

## Report Cover Page

<b>ACERA Project</b>		
0605		
<b>Title</b>		
Statistical Methods for Biosecurity Monitoring and Surveillance		
<b>Author(s) / Address (es)</b>		
David Fox, University of Melbourne		
<b>Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)</b>		
Final Report		
<b>Summary</b>		
<p>This report investigates the applicability of traditional methods of analysing surveillance data to biosecurity risks, and explores some more recent innovations designed to detect subtle trends and anomalous behaviour in data over space and time. In particular, it examines control charting and syndromic surveillance methods, and explores how useful they are likely to be in dealing with typical biosecurity disease and pest surveillance. It focuses on disease detection, and methods for optimising surveillance networks and robust methods for minimising levels of inspection.</p> <p>This work provides a proof of concept of these approaches. The case studies, while based on real contexts, are intended only to be illustrative. If the tools are considered to be potentially useful, the next stage would involve development of specific applications to trial their utility.</p>		
<b>ACERA Use only</b>	Received By:	Date:
	ACERA / AMSI SAC Approval:	Date:
	DAFF Endorsement: ( ) Yes ( ) No	Date:

## AUSTRALIAN CENTRE OF EXCELLENCE FOR RISK ANALYSIS

---

Project 06-05

# Statistical Methods for Biosecurity Monitoring and Surveillance



THE AUSTRALIAN CENTRE OF EXCELLENCE FOR RISK ANALYSIS

# Statistical Methods for Biosecurity Monitoring & Surveillance

---

**DAVID R. FOX**

© 2009 The University of Melbourne  
Parkville Victoria 3052  
Phone +61 3 8344 7253 • Fax +61 3 8344 6215  
Email: [david.fox@unimelb.edu.au](mailto:david.fox@unimelb.edu.au)

This report may not be reproduced in  
part or full by any means without the  
express written permission of the  
copyright owner.

**DISCLAIMER**

This report has been prepared by consultants for the Australian Centre of Excellence for Risk Analysis (ACERA) and the views expressed do not necessarily reflect those of ACERA. ACERA cannot guarantee the accuracy of the report and does not accept liability for any loss or damage incurred as a result of relying on its accuracy.

# Table of Contents

LIST OF FIGURES .....	VI
ACKNOWLEDGEMENTS.....	IX
EXECUTIVE SUMMARY .....	1

## CHAPTER 1: Basic Control Charting

1-1 INTRODUCTION .....	5
1-2 BACKGROUND.....	6
1-3 BASIC STATISTICAL CONCEPTS .....	9
1-4 STATISTICAL SIGNIFICANCE.....	17
1-5 CONTROL CHARTS .....	21
1-6 TIME BETWEEN EVENTS .....	27
1-7 TRANSFORMATIONS TO NORMALITY .....	32
1-8 CONTROL CHART FOR TIME-BETWEEN-EVENTS.....	36
1-9 DISCUSSION.....	39

## CHAPTER 2: Bayesian Control Charting

2-1 INTRODUCTION .....	41
2-2 FUNDAMENTALS OF BAYESIAN CONTROL CHARTING .....	43
2-3 A BAYESIAN CONTROL CHART FOR QUARANTINE INSPECTION.....	45
2-3-1 MATHEMATICAL DETAIL .....	46
<i>Updating the prior</i> .....	47
<i>Predictive distributions for <math>X_{t+1}</math> and <math>Y_{t+1}</math></i> .....	49
2-3-2 ADAPTIVE CONTROL LIMITS .....	50
2-4 EXAMPLE – AN ADAPTIVE CONTROL CHART FOR FOOD IMPORTS.....	51
2-4-1 UPDATING THE PRIOR.....	51
2-4-2 SETTING ADAPTIVE TRIGGERS .....	54
2-5 DISCUSSION.....	55

## CHAPTER 3: Space-Time Disease Spread

3-1 INTRODUCTION .....	57
3-2 A CELLULAR AUTOMATA MODEL FOR DISEASE SPREAD.....	58
3-2-1 MATHEMATICAL FORMULATION .....	63
<i>Models for transmission effectiveness</i> .....	65
<i>Initial conditions</i> .....	65

<i>Modelling the disease spread</i> .....	66
<b>3-3 INFERRING THE TIME AND LOCATION OF AN OUTBREAK</b> .....	69
<b>3-4 EXAMPLE</b> .....	72
<b>3-5 DISCUSSION</b> .....	75

## **CHAPTER 4: Sensor Network Optimisation**

<b>4-1 INTRODUCTION</b> .....	79
<b>4-2 A SURVEILLANCE NETWORK FOR EI</b> .....	80
<b>4-3 THE MAXIMAL COVERING LOCATION PROBLEM (MCLP)</b> .....	82
<b>4-4 THE GENERALISED MCLP (G-MCLP)</b> .....	85
4-4-1 A LOGISTIC MODEL FOR DISEASE PROBABILITY .....	87
<b>4-5 PROBLEM FORMULATION</b> .....	89
<b>4-6 EXAMPLE</b> .....	91
4-6-1 IMPLEMENTATION .....	95
<i>Scenario evaluation</i> .....	95
<b>4-7 DISCUSSION</b> .....	102

## **CHAPTER 5: Robust Methods for Biosurveillance**

<b>5-1 INTRODUCTION</b> .....	105
<b>5-2 SURVEILLANCE WITH IMPERFECT DETECTION</b> .....	107
5-2-1 PROBLEM FORMULATION .....	108
<b>5-3 AN INFO-GAP MODEL FOR SURVEILLANCE PERFORMANCE</b> .....	110
<b>5-4 ILLUSTRATIVE EXAMPLE</b> .....	113
5-4-1 COMPARISON WITH A BAYESIAN APPROACH .....	114
<b>5-5 DISCUSSION</b> .....	117
<b>6 REFERENCES</b> .....	119
<b>APPENDIX A : DATA USED IN CHAPTER 1 EXAMPLE</b> .....	126
<b>APPENDIX B: DERIVATION OF EQUATION 2-5</b> .....	127
<b>APPENDIX C: DERIVATION OF EQUATION 2-9</b> .....	129
<b>APPENDIX D : DATA USED IN CHAPTER 3 EXAMPLE</b> .....	133
<b>APPENDIX E : MATHCAD CODE FOR CELLULAR AUTOMATA MODEL</b> .....	135
<b>APPENDIX F : DATA USED IN CHAPTER 4 EXAMPLE</b> .....	138
<b>APPENDIX G: LINGO<sup>®</sup> CODE FOR OPTIMAL SENSOR CONFIGURATION</b> .....	142

## List of Figures

Figure 1. Histogram of number of imported food items inspected each day in the period 1/7/2006 to 29/6/2007.....	9
Figure 2. Graphical and numerical statistical summary number of weekday inspections. Curve overlaying the histogram is the best-fitting normal distribution.....	10
Figure 3. Graphical and numerical statistical summary number of weekend inspections. Curve overlaying the histogram is the best-fitting normal distribution.....	11
Figure 4. Daily inspection failure rate histogram for food items imported between 17/2006 and 30/6/2007. Blue curve is best-fitting normal distribution. ....	13
Figure 5. Boxplot for inspection failure rate for weekends and weekdays. ....	14
Figure 6. Time series plot of number of containers inspected per week (black line) and number of containers found to be of quarantine concern (red line). ....	16
Figure 7. Times series plot of overall failure rate for food imports. Solid blue line is a loess smooth. ....	16
Figure 8. Theoretical distribution for the true proportion of weekday inspection failure rate. ....	18
Figure 9. Theoretical distributions for an individual proportion (black) and the average of 52 proportions (red). ....	18
Figure 10. Assumed distribution for mean of $n=52$ sample proportions. One-tail, 5% 'critical region' identified by red shading. ....	19
Figure 11. Assumed distribution for mean of $n=52$ sample proportions. Two-tail, 5% 'critical regions' identified by red shading. ....	20
Figure 12. P-chart for inspection failure rate data. ....	21
Figure 13. Smoothing using block averaging. ....	22
Figure 14. Moving average scheme. A block or 'window' is stepped incrementally over the series and the block mean computed and plotted. ....	23
Figure 15. Moving average chart of weekday inspection failure rate. Sub-group size =1; MA length=5.....	24
Figure 16. Moving average for weekday inspection failure rate. Subgroup size defined by week of year (usually 5); MA length=4. ....	25
Figure 17. Comparison of exponentially declining weights (red bars) compared with equal-weighting scheme (blue rectangles) for $k=10$ . ....	26
Figure 18. EWMA chart for weekday inspection failure rate. Subgroup size defined by week of year (usually 5); EWMA weight=0.2.....	27
Figure 19. Time sequence of detection of quarantine threats. ....	28
Figure 20. Pattern of inter-arrival times as measured by the 'white space' between blue lines..	28
Figure 21. Histogram of inter-arrival times with smoothed version (red line) and theoretical normal distribution (black line) overlaid. The normal distribution provides a poor description of this data (evidenced by the both the shape and probability mass associated with negative values of days between detects).....	29
Figure 22. Histogram of days between detects. Smoothed histogram indicated by red curve, theoretical exponential distribution depicted by black curve.....	30
Figure 23. Empirical <i>cdf</i> for days between detects (red curve) and theoretical exponential <i>cdf</i> (blue curve).....	31
Figure 24. Chart of individual values of days between detects (I-Chart).....	31



Figure 25. Box-Cox profile plot for the days between detects. Optimal lambda is 0.24.....	33
Figure 26. Histogram of transformed days between detection with smoothed version (red curve) and theoretical normal (black curve) overlaid.....	33
Figure 27. Fitted normal distribution to transformed days between detects with upper 10% point indicated. ....	34
Figure 28. Theoretical negative exponential distribution for untransformed days between detects and upper 10% point indicated.....	35
Figure 29. I-Chart for transformed days between detects. ....	36
Figure 30. Performance characteristics (as measured by equation 1.9) for a one-sided, lower control chart.....	38
Figure 31. Performance characteristics (as measured by equation 1.10) for a one-sided, upper control chart.....	38
Figure 32. Illustrative non-informative priors for true failure rate, theta. ....	47
Figure 33. Daily consignment failure rate (red curve) and cumulative failure rate (blue curve) for imports between 4/7/2006 and 29/6/2007. ....	52
Figure 34. Top: original (subjective) prior density (blue curve) and posterior density (red curve) for true failure rate after 1 year. Bottom: Empirical cumulative failure rate (red line), overall mean failure rate (blue dotted line) and mean of posterior distribution after 1 year (green dashed line). ....	53
Figure 35. Average number of days per year that are conducive to persistence of FMD virus in aerosol (From Cannon and Gardner 1999).....	60
Figure 36. Zonation of EI infected regions in Queensland during the 2007/08 outbreak. <i>Source:</i> <a href="http://www2.dpi.qld.gov.au/extra/ei/maps/QLDInfectedCluster.gif">http://www2.dpi.qld.gov.au/extra/ei/maps/QLDInfectedCluster.gif</a> (accessed 25 January 2008).....	60
Figure 37. Illustration of grid representation used by authorities in managing EI outbreak. ( <i>Source:</i> <i>Source:</i> <a href="http://www2.dpi.qld.gov.au/extra/ei/maps/map8.gif">http://www2.dpi.qld.gov.au/extra/ei/maps/map8.gif</a> ) ....	61
Figure 38. The region of interest in Figure 36 with grid overlay that forms the basis of the CA modelling approach.....	62
Figure 39. The grid in Figure 38 showing an ‘infected’ cell (red shading) and associated spatial pattern for disease transmission.....	62
Figure 40. General CA situation with cell of interest (red shading) and neighbouring cell. $Z$ is a binary variable indicating cell’s disease status. Region bordered by heavy line depicts spatial extent of influence or impact of cell on its neighbours. ....	63
Figure 41. Illustration of a spatial probability model for disease spread for generic cell $\{i,j\}$ at fixed point in time. Grid represents region of interest. Height of surface is proportional to probability of spread of infection from cell $\{i,j\}$ to neighbouring cells. Each plot represents a different range of influence from highly localised (bottom right) to far-ranging (top left).....	66
Figure 42. Illustrative contours of probability representing likelihood of infection around cell having grid coordinates $\{12,8\}$ .....	67
Figure 43. 3-D depiction of progression of disease spread starting with initial outbreak pattern at $T=0$ and at three subsequent time periods. Vertical scale is probability of infection. ....	68
Figure 44. Likelihood profile plot for $k$ (number of time increments since outbreak). ....	69
Figure 45. General situation depicting region of interest with vertical bars depicting empirical rates of infection.....	70
Figure 46. Illustrative likelihood surface for outbreak location.....	72



## **Acknowledgements**

This report is a product of the Australian centre of Excellence for Risk Analysis (ACERA). In preparing this report, the author acknowledges the financial and other support provided by the Department of Agriculture, Fisheries and Forestry (DAFF), the University of Melbourne, Australian Mathematical Sciences Institute (AMSI) and the Australian Research Centre for Urban Ecology (ARCUE).

The author is grateful to the following people for their reviews of earlier versions of individual chapters of this report: Mark Burgman (University of Melbourne); Andrew Robinson (University of Melbourne); Rob Cannon (AQIS); and the anonymous referees who provided valuable feedback on individual chapters. The contributions and assistance of Colin Thompson (University of Melbourne) in the preparation of the material in chapter 5 is greatly appreciated.

## Executive Summary

Wagner et al. (2006) define *biosurveillance* as “a process that detects...outbreaks of disease...monitors the environment...and systematically collects and analyses data”. Clearly, statistics and statistical methods have a critical role to play in all aspects of biosurveillance: detection; monitoring; and data collection and analysis.

Early work on developing statistical tools for biosurveillance for the most part represented a re-working or adaptation of standard, pre-existing methodologies such as control charts and attribute sampling inspection schemes. While these methods are certainly applicable, there is growing recognition that the data and processes underpinning modern biosecurity and biosurveillance deviate from the industrial context in which they were originally developed (Shmueli and Burkom *in press*). Classical ‘frequentist’ statistical tools struggle with the nuances of biosecurity data which invariably exhibit some or all of the following analytical ‘curses’: data paucity; non-normality; non-stationarity; heterogeneous error structures; and over-dispersion. Other issues such as an inability to deal with data from multiple sources, and a model focus on natural/physical processes rather than ‘choice processes’ are also cited as reasons for the failure of traditional methods of monitoring and analysis. The peculiarities of biosurveillance systems demand ‘new’ statistical approaches to both data acquisition and analysis. Techniques that have been successfully applied to the analysis of *syndromic* and climatic data are candidates for biosurveillance.

This report summarises the outcomes of ACERA Project 05/06 which had as its two main aims:

1. Investigate the applicability of ‘traditional’ monitoring methods to the surveillance and detection of bio-security risks &/or threats; and
2. Undertake statistical research in ‘new’ and emerging areas of bio-surveillance that show promise for their ability to identify and predict trends and anomalous behaviour in both space and time.

Objectives 1 and 2 above have been met, and as the project evolved over the past three years, so too did its direction and emphasis. Chapter 1 of this report is associated with

activities under objective 1. Chapter 2 extends the traditional control charting methods and develops this within a Bayesian context for quarantine inspection while chapters 3 to 5 report on research undertaken as part of the second objective. Specifically, chapter 3 focuses on disease outbreak detection in time and space; chapter 4 examines methods for optimising surveillance network designs; and chapter 5 looks at robust methods for determining minimum levels of inspection.

During the course of this project a number of ancillary activities were undertaken. These included:

- Meetings with key U.S. researchers working in the area of *syndromic surveillance*. This led to an enhanced awareness and understanding of statistical methods used for analysing space-time clusters of ‘incidences’ (eg. disease outbreaks);
- Creating and fostering linkages with research groups at Harvard; N.Y. Department of Health and Rutgers University (DIMACS);
- Presentations (SRA conference, Melbourne; AQIS, Canberra; Rutgers University, New York; International Statistical Institute conference, Lisbon);
- Participation in “Workshop on Info-Gap Applications in the Life Sciences”, University of Houston, 11–15 Sept. 2006. An outcome from this workshop was preparation of a manuscript titled: “*Robust Profiling for Quarantine Inspection*” (Fox, D.R., Moilanen, A. and Beare, S.);
- Organising an international “Uncertainty and Surveillance” Workshop August 12–17, 2007 Hobart. The purpose of the workshop was to bring together individuals from a diversity of backgrounds to discuss surveillance and uncertainty in the context of bio-security. Topics considered by the working group were drawn from the following list:
  1. Surveillance for exotic diseases in animal / plant populations
  2. How to design monitoring/surveillance programs for events for which we have no data and hope to never have data eg. catastrophic events having unimaginable consequences?
  3. Attribute sampling and inspection – how many samples; where; and when in order to declare pest or disease-free status? Distinction between sampling zeros and structural zeros.
  4. Characteristics of a surveillance network that will make it most robust in discovering emerging animal diseases early while adhering to cost and performance criteria.

5. Resource Allocation (static). Inspection resources are limited. N sites could be inspected. There is uncertainty in any or all of: cost of inspection, probability of detection, consequence of missed detection.
6. Resource Allocation (dynamic). This is an extension of the previous problem. Here the temporal dimension enters. Inspection resources are limited. N sites could be inspected. There is uncertainty in any or all of: cost of inspection, probability of detection, consequence of missed detection. We assume the inspector learns from inspection process and adjusts his/her behaviour. This can be studied from a multitude of perspectives.
7. Search and evasion: This is a generalization of the previous problem. The main extension is that now we consider strategic behaviour on the part of the target: the target adjusts his behaviour in response to the behaviour of the inspector. That is, both inspector and target behave strategically.

The outcomes of this project should be viewed as a starting point for further exploration and analysis. We have developed to a 'proof of concept' stage a number of strategies and ideas which we believe have the potential to enhance biosecurity monitoring and surveillance. The need to refine these methodologies and 'road test' them within an actual quarantine inspection program is an essential next step. Early investigations were frustrated by a lack of data and resources. ACERA Project 08/04 provided much-needed data for this project and efforts to enhance the flow of data and information between ACERA and Client agencies is strongly encouraged.

*This page intentionally blank*

## **1-1 INTRODUCTION**

We review some basic statistical concepts as well as introducing some common control-charting techniques that have been successfully applied in areas as diverse as the manufacturing industries and veterinary epidemiology. The review focuses on the applicability of control charts for monitoring temporal trends and aberrations in bio-security related applications. Control charts are particularly well suited to the visualisation and assessing of moderate to large volumes of time-based data and as such would be expected to have greater utility for container inspection regimes say, than for detecting the occurrence (in space) of an invasive species. Control charts need to be viewed as just one method in a tool-kit of available techniques which can potentially assist field officers and quarantine risk assessors in identifying ‘unusual’ or ‘aberrant’ trends. For events having very low probabilities of occurrence (eg. exotic disease outbreak) the monitoring of ‘time between outbreaks’ is a potentially more useful quantity to be charting although as shown in this report, the statistical power (ability to correctly identify real ‘shifts’ in the mean time between events) of current charting techniques is relatively low.

The various forms of charting in the context of a quarantine inspection service are illustrated using imported food inspection data gathered under ACERA Project 08/04 “*AQIS Import Clearance Data Framework*”



## 1-2 Background

This chapter is intended to provide managers, field officers, and researchers who have some responsibility for biosecurity monitoring and surveillance with an introduction into the principles and procedures of statistical process control (SPC) and in particular, control chart techniques.

Statistical process control can be broadly defined as the (statistical) methodologies and tools by which ‘quality’ is monitored and managed. With this definition, the original context of SPC is clear – to ‘control’ the quality of manufactured items in an industrial process or setting, although these days the word *control* is de-emphasised and is usually either dropped<sup>1</sup> or replaced by *improvement*. The development of SPC techniques can be traced back to the First World War and shortly thereafter with the introduction of the Shewhart control chart in the 1920s. General acceptance and uptake of SPC tools in the West was relatively slow until it was realised that a major contributing factor to the high productivity *and* quality of Japanese manufactured goods was due to that country’s enthusiastic embrace of a total quality philosophy as espoused by leading (American) statistician and quality advocate– Edwards Deming.

The 1980s saw a resurgence of interest in SPC under the banner of ‘Total Quality Management’ or TQM. The ‘six-sigma’ philosophy was conceived during this time in response to Motorola’s desire to achieve a tenfold reduction in product-failure levels within five years. The Six Sigma methodology (based on the steps Define - Measure - Analyse - Improve – Control) underpins the objectives of process improvement, reduced costs, and increased profits.

While much good work was done in the ensuing years with numerous examples of demonstrable success attributed to the SPC/TQM paradigm, the trend attracted some poorly credentialed ‘experts’ offering radical transformations to new (and often times unrealisable) levels of profitability. Not surprisingly, there were some failures and residual ill-feeling towards TQM. For example, one large company found that two thirds of the

---

<sup>1</sup> The American Society for Quality Control (ASQC) changed its name to the American Society for Quality (ASQ) on July 1, 1997.

TQM programs it examined had been halted due to lack of results while a 1994 American Electronics Association survey showed that TQM implementation had dropped from 86 percent to 63 percent, and that reductions in defect rates were not being realised (Dooley and Flor, 1998).

Despite some negative experiences in the manufacturing sector, TQM made a substantial contribution to quality improvement. The issue these days is one of how much quality is enough? For example, would anyone be interested in driving a 50-year old vehicle even if it was still under warranty? Or is there any value in ensuring that a computer will run reliably for 5 years when generational change in the computer industry is measured in months?

Environmental applications of TQM and SPC techniques have only more recently been identified despite the need for robust and reliable monitoring and surveillance systems. Fox (2001) attributes this to a lack of cross-talk between the 'brown' (industrial) statisticians and the 'green' (environmental) statisticians. Whatever the reasons, the slow uptake of SPC tools for environmental monitoring meant that critical assessments about environmental condition and important decisions about responses were being made on the basis of often-times flawed statistical advice. The ecologists' statistical toolkit was generally standard issue – t-tests, ANOVA, ANOSIM, and MDS were invariably represented and much used while the relatively simple techniques of Xbar/S charts, EWMA charts, and capability analysis were virtually unheard of.

The Australian Guidelines for Fresh and Marine Water Quality (ANZECC/ARMCANZ 2000) advocated a more 'risk-based' approach to water quality monitoring and assessment and in particular, a reduced emphasis on binary decision-making (t-tests and ANOVA) and increased emphasis on early-warning systems underpinned by dynamic visualizations (control charting). Shortly after the release of the Guidelines the events of 9/11 and the subsequent discovery of high-grade anthrax sent via the U.S. Postal service elevated the importance of early warning systems. In response, U.S. State and Federal governments have invested hundreds of millions of dollars in developing advanced surveillance systems to detect, among other things, another anthrax attack. Recognising that existing monitoring systems which relied on the gathering and processing

of hospital records had unacceptable latencies, a new area of research emerged which aimed to provide close to real-time monitoring and warning of ‘aberrant’ events.

*Syndromic surveillance* is underpinned by a belief that signals of an emerging ‘syndrome’ such as a flu outbreak can be identified by an analysis of multiple time-series of ancillary variables such as absenteeism records and sales of non-prescription cold and flu medications together with an analysis of spatial clustering of outbreaks. Kulldorf (1997) developed a spatial scan statistic to help with the latter, while control charts were an obvious first candidate for the former. A major barrier at present is the difficulty in ‘proving’ that any of these new systems has made a difference or even do what they’re meant to. As noted by Mostashari and Hartman (2003), no syndromic surveillance system has provided early warning of bioterrorism, and no large-scale bioterrorist attack has occurred since existing systems were instituted.

While the use of syndromic surveillance for counter-terrorism (see for example, <http://www.bt.cdc.gov/surveillance/ears/>) is a recent development, similar systems have been used for some time now to detect outbreaks, patterns, and trends in diseases and epidemics (see for example, <http://www.satscan.org/>). These techniques do not appear to have had any appreciable uptake in Australia or elsewhere around the world in quarantine inspection and bio-security. While adoption and uptake of SPC techniques in the Australian Quarantine and Inspection Service has been low, a search of the Department of Agriculture, Fisheries and Forestry (DAFF) website ([www.daff.gov.au](http://www.daff.gov.au)) reveals two (publicly available) documents that discuss the use of simple control charts (Korth 1997, Commonwealth of Australia 2002). One of these documents (Commonwealth of Australia, 2002) devotes a chapter to the use of control charting as an effective means of detecting trends for meat hygiene assessments. Control charting has also been recommended for detecting spatial and temporal clusters in veterinary monitoring (Carpenter, 2001) as well as testing waste streams from wastewater treatment plants (Hall and Golding, 1998). Stark et al. (2006) discuss risk-based veterinary surveillance approaches to protecting livestock and consumer health although control-charting was not mentioned.

This chapter provides a review of basic SPC techniques (focusing on control charting methods) with a view to introducing our Client Agencies to ‘new’ or alternative approaches to monitoring that may offer enhanced anomaly detection capabilities. With a shared

understanding of the control charting principles and an understanding of the strength and weaknesses of these methods, it is hoped that opportunities for implementation and evaluation in an actual quarantine inspection / bio-surveillance environment will emerge.

### 1-3 BASIC STATISTICAL CONCEPTS

Statistics is concerned with random variation. Moreover, statistics is concerned with *random* variation. This does not mean that life for a statistician is totally unpredictable. The quantities exhibiting the random variation (the **random variables**) can be ‘predicted’ or described to the extent that in repeated ‘trials’ or observations, the values assumed by the random variable can be described by a *frequency distribution*. Figure 1 shows the **histogram** for the number of imported food items inspected per day between July 1 2006 and 30 June 2007.

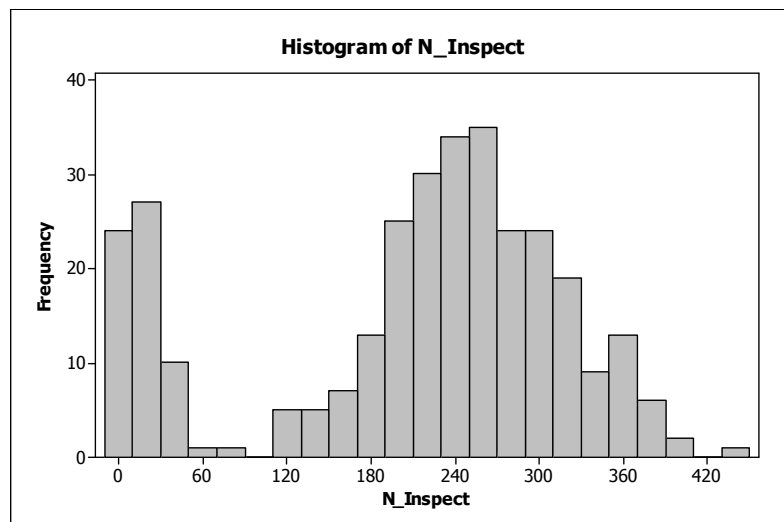


Figure 1. Histogram of number of imported food items inspected each day in the period 1/7/2006 to 29/6/2007.

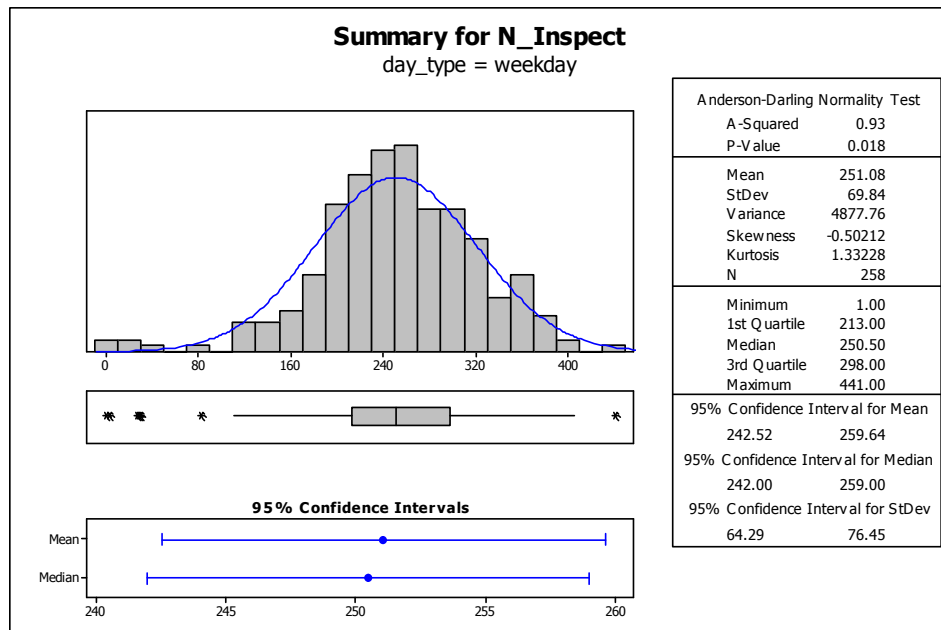
A number of features are immediately apparent from Figure 1: (i) the *range* of items inspected/day is from zero to around 440 ; (ii) the most frequent number of inspected items is approximately 250 - this is the *modal* value; (iii) the histogram exhibits bimodality (i.e. two ‘peaks’); (iv) an eyeball estimate of an *average* or *mean* value is about 200 items/day.

With respect to the bimodality, two groupings are evident: <100 inspections/day; >100 inspections/day. Further analysis reveals that this is simply a weekend / weekday dichotomy as highlighted by 2-way breakdown in Table 1<sup>2</sup>

**Table 1-1. Two-way breakdown of number of inspections according to weekday/weekend.**

	Fewer than 100 inspections/day	Greater than 100 inspections/day	Totals
<b>Weekend</b>	57	0	57
<b>Weekday</b>	6	252	258
<b>Totals</b>	63	252	315

As a companion to the graphical summary provided by the histogram, we generally compute relevant *statistics* for our data. Most software tools (such as MINITAB) combine both the graphical and numerical summaries. Individual summaries for the number of daily inspections are provided for both weekdays (Figure 2) and weekends (Figure 3).



**Figure 2. Graphical and numerical statistical summary number of weekday inspections. Curve overlaying the histogram is the best-fitting normal distribution.**

<sup>2</sup> As an aside, the data in Table 1 are conventionally analysed using *contingency table* methods and a chi-squared test of independence. The result of such an analysis when applied to Table 1 is highly significant ( $p < 0.0001$ ) suggesting that the number of items inspected is *not* independent of the weekday/weekend classification.

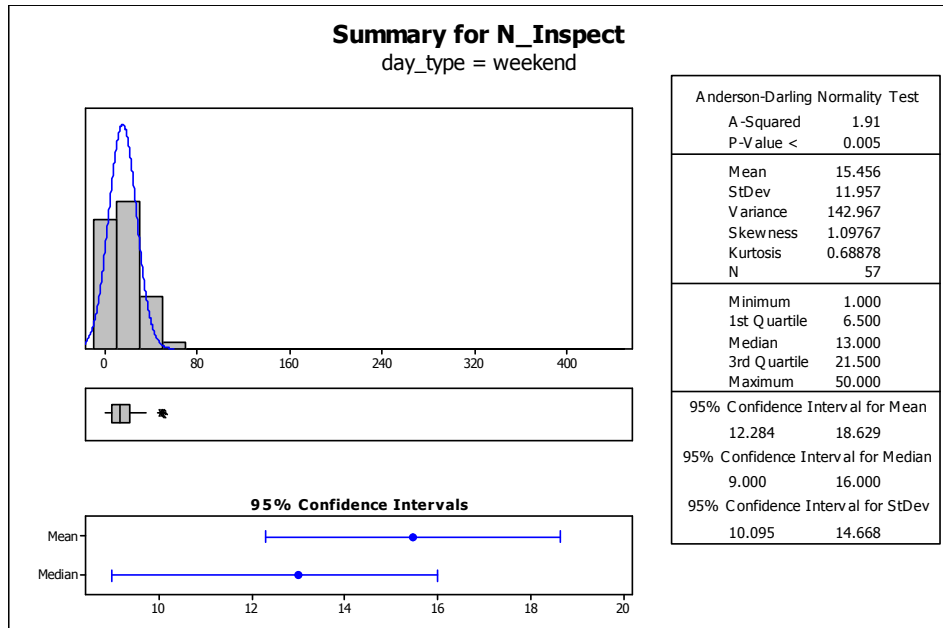


Figure 3. Graphical and numerical statistical summary number of weekend inspections. Curve overlaying the histogram is the best-fitting normal distribution.

The numerical summaries presented on the right-hand side of Figures 2 and 3 are explained below:

### Anderson-Darling Normality Test

Because many statistical procedures (including control charting) rely on an implicit assumption of underlying normality in the distribution of the quantity of interest, the best-fitting normal distribution for the data at hand has been identified and overlaid on the sample histogram. While this allows for a visual inspection of the plausibility of the normality assumption, a more formal statistical test can be performed. There are many such tests – some good, others not so good (see for example Stephens, 1974). One such test that has been shown to perform well under a wide range of conditions is the Anderson-Darling test. The implicit hypothesis being tested is that the sample data has been drawn (at random) from a much larger population of values which is normally distributed. In this case, the *test-statistic* is a value of  $A^2$  (0.14 in Figure 2). The numerical value is rather meaningless by itself. In order to gauge the *significance* of the result we look at the companion *p-value*. The rule-of-thumb is that small p-values lead to a rejection of the implicit hypothesis (otherwise known as the *null hypothesis*). So, how small is small? Convention dictates that p-values of less than 0.05 are ‘significant’. However you are cautioned against the unthinking adoption of this 0.05 norm. Since the p-value of 0.971 is well above the nominal 0.05, we do not reject the hypothesis of normality.

## Mean

This is the usual *arithmetic mean* or *average* of the data. Other means are available (geometric and harmonic) although they are not widely used.

## StDev and Variance

This is the *standard deviation* and is the most common measure of spread (or *dispersion*) of a data set. It is defined as the positive square root of the *variance*. Since the variance is computed by summing the square of differences formed by subtracting the (sample) mean from each data value, it (the variance) can only ever take on non-negative values. The reason for preferring the standard deviation as a measure of spread is that it has the same units as the original data values.

## Skewness

As the name suggests, this is a measure of *skewness* or *asymmetry*. Skewness can be *negative* (long-tail to the left) or *positive* (long-tail to the right). Symmetrical distributions have *zero* skewness.

## Kurtosis

This is not a quantity that is used often in its own right. It is a numerical measure of '*peakedness*' of a distribution. A flat distribution is said to have low kurtosis while a highly peaked distribution has high kurtosis. Different software packages will compute skewness slightly differently. The normal distribution has a skewness of 3. MINITAB and other software tools subtract 3 from the computed skewness so as to make the measure relative to a normal distribution. The skewness of -0.0834 in Figure 2 is very close to zero, implying that this sample data is about as peaked as a normal distribution.

## Minimum and Maximum

These are self-evident.

## First, second, and third quartiles

The *quartiles* are numerical values that divide the distribution into four equal parts. The first quartile ( $Q_1$ ) is such that 25% of all values are less than or equal to  $Q_1$ ; the second quartile ( $Q_2$ ) is such that 50% of all values are less than or equal to  $Q_2$ ; while the third quartile ( $Q_3$ ) is such that 75% of all values are less than or equal to  $Q_3$ .  $Q_2$  is also known as the *median* value. The difference  $Q_3 - Q_1$  is referred to as the *inter-quartile range (IQR)* and is a measure of spread or variation.

### Confidence Intervals: mean, median, and standard deviation

An explanation of a confidence interval would require considerably more than a few lines. However, a practical interpretation is that we may assert that the *true parameter value* lies within the stated interval with stated degree of confidence.

A quantity of potentially greater interest than the number of inspections performed is the *failure rate*,  $p$  defined as  $p = \frac{N_f}{N_I}$  where  $N_f$  is the number of failed items out of  $N_I$  inspected. The histogram of daily failure rates (Figure 4) is slightly positively skewed (i.e to the right) and is not well described by a normal probability model.

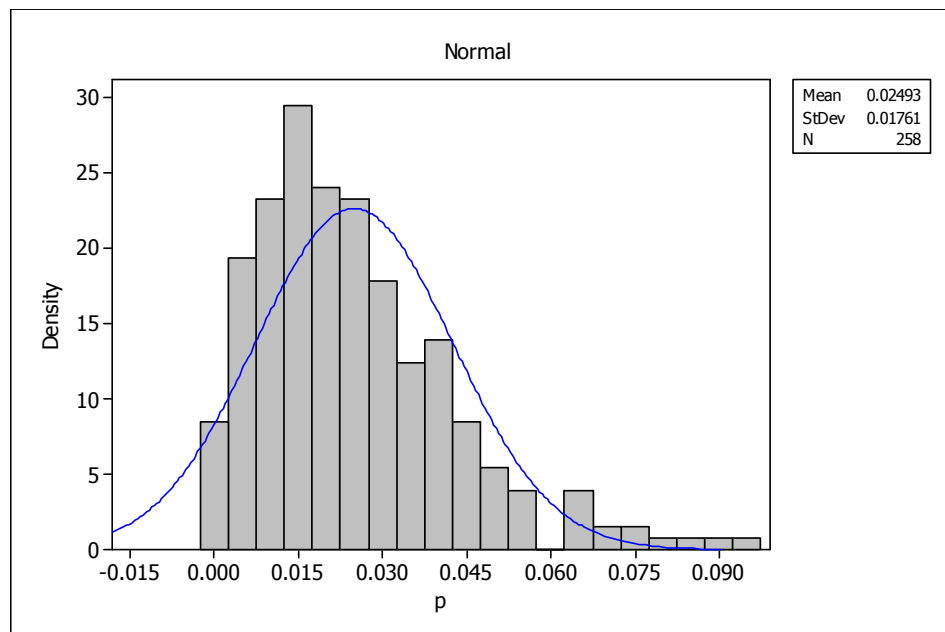


Figure 4. Daily inspection failure rate histogram for food items imported between 17/2006 and 30/6/2007. Blue curve is best-fitting normal distribution.

Descriptive statistics for  $p$  for weekdays and weekends are given in Table 1-2. It is seen that while far fewer inspections are undertaken on weekends, the average failure rate is essentially the same as for weekdays, although the variability in failure weekend failure rates is much greater.

Table 1-2. Numerical summaries of inspection failure rates for weekdays and weekends.

day_type	N	N (missing)	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
weekday	258	0	0.02493	0.0011	0.01761	0	0.01219	0.0211	0.0342	0.09272



<b>weekend</b>	57	0	0.02524	0.00669	0.05048	0	0	0	0.03399	0.24
----------------	----	---	---------	---------	---------	---	---	---	---------	------

Another way of presenting the data is the *box-plot*. The box-plot for the proportion data is shown in Figure 5. There are a number of variations on how the box-plot is constructed so you should check your computer software to be clear on the specifics. The information given by MINITAB (the statistical software used to produce Figure 5) is shown in Box 1 and general information on graphical summaries from SYSTAT is shown in Box 2.

A number of features are immediately apparent from Figure 5: (i) the distributions of failure rates are approximately symmetrical around a common average of about 0.02 although the weekend distribution is slightly positively skewed; and (ii) the weekend distribution shows greater variability than the weekday distribution as evidenced by the larger interquartile range (*IQR*) and a number of relatively large values.

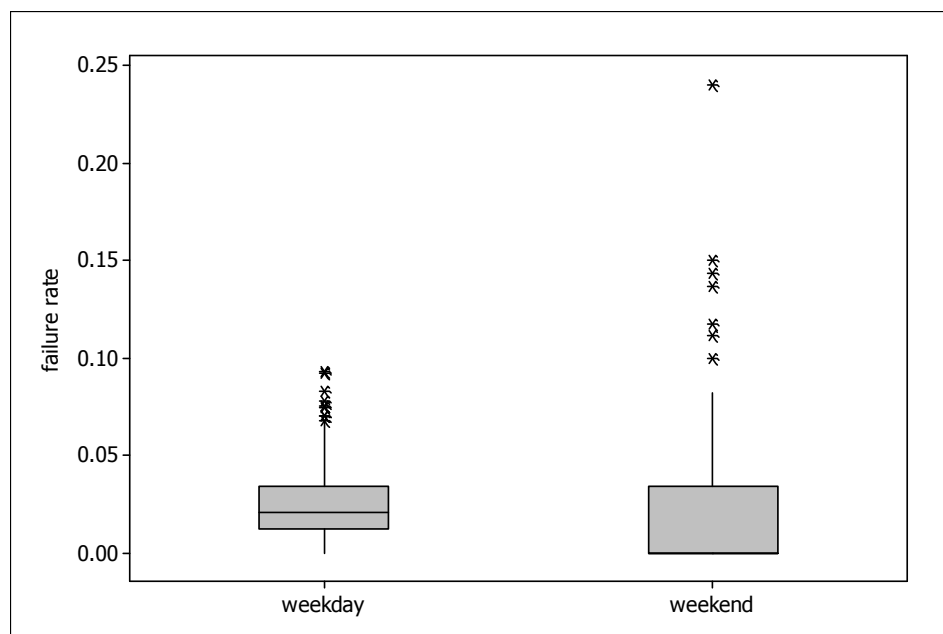



Figure 5. Boxplot for inspection failure rate for weekends and weekdays.

## Box 1. MINITAB's Help on Box-plots

**Graphical Summary**  
Boxplot (1 of 1)

Boxplots summarize information about the shape, [dispersion](#), and center of your data. They can also help you spot [outliers](#).

- The left edge of the box represents the [first quartile](#) (Q1), while the right edge represents the [third quartile](#) (Q3). Thus the box portion of the plot represents the [interquartile range \(IQR\)](#), or the middle 50% of the observations.
- The line drawn through the box represents the [median](#) of the data.
- The lines extending from the box are called [whiskers](#). The whiskers extend outward to indicate the lowest and highest values in the data set (excluding outliers).
- Extreme values, or outliers, are represented by dots. A value is considered an outlier if it is outside of the box (greater than Q3 or less than Q1) by more than 1.5 times the IQR.

Use the boxplot to assess the [symmetry](#) of the data:

- If the data are fairly symmetric, the median line will be roughly in the middle of the IQR box and the whiskers will be similar in length.
- If the data are [skewed](#), the median may not fall in the middle of the IQR box, and one whisker will likely be noticeably longer than the other.

In the boxplot of the precipitation data the median is centered in the IQR box, and the whiskers are the same length. This indicates that except for the outlier (asterisk), the data are symmetric. This is a good indication that the outlier may not be from the same population as the rest of the sample data.

## Box 2. SYSTAT's help on graphical summaries.

**Histogram, Box plot, Dot density, Density function**

---

The density of a sample is the relative concentration of data points in intervals across the range of the distribution. A histogram is one way to display the density of a quantitative variable; box plots, dot or symmetric dot density, frequency polygons, fuzzygrams, jitter plots, density stripes, and histograms with data-driven bar widths are others.

A histogram is the most familiar one among these displays. The word comes from a Greek word (*histos*) for a straight standing beam, like a mast or loom frame, and a word (*gram*) for a drawn picture. Thus, a histogram is a pictorial display of vertically standing bars. It is a crude density estimator because the shape of a histogram depends upon the choice of the number of bars. Most other graphical density estimation methods depend on subjective choices of parameters (or settings) as well, which is one reason the general field of density estimation is rather controversial (Wegman, 1982).

The software can use the sample mean and standard deviation to construct a normal curve (or cumulative normal curve) for comparison against the actual anomalies of the sample distribution. A kernel curve is also available for density and distribution curves.


Rather than comparing sample values to the normal distribution (mean, standard deviation, and so on), box plots show robust statistics (median, quartiles, and so on). Some complain that box plots or the choice of intervals for bars in a histogram can mask gaps or separations in the distribution. Dot histograms (dit) and symmetric dot displays (dot) answer this problem because they display every value in the sample. It is often useful to examine both a box plot and a dit display. A gap histogram is another alternative. Its bar widths vary across the range of the distribution--when there are gaps, the neighboring bar is made wider to include the gap.

Fuzzygrams superimpose a probability distribution on each bar of a histogram. Bars for histograms based on small samples are fuzzier than bars for large sample histograms. Jittered dot density displays points by calculating the exact locations of the data values and then, to keep points from colliding, jittering them randomly on a short vertical axis. These displays work better for large samples than small samples. Density stripes are vertical lines placed at the location of data values along a horizontal data scale and look like supermarket bar codes. For large samples, the stripes tend to collide, so you should consider a jitter dot density instead.

Bivariate densities can be displayed as 3-D histograms and as 2-D surfaces or contours constructed using normal theory sample statistics or a nonparametric kernel estimator.

These displays can be stratified across the levels of a grouping variable; for 2-D displays, if the grouping variable has only two values, a dual (or back-to-back) version is available.

The Dynamic Explorer is available for rotating 3-D displays and also for fine tuning all displays.



In addition to the graphical summaries already presented, it will be useful to display the time-series plot (ie. values of a variable plotted over time) for the monitored data. Figure 6 shows such a plot of the weekend and weekday data. Given that the number of failures is very much smaller than the number of items inspected, plotting both quantities on the same axes is of limited value. Given that both the number of containers inspected and the number of 'detects' vary with time, a more useful quantity to plot is the failure rate,  $p$  (Figure 7).

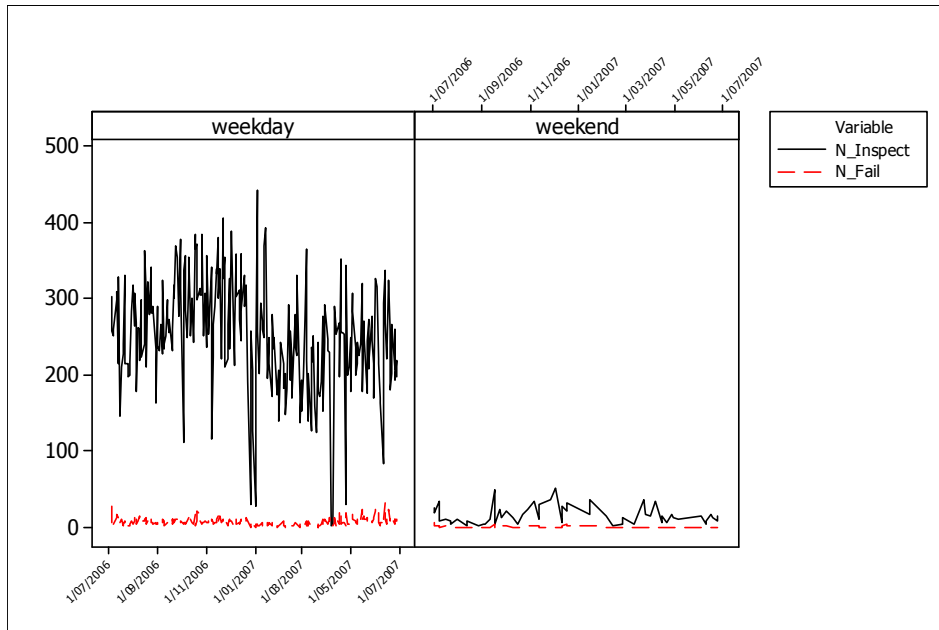


Figure 6. Time series plot of number of containers inspected per week (black line) and number of containers found to be of quarantine concern (red line).

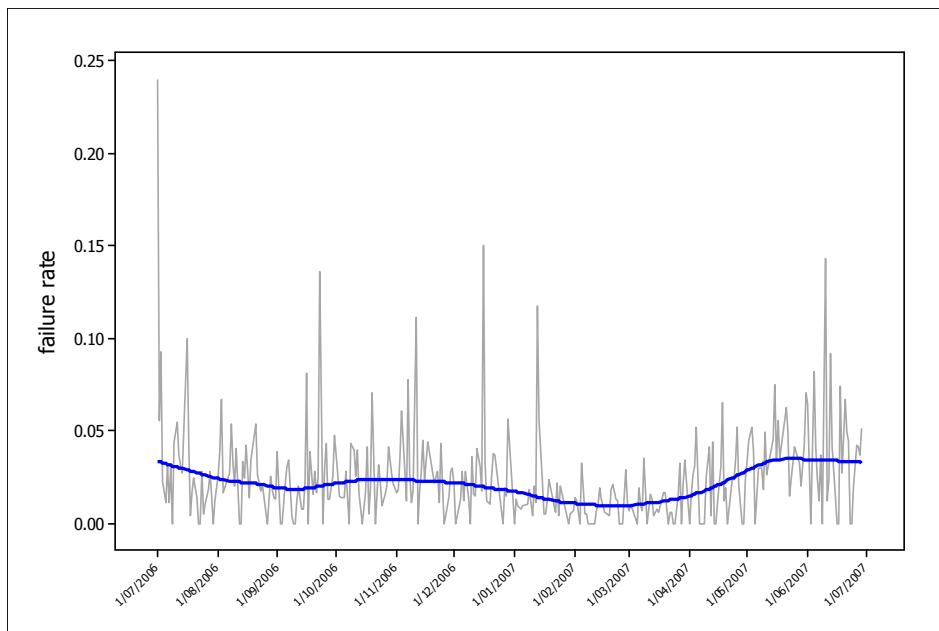


Figure 7. Times series plot of overall failure rate for food imports. Solid blue line is a loess smooth.

Figure 7 shows the failure rate generally fluctuates between zero and 5%. The blue line in Figure 7 is the result of applying a local smoothing procedure, called a ‘loess’ to the failure rate data. The idea behind the loess (and other procedures like it) is that the high frequency oscillations can be removed by sub-setting the data and replacing individual data points by some statistic (such as the median of the data in the sub-set). In this way, the relatively ‘noisy’ data are smoothed. The degree of smoothing can be varied so as to reveal different features in the time-series data. In the case of Figure 7, we see that the failure rate declined slowly between July 2006 and October 2006; was relatively constant between October 2006 and January 2007 and then increased slightly until April 2007. There was a significant increase between April and June 2007. Also evident in Figure 7 is a period of low variability between February and April 2007 followed by a period of substantially higher variability. Whether or not these observations correlate with other known facts or can be attributed to known causes is a matter for the relevant agencies. One way of helping address the ‘significance’ of these variations is through the use of *control charts*.

## 1-4 Statistical significance

While a comprehensive treatment of statistical inference is outside the scope of this introductory note, a brief discussion of the main ideas underpinning ‘statistical significance’ is warranted.

To place the discussion in context, suppose that it is known from lengthy monitoring that the proportion of imported food items that fail inspection is normally distributed with a mean of 0.02493 and a standard deviation of 0.01761. Statisticians write this in an abbreviated way as  $\theta \sim N(0.02493, 0.01761^2)$  where  $\theta$  denotes the true proportion. Figure 8 shows a plot of the *probability density function* (or *pdf*) for this particular normal distribution.

Now, suppose we collect  $n=52$  weekly proportions and looked at the average (arithmetic mean) proportion of containers classified as a quarantine risk. Statistical theory tells us that if  $\theta \sim N(0.02493, 0.01761^2)$ , then the average of  $n$  sample proportions is distributed as  $\bar{\theta} \sim N(0.02493, 0.01761^2/n)$ . For  $n=52$ , this distribution is shown in Figure 9 (together with the original distribution of Figure 8).

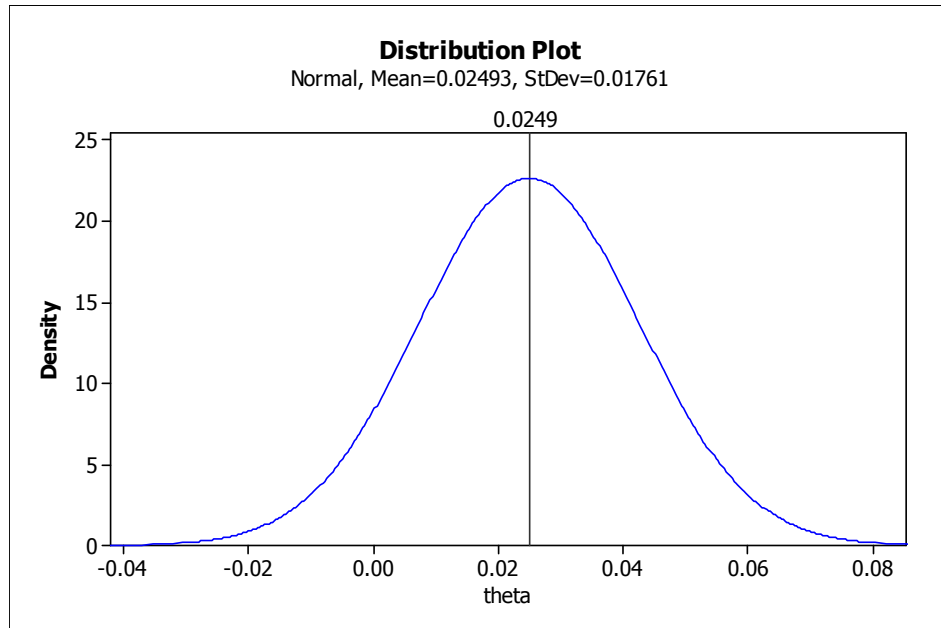


Figure 8. Theoretical distribution for the true proportion of weekday inspection failure rate.

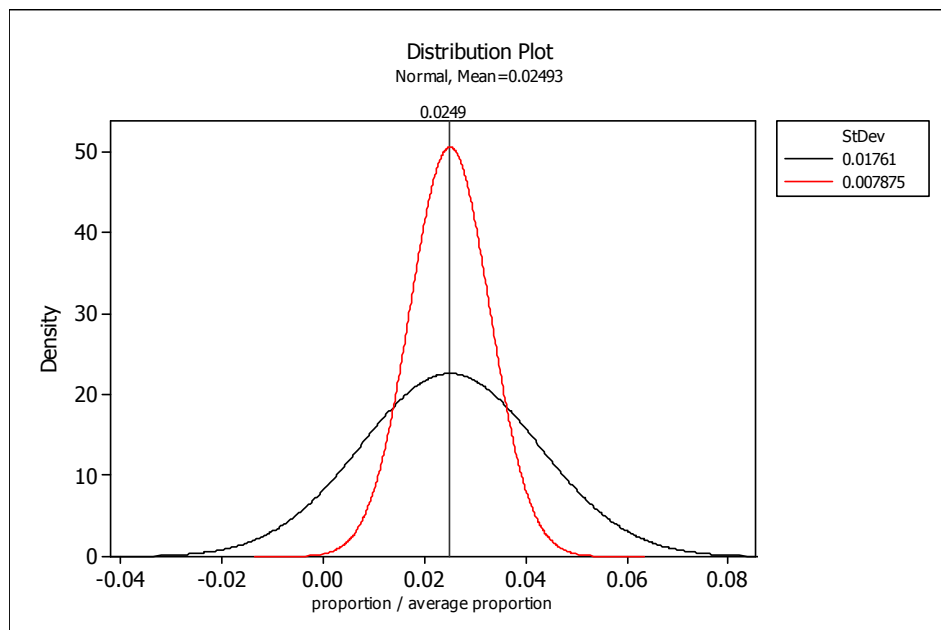


Figure 9. Theoretical distributions for an individual proportion (black) and the average of 52 proportions (red).

It is readily seen from Figure 9 that although both distributions are centred about the same mean value, the distribution of the average is more compactly so. Thus, while an

*individual* proportion of say, 0.045 or more is expected to occur quite frequently it is highly unlikely that the average of 52 readings would be this high if the individual values had been drawn from a population described by the red curve in Figure 9. In statistical parlance, we would say that an average (of  $n=52$ ) proportion of 0.045 or greater is a highly *significant* result. This immediately raises the question of ‘how significant is significant’ or where do we draw the line between statistical significance and non-significance? Convention dictates (and this is a potentially dangerous approach without understanding the ramifications) that we choose a cut-off value (call it  $\theta^*$ ) such that the probability  $P[\bar{\theta} > \theta^*]$  is ‘small’ – and small is taken to mean 0.05. For our example, we find that a  $\theta^*$  equal to 0.0379 has an associated ‘tail probability’ of 0.05 (Figure 10). So, if we are only interested in large proportions, then we will declare a sample average (of  $n=52$ ) to be *statistically significant* if it is numerically greater than 0.0379. If we are interested in both increases and decreases in the assumed proportion, we split the 0.05 area into two ‘tail’ areas each of 0.025. This generates a *two-sided* test of significance (Figure 11).

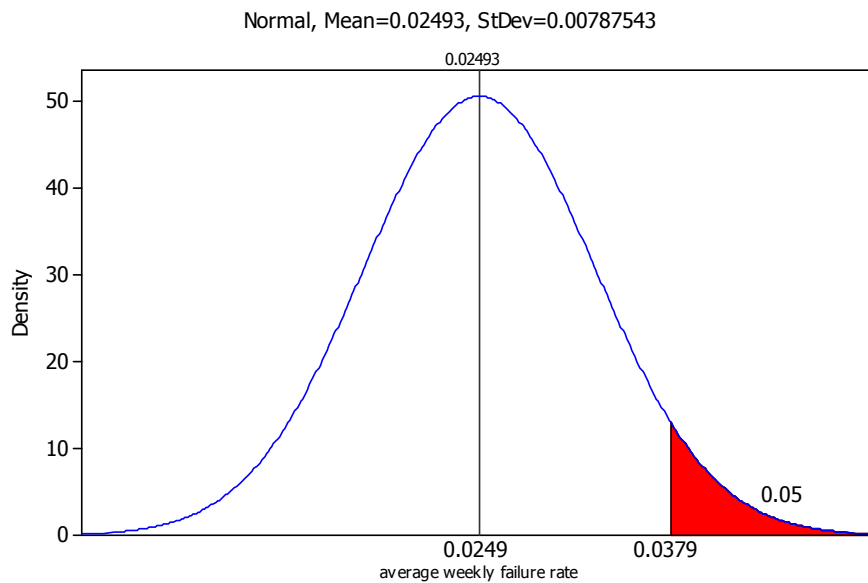
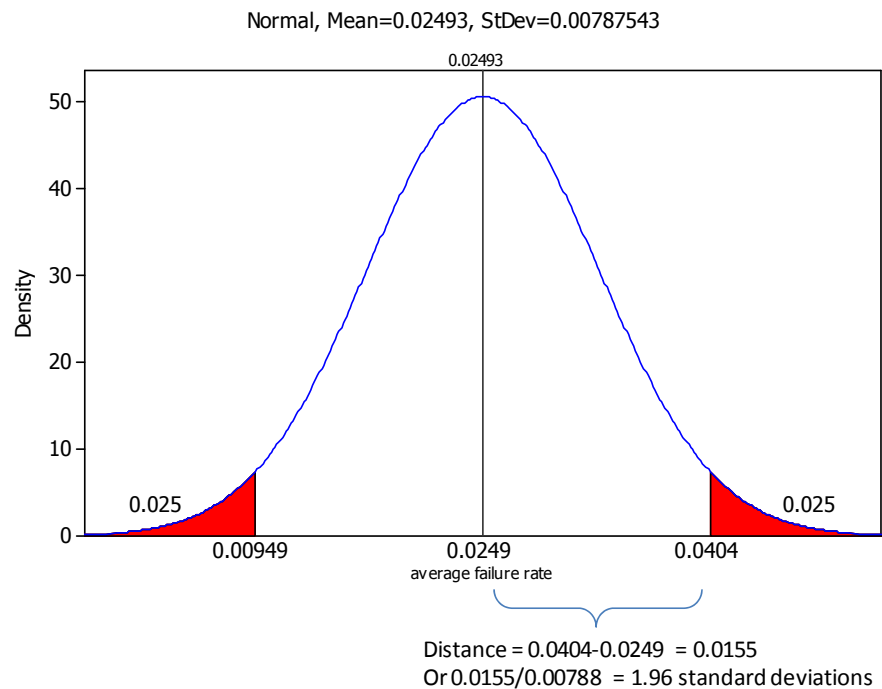


Figure 10. Assumed distribution for mean of  $n=52$  sample proportions. One-tail, 5% ‘critical region’ identified by red shading.



**Figure 11. Assumed distribution for mean of n=52 sample proportions. Two-tail, 5% 'critical regions' identified by red shading.**

It is also seen from Figure 11 that the upper limit of 0.0404 is equivalent to 1.96 standard deviations from the mean (it is easily verified that the lower limit is also 1.96 standard deviations from the mean, but obviously in the opposite direction). These  $\pm 1.96$  'sigma limits' or 'control limits' form the basis of 'early warning' or detection limits on control charts. The multiplier (1.96) determines the width of the limits which in turn determines important statistical properties associated with 'false triggering'. Thus, there is a trade-off to be struck: very narrow limits will give high sensitivity to shifts in the 'process' but at the expense of increased rates of false-triggering. By default, limits on control charts are placed at either 2 or 3 standard deviations from the mean (centre-line). Limits determined through non-statistical considerations (eg. biological or ecological significance) can also be used, although the implications for 'statistical significance' would need to be determined if the chart was to be used in an inferential mode.

## 1-5 Control Charts

### The Basic Shewhart Chart

The simplest control chart is essentially Figure 7 with the addition of upper and/or lower control limits. This could be done manually, although it is easier to have computer software do it. The output from the MINITAB statistical software package is shown in Figure 12.

Note that the upper and lower control limits in Figure 12 are not constant. This is because the denominator in the expression  $p = \frac{N_f}{N_I}$  is not constant (as noted by the warning message in the lower right corner of Figure 12). The red plotting symbols in Figure 11 indicate ‘violations’ or ‘excursions’ outside the control limits. By itself, this chart raises no particular concerns, other than there were two occasions when the proportion was ‘significantly’ high and three occasions when it was ‘significantly’ low. Various modifications and options are available in the presentation of control charts such as Figure 12. For example, if we think there are two ‘epochs’ as discussed earlier, we can provide separate limits in each epoch (Figure 12).

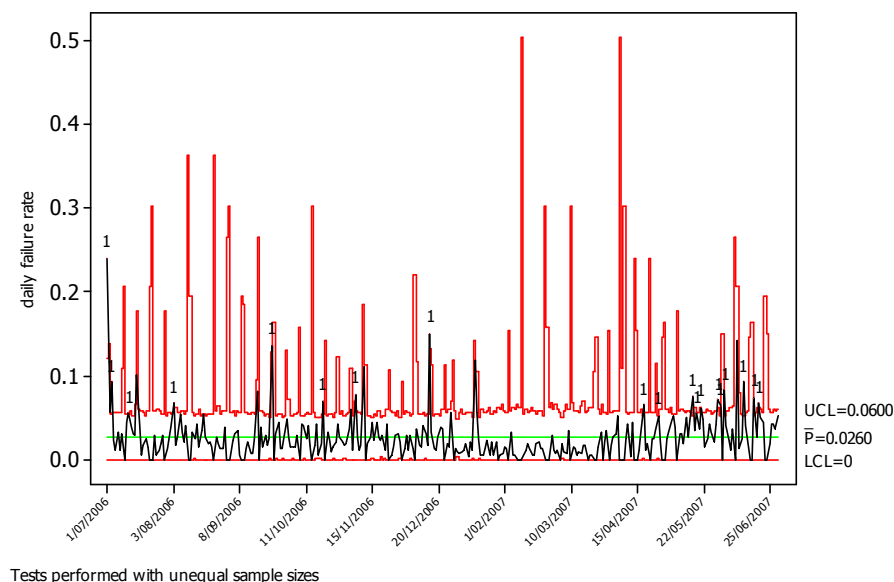


Figure 12. P-chart for inspection failure rate data.



For data that are collected over *time* a number of time-based control charts are available. Some of the more common/useful are described below.

### Time-based charts

The simplest way of smoothing over time is by ‘block-averaging’. Figure 13 shows a time series divided into a number of non-overlapping ‘blocks’ of constant width. The average of the data in each block is computed and plotted at the centre of the block and these points can be connected by straight line segments to reveal a smoother version of the series.

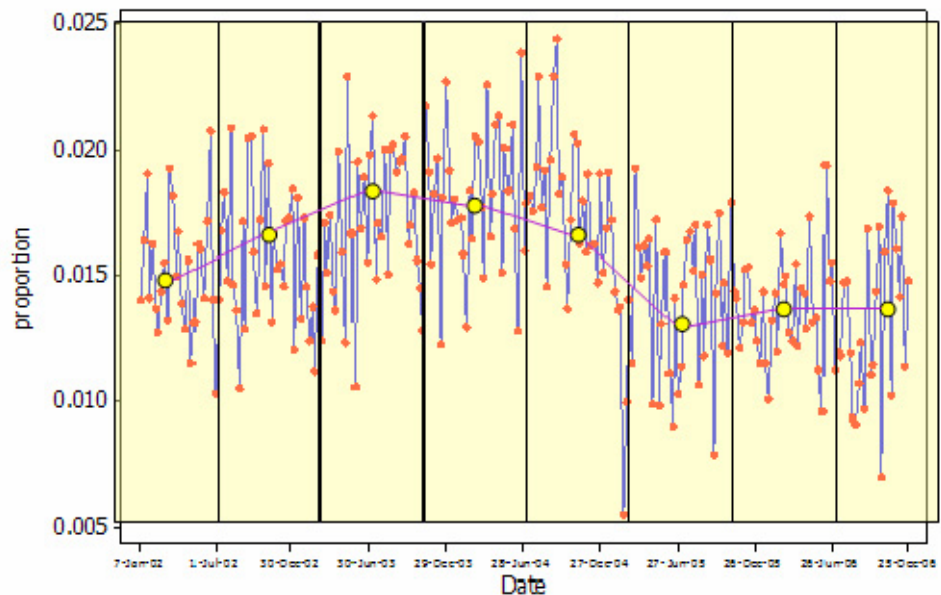


Figure 13. Smoothing using block averaging.

Block averaging is a relatively unsophisticated way of smoothing and has some potential difficulties – not least of which is that the mean of each block is computed without reference to the rest of the series. In other words there is no ‘history’ built in to the mean of an individual block and so the ‘smoothed’ series can still exhibit some erratic jumps. To overcome this, we can take the basic ‘block’ or ‘window’ and step it across the

series so that there is overlap. This is achieved by replacing the 'oldest'  $k$  observations with the most recent  $k$  observations. The block averages in this case result in a *moving average* (MA) of the original series (Figure 14). A MA plot for the weekday inspection failure rate data is shown in Figure 15. The MINITAB help for this procedure is given in Box 3.

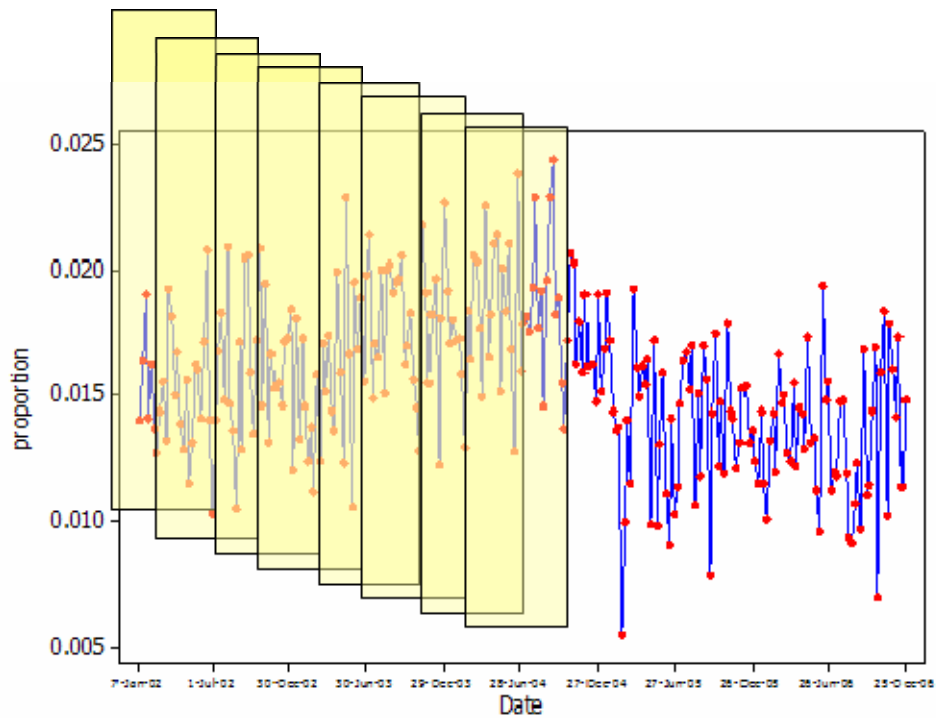


Figure 14. Moving average scheme. A block or 'window' is stepped incrementally over the series and the block mean computed and plotted.

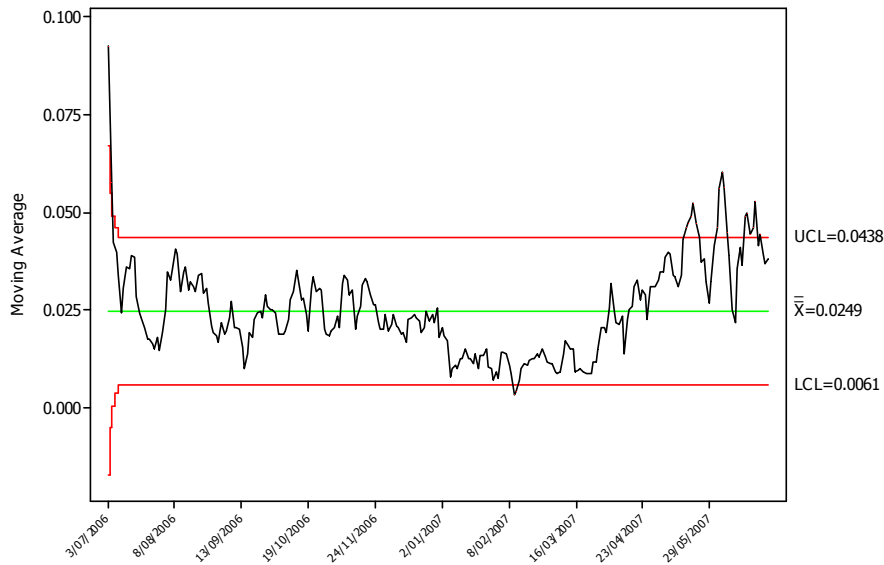



Figure 15. Moving average chart of weekday inspection failure rate. Sub-group size =1; MA length=5.

Box 3. MINITAB's Help on moving average chart

 **Moving Average Chart**  
[overview](#) [how to](#) [example](#) [data](#) [see also](#)

---

**Stat > Control Charts > Time-weighted Charts > Moving Average**

A **moving average chart** is a chart of **moving averages** – averages calculated from artificial **subgroups** that are created from consecutive observations. The observations can be either individual measurements or subgroup means. This chart is generally not preferred over an **EWMA chart** because it does not weight the observations as the EWMA does.

When data are in subgroups, the mean of all the observations in each subgroup is calculated. Moving averages are then formed from these means. By default, the process standard deviation,  $\sigma$ , is estimated using a pooled standard deviation. You can also base the estimate on the average of subgroup ranges or subgroup standard deviations, or enter a historical value for  $\sigma$ .

When you have individual observations, moving averages are formed from the individual observations. By default,  $\sigma$  is estimated  $\sigma$ , with  $\overline{MR} / d_2$ , the average of the moving range divided by an **unbiasing constant**. Moving ranges are artificial subgroups created from consecutive measurements. The moving range is of length 2, since consecutive values have the greatest chance of being alike. You can also estimate  $\sigma$  using the median of the moving range, change the length of the moving range, or enter a historical value for  $\sigma$ .

For more information, see [Control Charts Overview](#) and [Time-Weighted Control Charts Overview](#).

**Dialog box items**

**All observations for a chart are in one column:** Choose if data are in one or more columns, then enter the columns.

**Subgroup sizes:** Enter a number or a column of **subscripts**. If the subgroups are not equal, each control limit is not a single straight line but varies with the subgroup size. If the subgroup sizes do not vary much, you may want to force the control limits to be constant by specifying a fixed subgroup size using **MA Option > Estimate**.

**Observations for a subgroup are in one row of columns:** Choose if subgroups are arranged in rows across several columns, then enter the columns.

**Length of MA:** Enter the length of the moving averages. The value you enter is the number of subgroup means to be included in each average. If you have individual observations (that is, you specified a subgroup size of 1), Minitab uses them in place of the subgroup means in all calculations.

By adjusting the parameters of the moving average plot, different levels of smoothing can be achieved. For example, Figure 16 shows a MA plot for the proportion data using the averages of 3 weekly sub-groups. With this degree of smoothing, the three periods are clearer.

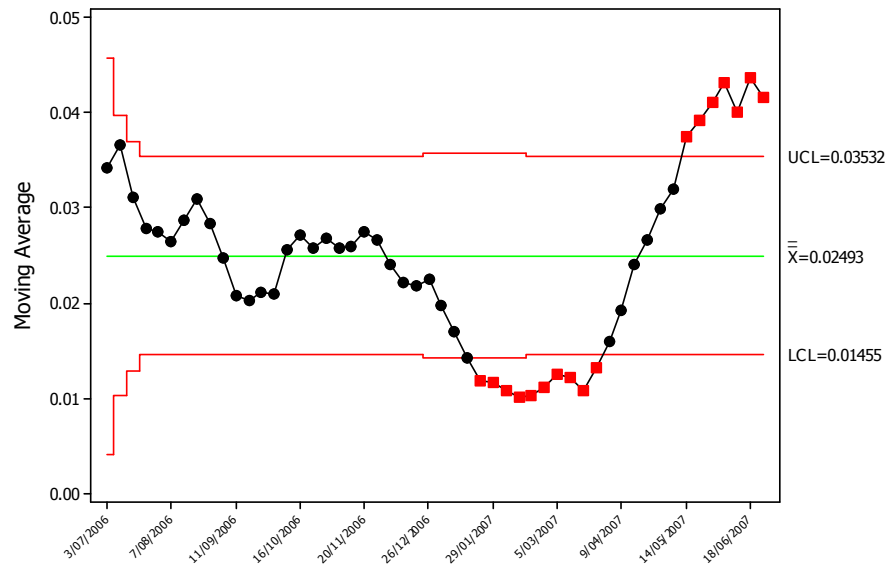


Figure 16. Moving average for weekday inspection failure rate. Subgroup size defined by week of year (usually 5); MA length=4.

### Time-weighted charts

The block or moving average charts just discussed give equal importance or weighting to all data in the current ‘window’. While this may be appropriate in some situations, it doesn’t accord with the usual notion that the greater the time separation, the less influential the data becomes. Time-weighted control charts such as the *Exponentially Weighted Moving Average* (EWMA) are more flexible in that the relative weightings given to recent and historical data can be specified.

By way of example, suppose we wish to form a weighted average of the current observation and the  $(k-1)$  most recent values. That is, we’re interested in forming the *weighted mean*  $\bar{X}_1 = \alpha X_1 + \alpha^2 X_2 + \dots + \alpha^k X_k$  where the weighting factor is  $\alpha$ . The requirement that  $\bar{X}_1$  is an unbiased estimator of the true mean imposes the constraint

$\sum_{i=1}^k \alpha^i = 1$  and for a given  $k$  the solution to this is the root of the equation  $\alpha^k - 2\alpha + 1 = 0$ .

For example, if  $k=10$ , we find  $\alpha = 0.5002$ . A plot of these weights compared to the simple arithmetic mean is shown in Figure 17.

The recursive formula for computing values of the EWMA chart is

$$EWMA_t = \alpha X_t + (1 - \alpha) EWMA_{t-1} \quad 0 < \alpha < 1$$

In other words, the current EWMA is a weighted average of the current data value and the EWMA in the preceding period. Figure 18 shows the EWMA chart for the weekday failure rate data with  $\alpha = 0.2$ .

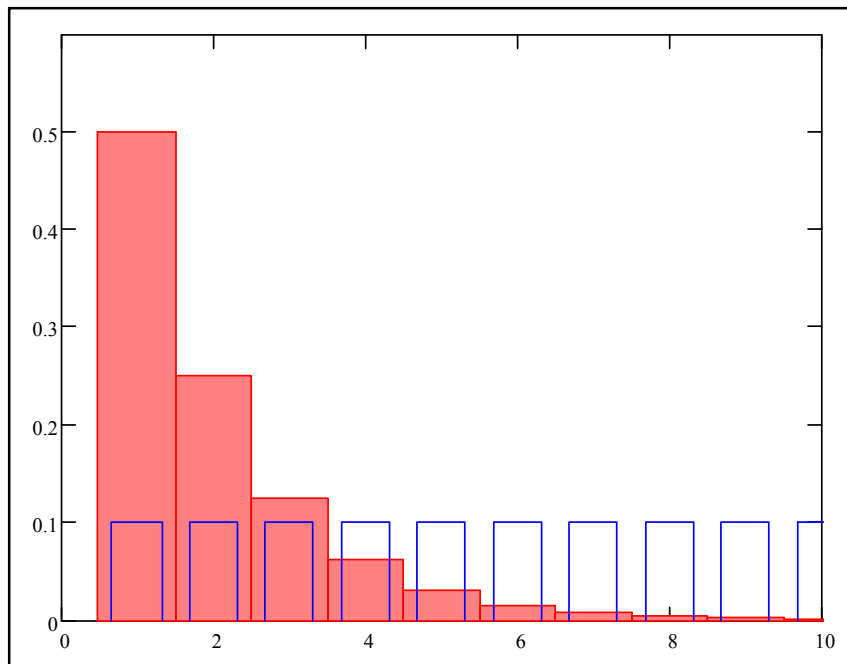


Figure 17. Comparison of exponentially declining weights (red bars) compared with equal-weighting scheme (blue rectangles) for  $k=10$ .

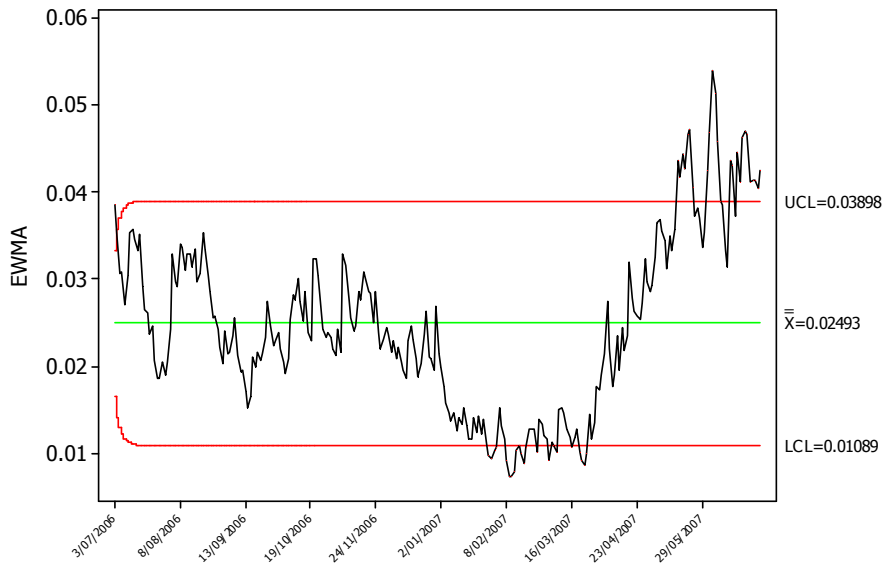


Figure 18. EWMA chart for weekday inspection failure rate. Subgroup size defined by week of year (usually 5); EWMA weight=0.2.

## 1-6 Time between events

We have seen how control charts can be used to monitor *variables* (such as the *number* of quarantine threats detected) or *attributes* (eg. the *proportion* of containers having a quarantine threat). When ‘events’ (eg. the ‘arrival’ of a quarantine risk at a port of entry) occur randomly in time, an alternative approach is to monitor the *inter-arrival time*. One advantage of this approach is that the inter-arrival time is available at the time of the arrival whereas if we are tracking the number of arrivals then these need to be aggregated over some time period before a meaningful analysis can be performed. However, some modifications to the standard charts are required to accommodate the fact that the distribution of *inter-arrival* times is usually (highly) non-normal. Details of the theoretical development can be found in Radaelli (1998). More recently, control charts for the number of ‘cases’ between events (so-called ‘g’ and ‘h’ charts) have been developed and applied to monitoring hospital-acquired infections and other relatively rare adverse health-related events (Benneyan 2001a, 2001b).

By way of example, consider Figure 19 which depicts the ‘arrival’ of a quarantine threat over time. Figure 20 shows a more detailed ‘slice’ through this pattern of arrivals. By

measuring the 'white-spaces' (i.e. computing the differences  $t_{i+1} - t_i$ ) in Figure 20 we obtain data on the inter-arrival times. The complete listing of data for this example is given in Appendix A.

Our analysis of the inter-arrival data commences with an inspection of basic distributional properties (Figure 21).

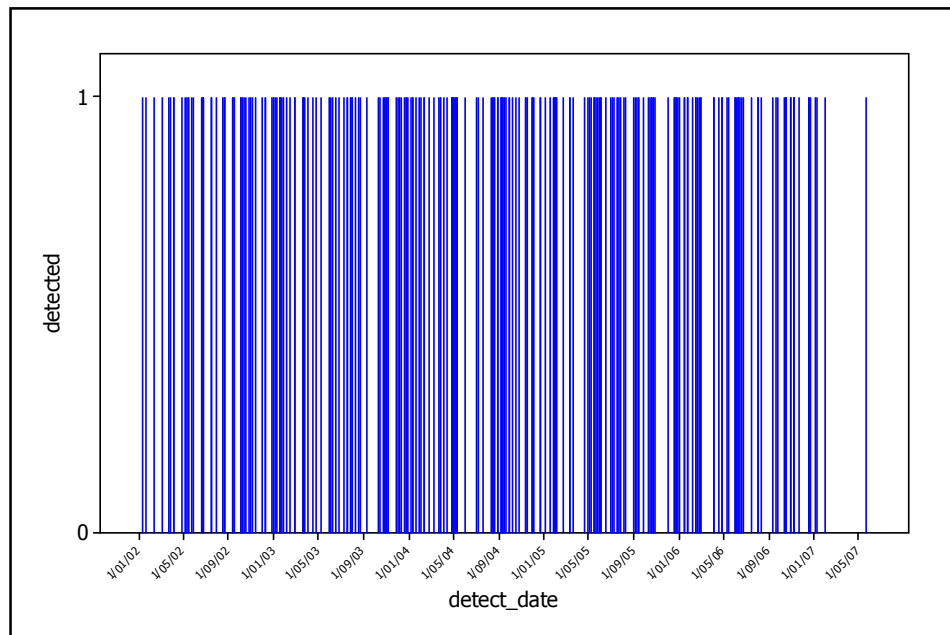


Figure 19. Time sequence of detection of quarantine threats.

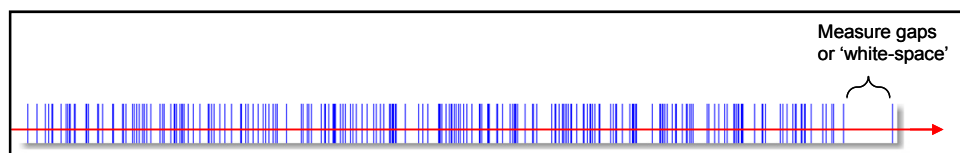


Figure 20. Pattern of inter-arrival times as measured by the 'white space' between blue lines.

It is clear from Figure 21 that these data are highly skewed and that the normal distribution is not an appropriate probability model.

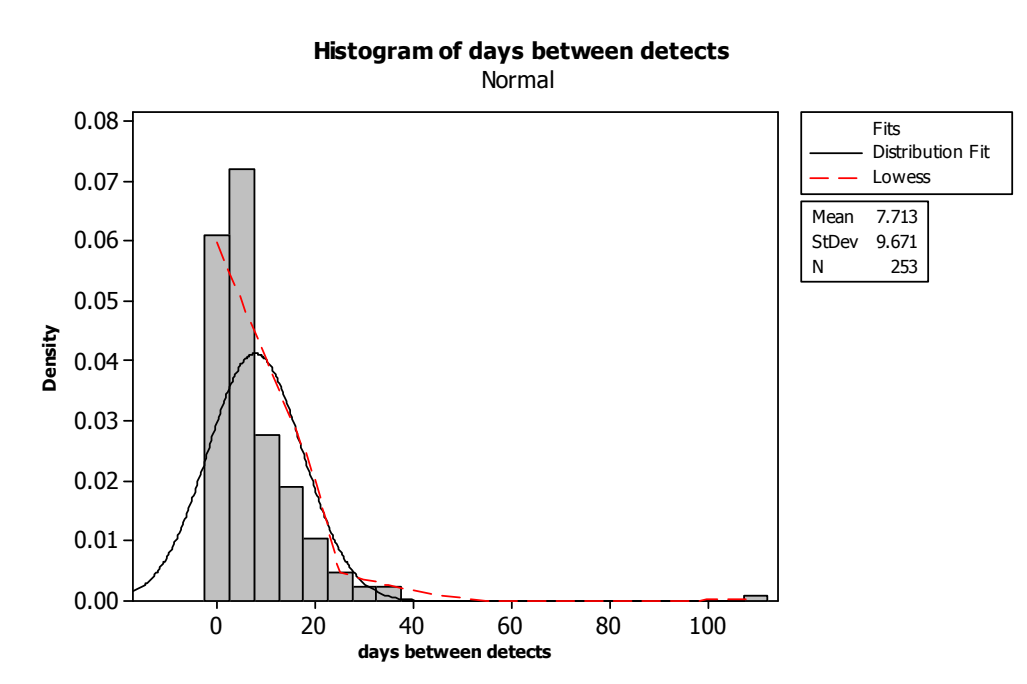


Figure 21. Histogram of inter-arrival times with smoothed version (red line) and theoretical normal distribution (black line) overlaid. The normal distribution provides a poor description of this data (evidenced by the both the shape and probability mass associated with negative values of days between detects).

The smoothed histogram in Figure 21 suggests a ‘J-shaped’ probability model is more appropriate. One such model is the *negative exponential* probability distribution. This choice is also supported by statistical theory which says that if events arrive randomly in time according to a *Poisson* probability model with an average rate of arrival of  $\lambda$  per unit time, then the distribution of the inter-arrival time is negative exponential with parameter  $\lambda$ . The probability density function (*pdf*) for the negative exponential is given by Equation 1.1 and the corresponding cumulative distribution function (*cdf*) is given by Equation 1.2.

$$f_X(x) = \lambda e^{-\lambda x}, \quad \lambda, x > 0 \tag{1.1}$$

$$F_X(x) = 1 - e^{-\lambda x}, \quad \lambda, x > 0 \tag{1.2}$$

For this distribution, the mean is  $\frac{1}{\lambda}$  and the variance is  $\frac{1}{\lambda^2}$ . Notice, that this immediately implies that the variance increases/decreases with an increasing /decreasing mean – in contradiction to many ‘conventional’ statistical techniques which assume constant variance.



One simple way of estimating the parameter  $\lambda$  is to equate the theoretical and sample means. In this case, we have  $\bar{y}_\lambda = 7.713$  and hence our estimate is  $\hat{\lambda} = 1/7.713 = 0.130$ . A plot of the histogram of the data with a negative exponential distribution having  $\hat{\lambda} = 0.130$  overlaid is shown in Figure 22. The adequacy of this fit is more readily seen by comparing the empirical and theoretical *cumulative distribution functions* (Figure 23).

The ‘false-triggering’ due to the non-normality of the data is evident in the I-Chart<sup>3</sup> of Figure 24. There are two ways of over-coming this. The first is to modify the control chart itself to account for the non-normality. The second approach is to *transform* the data so that the transformed data are normally distributed (or approximately so) and then apply standard control charting techniques to the transformed data. We consider each of these approaches in turn.

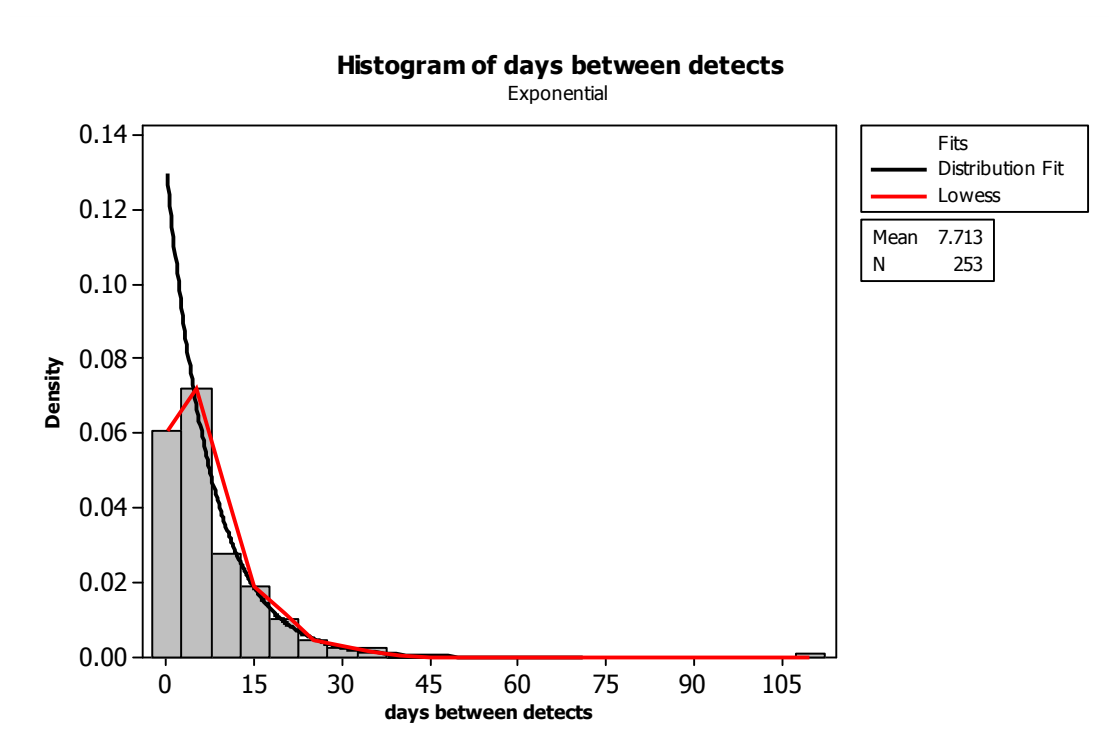


Figure 22. Histogram of days between detects. Smoothed histogram indicated by red curve, theoretical exponential distribution depicted by black curve.

<sup>3</sup> An “I-Chart” is simply a control chart for *individual* observations i.e. ungrouped data.

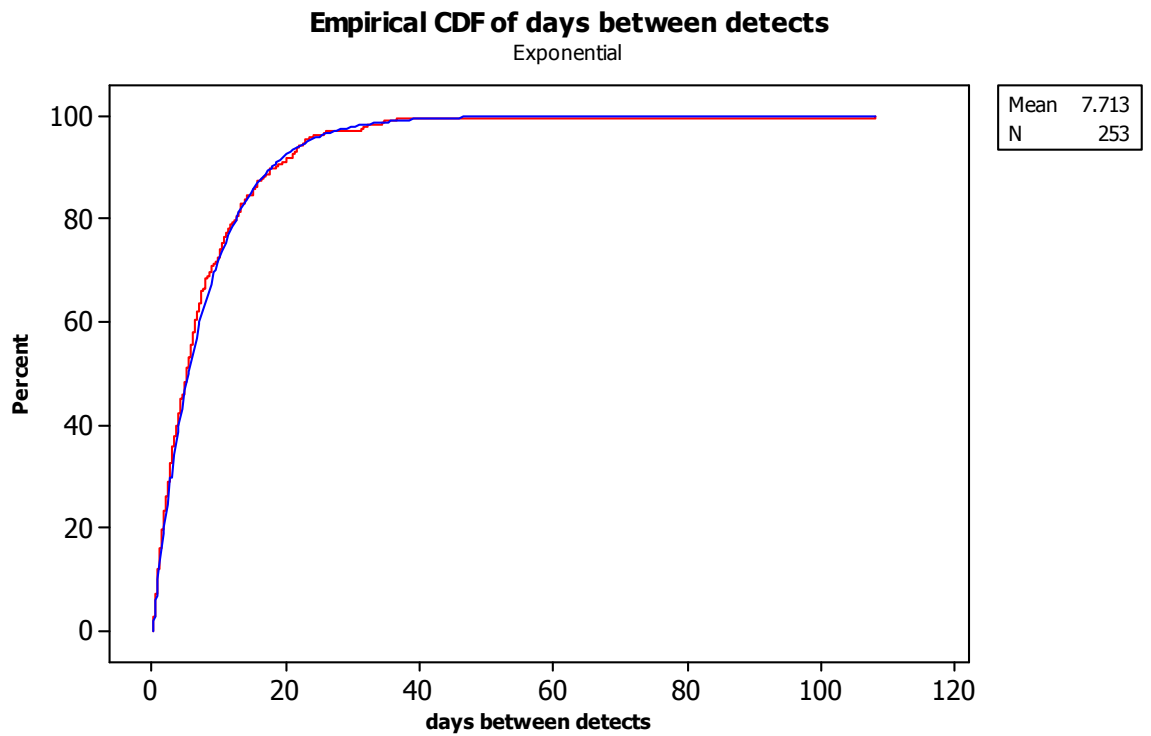


Figure 23. Empirical *cdf* for days between detects (red curve) and theoretical exponential *cdf*(blue curve).

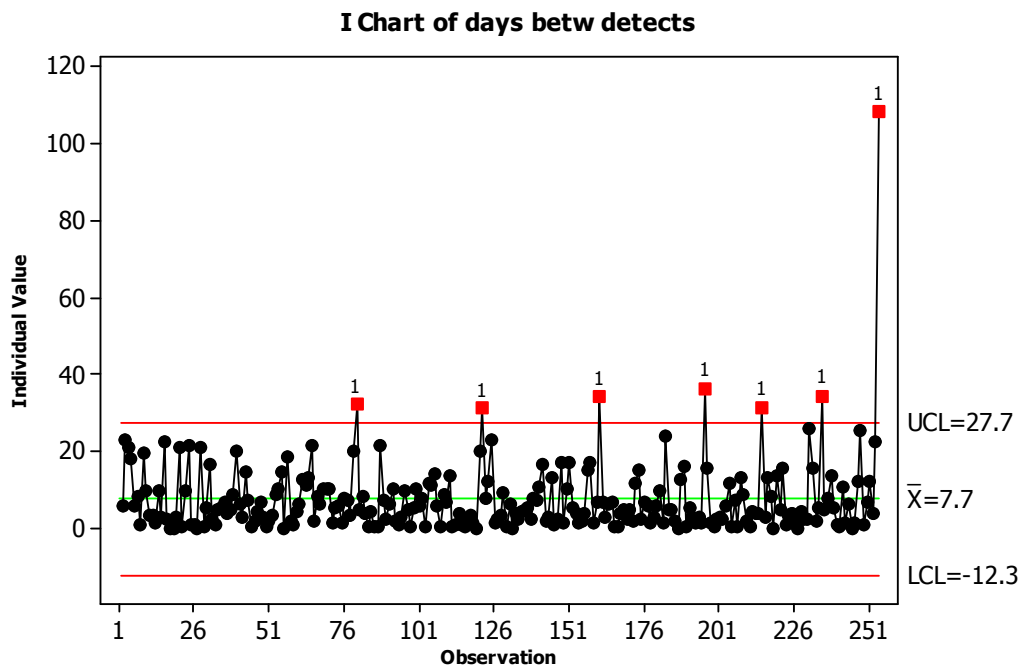


Figure 24. Chart of individual values of days between detects (I-Chart).

## 1-7 Transformations to Normality

In section 1-3 we looked briefly at the issue of statistical significance. The discussion focussed on a normally distributed random variable. Violations of the normality assumption will tend to invalidate the results of any statistical procedure which invokes this assumption<sup>4</sup>. One way to overcome this problem is to identify a mathematical transformation of the data so that the transformed data are normally distributed or approximately so. There is a tendency among practitioners to spend an inordinate amount of time on the identification of the ‘best’ transformation (best in the sense that the resulting data are most nearly normal). This is often wasted effort since many statistical procedures (including control charting) are relatively robust to mild to moderate departures from normality. The over-riding objective should be to identify a *simple* mathematical transformation that at least results in data that is approximately symmetrical. The identification process can be by trial and error or by some ‘automated’ procedure. An example of the latter is the so-called *Box-Cox* family of transformations. The idea is simple enough: we wish to find the value of the transformation parameter  $\lambda$  (not to be confused with the  $\lambda$  in equations 1.1 and 1.2) so that data transformed according to

$$Y = \begin{cases} \frac{X^\lambda}{\lambda} & \lambda \neq 0 \\ \ln(X) & \lambda = 0 \end{cases}$$

exhibit a greater degree of normality than the untransformed data (the  $X$ s). Having found this  $\lambda$  we proceed to work with the transformed values,  $Y$ . MINITAB and other software packages simplify the task of determining  $\lambda$  for a given data set. The ‘optimal’  $\lambda$  is identified as the abscissa value at the minimum on a Box-Cox ‘profile plot’ (Figure 25) – in this case we find  $\lambda = 0.24$ . With this value of  $\lambda$  we then transform the data and then use control charting methods on the *transformed* data.

---

<sup>4</sup> The severity of the violation cannot be anticipated in advance since it is a function of the degree to which the assumption is violated, the manner in which it is violated, and the robustness of the statistical procedure to such violations.

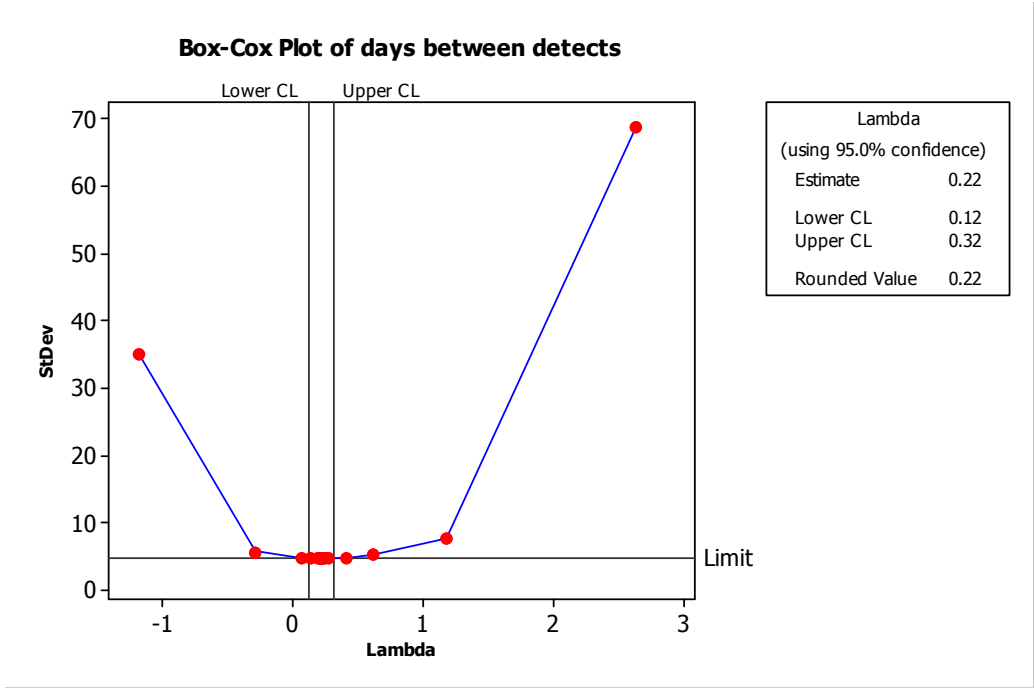


Figure 25. Box-Cox profile plot for the days between detects. Optimal lambda is 0.24.

The effectiveness of the transformation is evident from the histogram of the transformed data (Figure 26).

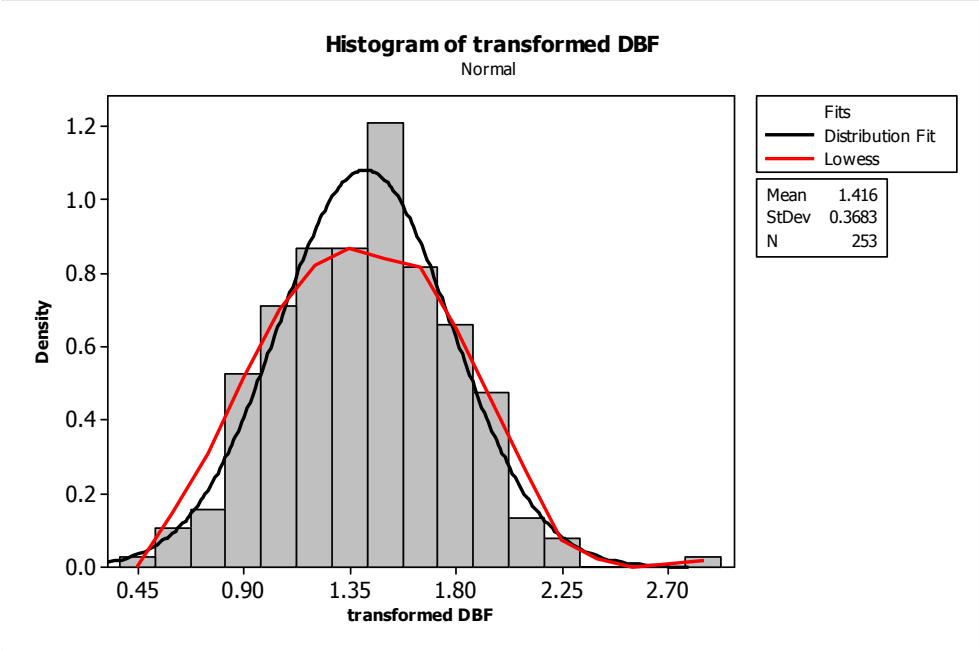


Figure 26. Histogram of transformed days between detection with smoothed version (red curve) and theoretical normal (black curve) overlaid.

To see how the transformation process works, consider setting an upper control limit on the days between detects such that this limit would only be exceeded 10% of the time when there has been no change in the underlying response-generating mechanism. In other words, on the *transformed scale*, we wish to find that value  $y^*$  which satisfies the following  $P[Y > y^*] = 0.10$ . From Figure 26, we see that the transformed data ( $Y$ ) are well described by a normal distribution having mean 1.416 and standard deviation 0.3683. Either using tables of the normal distribution or computer software (as in Figure 27) we determine that  $y^* = 1.89$ . We can ‘back-transform’ this  $y^*$  to determine an equivalent  $x^*$  on the *untransformed scale* by noting that

$$P[Y > y^*] = P[X^\lambda > y^*] = P[X > y^{*\frac{1}{\lambda}}]$$

That is,  $x^* = y^{*\frac{1}{\lambda}}$ . With  $y^* = 1.89$  and  $\lambda = 0.22$  we obtain  $x^* = 18.1$  which compares favourably to the theoretical result of 17.8 obtained directly from the negative exponential distribution (Figure 28).

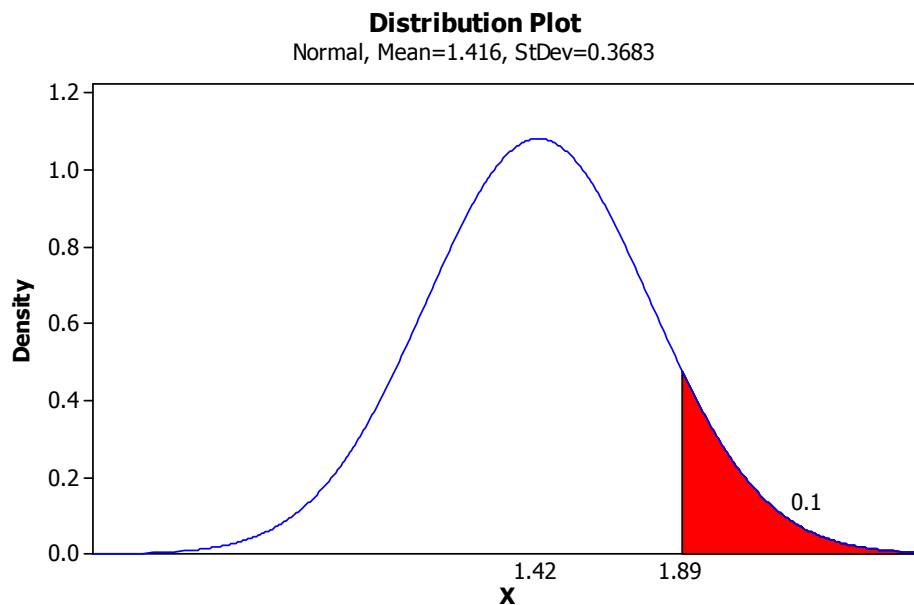


Figure 27. Fitted normal distribution to transformed days between detects with upper 10% point indicated.

Finally, we plot the I-Chart for the transformed days between detects and note that there is now only one 'out-of-control' situation indicated (Figure 29) compared with the previous seven (Figure 24).

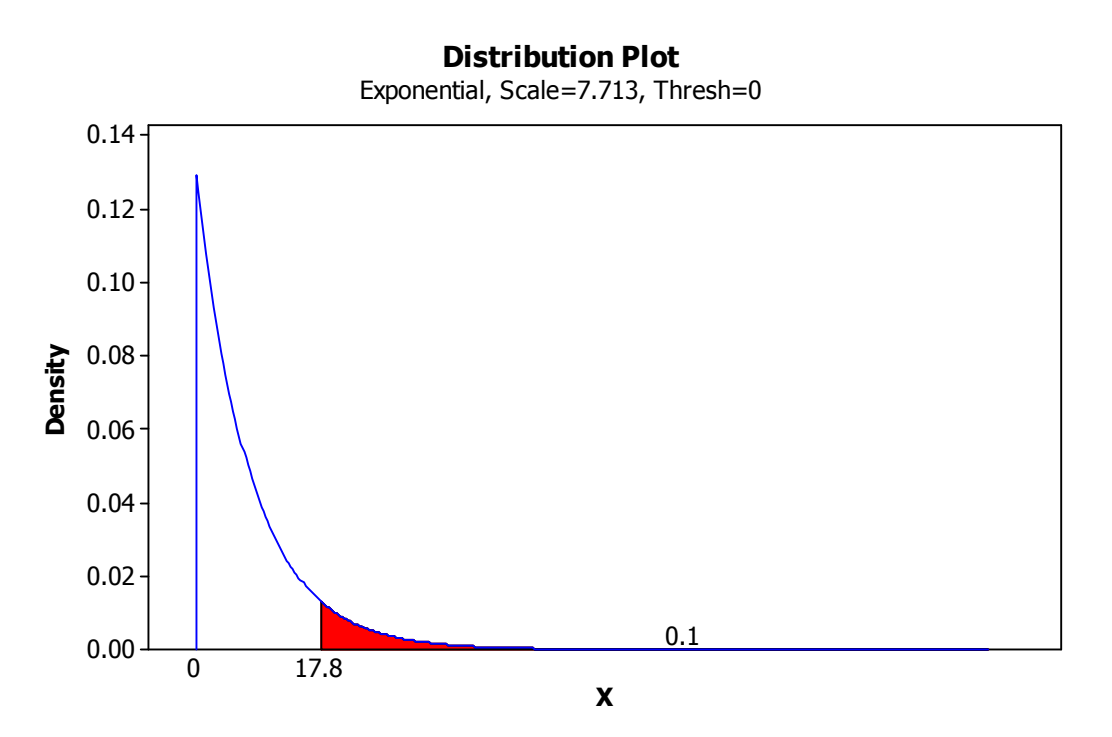


Figure 28. Theoretical negative exponential distribution for untransformed days between detects and upper 10% point indicated.

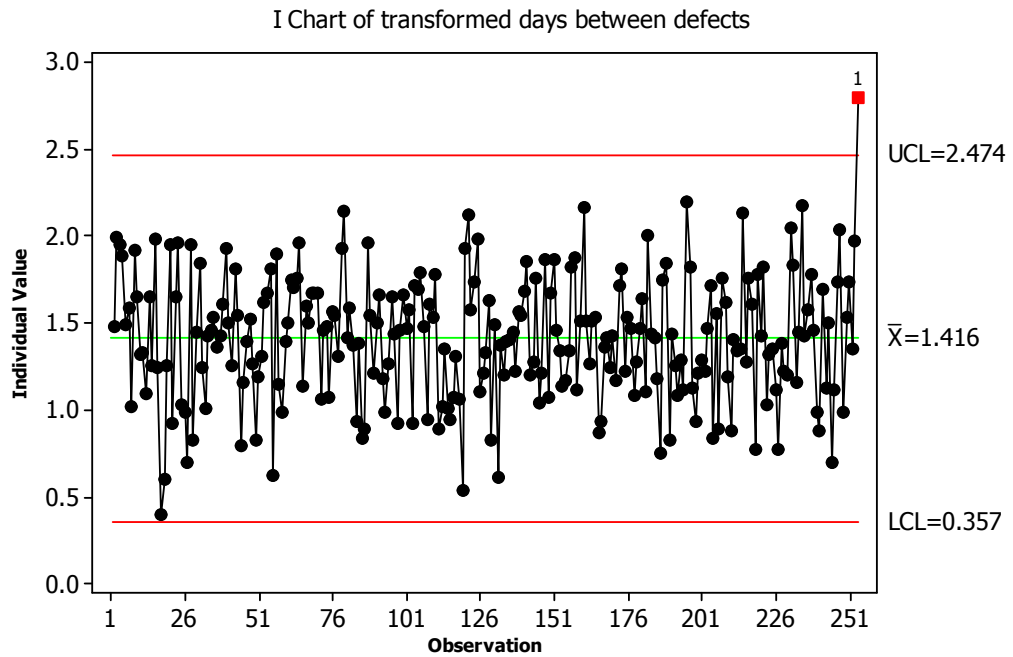


Figure 29. I-Chart for transformed days between detects.

## 1-8 Control chart for time-between-events

Rather than transform the data as described in the preceding section, alternative methods have been developed which modify existing control charts for use with untransformed (and non-normal) data. Radaelli (1998) describes procedures for setting control limits for both one and two-sided control charts for inter-arrival times. Only the one-sided case is considered here since we are generally only interested in tracking significant deviations in one direction (eg. where the inter-arrival time between quarantine risk detects is decreasing).

Let  $X_i$  be the  $i^{\text{th}}$  inter-arrival time. An ‘out-of-control’ situation is declared if  $X_i < T_L$  in the case of *decreasing* inter-arrival times (ie. increasing counts) or  $X_i > T_U$  in the case of *increasing* inter-arrival times (ie. decreasing counts) where  $T_L$  and  $T_U$  are suitably chosen positive constants. Suppose that an ‘in-control’ situation corresponds to a mean inter-arrival time of  $\lambda_0^{-1}$  (where  $\lambda$  is the parameter in equation 1) then using Equation 1.2, it can be determined that

$$P[X_i < T_L | \lambda = \lambda_0] = 1 - e^{-\lambda_0 T_L} \quad (1.3)$$

$$P[X_i > T_U | \lambda = \lambda_0] = e^{-\lambda_0 T_U} \quad (1.4)$$

Equations 1.3 and 1.4 are analogous to the Type I error in a hypothesis test: it's the probability of a false-positive. As in statistical hypothesis testing, the Type I error-rate ( $\alpha$ ) is set to be some arbitrarily small value (eg.  $\alpha = 0.05$ ). Thus, the upper and lower control limits can be determined by setting Equations 3 and 4 equal to  $\alpha$  and solving for either  $T_L$  or  $T_U$ . Thus we have:

$$T_L = -\lambda_0^{-1} \ln(1 - \alpha) \quad (1.5)$$

$$T_U = -\lambda_0^{-1} \ln(\alpha) \quad (1.6)$$

In addition to having a low  $\alpha$ , we also require our control chart to correctly signal an important deviation from 'in-control' conditions. Suppose we wish to detect a change from  $\lambda_0$  to  $\lambda_1$  with some high probability,  $(1 - \beta)$  where  $\lambda_1 = k\lambda_0$  ( $k > 1$  for a one-sided lower chart;  $k < 1$  for a one-sided upper chart). That is:

$$P[X_i < T_L | \lambda = \lambda_1] = 1 - e^{-k\lambda_0 T_L} = (1 - \beta) \quad (\text{lower chart}) \quad (1.7)$$

$$P[X_i > T_U | \lambda = \lambda_1] = e^{-k\lambda_0 T_U} = (1 - \beta) \quad (\text{upper chart}) \quad (1.8)$$

Substituting  $T_L$  and  $T_U$  in Equations 7 and 8 respectively, we obtain:

$$(1 - \beta) = 1 - e^{k \ln(1 - \alpha)} \quad (\text{lower chart}) \quad (1.9)$$

$$(1 - \beta) = e^{k \ln(\alpha)} \quad (\text{upper chart}) \quad (1.10)$$

The performance characteristics for both lower and upper one-sided charts are shown in Figures 30 and 31. Both of these figures show that the ability to detect even relatively large shifts (eg. a doubling or halving) in the mean inter-arrival time is low (typically less than 0.2) for values of  $\alpha$  less than 0.1. For example, using a 10% level of



significance (ie.  $\alpha = 0.10$ ), the one-sided chart of Figure 30 suggests that there is a less than 20% chance of detecting a doubling (ie.  $k=2$ ) of the mean arrival rate (or a halving of the inter-arrival time).

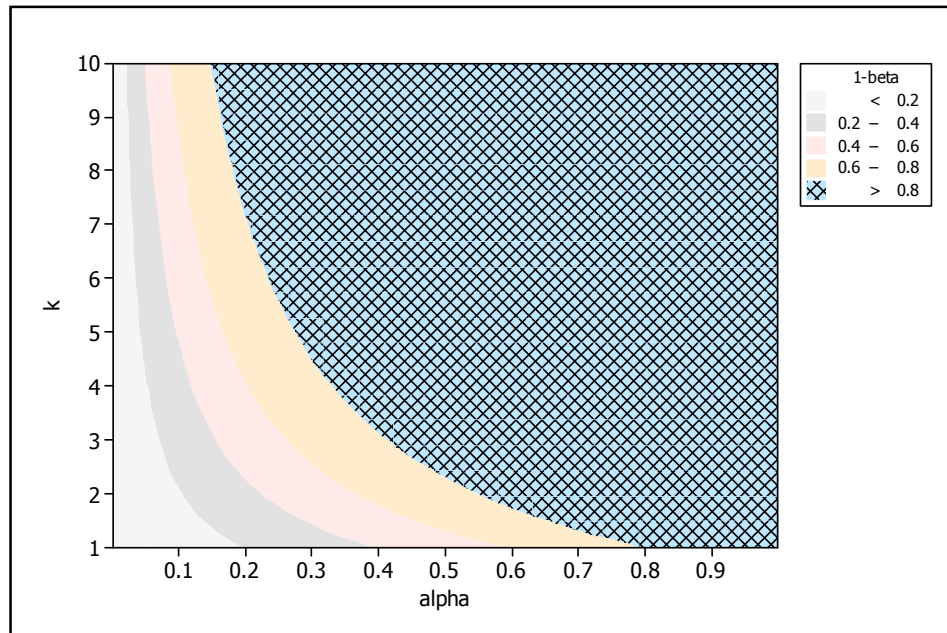


Figure 30. Performance characteristics (as measured by equation 1.9) for a one-sided, lower control chart.

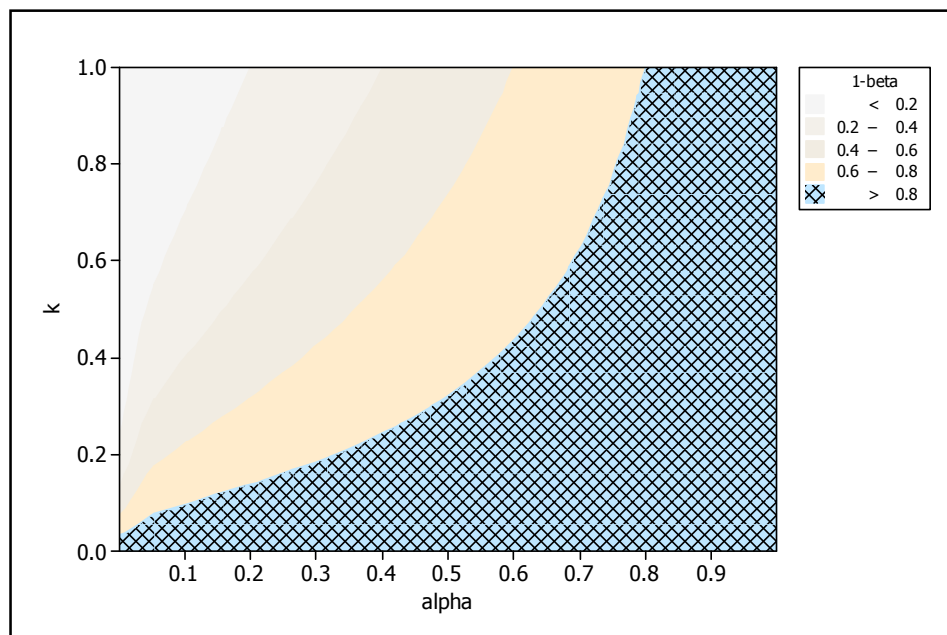


Figure 31. Performance characteristics (as measured by equation 1.10) for a one-sided, upper control chart.

## 1-9 DISCUSSION

In this chapter we have provided a review of basic statistical concepts as well as introducing some common control-charting techniques that have been advocated elsewhere (Carpenter 2001, Commonwealth of Australia 2002) as being particularly suited to monitoring for temporal trends and aberrations in bio-security related applications. Control charts are particularly well suited to the visualisation and assessing of moderate to large volumes of time-based data and as such would be expected to have greater utility for container inspection regimes say, than for detecting the occurrence (in space) of an invasive species. Control charts need to be viewed as just one method in a tool-kit of available techniques which can potentially assist field officers and quarantine risk assessors in identifying ‘unusual’ or ‘aberrant’ trends. For events having very low probabilities of occurrence (eg. exotic disease outbreak) the monitoring of ‘time between outbreaks’ is a potentially more useful quantity to be charting although as shown in this report, the statistical power (ability to correctly identify real ‘shifts’ in the mean time between events) of current charting techniques is relatively low.

Since the events of September 2001, there has been a substantial research push in the area of ‘syndromic surveillance’ with the accompanying development of new approaches and methods to detecting unusual patterns in a space-time continuum. Some of these techniques would appear to have direct applicability to the activities of Biosecurity Australia and AQIS.



## 2-1 INTRODUCTION

In this chapter we investigate some more specific aspects of control charting and in particular, focus on the use of Bayesian statistical methods. This work advances conventional control charting methods described in chapter 1 by adaptively updating the alerting mechanisms as well as explicitly incorporating prior belief about the state of monitored ‘system’. The methodology is developed in the context of routine quarantine inspection of imported foods although the potential applications extend to other areas of bio-surveillance where data is being gathered over time and ‘early warning’ triggers are required.

Detailed background information on current quarantine inspection practice can be found in the report by Robinson et al. (2008). Robinson et al. (2008) also provide details of a suggested risk-based framework for allocating scarce resources to the monitoring and surveillance effort.

The present study is concerned with the *implementation* phase of a specific risk-based approach to monitoring. Most quarantine surveillance and monitoring programs are candidates for a statistical approach since they invariably involve small sampling fractions and there overarching requirement to balance the cost of sampling with the probability of failing to detect a threat. Our focus in this chapter is on the *temporal* component of monitoring – that is, detecting important ‘shifts’ or ‘aberrations’ in monitored data in close to real time. Chapters three and four examine statistical methods associated with the *spatial* dimension of bio-surveillance.

A common requirement of statistical surveillance techniques is to detect important changes in a stochastic process at an unknown time, as quickly and as accurately as possible (Sonesson and Bock 2003). Many of the reported techniques use likelihood based methods

to detect step changes in a parameter of interest (eg. process mean or variance). While a number of papers have appeared recently on statistical surveillance in the context of epidemiology, public health, and syndromic surveillance (Doherr and Audigé 2001, Sonesson and Bock 2003, Marshall et al. 2004, Höhle and Paul 2008) relatively little has been published on quarantine inspection.

The utility of conventional statistical process control (SPC) tools such as the Shewhart chart, EWMA chart, and other control charting variants for quarantine inspection was covered in chapter one of this report. Control charting methods have a number of attributes which make them particularly well-suited to the task of identifying ‘abnormal’ trends in the detection of non-compliant shipments, such as an ‘early-warning’ capability and easily communicated visual displays of historical results. While some of these tools (such as the EWMA chart) have the ability to couple past history and present observations, they are conventionally data-driven approaches that do not readily accommodate expert opinion or existing understanding about the underlying response-generating process. This is potentially an important consideration, particularly when a new commodity or product is shipped into the country for which historical data does not exist or in cases where other ancillary information concerning the commodity becomes available (for example increased susceptibility to contamination at the point of manufacture).

Another difficulty with standard control-charting tools is that they are constructed on models which assume process parameters are known exactly and observations are *i.i.d.* (Tsiamirtzis and Hawkins 2007). This is problematic since parameter values are rarely known and secondly, the assumption of *i.i.d.* data is frequently violated – particularly for time-series data which often exhibit moderate to strong autocorrelation. More recently, Bayesian control charting methods have been developed to help overcome some of these limitations. Baron (2001) used the theory of optimal stopping of Markov sequences to develop efficient algorithms for the detection of a distributional change in sequentially collected data while Hamada (2002) used Bayesian tolerance interval control limits in the context of attribute sampling. Our approach to Bayesian control charting for quarantine

---

<sup>5</sup> independently and identically distributed

inspection has been motivated by Menzefricke (2002) who used Bayesian predictive distributions to derive rejection regions for various monitoring applications.

It is not the aim of this report to provide a comprehensive review on Bayesian control charting techniques. However, to facilitate the subsequent mathematical development we digress momentarily to explain some of the underlying concepts. Readers requiring a more comprehensive treatment of the topic may find the collection of papers in Colosimo and del Castillo (2007) a useful entry point.

## 2-2 Fundamentals of Bayesian Control Charting

The techniques presented in chapter one fall within the realm of ‘frequentist’ statistics. This mode of statistical thinking is by far the most common and underpins nearly every undergraduate course in statistics. The frequentist view of the world is one in which the only admissible probabilities are those that are expressible as the ratio of the number of outcomes which are favourable to the ‘event’ under consideration to the total number of outcomes, or alternatively, can be thought of as the limiting value of the *relative frequency* of some phenomenon – hence the term frequentist statistics. A point of clear demarcation between frequentist and Bayesian statistics is the role of *subjective* probability. There is no role for subjective probability in a frequentist framework; for Bayesians it is pivotal.

The term *Bayesian* derives from the Rev. Thomas Bayes (b. 1702, London - d. 1761). Bayes was not known as a mathematician and his only significant work "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), was published posthumously in the *Philosophical Transactions of the Royal Society of London*. Although a largely turgid piece of work, Bayes’ *essay* identified a fundamental proposition in probability. This was a profound insight and provided a logical and consistent way of updating *prior* belief or probability in the light of new evidence. The formula was named Bayes rule or theorem after him. The updated probability is referred to as the *posterior* probability. Unlike frequentist statistical inference which tests hypotheses or estimates (unknown) parameters on the basis of information contained in data alone, the Bayesian paradigm combines prior belief about unknown parameters with evidence from data using Bayes’ rule. More formally, the aim of Bayesian inference is then to make inferences about a parameter  $\theta$  or future observation  $\tilde{y}$  using

probability statements *conditional* on the data  $y$ . Both parameters and future observations are treated as random variables in a Bayesian framework and we talk of the *posterior density* of  $\theta$  [denoted  $p(\theta|y)$ ] and the *posterior predictive density* of  $y$  [denoted  $p(\tilde{y}|y)$ ].

The simplest version of Bayes' rule for two 'events' A and B says that the conditional probability that event A occurs given event B has occurred is given by the formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where the probability in the numerator is the joint probability (i.e.

the probability that both A and B occur). Bayes' theorem applies equally to probability density functions (*pdf*). Thus if  $y$  denotes data and  $\theta$  some parameter or vector of

parameters, then  $P(y|\theta) = \frac{P(y \cap \theta)}{P(\theta)} = \frac{P(y, \theta)}{P(\theta)}$  and the numerator is the joint probability

density function for  $y$  and  $\theta$ . The roles of  $y$  and  $\theta$  can be interchanged in this formula and we have immediately that:

$$P(\theta|y) = \frac{P(\theta \cap y)}{P(y)} = \frac{P(y, \theta)}{P(y)}$$

and a comparison of  $P(y|\theta)$  and  $P(\theta|y)$  reveals that

$P(y, \theta) = P(\theta)P(y|\theta) = P(y)P(\theta|y)$ . Finally, substituting this last expression for  $P(y, \theta)$  into the expression for  $P(\theta|y)$  gives Bayes' law for densities:

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)}.$$

This formula takes a prior density for  $\theta$  [i.e.  $P(\theta)$ ] and converts it into a posterior density  $P(\theta|y)$  via the term  $\frac{P(y|\theta)}{P(y)}$  called the Bayes factor. The denominator in the expression for the posterior density does not involve  $\theta$  and only serves to normalise the *pdf* (i.e. make it integrate to unity). Inference for  $\theta$  based on the posterior is therefore unaffected by working with  $P(\theta|y) \propto P(\theta)P(y|\theta)$  instead of the normalised posterior. Thus we see that the posterior distribution is proportional to the product of the prior times the likelihood of the

data. In frequentist inference, only the likelihood is used; in Bayesian statistics the likelihood is modified by our prior belief.

In the remainder of this chapter we describe a new/novel Bayesian control charting approach to quarantine inspection. The motivation in the present context is that conventional (i.e. non-Bayesian or Frequentist) approaches to control charting need to be ‘primed’ with hard data in the absence of known parameter values. While this might not be an issue in a manufacturing context where production data is both plentiful and continuous, it is problematic for the monitoring of processes for which little background data is available. This problem becomes particularly acute for the development of a surveillance program aimed at detecting a new threat for which there is *no* prior data. Similarly, in the case of monitoring a rare phenomenon, a paucity of data is inevitable. In such cases the monitoring data will be comprised of a string of zeros – corresponding to ‘no detect’ outcomes. Frequentist statistical methods will thus estimate the *true* rate of occurrence as zero with a standard error of zero. The Bayesian paradigm on the other hand commences with the specification of a *prior density* for the parameter of interest (such as the true rate of occurrence) and continually updates this as new data becomes available. The method is illustrated with application to food import data used in the previous chapter (Robinson et al. 2008).

## **2-3 A BAYESIAN CONTROL CHART FOR QUARANTINE INSPECTION**

The following development assumes attribute sampling whereby at time  $t$ ,  $n_t$  ‘units’ are selected from a total volume of trade comprising  $N_t$  units and the result of inspection is a binary outcome: “pass” or “fail”. The time index  $t$  will generally represent daily increments. On each sampling occasion two related control-charting questions are considered: (i) is the observed failure rate for the current sample within acceptable limits?; and (ii) is the cumulative failure rate for all samples inspected to date within acceptable limits? These objectives mirror the detection of ‘pulse’ and ‘press’ stresses in natural ecosystem management (Underwood, 1994). In answering questions (i) and (ii) we wish to



incorporate both historical monitoring data and prior information on the true failure rate,  $\theta$ . We do this through the use of a conditional probability distribution for the data given  $\theta$  and a prior probability model for  $\theta$ . These two elements can be combined to obtain a predictive distribution for a new sample. The mathematical detail is developed in the following section. Readers not interested in this can skip forward to section 2-4.

### 2-3-1 Mathematical detail

The problem as formulated leads us to consider the random variable  $X_t$  - the number of failed units in the sample of  $n_t$  taken at time  $t$ . Assuming independent Bernoulli trials for each inspected unit, the conditional distribution of  $X_t | \theta$  is binomial (we have dropped the time subscript to improve clarity where it is understood that all results pertain to the current sample unless otherwise indicated):

$$f_{X|\theta}(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}; \quad x = \{0, 1, \dots, n\}, 0 < \theta < 1 \quad (2.1)$$

Uncertainty in the true failure rate  $\theta$  is reflected in the prior distribution  $p(\theta)$ . A suitable choice for  $p(\theta)$  is the beta density:

$$p(\theta; a, b) = \frac{1}{\beta(a, b)} \theta^{a-1} (1-\theta)^{b-1}; \quad 0 < \theta < 1, a > 0, b > 0 \quad (2.2)$$

Initial values for the  $a$  and  $b$  parameters in equation 2.2 can be chosen according to various strategies depending on how much or how little we know about the true rate of risk for a particular commodity, country, test etc. Robinson et al. (2008) discuss some of these strategies in the context of food imports and recommended the use of a Jeffrey's prior corresponding to  $a = 0.5$  and  $b = 0.5$ . A plot of this Jeffrey's prior and two 'vague' or 'non-informative' prior densities are shown in Figure 32.

### Updating the prior

The underlying principal in the adaptive monitoring process is that our estimate of the true failure rate is constantly revised as new data is gathered. In the early stages of monitoring, our probability model for the true failure rate will be driven by prior information.

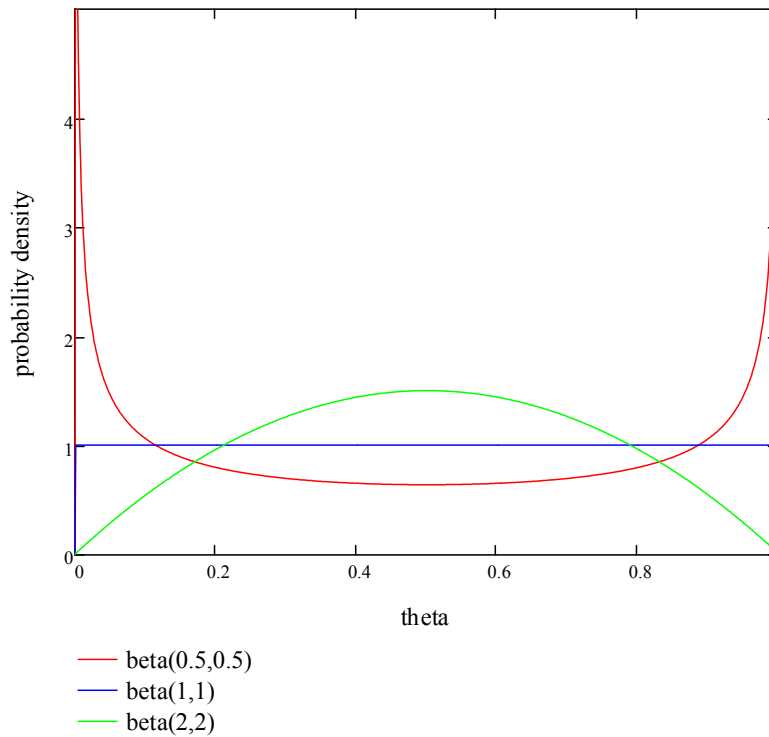


Figure 32. Illustrative non-informative priors for true failure rate,  $\theta$ .

At each time increment the current prior probability for  $\theta$  is updated using standard Bayesian methods to generate a posterior marginal *pdf*. The procedure is described below.

At time  $t$  we have available the history of observed failures up to and including the current observation – denote this as  $\{x_1, x_2, \dots, x_N\}$  where  $N = \sum_{i=1}^t n_i$  is the total number of sampled units. We let  $Y$  denote the total number of failures at time  $t$  i.e.  $Y = \sum_{i=1}^t X_i$ . For a

stable process, the distribution of  $Y$  is also binomial with parameters  $(N, \theta)$ . However  $N$  will rapidly become 'large' (i.e greater than  $\sim 30$ ) and provided  $\theta$  is not close to either zero or one the binomial distribution is well approximated by a Poisson distribution with mean  $N\theta$ .

Now the marginal posterior distribution for  $Y$  as a function of the parameters  $a$  and  $b$  is:

$$p(y|a, b) = \int_0^1 l(y|\theta) p(\theta|a, b) d\theta \quad (2.3)$$

where  $l(y|\theta)$  is the likelihood of the data ( $y$ ) given  $\theta$ . Thus, equation 2.3 can be written as:

$$p(y|a, b) = \int_0^1 \left[ \frac{e^{-N\theta} (N\theta)^y}{y!} \right] \frac{1}{\beta(a, b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \quad (2.4)$$

Equation 2.4 can be evaluated using numerical integration or alternatively computed using equation 2.5 (a derivation of equation 2.5 is provided in Appendix B).

$$p(y|a, b) = \frac{N^y}{y!} \prod_{j=0}^{y-1} \left( \frac{j+a}{j+a+b} \right) \left\{ 1 + \sum_{m=1}^{\infty} \left[ \prod_{r=0}^{m-1} \frac{y+a+r}{y+a+b+r} \right] \frac{(-N)^m}{m!} \right\}; \quad (2.5)$$

$$y = \{0, 1, \dots, N\}, \quad a, b > 0$$

At time  $t$  we have available the data  $\{(N_1, Y_1), (N_2, Y_2), \dots, (N_t, Y_t)\}$ . The likelihood is thus

$$l(a, b; y) = \prod_{i=1}^t p(X_i = y_i - y_{i-1} | a, b); \quad y_0 = 0 \quad (2.6)$$

The maximum likelihood estimates,  $\hat{a}$  and  $\hat{b}$  at time  $t$  are found by simultaneously solving equations 2.7a and 2.7b.

$$\hat{a} = \left\{ a : \frac{\partial l(a, b; y)}{\partial a} \Big|_{\hat{a}} = 0 \right\} \quad (2.7a)$$

$$\hat{b} = \left\{ b : \frac{\partial l(a, b; y)}{\partial b} \Big|_{\hat{b}} = 0 \right\} \quad (2.7b)$$

The updated distribution for  $\theta$  at time  $t$  is equation 2.2 with parameters  $\hat{a}$  and  $\hat{b}$ . We use this posterior to obtain the predictive distributions for the number of failures ( $X_{t+1}$ ) in the next sample of units to be inspected and the cumulative number of failures ( $Y_{t+1}$ ).

#### **Predictive distributions for $X_{t+1}$ and $Y_{t+1}$**

The predictive distribution for  $X_{t+1}$  can be written as

$$p[X_{t+1} | y_t] = \int_0^1 f(x_{t+1} | \theta) p(\theta | y) d\theta \quad (2.8)$$

where  $p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)}$ ;  $p(y) = \int_0^1 p(y | \theta) p(\theta) d\theta$  and  $p(\theta)$  based on the most

recent estimate using equation 2.2 with parameters  $\{\hat{a}, \hat{b}\}$ .

It can be shown (Appendix C) that  $p[X_{t+1} | y_t]$  is given by equation 2.9.

$$p[X_{t+1}|y_t] = \binom{n_{t+1}}{x_{t+1}} \frac{\beta(x_{t+1} + y_t + a, n_{t+1} - x_{t+1} + b)}{\beta(y_t + a, b)} \frac{\left\{ 1 + \sum_{m=1}^{\infty} \left[ \prod_{r=0}^{m-1} \frac{x_{t+1} + y_t + a + r}{n_{t+1} + y_t + a + b + r} \right] \frac{(-N_t)^m}{m!} \right\}}{\left\{ 1 + \sum_{m=1}^{\infty} \left[ \prod_{r=0}^{m-1} \frac{y_t + a + r}{y_t + a + b + r} \right] \frac{(-N_t)^m}{m!} \right\}} \quad (2.9)$$

We next consider the predictive distribution for  $Y_{t+1}$ . First, it can be seen that

$p(Y_{t+1} = s | Y_t = y) = p(X_{t+1} = s - y)$ . The unconditional distribution of  $X_{t+1}$ ,  $p(X_{t+1})$  is obtained as follows:

$$p(X_{t+1}) = \int_0^1 p(X_{t+1} | \theta) p(\theta) d\theta$$

where  $p(X_{t+1} | \theta) \stackrel{d}{\sim} \text{bin}(n_{t+1}, \theta)$  and  $p(\theta) \stackrel{d}{\sim} \text{beta}(a, b)$ . Thus,  $p(X_{t+1})$  is a beta-binomial distribution and the predictive distribution for  $Y_{t+1}$  is therefore:

$$p(Y_{t+1} = s | Y_t = y) = \binom{n_{t+1}}{s - y} \frac{\beta(s - y + a, n_{t+1} - s + y + b)}{\beta(a, b)} \quad (2.10)$$

We next discuss how the predictive distributions are used to set control limits for routine inspection programs.

### 2-3-2 Adaptive control limits

The idea of a control limit is to provide an early warning that the underlying response generating mechanism has departed from an assumed stable state. In the present context we wish to set two limits (designated as  $RL_1$  and  $RL_2$ ) on the number of failures in the next batch of sampled units.  $RL_1$  is set such that, when exceeded, it draws our attention to the fact that there is a higher number of failures in that particular sample of  $n$  units than would be expected. Exceedence of  $RL_2$  signifies an unusually high number of failures which would significantly increase the *cumulative* failure rate. Clearly the two limits are related since

triggering of  $RL_2$  implies a triggering of  $RL_1$ , although the converse is not necessarily true. Thus, the second limit will tend to be more liberal than the first.

$(1 - \alpha)100\%$  response levels  $RL_1$  and  $RL_2$  are obtained by solving equations 2.11 and 2.12 respectively.

$$RL_1(\alpha; n_{t+1}, y) = \left\{ r : \sum_{x=r}^{n_{t+1}} p(X_{t+1} = r | Y_t = y) \geq \alpha \wedge \sum_{x=r-1}^{n_{t+1}} p(X_{t+1} = r | Y_t = y) < \alpha \right\} \quad (2.11)$$

$$RL_2(\alpha; n_{t+1}, y) = \left\{ r : \sum_{x=r}^{n_{t+1}} p(Y_{t+1} = r | Y_t = y) \geq \alpha \wedge \sum_{x=r-1}^{n_{t+1}} p(Y_{t+1} = r | Y_t = y) < \alpha \right\} \quad (2.12)$$

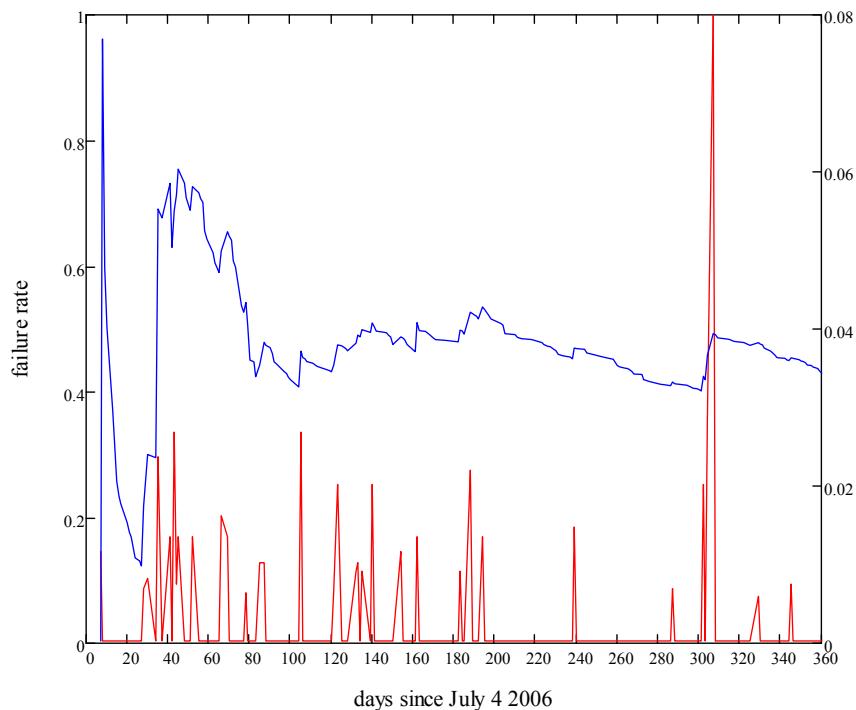
These ideas are illustrated with an example using the AQIS food import data set (*see Chapter 1*) for the period 1-Jul-2006 to 30-Jun-2007. Detailed information about the data collection methods can be found in Robinson et al. (2008).

## 2-4 EXAMPLE – AN ADAPTIVE CONTROL CHART FOR FOOD IMPORTS

Between 4/7/2006 and 29/6/2007 a total of 1,718 items were imported from a particular country. These were predominantly food items such as soy sauce, instant noodles, fish, pasta, and crabmeat. There are generally multiple consignments each day and the AQIS database records a “PASS/FAIL” result for each consignment.

### 2-4-1 Updating the prior

For the purpose of illustrating the proposed control charting methods, we have aggregated the results on a daily basis and simply noted the number of failures  $x_t$  out of  $n_t$  consignments on day  $t$ . A plot of the observed daily failure rate and cumulative failure rate is shown in Figure 33. Initial estimates of the failure rate are highly variable although ultimately converge to about 3% as evidenced by the blue trace in Figure 33.



**Figure 33. Daily consignment failure rate (red curve) and cumulative failure rate (blue curve) for imports between 4/7/2006 and 29/6/2007.**

We initially assumed a  $\text{beta}(4,20)$  distribution for the prior on  $\theta$  which has 99% of its probability mass between zero and 0.374, a mean of 0.167 and a modal value of 0.136. This choice reflects little or no prior knowledge about  $\theta$  other than we expect it to be less than 0.4. Using the methods of the previous sections we can update the prior at any point in time using all the available information available at that time. This could be as frequently as every day or say, once a month. Figure 34 shows the situation at the end of a year of monitoring.

The top panel in Figure 34 shows the initial  $\text{beta}(4,20)$  prior (blue curve) and the posterior density at the end of the 1 year period (red curve). The posterior is a  $\text{beta}(6.248,165.337)$  distribution which has a mean of 0.036, a median of 0.035, and a modal value of 0.031. 99% of the posterior distribution lies in the interval  $(0, 0.08)$ .

The lower panel of Figure 34 shows the cumulative failure rate as a function of time since monitoring commenced.

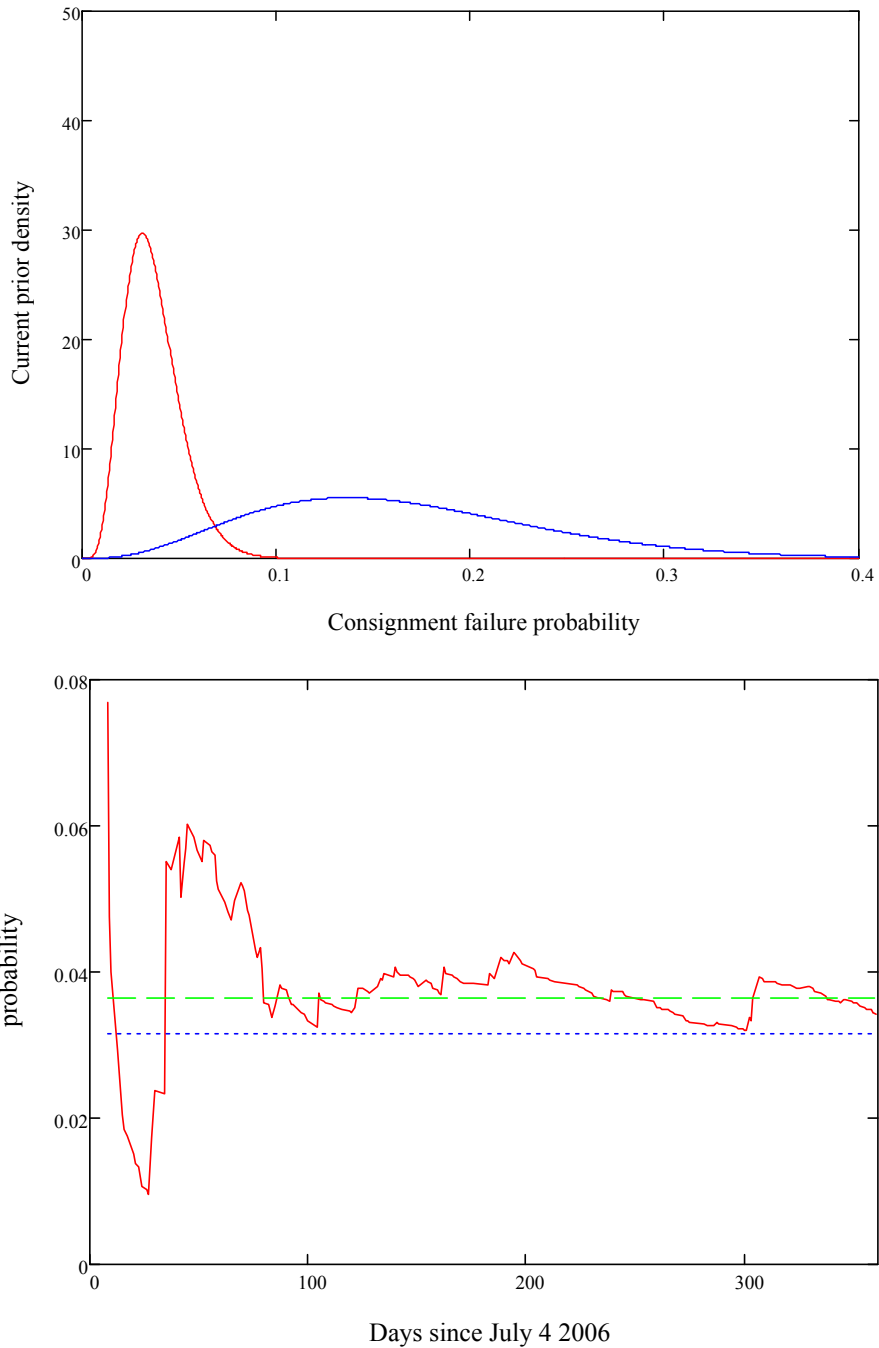


Figure 34. Top: original (subjective) prior density (blue curve) and posterior density (red curve) for true failure rate after 1 year. Bottom: Empirical cumulative failure rate (red line), overall mean failure rate (blue dotted line) and mean of posterior distribution after 1 year (green dashed line).



It is evident from Figure 34 and the related distributional summaries that the relatively vague prior has been considerably ‘sharpened’ after a year of monitoring. The posterior density is compactly centred about the overall failure rate of 0.03 and a Bayesian 95% highest posterior density credibility interval for the true consignment failure rate is readily determined to be  $\{0.011, 0.065\}$ . The upper limit of a 1-sided 95% credibility interval is 0.063 which suggests that a failure rate more than the equivalent of 1 in 16 is evidence of a significant increase in import failure. A more refined instrument for alerting to changing failure status has been provided in the form of the two triggers,  $RL_1$  and  $RL_2$ . We next illustrate how these operate with reference to the present example.

### **2-4-2 Setting adaptive triggers**

By way of example, suppose the current date is August 7, 2006 and we wish to place approximate 99% limits on the number of failures for imports for the following day. There were a total of 128 consignments from the country in question since the start of our monitoring period (we take this to be July 4 2006) – three of which failed inspection, giving a current failure rate of 2.3%. Solving equations 2.7a and 2.7b we obtain the maximum likelihood estimates  $\hat{a} = 3.805$  and  $\hat{b} = 167.819$ . There are 17 consignments on August 8 2006. With  $N = 128, y = 3, n = 17$  and  $\alpha = 0.01$  we use equation 2.11 to determine  $RL_1=2$  and equation 2.12 to determine  $RL_2=2$ . Note, because the outcome of inspection is a *discrete* random variable, equations 2.11 and 2.12 will generally not be able to be satisfied exactly. Thus in this case, the *actual* value for  $\alpha$  is 0.008 for  $RL_1$  and 0.01 for  $RL_2$  as distinct from the *nominal*  $\alpha = 0.01$ . As it turned out there were 5 failures on August 8 2006 and this outcome would have tripped both our triggers for further investigation.

We also note that in the early stages of monitoring,  $RL_1$  and  $RL_2$  will be quite close (in this case they’re identical) reflecting the fact that not much data has been gathered and a significant increase in failure rate on any one occasion has a relatively large impact on the cumulative failure rate. As monitoring progresses, there will tend to be greater separation between  $RL_1$  and  $RL_2$ , although the difference will still tend to be small given the relatively small sample sizes involved. By way of example, we now advance to June 28 2007. By this time, there have been 1,680 consignments resulting in 58 failures. The approximate

$\alpha = 0.01$  triggers for the following day's 15 consignments are  $RL_1=2$  and  $RL_2=3$ . The actual result was zero which is clearly acceptable.

## 2-5 Discussion

The adoption of a Bayesian framework has allowed us to extend traditional control charting methods discussed in chapter 1 to accommodate expert opinion and/or prior belief about the monitored process. Furthermore, the Bayesian approach provides some other important enhancements. For example, 'ignorance' about a new or previously undetected threat is readily accommodated and the intrinsic updating of prior information means that these methods are evolutionary, learning and adaptive. We believe these are important prerequisites for a successful biosecurity surveillance and monitoring system.

It is important to distinguish between monitoring activities that aim to predict or forecast future events with those whose primary objective is to alert or flag the existence of an abnormal event. A comprehensive biosurveillance monitoring strategy will incorporate both pro-active and reactive components. Control charting techniques are reactive, although depending on how they are constructed and implemented, they can provide a close to real-time monitoring capability. A difficulty with pro-active systems such as those used in syndromic surveillance is that forecasting (particularly rare events) is exceedingly difficult with success depending very much on model choice and parameterisation. Indeed, as noted by Burkhom et al. (2007) a critical issue for syndromic surveillance / forecasting systems is their sensitivity to "expected and unexpected data outliers". Burkhom et al. (2007) go on to further state that "for unexpected outliers, we have implemented automated outlier removal schemes to avoid baseline contamination for the adaptive regression, but such schemes can produce unexpected effects and need further study". We regard this as a flawed strategy for two reasons: (i) an "expected outlier" is an oxymoron; and (ii) the automated removal of observations that are, in some sense, aberrant is to be strenuously avoided. It was precisely because of the automated removal of 'outliers' that the hole in the ozone layer was initially undetected. It was only when the 'offending' data was reinstated and the time series data reanalysed that the seriousness of the problem became apparent. Given that the utility of forward looking systems is critically dependent on which

data is included/excluded in the modelling process, it seems to us that data screening tools such as control charts have an important role to play in the development of prospective, forecasting tools.

Hitherto, Bayesian methods have not been widely used in biosecurity / biosurveillance applications although a number of papers have appeared recently which suggest that there is a growing awareness of the potential utility of this statistical paradigm. Wong et al. (2005) used Bayesian networks to extend the Population-wide Anomaly Detection and Assessment (PANDA) algorithm for syndromic surveillance while Hogan et al. (2007) describe a Bayesian aerosol release detector (BARD) that combines medical surveillance and meteorological data to provide an early warning capability for the release of *B. Anthracis*.

In this chapter we have outlined a Bayesian approach to control charting within the context of quarantine monitoring and extension. We have provided a proof-of-concept evaluation of the method using AQIS food import data (Robinson et al. 2008) which demonstrates that the approach shows promise and warrants further development and evaluation. Future work could usefully focus on the elicitation of prior probability distributions as well as the incorporation of other covariates.

## **3-1 INTRODUCTION**

Australia is free of the world's worst animal diseases such as foot and mouth disease (FMD) and bird flu (avian influenza H5N1) although the list of potential threats is long (<http://www.daff.gov.au/animal-plant-health/pests-diseases-weeds/animal>). There are good reasons for taking whatever steps are necessary to ensure that this status is maintained. The 2001 FMD outbreak in the United Kingdom had disastrous consequences with the slaughter of over 4.2 million animals and substantial economic loss (Riley 2007).

Four outbreaks of what is thought to be FMD occurred in Australia in the nineteenth century however there have been no reported outbreaks for over 100 years. Equine influenza (EI) is another highly contagious disease afflicting horses, donkeys, mules, and zebras. Shortly after Animal Health Australia released its disease strategy for equine influenza (Animal Health Australia 2007) an EI outbreak was detected in the Sydney area. The disease spread rapidly through northern NSW into Queensland where it concentrated in the Brisbane region (DPI 2008). It wasn't until Christmas Day 2008 that Australia was officially declared EI disease-free.

In handing down his findings The Hon. Ian Callinan (AC) highlighted shortcomings in the Government's monitoring and surveillance protocols for biosecurity threats (Callinan 2008). Similarly, an analysis of the 2001 FMD outbreak in the United Kingdom revealed serious shortcomings in data collection, processing, and analysis activities in the initial stages of the outbreak (AusVet-CSIRO 2005). The AusVet-CSIRO report suggested that Australia accordingly review its data requirements and mathematical modelling in order to understand the quantitative aspects of animal disease outbreaks.

In a recent review of infectious disease outbreaks Riley (2007) noted a number of interesting phenomena in the spatial dynamics of disease propagation in human and animal populations. Examples of these included "spatial waves of infection" and the tendency of

disease incidence to occur in spatial clusters. The phenomenon of epidemic travelling waves is not new with historical examples provided by the European plague in the Middle Ages, the influenza pandemic in the early 20<sup>th</sup> century and the spread of cholera in Asia and East Europe during the 1960s (Fuentes and Kuperman 1999)

As with any disease, prevention is better than cure and continuous surveillance coupled with stringent border and pre-border controls is essential to the maintenance of Australia's disease-free status. However, *if* an outbreak of a highly contagious and economically devastating disease such as EI or FMD was to occur, the ability to *predict* the subsequent spread of the disease would greatly enhance the prospects of early control and containment. As in the 2007 EI outbreak, enhancements to the biosecurity network can only be made once the deficiencies are understood. To this end, an ability to pin-point the location of the initial outbreak is a critical first step.

This chapter details the outcome of investigations into the development of a cellular automata model to describe the spatial dynamics of infectious disease spread. Additionally, the output of the cellular automata model is combined with limited, spatially-temporally referenced empirical data on actual disease numbers to provide an estimate of the most likely location and time of the initial outbreak.

## **3-2 A CELLULAR AUTOMATA MODEL FOR DISEASE SPREAD**

There is a considerable literature on epidemic models although most population models are zero-dimensional and describe intrinsic epidemical features such as the existence of threshold values for the spread of an infection, the asymptotic solution for the density of infected individuals and density-dependent effects (Fuentes and Kuperman 1999). However, Riley (2007) suggests that spatial models of infectious disease transmission which integrate knowledge of the infection process are “the only plausible experimental system ... to investigate observed patterns and to evaluate alternative intervention options”.

Mathematical models describing population dynamics usually use either differential or difference equations depending on whether ‘time’ is treated as a continuous or discrete variable. An alternative to this approach are cellular automata (CA) methods having

discrete time increments and a matrix representation of a geographical network. A set of rules governs the evolution of the automata such that the state of an element at each time step is expressed in terms of its own state and those of its neighbours at earlier time steps. As noted by Fuentes and Kuperman (1999), CA methods enjoy a number of advantages over conventional differential-difference equation approaches such as considerably faster computational speeds and the ease with which certain epidemiological features can be incorporated as well as local and seasonal effects. CA models have been used to model a range of problems associated with pathogen and disease spread including rabies in fox populations (Benyoussef et al. 1999) and FMD in feral pigs in Queensland (Doran and Laffan, 2005).

CA models are potentially well-suited to modelling the spread of an infectious disease such as FMD which can exhibit long-range spatial dynamics as a result of airborne spreading (Cannon and Garner 1999). Indeed, it has been suggested that a 1981 FMD outbreak in the UK was initiated by windborne particles carried across the English Channel from France (Alexandersen et al., 2001; Donaldson, 1983; Sørensen et al. 2000). Unless the wind is erratic, it is reasonable to assume that disease incidence data would exhibit a relatively high degree of spatial continuity that was aligned with the predominant wind-direction. According to Cannon and Garner (1999), most wind-borne spread over land is less than 10 km although this can be up to 60 km. This suggests that an anisotropic spatial covariance model having a 10 km 'range of influence', say, could potentially be useful in describing the spatial correlation structure in disease spread. The risk of airborne spread of FMD in Australia was assessed by Garner and Cannon (1995). Figure 35 is taken from their report and indicates that conditions favourable for the survival of FMD in aerosols occur at least 50% of the time in most of south-east Australia and all of Tasmania. The use of a discrete grid as the basis for representing and modelling the spatial dynamics of disease spread is not only convenient but parallels the way in which management agencies convey risk to the broader community (Figure 37). Our approach is to take the region of interest (e.g. Figure 36) and overlay a grid whose spacing is commensurate with the phenomenon of interest (Figure 38).

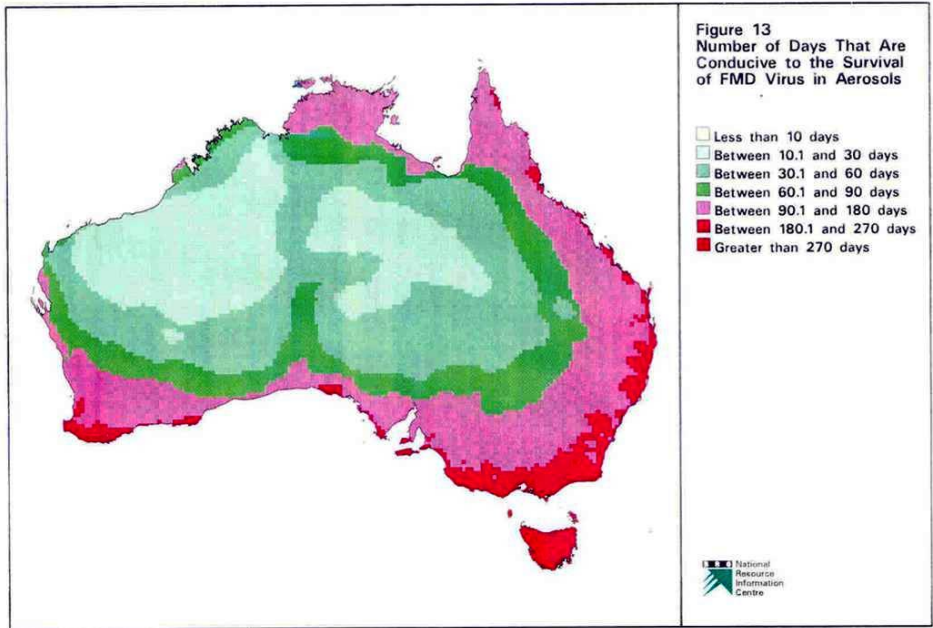


Figure 35. Average number of days per year that are conducive to persistence of FMD virus in aerosol (From Cannon and Gardner 1999).

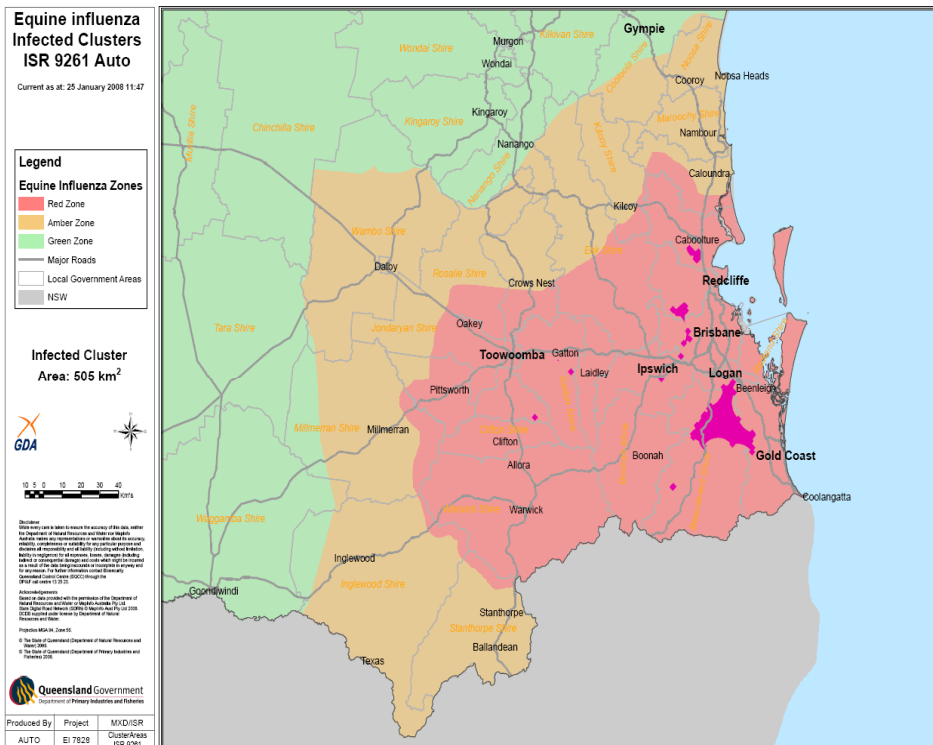


Figure 36. Zonation of EI infected regions in Queensland during the 2007/08 outbreak. Source: <http://www2.dpi.qld.gov.au/extra/ei/maps/QLDInfectedCluster.gif> (accessed 25 January 2008)

The starting point for the CA model is a probability model that describes the spatial extent and orientation of the likelihood that a ‘diseased’ cell (i.e. one in which the presence of the disease has been confirmed) will ‘infect’ neighbouring cells. The general situation is depicted in Figure 39.

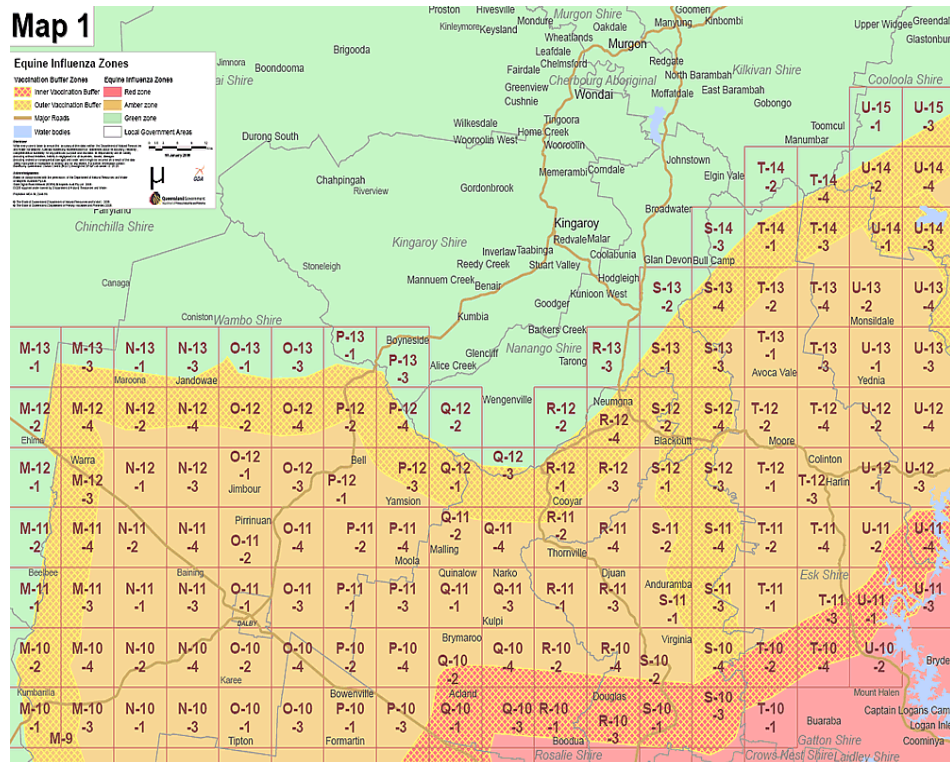


Figure 37. Illustration of grid representation used by authorities in managing EI outbreak. (Source: <http://www2.dpi.qld.gov.au/extra/ei/maps/map8.gif>)



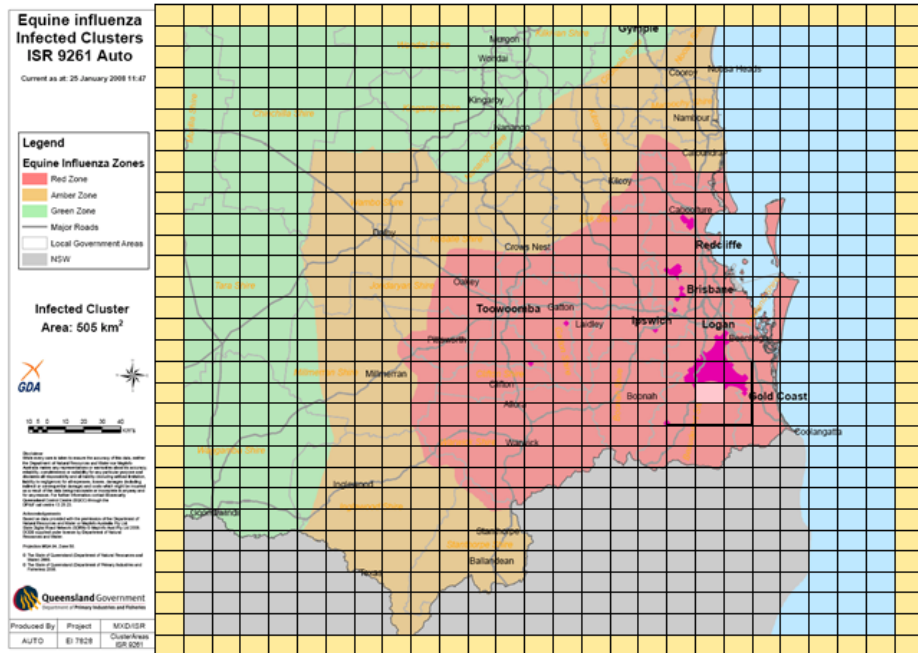


Figure 38. The region of interest in Figure 36 with grid overlay that forms the basis of the CA modelling approach.

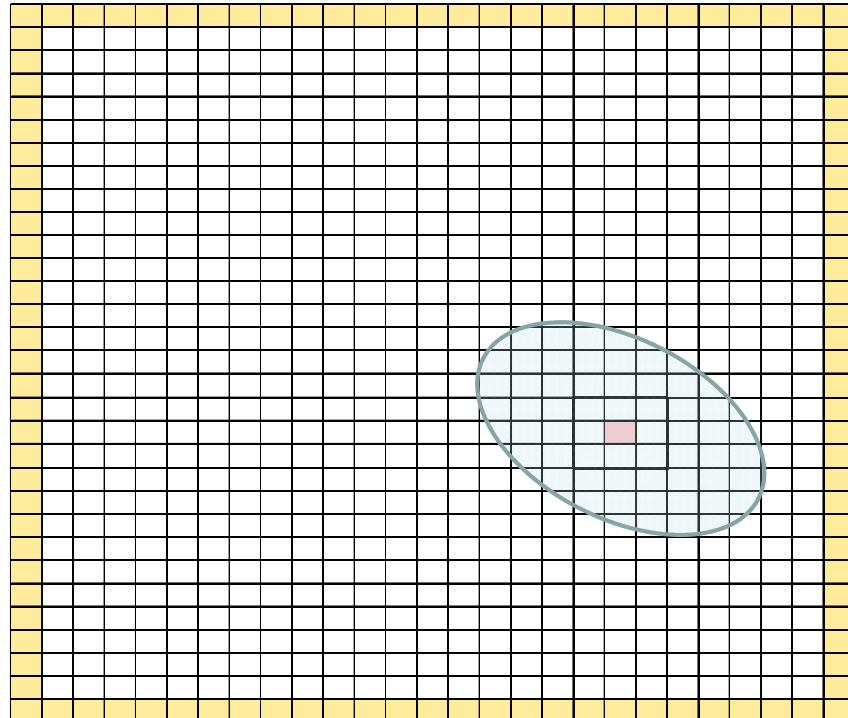


Figure 39. The grid in Figure 38 showing an 'infected' cell (red shading) and associated spatial pattern for disease transmission.

The mathematical detail underpinning the development of the CA model is presented in the next section.

### 3-2-1 Mathematical formulation

Our problem formulation is constructed around the general representation depicted in Figure 40. Our ‘target cell’ (ie. the one of interest) is located in row  $u$  and column  $v$  of a 2-D grid overlayed on the region of interest. Associated with each grid cell is a Bernoulli variable  $Z$  indicating disease status<sup>6</sup> viz:

$$Z_{i,j}^t = \begin{cases} 1 & \text{if cell } \{i,j\} \text{ is infected at time } t \\ 0 & \text{otherwise} \end{cases}$$

with  $P[Z_{i,j}^t = 1] = \theta_{i,j}^t$ .

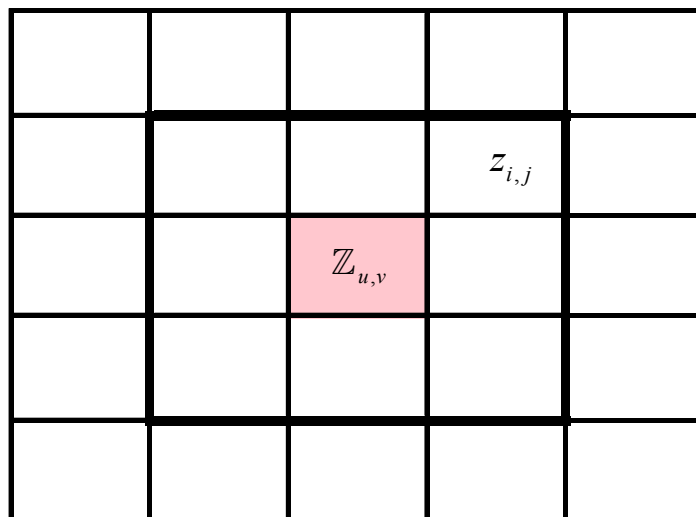


Figure 40. General CA situation with cell of interest (red shading) and neighbouring cell.  $Z$  is a binary variable indicating cell’s disease status. Region bordered by heavy line depicts spatial extent of influence or impact of cell on its neighbours.

We further make the following assumptions: (i) infection is an absorbing state, that is, once a cell is ‘infected’ it remains infected (at least for the period of interest); and (ii) the

---

<sup>6</sup> Clearly, it is not the cell *per se* which is infected – it is the occurrence of at least one infected individual within the cell.

risk of infection is a function of the disease status of neighbouring cells *and* the distance between the target cell and infected neighbouring cells.

We define the transmission ‘effectiveness’,  $\phi(r)$  as the probability that an uninfected cell at distance  $r$  from an infected cell will become infected during the period  $\{t, t+1\}$ . Without loss of generality, we denote  $r_{i,j}$  to be the distance between the centre of cell at grid location  $\{i,j\}$  and the centre of the target cell.

We next consider the updating step as time is incremented by 1 unit. Hence:

$$\theta_{u,v}^{t+1} = P[\mathbb{Z}_{u,v}^t = 1] + P[\mathbb{Z}_{u,v}^t = 0]P[\text{cell } \{u,v\} \text{ becomes infected in the interval } \{t, t+1\}] \quad (3.1)$$

For the sake of brevity and simplicity, we let  $I_{i,j}$  denote the event that target cell in row-column position  $\{u, v\}$  becomes infected by cell  $\{i, j\}$  in the interval  $\{t, t+1\}$  and  $\bar{I}_{i,j}$  its complement ie.  $P[\bar{I}_{i,j}] = 1 - P[I_{i,j}]$ . Thus,

$$P[\bar{I}_{i,j}] = [1 - \phi(r_{i,j})] \theta_{i,j}^t + (1 - \theta_{i,j}^t) \quad (3.2)$$

For the target cell *not* to become infected during the period  $\{t, t+1\}$  requires an unsuccessful transmission from every cell to the target. Assuming the effectiveness of transmissions are unrelated, we have:

$$P[\bar{I}_{i,j}] = 1 - \prod_{\substack{\text{all } i,j \\ \{i,j\} \neq \{u,v\}}} \left\{ [1 - \phi(r_{i,j})] \theta_{i,j}^t + (1 - \theta_{i,j}^t) \right\} \quad (3.3)$$

Substituting equation 3.3 for the last term in equation 3.1 gives the result:

$$\theta_{u,v}^{t+1} = P[\mathbb{Z}_{u,v}^t = 1] + P[\mathbb{Z}_{u,v}^t = 0] \left\{ 1 - \prod_{\substack{\text{all } i,j \\ \{i,j\} \neq \{u,v\}}} \left\{ [1 - \phi(r_{i,j})] \theta_{i,j}^t + (1 - \theta_{i,j}^t) \right\} \right\}$$

or

$$\theta_{u,v}^{t+1} = \theta_{u,v}^t + (1 - \theta_{u,v}^t) \left\{ 1 - \prod_{\substack{\text{all } i,j \\ \{i,j\} \neq \{u,v\}}} \left\{ [1 - \phi(r_{i,j})] \theta_{i,j}^t + (1 - \theta_{i,j}^t) \right\} \right\} \quad (3.4)$$

We next consider candidate models to describe the transmission effectiveness as a function of separation from an infected cell.

### Models for transmission effectiveness

It is assumed in this development that the transmission effectiveness is only a function of the distance to an infected cell – in other words, our model is *omnidirectional* or *anisotropic*. Although the exact form of the function describing this relationship needs to be informed by expert knowledge and disease-specific data, it is not unrealistic to assume a radial basis function for  $\phi(r)$  - such as a Gaussian model. One such possibility is given by equation 3.5.

$$\phi(r) = k e^{-\frac{r^2}{\lambda}} \quad (3.5)$$

Different choices for the constants  $k$  and  $\lambda$  in equation 3.5 give rise to different spatial patterns of transmission effectiveness (Figure 41).

### Initial conditions

Equation 3.4 is recursive and therefore requires the specification of an initial state for each cell, ie.  $\theta_{i,j}^0 \forall \{i, j\}$ . The process of supplying an initial guess for  $\theta_{i,j}^0$  followed by the updating step parallels a Bayesian analysis whereby a *prior* probability is transformed into a *posterior* probability via a likelihood function (see section 2-2 for a discussion). As in a Bayesian analysis,  $\theta_{i,j}^0$  can be chosen to reflect the degree of belief in the initial infection status. In the absence of any prior information, knowledge, or understanding (other than that captured in equation 3-5) we will assume an initial configuration that is the equivalent of a Bayesian non-informative prior. For example,  $\theta_{i,j}^0 = p \forall \{i, j\}$  where  $p$  is a constant

corresponds to the case where the initial outbreak is equally likely to occur (or have occurred) anywhere in the region of interest.

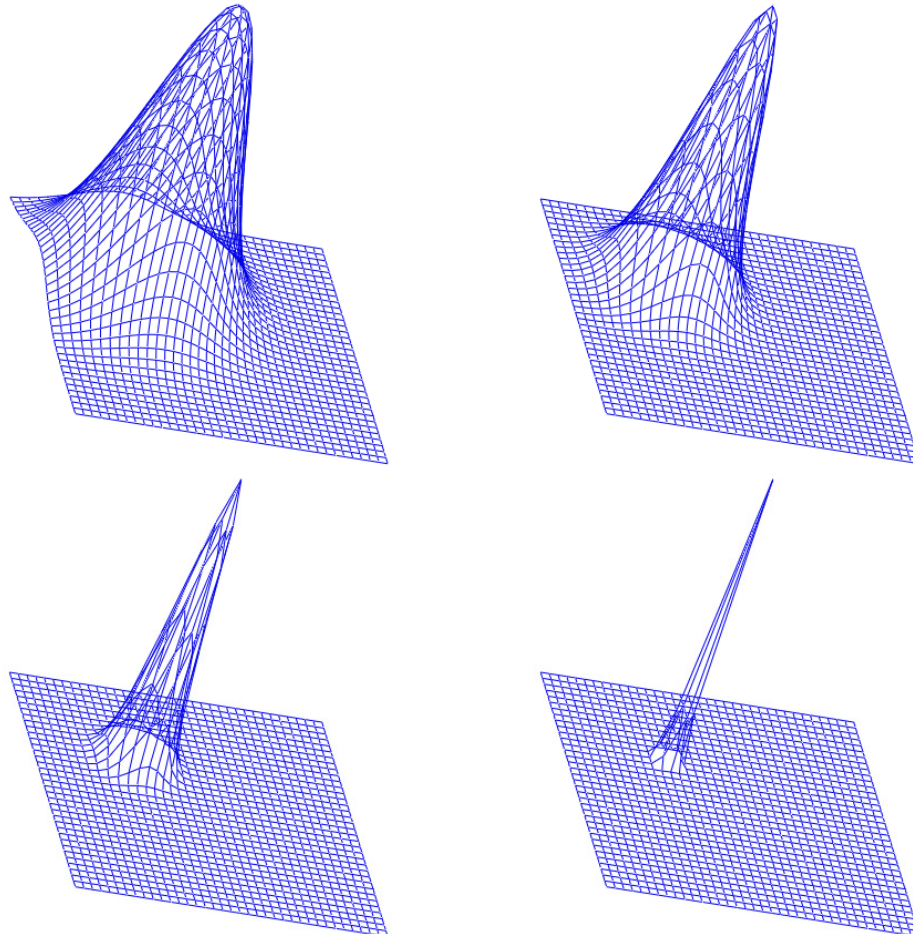
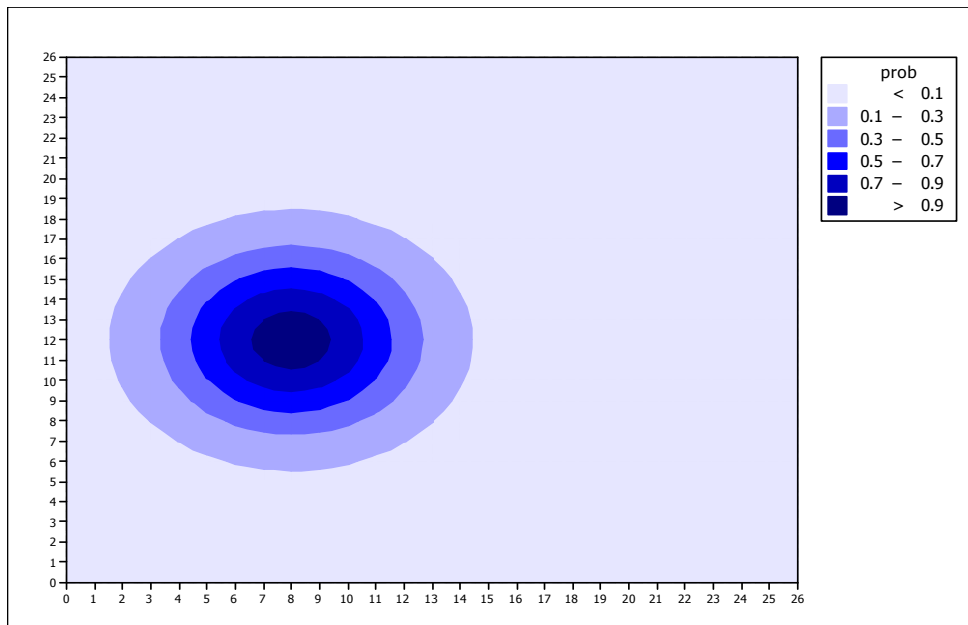


Figure 41. Illustration of a spatial probability model for disease spread for generic cell  $\{i,j\}$  at fixed point in time. Grid represents region of interest. Height of surface is proportional to probability of spread of infection from cell  $\{i,j\}$  to neighbouring cells. Each plot represents a different range of influence from highly localised (bottom right) to far-ranging (top left).

### Modelling the disease spread

To illustrate how equation 3.4 models the spread of a disease, we consider the case of equally-likely outbreak locations. At  $T=0$ , we assume an outbreak at a particular grid

location<sup>7</sup>  $\{r_0, c_0\}$  and make the assignment  $\theta_{r_0, c_0}^0 = 1$ . The probability that neighbouring cells become infected during the next time increment is given by  $\phi(r_{i,j})$  where  $r_{i,j}$  is distance from  $\{r_0, c_0\}$  to a neighbouring cell at grid location  $\{i,j\}$ . Figure 42 illustrates the spatial characteristics of the transmission effectiveness probability contours based on a particular parameterisation of equation 3-5 with  $\{r_0 = 12, c_0 = 8\}$ . At the next time increment, the probability of infection is updated for every grid cell using equation 3.4 and the process repeated  $N$  times.

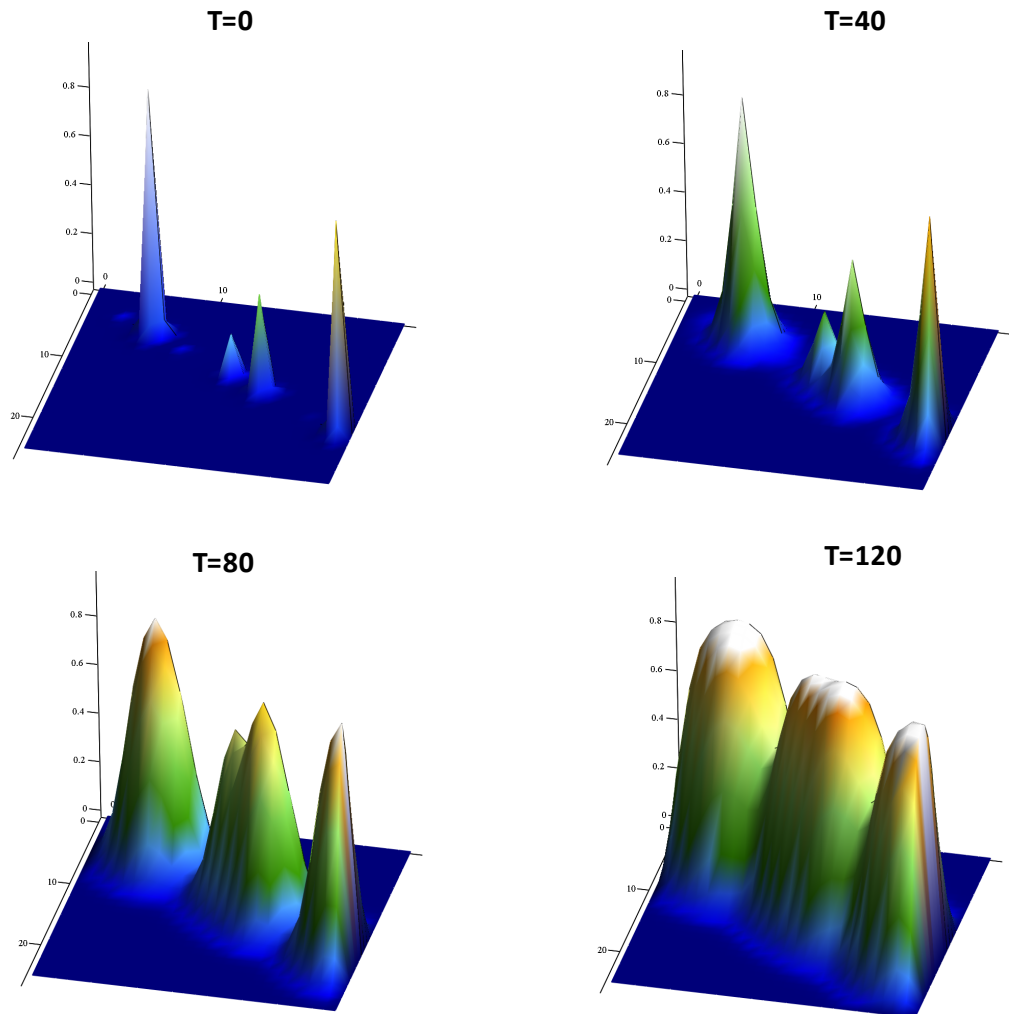


**Figure 42. Illustrative contours of probability representing likelihood of infection around cell having grid coordinates  $\{12,8\}$ .**

A slightly more complex scenario is shown in Figure 9 which depicts a non-uniform assignment of initial outbreak probabilities at time  $T=0$  together with ‘snapshots’ at  $T=40$ ,  $T=80$ , and  $T=120$ .

---

<sup>7</sup> Or set of locations



**Figure 43.** 3-D depiction of progression of disease spread starting with initial outbreak pattern at  $T=0$  and at three subsequent time periods. Vertical scale is probability of infection.

The attractive feature of Figure 43 (or more correctly, the underlying model) is it can be used as the basis of inference about the *time* of an outbreak. For example, we assume the outbreak was at grid cell  $\{r_0, c_0\}$  and that we have available sample data on disease incidence at some time  $T = T_0 + k\delta t$  where  $T_0$  is the outbreak time and  $k$  is the number of time increments each of length  $\delta t$  units. We then use a maximum likelihood estimation (*mle*) procedure to estimate  $k$  as the value (call it  $\hat{k}$ ) that maximises the joint probability function for the observed data using the spatial probability model of equation 3.4 evaluated at the  $k^{\text{th}}$  time step. An example of the likelihood function plotted against  $k$  is shown in Figure 44. This procedure can be repeated for all grid locations thereby generating a total of

$N = r \times c$  such curves, where  $r$  and  $c$  denote respectively the number of rows and columns of the grid. The estimated outbreak time *and location* is associated with the likelihood profile whose maximum is the largest among all  $N$  plots. The *mle* for the outbreak time is  $\hat{T}_0 = T - \hat{k} \delta t$ . This procedure is detailed more formally in the next section.

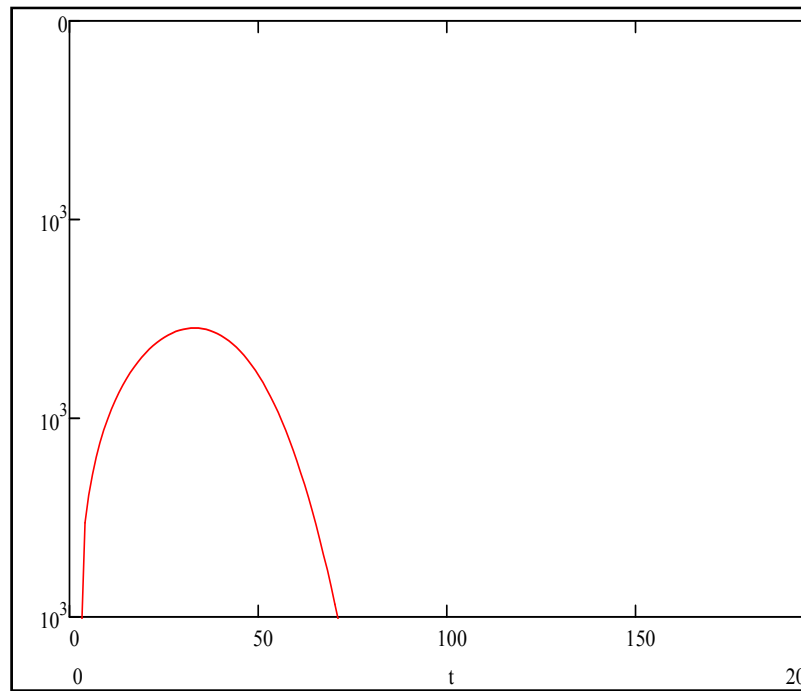


Figure 44. Likelihood profile plot for  $k$  (number of time increments since outbreak).

### 3-3 Inferring the time and location of an outbreak

The general situation was described in the previous section. We now develop the mathematical and computational detail associated with the maximum likelihood procedure for estimating the time and location of an outbreak.

We commence by assuming that at some time  $T = T_0 + k \delta t$  we observe for grid cell  $\{i, j\}$  a proportion,  $p_{ij}^{(T)}$  of infected units (animals, people, agricultural plots etc.) where



$p_{ij}^{(T)} = \frac{X_{ij}^{(T)}}{n_{ij}^{(T)}}$  and  $X_{ij}^{(T)}$  is the number of infected units in cell  $\{i,j\}$  out of a total  $n_{ij}^{(T)}$  inspected in grid cell  $\{i,j\}$  (Figure 45).

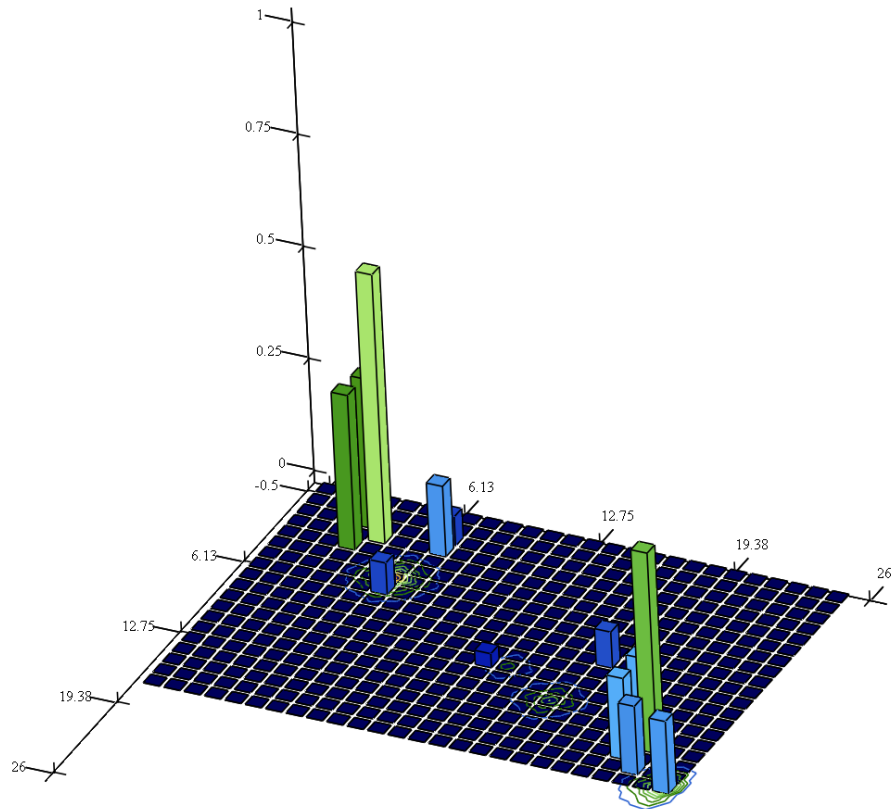


Figure 45. General situation depicting region of interest with vertical bars depicting empirical rates of infection.

The total number of units in grid cell  $\{i,j\}$ ,  $N_{ij}^{(T)}$  may or may not be known. In any event, we assume  $N_{ij}^{(T)} \gg n_{ij}^{(T)}$ . If we further assume that the inspected units represent a random sample from the population of susceptible units and that the observations on disease status are independent of each other, then a reasonable probability model for the  $X_{ij}^{(T)}$  data is the binomial distribution.

Hence

$$X_{ij}^{(T)} \sim \text{binom}\{n_{ij}^{(T)}, \theta_{ij}^{(T)}\} \quad (3.6)$$

with  $X_{ij}^{(T)}$  and  $n_{ij}^{(T)}$  defined above and  $\theta_{ij}^{(T)}$  given by equation 3.4.

We next define the set  $\mathbf{S}$  of cardinality  $m$  to be comprised of the row-column indexes of all *sampled* grid locations. For fixed  $T$  and given outbreak location in grid cell  $\{r_0, c_0\}$ , the log-likelihood function for the unknown  $\theta_{ij}^{(T)}$  is given by equation 3.7.

$$\ell\{\theta_{ij}^{(T)}; x_{ij}^{(T)}, r_0, c_0\} = \sum_{(i,j \in \mathbf{S})} \left[ x_{ij}^{(T)} \log(\theta_{ij}^{(T)}) + (N_{ij}^{(T)} - x_{ij}^{(T)}) \log(1 - \theta_{ij}^{(T)}) \right] \quad (3.7)$$

Note that the right hand side of equation 3.7 is a function of  $r_0$  and  $c_0$  by virtue of the relationship between  $\theta_{ij}^{(T)}$  and  $\phi(r_{i,j})$  with  $r_{i,j}$  the distance between grid locations  $\{r_0, c_0\}$  and  $\{i, j\}$ . Thus, for *fixed*  $T$ , equation 3.7 is evaluated for all  $N = r \times c$  possibilities obtained by varying  $\{r_0, c_0\}$  over the entire grid. It is then a simple matter to identify the location at which equation 3.7 attains its maximum (Figure 46).

We now proceed to illustrate the method with an example.

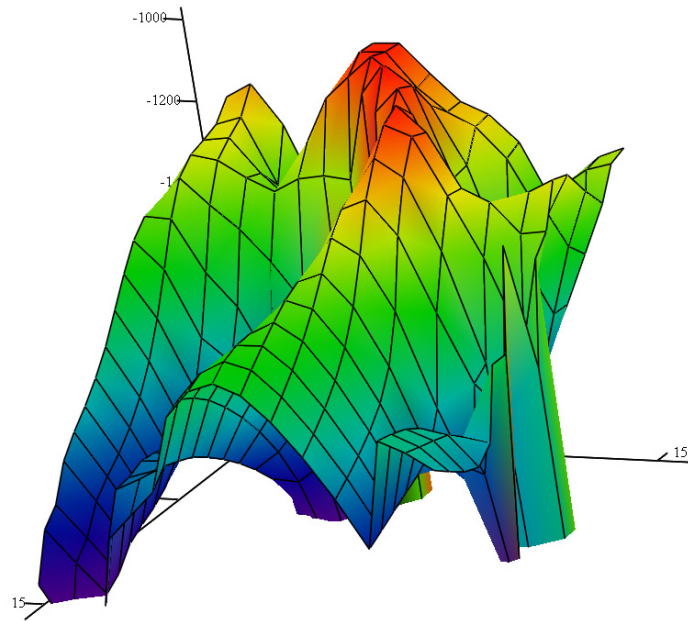


Figure 46. Illustrative likelihood surface for outbreak location.

### 3-4 Example

To test the efficacy of the estimation procedure, we generated  $\{x, n\}$  data pairs on a 26 x 26 grid. The outbreak occurred at time  $T = T_0$  in grid cell  $\{r_0 = 10; c_0 = 15\}$  and sample data was obtained  $78\delta t$  time units later. The transmission effectiveness model used is given by equation 3.8.

$$\phi(r_{ij}) = \exp \left\{ - \left[ \frac{(r_i - r_0)^2}{a} + \frac{-2(r_i - r_0)(c_j - c_0)}{\sqrt{ab}} + \frac{(c_j - c_0)^2}{b} \right] \right\} \quad (3.8)$$

with  $a=0.0025$  and  $b=0.0055$ .

A random sample of  $n=146$  cells was selected from these 676 pairs (see Appendix D for a listing of the data used). Figure 41 shows the sample proportions  $(x_{ij} / n_{ij})$  overlaid on a contour plot of the theoretical probability field from which they were generated. The maximum likelihood procedure described in the previous section was programmed within

the Mathcad 14<sup>®</sup> symbolic computing environment. A listing and description of the routines can be found in Appendix E. The likelihood was evaluated for 200 time increments for each of the 676 possible outbreak locations. The maximum was attained at grid location  $\{r_0 = 10; c_0 = 15\}$  (Figure 48) after 78 time increments (Figure 49) which corresponds exactly with the parameters used to generate the data. Given that we used the *true* probability model (equation 3.8) to compute the likelihoods, the perfect agreement between estimated and actual parameters is not totally unexpected. Nevertheless, we regard this example as a test of the integrity of the estimation procedure. Substantial discrepancies between actual and estimated parameters for this example would have cast doubt over the utility of the modelling and estimation procedures.

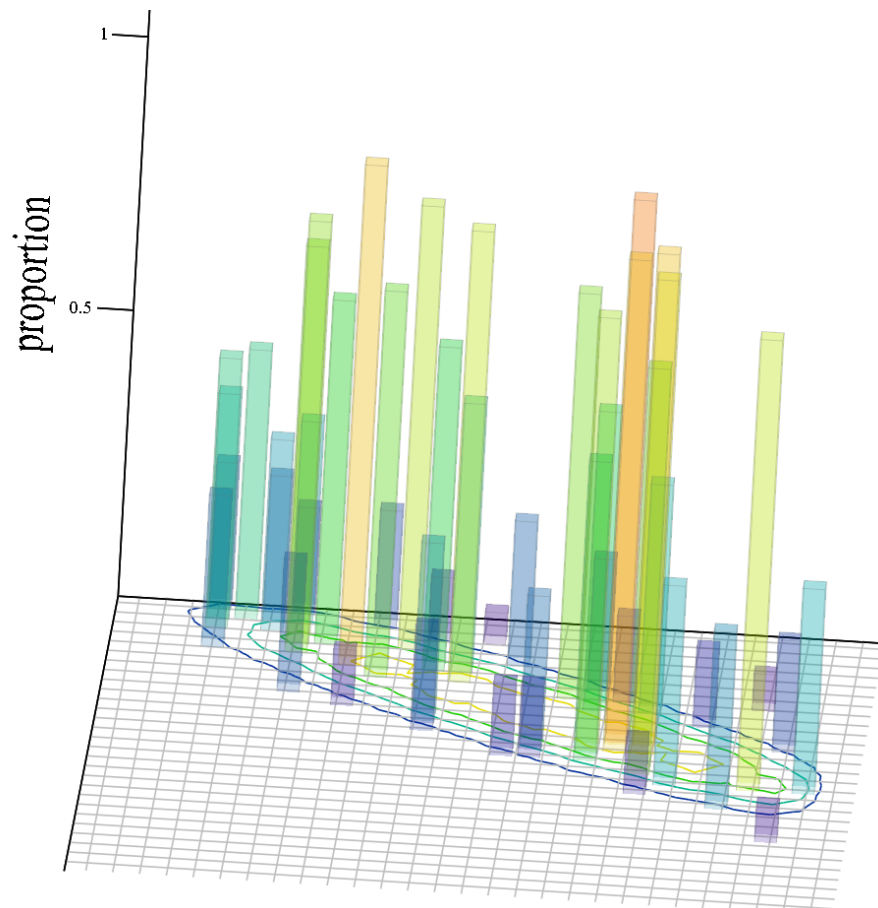


Figure 47. 3-D visualisation of empirical data (infection rates) used in the example. Underlying contour lines derived from theoretical response-generating model.

In reality, neither the functional form nor the parameter values for  $\phi(r_{ij})$  will be known. There are a number of strategies which could be employed in such cases. The simplest, would be to use expert opinion and prior studies to identify a suitable probability model. A more complex and computationally intensive approach would be to specify the functional form but leave the parameters as unknown. A relatively straightforward modification to the estimation algorithm could be made so that the unknown model parameters were estimated simultaneously with the outbreak location and time by maximising over all  $\{r_0; c_0\} \times \mathcal{K}$  space-time combinations within some defined parameter space,  $\Theta$ .

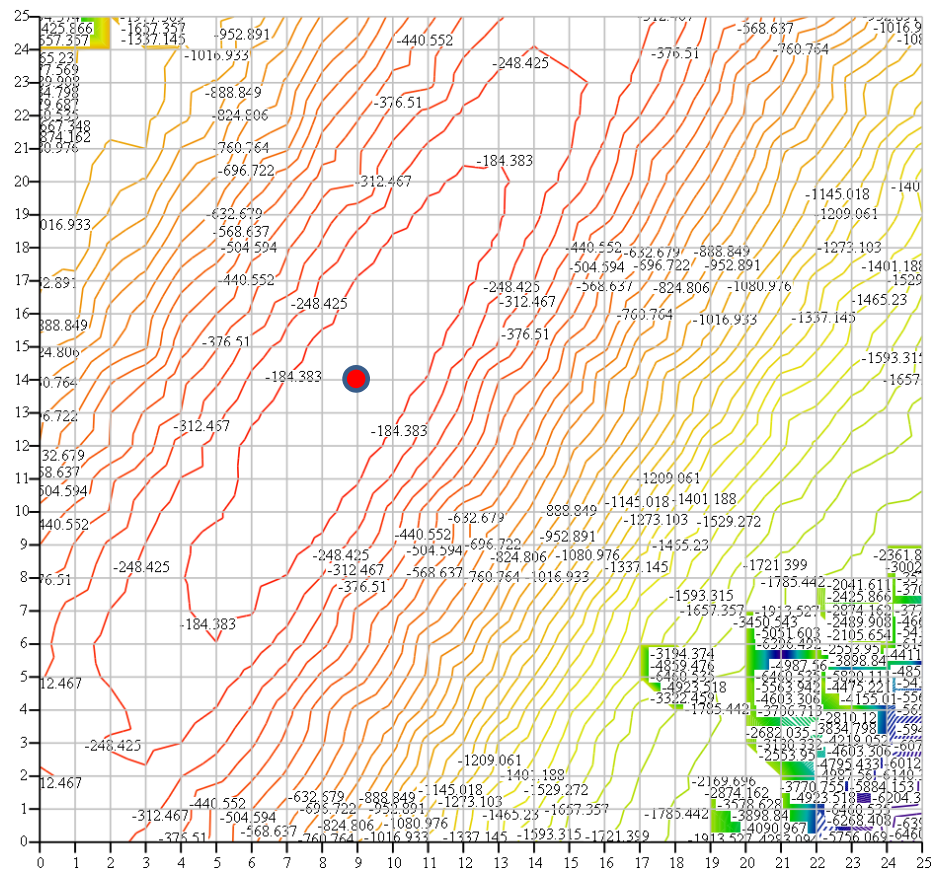
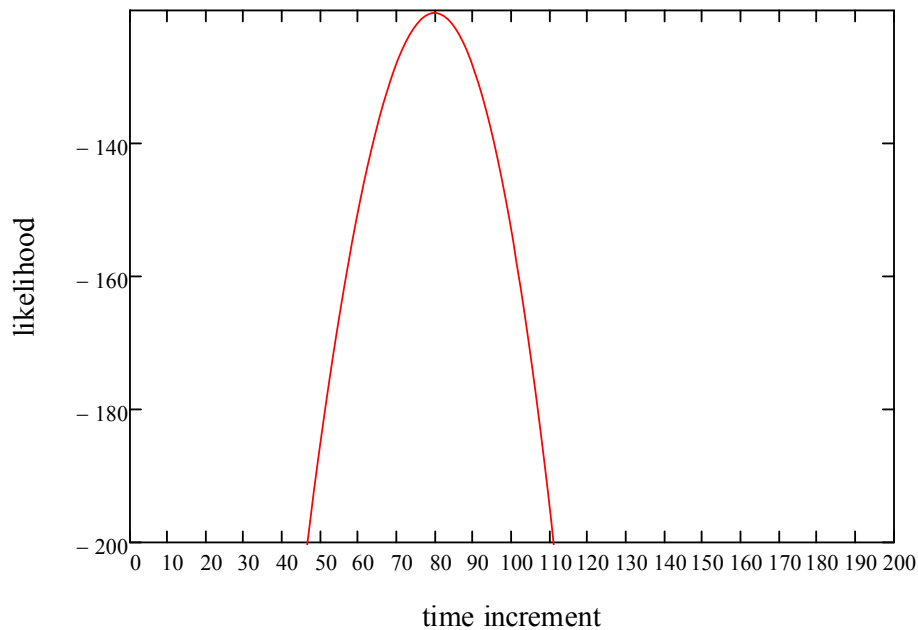


Figure 48. Profile plot showing contours of likelihood of outbreak location. Maximum attained at grid position located at row 14, column 9. Notes: (i) this plot is rotated 90° relative to Figure 47; (ii) plot is offset by 1 unit in each direction due to placement of origin at {0,0}, hence most likely grid location for disease outbreak is cell {10,15}.



**Figure 49.** Likelihood profile plot for temporal component of CA model. Most likely outbreak time is 78 time *prior* to time of sampling.

### 3-5 Discussion

In this chapter we have outlined an approach based on a cellular automata model coupled with empirical observations on incidence rates to estimate both the time and location of an initial disease outbreak. Traditional epidemiological modelling approaches typically focus on the temporal component using differential – difference equations to model the progression of disease spread through time as well as describing other (important) phenomena such as threshold and density-dependent effects. Being discrete, the CA model can be readily implemented on a spatial grid of arbitrary size and resolution and has the advantage of being able to model complex space-time interactions. When coupled with a probability model for the response-generating mechanism the CA model can be used to infer the initial outbreak time and location using maximum likelihood methods. While there is no doubt that once an outbreak of a disease has been detected, much of the subsequent monitoring effort would be focussed on tracking its progression in space and time with a view to containment and eradication. Nevertheless, an ability to

pinpoint the outbreak location and time has been acknowledged as an important capability for post-outbreak evaluation and follow-up activities – particularly if the outbreak was the result of a failure in procedures, process, or facilities. In other cases it may be important to understand whether a disease outbreak was of human or natural origin. For example, the official Soviet explanation of the 1979 outbreak of anthrax in Sverdlovsk, U.S.S.R. was that it was caused by contaminated meat. The U.S. intelligence community suspected the outbreak was due to the aerosol release of *B. Anthracis* from a military microbiology facility. Resolution of this issue was important in deciding whether or not the Soviet Union was in violation of international treaties to which it was a signatory (Hogan et al. 2007).

A number of spatial and spatio-temporal modelling tools for biosurveillance have used conventional statistical modelling approaches such as generalised linear mixed models (GLMMs). For example, the BioSense program run by the U.S. Center for Disease Control (<http://www.cdc.gov/BioSense/>) utilises a variant of GLMMs known as small area regression and testing (SMART) to enhance early detection and situational awareness of possible biologic terrorism attacks (Bradley et. al. 2005). However, as noted by Fricker (2008) the method in BioSense only uses spatial information to bin data into separate time series and is thus not strictly a spatial model. Most spatial models for biosurveillance utilise the scan statistic (Kulldorff, 1997) to detect disease clusters. The method presented in this chapter models the space-time trajectory of infectious state and couples this with a maximum likelihood method to infer initial outbreak time and place using monitored data.

While our approach has been developed to a proof-of-concept stage, further enhancements would increase the appeal of this methodology. For example, the transmission effectiveness model (equation 3-5) could be extended to incorporate ancillary information about the target and neighbouring cells such as the number of susceptible individuals for an animal disease or the size of area at risk for a plant disease. Thus, in considering the transmission between any two grid cells (denote them cell  $I$  and cell  $J$ ), a  $p$ -dimensional vector of ancillary values  $\underline{X}_{IJ} = [x_{IJ1}, x_{IJ2}, \dots, x_{IJp}]^T$  could be defined (where one of the  $x_{IJ}$  is the distance between  $I$  and  $J$ ). A candidate model for transmission effectiveness that utilises all the available information is the multivariate logistic function (equation 3.9).

$$\Phi_{ij} = \frac{\exp(X_{ij}^T \underline{\beta})}{1 + \exp(X_{ij}^T \underline{\beta})} + \xi_{ij} \quad (3.9)$$

where  $\underline{\beta}$  is a ( $p \times 1$ ) vector of parameters and  $\xi_{ij}$  is a stochastic error term with some assumed distribution (for example  $\xi_{ij} \sim N(0, \sigma_{\xi}^2)$ ). To be applicable, values for the parameter vector  $\underline{\beta}$  would need to be either supplied (based on knowledge of the system) or estimated using data from a separate calibration study.



