Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

When the number of measurements is insufficient

by Bernard Beauzamy
PDG, SCM SA

January 23rd 2008

English summary

This note is a methodological complement to several contracts we treated during these last years, in particular :

- Veolia Environnement, Région Ouest, 2005 - 2007 ;
- Direction Générale de l'Energie et des Matières Premières, 2006 - 2007 ;
- European Environment Agency, 2007.

In each case, a decision must be taken on the grounds of incomplete information: there are not enough measurements to obtain a sufficient knowledge of the phenomena. What mistakes should be avoided?

The question is not just, let us insist upon that, to reconstruct missing data from existing data, as we did about rivers (our book « reconstruction de données manquantes, by Bernard Beauzamy and Olga Zeydina). The question is about a situation where nothing is known: a whole zone, or a whole period, is totally without data.

Our conclusions are as follows :

- One should never try to complete the knowledge using a purely statistical basis, using only the data themselves. For instance, one should not use linear regressions, interpolations, extrapolations, and so on, because such reconstructions are totally fictitious. One must use, at least in a simplified manner, the physics behind the problem. But conversely one should not use very precise physical models, which are usually irrelevant.

For instance, assume that you want to reconstruct the flow of a river during one year ; you have only the data for another year, and no other river. You should not use a linear regression (because the flow is not linear). You should use the rain (data of rainfalls are usually available), but you should not try to build a very detailed physical model relating the flow and the rain, because such models require a lot of information (type of soils, and so on). You should try to identify the points where the rainfalls are best correlated with the flow of the river, using the year when the data are known, and then use these rainfalls in order to predict the flows for the missing year.

We see here a very simple and striking example : some physical information is used (the fact that the flow is linked with the rain), but in a very coarse and robust manner.

So, any method which does not use, at least in a simple manner, some physics from the problem, must be avoided.

- A net of sensors is often built in order to answer a social need : to know the necessary quantity of water in some city, the amount of pollution in a river, the need in energy, and so on. This social usefulness is local, limited in time and space, and often coarse, in the sense that it does not require measurements of high precision. One should not pretend that this social knowledge is sufficient for a scientific knowledge of the whole phenomenon.

For instance, the knowledge of temperature on Earth is dictated by preoccupations linked with production of energy, resistance of buildings, and information linked with travels. The existing net of sensors is more or less sufficient for this respect. We should not deduce that we know the temperature at each point of the Earth at each instant.

To define a net of sensors for scientific use is an entirely different topic. The questions: how many sensors? where should they be? which frequency of measurements ? must be addressed within the scope of scientific knowledge. The cost will of course be very high and the social benefit will be small.

In all situations we met, the social net of sensors hardly suffices in order to meet the social need for which it was created : the sensors do not work properly, or are not correctly maintained. In these conditions, to use this net of sensors for a precise and global scientific knowledge is totally fictitious, and generally dishonest.