An enhanced method for validating gridded data sets

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

Abstract
The purpose of this work is to propose a reliable method to detect outliers, continuing the work done in a previous article. The previous work was affected by a serious defect related to a rather unstable local estimate of a dispersion index. This paper relies on a robust estimation of a relevant parameter that, before being used for testing the data, is processed with a low-pass filter. The many advantages, obtained in this way, are highlighted in an experiment on a real D.T.M. dataset.

Introduction

The adopted method for outlier rejection in a gridded points data set is the following one: to validate the value $u(t_k)$, a window [whose dimension is $(2s+1)\delta_x$ x $(2s+1)\delta_y$] is opened around the position t_k; the observation points t_j falling in the window [whose number is $N_s = (2s+1)^2 - 1 = 4s\,(s+1)$] are considered in order to estimate the coefficients of an interpolating model. With the interpolating model (which, we underline, is tuned only on the surrounding points), an estimated value of $u(t_k)$ is compared with the observed $u(t_k)$ itself, leading to test the hypothesis (with significance level α)

$$H_0: E[\hat{u}(t_k) - u(t_k)] = 0 \quad .$$

If the local model is a constant function, there is only one coefficient to estimate: $\hat{u}(t_k)$ itself. In the past we proposed fitting polinomials with coefficients estimated by a Minimum Least Square criterion. As it is known (see [18]), we found that these methods have many disadvantages; in particular they are unreliable when there are more outliers in the window and in certain cases they are too much sensitive.
The enhanced method we propose now is robust against outliers, because we use

$$\hat{u}(t_k) = \underset{j \neq k}{Median} \{u(t_j)\} \quad ;$$

therefore t_k is considered an outlier if

$$|Z_{emp}(t_k)| > Z_{\alpha/2} \quad ,$$

where

$$Z_{emp}(t_k) = \frac{u(t_k) - \hat{u}(t_k)}{\sqrt{\left(1 + \frac{\pi}{2N_s}\right)\frac{\pi}{2}} \cdot MAE_k} \quad ,$$

$$MAE_k = \frac{1}{N_s} \sum_{j \neq k} \left| u(t_j) - \hat{u}(t_k) \right| \quad ,$$

while $Z_{\alpha/2}$ is the abscissa value of the standard gaussian distribution, corresponding to a right tail of area $\alpha/2$.[1]

The estimated σ of $[\hat{u}(t_k) - u(t_k)]$ is typical of robust estimation literature (cfr. [8], [6] and [17]). In our conservative approach the normal (gaussian) approximation is used, but – as discussed in [18] – this does not represent a problem.

It should be noted that $Z_{emp}(t_k)$, depends on the size of the moving window. Therefore an algorhytm to find an acceptable dimension for the windows has been developed, too. The size of the window depends essentialy on *s*, since δ_x and δ_y are led by the kind of acquisition process and the type of the cartographic projection of the area to which the points belong.
A good choice of *s* can be made by analyzing the quality of the interpolation model on a global scale. We performed this analysis with the index "global MAE":

$$\mathrm{gMAE}(s) = \frac{1}{N}\sum_k \left|u(t_k) - \hat{u}(t_k)\right| \quad .$$

Its behaviour (that depends on *s* since $\hat{u}(t_k)$ comes from the median of values that are enclosed in the moving window) is this: starting from *s*=1, gMAE decreases until, for a value *s*=*s**, it attains the minimum value gMAE_{min}; for *s*>*s**, gMAE increases indefinitely with *s*. Therefore this optimal value *s** can also be used for the computation of every local MAE_k.
Though we are not able to justify this result, we found experimentally that, when data are regularly gridded, then *s**=1 has always been the optimal value (gMAE is minimized for a window of size 3x3).

Inadequacy of the simple local MAE calculated as above
When the points t_j, around t_k , are well fitted by their median (better than the mean index for not being "outlier-prone" - see also [1], [2] and [3]), then a small difference between $u(t_k)$ and its estimation leads to consider t_k an outlier.
For instance, consider 8 points in a grid with exactly the same height and a central point with 5 cm height difference; of course, since the MAE in this case is exactly zero, this point will be flagged as an outlier. We notice already that this can happen exactly because here and there 8 points can really bear similar values; when we increase the number of surrounding points, this effect could become milder.
In any way the formula for outlier rejection needs a revision, in order to label as outliers only points whose values are sensibly different from their estimations.
Looking back at

$$Z_{emp}(t_k) = \frac{\left|\hat{u}(t_k) - u(t_k)\right|}{\sqrt{\left(1+\frac{\pi}{2N_s}\right)\frac{\pi}{2}}\cdot MAE_k} \quad ,$$

we find that, to decrease $Z_{emp}(t_k)$, under equal conditions, the value of local MAE_k should be increased, avoiding critical situations like that described above. This

[1] A partial demonstration of the above formulae is the asymptotic behaviour of Z_{emp}: when $N\to\infty$, Mean Absolute Error (MAE) is close to Standard Deviation (σ) and Median (MD) is close to Mean (μ). So, having defined $y=u(t_k)-MD$, we discover that $y=N[0,\sigma^2]$, $\mu[|y|]=(2/\pi)^{1/2}\,\sigma$ and therefore $\frac{y}{\sqrt{\frac{\pi}{2}}MAE_k} \approx Z$.

operation must be accomplished with caution, since the rule that is going to be introduced has to be valid under fairly general conditions, not case by case.
We decided that a suitable way to "modulate" the local MAE_k is to smooth it, as if it were a signal. From the general theory of signals, when a signal is smoothed its minimum increases, its maximum decreases, its mean is preserved and the variance decreseases. Moreover, these behaviours are stressed as far as the size of the moving average window increases.

The smoothing strategy
A simple way to smooth our "MAE signal" is using a traditional bi-dimensional moving average on a window of size (2*s*+1)x(2*s*+1). It is known that when a matrix **M** (nr)x(nc) is convolved with a kernel **K** (2*s*+1)x(2*s*+1), the result is a matrix **R** (nr+2*s*)x(nc+2*s*) because **M** is padded with a frame of zeros for s times. But the significant values of the operation are in the submatrix **S** (nr-*s*)x(nc-*s*) of **M**, since only the most internal values are calculated with all the elements covered by **K**. Naturally the choice of the dimension *s* in this case is very important.
In the next paragraph we show some experiments with real data from a D.T.M. that have given us some hints on this choice.
On the other hand at least two two things are clear: altough an optimal window to estimate the central value is often of dimension 3x3, yet the corresponding MAE (computed on that window only) has a by far too large probability of having a too small value, thus generating false outliers. On the other side, if we used the average of the local MAEs over the whole data area that would be too large to detect real outliers. So we decided to base our choice on experiments.

The experiment
In order to test the quality of the method of outlier rejection and to calibrate the one on which local MAEs have to be averaged, we chose an area (in the northern part of Italy) where we have a Digital Terrain Model characterized by a flat zone and a mountainous one. There is also a lake (Lake of Como) whose form is a "Y" upside down. The description of the region is the following (in degrees with their decimals):

```
north:      46.1326129
south:      45.491
east:       9.2928936
west:       9.043
e-w resol:  0.00277659556
n-s resol:  0.00194428152
```

therefore the matrix M is composed by 330 rows and 90 columns, as shown in figure 1.

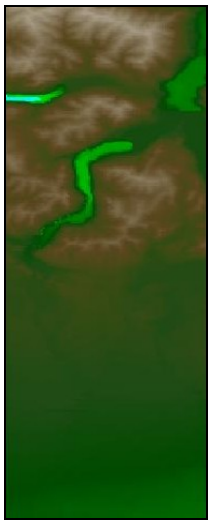

Figure 1 – The area of study

An outlier rejection performed with our method on the above dataset (with significance level of α=0.1%) and using local MAE indexes without any smoothing, leads to identifying 40 outliers.
It should be noted that, in order to avoid border effects, a frame of width 9 all around the boundary has been dropped in every case because the largest kernel is 19x19 (so s=9) and we wanted to have results capable of overlapping.
So instead of validating 330x90=29700 points, we effectively validated 312x72=22464 points.

By smoothing the local MAE with a simple kernel 3x3, the number of outliers found is 1; this shows that smoothing in a so restricted area provoked an increasing of the local MAE (in every point) and therefore to consider good points the ones previously labelled as outliers. Using wider and wider kernels, the number of found outliers goes around 50 (as can be seen in the table 1). This can lead to think that this is the reasonable number of outliers.

MAE smoothing's Window	# Outliers	Min MAE	Max MAE	Mean MAE	Var MAE
None	40	0	210,88	32,69	1306,8
3x3	1	0	162,06	32,69	1167,7
7x7	30	0	117,12	32,69	1023,4
11x11	46	0,535	101,665	32,68	945,9
15x15	48	0,551	94,96	32,64	900,1

Table 1 – Interesting indicators with different smoothing windows

Another way of perceive the process is by plotting an histogram of the entire MAE dataset in the different cases.
In figure 2 it can be seen that – smoothing with a wider and wider mask – the number of very small MAEs decreases while the number of MAE around the mean increases.

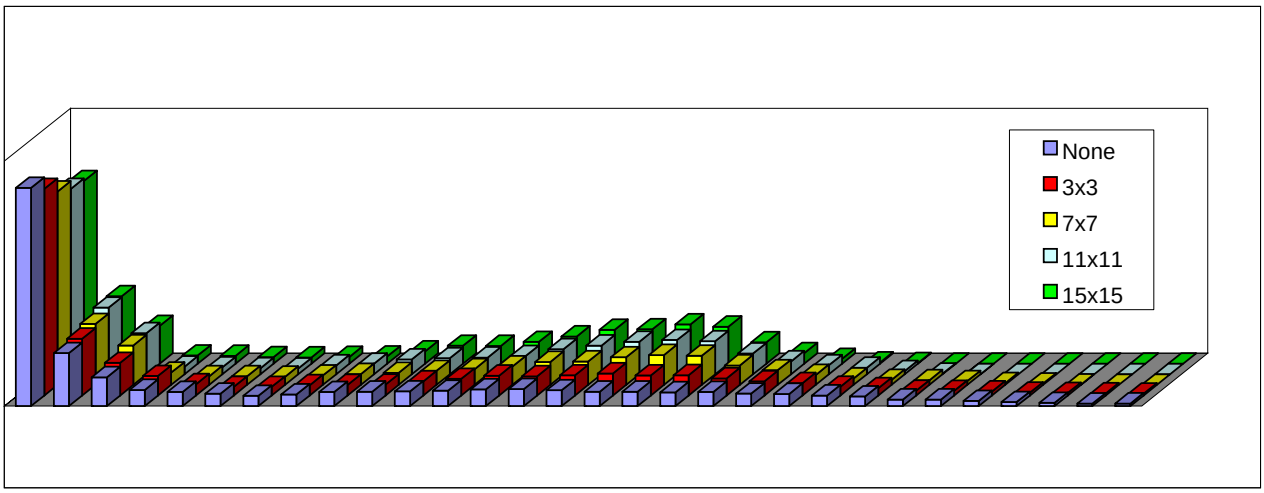

Figure 2 – Histogram of MAE dataset in different cases

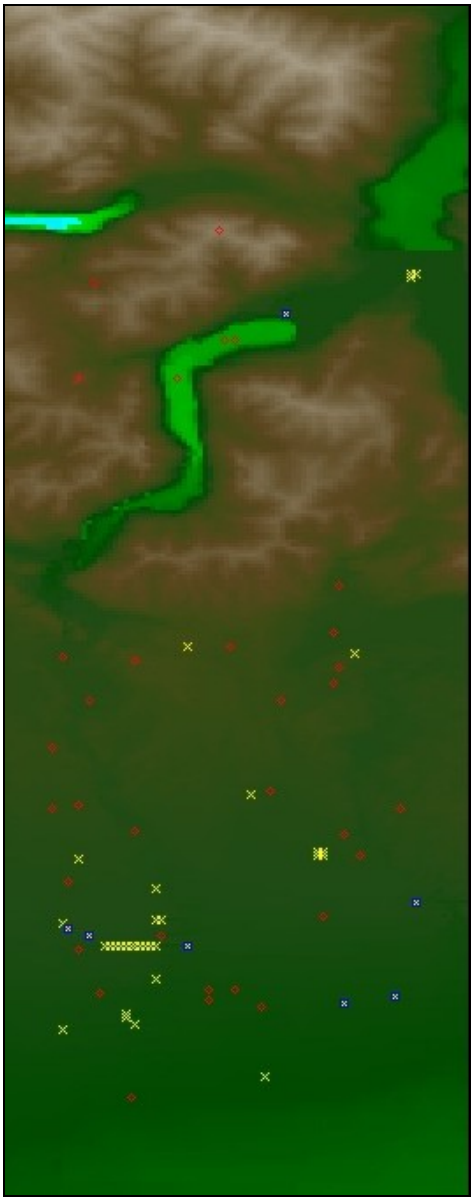

Figure 3 – Relevant points according to two approaches

As it can be seen from Figure 3 (where outliers according to original MAE are plotted with red asterisks, outliers according to mask 9x9 with smoothed MAE are plotted with yellow, and the 7 outliers in common with blue boxes), by smoothing the MAE, we are able to find "true" outliers (for example very interesting the row of 11 outliers in the south west, which gives the impression of some systematic error) and ignore many "false" ones, for example those that are in the flat zone.

The smoothing process of the local MAEs produces a positive effect that comes from the combination of two effects: the increasing of the minimum MAE (that leads to consider good points the ones that with calculated MAE were considered outlier, because of their small deviation from the estimation) and the decreasing of the maximum MAE (that leads to consider outlier points some that previously were considered good ones). The resulting effect is that a reasonable number of outliers is detected by a reliable process. This insight is confirmed by figure 4 (and figure 5, that is a zoom of figure 4), where – in a diagram local MAE vs absolute differences – with red circles (o) are shown outliers according to original MAE, while with blue crosses (x) are shown outliers according to smoothed MAE with mask 9x9.
Indeed, circles with very low MAEs and very low absolute differences[2] are ignored by smoothing the MAE; also circles with very high local MAE and very high absolute differences are not considered outlier, according to the enhanced process. The 7 outliers in common are perceivable because each of the 7 circles has its own cross on the same value of absolute difference.
It can also be noted that, in a gridded dataset, it is always $N_s=8$. So, having fixed $\alpha=0.1\%$, is easy to calculate the gradient of the dotted line that defines the limit over which we find outliers: since the theoretic value of Z is 2.5758,

$$\frac{|\hat{u}(t_k)-u(t_k)|}{\sqrt{\left(1+\frac{\pi}{16}\right)\frac{\pi}{2}\cdot MAE_k}} > Z_{\alpha/2} \Leftrightarrow |\hat{u}(t_k)-u(t_k)| > 3.53118\, MAE_k \quad .$$

In conclusion, the simple idea of smoothing the MAE is a powerful tool to detect outliers as described above. The smoothing window can be fixed to dimension 9x9 altough close by values would not dramatically affect the result. There is still work left to conceive a more general method suitable for not gridded datasets.

[2] With the term "absolute difference" we mean the absolute value resulting from the difference between the value of the observed point and its estimation (the median of the surrounding points).

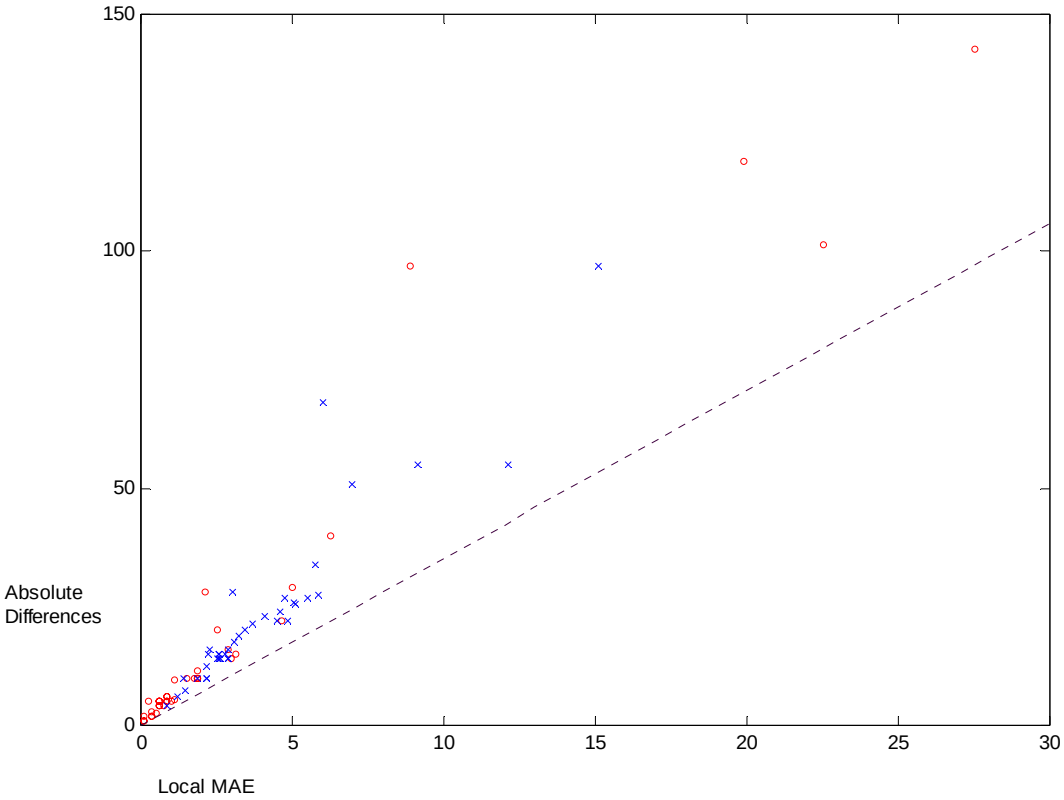

Figure 4 – Relevant points according to two approaches

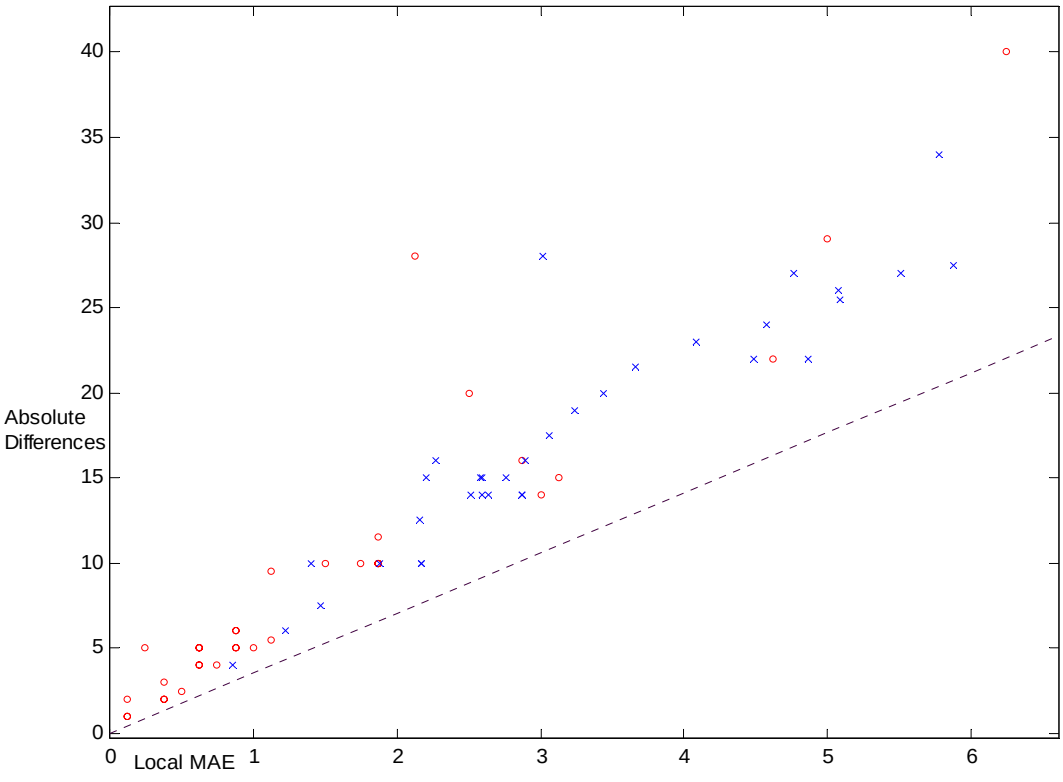

Figure 5 – Particular of figure 4

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] 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.

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

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

[6] H.A. David and H.N. Nagaraja (2003), Order Statistics, Third Edition, Wiley.

[7] A. Dermanis, A. Grün and F. Sansò (2000), Geomatic Methods for the Analysis of Data in the Earth Sciences, Springer – Verlag.

[8] E.R. Dougherty and J. Astola (1994), An Introduction to Nonlinear Image Processing, SPIE Optical Engineering Press.

[9] R.O. Duda and P.E. Hart (1973), Pattern classification and scene analysis, John Wiley & Sons.

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

[11] A. Jain (1989), Foundamentals of digital image processing, Prentice Hall.

[12] L. Kaufman and P.J. Rousseeuw (1990), Finding Groups in Data. An Introduction to Cluster Analysis, John Wiley & Sons.

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

[14] J.S. Lim (1990), Two-dimensional Signal and Image Processing, PTR Prentice Hall.

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

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

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

[18] D. Triglione, Reasoning about the presence of outliers in the European quasi-geoid data set, Same Number, In print.