CORRELATIONS IN LARGE SETS OF DATA

BRUNO CRIPPA
Dip. IIAR Politecnico di Milano
P.za Leonardo da Vinci, 32 - 20133 Milano - Italy
E-mail: bruno@ipmtf4.topo.polimi.it

ISPRS Commission VI, Working Group 3

KEYWORDS: Correlation, Data Processing, Least Squares.

ABSTRACT. A common assumption in processing large, or sometime small, sets of data is that correlations inside the data are either. This hypothesis on the observations isn't always true and could lead to wrong results (often the results seem better!). The correlation inside the observations is highlight through the computation of the autocovariance function

1. INTRODUCTION

Processing large, or sometime small, sets of data (observations) pose an interesting problem about the presence of correlations inside the data. Indeed when we apply a statistical procedure we suppose that observations are incorrelated. This hypothesis on the observations isn't always true and could lead to wrong results (often the results seem better!).
The correlation inside the observations, or on residuals after the removal of trend with a deterministic model, is highlight with the computation of the autocovariance function.
The autocovariance function takes into account all effects that aren't considered in the deterministic model and therefore they are treated as stochastic process. The autocovariance function give us the degree of correlation between the value (i.e. grey level of an image) in a point and the value to another point, related to the distance between the two points. In general nearly points have higher correlation with respect to points far apart.
Before applying the algorithm to compute the autocovariance function it is needed to remove any trend, if it is present, inside the data, since they are assumed to come from a stationary and isotropic (for 2D data) stochastic process.

2. EXAMPLES (Crippa, Mussio, 1987)

We have considered some data sets chosen in different fields:

- geodesy (GPS)
- satellite image (SAR, SPOT)
- digital aerial image (scanned aerial photo image)
- close range image (scanned or taken with CCD camera)
- numerical cartography (digitized map).

We have taked the following examples:

- A session of GPS measurements (in this case we have applied a ionosphere correction)
- A portion of SAR filtered interferogram (400x100 pixels)
- A portion of a SPOT image (200x200 pixels)
- A portion of an aerial image (301x301 pixels)
- A portion of close range image (200x200 pixels)
- A digitized set of data in Lissone area (Italy) (we have chosen the points referred to buildings and urban streets).

The GPS data come from measurements done in Noto (Italy), with Rogue equipment the data are stored in a file with RINEX format, the time interval between a two measures was 15 seconds.
Inside the file of the measurements there was cycle slips than it is necessary to preprocess the file of measurements in the following way:

- extract from the RINEX file the measures: time, codes P, C/A and carrier phases L1 and L2 , this must be done for each satellite
- look for in the time serie the present of cycle slip if it is present cut the time serie
- apply to each time serie, without cycle slip, the ionosphere correction.

The GPS session had 25 satellites, we present three different results of autocovariance functions:

- satellite 21 the autocovariance function doesn't shows correlation (fig. 1,2)
- satellite 6 the autocovariance function shows a weak correlation (fig. 3,4)
- satellite 12[1] the autocovariance function shows a strong correlation (fig. 5,6)

Note that the satellite with strong correlation has a low variance (one order of magnitude) with respect to the other satellites (Tab. 1).

	Sat 21	Sat 6	Sat 12
No. obs	617	343	673
average D1	6.566	0.972	12.124
sqm D1	0.295	0.774	0.075
average D2	9.831	1.438	19.923
sqm D2	0.885	1.301	0.054

Tab 1 - Average and sqm of the measurements.

The autocovariance function has been computed on the data after an ionospheric correction.

[1] This satellite had some problem in the clocks equipment: this can be the reason of the high correlation.

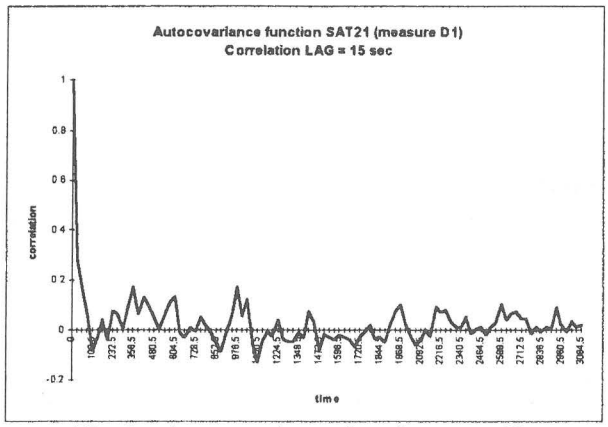

Fig. 1 - Autocovariance function on sat21.

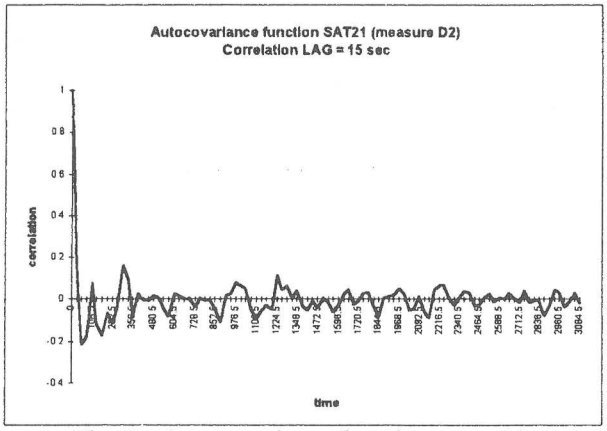

Fig. 2 - Autocovariance function on sat21.

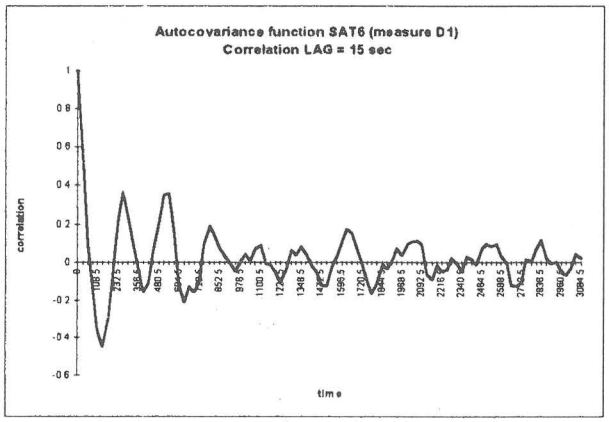

Fig. 4 - Autocovariance function on sat6.

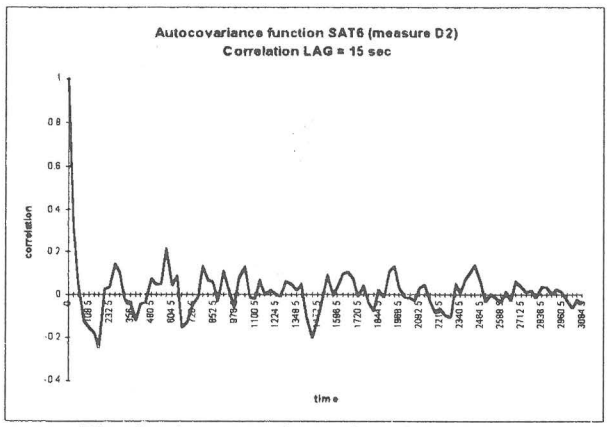

Fig. 5 - Autocovariance function on sat6.

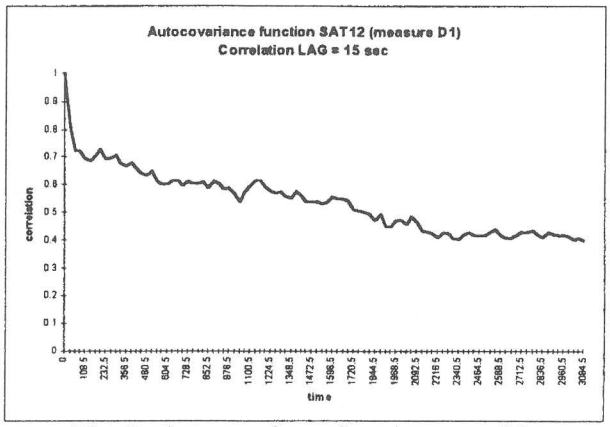

Fig. 5 - Autocovariance function on sat12.

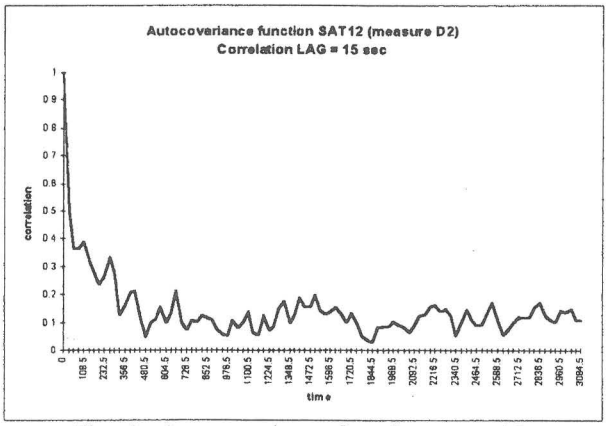

Fig. 6 - Autocovariance function on sat12.

The SAR filtered interferogram (fig. 7) has been generated with ISAR packages. We have computed the autocovariance function on phases (difference of the phases of the two SAR images) and the process parameter, which have been used, is the distance (on the interferogram image) between the phases (realization of 2-dimensional stochastic process). As we can see (fig. 8) the correlations between the phases are negligible for pixels which are about ten pixels apart, therefore the phase value in a point depends (statistically) of the phase values in a neighbour of ten pixel size.

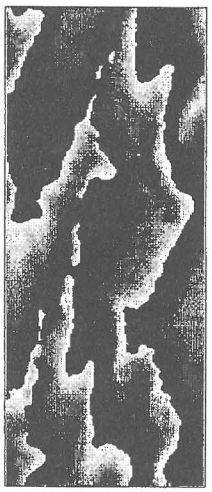

Fig. 7- SAR filtered interferogram

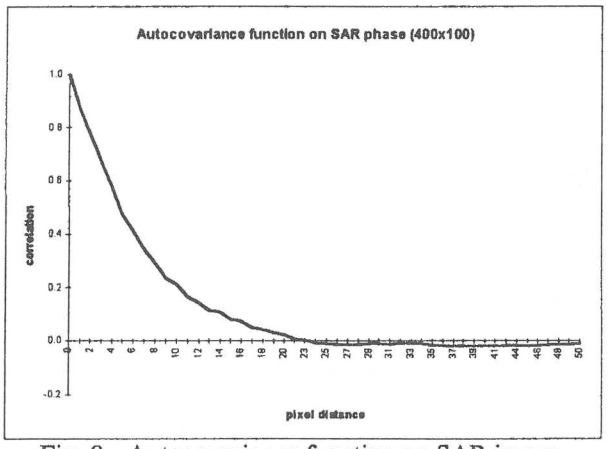

Fig. 8 - Autocovariance function on SAR image.

A portion (200x200 pixels) of a SPOT image (fig. 9) has been processed, in this case we have computed the autocovariance function (fig. 10) on grey levels and the process parameter, which have been used is the distance between the pixels (realization of 2-dimensional stochastic process).

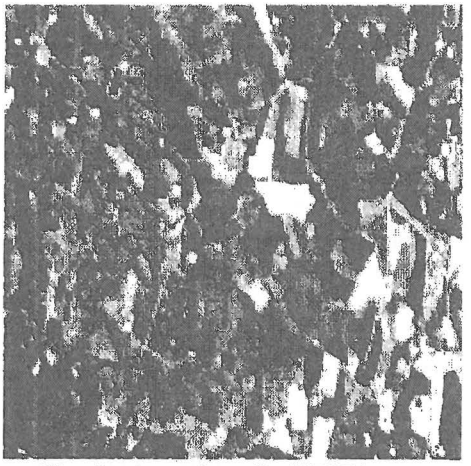

Fig. 9 - A portion of a SPOT image.

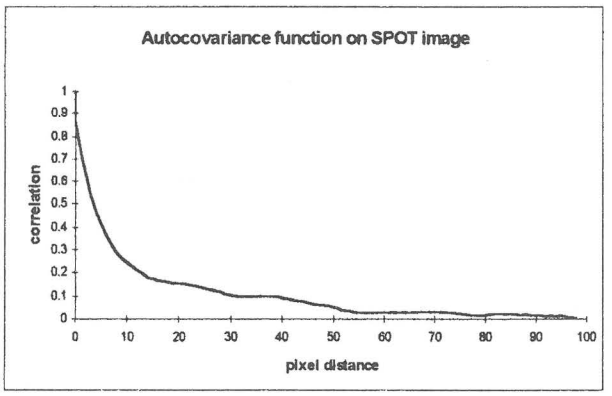

Fig. 10 - Autocovariance function on SPOT image

A portion (301x301 pixels) of an aerial scanned image (fig. 11) has been processed, in this case we have computed the autocovariance function (fig. 12) on grey levels and the stochastic parameter, which have been used is the distance between the pixels (realization of 2-dimensional stochastic process).

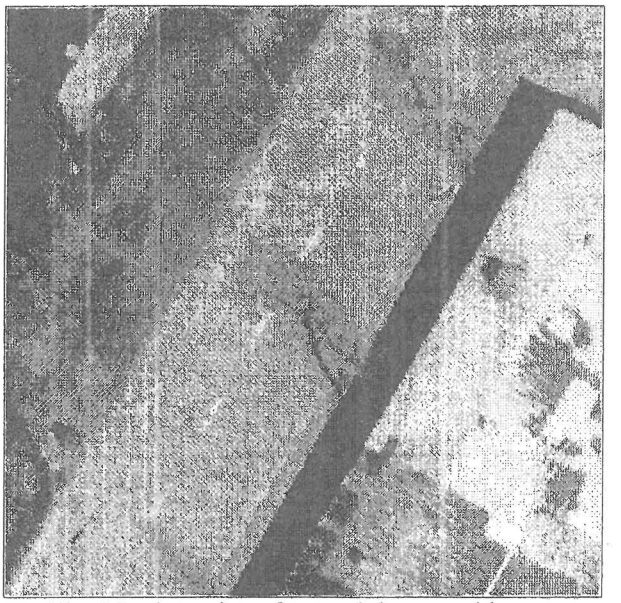

Fig. 11 - A portion of an aerial scanned image.

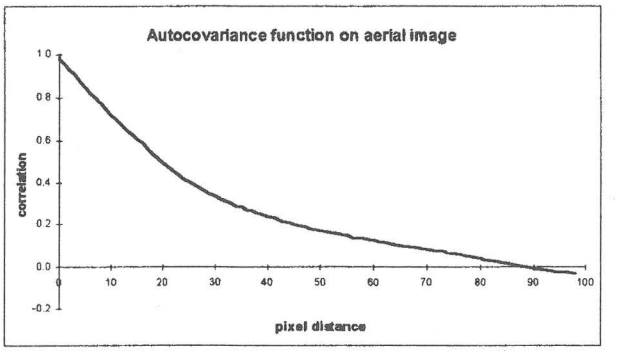

Fig. 12 - Autocovariance function on SAR image.

A portion (200x200 pixels) of a close range image (fig. 13) has been processed, in this case we have computed the autocovariance function on grey levels and the stochastic parameter, which have been used is the distance between the pixels (realization of 2-dimensional stochastic process.

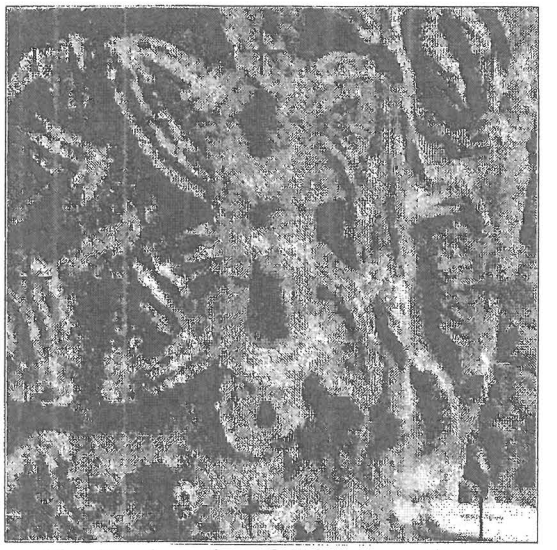

Fig. 13 - A portion of a close range image.

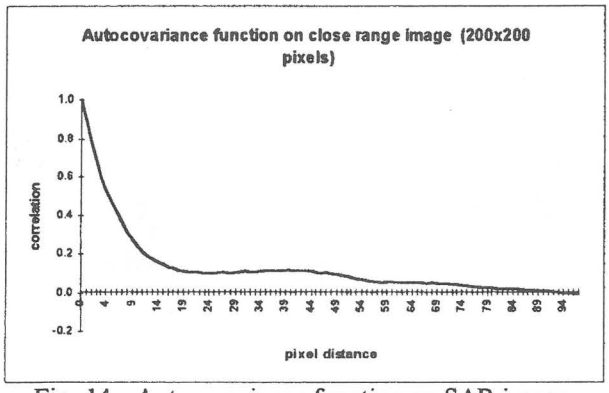

Fig. 14 - Autocovariance function on SAR image.

When the autocovariance function is computed on data distributed in a regular grid, (as it is the case with SAR, SPOT, aerial image and close range image) it is possible to use a fast algorithm which takes into account the high regularity of the data.
We have extracted from a cartographic database two sets of digitized points, respectively referred to urban streets and buildings. For each set we have counted the points which fall in each mesh (class) of the grid superimposed on the area. Furthermore, we have surrounded the boundary of the classes in the area with a polygon. In this case we have computed the autocovariance function (fig. 15, 16) on the numbers of points which belong to a classe (frequency) and the process parameter, which have been used is the distance between the centres of the classes.

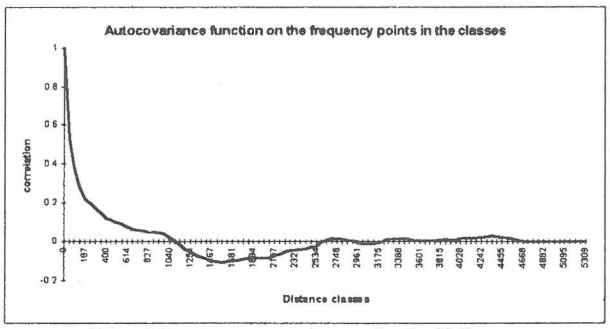

Fig. 15 - Autocovariance for buildings.

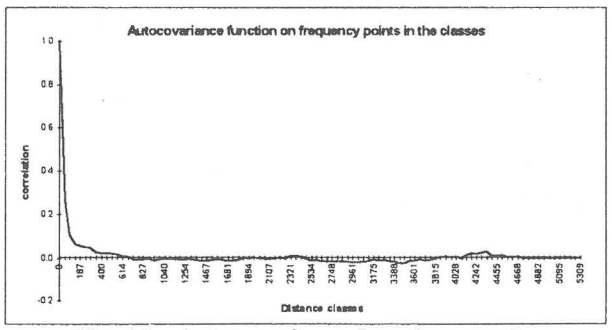

Fig. 16 - Autocovariance for urban streets.

3. CONCLUSION

The aim of this work is to show that the hypothesis of independence of the measurements, in a set of data, it is not always true and this must be taken into account when the measures are treated with a statistical procedures.

For example when applying the least squares procedure to a set of measures (residual after a removing trend) it is need verify the hypothesis of indipendence for use a diagonal weight matrix.
The presence of correlation inside the data can be a first alarm to verify the correctness of the functional model in the least squares procedures.
A special remark concerns the use of the autocovariance function regarding to the map generalization. Indeed when two lines are distinct already at a distance lower than a high autocorrelation value (e.g. $\rho = 0.5$), the generalization could be done substantially without loss of information. On the contrary, when two lines are distinct only at a distance greather than a high autocorrelation value (e.g. $\rho = 0.5$), the generalization should be done suitably selecting the most important features.

APPENDIX IONOSPHERE CORRECTION ON GPS MEASURES[2]

As just explained above in the GPS measures must be removed the ionosphere effects. In the follow we present the expression to achieve this goal:
1) Ambiguity of L1:

$$N1 = C1 - \lambda_1 * L1$$

(at first measure; used P1 when available)
Ambiguity of L2:

$$N2 = P2 - \lambda_2 * L2$$

(at first measure)

where λ_1 and λ_2 are the wave lengths of L1 and L2 GPS phase measures

2)
$$\Phi_1 = \lambda_1 * L1 + N1$$
$$\Phi_2 = \lambda_2 * L2 + N2$$

(Φ represent the distance station-satellite computed with phase measures)

3)
$$I_1 = (\Phi_1 - \Phi_2) * f_2^2 / (f_1^2 - f_2^2)$$
$$I_2 = (\Phi_1 - \Phi_2) * f_1^2 / (f_1^2 - f_2^2)$$

where $f_1 = 1575.42$ MHz and $f_2 = 1227.60$ MHz are the frequencies of L1 and L2 GPS phase measures

4)
$$C1_{correct} = C1 - I_1 \qquad P2_{correct} = P2 - I_2$$
$$\Phi_{1correct} = \Phi_1 + I_1 \qquad \Phi_{2correct} = \Phi_2 + I_2$$

5)
$$D1_{correct} = C1_{correct} - \Phi_{1correct}$$
$$D2_{correct} = P2_{correct} - \Phi_{2correct}$$

The empirical autocovariance functions are computed over D1 and D2.

REFECENCES

Crippa, B., Mussio, L., 1987. The new ITM System of the programs MODEL for digital modelling. Proc. of the Int. Colloquium on Progress in Terrain Modelling, Copenhagen 20-22 maggio 1987, O. Jacoby and P. Frederiksen (Eds.), Technical University of Denmark, pp. 75-87.

[2] This formulas have been given by dr. ing. Mattia Crespi of D.I.T.S. - Univ. di Roma 'La Sapienza'.