

An adaptive local procedure to approximate unevenly distributed data

M. Bozzini¹ and L. Lenarduzzi²

¹ *Dipartimento di Matematica Applicata,
Università di Milano Bicocca, via Cozzi 53, 20125 Milano
mira.bozzini@unimib.it*

² *IMATI CNR, via Bassini 15, 20133 Milano
licia.lenarduzzi@mi.imati.cnr.it*

Abstract

We propose an adaptive local procedure, which uses the modified Shepard's method with local polyharmonic interpolants. The aim is to reconstruct, in a faithful way, a function known by a large and highly irregularly distributed sample. Such a problem is generally related to the recovering of geophysical surfaces, where the sample is measured according to the behaviour of the surface.

The adaptive local procedure is used to calculate, by an efficient algorithm, an interpolating polyharmonic function, when a very large sample is assigned.

When we consider a sample of size $N < 10^4$, we propose an approximating polyharmonic function obtained by combining adaptively a global interpolant, relevant to a subset of the data, with local adaptive interpolants.

The goodness of the approximating functions in two different cases is shown by real examples.

Keywords: Polyharmonic functions, Shepard's method, interpolation, adaptivity, unevenly distributed data.

1. Introduction

One of the current problems in numerical approximation is the reconstruction of surfaces known by a large sample of data, whose locations within a domain $\Omega \subset \mathbb{R}^2$ are highly irregularly distributed.

Such a problem is generally related to the reconstruction of geophysical surfaces, where the sample is measured according to the behaviour of the surface. This means that there is high sparsity of points where the function is flat; on the contrary, the density of the data is large where the function is rapidly varying.

Large sets of data, extracted from a strongly nonhomogeneous density, are treated by several authors with different approaches, all involving the determination of the suitable choice of a parameter, in order to obtain a correct behaviour of the approximating function (see, for example, the papers [1], [8] and the recent manuscript [2]).

In this paper, we present two adaptive techniques based on the modified Shepard method with radial basis functions: the former for a sample of very large size, the latter for not too large size, $N < 10^4$.

The corresponding algorithms are automatic and they determine approximating functions with a faithful graphical reproduction.

The goodness of the solution is ruled by the adaptivity of the local neighbourhoods and by the choice of the polyharmonic functions as a basis.

The paper is organized as follows.

After giving the main notations in §1.1, in §2, for convenience of the reader, we recall the modified Shepard method with radial bases [8]; in §3 we present the adaptive procedure, while in §4 we describe the two techniques. Lastly, in §5 we show the effectiveness of the procedures by examples relevant to real data and treated in the literature about the topic.

1.1. Notations.

- The polyharmonic function is defined as:

$$\phi_m(x) := \|x\|_2^{2\beta} \log \|x\|_2, \quad x \in \mathbb{R}^2, \text{ with } \beta \in \mathbb{N}.$$

$\phi_m(\cdot)$ is conditionally positive of order $\beta + 1 = m$.

- We denote with $F(X_l) := \{f(x_1), \dots, f(x_L)\}$, the set of functional values obtained from an unknown function $f : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ on a finite set of points $X_l = \{x_1, \dots, x_L\} \subset \Omega$ and we call $I_l = I(X_l) := \{1, \dots, L\}$ the corresponding set of the indices.

- We denote

$$I_l f(x) = \sum_{j \in I_l} c_j^l \phi_m(\|x - x_j\|_2) + p_\beta^l(x)$$

the interpolant of the set $F(X_I)$, where $p_\beta^l(x)$ is the polynomial of degree β , subject to the constraint

$$\sum_{j \in I_I} c_j p_\beta^l(x_j) = 0.$$

As it is well known, $If_I(x)$ achieves the property of the minimal energy. That is to say that, among all the functions that interpolate the set $F(X_I)$ and that belong to the Beppo Levi space of order m , If_I minimizes the seminorm on the space.

In the case $m = 2$, the energy is the bending energy and $\phi_m(\cdot)$ is named thin plate spline.

2. Modified Shepard's method with the polyharmonic basis.

Let be assigned a set of scattered data $\{X_N, F(X_N)\}$ and we denote with $I_N = \{1, \dots, N\}$ the set of the corresponding indices.

For each point $x_i \in X_N$, we consider a circular neighbourhood \mathcal{U}_i that includes the n points of X_N nearest to x_i according to the Euclidean norm, and we denote I_i the set of their indices.

We construct the local polyharmonic interpolant

$$If_i(x) = \sum_{j \in I_i} c_j^i \phi_m(\|x - x_j\|_2) + p_\beta^i(x).$$

The function that interpolates the data $(X_N, F(X_N))$ is obtained by a convex combination of the interpolants $\{If_i(x)\}_{i=1}^N$ with weights $\{\bar{w}_i(x)\}$ defined as follows.

We choose an integer $n_w < n$ and, for any neighbourhood $\mathcal{U}_i, i \in I_N$, we indicate with $r(I_i^w)$ the radius of the circular neighbourhood, centered at the point $x_i \in X_N$, containing the n_w points closest to x_i and belonging to the set $X_N \cap \mathcal{U}_i$. We put $r_i(x) = \|x - x_i\|_2$ and we define

$$(1) \quad w_i(x) = [(r(I_i^w) - r_i(x))_+ / (r(I_i^w) * r_i(x))]^2.$$

Let

$$(2) \quad \bar{w}_i(x) = w_i(x) / \sum_{j \in I_N} w_j(x)$$

be the weights which are a partition of the unity.

Hence the interpolant of $(X_N, F(X_N))$ is defined as the convex combination:

$$s_{\phi_m}(x) := \sum_{i \in I_N} \bar{w}_i(x) I f_i(x).$$

In addition, we recall that Iske proves that, if $f \in C^m$, the approximation order of *local polyharmonic spline* interpolation, using ϕ_m , is m , [7].

3. Construction of the adaptive neighbourhood.

Here and in what follows, we consider $m = 2$; we assume $f \in C^2(\Omega)$, where Ω is a domain that includes the unitary square Q . Let be given a set of points X_N over Q , sampled according to a continuous distribution function $P_{\mathbf{x}}$ depending on the behaviour of the function f . Let N be sufficiently large.

The construction of the local adaptive neighbourhood arises from two considerations exposed in §3.2, and it makes use of the restriction of the Delaunay triangulation within the minimal polygonal containing X_N ; restriction that we denote $\mathcal{D}_0(X_N)$.

3.1. Construction of the triangulation $\mathcal{D}_0(X_N)$.

We consider the extension X^* of X_N outside Q by copying the set X_N by even symmetry with respect to $x = -\rho$ at first; successively with respect to $x = 1 + \rho$, to $y = -\rho$ and to $y = 1 + \rho$, where ρ is an arbitrary positive real number. We calculate the Delaunay triangulation, [9], of X^* and then we restrict the triangulation to the sole triangles whose vertices belong to X_N .

In Fig. 1 we show the classical Delaunay triangulation of a set X_N and in Fig. 2, for the same set X_N , we show $\mathcal{D}_0(X_N)$.

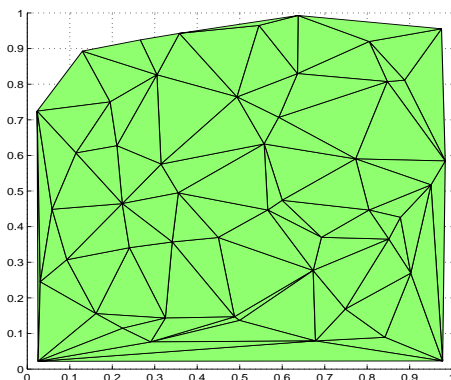


Fig. 1. Delaunay triangulation.

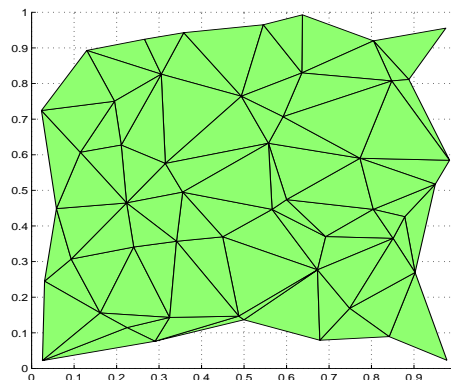


Fig. 2. $\mathcal{D}_0(X_N)$.

3.2. Two considerations.

- We consider a point $x_i \in X_N$ and its circular neighbourhood \mathcal{U}_i that includes n points $x_j \in X_N$; let I_i be the set of the indices relevant to the n points.

We construct on \mathcal{U}_i the local interpolant $If_i(x)$. Such a function is defined on the whole Ω and here it is differentiable. Under the hypothesis that N is sufficiently large, the local interpolant enjoys a good accuracy for all $x \in \mathcal{U}_i$, see §2.

Let $x_k \in X_N \cap \mathcal{U}_i$ be a point close to the boundary of \mathcal{U}_i and give our attention to the closed interval $[x_k, x_l]$, where $x_l \in X_N \setminus (\mathcal{U}_i \cap X_N)$ is one of the points contiguous to x_k within the triangulation $\mathcal{D}_0(X_N)$.

We evaluate $e(x_l) = |f(x_l) - If_i(x_l)|$.

Because of the differentiability of the two functions, it is:

$$f(x_l) = f(x_k) + (x_l - x_k) \cdot \text{grad } f(\theta_1),$$

$$If_i(x_l) = If_i(x_k) + (x_l - x_k) \cdot \text{grad } If_i(\theta_2),$$

with $\theta_1, \theta_2 \in [x_k, x_l]$; it follows that

$$e(x_l) = |(x_l - x_k) \cdot (\text{grad } f(\theta_1) - \text{grad } If_i(\theta_2))|.$$

Therefore, a discrepancy between the gradient of the function and that of the interpolant is plain when the error $e(x_l)$ turns out to be greater than $K \|e_i(x)\|_\infty$, where $e_i(x)$ indicates the error function within \mathcal{U}_i .

In order to improve the similarity of behaviour between $\text{grad}f(x)$ and $\text{grad}If_i(x)$ along the segment $[x_k, x_l]$, it is opportune to include the point x_l in the set of the data to be locally interpolated.

- The median is the index of central tendency that best individuates the outliers. When the values of a random variable \mathbf{z} are uniformly distributed in an interval, the median falls in the center of the interval.

Let be assigned a sample of size q and we denote $\{z_1, \dots, z_q\}$ the increasingly ordered sample. We consider the estimator M of the median that is unbiased and consistent. In addition, we know that a confidence interval of the median is given by $[z_1, z_q]$ with probability $1 - (1/2)^{q-1}$.

Therefore, in the case in which the values of the sample are roughly uniformly distributed, the estimator M will not be far from the value

$$(z_1 + z_q)/2 := \bar{z}.$$

On the contrary, in presence of outliers, the value of M will be far from \bar{z} .

3.3. Construction of the local adaptive neighbourhood.

From what above, the construction of the adaptive neighbourhood is done according to the following procedure:

- For $x_i \in X_N$, we consider the circular neighbourhood \mathcal{U}_i that includes the n closest data sites $x_j \in X_N \cap \mathcal{U}_i$ and we calculate the interpolant $If_i(x)$.

- We consider the set of the points $x_l \in X_N \setminus (X_N \cap \mathcal{U}_i)$ contiguous to the points $x_j \in \mathcal{U}_i$ with respect to $\mathcal{D}_0(X_N)$; let n_i be their number and $\delta\mathcal{U}_i$ a region that contains only them. We calculate the vector whose increasingly ordered components

$$\{e(x_l) := |If_i(x_l) - f(x_l)|\}_{l=1}^{n_i},$$

are the result of the sampling of the random variable \mathbf{e}_δ depending on the distribution $P_{\mathbf{x}}$ and defined in $\delta\mathcal{U}_i$.

- We consider the following hypothesis test:

$$H_0 : M = \bar{e} = (e(x_1) + e(x_{n_i}))/2$$

against

$$H_a : M \neq \bar{e}.$$

The critical region is $R_c : [2M, \infty)$.

If for some l , $e(x_l)$ belongs to R_c , then we reject the null hypothesis. We calculate the interpolant $If_i(x)$ on the points x_j , $j \in I_i \cup I_l$ where I_l is the set of the indices relevant to the sites $x_k \in \delta\mathcal{U}_i \cap X_N$ for which the inequality

$$e(x_k) \geq 2M$$

is satisfied.

4. Construction of the approximating function.

4.1. Sample with size $N > 10^4$.

In the case of very large samples ($N > 10^4$) the approximating function is provided by the interpolant constructed as a convex combination of local interpolating functions defined on neighbourhoods determined adaptively as said in the previous section.

We denote these functions $\{If_i(x), i \in I_N\}$, and we refer to them as local adaptive interpolants. Let

$$s(x) = \sum_{i \in I_N} \bar{w}_i(x) If_i(x)$$

be the interpolant obtained by the weights defined in (1) and (2) of §2 .

Proposition 4.1. *Let be $f \in C^2(\Omega)$ and let X_N be the N sampled data sites in Q from a continuous distribution function $P_{\mathbf{x}} = P_{\mathbf{x}}(f)$. We measure the local distribution of the data sites by $h_i = \max_{x \in \mathcal{U}_i^*} \min_{j \in I_i \cup I_l} \|x - x_j\|_2$, where \mathcal{U}_i^* is the adaptive neighbourhood. Let h_i vanish as $N \rightarrow \infty$. Then for each $x \in Q$*

$$e(x) = |f(x) - s(x)| \leq Ch^2, \quad h = \max_{i \in I_N} \{h_i\}.$$

Proceeding along the lines of the proof given by Iske, [7] , we express the local adaptive interpolant $If_i(x)$ in Lagrangian form:

$$If_i(x) = \sum_{j \in I_i \cup I_l} f(x_j) l_j(x),$$

that has bounded Lebesgue constant, because the set $I_i \cup I_l$ is of moderate size due to the construction.

So we deduce, as Iske, that the approximation order of the local adaptive interpolant is 2. That is

$$e_i(x) = |f(x) - If_i(x)| \leq C_i h_i^2.$$

Therefore for any $x \in Q$, it is:

$$\begin{aligned} e(x) = |f(x) - s(x)| &= \left| \sum_{i \in I_N} \bar{w}_i(x) \{f(x) - If_i(x)\} \right| \\ &\leq \sum_{j \in I_x} \bar{w}_j(x) |f(x) - If_j(x)| \leq Ch^2 \end{aligned}$$

where I_x is the set of indices of the weights not vanishing at x .

4.2. Sample with size N which is at most of order 10^4 .

We observe that the methods of global type provide better approximations with respect to those of local type. However, the more the distribution of the data points is irregular and far from uniform, the more the global interpolant can lead to approximating functions which present undue oscillations.

This leads us to define a technique that coniugates a method of global type, which makes sense since we consider a not large subset of the sample, with one of local type and adaptive. In such a way, the global interpolant, constructed on a subsample of size $n_G \ll N$, renders the features while the local interpolant improves the final result.

In the problems relevant to the geophysical surfaces and to scattered data, in general there is one of two possible configurations:

- a) data scattered according to the behaviour of the surface,
- b) data distributed along contours.

In the following, we describe the procedure relevant to the case a) because it is the same for the case b), except that, for the construction of the global interpolant, the data are chosen by adaptive thinning, applied contour by contour.

- Procedure for case a)

- Construction of the set X_G . Having fixed a tolerance T_0 by the adaptive thinning described in [5], we individuate a set X_G^N of significant data points that includes also the boundary points of the convex hull of X_N . On the set X_G^N we construct the global interpolant and we evaluate the differences between the interpolant and the sample data not belonging to X_G^N . If for some point $x_l \in X_N \setminus X_G^N$ the difference is greater than T_0 , we enlarge the set X_G^N to a new set X_G including the points $\{x_l\}$. We calculate the global interpolating function $If_G(x)$ on $X_G \supseteq X_G^N$.

- Construction of the local adaptive interpolants

We evaluate the vector E with components

$$e(x_j) = |f(x_j) - If_G(x_j)| \quad \forall x_j \in X_N \setminus X_G$$

and determine two tolerances T_1 and T_2 :

$$T_1 = 0.5 \|E\|_\infty$$

$$T_2 = \alpha T_1 \quad \alpha \in (0, 1).$$

- Let I_1 be the set of the indices relevant to the points $\{x_i\}$ for which

$$e(x_i) > T_1$$

and I_2 the set of the indices relevant to the points for which

$$T_1 > e(x_k) > T_2.$$

- For each $i \in I_1$ and for each $k \in I_2$, we calculate the local adaptive interpolants that we denote respectively $If_i^1(x)$ and $If_k^2(x)$.

- Construction of the approximating function.

The approximating function is constructed as a convex combination of different nodal functions, in order to pass from a region of the domain to another gently. Precisely:

- $R_i(x) = If_i^1(x) \quad \forall i \in I_1.$

· We define $\omega_k = (e(x_k) - T_2)/(T_1 - T_2) \quad \forall k \in I_2$ and we put

$$R_k(x) = If_k^2(x) \cdot \omega_k + (1 - \omega_k) \cdot If_G(x) |_{\mathcal{U}_k^*},$$

where $If_G(x) |_{\mathcal{U}_k^*}$ is the restriction of the global interpolant to the adaptive neighbourhood \mathcal{U}_k^* .

· For each $l \in I_N \setminus (I_1 \cup I_2)$, let $R_l(x) = If_G(x)$ be the restriction of the global interpolant to the circular neighbourhood \mathcal{U}_l .

The approximating function is obtained by the convex combination with the weight defined in (1) and (2) of §2:

$$s_R(x) = \sum_{j \in I_N} \bar{w}_j(x) R_j(x).$$

5. Examples.

We show three examples. The first relevant to a sample of very large size. For it we have constructed the interpolating function by local adaptive interpolants, having fixed $n = 14$ and $n_w = 10$. The other two are relevant to samples of size $N < 10^4$.

These examples were considered by other authors too. Therefore, due to the limited number of pages, for each example, we give the corresponding references in order to make a comparison.

Example 5.1. This example refers to the data of the Black forest described in [4], $N = 15885$.

In Fig. 3, we show the reconstruction of the part that is most difficult to be rendered well; it is a detail of the subregion exemplified in [2], while for the whole subregion we give the contours of the interpolant in Fig. 5.

The figures confirm that our procedure provides surfaces of high quality from difficult terrain data; on the contrary in the recent parametric approach of [2], one observes only a substantial reduction of the undue artifacts.

In Fig. 4 we present the reconstruction obtained by following the hints of [8] for unevenly distributed data. Precisely, the interpolating function is constructed by the modified Shepard's method with the C^2 Wendland's function, whose support, for any circular neighbourhood \mathcal{U}_i , is proportional to the local density of the data. It is possible to observe that such a strategy does not succeed in removing undue oscillations.

Example 5.2. We refer to the bay data set, from [6], $N = 1669$. This data set is difficult because of the sharp peak and of the shallow area, both with few data sites.

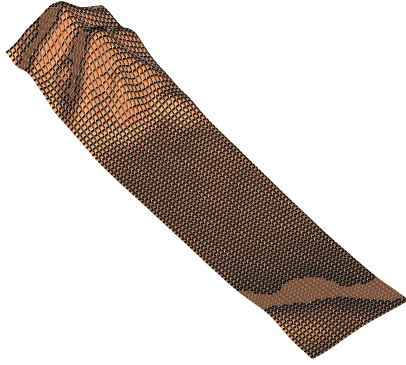


Fig. 3. $s(x)$; detail.

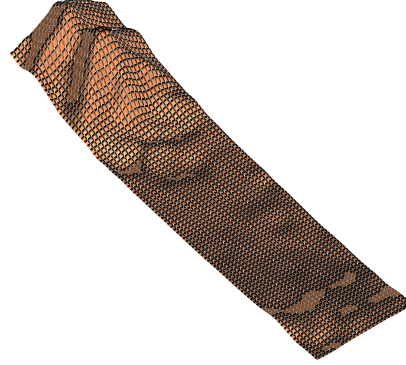


Fig. 4. Wendland's basis; detail.

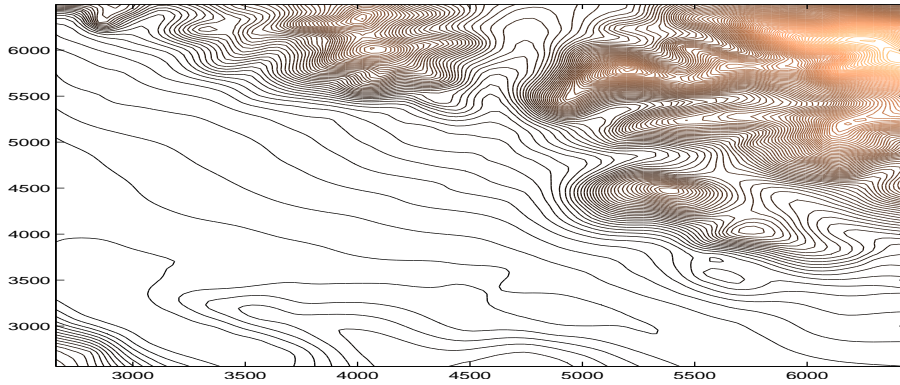


Fig. 5. $s(x)$; contours on the subregion.

We put $T_0 = 1\%$ and we obtain for the set X_G the size 316.

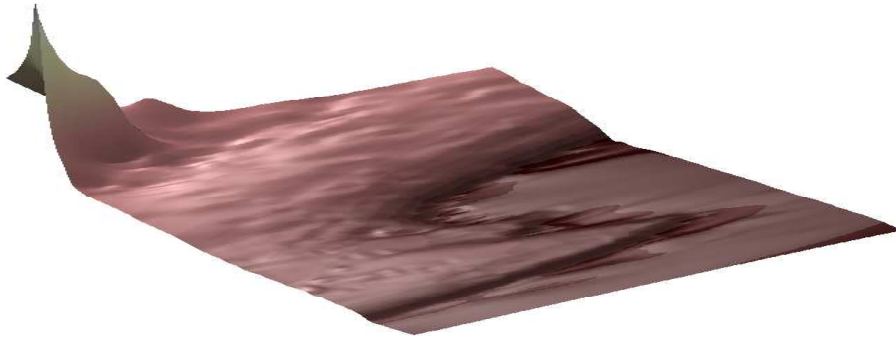
We have prescribed $n = 14$, $n_w = 10$, $T_1 = 0.005$ and $T_2 = 0.0025$. This leads us to calculate 598 local nodal functions.

The final result is shown in Fig. 6 that we can compare with the figures in [10]. Moreover the errors of $s_R(x)$ at the data points are $e_2 = 0.0014$ and $e_\infty = 0.0025$.

Example 5.3. It refers to the glacier data of [6], $N = 8345$ which have been treated in [11] and in [3].

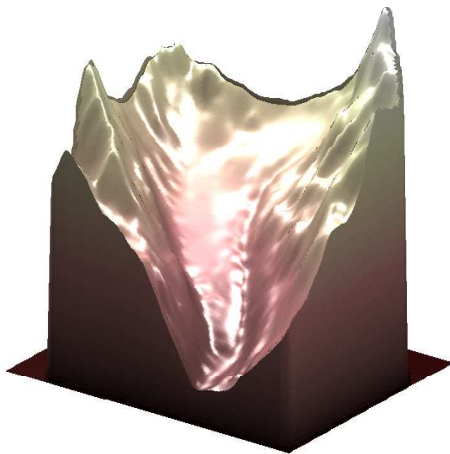
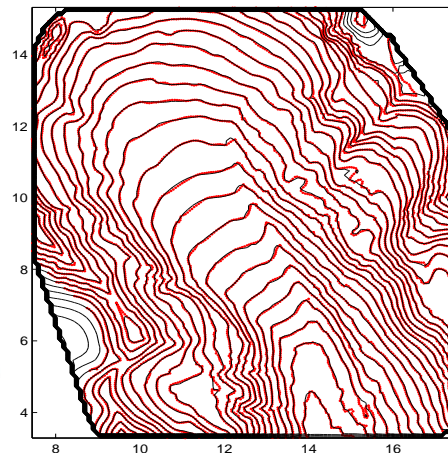
There is a strongly nonhomogeneous configuration of the data because they are dense along some contours;

The cardinality of X_G with $T_0 = 1\%$ is 992 while the number of nodal functions with $n = 60$, $n_w = 40$ and $T_1 = 0.14\%$, $T_2 = 0.1\%$ is 3246.

Fig. 6. $s_R(x)$

The 3d graphical representation of the approximation is shown in Fig. 7 while in Fig. 8 we can see the contours of $s_R(x)$ (lines) together with the data points (dots); the range of the data is from 1200 m to 2200 m . The errors at the data points are $e_2 = 0.965 m$ and $e_\infty = 2.5 m$. Notice that the best errors by the method [3] are $e_2 = 1.74 m$ and $e_\infty = 15.4 m$.

The different accuracy determines different accuracy of the graphical reproduction, as shown in the figures reproduced in [3] and in [11].

Fig. 7. $s_R(x)$ Fig. 8. Contours of $s_R(x)$ and data over-imposed as dots.

6. Conclusions.

The adaptive procedure here presented is fast and robust. It provides high quality surfaces from data with strong variations in the data density

according to strong changes in the gradient of the surface.

The core of our adaptive procedure is that it provides a local interpolant whose gradient follows the behaviour of the gradient of the function.

This procedure, together with the use of an approximation of global type, allows a function approximation of high accuracy and of graphical quality with a good computational cost for a sample of not too large size.

REFERENCES

1. A. Bouhamidi and A. Le Méhauté, Radial basis functions under tension. *J. of Approx. Theory* **127** (2004), pp. 135–154.
2. O. Davydov and C. Manni, A tension approach to controlling the shape of cubic spline surfaces on FVS triangulations. Preprint 2007
3. O. Davydov, R. Morandi and A. Sestini, Local hybrid approximation for scattered data fitting with bivariate splines. *Comput. Aided Geom. Design* **23** (2006), pp. 703–721.
4. O. Davydov and F. Zeilfelder, Scattered data fitting by direct extension of local polynomials to bivariate splines, *Advances in Comp. Math.* **21** (2004), pp. 223–271.
5. N. Dyn, M. Floater and A. Iske, Adaptive thinning for bivariate scattered data, *J. Comput. Appl. Math.* **145** (2002), pp. 505–517.
6. www.math.nps.navy.mil/~rfranke
7. A. Iske, *Multiresolution Methods in Scattered Data Modelling*, Springer Verlag, 2004.
8. D. Lazzaro and L. Montefusco, Radial basis functions for the multivariate interpolation of large scattered data sets. *J. Comput. Appl. Math.* **140** (2002), pp. 521–536.
9. Preparata, F. P., Shamos, M. I., *Computational Geometry*. Springer Verlag (1988).
10. R. Schaback and H. Wendland, Adaptive greedy techniques for approximate solution of large RBF systems, *Numerical Algorithms* **24** (2000), pp. 239–254.
11. H. Wendland, *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2005.