

# **Interpolation and error propagation models for raster datasets based on map algebra in master planning of Markazi province**

Hooshang Eivazy

Faculty member of Iran university of Science and Technology

Tel (+98) 861-3675604

[h\\_eivazy@yahoo.com](mailto:h_eivazy@yahoo.com)

Mohamad Mehdi Khoshgoftar

Ms.c students of Khajenasir University

Tel (+98) 918-3637383

[Mahdimk61@yahoo.com](mailto:Mahdimk61@yahoo.com)

## **Abstract**

Nowadays, majority of spatial and non-spatial data that is commonly used in public and private sectors, for example topography and other related attributes such as soil PH etc, is generated via the traditional software. Despite the error coherent with the nature of the interpolation, in so many cases, error volume exceeds the eligible error and fails the results. In addition lack of a suitable criteria, doesn't let to evaluate dataset results. In the other hand, provided data, is the source for many basic studies in countries. These studies significantly form the infrastructure studies and are the milestone for sustainable development in countries. Therefore any long, mid and short term planning is designed by direct effect of such studies.

Hence to avoid any misdeed or deflection in country plans, it is necessary to clean, evaluate and even override spatial data especially for those which are generated via the interpolation models.

The paper focuses on data exploratory analysis and methods to filter data based on powerful tools of ArcGIS. Besides, regarding to extension of deploying spatial data analysis and map algebra models, a new solution has been proposed to derive and compute error propagation models and assurance raster dataset surface at the certain confidence level due to analysis model for raster datasets.

## **Introduction**

Nowadays the large spectrum of spatial, non-spatial, quantitative and qualitative data are prepared in the different part of topographical, geological, meteorological and ecological, disaster management, environment and etc by interpolation. The provided data via the interpolation, are settled as the basis of a lot of substructure or executive and engineering studies. Since the interpolation is essentially linked with error, these errors are propagated in calculations and processes. It causes deflection of responses. These errors may cause the wrong estimation in the initial studies and even the incorrect designs in the executive projects, which finally lead to financial damage.

## **Preparation of topographic data**

Topographic data and its results such as DSM and DEM are one of the most utilized products of interpolation. At present, there are three current main methods to collect

the required data for interpolation: Aerial photogrammetry, Ground method and Laser Scan. Each method has its own weak and strong points that are mentioned below:

**Aerial Photogrammetry:** This method is based on 1:10000 to 1:40000 photography and direct extraction of contours. Altimetric precision is depended on straw operations precision and also extractor operator precision that varies from a few centimeters to some meters. The process of interpolation is done on the whole or a part of contour points in this method. This method has the suitable altimetric precision. On the other hand, it leads to make the surfaces, have a good relative correspondence with the environment. However, its absolute precision is low and irrespective the high expanses of aerial photogrammetry and its needful time, considering the high rate of the level of operator actions. (In addition operators are more satisfied with taking single height point). On the other hand, the automatic algorithm in digital photogrammetry are not able to generate altimetric model with the suitable precision needed for users.

**Ground surveying:** This method is based on surveying the altimetric points as stadimetry and or digital from the area. The most important property of this method is how economical the ground survey is for the limited areas. However, putting the low price and simplicity aside, the method is not economical in vast areas. In spite of generating the vertical points with the high precision in a small search radius, the resulting altimetric model initiates the less relative precision in comparison with aerial model due to the fewer number of points.

**Ground Laser Scan:** In this method that it is based on taking extra points X,Y,Z by the laser pulses. It is possible to considerably decrease costs and the time. Considering the low radius of search of the operation system, using ground type will no longer be economical. The height precision of the result model points, according to be unsupervised the taking method, will essentially be lower than the ground methods.

### **The altimetric data modeling**

In addition to method of taking the required altimetric points for interpolation, main part of errors are exerted in the modeling process. For example surveying companies, often make some sharp contours and data by taking advantage of the softwares like Inroad, SDRMap, Land, Surfer after taking the altimetric points based on ground surveying. They turn to smoothing the contours with a manual and time consuming cartographic process. In this way, model is no longer useful for processing system like GIS, considering that the contours are one of the methods of flaunting the heights. Secondly, what is obtained as the model gradually keeps distance from the reality and by no means would be a virtual model of reality. This is completely obvious, particularly in difference between the results of estimating the soil volume contributing the current commercial software and their actual measures.

These problems mainly are arisen from unsuitable sampling and modeling method. Hence we will turn to some of them.

Principally, every quantity like height, temperature, pressure, amount of the rain, soil PH follow a specific trend of the nature quantity generator. Hence, the quantity variation of position is along with the slope of quantity model generator. Therefore, recognition the model generator of the nature quantity will be obligatory for sampling the base points that determine the model parameters. In the next section, first we turn to different models of interpolation and the methods of using these models for simulating the nature of a quantity.

### **Triangulated Irregular Networks (TIN)**

TIN (Triangulation Irregular Networks) datasets can be used to display and analyze surfaces. They contain irregularly spaced points that have x,y coordinates describing their location and a z-value that describes the surface at that point. The surface could represent elevation, precipitation, or temperature. A series of edges join the points to form triangles. The resulting triangular mosaic forms a continuous faceted surface, where each triangle face has a specific slope and aspect.

The triangulation method uses Renka's algorithm to carry out a Delaunay triangulation of the observation points. The purpose is to identify a neighborhood of nearby observation points to be used in the interpolation.

Two points are available to perform the interpolation. The simplest method is a linear interpolation of the vertices of the triangulation. In this case, the barycentric coordinates of the point are used to take a weighted average of three observation values (associated with the three vertices of the triangle). The resulting interpolation function is continuous, but not differentiable across an edge.

The second triangulation method uses a polynomial fit within each triangle in the triangulation. This method involves an intermediate step of fitting a cubic spline on each edge of the triangulation. The interior values are taken as weighted averages of the edge values, based up on the barycentric coordinates of a point. In the intermediate step of fitting the spline, approximations to the gradient at each point must be computed. The user has the option to specify which of two gradient estimation methods is to be used. The local method uses only a few neighboring points, while the global method uses all points. The latter method may not be possible if the number of points is larger than a few hundred.

A good method to honor data collected along ridges, drains, and significant break lines of changing slope on the terrain surface.

The initial surface generated will appear peaked at the data points and have angular sharp corners.

Location of the data points on the terrain should be assessed before any application of smoothing is used.

Topology of triangles allows for a variety of applications, such as slope, shading, aspect, and contours.

## **Spline**

The basic form of the minimum-curvature spline interpolation imposes the following two conditions on the interpolation:

- 1) The surface must pass exactly through the data points.
- 2) The surface must have minimum-curvature the cumulative sum of the squares of the second derivative terms of the surface, taken over each point on the surface, must be a minimum.

The basic minimum-curvature technique is also referred to as thin plate interpolation. It ensures a smooth (continuous and differentiable) surface, together with continuous first derivative surfaces. Rapid changes in gradient or slope ( the first derivative) may occur in the vicinity of the data points; hence this model is not suitable for estimating second derivative (curvature).

The thin-plate spline method fits a spline function to the observations. The fitted function agrees with the observation values at the observation points. The coefficients of the functional fit can be stored for later use. For  $n$  observation points, this requires the solution of  $n$  simultaneous equations and the inversion of an  $n*n$  matrix. As a result, a spline fit is limited to a reasonable number of points. In spherekit, if more than 500points are entered, the domain is broken up into overlapping regions, and a

surface fit is performed on each region. Points lying in more than one region are assigned a value that is a weighted average of the multiple estimates.

The thin-plate spline function can be any of the following forms:

$$f(d) = d^2 \log d$$

$$f(d) = d^2 \log d^2$$

$$f(d) = d^2 (\log d - 1)$$

Where  $d$  is the spherical distance.

Spline functions in spatial interpolation are formed by piecing together successive curve segments involving relatively few points at a time and should be closely related to the value being interpolated; they are analytic; and are flexible. When applied to a spline surface, successive rectangular surface patches form a composite surface similar to a patchwork quilt.

Advantageous for dense or rectangular data, can introduce anomalies not in the original surface depending on the interpolation and blending methods used.

## Kriging

Kriging assumes that the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface. Kriging fits a mathematical function to a specified number of points, or all points within a specified radius, to determine the output value for each location. Kriging is a multi step process; it includes exploratory statistical analysis of the data, variogram modeling, creating the surface, and (optionally) exploring a variance surface. **Kriging** is most appropriate when you know there is a spatially correlated distance or directional bias in the data. It is often used in soil science and geology.

Kriging is similar to IDW in that it weights the surrounding measured values to derive a prediction for an unmeasured location. The general formula for both interpolators is formed as a weighted sum of the data:

$$\hat{Z}(s_0) = \sum_{i=1}^N \hat{\lambda}_i z(s_i)$$

where

$Z(s_i)$  is the measured value at the  $i^{th}$  location;

$\lambda_i$  is an unknown weight for the measured value at the  $i^{th}$  location;

$s_0$  is the prediction location;

$N$  is the number of measured values.

There are two Kriging methods: Ordinary and Universal.

Ordinary Kriging is the most general and widely used of the Kriging methods and is the default. It assumes the constant mean is unknown. This is a reasonable assumption unless there is some scientific reason to reject this assumption.

Universal Kriging assumes that there is an overriding trend in the data (for example, a prevailing wind), and it can be modeled by a deterministic function, a polynomial. This polynomial is subtracted from the original measured points, and the autocorrelation is modeled from the random errors. Once the model is fit to the random errors and before making a prediction, the polynomial is added back to the predictions to give you meaningful results. Universal Kriging should only be used when you know there is a trend in your data and you can give a scientific justification to describe it.

## Average Nearest Neighbor Distance

The average nearest neighbor distance measures the distance between each feature centroid and its nearest neighbor's centroid location. It then averages all of these nearest neighbor distances. If the average distance is less than the average for a hypothetical random distribution, the distribution of the features being analyzed are considered clustered. If the average distance is greater than a hypothetical random distribution, the features are considered dispersed. The index is expressed as the ratio of the observed distance divided by the expected distance (expected distance is based on a hypothetical random distribution with the same number of features, covering the same total area). Hence if the index is less than 1, the pattern exhibits clustering; if the index is greater than 1, the trend is toward dispersion.

## Inverse Distance Weighting (IDW)

Inverse distance weighting is the simplest interpolation method. A neighborhood about the interpolated point is identified and a weighted average is taken of the observation values within this neighborhood. The weights are a decreasing function of distance. The user has control over the mathematical form of the weighting function, the size of the neighborhood (expressed as a radius or a number of points), in addition to other options.

Weighting function

The simplest weighting function is inverse power:

$$w(d) = 1/d^p$$

with  $p > 0$ . The value of  $p$  is specified by the user. The most common choice is  $p=2$ . For  $p=1$ , the interpolated function is "cone-like" in the vicinity of the data points, where it is not differentiable.

The neighborhood size determines how many points are included in the inverse distance weighting. The neighborhood size can be specified in terms of its radius (in km), the number of points, or a combination of the two. If a radius is specified, the user also can specify an override in terms of a minimum and/or maximum number of points. Invoking the override option will expand or contract the circle as needed. If the user specifies the number of points, an override of a minimum and/or maximum radius can be included. It also is possible to specify an average radius based upon a specified number of points. Again, there is an override to expand or contract the neighborhood to include a minimum and/or maximum number of points.

## Multi-quadrics

The multi-quadric method fits a set of quadric (i.e. hyperbolic or conical) functions to the observations. The fitted function agrees with the observation values at the observation points. The coefficients of the functional fit can be stored for later use.

The estimate at point  $P$  is:

$$F(b) = \sum_{k=1}^N A_k \sqrt{1 + R^2 - 2R \cos(s_k)}$$

where  $R$  is a user-specified tension parameter and  $s_k$  is the angular distance between the interpolation point and the  $k^{th}$  observation value. The weights  $A^k$  are computed so that the estimated function agrees with the observations at the observation points. For

$n$  observation points, this requires the solution of  $n$  simultaneous equations and the inversion of an  $n*n$  matrix. As a result, a multiquadric fit is limited to a reasonable number of points.

## Wavelet theory and its properties

### Introducing on wavelets

In the furrier analysis, one function or signal are transformed to frequency space with assistance of complex exponential function and integral operator, as one of the spectrum analysis methods and then its different frequencies are separated. After recognizing and separating the frequencies, it is possible to turn to sigma analysis or preparing a suitable filter to eliminate the present noises. Since the base function is as a sinusoidal wave with the infinite amplitude, so this process is not able to reply to this question when or where the present frequencies happen, therefore, it leads us to a transformation with such as ability that called wavelet transformation in which our question can be replied by applying the wavelets as waves with focused energy in time (or location).

### Wavelet transformation

If a function such as  $f(t)$  can be written as a linear combination  $f(t) = \sum a_i \psi_i$  in which  $a_i$ 's are real coefficients called the expansion coefficients and  $\psi_i$ 's are also real functions named the expansion functions, then analyzing, describing and processing signal will be performed more noticeably and capably and this will be the basis of the spectrum analysis. If the set of the expansion coefficients for a function be exclusive, the expansion functions are called a basis for the space including  $f$  and whereas these functions are the base of ortho-normal, it can be written:

$$\langle \psi_k(t), \psi_l(t) \rangle = \int \psi_k(t) \psi_l(t) dt = \delta_{k,l}$$

and

$$a_k = \langle f(t), \psi_k(t) \rangle = \int f(t) \psi_k(t) dt$$

In a furrier expansion, the basis orthogonal functions of the system are as an exponential function and these function are non-orthogonal and as the set of polynomials in tailor expansion.

In the wavelet expansion, transmission is performed from the argumant space to the frequency space by a set of wavelets as the base functions, as the following:

$$f(t) = \sum_k \sum_j a_{j,k} \psi_{j,k}(t)$$

In this formula,  $\psi_{j,k}$ 's are the same wavelets and with the same perception as the furrie transformation, the expansion's coefficients are in fact the signal image in the frequency space (by the integral operator) and are called the discrete wavelet transformation of the  $f$  function (DWT) and the above formula is called the inversed discrete wavelet transformation (IDWT). Therefore, it is possible to consider to the both cases of location and frequency (scale) in the wavelet analysis simultaneously.

The set of functions that can be used as wavelet are not exclusive and after providing the present conditions if necessary, there will be a number degree of freedom design the base functions but anyhow. In any wavelet system, the base functions set of the system are achieved owing to scale exerting and also displacement of a primitive generator function called the mother wavelet:

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k)$$

### The scale and wavelet functions

In expanding a function by Taylor series, the main part of the original function is carried on by the first term and the details are explained by the following terms. Similarly, it is possible to obtain the same result by taking advantage of two functions called scale and wavelet in wavelet analysis.

In the beginning, defining a generator function called the primary scale function ( $\varphi(t)$ ) and transforming along time or location axis, an order of functions are obtained as  $V_0$  what we call the subspace of  $L^2(R)$  :

$$\varphi_k(t) = \varphi(t-k)$$

$$V_0 = \overline{\text{span}\{\varphi_k(t)\}}$$

Now exerting the scale factor  $j$  the scale function is minimized or maximized:

$$\varphi_{j,k}(t) = 2^{j/2} \varphi(2^j t - k)$$

$$V_j = \overline{\text{span}\{\varphi_{j,k}(t)\}}$$

Changing the scale, the set of  $V$  spaces are resulted with the following relevance:

$$\begin{cases} V_0 \subset V_1 \subset V_2 \subset \dots \subset L^2 \\ V_j \subset V_{j+1} \end{cases}$$

The complementary orthogonal of  $V_j$  space will include wavelets in  $V_{j+1}$  space and it can be written:

$$V_{j+1} = V_j \oplus W_j$$

$$L^2 = V_0 \oplus W_0 \oplus W_1 \oplus W_2 \oplus \dots$$

According to above, any optional function from the  $L^2(R)$  space such as  $g$  can be written as sum of the generalities and details:

$$g(t) = \sum_{k=-\infty}^{+\infty} C_k \varphi_k(t) + \sum_{j=0}^{+\infty} \sum_{k=-\infty}^{+\infty} d_{j,k} \psi_{j,k}(t)$$

That are the expansion coefficients just DWT for the  $g$  function.

Meanwhile above subjects are generalized for the 2-dimensional state, as well.

### Interpolating the imperfect grational data by the wavelet transformation:

In this section, a network of the sample gravitational anomaly information involving 144 points (12\*12) with 5 minute distance in length and width has been used as high density information, a large amount of information were eliminated intentionally and instead zero was supplanted. The real complete data of that area have been applied to interpolate the imperfect information in 10 minute distance side by side.

The imperfect dense data have been transferred to the frequency space with the 2-dimensional discrete wavelet transformation and the simplest kind of the base functions i.e. the Haar system. Whereas the signal image is separated to 2 parts, details and generalities, consequently the related matrix in argument space can be rebuilt with the real complete information but with less density and finally these restored generalities with the details resulted the wavelet transformation are mixed in 21-dimensional inverse wavelet transformation and final interpolated information are computed.

### Fitting a model for specified nature

In previous sections, various types of interpolation model were considered. As one know more about the mathematical nature of a model, should understand how to select and use a model for simulation?

In the following section, a method will be discussed that contains some five steps.

- Primary data evaluation
- Selecting a proper model and testing model against data
- Initializing model parameters
- Evaluation of parameters and optimum selection
- Providing assurance graph of interpolated data

## Primary data evaluation

### Selecting a proper model and testing model against data

As one encounter a nature, should consider it to explore its mathematical behavior. For an example, if the nature is the topographic status of a field, one should test various models such as TIN, Wavelet, IDW and .... Visual overview would be the easiest way to evaluate the model. But using the check points would be the better idea. This method checks different models with predefined values against finite number of check points. Control points should not be deployed in interpolation algorithms. Then, error will be equal to difference of interpolation results with their real values that have not attended in interpolation. To present an income of error, Least Square would be a suitable solution in whole dataset. This test, is implemented several times, by small shifts in parameters and different models, to give the best fit.

### Primary Data evaluation

In this section, data is interrogated to remove special part of random error that has been provided by dependency among data in a dataset. This error is recognized by auto-covariance function. This error would be recognized by different tools. Among those tools, Histogram tool, is a useful one. Using such tool, according to data skewness, it is possible to asset data dependency.

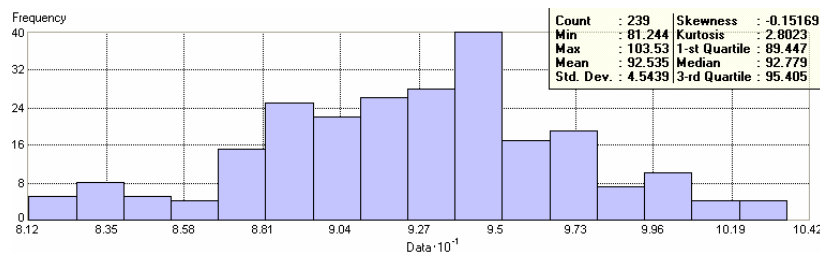


Figure- Data histogram of a topographic dataset

Although drawing QQPLOT reconfirm existence of dependency and deflection in a dataset. This graph compares the quintile value of the sample dataset against the quintile value of normal samples.

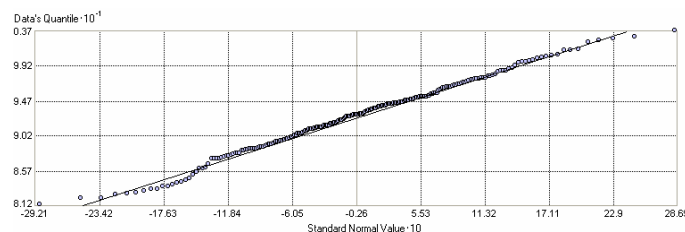


Figure- QQPLOT of a dataset which shows data deflection against the normal distribution

An other tool to explore dependency and deflection, is to explore trend in dataset via the trend analysis. For example, following figure is result of trend analysis for the same topographic dataset.

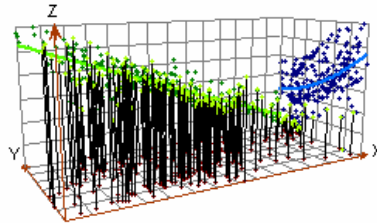


Figure- explored trend in a dataset

Finally, it is possible to reveal deflection of a dataset by considering its covariance and semi-variogram. These methods let to detect any directional dependencies too.

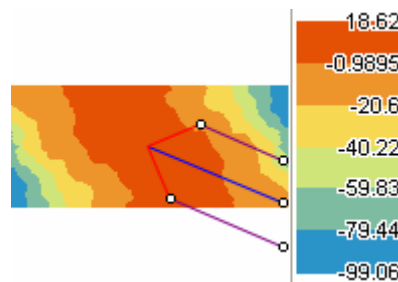


Figure- directional dependency in a dataset

As dependencies in a dataset are explored, it would be possible to select suitable model for dataset.

### Determining Model Parameters

Despite the model selection, determining model parameters is too important. In order to find out the best model parameters, it is possible to construct a cycle of an educative array and control points array. In each iteration, a comparison criteria such as L1 norm, measures deflection to the control array. Then it would be easy to find out the best parameters by minimizing the norm.

For example assume an IDW model. In the first step, it is possible to consider ellipse error of search radius to optimum the number of neighbors that should be used for interpolation.

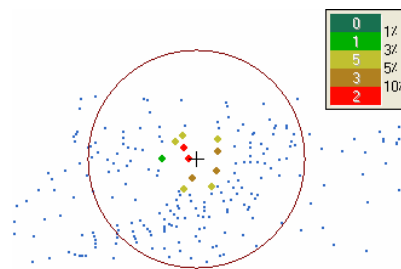


Figure- ellipse error of at least 10 and at most 11 neighbors in IDW model for a dataset

In the next step, smoothness of model would be optimized using its ellipse of error.

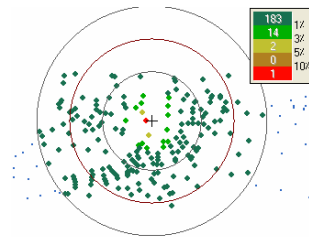


Figure- size of ellipse error for smoothness of 0.4

Finally it is possible to predict the model error based on defined parameters.

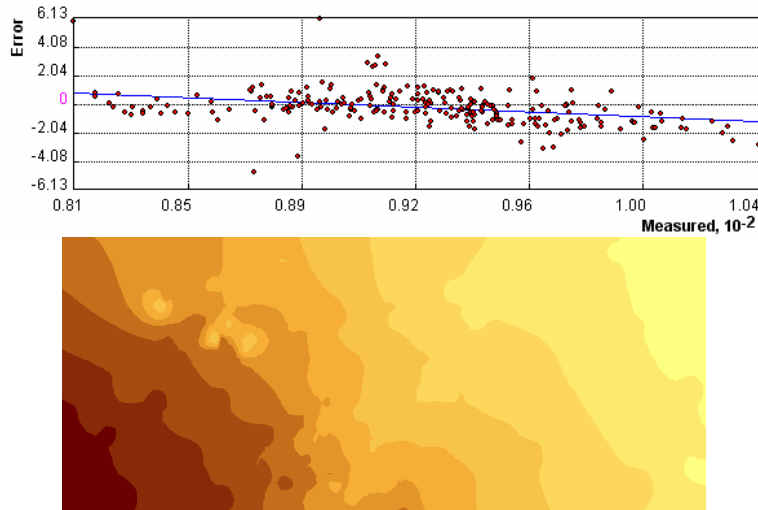


Figure- Interpolated model and its predicted error for defined model parameters

### Conclusion

As it was mentioned earlier, by expansion of traditional software usage among naïve users, production of spatial dataset has been significantly increased. An important point is the accuracy of such dataset. Our investigation on at least 30 different datasets in several fields revealed that, these datasets are not accurate. In most of the cases, error volume exceeded eligible border in the manner that made datasets useless.

Data set	Samples	Test points	Relative Accuracy	Result
Water PH of well in Delijan	104 points	15	16%	Failed
Soil quality of Saveh	86 points	10	25%	Failed
E.C in Komijan	63 points	8	31%	Failed
Fill and Cut volume, Qom	1087 points	50	19%	Failed
Temperature, Markazi province	16 points	4	41%	Failed

Table- some sample datasets and their relative accuracy

Despite the tremendous amount of error, these datasets are commonly used in different organization in Iran. Consequently such datasets form the basic milestone of spatial decisions and several levels of management in country. Therefore, resulted errors, in spatial datasets, deflects spatial management and gradually dismiss the managers with GeoSpatial Systems.

Nevertheless, using mentioned simple methodology and based on dominant ArcGIS toolbars, it would be possible to decrease dataset error.

Data set	Samples	Test points	Relative Accuracy	Result
Water PH of well in Delijan	104 points	15	7%	Approved
Soil quality of Saveh	86 points	10	5%	Approved
E.C in Komijan	63 points	8	15%	Approved
Fill and Cut volume, Qom	1087 points	50	6%	Approved
Temperature, Markazi province	16 points	4	16%	Approved

Table- Sample datasets and their evaluation results based on modified interpolation model

Significant growth in result of modified models, depicted the usefulness of method. Nevertheless, modification of models, is not enough lonely. An other idea is to deployment of error propagation analysis in spatial models. Error propagation analysis, helps to assign a strict accuracy criteria on datasets. This would accompany data, in its context, via a regulated metadata.

## References:

- Hardy, R.L & W.M. Gofert, 1975, Least square prediction of gravity anomalies, geoidal undulations, and deflections of the vertical quadric harmonic functions
- Pottmann, H & M. ECK, 1990, modified multi quadric method for scattered data interpolation over a sphere
- Franke, 1982, harder and Desmarais, 1972 and Sandwell 1987
- Robbert DE Wal, Digital elevation data and GIS projects
- ArcGIS Manual
- Cressman, 1968
- Fisher, N.I, T. Lewis, and B.J.J.EMBlton, 1987, Statistical analysis of spherical data
- Sheperad, D. 1968, Aid to dimensional interpolation function for irregular space data

This document was created with Win2PDF available at <http://www.daneprairie.com>.  
The unregistered version of Win2PDF is for evaluation or non-commercial use only.