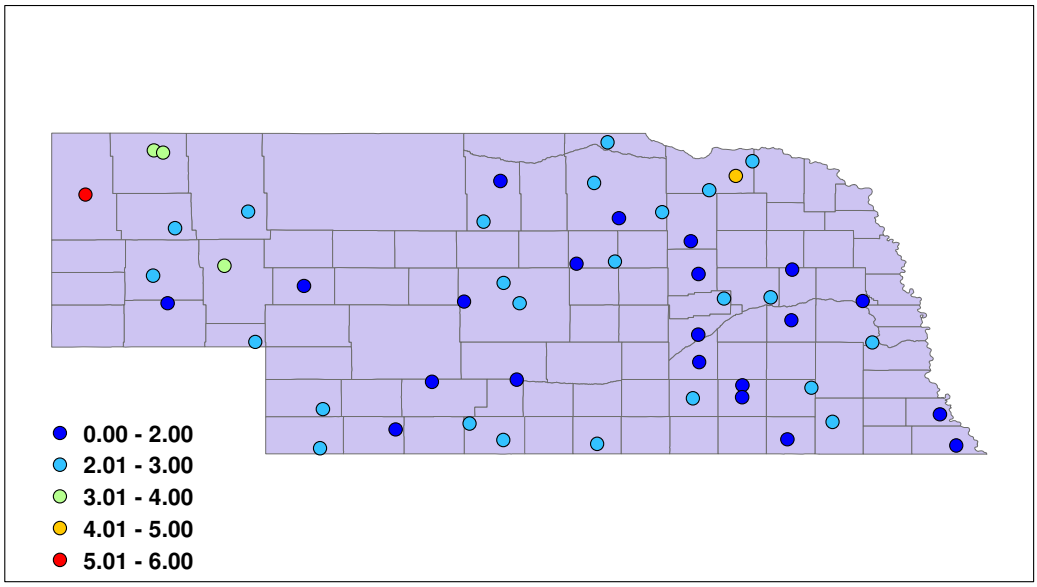

Figure 7.16: MAE of weather stations in Nebraska using ASTS over May to August 2002

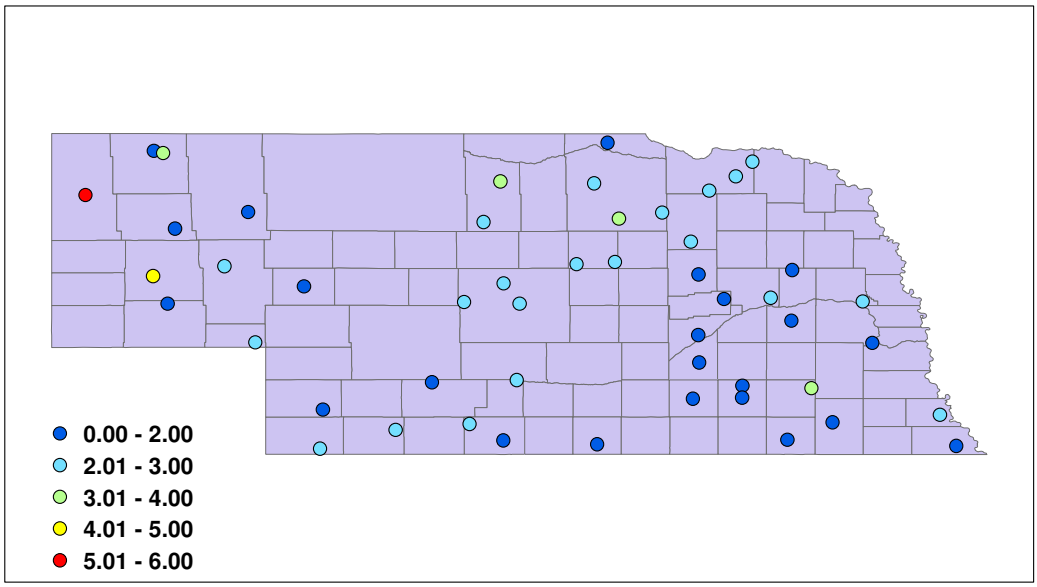

Figure 7.17: MAE of weather stations in Nebraska using ASTL over May to August 2002

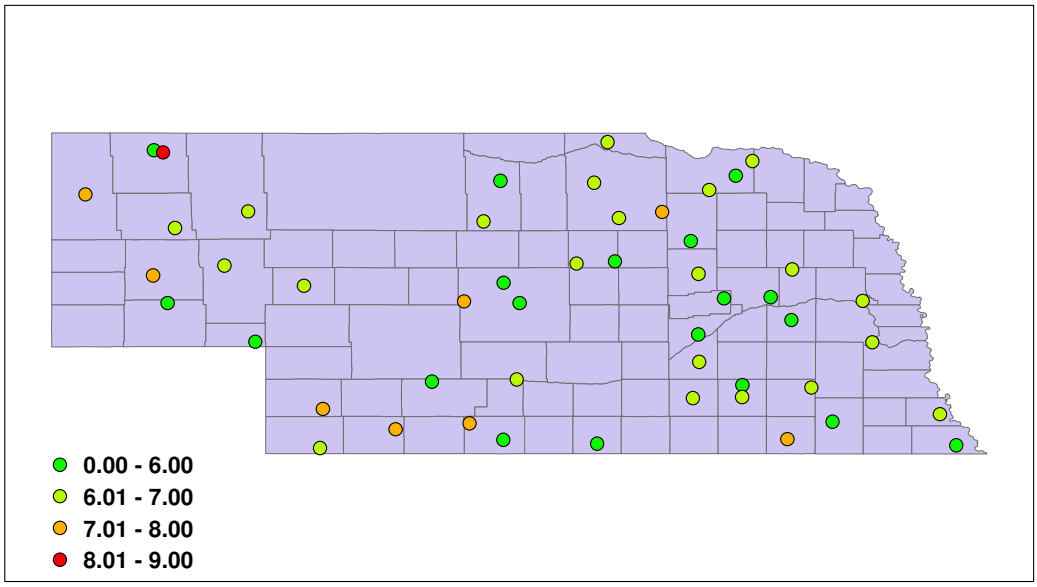

Figure 7.18: MAE of weather stations in Nebraska using LT over May to August 2002

lower minimum values (1.42, 1.43, or 1.53) than the LT method has (5.07). The trend of maximum values is similar to that of minimum values. The first three methods have lower maximum values than the LT method has.

Figures 7.19 to 7.21 show the RMSE of 50 Nebraska weather stations using the IDW, ASTS, and ASTL methods, respectively. Here the legend is the following. The color blue means the range of MAE is [0.00, 2.00], light blue [2.01-3.00], green [3.01-4.00], yellow [4.01-6.00], and red [6.01-8.00]. Compared with the results in Colorado, the value of RMSE is smaller, with the maximum RMSE is less than 8.00. Similarly to the result of MAE, for all three methods, most of the stations have RMSE less than 3.00.

Similarly to the results of MAE in Nebraska, the LT method still yields the worst performance, and the difference between the LT method and the other three methods is more obvious. Figure 7.22 shows the RMSE of 50 Nebraska weather stations using

Table 7.5: Minimum, average, and maximum values of MAE and RMSE using four methods in Nebraska

	MAE			RMSE		
	Min	Avg	Max	Min	Avg	Max
IDW	1.42	2.23	5.46	1.90	2.95	7.24
ASTS	1.53	2.26	5.14	2.05	2.96	6.81
ASTL	1.43	2.27	5.22	1.92	2.98	6.74
LT	5.07	6.26	8.41	4.23	7.60	10.24

the LT methods. Using the temporal method, only one station has RMSE less than 6.00, and 33 out of 50 stations have RMSE between 6.00 and 8.00, and the other 16 stations have 8.00-11.00. We can see that for the IDW, ASTL, and ASTS methods, all the stations have RMSE less than 8.00, furthermore, most of the stations have RMSE less than 3.00.

Table 7.5 records the minimum, average, and maximum values of RMSE using the four methods in Nebraska. It has almost same trend with that of MAE values. The first three methods still have much smaller minimum and maximum values than the LT method.

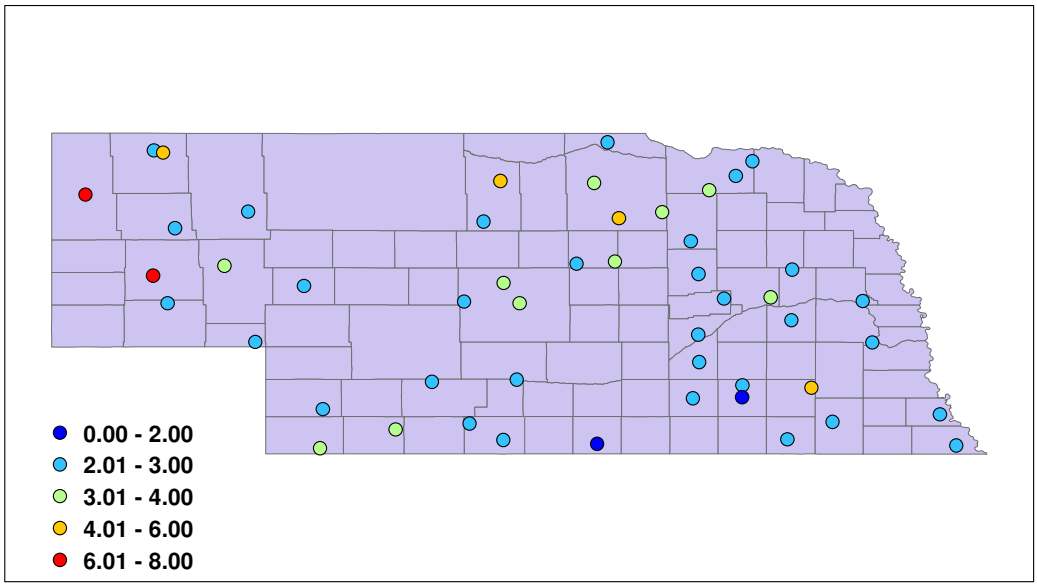

Figure 7.19: RMSE of weather stations in Nebraska using IDW over May to August 2002

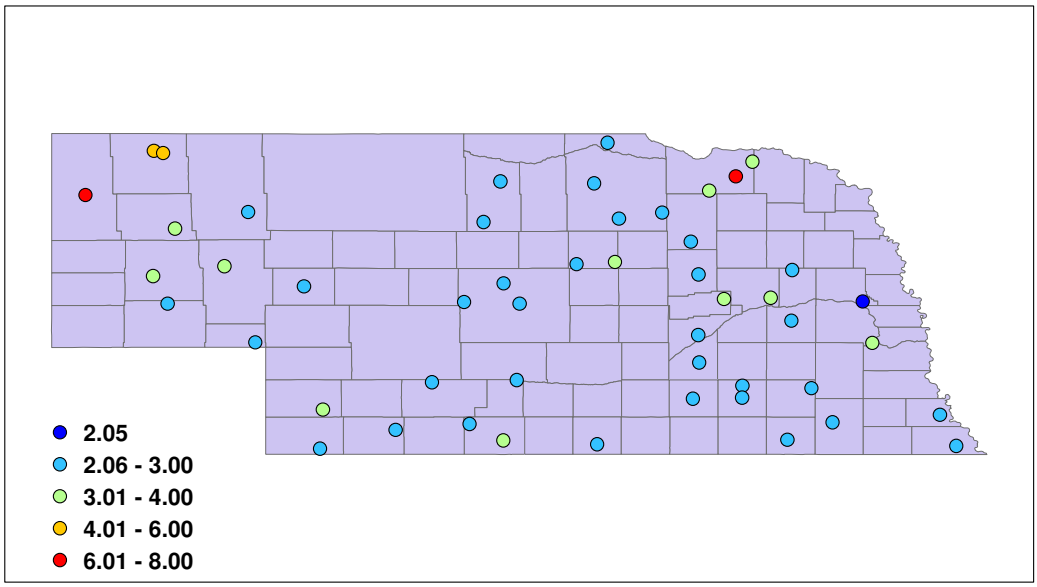

Figure 7.20: RMSE of weather stations in Nebraska using ASTS over May to August 2002

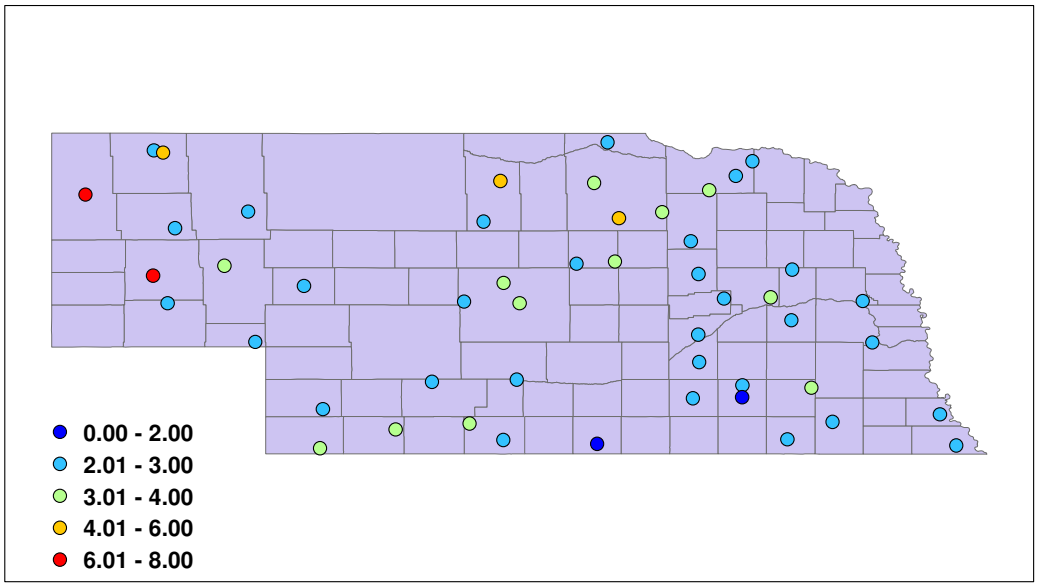

Figure 7.21: RMSE of weather stations in Nebraska using ASTL over May to August 2002

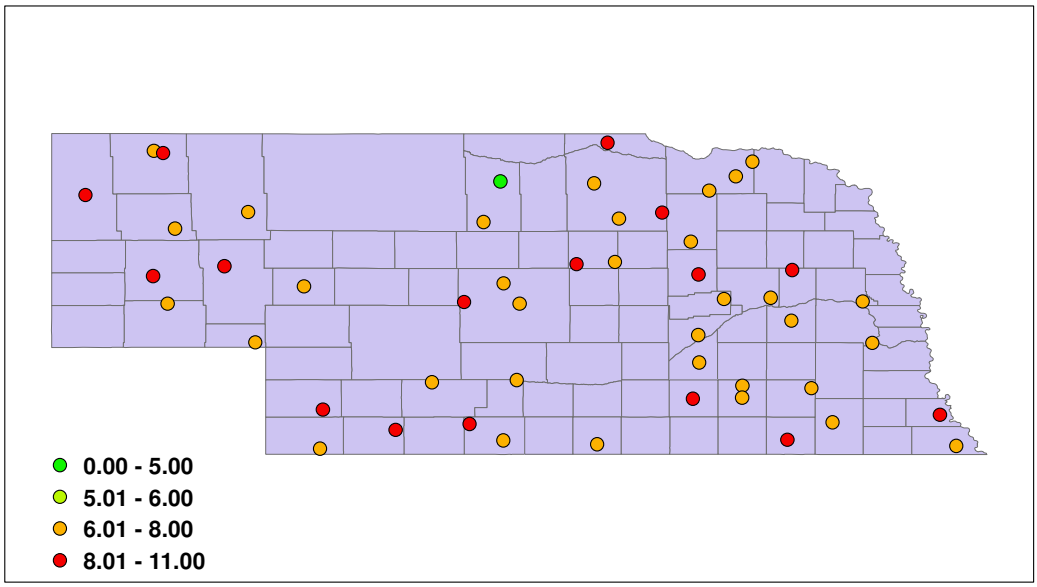

Figure 7.22: RMSE of weather stations in Nebraska using LT over May to August 2002

Chapter 8

USA Presidential Election Prediction

As describe in Chapter 3 the USA presidential election data is a typical temporal-dominated data set. We apply our adaptive interpolation method to this election data set. In interpolating the percentage vote for a given party in some county C for which we do not have information, we would naturally like to rely on its historical results in previous election and the percentage votes in its neighboring counties if those values are known. For election voting, it is very unlikely that previous voting result can not be found. And what people are most interested in is who will win in the coming election. Therefore, instead of doing a interpolation, we use our method to do a prediction.

As stated in Section 2.4, for many presidential election forecasting models one common limitation is the choice of factors to include in the model. We aim to keep the number of variables to a minimum. Therefore, our method focuses on the historical election data itself and uses it as the basis of the spatiotemporal interpolation without any additional variables.

8.1 Determination of E_t

We apply the two temporal interpolation approaches, *inverse linear temporal method* and *exponential decay temporal method* which are proposed in Chapter 6, to calculate the temporal interpolation values.

8.2 Determination of E_s

Since the USA presidential election data is a region-based data, we apply the two approaches, *IDW with uniform weights* and *IDW with centroid distance weights* which are proposed in Chapter 5 to calculate the spatial interpolation values.

However, when we use IDW to do the prediction instead of interpolation, a problem arises. For 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. It is not reasonable to use the actual votes in Bradford, Clay, Columbia, Gilchrist, Levy, Marion, Putnam, and Union, which are the neighboring counties of Alachua, 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 exponential decay temporal methods.

An interesting aspect is that in states with long and narrow shapes, such as Florida, there are fewer neighbors on average for each county than in counties with a more round shape such as Ohio. Therefore, we were concerned that the overall shape of a state can influence heavily the accuracy of our spatiotemporal interpolation method. Hence we choose three states, that is, California, Florida, and Ohio, with very different shapes as our test cases.

In each of our three test states, there are counties that have additional neighbors in other states. For example, some counties in Florida are neighbors of some counties

in Georgia. However, we did not count neighbors in other states, because we did not have available data for them. Presumably the accuracy of our interpolation methods can be further improved by counting those neighbors too.

8.3 Determination of α and β

We adopt the step function-based combination and the linear function-based combination which are proposed in Section 4.3 to determine α and β. Therefore, in this application of election data prediction, we have the following methods: our adaptive spatiotemporal interpolation method with step function-based combination (ASTS), adaptive spatiotemporal interpolation method with linear function-based combination (ASTL), the IDW with uniform weights interpolation method (IDWU), the IDW with centroid distance weights interpolation method (IDWC), the inverse linear temporal interpolation method (LT), and the exponential decay temporal method (EDT).

According to the definitions of $I_{i,t}^{M}$ and $O_{i,t}$ in Section 4.3, in this application $I_{C,t}^{M}$ is the interpolated value for county C at time t using method M and $O_{C,t}$ is the original value for county C at time t. In particular, we have $I_{C,t}^{IDWU}$, $I_{C,t}^{IDWC}$, $I_{C,t}^{ASTS}$, $I_{C,t}^{ASTL}$, $I_{C,t}^{LT}$, and $I_{C,t}^{EDT}$.

According the definition of $D_{i,t}^{M}$ in Section 4.3, in this application $D_{C,t}^{M}$ is the absolute difference between the interpolated value using method M and the original value at county C at time t. In particular, we have $D_{C,t}^{IDWUorIDWC}$ and $D_{C,t}^{LTorEDT}$.

Let σ_C be the changes in the vote percentages of all pairs of subsequent presidential elections for a county C (See Section 8.4 for an example about the calculation of the σ_C for a county C), that is,

$$\sigma_C = \frac{\sum_{i=1}^{N}(P_i - P_{i+1})}{N} \tag{8.1}$$

where P_i is the vote percentage of ith election elevation in a county C.

The constant parameter θ can be fixed as 1%, 2% and so on.

Intuitively, a smaller σ_C means that the values in a county C are more consistent over time, hence we can rely more on the temporal interpolation method, which means that we should decrease α_C and increase β_C. Following this intuition, when we consider a step function, if $\sigma_C < \theta$, then we set $\alpha_C = 0$ and $\beta_C = 1$, which enforces that we use the temporal interpolation method; and if $\sigma_C \geq \theta$, then we set $\alpha_C = 1$ and $\beta_C = 0$, which enforces that we use the spatial interpolation method. We can describe it in the Equation 8.2, which is a reverse of the step function described in Section 4.3.1.

$$\begin{cases} \alpha_C = 0,\ \beta_C = 1 & \text{if } \sigma_C < \theta \\ \alpha_C = 1,\ \beta_C = 0 & \text{if } \sigma_C \geq \theta \end{cases} \tag{8.2}$$

For some θ, if most counties with $\sigma_C < \theta$ have $D_{C,t}^{LTorEDT} < D_{C,t}^{IDWUorIDWC}$, while most counties with $\sigma_C \geq \theta$ have $D_{C,t}^{LTorEDT} \geq D_{C,t}^{IDWUorIDWC}$, then the step function of Equation 8.2 would be a good choice.

We also experimented with the linear function-based combination 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 methods (LT or EDT) and the spatial methods (IDWU or IDWC) give similar variations for most counties, that is, when the temporal interpolated value is higher (or lower) than the original data, then the spatial interpolated value is also higher (or lower). That makes it difficult to find a good linear function.

8.4 Determination of σ in Election Data

Suppose we would like to predict the outcome of the USA presidential election of 2004 in Alachua, Florida. Let us look at how to calculate $\sigma_{Alachua}$.

Let P_{year} be the percentage vote for the democratic candidate in the given year in Alachua and use P_{00} instead of P_{2000} and so on. We have $P_{00} = 55.249682$ %, $P_{96} = 53.896139$ %, $P_{92} = 49.608382$ %, $P_{88} = 48.827313$ %, $P_{84} = 46.423513$ %, and $P_{80} = 52.287084$ %.

Let d be the absolute difference between two continuous USA presidential elections, then $d_1 = |P_{00} - P_{96}|, \ldots, d_5 = |P_{84} - P_{80}|$. That is, $d_1 = |55.249682\ \% - 53.896139\ \%| = 1.353543$ %, $d_2 = 4.287756$ %, $d_3 = 0.781069$ %, $d_4 = 2.4038$ %, and $d_5 = 5.863571$ %.

Hence we get:

$$\sigma_{Alachua} = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5} = 2.937948\ \%$$

Table 8.1 gives d_i and σ_C of six counties of the state of Florida. We calculated similarly the σ_C for the remaining 61 counties in Florida, but we do not show them for space limitations.

8.5 Evaluation

8.5.1 USA Presidential Election Data Sets

As stated before, in order to test our idea, we used the USA presidential election data for the states of California, Florida, and Ohio. For Florida, the data is obtained from the official web site [68], which is maintained by the Florida Division of Elections and contains a comprehensive USA presidential voting data for 67 different counties in

Table 8.1: d_i and σ_C of 67 counties of Florida, USA

d_i in each county	00/96	96/92	92/88	88/84	84/80	σ_C
Alachua	1.353543	4.287756	0.781069	2.403800	5.863571	2.937948
Baker	4.927593	5.031435	0.896374	0.045522	24.17032	7.014248
Bay	0.951147	4.895554	1.633311	2.241453	11.67629	4.279550
...						
Wakulla	2.071604	8.078953	1.023610	1.278107	16.490388	5.7885324
Walton	3.626663	5.728941	1.235268	3.937664	20.734451	7.0525974
Washington	3.204958	5.745716	0.326218	3.265001	18.464455	6.2012696

Florida between 1980 and 2004. Table 8.2 shows a part of the post-calculated data. For California and Ohio, the data is obtained from [36], for the time period between 1972 and 2004. We estimated the votes for the 2004 democratic candidate for USA president (John Kerry) in those three states using our new method and compared them with the actual votes.

8.5.2 Prediction Procedures

We tried out the inverse linear and the exponential decay temporal methods to get the temporal estimates. We used the IDWU and IDWC to get the spatial estimates.

Once we get the temporal and spatial interpolation values, we apply Equation 4.1 to calculate the final estimation value. We tested various step functions to find the best estimation parameters α, β, and θ. For the threshold parameter θ we tried the ten values 1%, 2%, 3%, ..., 10%.

Table 8.2: Votes for 2000 USA presidential election in 67 counties of Florida, USA

County name	Total votes	Votes for Republican candidate	Votes for Democratic candidate
Alachua	85,757	34,135	47,380
Baker	8,155	5,611	2,392
Bay	58,876	38,682	18,873
...			
Wakulla	8,587	4,512	3,838
Walton	18,323	12,186	5,643
Washington	8,026	4,995	2,798

8.5.3 Evaluation Methods

In order to analyze the quality of interpolation we conduct the experiments based on three measures comparing the accuracy of interpolation methods, MAE, 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.

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.

$$VPstate_e = \frac{\sum E_i \times V_i}{\sum V_i} \tag{8.3}$$

where

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) as follows.

$$TE = |VPstate_e - VPstate_a| \tag{8.4}$$

Example 8.5.1 Assume that a state S has three counties A, B, and C. For some election the numbers of voters in counties A, B, and C are 1000, 2000, and 3000, respectively. The estimated vote percentages for a given party in counties A, B, and C are 40%, 50%, 60%, respectively. And the actual vote percentage for a given party in state S is 58%. We can calculate that:

$$TE = \left|\frac{40\%\times1000+50\%\times2000+60\%\times3000}{1000+2000+3000} - 58\%\right| = 4.7\%$$

8.5.4 Experimental Results

Table 8.3 records our experimental results. We can see that the performance of the ASTS and EDT methods is the best, getting comparatively precise predictions, especially in predicting the 2004 USA presidential election in Florida. ASTS (with $\theta = 7\%$) predicts 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 [69], i.e., it had a TE of more than 5%. Let us look at the results of the presidential election forecasting models. For example, predictions of republican votes in Florida based on the polls of the week between Oct 25 and Nov 1 give an average TE of 2.2% [11].

The experimental results for California and Ohio are also impressive. The EDT method shows slightly better performance, TE is 3.46 and 3.18 in California and Ohio, respectively. Let us look at the prediction of republican votes in the exit polls. TE is 1.5 in California and 4.4 in Ohio, respectively [11]. For all three states, MAE and RMSE are reasonably low.

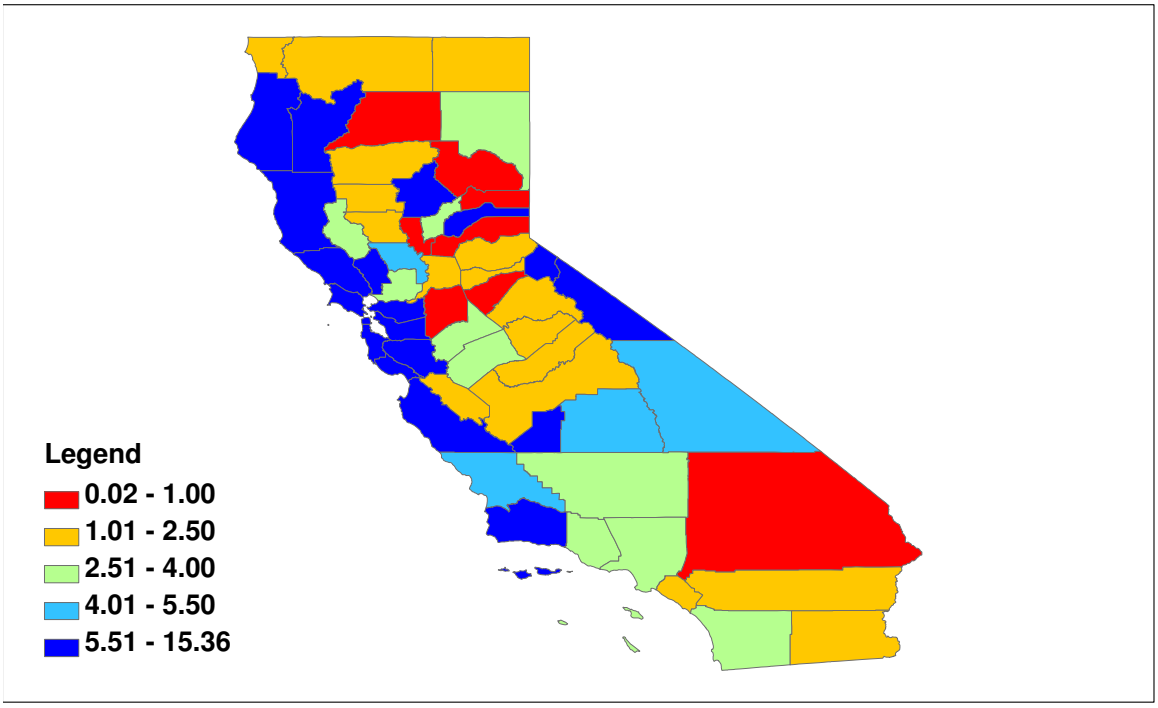

Figure 8.1: Prediction accuracy in California, USA

The table shows that the difference between IDWU and IDWC are extremely small.

Figures 8.1-8.6 illustrate the voting prediction results of the 2004 USA presidential election in the states of California, Florida, and Ohio at the county level. Figures 8.1, 8.2, and 8.3 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 8.4, 8.5, and 8.6 the dashed line shows the actual vote percentage in each county and the solid line describes the estimated vote percentages using ASTS 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.

Table 8.3: Comparison of ASTS, LT, EDT, IDWU, and IDWC methods

Method	California 2004			Florida 2004			Ohio 2004		
	TE	MAE	RMSE	TE	MAE	RMSE	TE	MAE	RMSE
Using IDWU									
IDWU	8.65	11.60	9.67	4.88	7.98	9.05	8.75	11.31	7.60
ASTS ($\theta = 7\%$)	3.49	4.51	6.26	1.09	2.40	5.18	3.57	4.37	3.57
ASTS ($\theta = 8\%$)	3.55	4.77	6.38	1.10	2.40	4.72	3.89	4.66	3.88
ASTS ($\theta = 9\%$)	3.49	4.51	6.26	1.10	2.39	4.61	3.27	4.05	3.14
Using IDWC									
IDWC	8.02	11.33	9.33	3.51	6.62	8.64	8.83	11.27	7.45
ASTS ($\theta = 7\%$)	3.58	4.63	6.83	1.10	2.39	4.84	3.45	5.06	4.88
ASTS ($\theta = 8\%$)	3.54	4.54	6.32	1.11	2.39	4.69	3.78	4.56	3.71
ASTS ($\theta = 9\%$)	3.50	4.51	6.03	1.11	2.39	4.59	3.25	4.03	3.10
LT	5.46	6.66	7.25	2.68	3.81	5.12	4.10	5.09	3.74
EDT	3.46	4.48	6.01	1.10	2.39	4.59	3.18	3.99	3.10

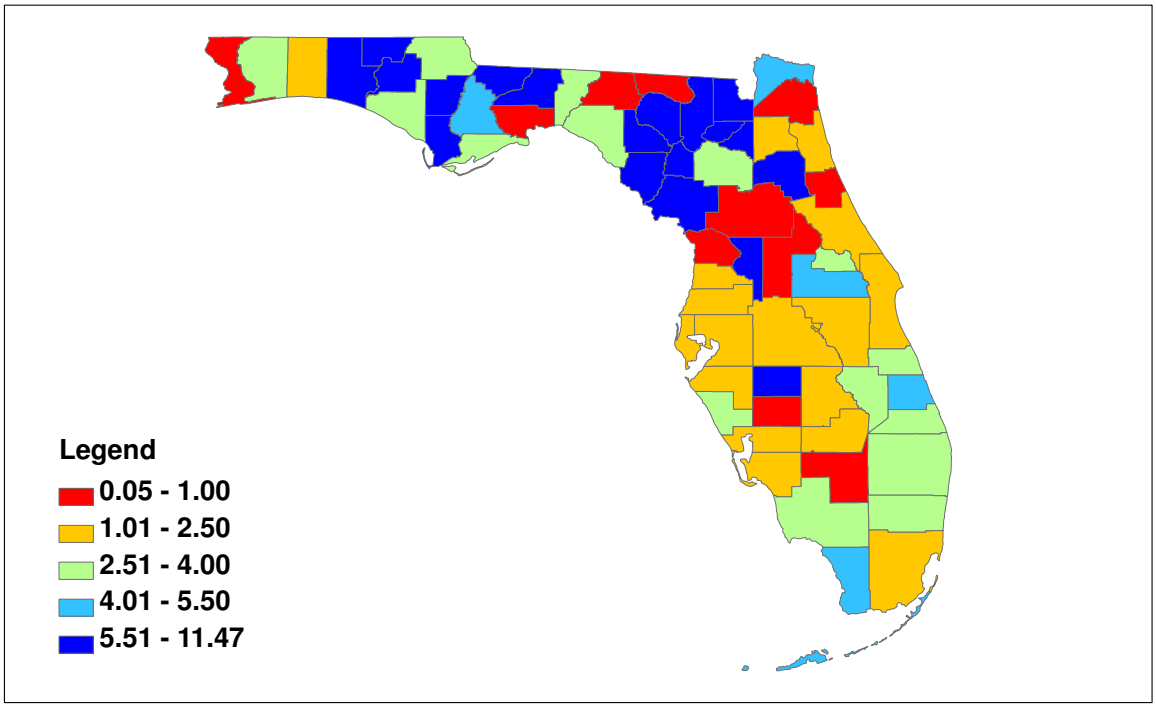

Figure 8.2: Prediction accuracy in Florida, USA

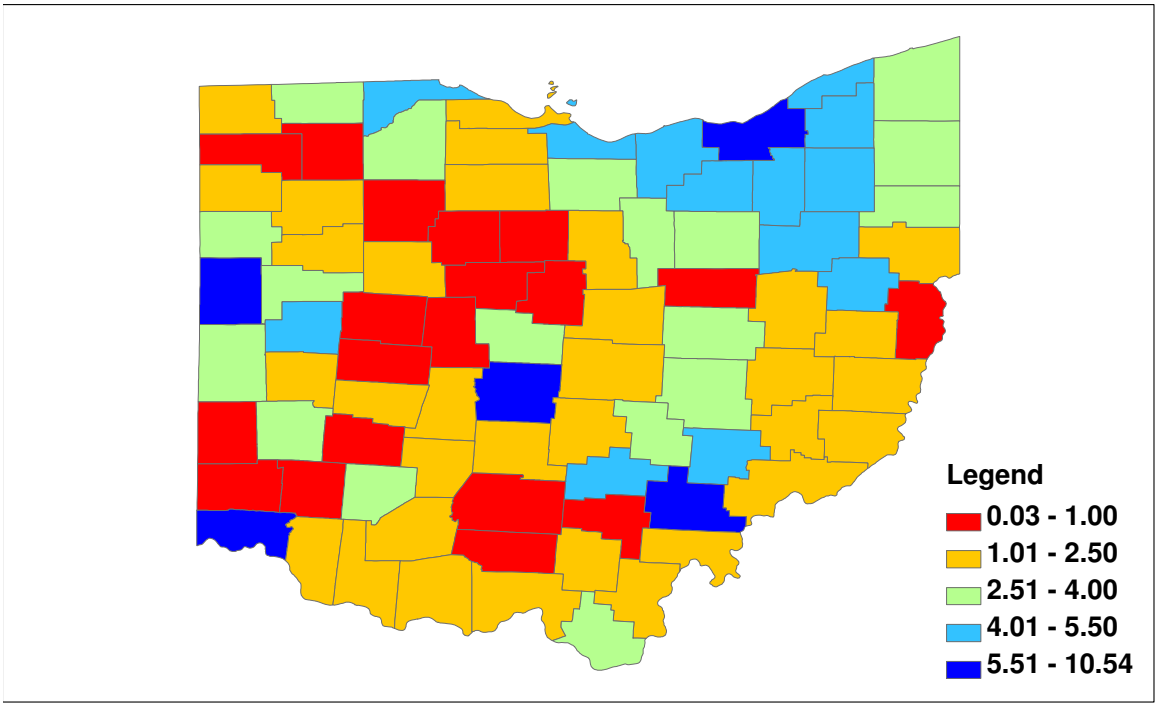

Figure 8.3: Prediction accuracy in Ohio, USA

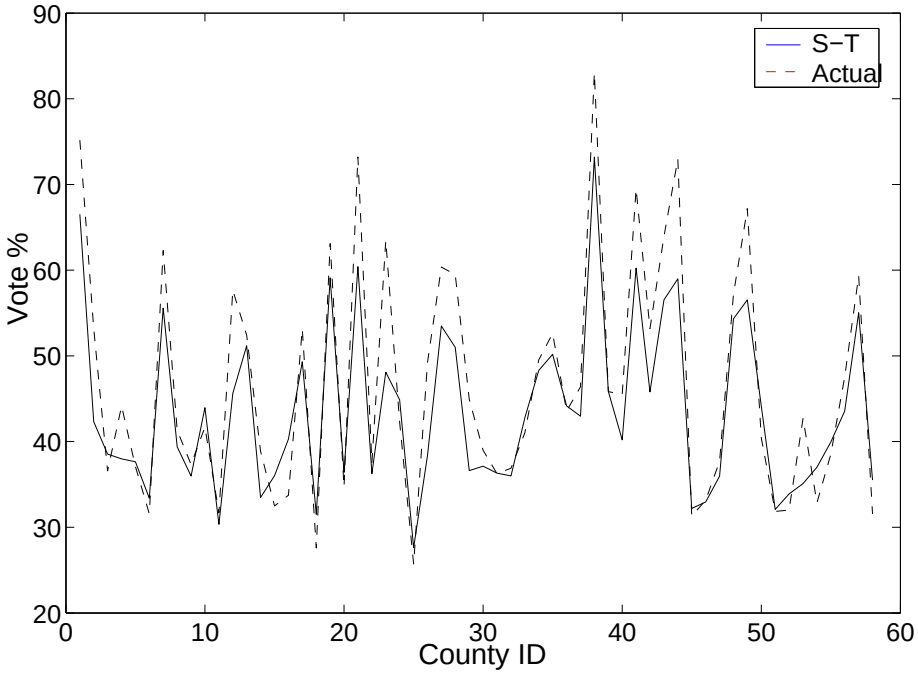

Figure 8.4: Predicted and actual voting in California

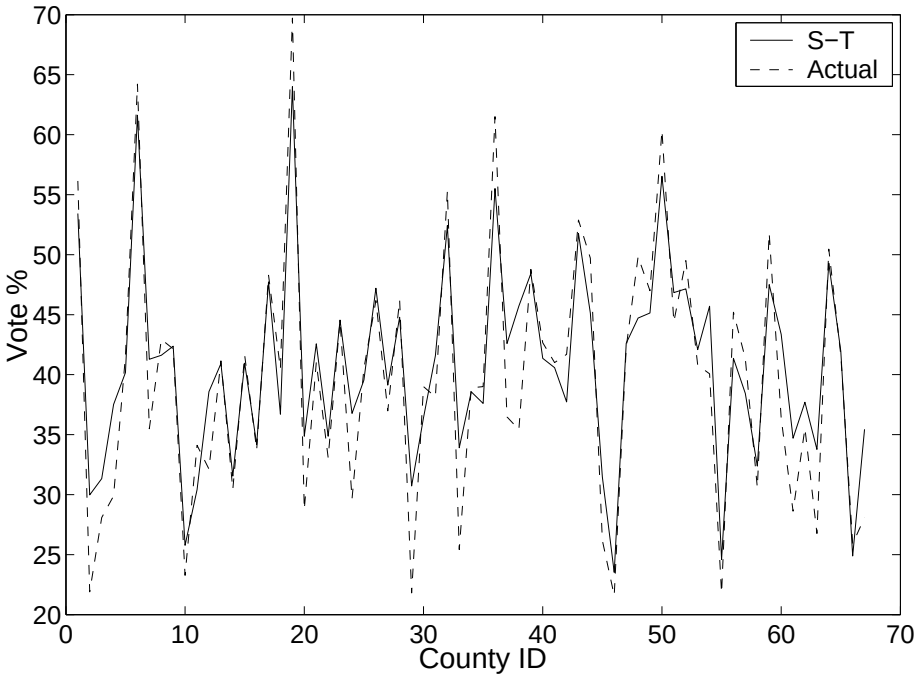

Figure 8.5: Predicted and actual voting in Florida

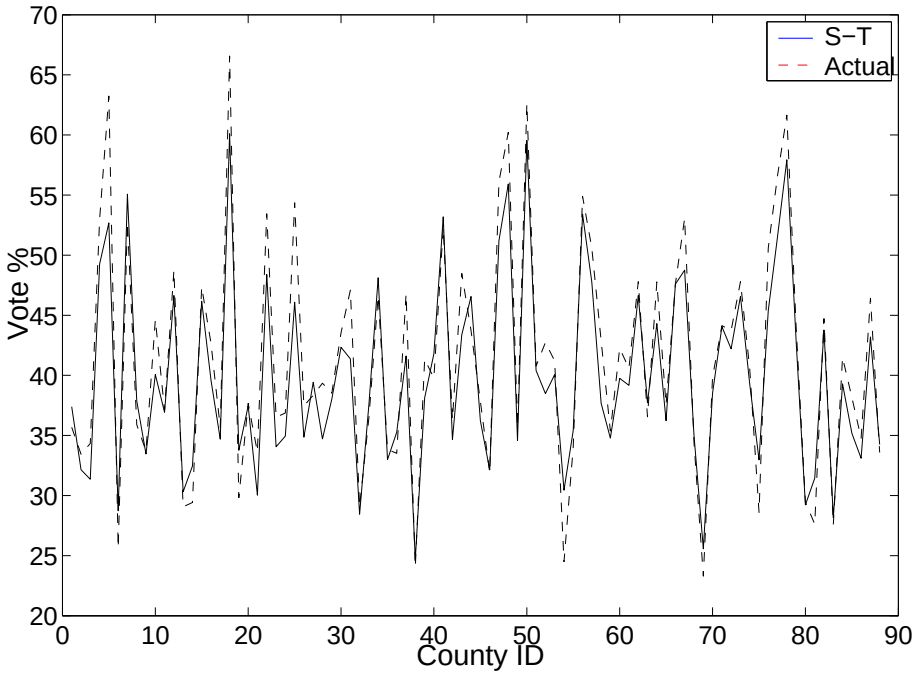

Figure 8.6: Predicted and actual voting in Ohio

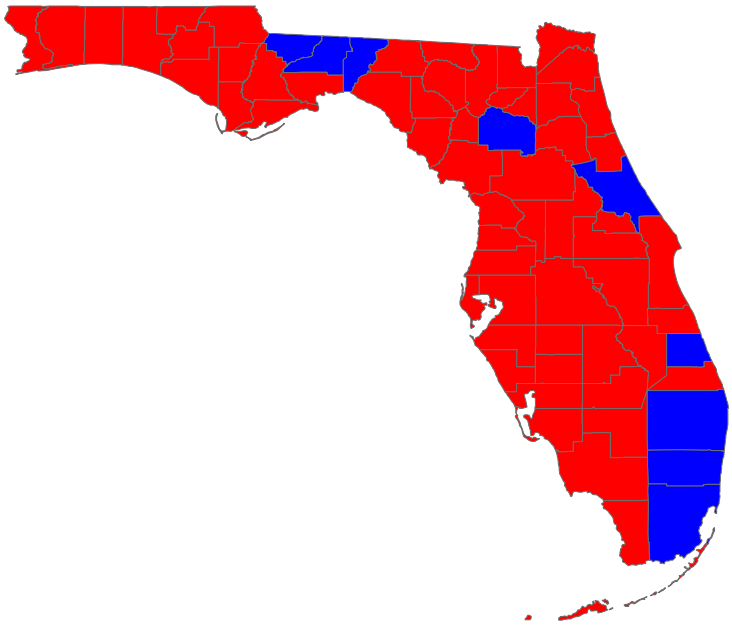

Figure 8.7: Actual results: red and blue counties in Florida, USA

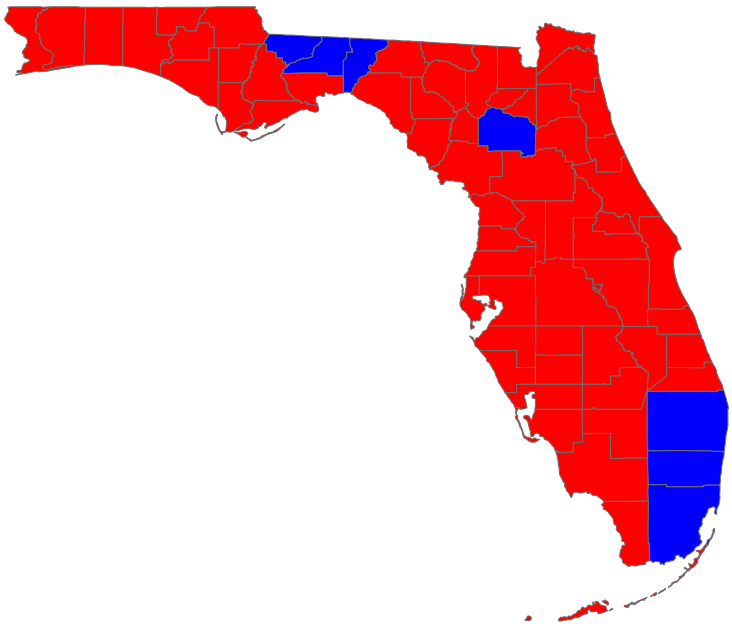

Figure 8.8: Interpolated results: red and blue counties in Florida, USA

In Figures 8.7 and 8.8 red counties vote for Republican candidate and blue counties vote for Democratic candidate in Florida for the 2004 USA presidential election. Figure 8.7 is based on the actual results while Figure 8.8 shows the ASTS interpolated results. We can see that only two out of 67 counties are different in the two figures.

Chapter 9

Conclusion and Future Work

The work can be extended into different directions.

First of all, more data sets should be used to test our spatiotemporal interpolation methods. In the future, we plan to apply our method to other climatic variables like precipitation, average temperatures, or maximum temperatures. We also plan to look at other problems that require a single value as the outcome of the interpolation problem. For example, an aggregate health statistics, such as the number of persons infected with various specific diseases in a state or country would be another natural problem to look at. Another example would be to predict human population changes in a country or worldwide. Both of these are known to be hard problems. For example, there are widely different values for the total number of AIDS cases predicted using different models or the predicted total human population in the world. By improving the estimation accuracy of these and similar types of problems, we can help governments and international health and environmental agencies to be better prepared in the future.

Second, we plan to look into other spatial methods like polynomial regression interpolation and develop spatiotemporal methods based on regression. Encouraged

by results on election prediction, we plan to exploit it by factoring more variables and fine-tune our spatiotemporal interpolation methods.

The results also show that our adaptive 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, such as a candidate's expenditures, gender, incumbency, and the interaction affects of those parameters. However, it is extremely interesting and encouraging that by combining a temporal and a spatial interpolation method, which in themselves are not too sophisticated, already yields prediction values that are more accurate than the results – published in various newspapers in the run-up to the elections – of much more sophisticated prediction systems. Hence our vote prediction system has a significant potential that we plan to exploit by factoring in more variables. Furthermore, as this approach produced both county-level and state-level results, it can be used by election agencies in election data verification for effective government. We can compare the collected election results with the estimates at the county-level and identify possible suspected data when there is significant difference between them.

One interesting topic to explore in the future is the relationship between visualization and interpolation. In many applications there is a need to represent numeric data in a form which has more visual impact [27]. Visualization is a powerful way to facilitate data analysis [49]. 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 [66].

While visualization can be highly effective in the recognition of patterns and trends, poor handling of missing data might lead to misleading data interpretation [14]. There are numerous sources for missing data, such as broken instruments, data-entry errors and data-processing mistakes. Given the intrinsic collection and pre-

sentation 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 [14].

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 [50]. 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 and efficiently computable 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 see patterns that would be too complex to capture mathematically. Occasionally, an interpolation technique may also be detrimental and hide more naturally emerging patterns. Therefore, one may try to generate visualizations both with and without the use of a preprocessing 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.

Finally, this dissertation did not consider periodic spatiotemporal objects, which are considered for example in [51]. It remains an interesting open question how to interpolate periodic spatiotemporal data.

Bibliography

[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] 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.

[3] D. G. Brown, G. Elmes, K. K. Kemp, S. Macey, and D. Mark. Geographic Information Systems. In G. Gaile and C. Willmott, eds. *Geography in America at the Dawn of the 21st Century*, pages 353-375, Oxford University Press, 2004.

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] A. Cuzán, S. Armstrong, and R. Jones. Combining Methods to Forecast the 2004 Presidential Election: The Pollyvote. *mimeo*, University of Pennsylvania, 2005.

[12] E. De Luca and M. Fahle. Learning of interpolation in 2 and 3 dimensions. *Vision Research*, 39(12):2051-2062, 1999.

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

[14] 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.

[15] S. Eberly, J. Swall, D. Holland, B. Cox, and E. Baldridge. Developing Spatially Interpolated Surfaces and Estimating Uncertainty. Environmental Protection Agency. Report No.(s): PB2005-103146, EPA-454/R-04-00, 2004.

[16] J. K. Eischeid, P. A. Pasteris, H. F. Diaz, M. S. Plantico, N. J. Lott. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *Journal of Applied Meteorology*, 39(9):1580-1591, 2000.

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

[18] J. Gao and P. Revesz. Differences in the Spatio-Temporal Interpolation between Plain and Mountainous Regions. In *Proc. of the Nebraska Academy of Sciences*, vol.115, pages 25, 2005.

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

[20] J. Gao and P. Revesz. Voting Prediction Using New Spatiotemporal Interpolation Methods. In *Proc. of the 7th Annual International Conference on Digital Government Research*, pages 293-300, 2006.

[21] J. Gao. Adaptive Interpolation Algorithms for Temporal-Oriented Datasets. In *Proc. of the 13th International Symposium on Temporal Representation and Reasoning*, IEEE Press, pages 145-151, 2006.

[22] J. Gao and P. Revesz. Visualization of Temporal-Oriented Datasets. In *Proc. of the International Conference on Geometric Modelling and Imaging*, IEEE Press, pages 57-62, 2006.

[23] J. Gao and P. Revesz. Adaptive Interpolation Methods for Spatiotemporal Data. To be submitted to *GeoInformatica*.

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

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

[26] M. F. Hutchinson. Interpolation of mean rainfall using thin-plate splines. *International Journal of GIS*, 9:385403, 1995.

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

[28] J. Jaffar and J. L. Lassez. Constraint logic programming. In *Proc. of 14th ACM Symposium on Principles of Programming Languages*, pages 111119, 1987.

[29] K. Johnston, J. M. V. Hoef, K. Krivoruchko, and N. Lucas. *Using ArcGIS Geostatistical Analyst*. ESRI Press, 2001.

[30] P. C. Kanellakis, G. M. Kuper, and P. Revesz. Constraint query languages. In *Proc. of ACM Symposium on Principles of Database Systems*, pages 299313, 1990

[31] P. C. Kanellakis, G. M. Kuper, and P. Revesz. Constraint query languages. *Journal of Computer and System Sciences*, 51(1):2652, 1995.

[32] G. M. Kuper, L. Libkin, and J. Paredaens, editors. *Constraint Databases*, Springer-Verlag, 2000.

[33] N. S. Lam. Spatial interpolation methods: A review. *American Cartography*, 10:129-149, 1983.

[34] G. Langran. *Time in Geographic Information Systems*, Taylor and Francis, London, 1992.

[35] D. R. Legates and C. J. Willmont. Mean seasonal and spatial variability in global surface air temperature. *Theoretical Application in Climatology*, 41:11-21, 1990.

[36] D. Leip. Dave Leip's Atlas of U.S. Presidential Elections. http://www.uselectionatlas.org, 2005.

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

[38] 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.

[39] B. Li, R. W. McClendon, and G. Hoogenboom. Spatial Interpolation of Weather Variables for Single Locations Using Artificial Neural Networks. *Transactions of the ASAE*, 47(2):629-637, 2004.

[40] 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.

[41] L. Li and P. Revesz. The Relationship among GIS-Oriented Spatiotemporal Databases. In *Proc. of the 3rd National Conference on Digital Government Research*, pages 375-378, 2003.

[42] L. Li. Spatiotemporal Interpolation Methods in GIS. *Ph.D Thesis*. University of Nebraska-Lincoln, 2003.

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

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

[45] B. A. Malmgren and A. Winter. Climate Zonation in Puerto Rico Based on Principal Components Analysis and an Artificial Neural Network. *Journal of Climate*, 12(4):977-985, 1999.

[46] A. Matera and F. G. B. De Natale. Video error concealment using spatio-temporal interpolation with snakes. In *Proc. of the 1st International Symposium on Control, Communications and Signal Processing*, pages 83-86, 2004.

[47] 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.

[48] E. J. Miller. Towards a 4D GIS: Four-dimensional Interpolation Utilizing Kriging. In Z. Kemp, editor, *Innovations in GIS 4*, pages 181-197, Taylor & Francis, London, 1997.

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

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

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

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

[53] B. Ripley. *Spatial Statistics*. New York: John Wiley & Sons, 1981.

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

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

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

[57] B. A. Smith, R. W. McClendon, and G. Hoogenboom. Improving Air Temperature Prediction with Artificial Neural Networks. *International Journal of Computational Intelligence*, 3(3):179-186, 2006.

[58] S. E. Snell, S. Gopal and R. K. Kaufmann. Spatial Interpolation of Surface Air Temperatures Using Artificial Neural Networks: Evaluating Their Use for Downscaling GCMs. *Journal of Climate*, 13(5):886-895, 2000.

[59] C. Stallings., R. L. Huffman, S. Khorram, and Z. Guo. Linking Gleams and GIS. In *Proc. American Society of Agricultural Engineers*, 1992.

[60] 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.

[61] Y. Tremblay, S. A. Shaffer1, and et al. Interpolation of animal tracking data in a fluid environment. *Journal of Experimental Biology*, 209:128-140, 2006.

[62] C. Vazquez, E. Dubois, and J. Konrad. Reconstruction of nonuniformly sampled images in spline spaces. *IEEE Transactions on Image Processing*, 14(6):713-725, 2005.

[63] D. Watson. The natural neighbor series manuals and source codes. *Computers and Geosciences*, 25(4):463-466, 1999.

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

[65] E. W. Weisstein. Spherical Coordinates. From *MathWorld*–A Wolfram Web Resource.

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

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

[68] http://election.dos.state.fl.us/

[69] http://www.cnn.com/2004/ALLPOLITICS/10/25 /florida.poll/