Workshop 3, SECURE Launch Event, 2015; led by Dr Mark Brewer

Shaping the data sets of the future

1. There is a scientific challenge in the types of data which are collected, especially in the light of current funding constraints. There appears to be a shift (for example, with SEPA) where monitoring data will be collected in a more focussed manner at the individual level, often by industrial organisations themselves. This creates regulatory issues, and modelling is required to provide feedback for improving monitoring.

2. It was helpful to think about three different classes of environmental monitoring data: (a) surveillance data, often for long-term monitoring, used for studying long-term trends; (b) operational data, collected by a regular monitoring scheme, often under some regulatory framework and the aim of enabling decisions to be made on the basis of evidence; and (c) short-term investigative/episodic monitoring, often used to determine why or how a particular event occurred. Different types of data have different kinds of uses, and it is important for statisticians and modellers to be aware of these differences, and that methodology is needed for all three types.

3. Leading on from 1 and 2, with a shift to more focussed data monitoring, it is vital that statisticians are made aware of how data were collected and for what purpose - i.e. what the questions were behind the decision to collect the data in a particular way. It was important to consider "repurposability", both when designing monitoring schemes (if possible) and when embarking on an analysis of the data.

4. While statistical design is important with respect to monitoring schemes, there are also statistical issues with regard to the "mechanics" of data collection. One example relates to sensors which (in some cases) can be tuned in different ways - for example, to cover different parts of the spectrum of visible light. At present, this tuning is done in an ad hoc way, but robust statistical methods for doing this "automatically" would be a boon.

5. Whatever data is collected, there are always issues of error and bias. These should be understood, and be part of any communication involving or referring to the data set. Improved methods for bias-correction and calibration are needed urgently, especially for climate data, for example.

6. The issue of "data knowledge". This would be an excellent topic for a workshop (or part of a workshop). Many environmental data sets require a considerable amount of processing, checking, validation and interpretation. Often, individual researchers will spend from hours to months creating usable, useful data sets from the raw sensor (for example) data sets. There is a suspicion that all too often, that knowledge gained by the person/people working on the

data is lost once the usable data have been produced, and that another group wanting to use the data for a similar purpose will themselves need to start from scratch (whether the same data set or a different, but related, data set). This is clearly horribly inefficient, and with current moves to make more large environmental data sets open access, the situation may arise more often. What is needed, therefore, is a way of ensuring that all this data processing effort is somehow made available to the scientific community; with a move to attach DOIs to data sets, there would seem to be a route for a framework in the offing. Some kind of platform is required for storage of vital information, such as: the raw data set itself; a complete, step-by-step description of the data processing stages, each with justification - and ideally with code for doing the processing, which might then be usable on updated data sets or other data sets with the same structure and issues; the processed, usable data set; and any reports or published papers making use of the usable data set.