DOI 10.4430/bgo00423

Exploring error estimation methods for natural neighbour interpolation: preliminary research and analysis

M. Iurcev and F. Pettenati

National Institute of Oceanography and Applied Geophysics - OGS, Trieste, Italy

(Received: 6 March 2023; accepted: 12 May 2023; published online: 20 July 2023)

ABSTRACT Interpolation of scalar data, in the 2D space, is an important topic in many fields of environmental and geoscience studies, and uncertainty assessment is as important as interpolation itself. An example of this is the Kriging method, which is well-established in geostatistics and enables the automatic evaluation of uncertainties by solving a linear equation, taking into account the bivariate spatial continuity of the data. The Sibson interpolation method (natural neighbour) has the important property of providing unambiguous and reproducible results. However, since it is fundamentally a deterministic method, it does not have qualitative and/or quantitative control of the uncertainty based on the sampling spatial distribution geometry. In this paper, we show the different steps leading to an analytical approach to evaluate the uncertainties of the Sibson method. After a series of tests with a synthetic data set and a surface with a known differentiable function, we show an example using the data set of accelerometric data from the *M* 6.5 Norcia earthquake of 30 October, 2016.

Key words: bivariate interpolation, natural neighbour, voronoi polygons, gradient estimation, ordinary least squares.

1. Introduction

In his early work, Sibson (1980) showed that the principle of natural neighbour (hereafter NN), based on convex Voronoi polygons (Aurenhammer *et al*., 2013), can be used to uniquely determine the coordinates of a point not included in input data. In the subsequent article (Sibson, 1981), he applied the NN method to spatial interpolation and demonstrated its unique and reproducible feature. However, what this method lacks is qualitative and/or quantitative control of uncertainties based on the sampling spatial distribution geometry, relative to the real data surface. The method is an exact interpolator, in the sense that the original data values are preserved at the reference data points.

A first approach to evaluate uncertainties in the NN method is based on cross-validation errors (Etherington, 2020). This approach involves the calculation of the mean absolute error (cross-validation) of estimate value f^* over data f_i, by taking into account the mutual distances between these points. What is important about Etherington's study is that he outlines the method's properties. Namely, the method: i) is an exact interpolator; ii) it creates a smooth surface; iii) it is entirely local; iv) it is spatially adaptive; v) no requirement is needed to make statistical assumptions; vi) it can be applied to very small data sets as it is not statistically based; and above all vii) it is parameter free. Thus, a previous attempt, based on fitting statistical uncertainty

models (Ghosh *et al*., 2012), contradicts properties v, vi, and vii of NN interpolation.

Early categories of interpolation methods, concerning the evaluation of uncertainties, are those based on objective analyses (Barnes, 1964; Gandin, 1965), and mainly used in the field of meteorology. For each grid point, the unknown function is computed using a series of Gaussian functions, given by a weighted distance. For the interpolation and variance estimation, an objective method, based on the spatial variability structure of the data, is used.

A recent article, which only deals with the uncertainty issue and applies a non-parametric approach, is that by Thiesen and Ehert (2022). For spatially distributed data, Kriging methods are the most important category of interpolation tools in geostatistics. The various Kriging approaches enable the estimation of uncertainties to estimate the confidence interval (Goovaerts, 1997). Variograms (Chilès and Delfiner, 1999), the basis in the theory of geostatistics, are empirical functions expressing data dissimilarity as a function of the distance between data themselves. Variograms are used to derive the covariance matrix of the likelihood function in order to solve the linear equation system of Kriging models, such as Ordinary Kriging, Universal Kriging, and so on. In practice, the theoretical semivariance, taken as the covariance function, is used as the second order moment. The semivariance is estimated by fitting an empirical variogram (Chilès and Delfiner, 1999). Kriging methods have the advantage of intrinsically estimating variance (known as the Kriging variance), or, rather, its square root (also called the standard deviation, or standard error for short).

Moreover, Kriging has a strong link with objective analysis, as described in Herzfeld (1996).

This paper outlines the steps of our research to evaluate the uncertainties of the Sibson (1980, 1981) interpolation theory based on Voronoi tessellation. In the initial phase, we attempted to combine the Sibson (1981) scheme with variograms on a geostatistical point of view (see Appendix A), and subsequently explored a deterministic approach. The approach presented here, complies with the rules of Etherington (2020), in particular with properties iii and vii. Based on the Mean Value Theorem (MVT), the proposed methods use the gradient, which is useful "to increase the accuracy of local interpolants" (Belward *et al*., 2008) and is an important element in uncertainty evaluation, as in our case. However, the various numerical problems encountered are described in the following.

All diagrams and tests in this paper were generated with original Python code, using some standard libraries such as NumPy, SciPy and scikit-learn. NN interpolation was implemented with Python software using the methods described in Iurcev *et al*. (2021).

2. Sibson interpolation

The interpolation of an unknown function, f: $\mathrm{R}^2 \Rightarrow \mathrm{R}$, can be seen as the problem of reconstructing value $f(\boldsymbol{x}^*)$, assumed that we have a set of measured values, $\{f_i = f(\boldsymbol{x}_i)\}$, in points $\{\boldsymbol{x}_i\} = X$ (our data set). The Voronoi tessellation of our data set is uniquely defined as a set of convex polygons, $\{V_i\}$, partitioning the plane and having the following property:

$$V_i = \{\boldsymbol{p} \in \mathbb{R}^2 : \forall j \neq i, \|\boldsymbol{p} - \boldsymbol{x_i}\| < \|\boldsymbol{p} - \boldsymbol{x_j}\|\}. \quad (1)$$

The edges and vertices of the Voronoi tessellation, $\mathrm{U}_i \{\partial V_i\}$, are the locus of points equidistant from two data points (edges) or more data points (vertices). The Sibson interpolation method (Sibson, 1981) is based on the uniquely defined set of NNs of interpolation point $\boldsymbol{x}^*$. Considering

the interpolation point as a new data point, the new Voronoi V^* has a non-empty intersection only with some V_i: the n NNs of $\boldsymbol{x}^*$. If A_i is the intersection area, then the Sibson interpolation is defined as follows:

$$\tilde{f} = \sum_{i=0}^{n-1} w_i f_i \tag{2}$$

where weights are:

$$w_i = \frac{A_i}{A^*} = \frac{A_i}{\sum_j A_j}\,. \tag{3}$$

Additionally, the interpolation is a convex combination, as

$$\sum_i w_i = 1; w_i \geq 0. \tag{4}$$

The geometry of the NN method is displayed in Fig. 1.
Some variables are defined for convenience purposes:

$$\boldsymbol{d}_i = (\boldsymbol{x}_i - \boldsymbol{x}^*) \tag{5}$$

$$d_i = \|\boldsymbol{x}_i - \boldsymbol{x}^*\|\,,\ d_{ij} = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|. \tag{6}$$

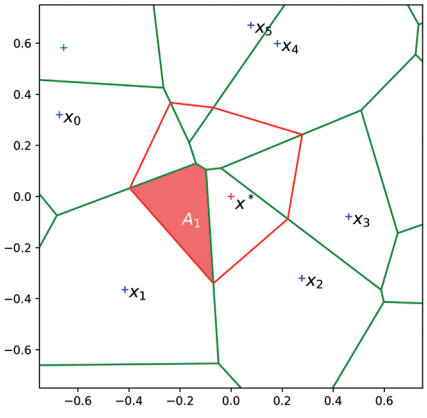

Fig. 1 - The Sibson interpolation method (NN).

Eqs. 2 and 4 are shared not only with the Sibson method but also with many other interpolation methods, for example, Inverse Distance Weighting (IDW). The latter and the Sibson method are both based on a convex linear combination and share the following property:

$$\boldsymbol{x}^* = \sum w_i \boldsymbol{x}_i. \tag{7}$$

A significant difference between these two methods consists in the choice of data used for the interpolation. IDW uses the whole X data set, or all the points within a certain radius, thus offering the possibility of more complex search criteria (e.g. quadrant search, anisotropic

windows, etc.). The NN method only uses the NNs of $\boldsymbol{x}^*$, or the subset $v(\boldsymbol{x}^*) \subset X$, as Etherington (2020) suggests in his third property.

The cardinality of v can be approximated under the hypothesis of the data set as a stationary Poisson point process with density λ (i.e. the spatial density of the sampling network), as every NN of $\boldsymbol{x}^*$ corresponds to one edge of a Voronoi polygon relative to $\boldsymbol{x}^*$, thus $\mathbf{E}(\#v(\boldsymbol{x}^*)) \simeq 6$, which is the first order moment of the number of edges for a Voronoi polygon (Okabe *et al.*, 2000). If we choose the data subset using a circle of radius r, the cardinality of $v_r = \{\boldsymbol{x} \mid \boldsymbol{x} \in X \text{ and } \|\boldsymbol{x} - \boldsymbol{x}^*\| \leq r\}$ can be estimated as $\mathbf{E}(\#v_r(\boldsymbol{x}^*)) = \lambda\pi r^2$.

With the NN method, every specific data set $\{\boldsymbol{x}_i\}$ defines a partition of R^2 as in Fig. 2. Every coloured region includes the points sharing exactly the same subset of NNs, i.e. $v(\boldsymbol{x})$.

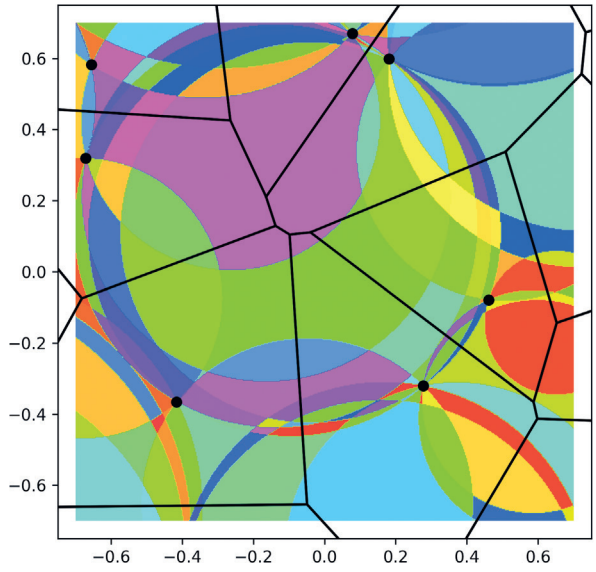

Fig. 2 - Partition of the plane into areas with the same v.

3. The gradient method

We assume the unknown function, f, to be differentiable, at least in proximity of the interpolation point. If we consider segment $S = [\boldsymbol{x}^*, \boldsymbol{x}_i]$ and apply the Mean Value Theorem (also known as the Lagrange theorem) extended to R^2, we obtain:

$$\exists \boldsymbol{\xi}_i \in S: (f_i - f^*) = \nabla f(\boldsymbol{\xi}_i) \cdot \boldsymbol{d}_i. \tag{8}$$

By substituting Eq. 2, we can evaluate the difference between interpolated value and real value:

$$\tilde{f} - f^* = f^*(\textstyle\sum_i w_i - 1) + \sum_i [w_i \nabla f(\boldsymbol{\xi}_i) \cdot \boldsymbol{d}_i]. \tag{9}$$

Since the weights sum to one, then the interpolation error is:

$$\tilde{f} - f^* = \textstyle\sum_i [w_i \nabla f(\boldsymbol{\xi}_i) \cdot \boldsymbol{d}_i] \tag{10}$$

considering the properties of the absolute value and the Cauchy-Schwarz inequality:

$$\left|\tilde{f} - f^*\right| \leq \sum_i w_i \|\nabla f(\boldsymbol{\xi}_i)\| d_i. \tag{11}$$

The last formula expresses the absolute interpolation error as a function of the gradient magnitude and geometrical distribution of our data set, in proximity of $\boldsymbol{x}^*$. All the sums are for $i = 0 \ldots n\text{-}1$, where n is the number of NNs in $\boldsymbol{x}^*$. As mentioned above, the formulae are valid for a wide range of methods, under the hypothesis of convex combination (Eqs. 2 and 4).

3.1. Gradient estimation

There are two issues with Eq. 10: the gradient and points $\boldsymbol{\xi}_i$ are unknown.

We can initially assume that the gradient is locally a constant vector. This assumption brings to a pointless result, since a locally constant gradient means that $f(\boldsymbol{x})$ is linear, the NN interpolation of a linear function is exact, and, therefore, the error is always zero. This can also be proved considering Eqs. 7 and 10 with $\nabla f = \mathbf{g}$ (constant):

$$\left|\sum_i (w_i \nabla f(\boldsymbol{\xi}_i) \cdot \boldsymbol{d}_i)\right| = \left|\boldsymbol{g} \cdot \sum_i w_i \boldsymbol{d}_i\right| = \left|\boldsymbol{g} \cdot \left(\sum_i w_i \boldsymbol{x}_i - \boldsymbol{x}^* \sum_i w_i\right)\right| = 0. \tag{12}$$

Although the hypothesis of a locally constant gradient can be used for the weaker scalar inequality of Eq. 11, it is still necessary to estimate the gradient magnitude in the neighbourhood of each interpolation point.

One possible approach, which is quite common in literature, is to approximate the gradient using finite differences, by superimposing a regular grid in which the function value is known or estimated. We have also tested this method, which, however, introduces an additional level of uncertainty, as the function must be interpolated through the grid.

Another approach is the local least-squares plane approximation of the unknown surface, as presented in Stead (1984) or De Keyser (2006).

The ordinary least-squares (OLS) approximation requires a subset of points $\boldsymbol{x}_i$, $f(\boldsymbol{x}_i)$ in the neighbourhood. With at least three non-collinear points in R^3 space, linear regression defines a plane whose slope is a possible gradient estimator, as in Fig. 3. In this context, two different least-squares strategies, for computing the gradient for bivariate surface interpolation, were

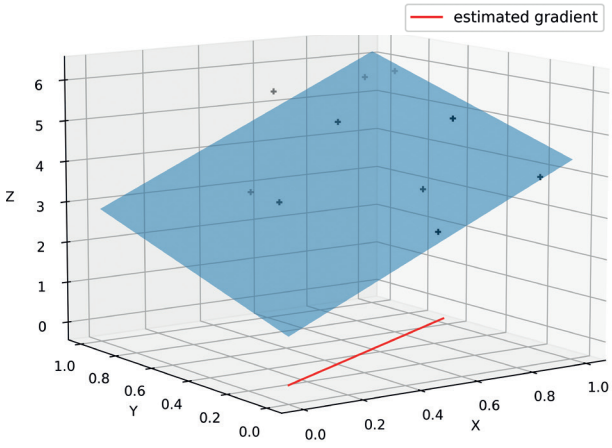

Fig. 3 - Gradient estimation with the OLS method.

investigated by Belward *et al*. (2008). The two methods are based on the generalisation of Moving Least Squares. In the former, a classic method based on a linear system of equations, the gradient is derived by a truncated Taylor expansion at the second order; while in the second method the gradient is a consequence of the Finite Volume Method solution used for solving diffusion equations. Belward *et al*. (2008) show that "the uniqueness of the gradient estimates (using both methods) is not a result of the analytical properties of the approximation processes, it is a consequence of the method of linear least squares".

There are many possible choices for the subset of points, for instance $v(\boldsymbol{x}^*)$ or $v_r(\boldsymbol{x}^*)$. We define the former *n* estimation (based on NNs), and the latter *r* estimation (using distance within a fixed radius). For sake of simplicity, let us define $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ and $G(\boldsymbol{x}) = \|\nabla f(\boldsymbol{x})\|$. The estimated gradients are then $\boldsymbol{g}_n$ (with NNs) and $\boldsymbol{g}_r$ (fixed radius); their magnitudes are G_n and G_r.

Of course, the choice of the radius with the *r*-method is quite arbitrary, whereas the *n*-method is uniquely defined. If the radius is too small, the subset used for the OLS estimator is possibly empty for many interpolation points. If the radius is too large, the gradient estimate is very poor. The best choice for the fixed radius depends on the local density of the data set. As described in De Keyser *et al*. (2007), the method is valid with a least an intrinsic stationary condition (Chilès and Delfiner, 1999).

4. Data sets

In order to investigate the validity of an uncertainty assessment, a first experiment with a synthetic data set, using a non-polynomial function proposed by Franke (1979), is presented in this paper. A representation of the Franke function can be found in Iurcev *et al*. (2021).

The second data is a set of 164 surveys of Peak Ground Acceleration (*PGA*) in cm/s^2 of the *M* 6.5 Norcia earthquake (central Italy) of 30 October, 2016.

4.1. The Franke test function

The Franke function is a differentiable function that is often used as a test function in literature. The surface has two Gaussian peaks and a narrow minimum superimposed fold on a surface sloping towards the first quadrant. Closed-form expressions for the gradient vector and its magnitude were obtained using the Python symbolic library SymPy.

The test data set was modelled as a Poisson point process, defined (Okabe *et al.,* 2000) by the process in which, for any point, subset A is:

$$P(\#A = k) = \frac{\lambda|A|e^{-\lambda|A|k}}{k!}, k \in \mathbb{N}. \tag{13}$$

Parameter λ is the density, or process intensity, and defines the expected number of points in a unitary area.

It is possible to improve the test by considering the impact of data errors (e.g. introducing some noise) or using other synthetic functions.

4.2. Franke function gradient estimation

We used a random data set of 500 points in the unitary square $(0.1)^2$. For every interpolation point, gradients $\boldsymbol{g}_r$ and $\boldsymbol{g}_n$ were estimated with the OLS method. The fixed radius for the *r*-method

was set to 0.08, which is approximately $2/\sqrt{\lambda}$.

Fig. 4 shows the results of an OLS gradient estimation, for a data set of 500 points. Fig. 4d shows the exact magnitude, whereas Fig. 4e the 'bubble' pattern displayed in Fig. 2, since all interpolation points, in each subset N_i, have the same NNs and, thus, the same OLS gradient estimator. In Fig. 4f, the gradient estimate with fixed radius can be seen.

We, then, examined the error in the gradient estimation (Fig. 5), both for the gradient magnitude and gradient phase. The approximation error for the gradient magnitude was higher near the edges, due to typical interpolation problems, such as the Runge effects. The phase error for $\boldsymbol{g}_n$ was plotted considering the metric h_n, a kind of pseudo-error, based on the cosine of the angle:

$$h_n = 1 - \frac{\boldsymbol{g}_n \cdot \boldsymbol{g}}{G_n G}. \tag{14}$$

This value is zero if the two vectors have the same direction, 0.5 if they are orthogonal, one if they point in opposite directions. The same definition applies to h_r. Table 1 contains a comparison between the two estimates of the gradient magnitude.

4.3. Franke function: interpolation error estimate

Both Eqs. 10 and 11 provide an estimate for the NN interpolation error. The former is a vectorial method, the latter a scalar one, which provides only the absolute error. For consistency reasons, only absolute estimated errors are compared.

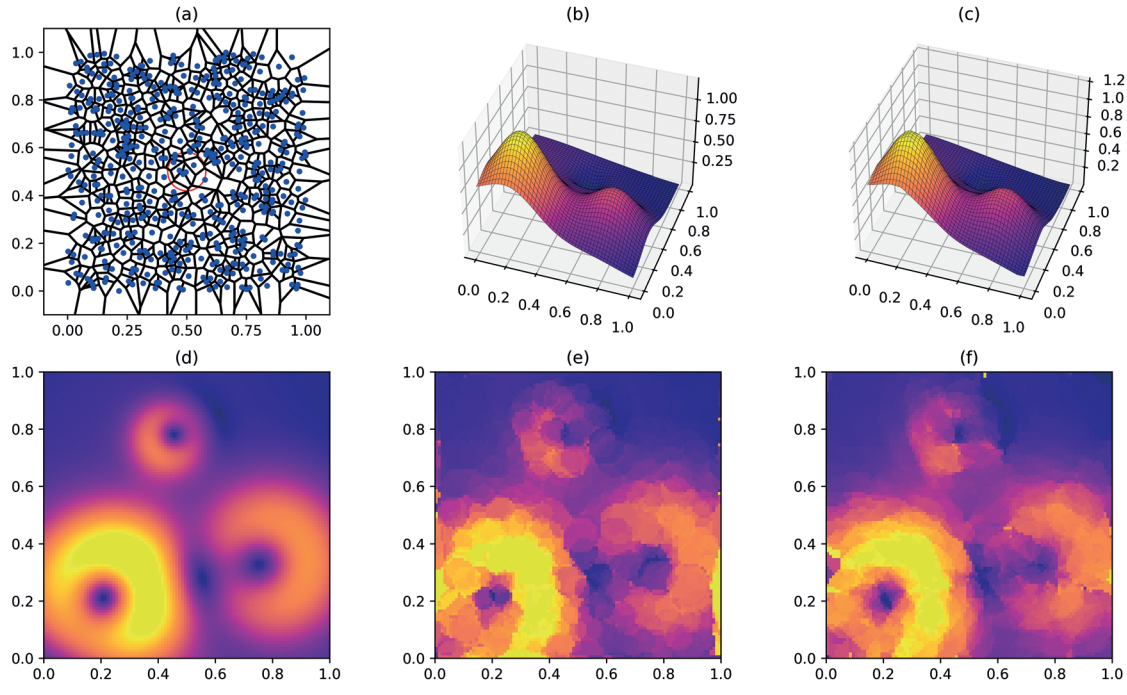

Fig. 4 - The Franke function on a data set of 500 random points: a) Voronoi tessellation; b) exact function; c) NN interpolated function; d) exact gradient magnitude $G(\mathbf{x})$; e) estimated gradient $G_n(\mathbf{x})$ with NNs; f) estimated gradient $G_r(\mathbf{x})$ with fixed radius.

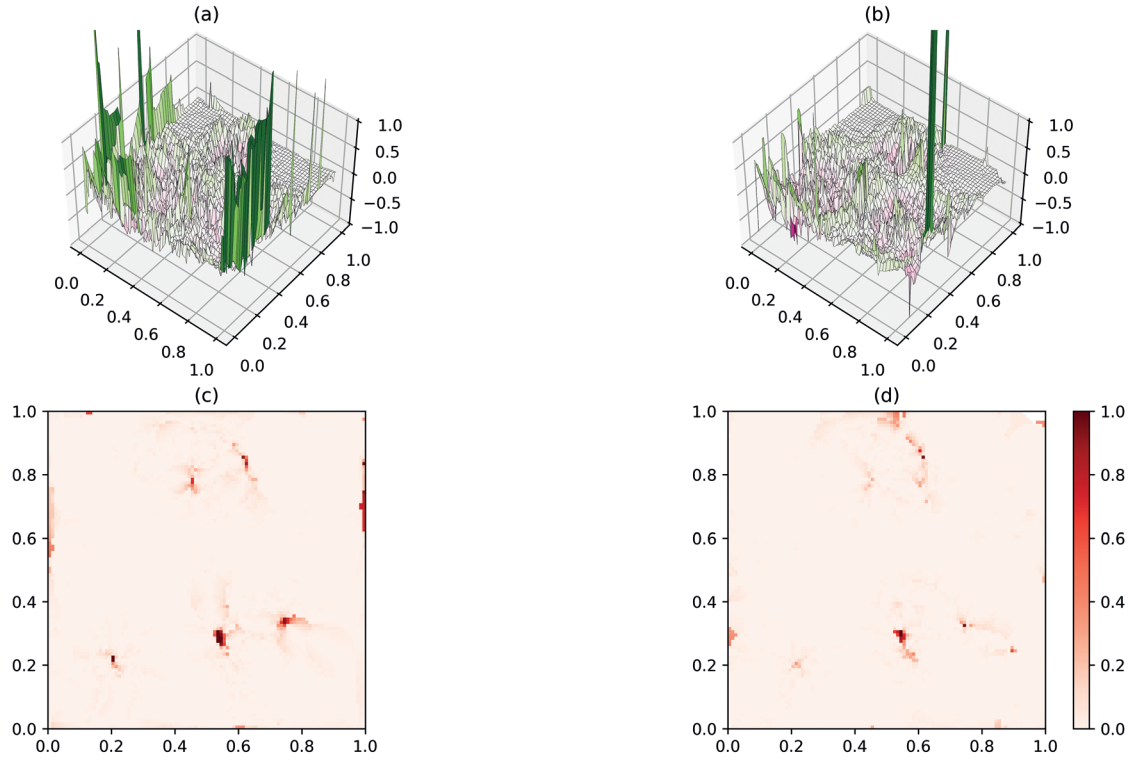

Fig. 5 - Errors for gradient estimates (Franke function, OLS method): a) (G_n-G) magnitude error for n estimator; b) (G_r-G) magnitude error for r estimator; c) h_n phase error for n estimator; d) h_r phase error for r estimator.

Table 1 - Magnitude gradient estimate errors with the OLS methods.

	***n* estimator (NNs) error G_n-G**	***r* estimator (fixed radius) error G_n-G**
Errors		
Mean	0.001874	-0.02568
Variance	0.08162	0.06519
RMSE	0.2857	0.2566
Skewness coefficient	180.1	831.8
Absolute errors		
Quartile Q1	0.02389	0.02528
Median Q2	0.07959	0.08112
Quartile Q3	0.1837	0.1711
Q3-Q1	0.1598	0.1458
min	2.52e-07	6.382e-07
Max	3.958	8.076
Bivariate statistics		
Pearson's correlation coefficient	0.9444	0.9543
Moment of inertia	4.081	3.292
Linear regression offset	0.05616	0.05052
Linear regression slope	0.9589	0.9423

There are many possible ways to combine gradient estimation and our equations. Since the Franke function is known, it is also possible to obtain a 'semi-exact estimator' by using the exact gradient. The only issue is given by the true location of points ξ_i, which must be approximated by $\boldsymbol{x}^*$, $\boldsymbol{x}_i$, the midpoints, or by other means. Table 2 lists the proposed methods.

Table 2 - Estimated interpolation errors.

Symbol	Method description	Approximation	Ref. formula	Estimated error
e_0	Exact error		$(\tilde{f} - f^*)$	
e_m	Semi-exact, midpoint	$\nabla f(\boldsymbol{\xi}_i) \cong \boldsymbol{g}\left(\frac{\boldsymbol{x}_i + \boldsymbol{x}^*}{2}\right)$	Eq. 10	Relative
e_i	Semi-exact, x_i	$\nabla f(\boldsymbol{\xi}_i) \cong \boldsymbol{g}(\boldsymbol{x}_i)$	Eq. 10	Relative
e_r	Vectorial estim., x_i	$\nabla f(\boldsymbol{\xi}_i) \cong \boldsymbol{g}_r(\boldsymbol{x}_i)$	Eq. 10	Relative
e_s	Scalar estim., x_i	$\|\nabla f(\boldsymbol{\xi}_i)\| \cong \boldsymbol{G}_r(\boldsymbol{x}_i)$	Eq. 11	Absolute
e_c	Scalar estim., x^*	$\|\nabla f(\boldsymbol{\xi}_i)\| \cong \boldsymbol{G}_r(\boldsymbol{x}^*)$	Eq. 11	Absolute

We cannot use an *n*-method for the gradient estimate, because the NNs of an NN are coincident with the NN itself, so the subset for the OLS estimation would be useless. Therefore, we only applied an *r*-method, generating the e_r estimated error. The results are compared in Fig. 6.

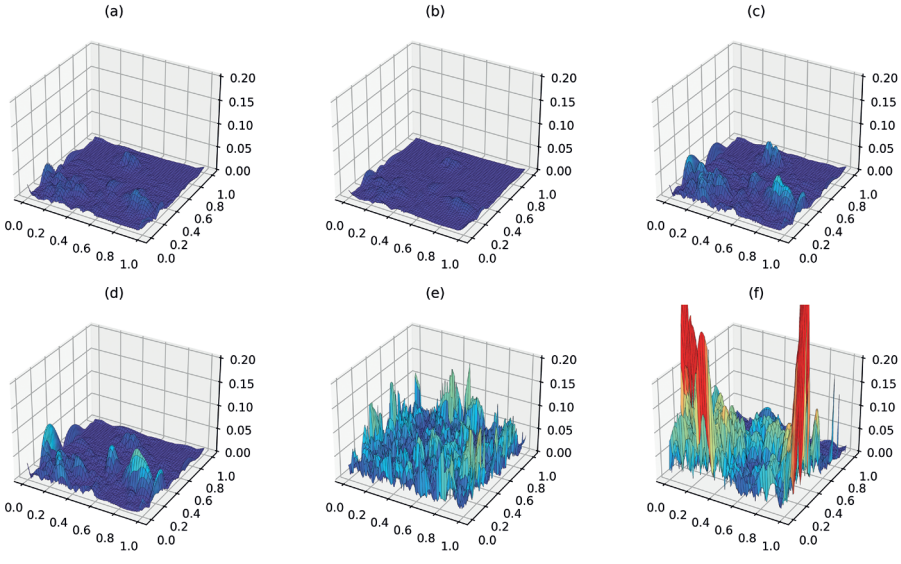

Fig. 6 - Interpolation errors obtained with different estimation techniques: a) exact error e_0; b) error e_m; c) error e_i; d) error e_r; e) error e_s; f) error e_c. Errors e_c and e_s are absolute errors, while the others are relative errors.

A statistical analysis between exact interpolation errors and estimated interpolation errors provides more details, as shown in Fig. 7 and Table 3.

4.4. Norcia test data set

To show an example of uncertainty estimation using the gradient method in a real-world scenario, we used the *PGA* surveys of the *M* 6.5 Norcia earthquake. This event occurred in central

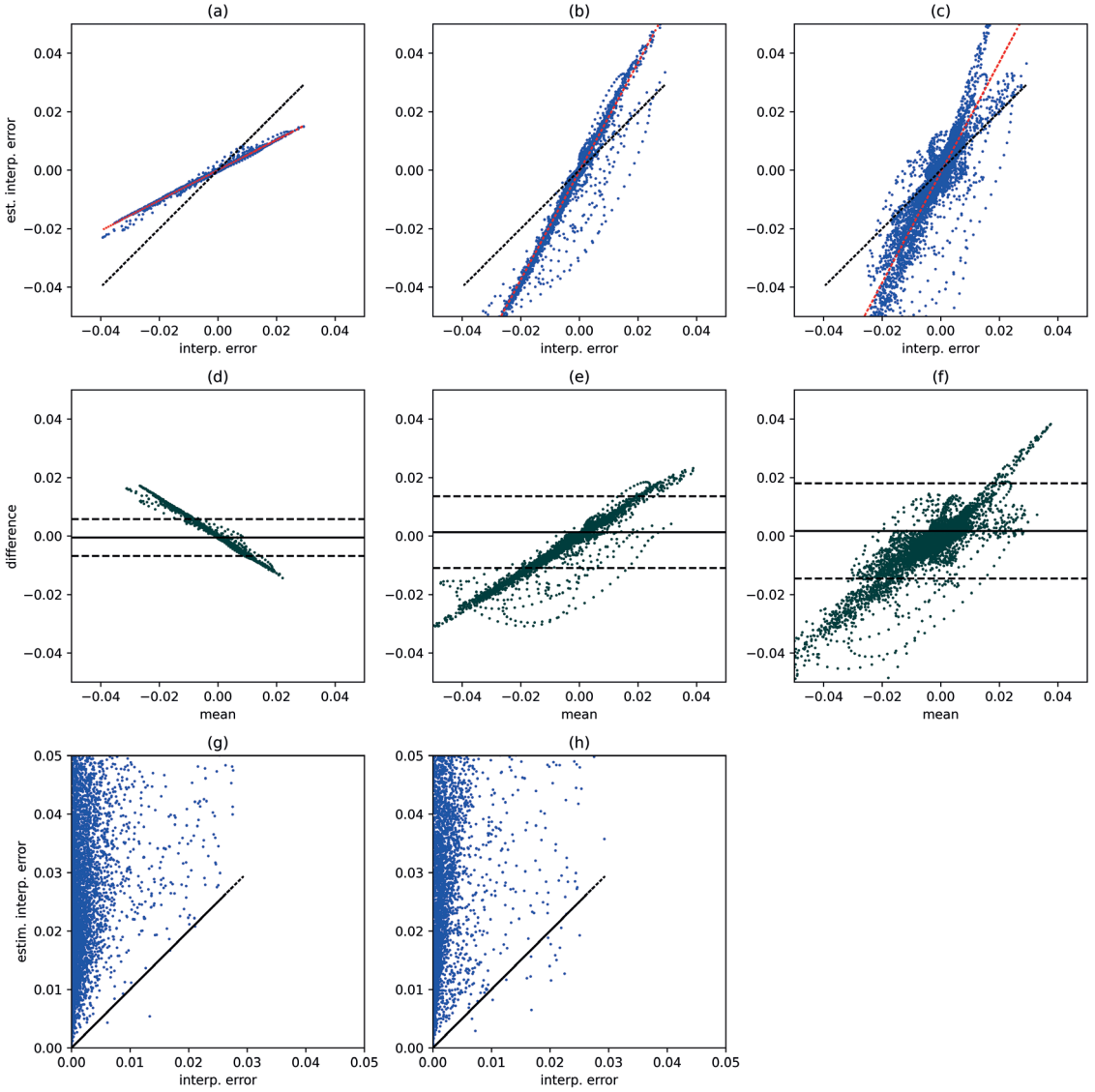

Fig. 7 - Scatter plot of estimated interpolation errors vs. exact errors, where the red line is the linear regression and the black line the exact estimation: a) plot for e_m; b) plot for e_i; c) plot for e_r; g) plot for e_s; h) plot for e_c. Bland-Altman plot (dashed lines show ±1.96 σ): d) plot for e_m; e) plot for e_i; f) plot for e_r.

Italy on 30 October 2016. The ITACA database [version 3.2: Russo *et al.* (2022)] provides a total of 296 registrations.

An amount of 164 data, within a distance of 200 km from the epicentre (at coordinates 42.832° N; 13.111° E), was selected. The nucleation, at a depth of 9.2 km, was located 5 km NE from Norcia. It was a normal mechanism, with strike in the direction of the central Italian axes (Apennine trends: NNW - SSE). The *PGA* range spans from 2 to 1,000 cm/s^2 (temporary station MZ24 - east component), but the data are presented here as $\log_{10}(PGA)$.

The WGS84 coordinates are projected into UTM33. Of utmost importance is the use of a metric projection or the application of other methods in order to take into account the anisotropy underlying latitude/longitude representation, as described in Iurcev *et al.* (2021).

Table 3 - Statistics of estimated interpolation errors vs. exact errors: e_s^* is an absolute error, while the others are relative errors.

	e_m	e_i	e_r	e_s^*
Errors				
Mean	0.00052	-0.0013	-0.0018	0.033
Variance	1e-05	3.9e-05	6.9e-05	0.00049
RMSE	0.0033	0.0064	0.0085	0.039
Skewness coefficient	3.6e+07	-5.5e+06	-3.8e+06	1.4e+05
Absolute errors				
Quartile Q1	-0.0007	-0.002	-0.0023	0.017
Median Q2	-3.9e-05	3.1e-05	-8e-05	0.027
Quartile Q3	0.00098	0.0012	0.0011	0.043
Q3-Q1	0.0017	0.0032	0.0034	0.027
Min	-0.014	-0.032	-0.058	0.00015
Max	0.017	0.023	0.038	0.17
Bivariate statistics				
Pearson's correlation coeff.	1	0.98	0.91	-0.12
Moment of inertia	0.00053	0.0021	0.0036	0.078
Linear regression offset	8.9e-06	-0.00039	-0.00081	0.031
Linear regression slope	0.52	1.9	1.9	-0.37

Fig. 8a shows the distribution of stations with the Apennine alignment clearly visible. Fig. 8b shows the 3D NN interpolated surface of $\log_{10}$ *PGA*, with the typical decay with distance. Finally, Figs. 8c and 8d show the distribution of uncertainties calculated with the fixed radius OLS (e_r) estimation method. The maximum errors (2.5) near the border are close to the upper range limit for $\log_{10}$ *PGA* values, while in the middle the errors do not exceed 0.5.

In Fig. 8a, the data set is represented with blue and red dots. The red dots are data where the OLS gradient approximation provides very poor results due to the fact that the least-squares matrix has a rank <2. In the peripheral regions, where the data set is sparser and the Voronoi polygons are larger, both interpolation and error estimation tend to be unreliable.

The implemented method for error estimation was e_r, based on the OLS gradient estimate with a fixed radius r = 25 km, applying the vectorial formula in Eq. 10 with $\nabla f(\xi_i) \cong g_r(x_i)$.

5. Discussion

Interpolation error assessment for the Sibson method (NN) is not a straightforward process. The most difficult aspect is potentially the geometric definition of weights, w_i, which is hard to analytically express and depends solely on the spatial distribution of the Voronoi polygons. Moreover, the properties of the interpolated function are generally unknown, so hypotheses related to them must be assumed. If differentiability is assumed, a possible deterministic approach is to apply the mean value theorem in the neighbourhood of the interpolation point.

A completely different approach is based on a geostatistical point of view, and takes into account the statistical properties of the function inferred from the data set. This is what the Kriging method does and this possible approach is mentioned in the Appendix.

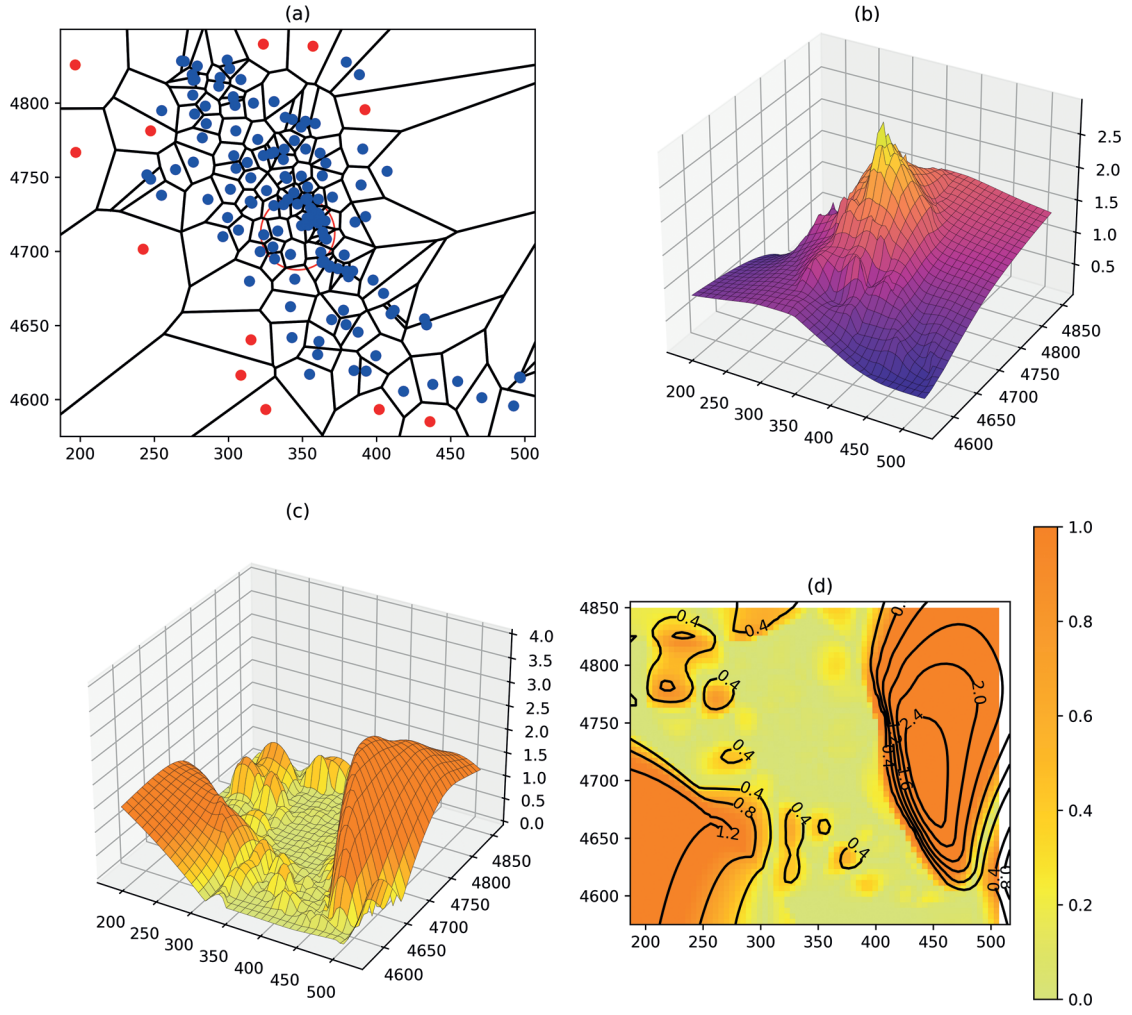

Fig. 8 - NN interpolation error estimation for Norcia data set: a) data set and Voronoi; b) interpolated surface; c) and d) estimated errors e_r.

The deterministic methods proposed here, based on the MVT, raise two major issues. The first problem is the unknown location of points ξ_i along the line between the interpolation point and the *i*-th NN. These points can be approximated by means of the NNs themselves or a point at half distance, etc.

The second issue concerns gradient estimation. Here, we have applied a method that employs OLS linear approximation, exploring two different techniques for choosing the data subset: a fixed radius or the NN. The former raises an additional problem, since the optimal radius is strictly dependent on the local spatial density of the data set. The latter bypasses this problem, but, at the same time, is not feasible if we approximate points $\boldsymbol{\xi}_i$ with relative NNs, since the vectorial expression of Eq. 10 becomes zero.

All these possibilities define several sub-techniques for deterministic error estimation; however, we have implemented and tested only some of them here. The test bed was the Franke test function, with a random data set generated as a Poisson point process. This is, to some extent, a Montecarlo approach for an assessment and comparison relying upon statistics.

Some interesting results, highlighting possible caveats and applications, can be observed.

The midpoint technique (e_m) seems to underestimate the interpolation error, while the other methods, on the contrary, seem to overestimate it.

The vectorial method, using OLS gradient approximation with fixed radius (e_r), gives better results, while the scalar methods (using Eq. 11) are far too unstable to be considered. The scatter plots exhibit a moderate degree of correlation with an offset close to zero, in the best cases. However, there is a significant difference between the linear regression slope and the ideal correlation line, which seems to imply a multiplicative factor.

It would appear that the fixed radius method is the most promising, although dependence on λ must be taken into account and should not be based on an arbitrary choice as in our examples.

Finally, we made an attempt at error estimation in a real-world scenario by applying the method to a small data set related to the Norcia earthquake. The data set has a spatial distribution, which is closely aligned with the Apennine trend (NNW-SSE). It was interpolated using the NN method and the error estimate was better in the central region where λ was higher and the Voronoi polygons denser.

6. Conclusions

We attempted to confront the bivariate NN interpolation error estimation problem with an irregular and, possibly, sparse data set. We explored some possible techniques based on the MVT, under the hypothesis of differentiability of the unknown function, f.

Some preliminary results of our investigations suggest that the OLS gradient estimation with a fixed radius can provide reasonable estimates. Many interesting issues that have arisen, require further investigation from both theoretical and experimental perspectives.

In the Appendix, we have also attempted to outline a geostatistical approach relative to error assessment or convex interpolation methods, like NN, and restricted to simple semivariogram models.

Acknowledgments. We are grateful to Daniele Del Santo of the University of Trieste and Omar Lakkis of the University of Sussex, for the discussion on the deterministic approach. We also thank Maximilian Majstorović, student at the University of Trieste, for his contribution in numerical problem solving and code implementation during his internship at the National Institute of Oceanography and Applied Geophysics - OGS, Trieste, Italy within his master's degree program. We acknowledge the contributions of Alberto Tamaro in providing GIS support. Finally, we wish to thank Sebastiano Trevisani of the IUAV University of Venice, for sharing important information on geostatistics.

REFERENCES

Aurenhammer F., Klein R. and Lee D.T.; 2013: *Voronoi diagrams and delaunay triangulation*. World Scientific Publishing Co. Pte. Ltd., Munich, Germany, 348 pp., doi: 10.1142/8685.

Barnes S.L.; 1964: *A technique for maximizing details in numerical weather-map analysis*. J. Appl. Meteorol. Climatol., 3, 396-409, doi: 10.1175/1520-0450(1964)003<0396:ATFMDI>2.0.CO;2.

Belward J.A., Turner I.W. and Ilić M.; 2008: *On derivative estimation and the solution of least squares problems*. J. Comput. Appl. Math., 222**,** 511-523, doi: 10.1016/j.cam.2007.11.022.

Chilès J.P. and Delfiner P.; 1999: *Geostatistics*: *modeling spatial uncertainty, 1st ed.* John Wiley & Sons Inc., New York, NY, USA, 695 pp., doi: 10.1002/9780470316993.

De Keyser J., Darrouzet F., Dunlop M.W. and Décréau P.M.E.; 2007. *Least-squares gradient calculation from multi-point observations of scalar and vector fields: methodology and applications with Cluster in the plasmasphere*. Ann. Geophys., 25, 971-987, doi: 10.5194/angeo-25-971-2007.

Etherington T.R.; 2020: *Discrete natural neighbour interpolation with uncertainty using cross-validation error-distance fields*. PeerJ. Comput. Sci., 6, e282, 16 pp., doi: 10.7717/peerj-cs.282.

Franke R.; 1979: *A critical comparison of some methods for interpolation of scattered data*. Naval Postgraduate School, Monterey, CA, USA, Technical Report, NPS-53-79-003, **??? pp.,** <hdl.handle.net/10945/35052>.

Gandin L.S.; 1965: *Objective analysis of meteorological fields*. U.S. Department Commerce and National Science Foundation, Washington, D.C., USA, 242 pp.

Ghosh S., Gelfrand A.E. and Mølhave T.; 2012: *Attaching uncertainty to deterministic spatial interpolations*. Stat. Methodol., 9, 251-264, doi: 10.1016/j.stamet.2011.06.001.

Goovaerts P.; 1997: *Geostatistics for natural resources evaluation*. Oxford University Press, Oxford, UK, 497 pp.

Herzfeld U.C.; 1996: *Inverse theory in the earth sciences - An introductory overview with emphasis on Gandin's method of optimum interpolation.* Math. Geol., 28, 137-160, doi: 10.1007/BF02084210.

Iurcev M., Pettenati F. and Diviacco P.; 2021: *Gridding, boundary definition and interpolation methods for near real-time spatial data.* Bull. Geoph. Ocean., 62, 427-454, doi: 10.4430/bgta0360.

Okabe A., Boots B., Sugihara K. and Chiu S.N.; 2000: *Spatial tessellation: concepts and applications of Voronoi diagrams, 2nd ed*. John Wiley & Sons Inc., New York, NY, USA, 700 pp.

Russo E., Felicetta C., D'Amico M.C., Sgobba S., Lanzano G., Mascandola C., Pacor F. and Luzi L.; 2022: *ITalian ACcelerometric Archive (ITACA) version 3.2*. Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy, doi: 10.13127/itaca.3.2.

Sibson R.; 1980: *A vector identity for the Dirichlet tessellation*. Math. Proc. Camb. Phil. Soc., 87, 151-155.

Sibson R.; 1981: *A brief description of natural neighbor interpolation*. In: Barnett V. (ed), Interpreting Multivariate Data, John Wiley & Sons Inc., New York, NY, USA, pp. 21-36.

Stead S.E.; 1984: *Estimation of gradients from scattered dat*a. Rocky Mt J. Math., 14, 265-279.

Thiesen S. and Ehret U.; 2022: *Assessing local and spatial uncertainty with nonparametric geostatistics*. Stochastic Environ. Res. Risk Assess., 36, 173-199, doi: 10.1007/s00477-021-02038-5.

Corresponding author: Franco Pettenati
Istituto Nazionale di Oceanografia e di Geofisica Sperimentale - OGS
Borgo Grotta Gigante 42c, 34010 Sgonico (TS), Italy
Phone: +39 040 2140???; e-mail: fpettenati@ogs.it

Appendix: A geostatistical approach

A geostatistical approach considers the unknown function, $f(.)$, as the realisation of a random process. Thus, we may think in terms of variance and covariance, where the latter is:

$$\mathrm{cov}\left[f(\mathbf{x_i}),\, f\left(\mathbf{x_j}\right)\right] \;=\; \mathrm{cov}\left(f_i,\, f_j\right) \;=\; \mathbf{E}\left\{\left[f_i \;-\; \mathbf{E}(f_i)\right]\left[f_j \;-\; \mathbf{E}\left(f_j\right)\right]\right\}. \tag{A1}$$

Let us adopt a temporary notation for the indices: if $\{\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n\}$ are the n NNs, $\boldsymbol{x}_0 = \boldsymbol{x}^*$ is the interpolation point, and $\boldsymbol{w}_0 = -1$ is defined, then the variance of the interpolation error ϵ is:

$$\mathrm{var}(\epsilon) = \mathrm{var}\left(\tilde{f} - f_0\right) = \mathrm{var}\left(\textstyle\sum_{i=0}^{n} w_i f_i\right) \tag{A2}$$

and for the properties of the variance of a linear combination:

$$\mathrm{var}(\epsilon) = \mathrm{var}(f_0) + \sum_{i=1}^{n} w_i^2 \mathrm{var}(f_i) - 2\sum_{i=1}^{n} w_i \mathrm{cov}\,(f_i, f_0) + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} w_i w_j \mathrm{cov}\,(f_i, f_j). \quad \text{(A3)}$$

Reverting to our original index notation and assuming the random process stationarity, which entails a constant variance, the following is obtained:

$$\mathrm{var}(\epsilon) = \mathrm{var}(f)(1 + \sum_{i=0}^{n-1} w_i^2) - 2\sum_{i=0}^{n-1} w_i \mathrm{cov}\,(f_i, f^*) + \sum_{i=0}^{n-1}\sum_{j=0, j\neq i}^{n-1} w_i w_j \mathrm{cov}\,(f_i, f_j). \quad \text{(A4)}$$

Incidentally, the stationarity hypothesis also ensures that the first moment of the error is null, because, if **E**(*f*) = constant, then:

$$\mathbf{E}(\epsilon) = \mathbf{E}(\sum_{i=0}^{n-1} w_i f_i - f^*) = \mathbf{E}(f)(\sum_{i=0}^{n-1} w_i - 1) = 0. \quad \text{(A5)}$$

For the stationarity, a theoretical semivariance, which is a function of the distance, can also be defined as follows:

$$\gamma(d) = 2\,\mathrm{var}[f(\mathbf{x}) - f(\mathbf{x} + d)] \quad \text{(A6)}$$

and:

$$\gamma(d) = \mathrm{cov}(0) - \mathrm{cov}(d) = \mathrm{var}(f) - \mathrm{cov}(d). \quad \text{(A7)}$$

Subsequently, since cov(f_i, f^*) = cov(d_i) and cov(f_i, f_j) = cov(d_{ij}), from Eqs. 4, A4, and A7, it can be inferred that (the summation indices are always from 0 to *n*-1):

$$\mathrm{var}(\epsilon) = 2\sum_i w_i \gamma(d_i) - \sum_i \sum_{j\neq i} w_i w_j \gamma(d_{ij}). \quad \text{(A8)}$$

A1. Linear model

Considering a linear semivariogram $\gamma(d) = (c_0 + c_1 d)$, which is a valid model in many practical cases and especially if *d* is small, and taking into account that in our case $d \leq \max_i (d_i)$, then from Eq. A8:

$$\mathrm{var}(\epsilon) = c_0(1 + \sum_i w_i^2) + c_1(2\sum_i w_i d_i - \sum_i \sum_{j\neq i} w_i w_j d_{ij}) = c_0 r_0 + c_1 r_1. \quad \text{(A9)}$$

The r_0 and r_1 functions, depending only on the geometrical disposition of our data set, are both dimensionally equal to a distance and are functions of the interpolation point. In Fig. A1, $r_1(\boldsymbol{x}^*)$ is represented over a Voronoi tessellation sample, in a Sibson interpolation (NN). The plot represents the error with $c_0 = 0$; $c_1 = 1$.

The r_1 function has its zeroes on data set points $\boldsymbol{x}_i$, where it is also not differentiable. Elsewhere, it represents how big the interpolation error is, assuming the stationarity condition and a local linear semivariogram with null nugget ($c_0 = 0$).

If we consider a non-negative nugget ($c_0 \geq 0$), then $r_0(\boldsymbol{x}^*)$ can be plotted as in Fig. A2. The plot represents the error with $c_0 = 1$; $c_1 = 0$.

The total error variance is the linear combination of r_0 and r_1, as in Eq. A9.

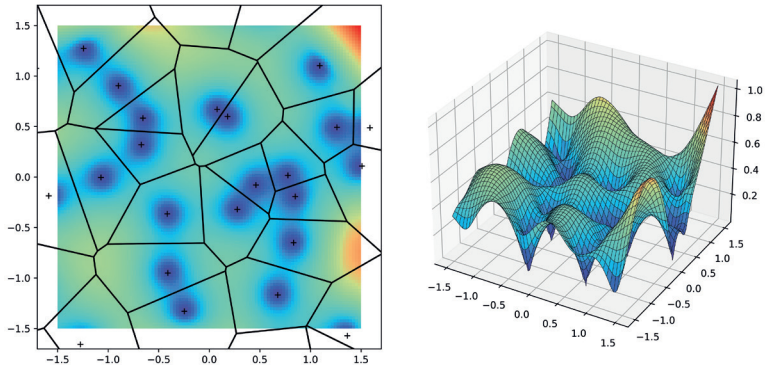

Fig. A1 - Plot of $r_1(\boldsymbol{x}^*)$ for NN interpolation.

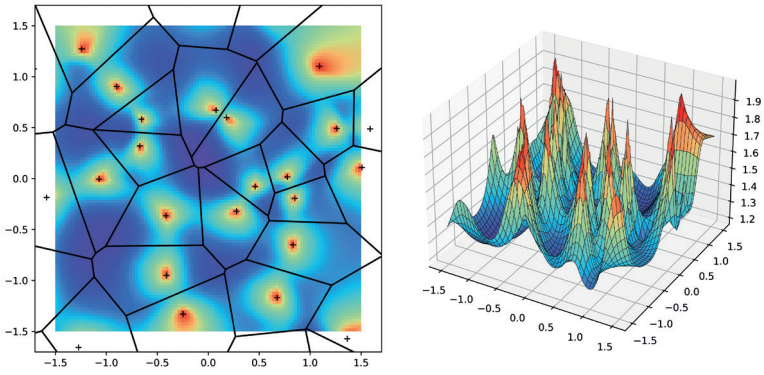

Fig. A2 - Plot of $r_0(\mathbf{x}^*)$ for NN interpolation.

A2. Exponential model

An exponential semivariogram model is in the form:

$$\gamma(d) = 1 - e^{-\frac{d}{c}}. \tag{A10}$$

Considering its Taylor expansion around $d = 0$:

$$\gamma(d) = \sum_{k=1}^{\infty} \frac{d^k(-1)^{k+1}}{c^k k!}. \tag{A11}$$

Substitution in Eq. A8 results in:

$$\mathrm{var}(\epsilon) = \sum_{k=1}^{\infty} \frac{1}{c^k} \frac{(-1)^{k+1}}{k!} \left(2\sum_i w_i d^k - \sum_i \sum_{j \neq i} w_i w_j d_{ij}^k\right) = \sum_{k=1}^{\infty} \frac{r_k}{c^k}. \tag{A12}$$

If we define:

$$r_k = \frac{(-1)^{k+1}}{k!} \left(2\sum_i w_i d^k - \sum_i \sum_{j \neq i} w_i w_j d_{ij}^k\right) \tag{A13}$$

and r_0, r_1 corresponds to the results for the linear model.