Foreword

One of the most successful application areas of data mining is in surveillance - that is, in monitoring ongoing situations to detect sudden changes or unexpected events. Such methods have extremely widespread application, from detecting epidemic disease outbreaks as quickly as possible, through fraud detection in banking, to detecting sudden changes in the condition of intensive care patients, as well as to detecting the imminent departure of manufacturing processes from acceptable operating limits, warning of potential terrorist atrocities, and the automatic analysis of video footage to detect suspicious behaviour. The aim in all such problems is to process data dynamically, as quickly as possible, to act as an early warning system so that an alert to the imminent change can be given and appropriate action can be taken.

Characteristic of such problems is that the data are *streaming data*: they keep on coming, are often multivariate, and require on-line processing. This is very different from the 'classical' statistical problem of batch mode data, which can be analysed and re-analysed in one's laboratory at leisure. This means that adaptive, sequential, *learning* algorithms are needed, and that often one will get only one chance to look at the data. The analysis has to be done immediately, and then the data are gone, and the system has to look for the next potential event.

Furthermore, surveillance problems are often characterised by *large data sets* - which makes them a very modern problem. At the extreme, the word petabyte (10^{15}) is not unusual: the Large Hadron Collider produces about 15 petabytes of data per year, which needs to be monitored for unusual data configurations, and AT&T, which uses surveillance methods to detect theft of telecomms resources, transfers some 16 petabytes per day.

A third characteristic is that the aim is to provide an *early warning*, so that a timely intervention can be made. A surveillance system to detect credit card fraud which raised an alert some three months after the transaction had occurred would be useless - even if it successfully detected all frauds and never raised a flag on legitimate transactions.

These three features of the data - its dynamic and ongoing nature, the sizes of the data sets, and the aim of providing an early warning - pose particular theoretical and practical challenges. This makes it a rich, as well as an increasingly important area, for research.

There are various approaches to surveillance. In some situations, the type of anomaly being sought is known. In such cases one can use *supervised* methods, in which one builds a system which is effective at distinguishing between data structures with the known characteristics of the anomaly and other data structures. So, for example, certain kinds of behaviour are known to be indicative of possible credit card fraud, and a system can be trained to look for such behaviour.

In contrast, in other situations, the system may hope to detect configurations which depart from the norm in various, but not completely specified ways. Outlier detection is an example, where all we

know is that the observation is extreme, without being able to say in what way (on what variables) it is extreme. In such cases, *unsupervised* methods are necessary, which simply compare data points, or data configurations with the norm, to detect unusual patterns. An example would be signs of imminent disease outbreaks arising from unexpected local clusters of cases.

A wide variety of statistical tools are applied in surveillance problems, including change point analysis, forecasting methods, scan statistics, and filtering. But in some sense, the area is a relatively new one: modern data capture technology has opened up a wealth of possibilities for analysing systems as they operate, to detect unusual or dangerous events. The area is one of increasing importance, over an increasing number of domains. As a consequence, a wide range of readers will find this book of interest. The book has captured this breadth by presenting studies from a number of different areas, illustrating the range of applications and the diversity of methods which are used. It is a welcome addition to the literature.

David J. Hand
Imperial College, London

David J. Hand *is Professor of Statistics at Imperial College, London. He studied mathematics at the University of Oxford and statistics and pattern recognition at the University of Southampton. His most recent books are Statistics: a Very Short Introduction, and ROC Curves for Continuous Data. He launched the journal Statistics and Computing, and served a term of office as editor of Journal of the Royal Statistical Society, Series C. He is currently President of the Royal Statistical Society. He has received various awards and prizes for his research, including the Guy medal of the Royal Statistical Society, a Research Merit Award from the Royal Society, and the IEEE-ICDM Outstanding Contributions Award. He was elected a Fellow of the British Academy in 2003.*