Data quality issues in visualization

Alex Pang, Jeff Furman

Computer and Information Sciences Board
University of California, Santa Cruz, CA 95064

Wendell Nuss

Department of Meteorology
Naval Postgraduate School, Monterey, CA 93943

ABSTRACT

Recent efforts in visualization have concentrated on high volume data sets from numerical simulations and medical imaging. There is another large class of data, characterized by their spatial sparsity with noisy and possibly missing data points, that also need to be visualized. Two places where these type of data sets can be found are in oceanographic and atmospheric science studies. In such cases, it is not uncommon to have on the order of one percent of sampled data available within a space volume. Techniques that attempt to deal with the problem of filling-in-the-holes range in complexity from simple linear interpolation to more sophisticated multiquadric and optimal interpolation techniques. These techniques will generally produce results that do not fully agree with each other. To avoid misleading the users, it is important to highlight these differences and make sure the users are aware of the idiosyncrasies of the different methods. This paper compares some of these interpolation techniques on sparse data sets and also discusses how other parameters such as confidence levels and drop-off rates may be incorporated into the visual display.

1 INTRODUCTION

Most visualization algorithms involve an interpolation or filtering step. These can be seen in techniques ranging from iso-surface extraction, ray-casting for volume rendering, to splatting and particle tracing. These operations are usually taken for granted since results are often quite acceptable when these visualization algorithms are applied to very dense data sets such as those obtained from medical imaging or computational fluid dynamics. However, when dealing with sparse data sets, the basic assumption of continuity or homogeneity between sample points may not be valid anymore. The situation is compounded when data sampling is not sufficient to accurately capture the physical phenomena happening between data points.

This paper is concerned with the effects of different interpolation strategies when visualizing sparsely sampled data. Sparsity of data may arise due to the physical constraints of data collection. Among these are the inaccessibility and excessive interference introduced by the sensors in the field they are trying to measure. For example, in invasive heart potential measurements or wind tunnel measurements, the presence of too many sensors may influence the readings. Another common reason given for sparse data falls under financial constraints. For example, the cost of populating every square mile of the country with a weather station (often called met-stations) is too high. In fact, in some instances, it would also be difficult to scientifically justify (e.g. over a homogeneous terrain). On the other hand, this may not be enough resolution for studies dealing with smaller time and space scales.

Another characteristic of sampled data is their imperfect nature. When dealing with measured data, one often have to deal with noisy (spurious or missing) data. The measuring devices may also drift and need to be recalibrated regularly. Some instruments may have different accuracy characteristics depending on angle and distance, and may also have a sharp drop-off in reliability. Data readings may also be affected by variability due to environmental conditions

when the data was taken (e.g. rain, fog, etc.).

Faced with the problems of sparse and imprecise data, it is hard to form an accurate picture of the world that we are observing. Nevertheless, a number of methods exist which try to compensate for the inadequacies in the available data. Making certain assumptions about the homogeneity data, these methods try to make up for the often large gaps in the data. Other, more elaborate, methods bring into consideration confidence levels of various sensors.

The rest of the paper is organized to give a brief overview of different interpolation techniques and the application domain in the environmental sciences. We then present and examine two classes of interpolation strategies (Shepard's interpolation and Hardy's multiquadrics). Finally, we consider slight modifications to these strategies to include uncertainty and drop-off rates.

2 BACKGROUND

Since data is not available everywhere, one has to resort to interpolation or approximation to fill in the missing areas. These methods have roots from different areas: statistical data analysis and approximation theory where the most commonly encountered would be least squares approximation; and surface modeling with splines and different forms of basis functions. With sparse data sets, one must be careful when applying these methods as there may be some physical dynamics going on between data points that are not incorporated into these general interpolation methods. It is thus important to understand the nature and applicability of the different methods.

One can categorize the numerous interpolation and approximation alternatives either by the effects of the algorithms or by the methods employed by the algorithms. Below, we summarize the effects-based taxonomy used by Schumaker[1] on the problem of finding a function f which reasonably approximates F from a given set of points $F_i = F(x_i, y_i)$, $i = 1, 2, ..., N$, located in the (x,y) plane domain D. Note that this could be extended to higher dimensions as well. Schumaker grouped an extensive but non-exhaustive list of methods to five categories: (a) global interpolation, (b) global approximation, (c) local interpolation, (d) local approximation, and (e) hybrid (two-stage) methods. Interpolation methods will construct the function f such that it fits the data exactly at the given points, while approximation methods will only approximately fit the data at the same points. All the data points will contribute to the construction of the function f in global methods, while only the neighboring data points will contribute to f in local methods. In general, global methods involve solving a large linear system while local methods solve a possibly large number of smaller systems of equations. Many of the global methods can be made local by partitioning the data space into smaller subsets and taking care at partition boundaries. Schumaker also recommends the use of interpolation schemes when the data points are known to high precision and approximation schemes when data are subject to inaccurate measurements or errors. Below is a short list of methods falling under the different categories:

1. *Global interpolation.* List includes polynomial interpolation of scattered and gridded data, spline interpolation of scattered and gridded data, and the Shepard's method for arbitrarily space data.

2. *Local interpolation.* Triangular and rectangular partitioning strategies over scattered and regular grids. Parametric representations (e.g. Coon's surfaces). Localized Shepard's methods.

3. *Global approximation.* Polynomial least squares or multi-dimensional regression, spline smoothing of scattered and gridded data, and discrete and continuous least squares.

4. *Local approximation.* Adaptive patch methods and direct local methods or quasi-interpolants.

5. *Two-stage methods.* An approximation g is calculated from scattered data during the first stage which is then refined to the surface f in the second stage. Combinations include interpolation-interpolation, approximation-interpolation, and approximation-approximation.

Any of these methods can be used, with varying degrees of success, to fill in the space between sparse data points. The problem with most of these methods is that few of them take into consideration anything about the dynamics of

the system that they are fitting. Obviously, a domain-dependent or a physically based interpolation method would provide a more believable picture. The next section describes our problem domain.

3 ENVIRONMENTAL DATA

The application we are interested in is the nowcasting of meteorological and oceanographic events on the scale of the Monterey Bay.[2] In contrast to forecasting, nowcasting is more near term. At the limit, we are interested in finding the current picture of the environment based on the limited number of sensor readings.

There are two complicating factors that make this problem more difficult. The first is the spatial scale of the domain. Unlike global climate modeling or weather forecasting where the focus of interest is in large scale structures, regional or local models need to resolve the fine structures that are important at this scale. The second difficulty is the immediacy of the forecast. We cannot afford to run forecasting models into the future when we need to know what the current situation looks like. In addition, most of the forecasting models require the past or current pictures to initialize them. There is actually some work in the area of data assimilation which attempts to continuously update simulation runs with the most recent data gathered from sensors. However, the problem at hand is how does one fill in the void of space where there are no measured data? Do scientists sit back and stare at the sparse data and form some kind of mental interpolation? If so, what kind and can we formalize it?

To understand this process, we also need to know the characteristics of the data as well as the sensors. For example, some fields such as pressure tend to be more uniform and vary smoothly over space. Changes to these fields normally do not happen suddenly and take time to develop and spread. On the other hand, wind fields tend to be more erratic. It can vary from place to place and can change direction and speed at any instant. Thus, depending on the data field, it can be assumed to be either smooth or bumpy. But in all cases, the data is assumed to be continuous and differentiable. These environmental data are obtained from a variety of sensors such as met-stations and buoys which obtain in-situ measurements. There are also a number of more exotic instrumentations such as vertical wind profilers and codars (which measures the ocean surface current). Measurements of the environment are constrained by battery supply, limited range/accuracy of equipment, prevailing conditions, drift in equipment calibration, etc. It is therefore common practice to attach a range or uncertainty factor to a reading.

Thus, the dilemma faced by the scientist is to integrate all these factors into a coherent mental picture of the current scenario. The task of the visualization specialist is to aid the scientist analyze the data set. At first glance, it is quite easy to generate an interpolated image of the sparse data sets. However, this may contain misleading information or artifacts that are not present in the data. Furthermore, there are several surface fitting methods to choose from. Which method provides the best estimate for the particular type of data field?

The next section studies two methods as applied to this problem domain in more detail. While the methods can be extended to 3D, the data are assumed to be located at sea level and hence on a 2D plane.

4 SURFACE FITTING

We study and compare the Shepard's methods[1,3,4] as well as some variations of Hardy's multiquadric methods.[5–7] Both methods have been around for over two decades and used in areas such as topography, geography, meteorology and computational fluid dynamics, but are relatively unknown in the visualization community. These two methods fall under Schumaker's global interpolation category but can also be made local. In the following sections, we present the descriptions and performance characteristics of both methods on different data sets and sampling distributions.

4.1 Shepard's interpolation

Perhaps the most familiar member of the Shepard's interpolation is the inverse square technique. The idea with this method is that for each unknown $f(x, y)$ in the field that is to be estimated, a weighted average is performed from

all of the known data points F_j, $j = 1, 2, ..., N$. The weight contribution of each data point is calculated as the inverse square of the distance from (x, y) to the data point. Using this rule, a data point will have the strongest influence to the function f around its neighborhood.

Shepard's formulation generalizes the inverse square method to include different powers of distance.

$$f(x,y) = \begin{cases} F_j & at\ data\ points \\ \frac{\sum_{j=1}^{N} F_j / d_j^u}{\sum_{j=1}^{N} 1/d_j^u} & otherwise \end{cases} \quad (1)$$

The value of u plays an important role in the shape of the surface specially in the vicinity of the data points. When $0 < u < 1$, $f(x, y)$ has cusps at the data points. These cusps turn into corners when $u = 1$ and flatten out when $u > 1$.[3] Another useful property of Shepard's interpolation is that the surface stays within the extrema of the data points.[1] That is, $\min(F_j) \leq f(x, y) \leq \max(F_j)$ for $j = 1, 2, ..., N$. Figure 1 illustrates the behavior of Shepard's interpolation using different distance power values.

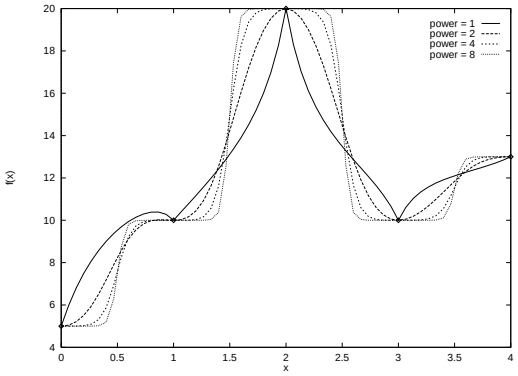

Figure 1: Effects of different power values on the Shepard's interpolation curve.

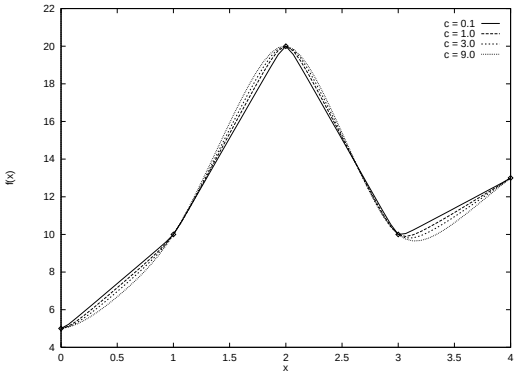

Figure 2: Effects of different c values on the multiquadric interpolation curve.

4.2 Hardy's multiquadrics

Another popular interpolation technique for sparse data sets is Hardy's multiquadric interpolation. The interpolated surface is defined by a sum of weighted radial hyperbolic basis functions. This is formalized in the equation below.

$$f(x,y) = \sum_{j=1}^{N} \alpha_j Q(x, y, x_j, y_j) \quad (2)$$

where the radial basis functions Q are defined as:

$$Q(x, y, x_j, y_j) = \sqrt{(x - x_j)^2 + (y - y_j)^2 + c^2} \quad (3)$$

c is an input parameter which can influence the shape of the surface. Variations on Hardy's multiquadrics arise primarily from how the basis functions are defined. For example, a simple change of removing c, would change the basis functions from hyperboloids to cones. Equation 2 can be extended to higher dimensions by simply incorporating the extra dimensions into Q. To solve equation 2, the coefficients α_j must be determined by solving a set of linear equations

$$F_j = f(x_i, y_i) = \sum_{j=1}^{N} \alpha_j Q(x_i, y_i, x_j, y_j) \quad (4)$$

Intuitively, larger c values in Q give rise to flatter basis functions and hence the interpolated surface will also be flatter or smoother. Figure 3 shows two Q basis functions with different c values. Figure 2 shows the behavior of the multiquadric interpolation using different c values.

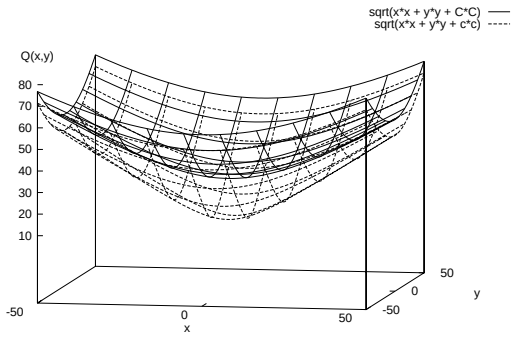

Figure 3: The flatter Q basis function on top has a larger $C = 30$ value compared to the $c = 10$ surface below.

4.3 Evaluation

The two methods discussed above were examined and compared to determine their strengths and weaknesses. To do this, we developed a test suite of four data sets, a set of parameters for each method, and different number and distribution of sampling points. The different data sets represent different distributions of data values and are included to see whether a method performs better for a given type of distribution. These data sets are described in more detail below. Each method has a free parameter that can be tuned to the range or distribution of data values. We test both methods using different sets of these parameters values. Varying the number and distribution of sampling points can greatly influence the performance of both methods. Hence, we also study how the methods perform as we vary the number of sample points.

The test procedures involved sampling each data set with a fixed set of points and guessing the rest of the field based on the values at those points. Then, comparisons are made between the interpolated field and the target data field in the form of difference images and root-mean-square (RMS) error calculations. Both methods were subjected to this procedure using different number of sampling points and different interpolation parameters.

Test data

We used four data sets: two were synthetically created while the other two were from numerical simulations. The synthetic data sets were created from an arbitrary combination of sine and cosine functions. The simulation data sets come from the NPS-NRL mesoscale model output of a CRAY Y-MP EL98 and centers around California. They contain 103 (East-West) by 91 (North-South) grid points spaced 1/6 degrees (approximately 18.5 km) apart on a Mercator projection that runs from 28N-43N and from 130W-113W. We further classify the data sets as *smooth* or *bumpy*. Below is a brief description of each:

1. *smooth synthetic:* Dimensions: 100x100. $f(x, y) = 10sin(\frac{x}{10}) + 7cos(\frac{y}{7})$; $x, y = 0..20$.
2. *bumpy synthetic:* Dimensions: 100x100. $f(x, y) = 10sin(\frac{x}{4}) + 15cos(\frac{y}{5}) + 7cos(\frac{1}{3}x + \frac{1}{2}y)$; $x, y = 0..20$.
3. *smooth simulation:* Dimensions: 103x91. Sea level pressure.
4. *bumpy simulation:* Dimensions: 103x91. Temperature.

Shepard's interpolation tests

In Shepard's method, each interpolated point is determined by calculating a distance weighted average of all the sample points available. As can be noted in Figure 1, lower distance power values tend to emphasize the sample

points, while higher distance power values produce smoother approaches near the sample points. It can also be noted that as the distance power is increased further, the gradient halfway between sample points tend to increase. The corresponding effects on 2D images can be seen in Figure 4. In particular, peaks are noticeable near sample points for lower distance powers and boundaries between "basins" are noticeable for higher distance powers. Hence, the sample value becomes the dominant contributor within its basin when the distance power is high. This suggests some strategies for incorporating uncertainty parameters in the interpolation (see section 5). Figure 5 shows how Shepard's interpolation performs on different data sets. It also suggests that while using higher distance power values may reduce the RMS error, using the inverse square may be sufficient for most purposes.

Multiquadric interpolation tests

One can see that as the multiquadric parameter c is increased, the interpolated curves (see Figure 2) become smoother. However, as c is driven higher, the interpolated curves tend to under or overshoot in the vicinity of the sample points. This may lead to a higher RMS error as evidenced in Figure 6.

The selection of c depends on two things: (a) the absolute magnitude and (b) the relative magnitude (gradient) of data values. Typically, one would choose a larger c value for data sets with large values. For example, you would choose a smaller c if data values range from 0 to 10 than if they range from 900 to 910. However, too high a c value would also cause the Q function (Equation 3) to be dominated by c. This would in turn cause the matrix, representing the system of linear equations, to be poorly conditioned. Hence, we have introduced an intermediate step where all the sample points are *shifted* down by $\min(F_j)$, the smallest sample value. This allows us to use a smaller c value. After an interpolated surface is fitted on the down-shifted values, the surface is raised back up by the same amount.

Another important factor in choosing c is the relative magnitude of the sample values. If the ratio of any two sample values is close to one, then c must be higher than if the ratio was further from one. This implies that smoother data require higher c values. Figure 8 shows higher c values consistently giving better guesses for the smooth surface. We can make the same observation when looking at the shape of the Q basis functions in Figure 3. So, sensor values of 1000, 1050 and 1200 need a much larger constant to produce a comparable surface than if the sensor values were 100, 150 and 300. This is the case, even though the shifted (subtract 900 from each of the larger values) sensor values have the same absolute magnitudes.

Recall that as c gets larger, the basis function gets flatter. While this may give us a smoother surface, it also has the drawback of not being able to handle steep gradients in the data (see Figures 9 and 11). Therefore, the statement that multiquadrics perform better in medium to steep gradients[6] need to be qualified with the need for an appropriate value of c.

Comparison

We compare the results qualitatively through images in and quantitatively using RMS values. The RMS values are based on the median of 11 different runs. Each run is uses the same number of sampling points but distributed differently. The median values are used instead of the average since the interpolated surfaces are also very sensitive to the clustering of the sampling points. The comparisons can be seen in Figures 8 – 11 with the multiquadrics with appropriate c values doing generally better than Shepard's. These figures also provide some guidelines for the relative benefit of adding more sampling points. Unfortunately, it does not help in telling us where to put additional sampling points. A qualitative look at the interpolated surfaces (Figures 4 and 7) may help.

The RMS values assign an aggregate score to an entire image but does not say how well the method performed in local regions near sample points. Looking at the images in Figures 4 and 7 gives us a better understanding of where the methods obtained a good/poor match in local regions. For example, it is clear that the multiquadrics interpolation produce smoother images with less artifacts (e.g. basins) than the Shepard's interpolation. Another observation is that when sampling points happen to cluster near each other, they collectively form a strong weight in that vicinity using Shepard's. These points act as if they were a single point with many times the weight of the other sampling points resulting in distorted images. In contrast, output from multiquadric interpolation is more independent from sampling locations.

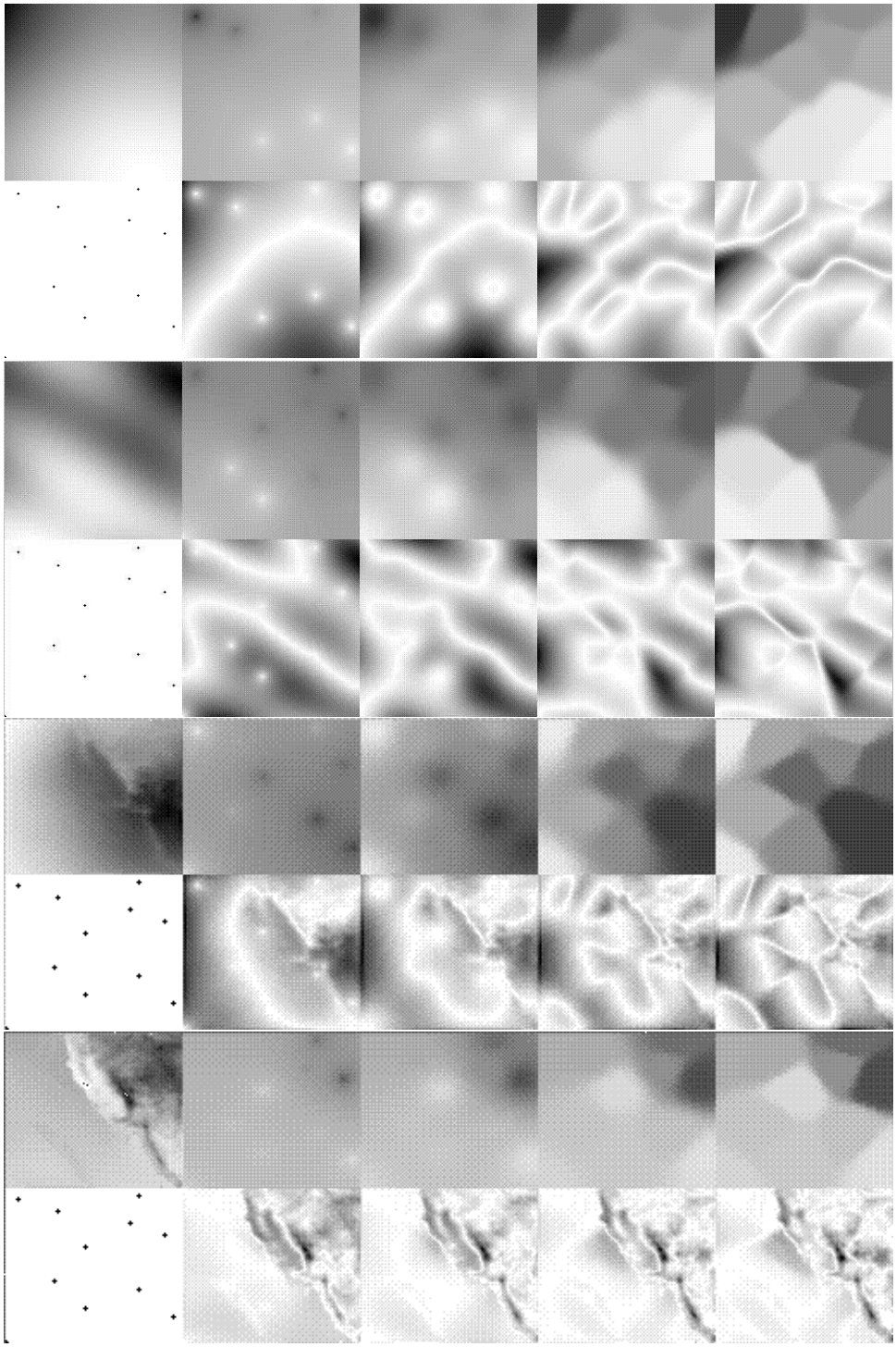

Figure 4: Top to bottom: Interpolated (guesses) and difference images of smooth, bumpy, pressure and temperature data sets using Shepard's interpolation. Target (correct) fields and locations of sampling points are shown on first column. Columns 2-5 use distance power values of 1, 2, 4, 8 respectively.

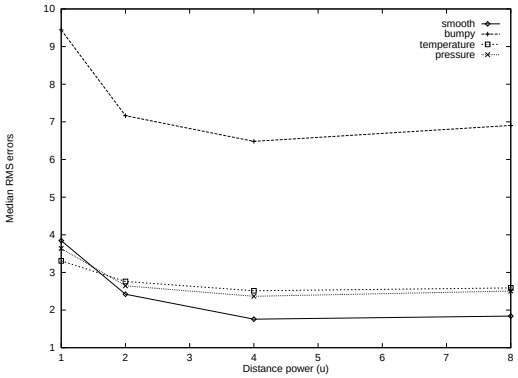

Figure 5: Effects of adjusting Shepard's power parameter on different data sets. Data points are the median values of RMS errors from 11 different distributions of 11 sampling locations. Graph shows need for selecting different parameters for different data sets.

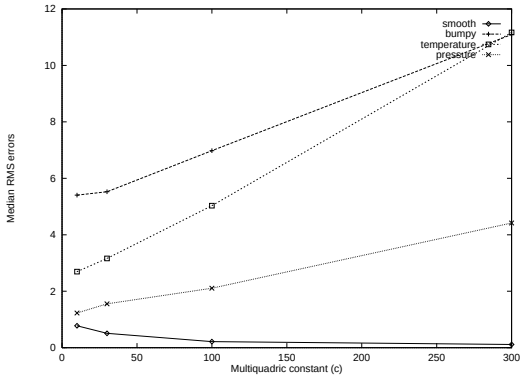

Figure 6: Effects of adjusting MQ parameter on different data sets. Data points are the median values of RMS errors from 11 different distributions of 11 sampling locations. Graph shows need for selecting different parameters for different data sets.

5 SURFACE FITTING WITH UNCERTAINTY

Compounding the problem of forming a mental image of the state of the system from a small set of scattered sampling points, is the fact that values from the samples may not be very accurate. As mentioned in section 3, data values may be degraded due to sensor range and conditions in which reading was taken. Furthermore, the utility of the reading may drop-off rapidly away from the sensor if the field is naturally more dynamic (e.g. wind). In this section, we describe some extensions to Shepard's interpolation that allow data quality parameters to be incorporated.

Data quality or confidence level has an opposite relationship to uncertainty. When data quality is high, uncertainty is low and vice versa. One way to incorporate uncertainty into existing methods is to attach a weight to each sample point. The higher the weight, the stronger the confidence on the reading. This can be described by the equation below.

$$f(x,y) = \begin{cases} F_j & at\ data\ points \\ \dfrac{\sum_{j=1}^{N} F_j W_j / d_j^u}{\sum_{j=1}^{N} W_j / d_j^u} & otherwise \end{cases} \tag{5}$$

The relationship above does not include how fast confidence levels drop-off with distance. This can be accounted for by replacing the W_j's in Equation 5 with distance reduced weights w_i's described by:

$$w_i = \max(W_j \frac{R_j - d_i}{R_j}, 0) \tag{6}$$

That is, the confidence of sample j is linearly reduced to zero as we move a distance R_j away. And the surface point at $f(x_i, y_i)$ is a distance weighted term of the different w_i's. Note that instead of the linear drop-off in confidence level, the drop-off could be modeled some other way.

Yet another way of incorporating uncertainty into the surface is by allowing the surface to be "lifted" or "lowered" in places where the certainties are not 100%. This would allow a surface to interpolate perfect sample points and only approximate imperfect sample points. One such function is described below, and its effects can be seen in Figure 12.

$$1/(d_j^u + U) \tag{7}$$

Notice that when the uncertainty term is 0, this reduces to Shepard's distance weights. However, as U increases to infinity, the contribution of F_j drops to 0. We relate the uncertainty term U to the weight W using a reciprocal relationship: $U = \frac{1-W}{W}$ (see Figure 12). Note that to be effective, the value of U must be comparable to the value of

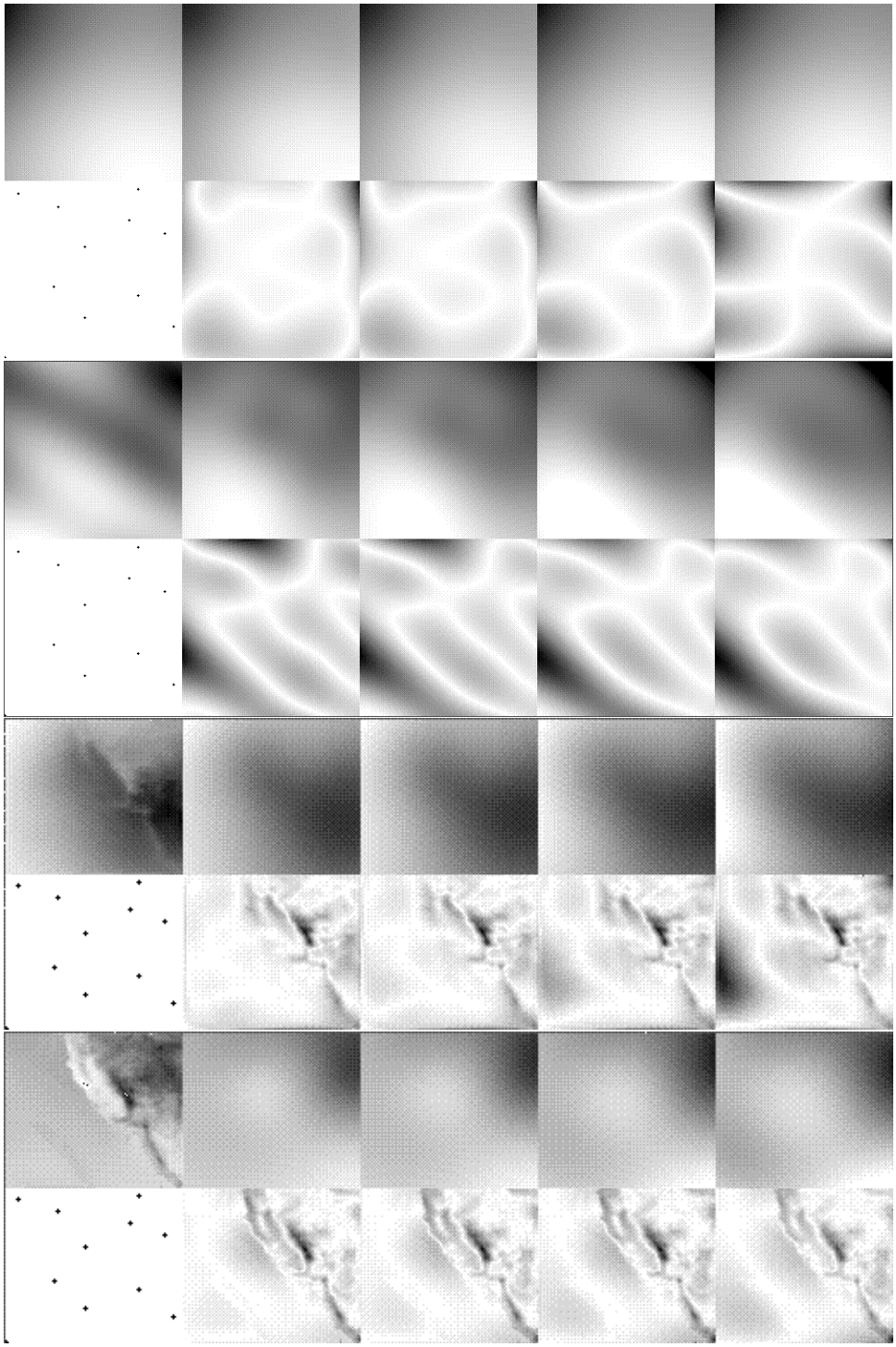

Figure 7: Top to bottom: Interpolated (guesses) and difference images of smooth, bumpy, pressure and temperature data sets using multiquadric interpolation. Target (correct) fields and locations of sampling points are shown on first column. Columns 2-5 use constant c values of 10, 30, 100, 300 respectively.

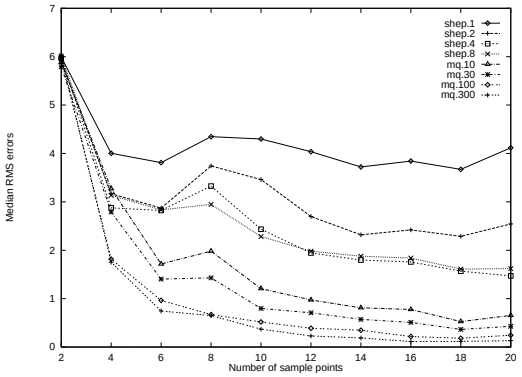

Figure 8: Comparison of methods on smooth data. Multiquadrics as a group performed better than Shepard's. Multiquadric with $c = 300$ gave the best result.

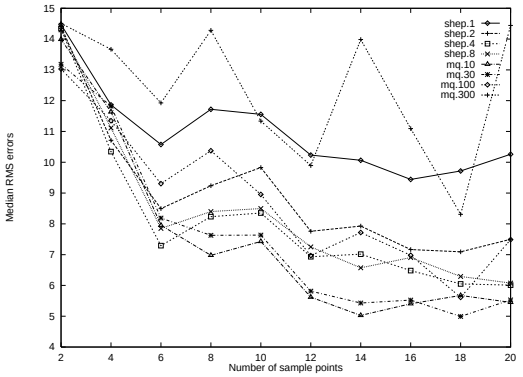

Figure 9: Comparison of methods on bumpy data. Multiquadric with $c = 10$ or 30 gave the best results. But multiquadric with $c = 300$ gave the worst result.

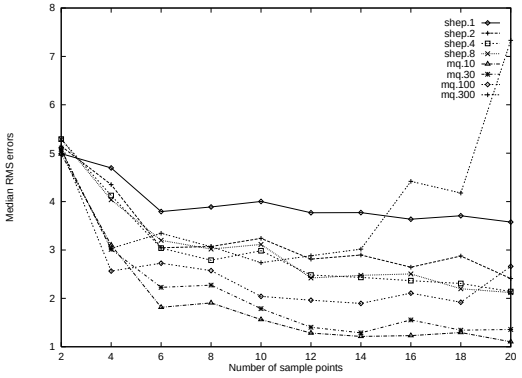

Figure 10: Comparison of methods on pressure data. Except for multiquadric with $c = 300$, the other multiquadric interpolations gave better results than Shepard's.

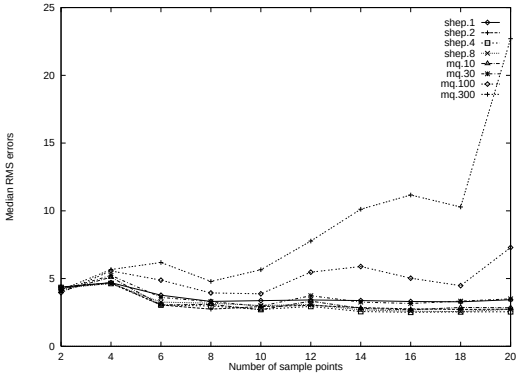

Figure 11: Comparison of methods on temperature data. Multiquadrics with $c = 100$ and $c = 300$ gave the worst results. The rest gave comparable results.

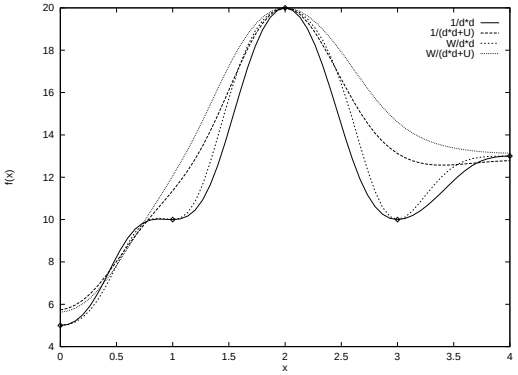

Figure 12: Different ways of incorporating uncertainty to Shepard's inverse square interpolation. Solid curve is from standard inverse square. Note that some curves are non-interpolating when confidence level is less than 100%.

d^u. Since the distance terms will be proportional to the dimensions of the entire surface, we use a modified version for the images in Figure 13. For images, we use $U = Xdim(\frac{1-W}{W})$.

The three strategies for including uncertainty and drop-off rates are interchangeable. Figure 13 compares three combinations using power $u = 4$: (a) Equation 5, (b) Equation 7, and (c) combination of the two (i.e. $W_j/(d_j^u + U)$). The bottom row shows the same surfaces with uncertainty but with a grid in the background. The interpolated weight surface is mapped to transparency such that areas with higher confidence cover the grid better than areas with lower confidence levels. Specifically, the amount of transparency at location i is determined by: $t_i = 1 - \max(w_i)$ over all N sampling points. This allows areas where no information is available or where the confidence level is zero to be become transparent.

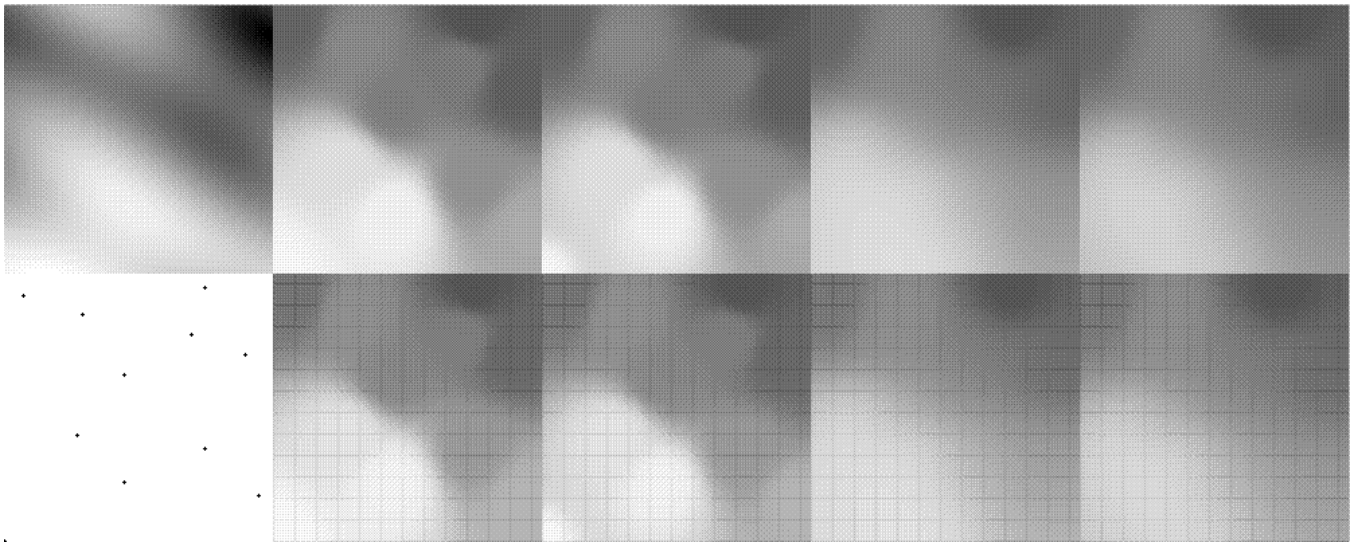

Figure 13: Surface fitting with uncertainty. Interpolated surfaces use distance power 4. Columns show the target, standard Shepard's, weights using Equation 5, weights using Equation 7 and combined U and W weights respectively. The drop-off distances R and the weights W of the sampling points are increasing from left to right, bottom to top. The bottom row maps uncertainty to transparency. Areas where the grid shows through are places where uncertainty is higher.

6 CONCLUSION

The dilemma facing scientists who need to do nowcasting, is finding a physics based interpolation model suited for a particular scale and locality. In its absence, the problem is which interpolation method to use and which results to believe. This paper studied and compared two popular interpolation methods used with sparse, scattered data sets. Based on the qualitative results, users will hopefully become more aware of each method's behavior and potential surface artifacts. Between the two methods, multiquadric interpolation gives better performance when provided an appropriate c value.

We have also introduced slight modifications to the Shepard's method to account for uncertainty and drop-off rates. Some displays using transparency to represent the level of confidence in data are also presented.

7 ACKNOWLEDGEMENT

We would like to thank Dr. Teddy Holt and Dr. Paul Hirschberg for kindly providing us the temperature and pressure data sets used in the comparisons. This work is partly funded by ONR grant N00014-92-J-1807.

8 REFERENCES

[1] L.L. Schumaker. Fitting surfaces to scattered data. In C.K. Chui G.G. Lorentz and L.L. Schumaker, editors, *Approximation Theory II*, pages 203–268. Academic Press, 1976.

[2] D.E. Long *et al.* REINAS: Real-time environmental information network and analysis system: Concept statement. Technical Report CRL-93-05, UCSC, January 1993.

[3] R.E. Barnhill, R.P. Dube, and F.F. Little. Properties of Shepard's surfaces. *Rocky Mountain Journal of Mathematics*, 13(2):365–382, 1983.

[4] R.E. Barnhill and S.E. Stead. Multistage trivariate surfaces. *Rocky Mountain Journal of Mathematics*, 14(1):103–118, 1984.

[5] R.L. Hardy. The application of multiquadric equations and point mass anomaly models to crustal movement studies. Technical Report NOS 76 NGS 11, NOAA, November 1978.

[6] E.J. Kansa. A comparative study of finite difference and multiquadric schemes for the euler equations. *Simulation*, pages 180–183, 1988.

[7] W.A. Nuss and D.W. Titley. Use of multiquadric interpolation for meteorological objective analysis. *submitted to Monthly Weather Review*, 1993.