Reasoning about the presence of outliers in the European quasi-geoid data set

Damiano Triglione
Politecnico di Milano, Polo Regionale di Como, Italy
damiano.triglione@polimi.it

Abstract

The purpose of this work is studying the (possible) relation between the distribution of the outliers found in the European quasi geoid data set and the magnitude of the slope (gradient) in that points, either in the same signal or in the topographic heights' signal. As we will see, this relation is highlighted by a Montecarlo method. A particular attention has been paid to the fact that, in a so wide area, the east-west step (used to calculate the gradient magnitude in the gridded data) has to be adapted to the changing curvature of the ellipsoid.

A definition of "Outlier"

An intuitive definition of outlier (taken from [14]) can be: "an observation that stands so away from the other ones that it leads to suspect that it could have been generated by a different mechanism" (another insight is presented also in [2]).

Therefore the "evidence" of the deviation presented by the erroneous data is the indicator of their probable incompatibility with the remaining data. Historically the error rejection was performed by people who were responsible of data preprocessing: watching the globality, they were able to identify the suspected minority. Nowadays an automatic procedure is desirable.

A deep exam of anomalous data is imposed by the awareness that just one outlier can "pollute" a big set of measures.

Introduction

In IGeS archives, several geoid data sets are collected. For the european area (North=77°, South=25°, East=67.5°, West=-35°), there are available two kinds of data: QuasiGeoid and Geoid (see for reference [9],[10],[26],[27]). Since a new project for the European Geoid re-computation is under definition, it was considered as particularly interesting to examine the old geoid data, chasing for outliers. We have then decided to apply our own software (hereafter described) to the two datasets, finding that about 90% of the outliers in each data set are common. So we decided to focus our attention only on the QuasiGeoid.

The software "r.outdet", used for the current outlier rejection, is an evolution of "r.outldetect", developed a few years ago. Both the modules were devoleped by us in the last years, exploiting also interesting routines offered in [19]. The actual release extends the capabilities of the open source GIS G.R.A.S.S., a software environment full of tools for processing spatial distributed data of any kind.

The software's core method of processing is placed in the analysis of every point of the data set, around which a moving window is opened. The size of this window is user-defined, and determines how many surrounding points will be involved in the test; the target is to decide whether the central point is an outlier.

More precisely: for a rectangular window of side $(2k+1)\delta_x$ x $(2k+1)\delta_y$ (of course it should be an odd number, while k is any integer larger than zero), when the data are gridded, there are

$$N_k = (2k+1)^2 - 1 = 4k\,(k+1)$$

observations $h_{obs,i}$ around the central h_{obs}. With an appropriate interpolation model, and using only the surrounding values, the central value $\hat{h}$ is estimated and compared with the central observation h_{obs}, in order to implement a statistical test on the hypothesis

$$H_0: E[\Delta h] = E[\hat{h} - h_{obs}] = 0$$

The statistics used to perform this test is obtained by a Least Square Estimation approach for the coefficients of a polynomial that constitutes the interpolation model. For example, if the user chooses the constant function as local model (see [15]), there is only one coefficient to estimate:

$$h_{model}(x,y) = a_0$$

If every observation $h_{obs,i}$ is assumed, as independent, with normal distribution (mean a_0 and variance s_0^2), we have

$$\hat{a}_0 = \frac{1}{N_k}\sum_{i=1}^{N_k} h_{oss,i} \ \sim\ N[a_0, s_0^2 / N_k]$$

A correct estimator of s_0^2 is

$$\hat{s}_0^2 = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}(h_{oss,i} - \hat{a}_0)^2 \sim \frac{s_0^2}{N_k - 1}\chi^2_{(N_k - 1)}$$

so, since

$$\Delta h = h_{oss} - \hat{a}_0 \ \sim\ N[0, s_0^2 (1 + 1/N_k)]$$

we have that the test is performed by

$$\sqrt{\frac{N_k}{N_k + 1}}\frac{\Delta h}{\hat{s}_0} \sim \frac{Z}{\sqrt{\frac{\chi^2_{N_k - 1}}{N_k - 1}}} = t_{(N_k - 1)}$$

Remark

Indeed real errors with respect to simple models like these are not expected to be neither independent nor normally distributed. The experience, however, says that traditional tests do tend to be conservative, i.e. to identify as outliers data that could not be so with respect to more realistic distributions. Since outliers as such are not “false” data but just data with a different statistical signature and have to be more closely analyzed, we think that the proposed procedure is still acceptable.

See [4] to read about another conservative approach to outlier detection, while the subjective nature of outlier rejection procedures is in [7].

The outlier rejection strategy

Actually, our outlier rejection is based on the following polynomials, well known in literature on approximation and reliability, that are a straightforward generalization of the above elementary example.

- Bilinear (4 coefficients):

$$h_{bil}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy$$

- Bicubic (16 coefficients):

$$h_{bic}(x,y) = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot xy + a_4 \cdot x^2 + a_5 \cdot y^2 + a_6 \cdot x^2 y + a_7 \cdot xy^2 + a_8 \cdot x^2 y^2 +$$
$$+ a_9 \cdot x^3 + a_{10} \cdot y^3 + a_{11} \cdot xy^3 + a_{12} \cdot x^3 y + a_{13} \cdot x^2 y^3 + a_{14} \cdot x^3 y^2 + a_{15} \cdot x^3 y^3$$

Since it is important to keep a high overdetermination, for the first case we chose a window size of 3x3, while for the second one a size of 7x7.

The test procedure is then performed by observing that, in both cases,

$$\hat{h} = \hat{a}_0$$

Since $\hat{a}_0$, in the case of gridded data, is still given by

$$\hat{a}_0 = \frac{1}{\mathrm{N}_k} \sum_{i=1}^{\mathrm{N}_k} \mathrm{h}_{oss,i}$$

for the bilinear interpolator and by a more complicated formula – coming from the Least Square Adjustment – for the bicubic interpolator, by exploiting a standard testing theory, we can write

$$\Delta\mathrm{h} = \mathrm{h}_{oss} - \hat{h} \sim \mathrm{N}[0, s_0^2 (1+q)]$$

where q depends on the number of unknown parameters (in other words, the degree of the polynomials plus 1) and the size of the moving window. It can be found that for both the methods (bilinear with 3x3 and bicubic with 7x7) the value of q is 1/8.

Therefore the test statistics applies (see for instance [3])

$$\frac{\mathrm{h}_{oss} - \hat{h}}{\hat{s}_0 (9/8)} \sim t_n$$

where t_ν denotes the student's t distribution with ν degrees of freedom (the number of observations reduced by the number of unknown parameters). For the bilinear case (with window size 3) ν=8-4=4, while for the bicubic case (window size 7) ν=48-16=32.

In table 1 there are the results, according to different significance levels α.

It shows the number of suspected outliers for the present test, to be compared with the number of suspected outliers expected because we fix a certain α risk of the first kind (for a normal distribution, given α, we expect αN data to be suspected).

The results point to a significant non normality of the data; we have just seen that in the 3x3-bilinear interpolation the redundancy (number of degrees of freedom, ν) is 4 while in the 7x7-bicubic interpolation is 32; therefore, with α=1%, the corresponding t values are 4.60 and 2.75. This explains why we find more outliers in the bicubic case.

The choice of the α value to be used (in our case α=0.1%) has been performed by considering that the suspected outliers should be in any way more than the expected, but not too many.

Method	***size***	***Alfa***	***"Outliers"*** (out of 127920 data)	***Expected "Outliers"*** (out of 127920 data)
Bicubic	7x7	5%	15061	6396
Bicubic	7x7	1%	2290	1279
Bicubic	7x7	0.1%	179	128
Bicubic	7x7	0.05%	79	64
Bilinear	3x3	1%	290	1279
Bilinear	3x3	0.1%	17	128

Table 1

Focusing on the 179 outliers found with the third option, their distribution is shown in figure 1; figure 2, instead, shows them against the corrected gradient of the signal.

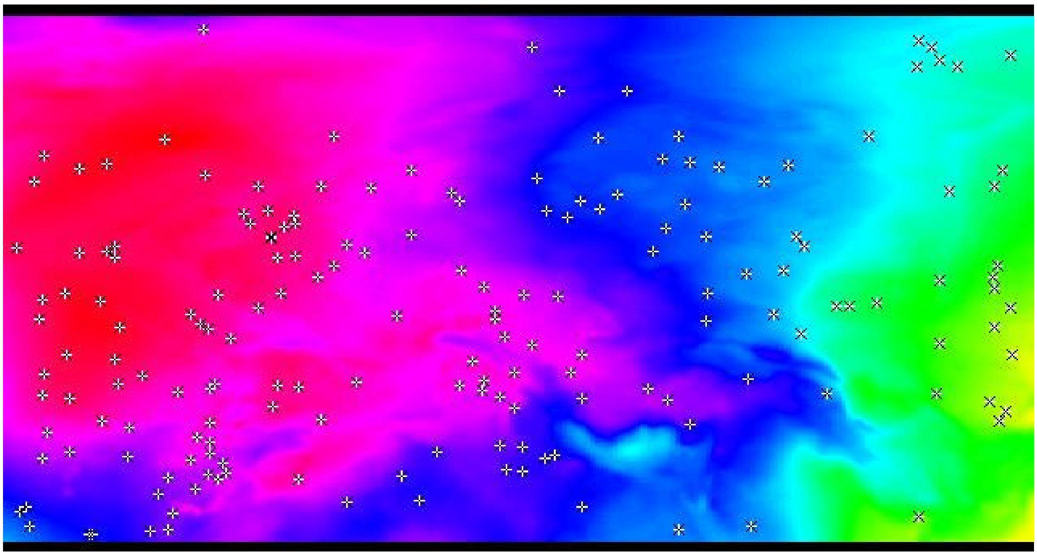

Figure 1 – 179 outliers against the QuasiGeoid signal

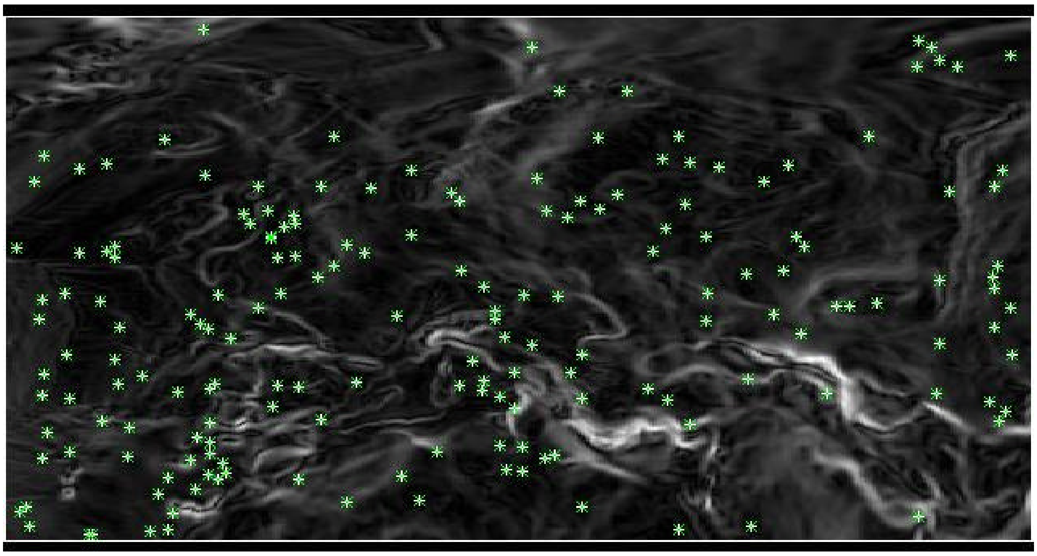

Figure 2 – 179 outliers against the gradient magnitude of the QuasiGeoid signal

It should be noted that the corrected gradient is computed by the following formula

$$\text{Grad* } G(j_{\text{i}},\lambda_{\kappa}) = \sqrt{\left(G^{*}_{NS}\right)^{2}+\left(G^{*}_{EW}\right)^{2}} \cong \sqrt{\left(G_{NS}\right)^{2}+\frac{1}{\cos^{2}(j_{i})}\left(G_{EW}\right)^{2}}$$

where

$G(j,\lambda)$ means the geoid signal evaluated in (j,λ);

* means approximately corrected according to the local curvature of the ellipsoidal coordinate lines;

$G_{\text{NS}} = 1/(8\ \Delta j)\ \cdot$

$$\{G(j_{\text{i-1}},\lambda_{\kappa-1})+2G(j_{\text{i-1}},\lambda_{\kappa})+G(j_{\text{i-1}},\lambda_{\kappa+1})–G(j_{\text{i+1}},\lambda_{\kappa-1})–2G(j_{\text{i}},\lambda_{\kappa})–G(j_{\text{i+1}},\lambda_{\kappa+1})\}$$

$G_{\text{EW}} = 1/(8\ \Delta l)\ \cdot$

$$\{G(j_{\text{i-1}},\lambda_{\kappa+1})+2G(j_{\text{i}},\lambda_{\kappa+1})+G(j_{\text{i+1}},\lambda_{\kappa+1})–G(j_{\text{i-1}},\lambda_{\kappa-1})–2G(j_{\text{i}},\lambda_{\kappa-1})–G(j_{\text{i+1}},\lambda_{\kappa-1})\}$$

It’s worth to investigate what kind (if any) of relationship there is between the location of the “outliers” and the gradient magnitude. At first sight, the geographical distribution of the outliers does not follow any kind of criterion based on the slope of the geoid. One way to have a confirmation of this statement is to proceed with a Montecarlo Method (for reference, [13]), by creating a suitable index that shows the general behaviour of the gradient magnitude against the behaviour of the same value when applied to the outliers. More precisely, we used the index

$$\Phi^{\mathbf{k}} = \sum_{\mathrm{i=1}}^{\#\,\mathrm{of\ outliers}} \log\left(1+\left|\nabla N(P_i)\right|\right)$$

where in this case number (#) of outliers is 179, while k (the repetition index) runs from 1 to 10000 and every point P_i is sampled from a uniform distribution on the knots of the grid. As it is evident, the index increases when points have systematically higher and higher gradient magnitude. The reader could ask why we did not choose a simpler expression, such as for instance

$$\Phi^{\mathbf{k}}_{simple} = \sum_{\mathrm{i=1}}^{\#\,\mathrm{of\ outliers}} \left|\nabla N(P_i)\right|$$

The reason stands in the limit of the “double precision” representation of float numbers: a computer cannot store, in its memory, big numbers having an arbitrary size. Using the logarithmic mapping, instead, we reduce the amplitude of the sum, without loosing any information because an increasing monotonic map preserves the ordering of the values. More over, adding one is required to make every addendum positive. In any way, since the boot-strap method applied here works with any distribution, we can in principle use any function.

The frequency distribution of Φ is compared with the single value

$$\Phi^{*} = \sum_{\mathrm{i=1}}^{\#\,\mathrm{of\ outliers}} \log\left(1+\left|\nabla N(P_i^{*})\right|\right)$$

where $\{P^*_i\}$ denotes the set of points where the outliers were found.

As it can be seen in figures 3a, 3b, 3c, 3d, the index Φ^* does not seem to be correlated with the gradient magnitude (in figures 3c and 3d it is close to the average of the empirical distribution) or slightly correlated, but in the opposite sense we expected (in figures 3a and 3b it is on the left, as if a smooth signal could lead to find more outliers).

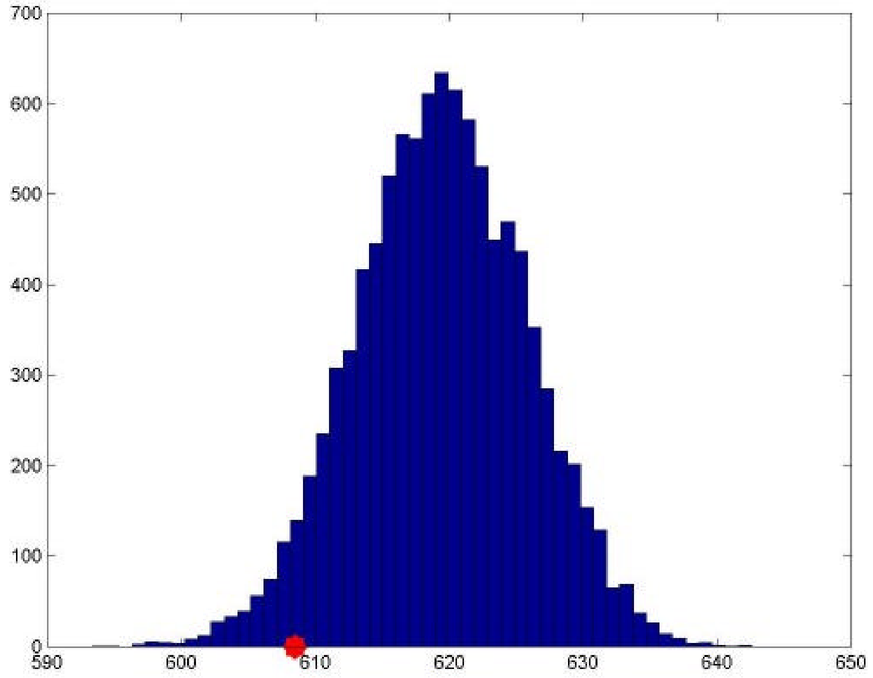

Figure 3a – Φ* with: bicubic interpolation, 7x7 window size, significance level 0.05%

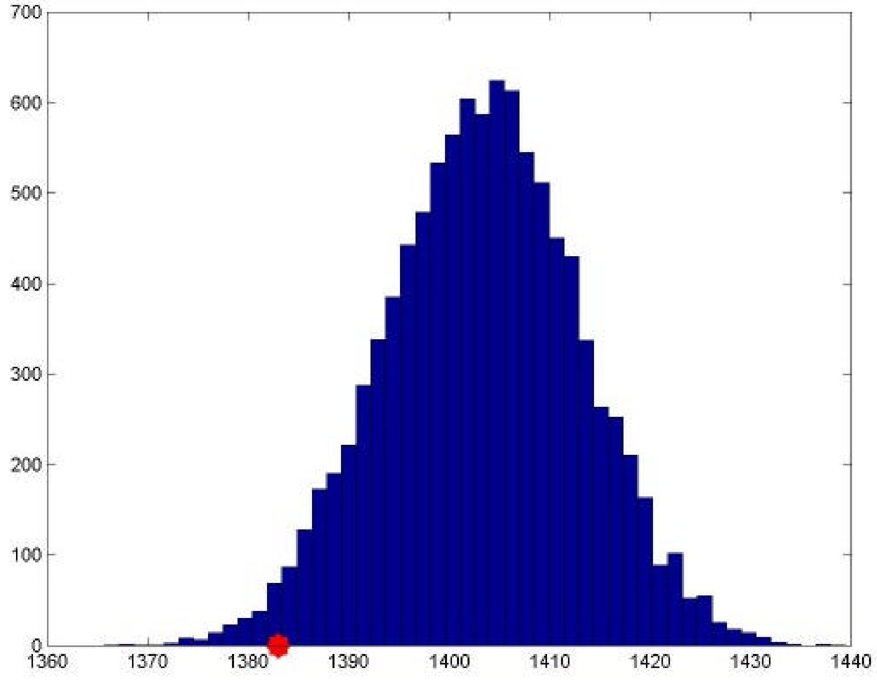

Figure 3b – Φ* with: bicubic interpolation, 7x7 window size, significance level 0.1%

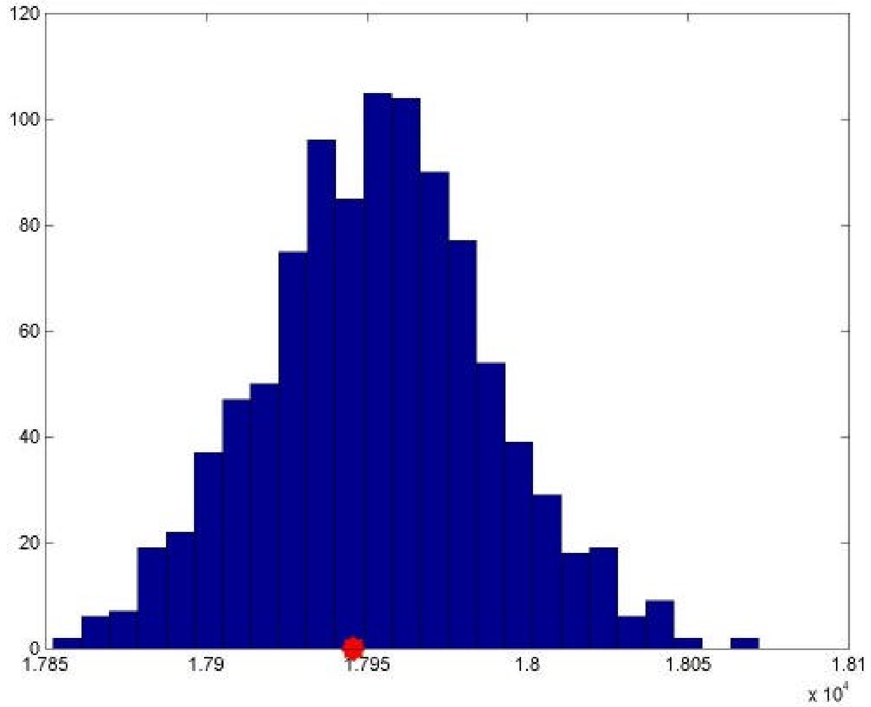

Figure 3c – Φ* with: bicubic interpolation, 7x7 window size, significance level 5%

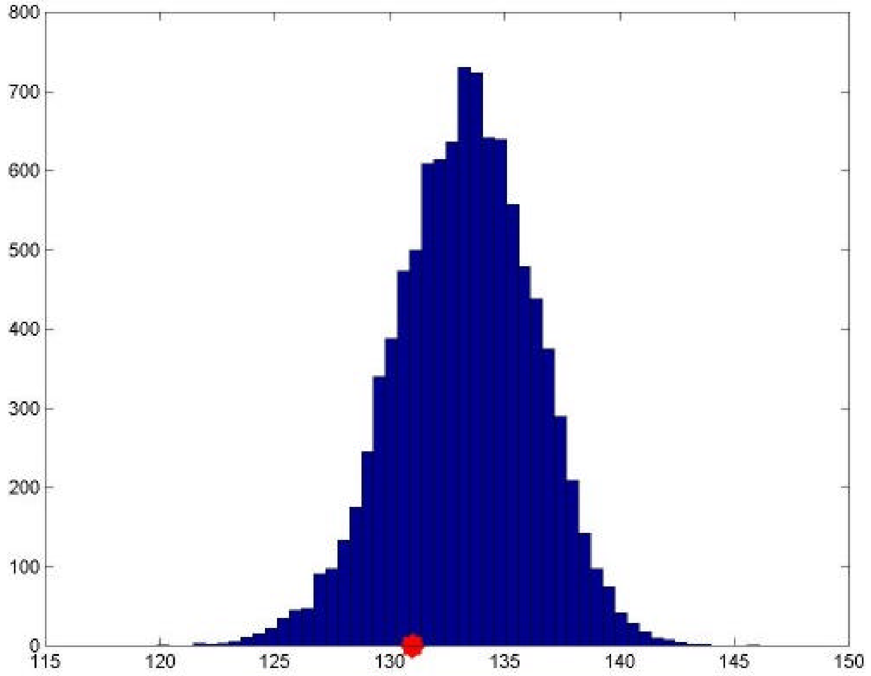

Figure 3d – Φ* with: bilinear interpolation, 3x3 window size, significance level 0.1%

Going deeper into the rejection criterion

This apparent lack of correlation, contrary to our intuition, can be studied more deeply. In fact we suspect that our “outliers” fall into two cathegories; in one we have outliers with large absolute residuals (we can call these “relevant outliers”), which we can consider as true outliers, in the other we have outliers just because the $\hat{s}_0$ resulting from the adjustment of the bicubic surface to the surrounding values is too small. Then we decided that we were more interested in looking into the relation between relevant outliers and inclination of the geoid. So, after inspecting the histogram of the “modulus” of the outliers (see figure 4), we decided to compare the distribution of the Φ index with the value of Φ for the most relevant ones. More precisely, we compared the value of Φ for the 10 “extreme outliers” (characterized by largest absolute residuals) with the distribution of Φ for 10 points randomly selected, on a 10000 sample. The result is shown in figure 5, where finally the exceptionality of the index Φ for the outliers is quite evident.

This has convinced us that the pure statistical rejection criterion should be modified; as a matter of fact the other 169 outliers have a smaller residual absolute value and they are seen as outliers by the testing procedure only because the bicubic model was particularly good for that specific window. This is also clarified by looking (figure 6) at the distribution of $\hat{s}_0$ on the 179 suspected where a relative increase of the frequencies is verified above the threshold of 15 cm and in fact our extreme outliers are all above 20 cm. All that points towards the need of implementing a new s_0^2 estimator based on an average value in the area. In any way, in order to confirm our guess that the 10 “largest” outliers are really related to a physical phenomenon, we have investigated also their connection to the gradient of the undelying topographic surface, giving some graphical representations in figures 7, 8, 9, 10.

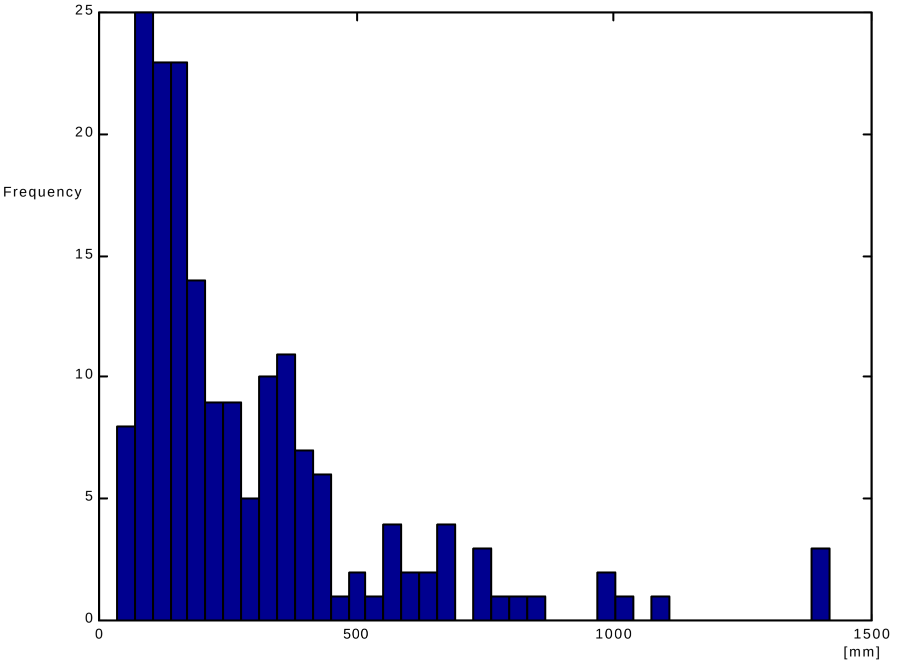

Figure 4 – Histogram of absolute residuals limited to the 179 points classified as outlier

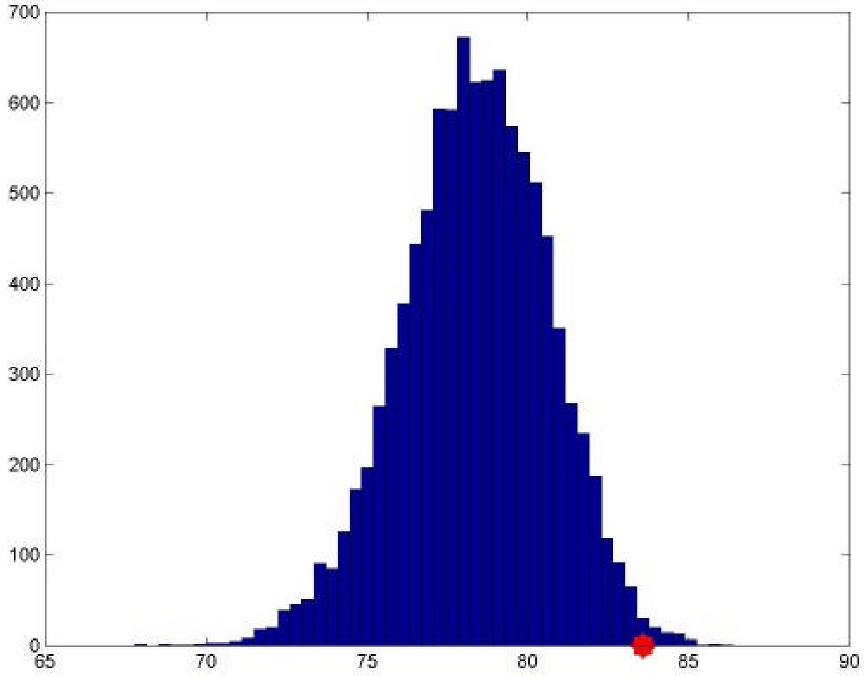

Figure 5 – **F** * for the 10 extreme outliers

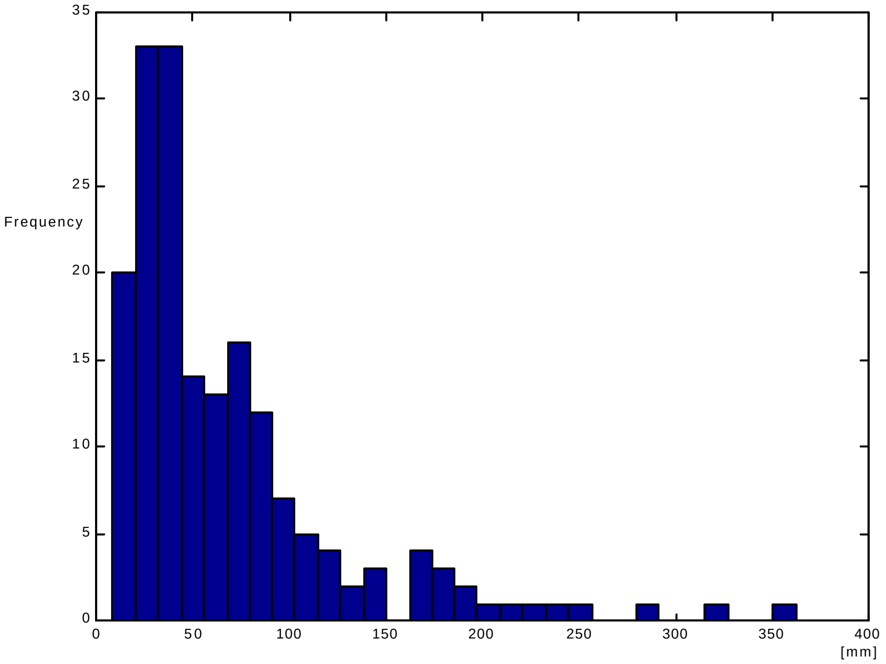

Figure 6 – Histogram of $\hat{s}_0$ limited to the 179 points classified as outlier

Conclusions

Our first conclusion is that the software implemented for the outlier rejection has to be improved, regarding to the s_0^2 estimation, which has to be estimated more specifically to avoid the detection of an unreasonable number of "irrelevant outliers". By the way, the Monte Carlo method, proposed to compare pointwise quantities (like outliers) with area distributed fields, seems to give significant results.

In particular, the relation, between the relevant outliers and large values of geoid gradients, is well illustrated in figure 8: most of the relevant outliers fall in areas characterized by high gradient magnitude and a complicated pattern.

Another conclusion is that these outliers do not seem to be so much related to mountainous areas, but rather to coastal areas with sudden jumps of sea floor. This points towards a good functioning of our terrain correction algorithm in rugged areas, but a poor functioning in coastal areas with depth contrasts, as anybody who has computed a gravimetric geoid knows by experience.

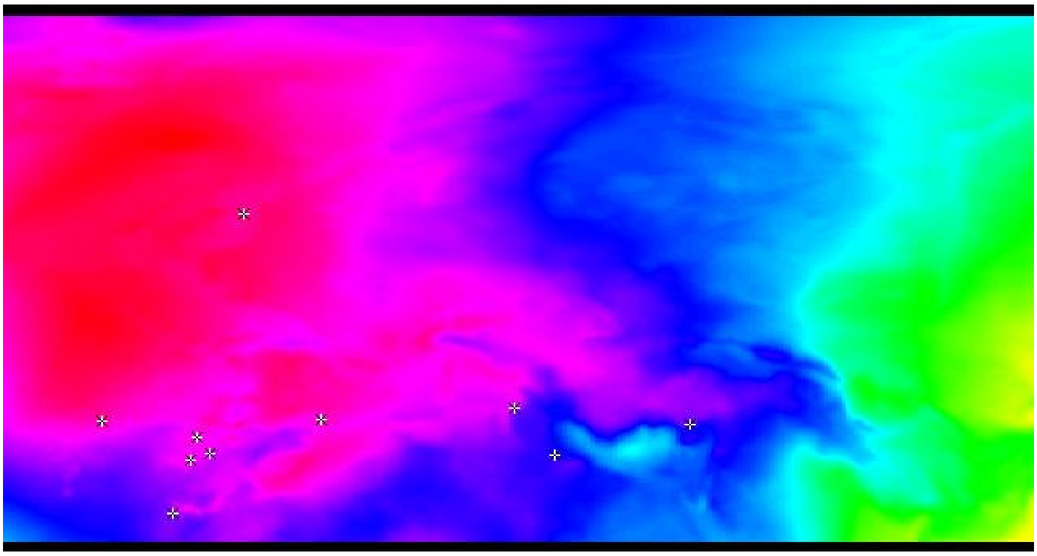

Figure 7 – 10 outliers against the QuasiGeoid signal

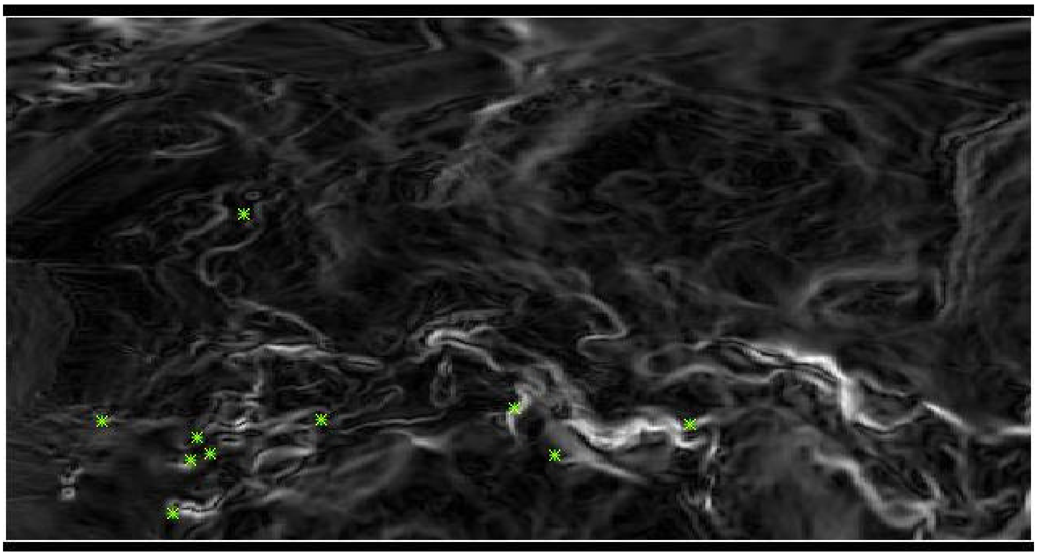

Figure 8 – 10 outliers against the gradient magnitude of the QuasiGeoid signal

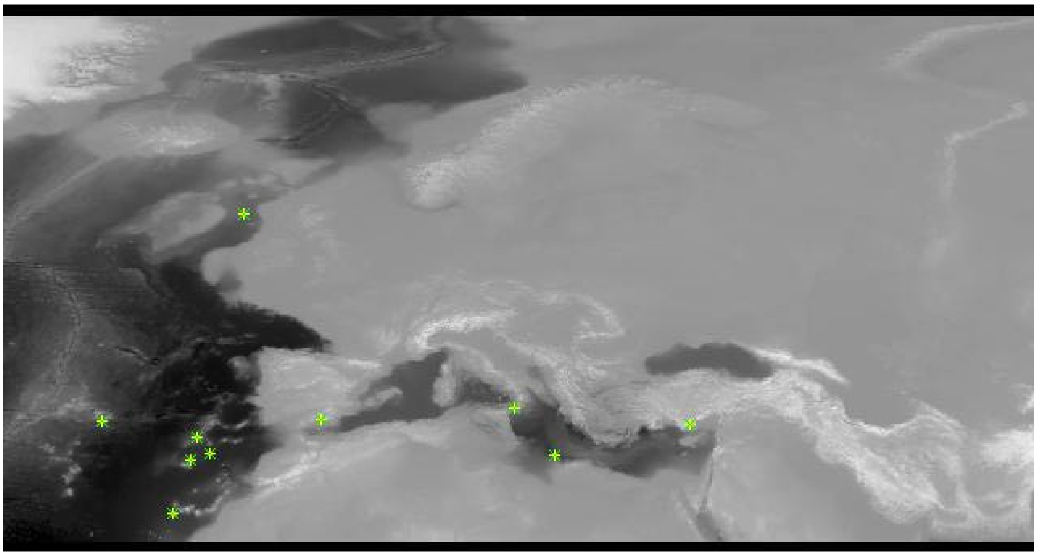

Figure 9 – 10 outliers against the Topographic heights

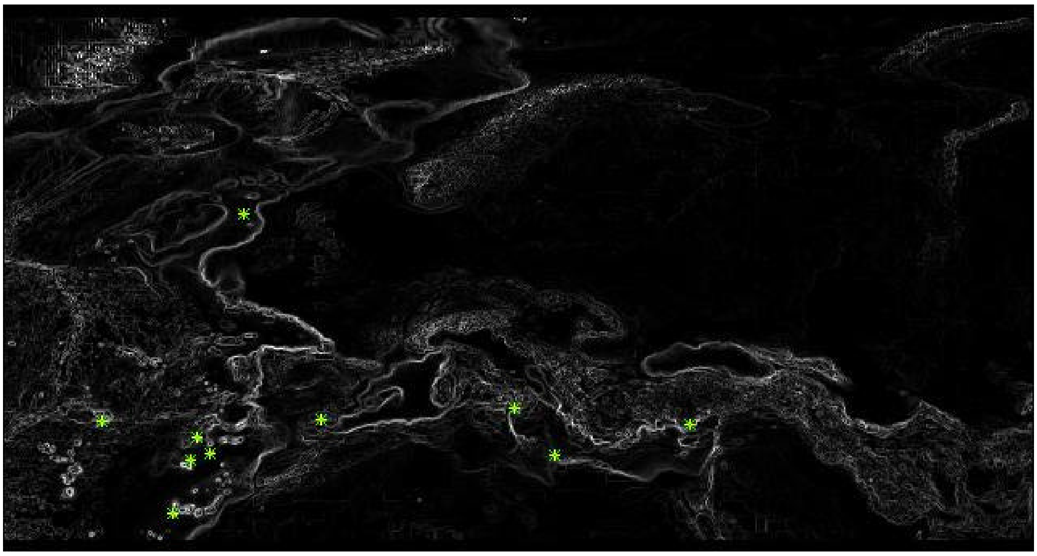

Figure 10 – 10 outliers against the gradient magnitude of the Topographic heights

Bibliography

[1] M. Barbarella, L. Mussio (1985), A strategy for Identification of Outliers in Geodetic Sciences, Statistic and Decisions, Supplement Issue n.2, R. Oldenbourg Verlag, Monaco.

[2] V. Barnet and T. Lewis (1978), Outliers in statistical data, John Wiley and Sons, New York.

[3] T. Bašiæ, R.H. Rapp (1992), Oceanwide prediction of gravity anomalies and sea surface heights using Geos-3, Seasat, and Geosat altimeter data and ETOPO5U bathymetric data. Dept. Geod. Sci. aNd Surv., Rep. 416, Columbus.

[4] B. Benciolini, L. Mussio, F. Sansò (1982), An Approach to Gross Errors Detection more Conservative than Baarda Snooping”, Int. Archives of Photogrammetry and Remote Sensing, Vol. 24, Helsinki.

[5] P.A. Burroghs (1986), Principles of Geographic Information Systems for land resources assesment, Oxford University Monographs on Soil and Resources Survey, n. 12, Clarendon Press, Oxford.

[6] G. Casella (1990), R.L. Berger, Statistical Inference, Wadsworth & Brooks, Pacific Grove.

[7] D. Collett, T. Lewis (1976), The subjective nature of outlier rejection procedures, Applied Statistics, 25, 228-237.

[8] N. Cressie (1991), Statistics for Spatial Data, Wiley, New York.

[9] H. Denker, D.Behrend, W.Torge (1994), European Gravimetric Geoid: Status report 1994, Proceed. IAG Symp. No. 113, Gravity and Geoid, Graz, Austria, Sept. 11-17, 1994, Springer.

[10] H. Denker, D.Behrend and W.Torge (1995), The European Gravimetric Quasigeoid EGG95, International Association of Geodesy, IGeS Bulletin 4

[11] W.J. Dixon (1953), Processing data for outliers, Biometrics.

[12] J.H. Ellenberg (1973), The joint distribution of the standardized least squares residuals from a general linear regression, Journal of the American Statistical Association, 68, No.344, 941-3.

[13] D. Gamerman (1997), Markov Chain Monte Carlo, Chapman & Hall.

[14] D.M. Hawkins (1980), Identification of Outliers, Chapman and Hall.

[15] K. Koch (1987), Parameter Estimation and Hypothesis Testing in Linear Models, Springer-Verlag.

[16] G.S. Lingappaiah (1976), Effects of outliers on the estimation of parameters, Metrika.

[17] R.B. Murphy (1951), On tests for outlying observations, PhD thesis, Princeton University.

[18] A.J. Pope (1975), The Statistics of Residuals and the Detection of Outliers, presented paper at IUGG XVI General Assembly, Grenoble.

[19] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery (1996), Numerical Recipes in C, Cambridge University Press.

[20] C.P. Robert and G.Casella (1999), Monte Carlo Statistical Method, Springer.

[21] L. Sachs (1984), Applied Statistic – A Handbook of Techniques, Second Edition, Springer-Verlag, New York.

[22] F. Sansò (1988), Il trattamento statistico delle misure, CLUP, Milano.

[23] F. Sansò (1991), Il trattamento statistico dei dati, CittàStudi.

[24] F. Sansò (1996), Quaderni di trattamento statistico dei dati, CittàStudiEdizioni.

[25] M.A. Tanner (1996), Tools for Statistical Inference, Springer.

[26] W. Torge, G.Weber, H.G. Wenzel (1982), Computation of a high resolution European gravimetric geoid. Proc. 2nd Int. Symp. On the Geoid in Europe and Mediterranean Area, Rome.

[27] W. Torge and H. Denker (1990), Possible Improvements of the Existing European Geoid, Proceed. IAG Symp. No. 106, Determination of the Geoid - Present and Future, Milan, Italy, June 11-13, 1990, Springer.

[28] H.Wackernagel (1995), Multivariate geostatistics, Springe-Verlag, New York.

[29] V. Vapnik (1982), Estimation of Dependences Based on Empirical Data, Springer-Verlag, New York.