WORKING NOTE AC62

Alan Creak
22 April 1987

SOME NOTES ON ANALYSING THE BOAT DATA, PRESENTED WITH THE INTENTION OF STIMULATING DISCUSSION.

WHAT WE'RE DOING : we are looking for regularities in a given collection of data, consisting of measurements made on certain data streams at known times.

APPROACHES : we are not the first people to try to impose order onto a large collection of data, and we should at least know about some other methods which have been used. There are statistical methods (see Multiple Classification Analysis), clustering (see ?), lots of work on pattern recognition, etc. ? We should know enough about these to be able to justify our approach.

EVENTS : a definition ? An event is something noteworthy connected with a data stream. More precisely, an event is a sudden change in a data value or one of its derivatives. Its structure is something like { Datastream, Whichderivative, Magnitudeofchange }.

OBJECTS : a useful idea, perhaps. An object is a particular instance of an event – it is a pair { Event, Time }.

PATTERNS : a pattern is a repeated instance of { Event1, Event2, Timeinterval }. The two events can be in the same or different data streams. The pattern exists even if not all Event1 objects are associated with corresponding Event2 objects – there is presumably some notion of the reliability of a pattern which quantifies this relationship.

That suggests two components to what we're doing :

- Identifying events. An event is a temporal pattern in a single data stream. To detect events, we need to be able to look at the raw data of a data stream over a period of time, get rid of the noise to a reasonable degree (which means we have to know what's noise and what's data : perhaps noise is patterns too. Not necessarily easy – the cutting on the right reports on solar phenomena discovered by reclassifying some "instrumental errors" as data), differentiate the signal, and look for recurring features in the result.

- Looking for patterns. Pick any two events from the many data streams, and and see whether the same pair of events separated by the same time interval (± a bit – how much ?) turns up again fairly often. That's something we may reasonably leave to ID3, or a variant thereof.

The sun has lent its name to a new field of astronomical research – helioseismology.

During the past two decades it has become clear that what were at firsts thought to be instrumental errors are real and that the sun oscillates over a wide range of frequencies.

Just as earthquake shock waves allow us to map the interior of the Earth, so these solar shock waves will tell us something about the solar interior.

New Zealand Herald, 21 April 1987

This approach MISSES things like :

- "quality" of data – an equally significant event might be a change from steady to varying values in (say) wind direction. Does that mean we also want a data stream called "wind steadiness", or something ? In two-dimensional work on pattern recognition in pictures, this is called *texture*, and is found to be useful.

- threshold values : for example, turbulence could be important for wind speeds exceeding X.

- dependent variables : the important variable may not be one which is measured directly, but some combination of the measured values – such as the component of wind velocity parallel to the boat's direction. That's the sort of thing that Bacon does, though one usually thinks of Bacon as working in a less noisy context.