

# New Approaches to monitoring for biosecurity risk

David R. Fox



# What is risk?

1. risk = hazard x exposure
2. Duckworth (1998):
  - is a qualitative term
  - cannot be measured
  - is not synonymous with probability
  - “to ‘take a risk’ is to allow or cause exposure to the danger”
3. is the chance, within a specified time frame, of an adverse event with specific (negative) consequences

- Development and adoption of a ‘standard’ risk metric seems a long way off (never?);
- Bayesian methods are becoming increasingly popular, although acceptance may be hampered by biases and lack of understanding;
- More attention needs to be given to appropriate statistical *modelling*. In particular:
  - model choice
  - Parameter estimation
  - Distributional assumptions
  - ‘Outlier’ detection and treatment
  - robust alternatives (GLMs, GAMs, smoothers etc).

## What's an anomaly?

- A specific form of change that is localised in the space-time domain
- Identification of an anomaly can be either 'reactive' or 'proactive'
  - Monitoring is typically reactive
  - Surveillance aims to be pro-active (or at least 'real-time').

## What's a control chart?

- Is a device to detect the non-random, 'out-of-control' state of a process.
- Use in environmental sciences becoming more common (eg. NWQMS)
- Two main types: *Variables* or *Attributes*
  
- Control charts not new – 1924 Shewhart
- Widespread use in industrial ('brown') statistics post WWII
- 1950s and 1960s – development of QC (Shewhart, Deming, and others)
- 1960s and 1970s – Japanese manufacturing industry (Taguchi)

# Control Charting for anomaly detection

- Simplest form of alerting based on 'warning' and 'action' limits

warning

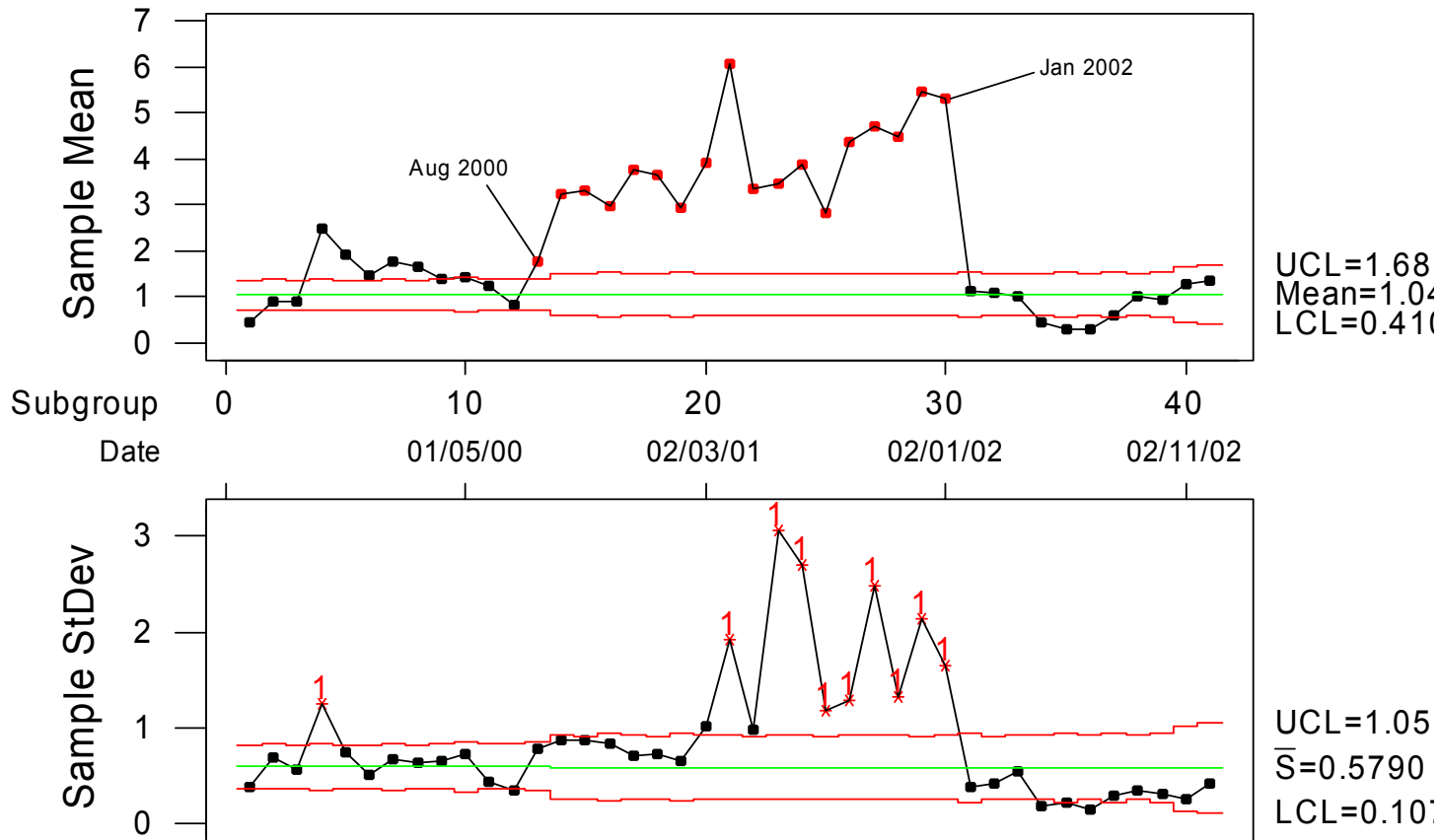
$$UCL = \bar{X} + 2\sigma_X \quad LCL = \bar{X} - 2\sigma_X$$

action

$$UCL = \bar{X} + 3\sigma_X \quad LCL = \bar{X} - 3\sigma_X$$

# Control Charting for anomaly detection

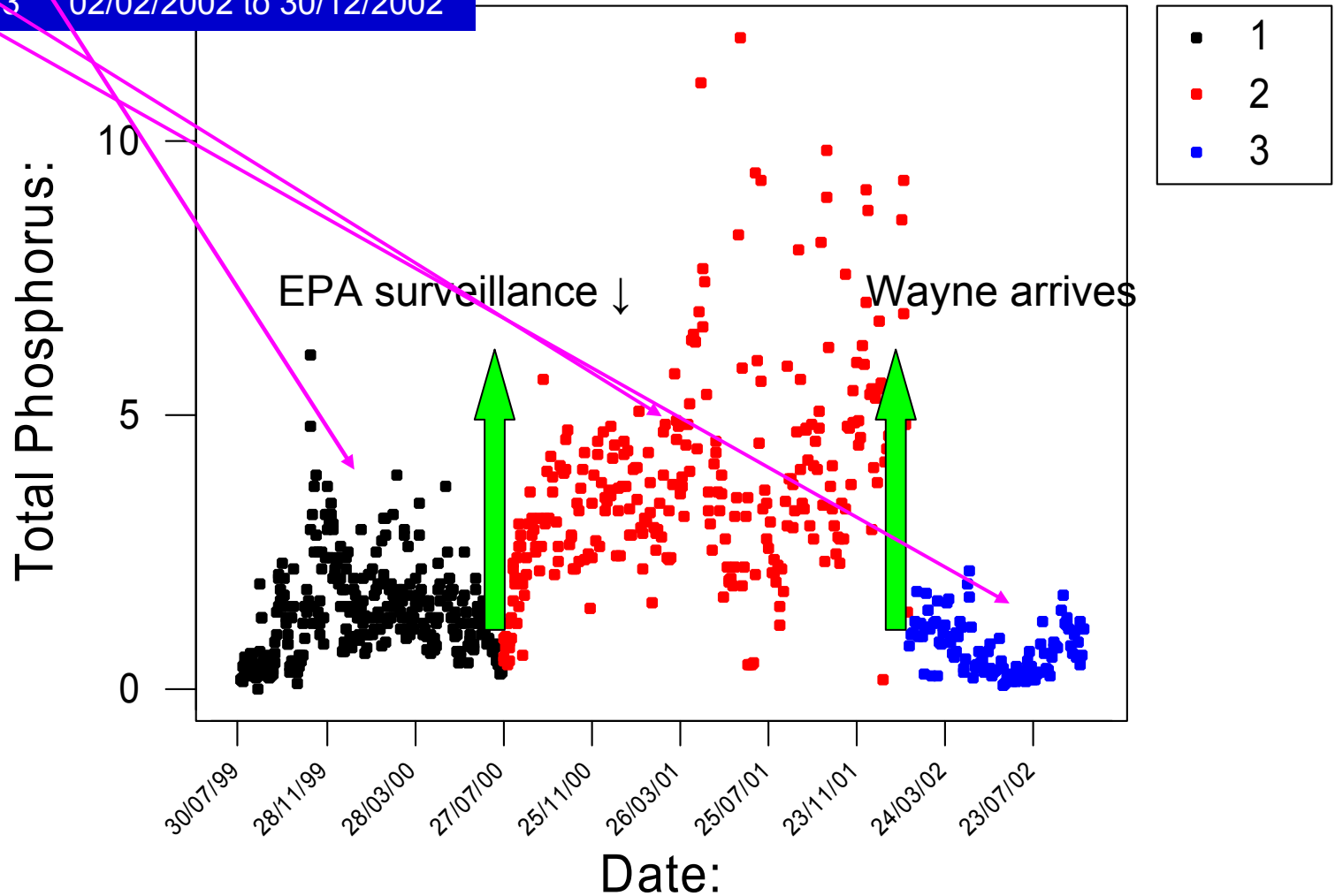
## Xbar/S Chart for Total P



## Control Charting for anomaly detection

## TP epochs

Epoch 1	01/08/1999 to 25/07/2000
Epoch 2	26/07/2000 to 01/02/2002
Epoch 3	02/02/2002 to 30/12/2002



# Exponential Smoothing

- E.S. – is “an elementary model of how a person learns”

$$\bar{X}_k = \bar{X}_{k-1} + w \left( X_k - \bar{X}_{k-1} \right) \quad 0 < w < 1$$

$$w = 0 \quad \Rightarrow \quad \bar{X}_k = \bar{X}_{k-1} \quad \text{ie. all history – no new data}$$

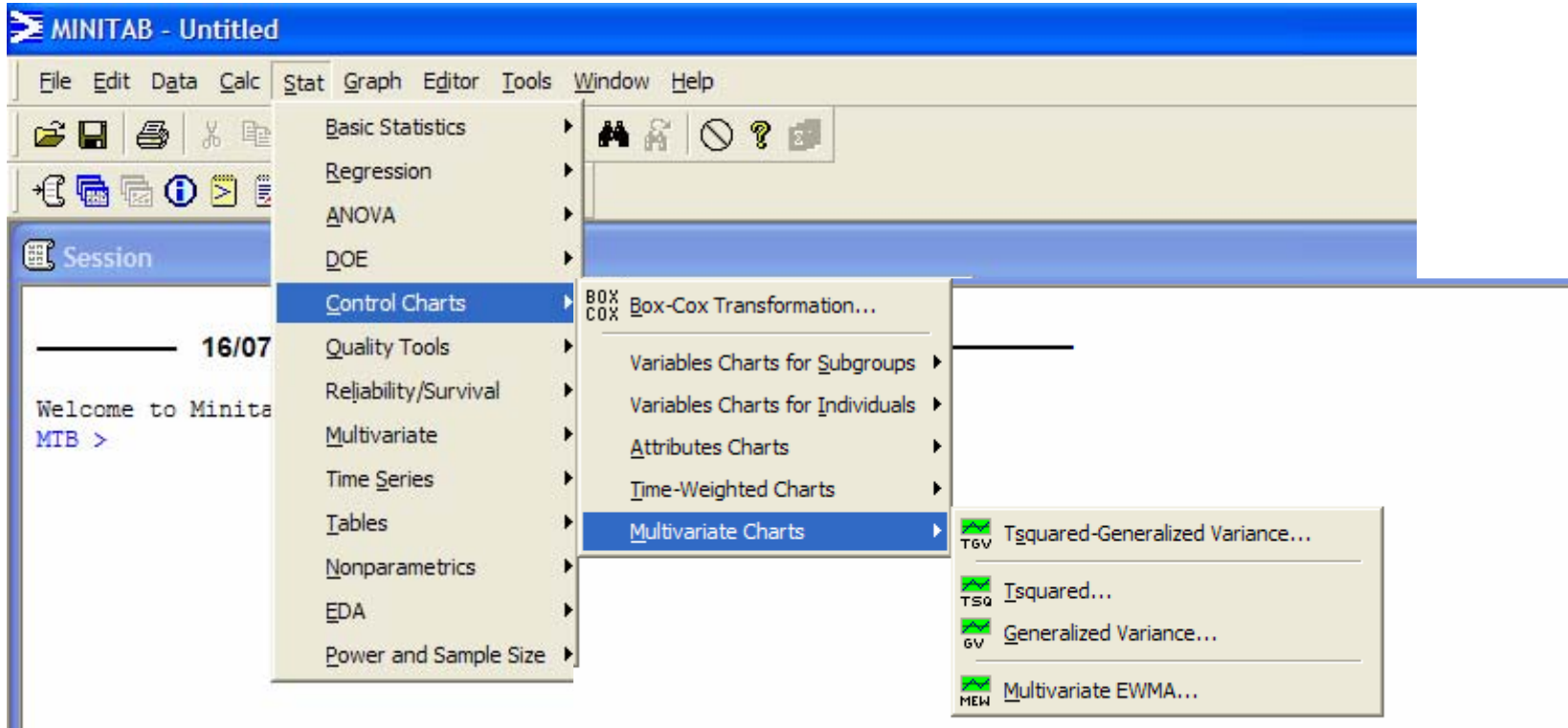
$$w = 1 \quad \Rightarrow \quad \bar{X}_k = X_k \quad \text{ie. no history – only current observation}$$

- Multiple variables &/or multiple locations

Multivariate analog based on:

$$T^2 = (X - \bar{X})^T S_X^{-1} (X - \bar{X})$$

- Predicated on normally-distributed responses
- Integrity compromised by:
  - Non-normality
  - Heterogeneous error structures
  - Non-stationarity



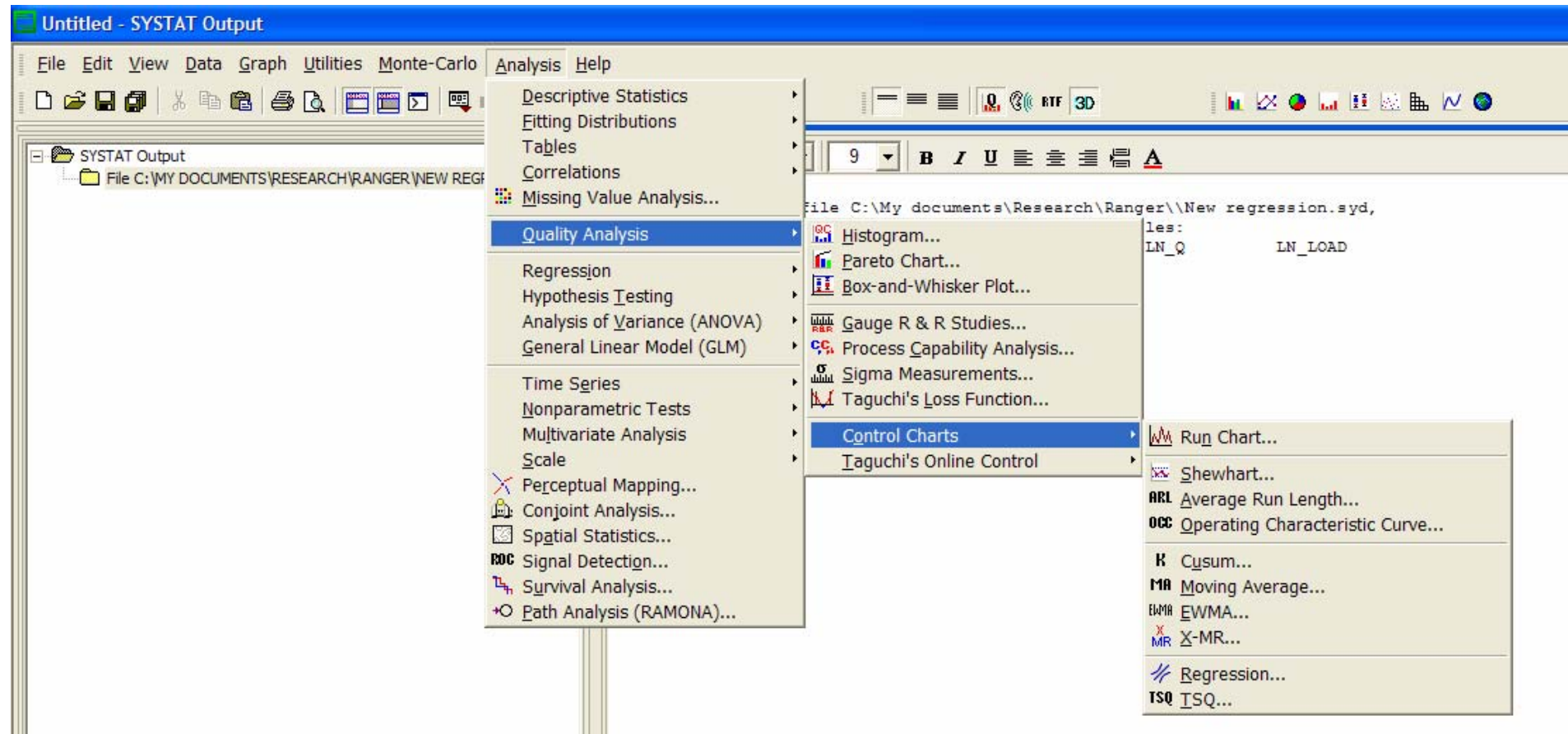
MINITAB - Untitled

File Edit Data Calc Stat Graph Editor Tools Window Help

Basic Statistics  
 Regression  
 ANOVA  
 DOE  
**Control Charts**  
 Quality Tools  
 Reliability/Survival  
 Multivariate  
 Time Series  
 Tables  
 Nonparametrics  
 EDA  
 Power and Sample Size

Box-Cox Transformation...  
 Variables Charts for Subgroups  
 Variables Charts for Individuals  
 Attributes Charts  
 Time-Weighted Charts  
**Multivariate Charts**  
 TGV Tsquared-Generalized Variance...  
 TSQ Tsquared...  
 GV Generalized Variance...  
 MEW Multivariate EWMA...

Session  
 16/07  
 Welcome to Minitab  
 MTB >



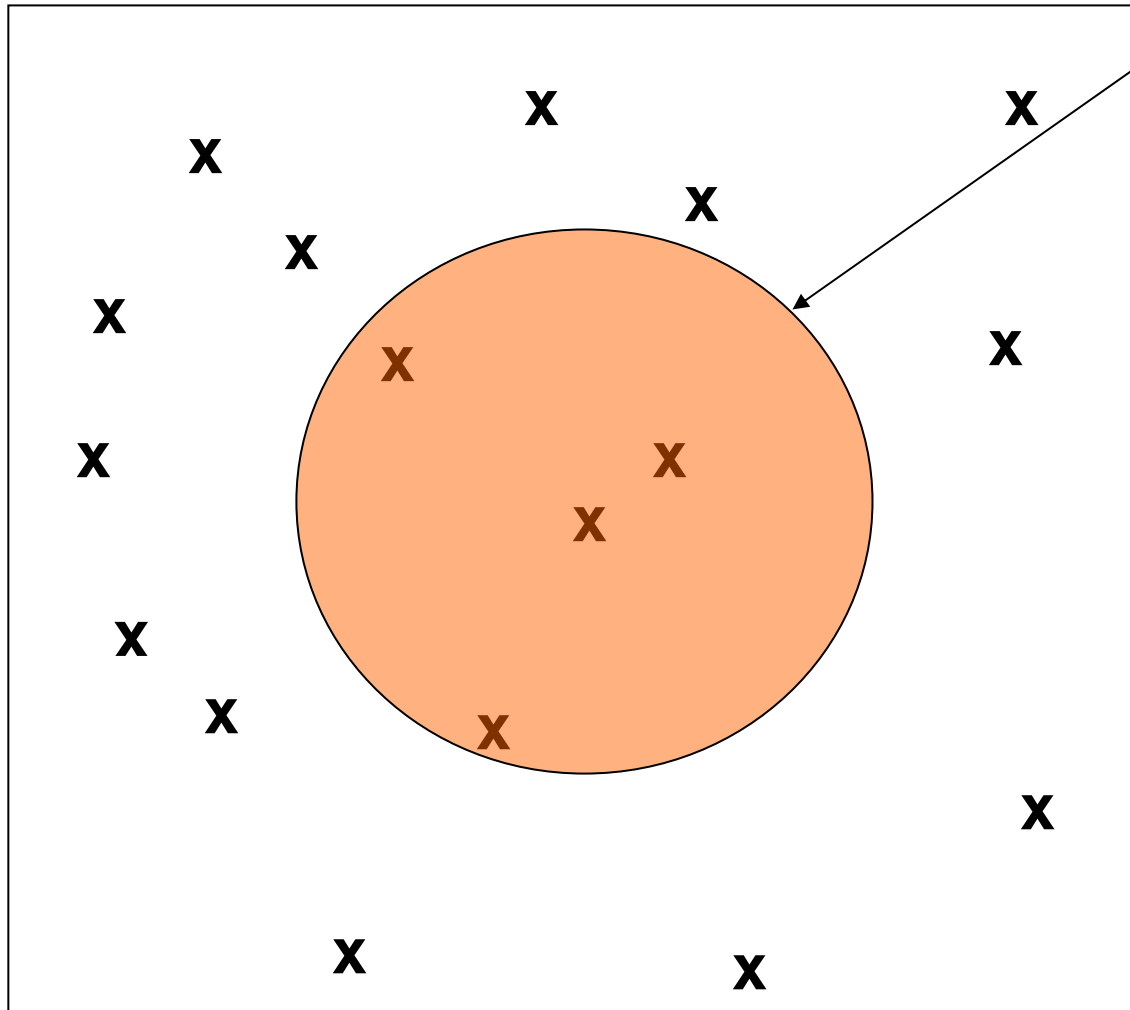
The screenshot shows the SYSTAT software interface. The title bar reads "Untitled - SYSTAT Output". The menu bar includes File, Edit, View, Data, Graph, Utilities, Monte-Carlo, Analysis, and Help. The Analysis menu is open, showing options such as Descriptive Statistics, Fitting Distributions, Tables, Correlations, Missing Value Analysis..., Quality Analysis (highlighted), Regression, Hypothesis Testing, Analysis of Variance (ANOVA), General Linear Model (GLM), Time Series, Nonparametric Tests, Multivariate Analysis, Scale, Perceptual Mapping..., Conjoint Analysis..., Spatial Statistics..., Signal Detection..., Survival Analysis..., and Path Analysis (RAMONA).... The Quality Analysis sub-menu is also open, listing Histogram..., Pareto Chart..., Box-and-Whisker Plot..., Gauge R & R Studies..., Process Capability Analysis..., Sigma Measurements..., Taguchi's Loss Function..., Control Charts (highlighted), and Taguchi's Online Control. The Control Charts sub-menu is further expanded, showing Run Chart..., Shewhart..., ARL Average Run Length..., OCC Operating Characteristic Curve..., K Cusum..., MA Moving Average..., EWMA EWMA..., X-MR..., Regression..., and ISQ... The main window displays a file path "File C:\My documents\Research\Ranger\New regression.syd," and a list of variables: "les:", "LN\_Q", and "LN\_LOAD".

## Scan Statistics

- A tool for detecting clusters in a spatial point process
- Move a window  $[t, t+w]$  of size  $w < b-a$  over a time interval  $[a, b]$
- Over all possible values of  $t$ , record the maximum number of points in the window
- Compare this number with cut off points under the assumption of a purely Poisson Process

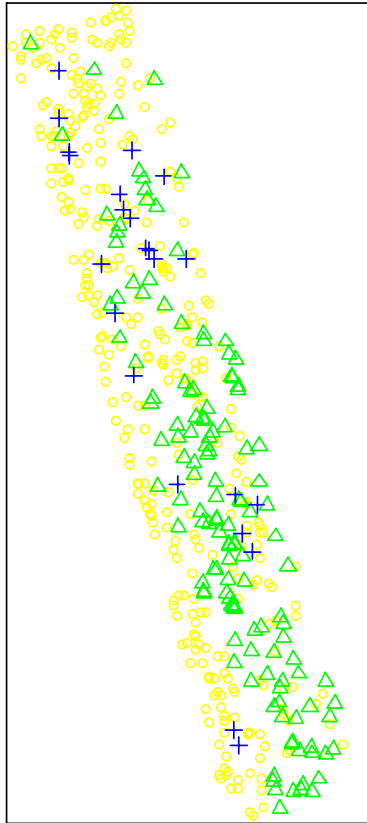
# Measures of anomalous behaviour in space – the scan statistic

## Surveillance Region



**Candidate cluster:**  
The scan statistic gives a measure of: “how unlikely is the number of cases *inside* relative to the number *outside*, given the expected spatial distribution of cases”

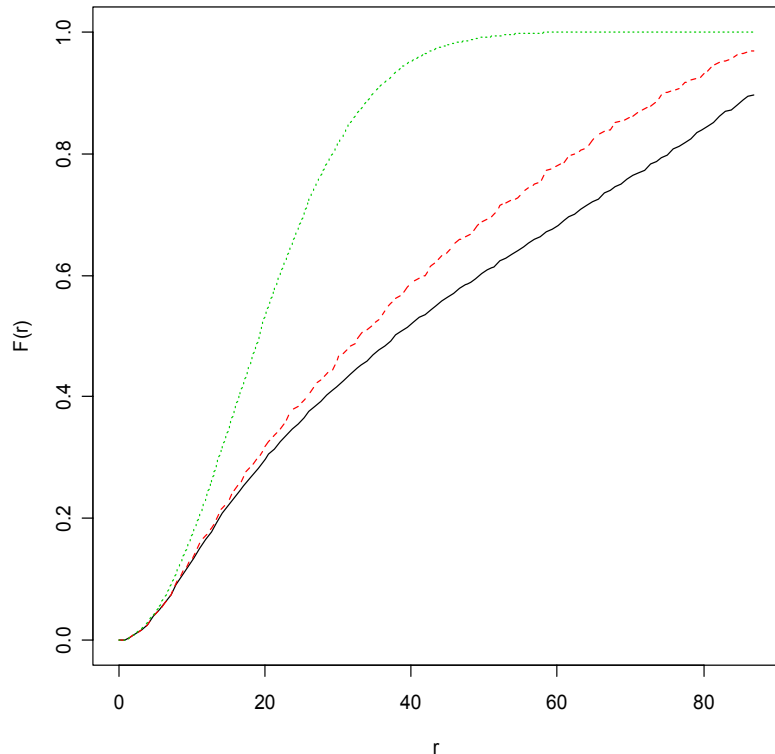
ma.ppp



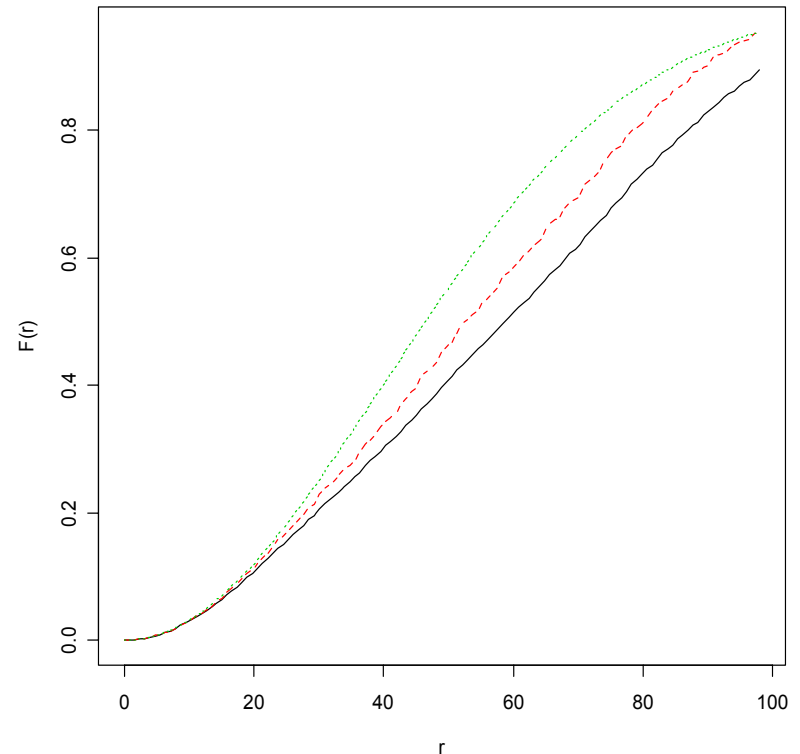
Plot of locations of the three cover classes.

# Measures of anomalous behaviour in space – the F, G, K, J functions

F statistic is measure of non-randomness of inter-point gaps



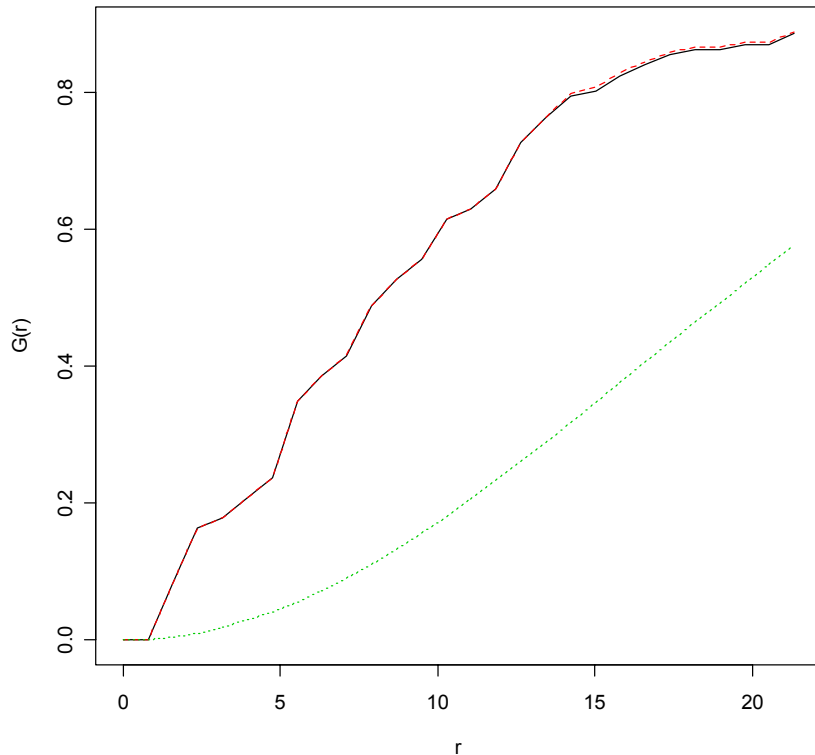
F-function plot for macroalga #2.



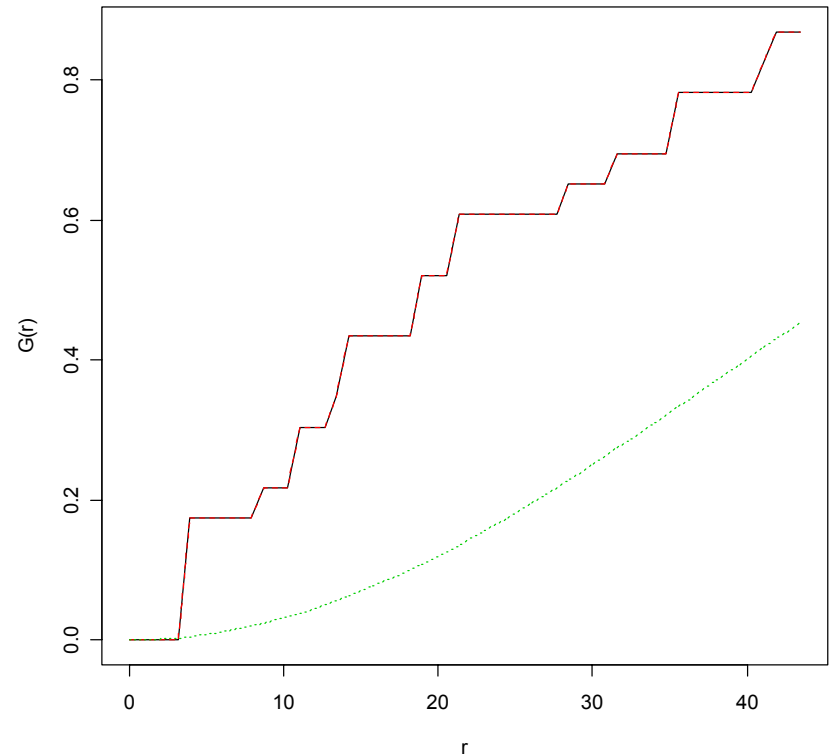
F-function plot for macroalga #3.

Black line = F; green dotted line=theoretical

G statistic is measure of clustering of pairs of points separated by distance  $r$



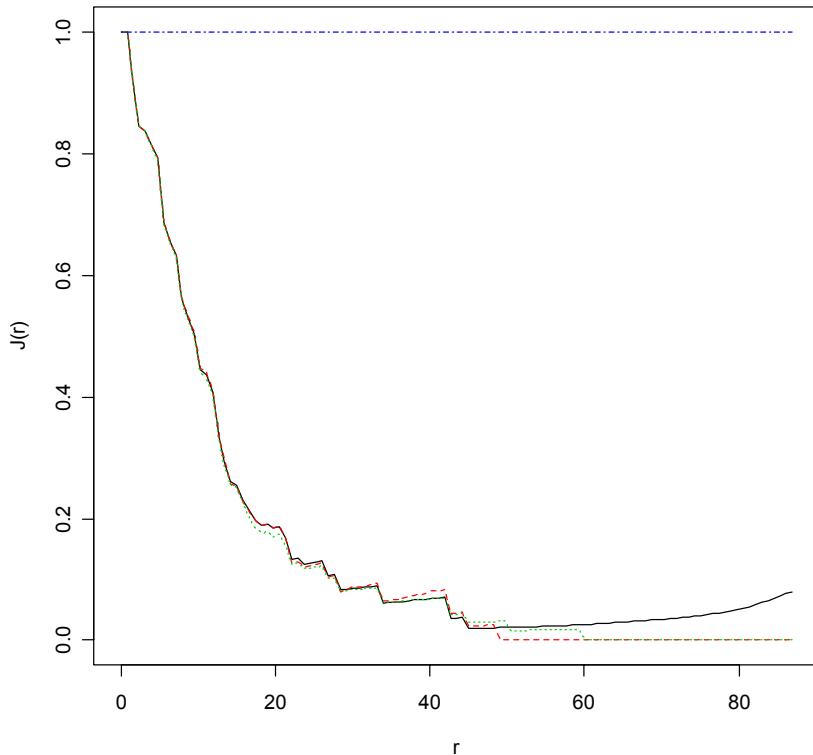
G-function plot for macroalga #2.



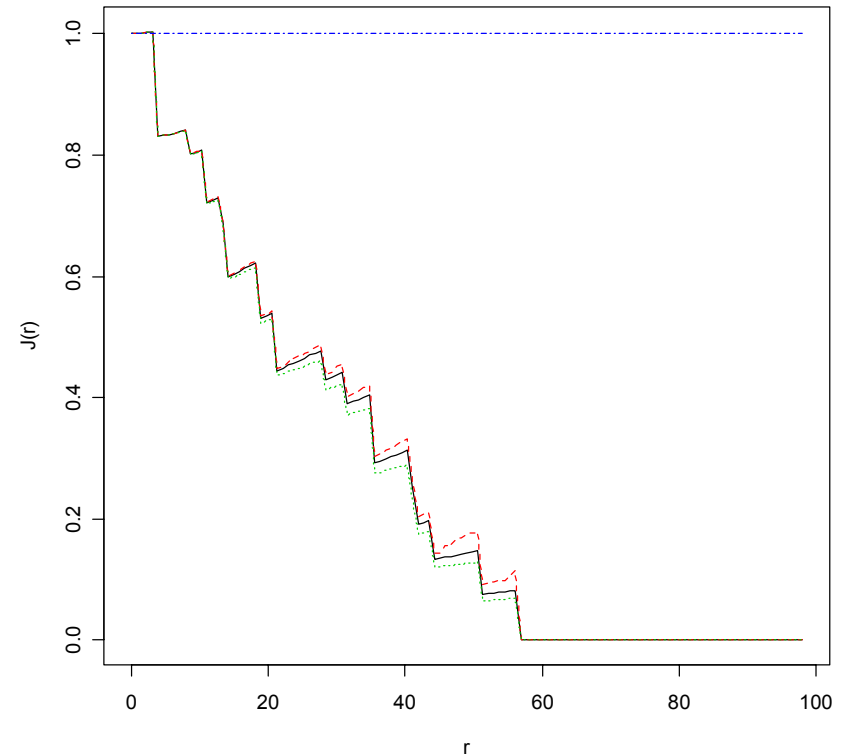
G-function plot for macroalga #3.

Black line = F; green dotted line=theoretical

J statistic is comparison (ratio) of F and G (= 1 for Poisson process)



J-function plot for macroalga #2.

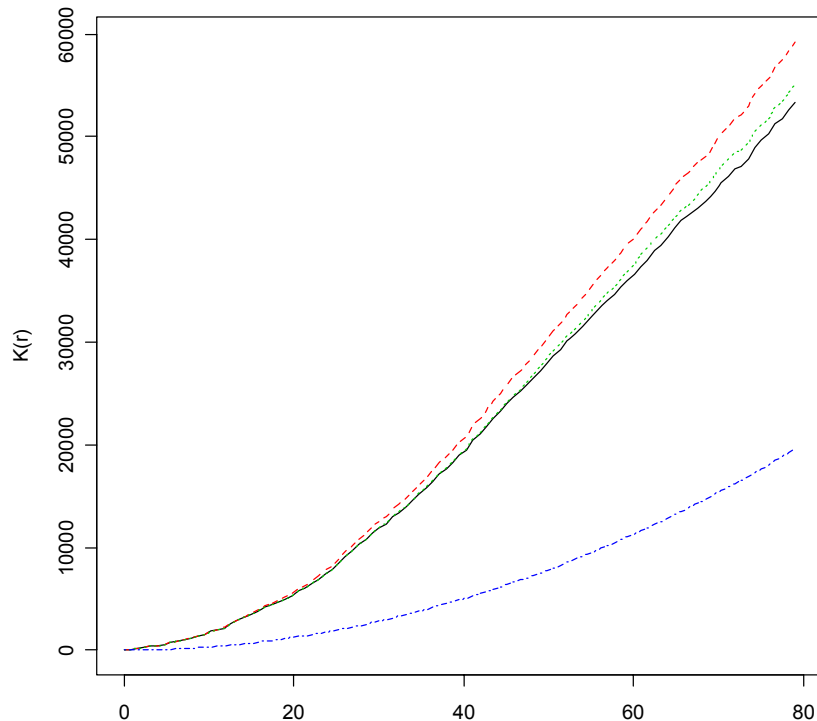


J-function plot for macroalga #3.

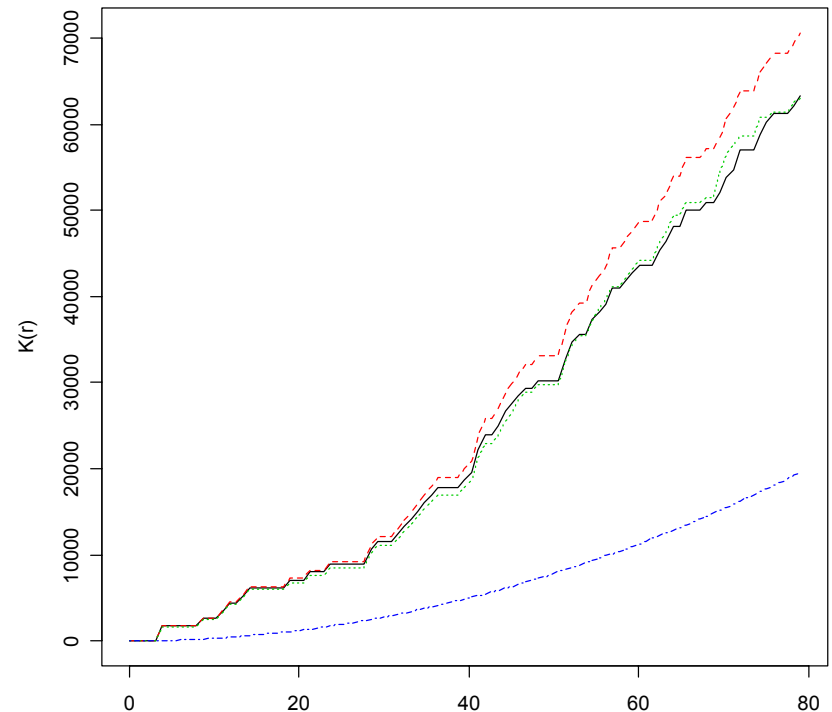
Black line = F; green dotted line=theoretical

# Measures of anomalous behaviour in space – the F, G, K, J functions

K statistic is a measure of association (spatial dependency) for pairs of points at distance  $r$



**K-function plot for macroalga #2.**



**K-function plot for macroalga #3.**

Black line = F; green dotted line=theoretical

**Biosurveillance** – monitoring with the aim of detecting the outbreak of an epidemic.

- Traditional approach – collect and analyse medical / public health data → verify existence of the disease
- Modern biosurveillance – based on notion of “syndromic” data → early detection

## Example

- Over the counter (OTC) medication sales;
- Calls to medical / nurse hotlines;
- School attendance records;
- Presenting illnesses at hospital emergency rooms

## Issue

- Impact of disease or bio-terror attack on syndromic data is unknown;
  - eg. how would anthrax attack manifest itself in ambulance dispatches or in sales of cough medicines?

- Precise definition of what constitutes ‘anomalous behaviour’ / ‘normal’ behaviour?
- Role of covariates
- Accounting for short-term trends / cyclic features
- Model adequacy – how do we know? Role of conventional goodness-of-fit statistics (eg.  $R^2$ )
- Avoiding excessive false triggering
- What to do after a trigger has been tripped?

## Fundamental requirements of a surveillance system

Sensitivity ↔ statistical power

Specificity ↔  $1 - \alpha$

Timeliness ↔ change-point detection

# Other approaches

- Kalman filtering
- Wavelets
- Bayesian techniques