UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Spatial cluster detection

Geoffrey T. Fosgate

BS, DVM, PhD, DACVPM
Professor
Department of Production Animal Studies
University of Pretoria
Co-Editor-in-Chief for *Preventive Veterinary Medicine*

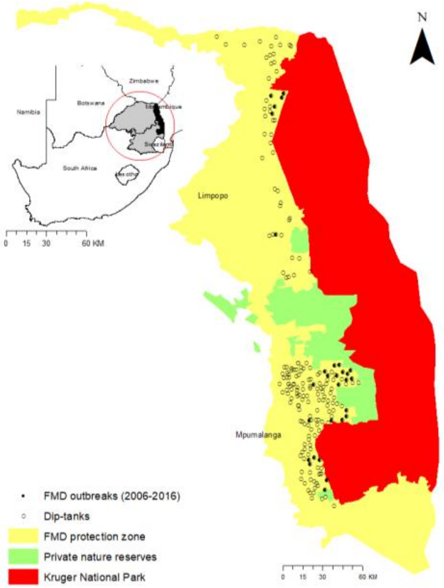
GF-TADs
Foot and Mouth Disease
Risk Assessment Training Workshop
Johannesburg, South Africa 19-21 September 2023

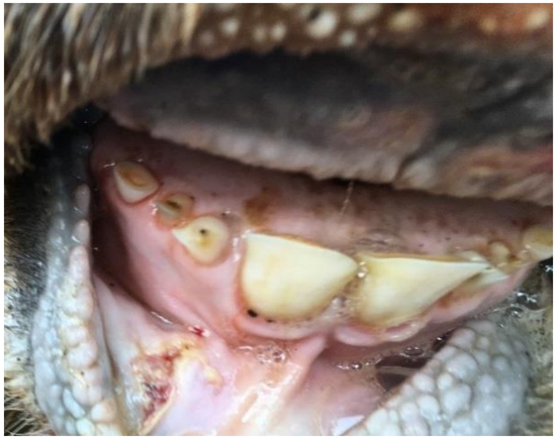
World Organisation for Animal Health
Founded as OIE

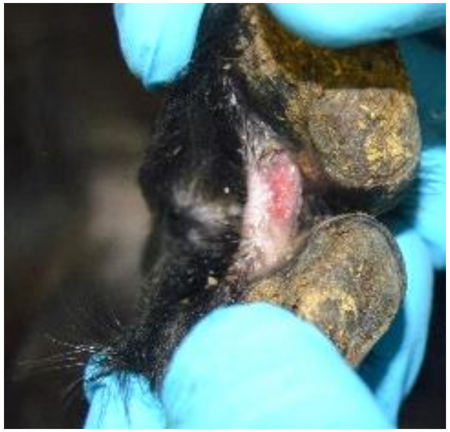
Spatial epidemiology

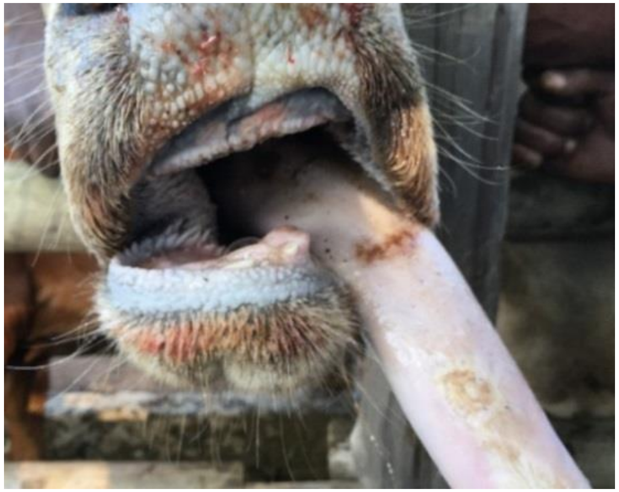
- Introduction to spatial epidemiology
- Cluster definition
- Descriptive presentation
  - Spot maps
  - Choropleth maps
- Risk mapping
  - Inverse distance weighting
  - Kriging
- Cluster detection
  - Temporal
  - Spatial
  - Temporospatial

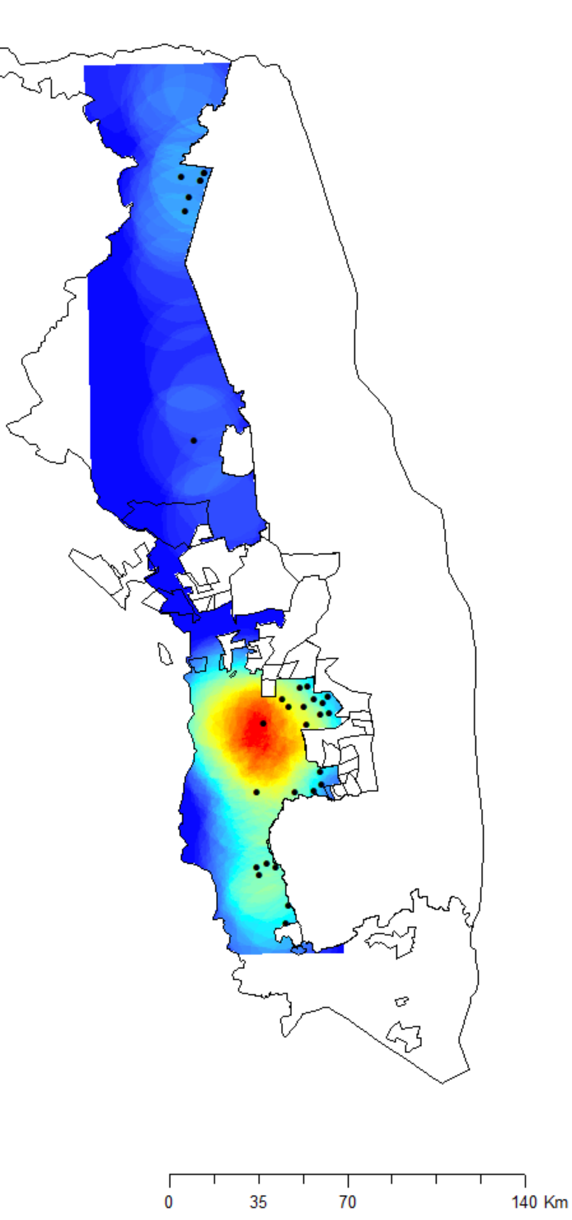
Spatial epidemiology











Spatial epidemiology

- The study of the spatial distribution of health-related states and health determinants in populations
- Spatial epidemiology provides a framework to examine the influences of space and place on health
  - Describe and analyze patterns of disease
  - Explore and analyze spatial patterns
  - Hypothesize about possible causal relationships
- Methods
  - Descriptive disease mapping
  - Risk mapping
  - Cluster detection
- Place can be used as surrogates for influences on disease
  - Exposure to environmental hazards
  - Animal movement networks
  - Management factors

Spatial epidemiology

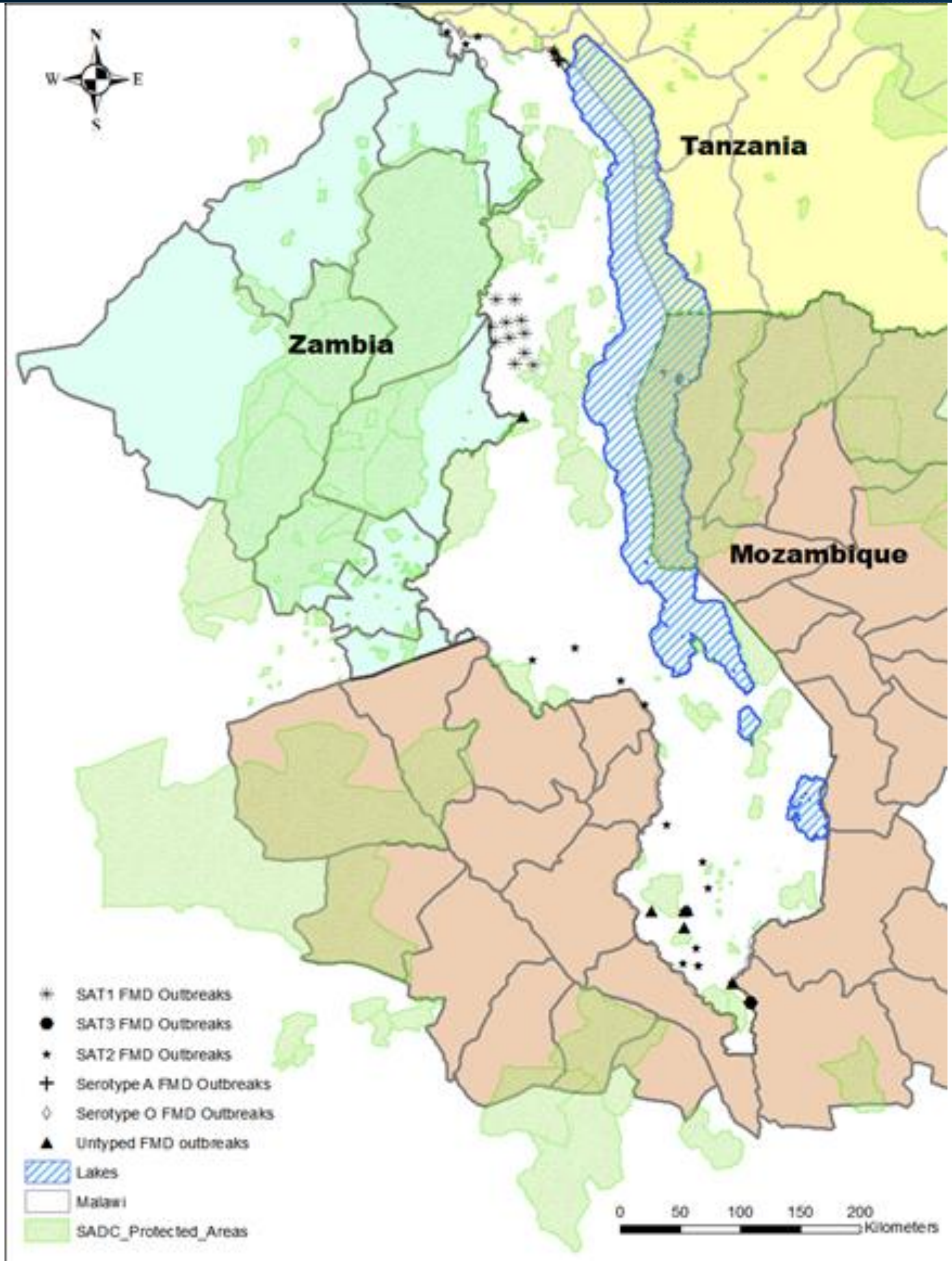
- An epidemic is an increase, often sudden, in the number of cases of a disease above what is normally expected in the population of that area
- An outbreak is also a sudden rise in the incidence of disease but is often used for limited geographical distributions
- A cluster is an aggregation of cases in place and time that are greater than expected

https://madison.com/ct/news/opinion/column/cartoons-of-the-week/collection_d62e61e8-61ff-11e9-8ac9-27446eda148d.html/

World Organisation for Animal Health
Founded as OIE

100 1908 - 2008
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science

Descriptive presentation

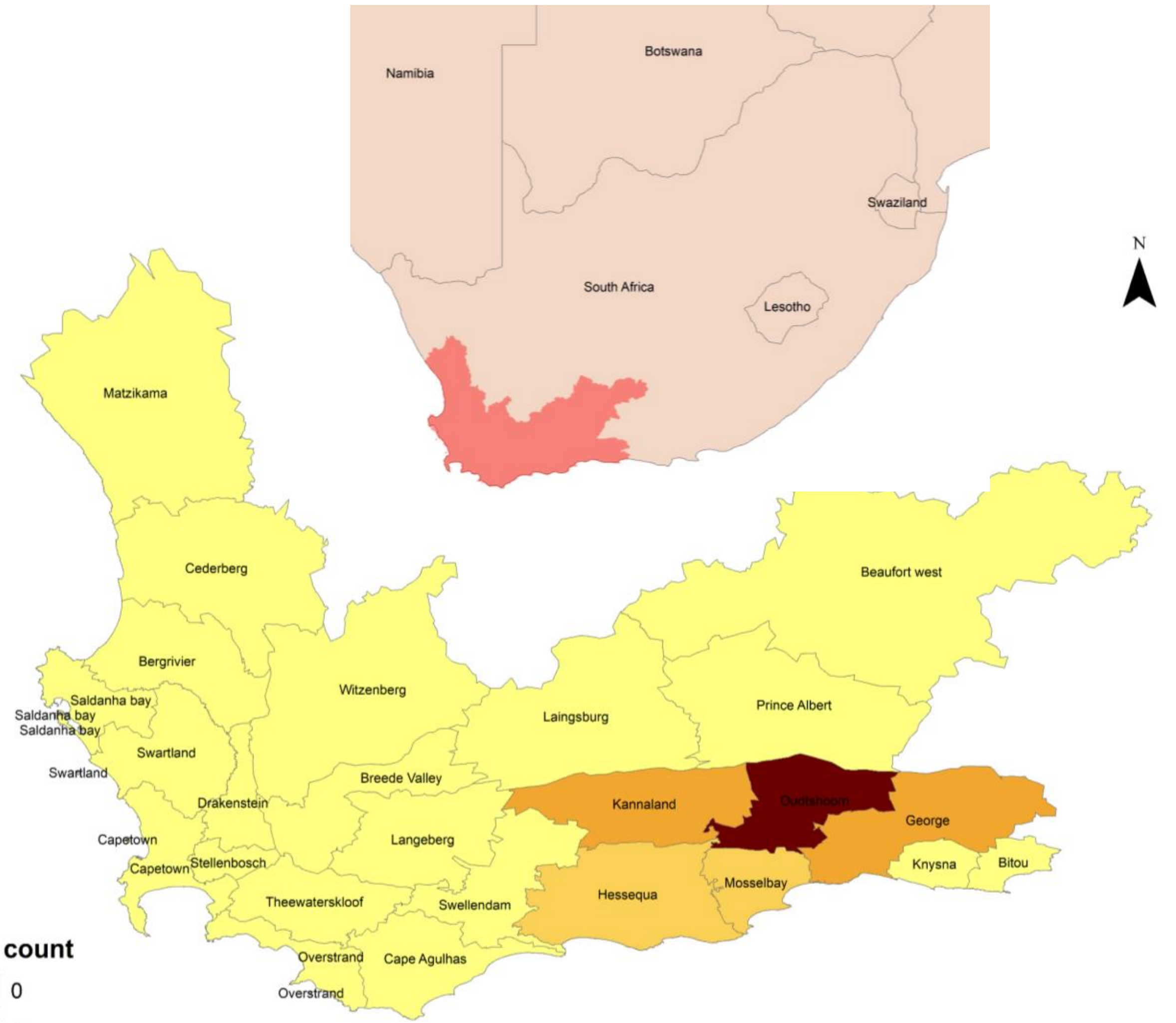


- **Spot maps provide a simple distribution of case reports**
- **Do not account for underlying population at risk**
- **Can give an indication of high-risk areas and possible risk factors**

Chimera ET, Fosgate GT, Etter EMC, Jemberu WT, Kamwendo G, Njoka P. Spatio-temporal patterns and risk factors of foot-and-mouth disease in Malawi between 1957 and 2019. *Prev Vet Med* 2022;204:105639.

Descriptive presentation

- Choropleth maps are aggregated for geopolitical regions
- When individual locations are not available
- Demarcations are arbitrary and unrelated to epidemiological factors
- Could account for population at risk within area units



Case count

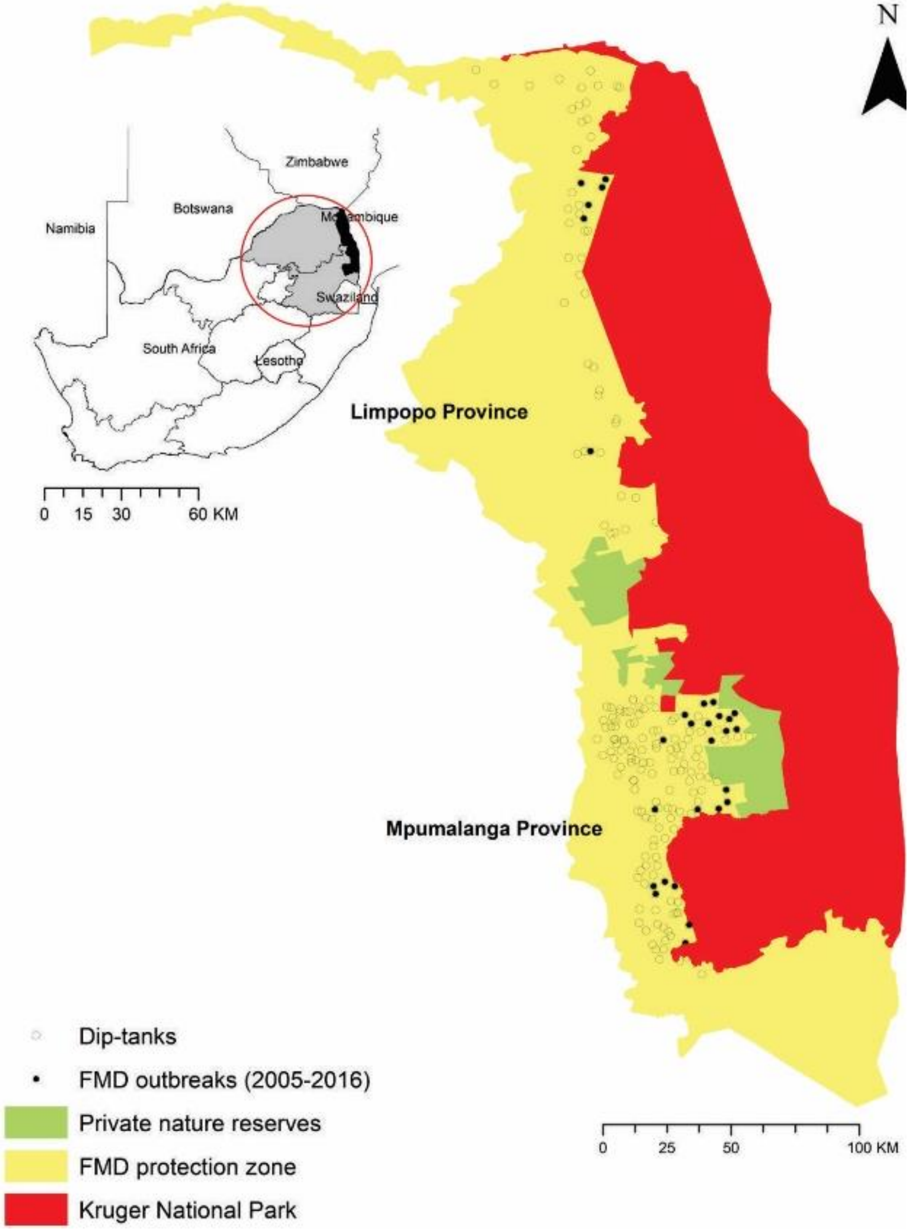
🟨 0
🟨 1
🟧 2 - 4
🟫 5 - 10
🟥 11 - 35

Marimwe MC, Fosgate GT, Roberts LC, Tavornpanich S, Olivier AJ, Abolnik C. The spatiotemporal epidemiology of high pathogenicity avian influenza outbreaks in key ostrich producing areas of South Africa. *Prev Vet Med* 2021;196:105474.

World Organisation for Animal Health
Founded as OIE

Descriptive presentation

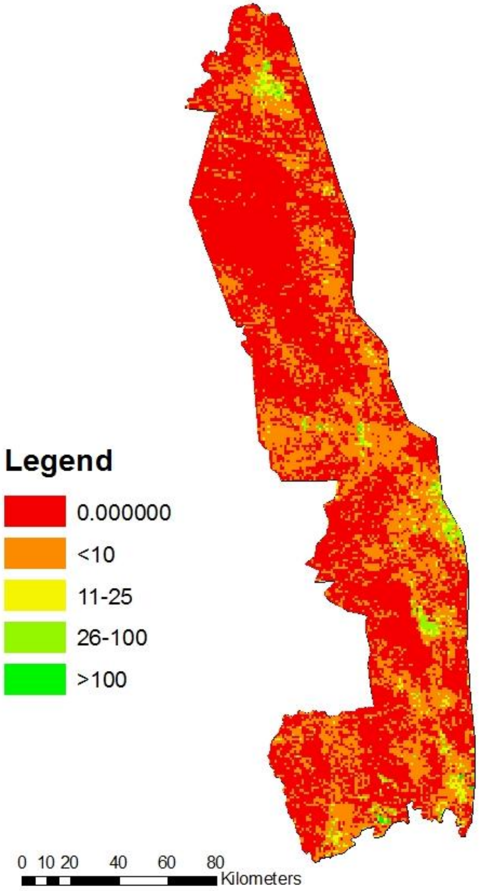
- Proximity to specific features can be evaluated
- Is the simplest form of spatial exposure assessment
- Assumes all individuals within a specific distance to a source have the same exposure
- Commonly used for measuring access to resources and exposure to environmental hazards
- Distance to the disease control fence



Sirdar MM, Fosgate GT, Blignaut B, Mampane RL, Rikhotso O, Du Plessis B, Gummow B. Spatial distribution of foot-and-mouth disease (FMD) outbreaks in South Africa (2005-2016). *Trop Anim Health Prod* 2021;53:376.

Risk mapping

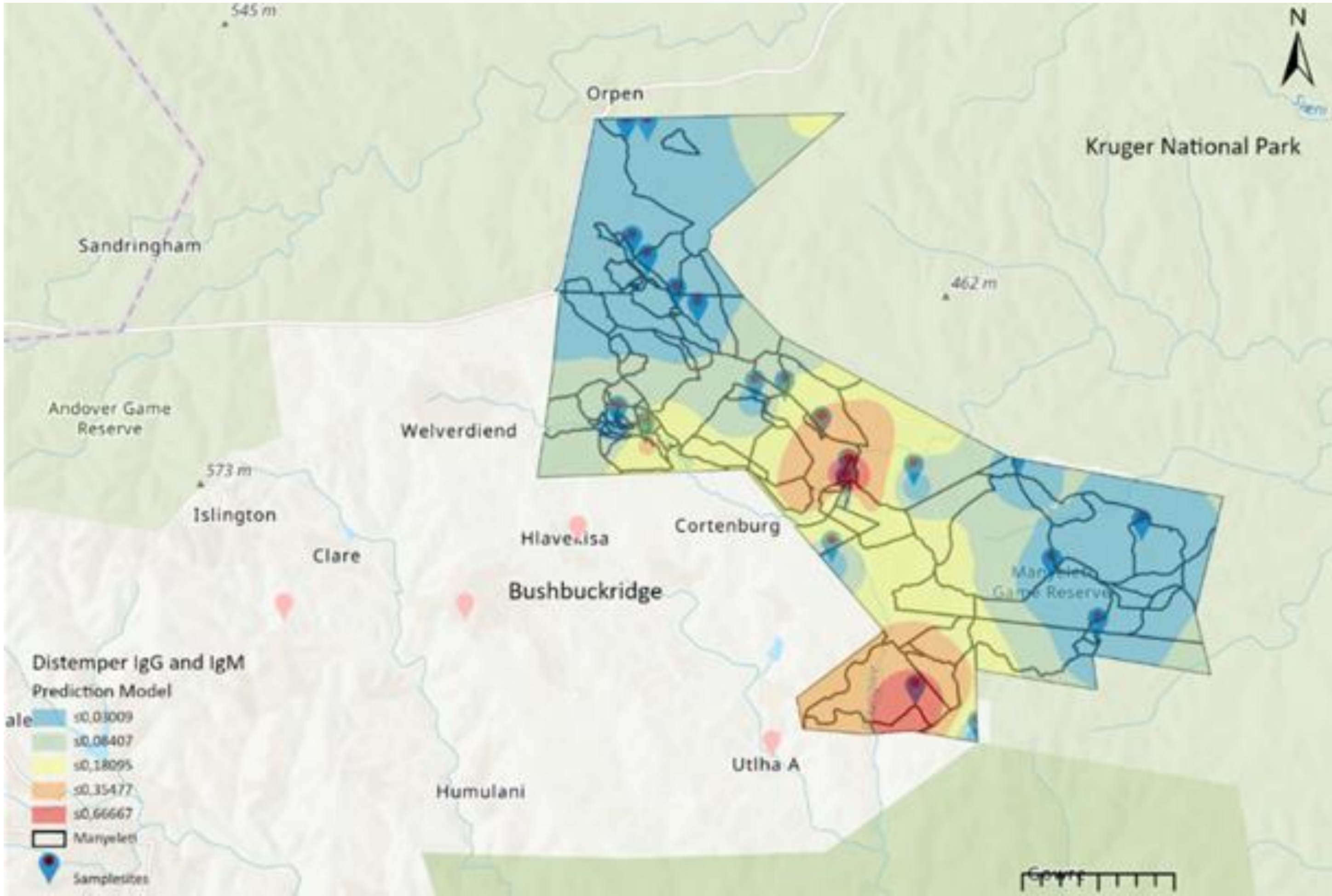
- Spatial interpolation is used to estimate a value of a variable at an un-sampled location from measurements made at other sites
- Spatial interpolation is based on the notion that points which are close together in space tend to have similar attributes
- Many different methods available:
  - Exact or approximate
  - Deterministic or geostatistical
  - Local or global
  - Gradual or abrupt



Hughes K, Fosgate GT, Budke CM, Ward MP, Kerry R, Ingram B. Modeling the spatial distribution of African buffalo (*Syncerus caffer*) in the Kruger National Park, South Africa. *PLOS ONE* 2017;12:e0182903.

Risk mapping

- **Deterministic techniques**
  - **Polynomial interpolation**
  - **Inverse distance weighting**

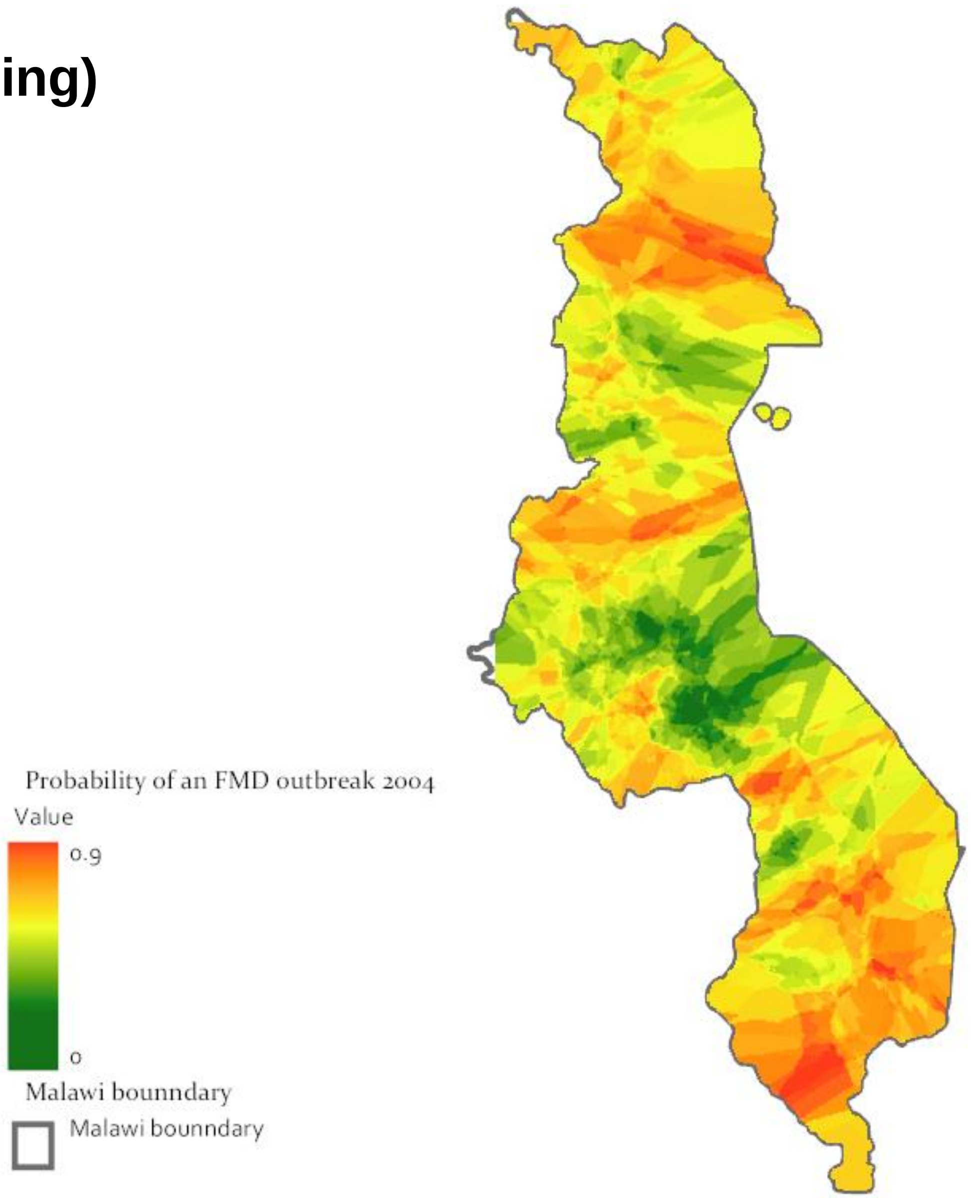


Risk mapping

- **Geostatistical approaches (Kriging)**
  - **Ordinary kriging**
  - **Simple kriging**
  - **Universal kriging**
  - **Empirical Bayesian Kriging**

$$d(x,y) = \sum_{i=1}^{n} w_i d_i$$

$$\gamma(h) \ = \Sigma\ (z(x) - z(x+h))^2/2n$$



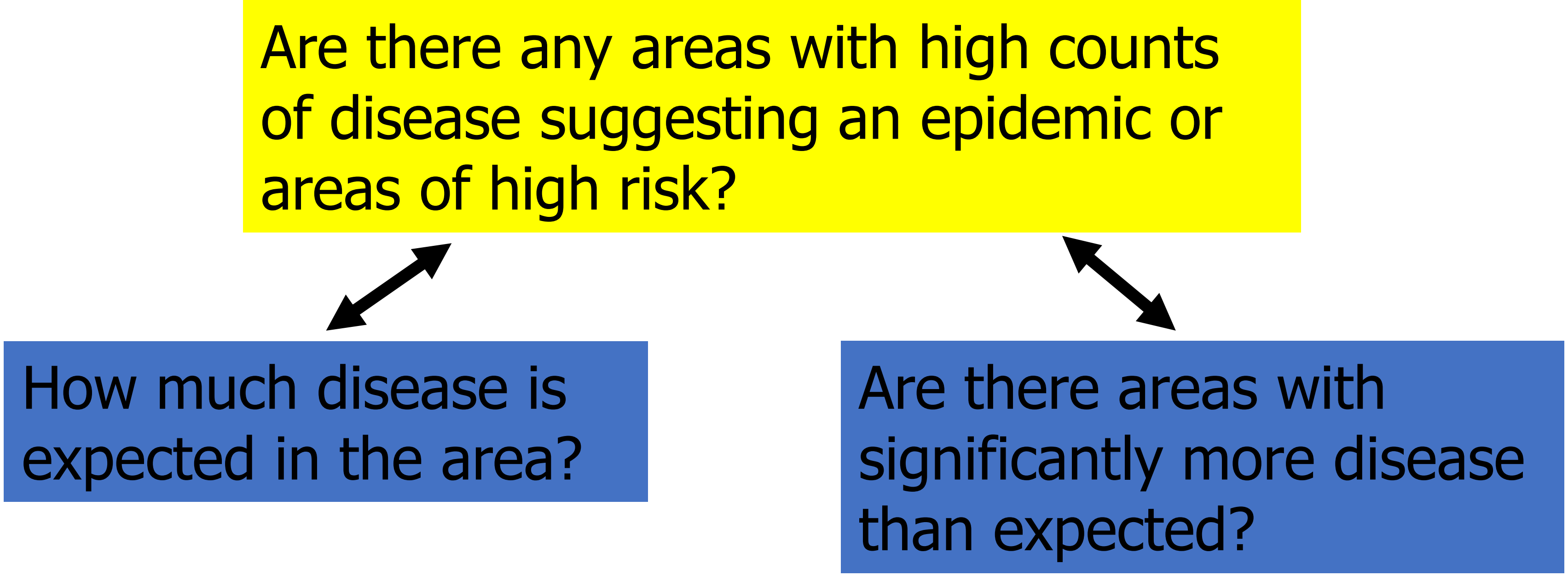
Chimera ET, Fosgate GT, Etter EMC, Jemberu WT, Kamwendo G, Njoka P. Spatio-temporal patterns and risk factors of foot-and-mouth disease in Malawi between 1957 and 2019. *Prev Vet Med* 2022;204:105639.

Cluster detection

- Clusters are geographically and/or temporally bounded groups of occurrences of sufficient size and concentration unlikely to have occurred by chance
- Clusters are either related to each other through some social or biological mechanism or they have a common relationship with some other event or circumstance
- Animals with similar characteristics tend to aggregate and their shared characteristics explain in part the disease and place association
- Environmental attributes influence whole groups and affect disease over and above aggregate individual characteristics

Cluster detection

- Identify the locations, shapes, and sizes of potentially anomalous spatial regions
- Determine whether each of these potential clusters is more likely to be a “true” cluster or a chance occurrence
- Is anything unexpected going on, and if so, then where?



Are there any areas with high counts of disease suggesting an epidemic or areas of high risk?

How much disease is expected in the area?

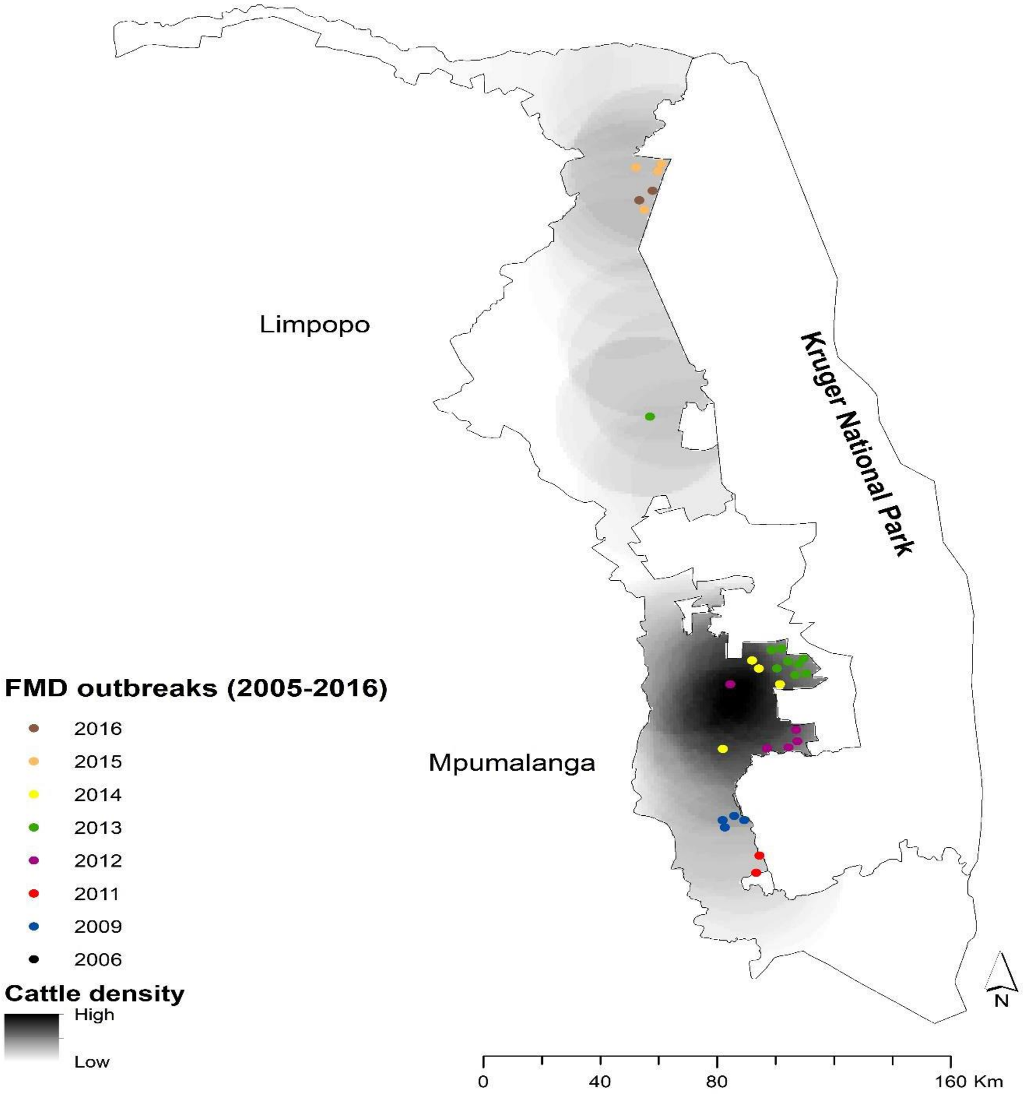
Are there areas with significantly more disease than expected?

Cluster detection

- Global cluster tests search for spatial clusters anywhere in a study area but do not necessarily identify where the clusters occur, and are used to identify departures from spatial randomness when overall spatial pattern is considered
- Local cluster tests identify locations at which there is some excess/deficit—a hot/cold spot—anywhere within a study area
- Temporal-only data
- Spatial-only data
- Case-only data
- Case-control (or cross-sectional) data
- Continuous predictors
- Combined spatial and temporal analysis

Does FMD cluster?

Sirdar MM, Fosgate GT, Blignaut B, Mampane RL, Rikhotso O, Du Plessis B, Gummow B. Spatial distribution of foot-and-mouth disease (FMD) outbreaks in South Africa (2005-2016). *Trop Anim Health Prod* 2021;53:376.

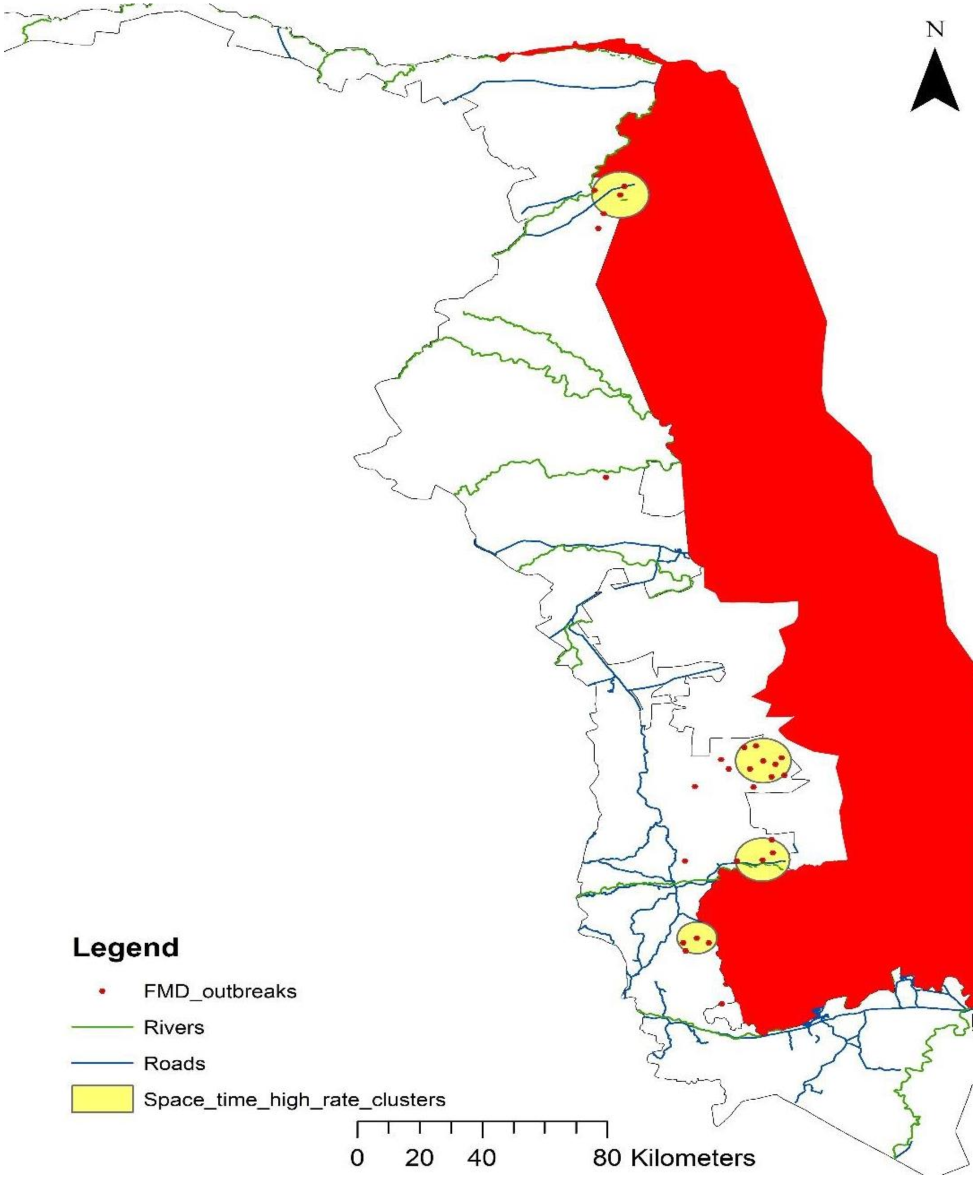


World Organisation for Animal Health
Founded as OIE

Universiteit van Pretoria
University of Pretoria
Yunibesithi ya Pretoria
Faculty of Veterinary Science

Does FMD cluster?

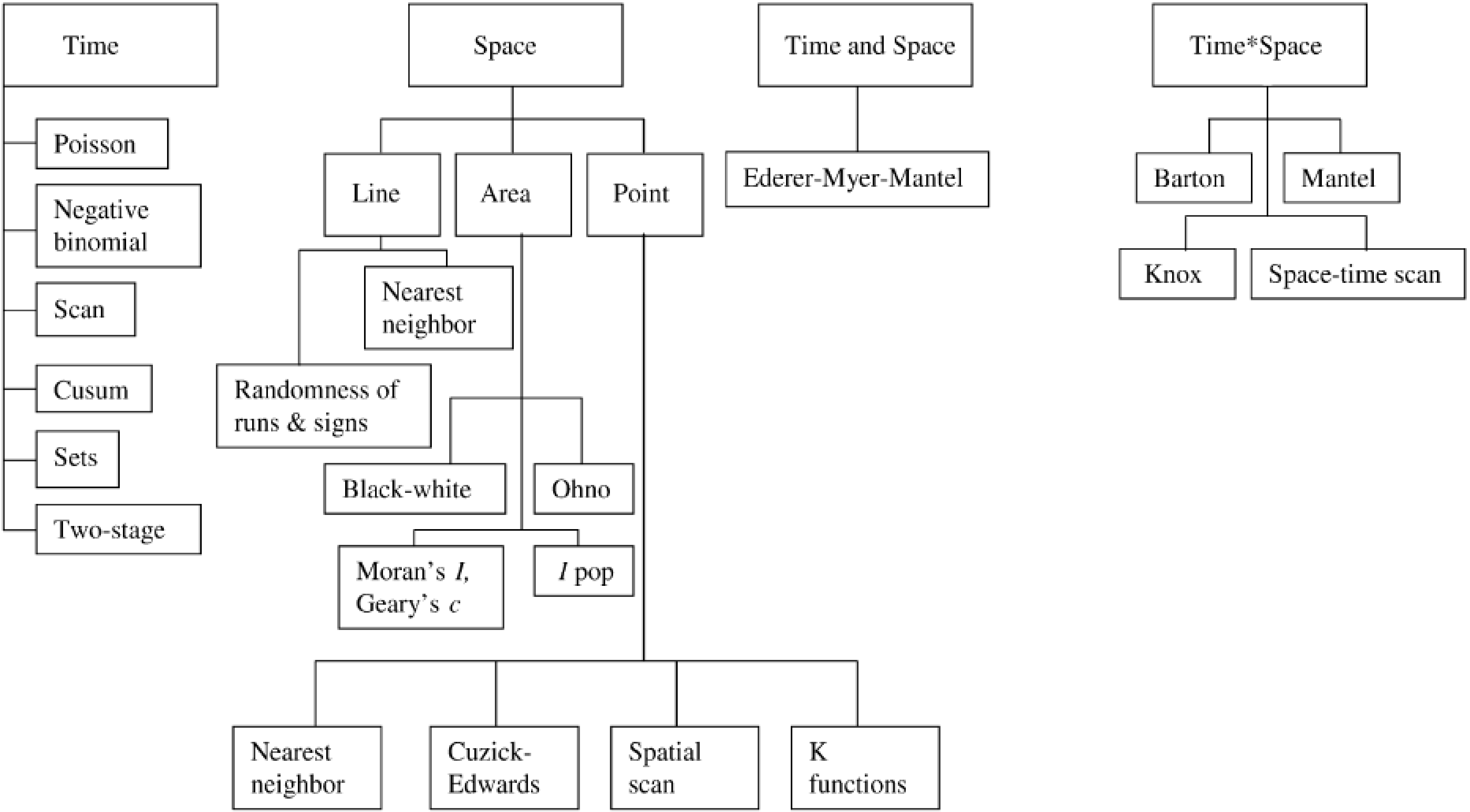
Sirdar MM, Fosgate GT, Blignaut B, Mampane RL, Rikhotso O, Du Plessis B, Gummow B. Spatial distribution of foot-and-mouth disease (FMD) outbreaks in South Africa (2005-2016). *Trop Anim Health Prod* 2021;53:376.



World Organisation for Animal Health
Founded as OIE

100
1908 - 2008
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science

Cluster detection



World Organisation for Animal Health
Founded as OIE

Carpenter TE. Methods to investigate spatial and temporal clustering in veterinary epidemiology. *Prev Vet Med* 2001;48:303-320.

100
1908 - 2008
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science

Cluster detection

Spatial		
Global	Local	Temporospatial
Ripley's K function	Kulldurff's scan statistic	Ederer, Myers, and Mantel (EMM) test
Cuzick-Edwards test	LISA	Kulldurff's scan statistic
Moran's I		
Ipop		

Global spatial clustering

- Ripley’s k-function
  - Analyzes point data related to distances between affected and unaffected locations
  - Assesses clustered and dispersed distributions
- Cuzick-Edwards test
  - Analyzes point data related to cases and controls
  - Identifies the nearest neighbor rather than actual distances
  - Can assess clustering and dispersed distributions
- Moran’s I
  - Analyzes point or areal data for quantitative outcomes
  - Can assess for clustered or dispersed distributions
- Ipop
  - Modification of Moran’s I to account for the population at risk

Ripley's K function

- **Used to analyze the spatial pattern of incident point data**
- **Summarizes spatial dependence (clustering or dispersion) over a range of distances**
- **Ripley's K-function can be used to assess how the spatial clustering or dispersion changes when the neighborhood size changes**

$$\widehat{K}(t) = \lambda^{-1} \sum_{i \neq j} \frac{I(d_{ij} < t)}{n}$$

- **Where d_{ij} is the Euclidean distance between the ith and jth points in a data set of n points, t is the search radius, λ is the average density of points (generally estimated as n/A, where A is the area of the region containing all points) and I is the indicator function (1 if its operand is true, 0 otherwise)**

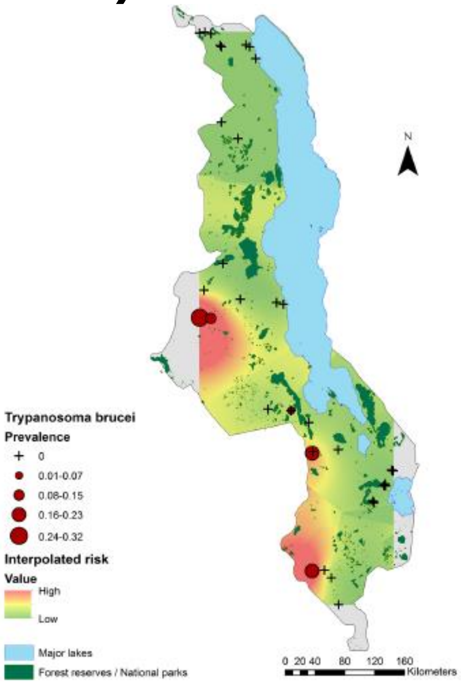
Cuzick-Edward's test

- The total number of case-case pairs are summed and compared to the expected number based on a hypergeometric distribution

- $E[m] = np = n_1(n_1-1)/n(n-1)$

- p is the probability of a case/case pair occurring (nearest neighbors)
- n is the total number of observations (cases and controls)
- n_1 is the total number of cases

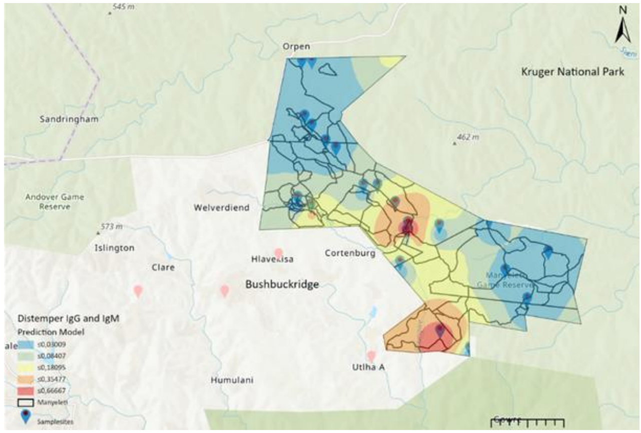
- Test statistic is a typical Z-test
- $Z = (m+0.5-E[m])/\sqrt{(np(1-p))}$



Chimera ET, Fosgate GT, Etter EMC, Boulangé A, Vorster I, Neves L. A One Health investigation of pathogenic trypanosomes of cattle in Malawi. *Prev Vet Med* 2021;188:105256.

Moran’s I

- **Used for the evaluation of continuous data**



Moran’s (I) coefficient

$$= \frac{n\sum(x_i - \bar{x})(x_j - \bar{x})}{J\sum(x - \bar{x})^2}$$

n = number of areas under study

J = total number of adjacencies

x_i & x_j are adjacent area values (either side of link)

X = area value & $\bar{X}$ = mean of all values (areas)

E_I = - 1 / n - 1

$$\sigma_I = \sqrt{\frac{n\left[J(n^2+3-3n)+3J^2-n\sum L^2\right]-k\left[J(n^2-n)+6J^2-2n\sum L^2\right]}{J^2(n-1)(n-2)(n-3)}}$$

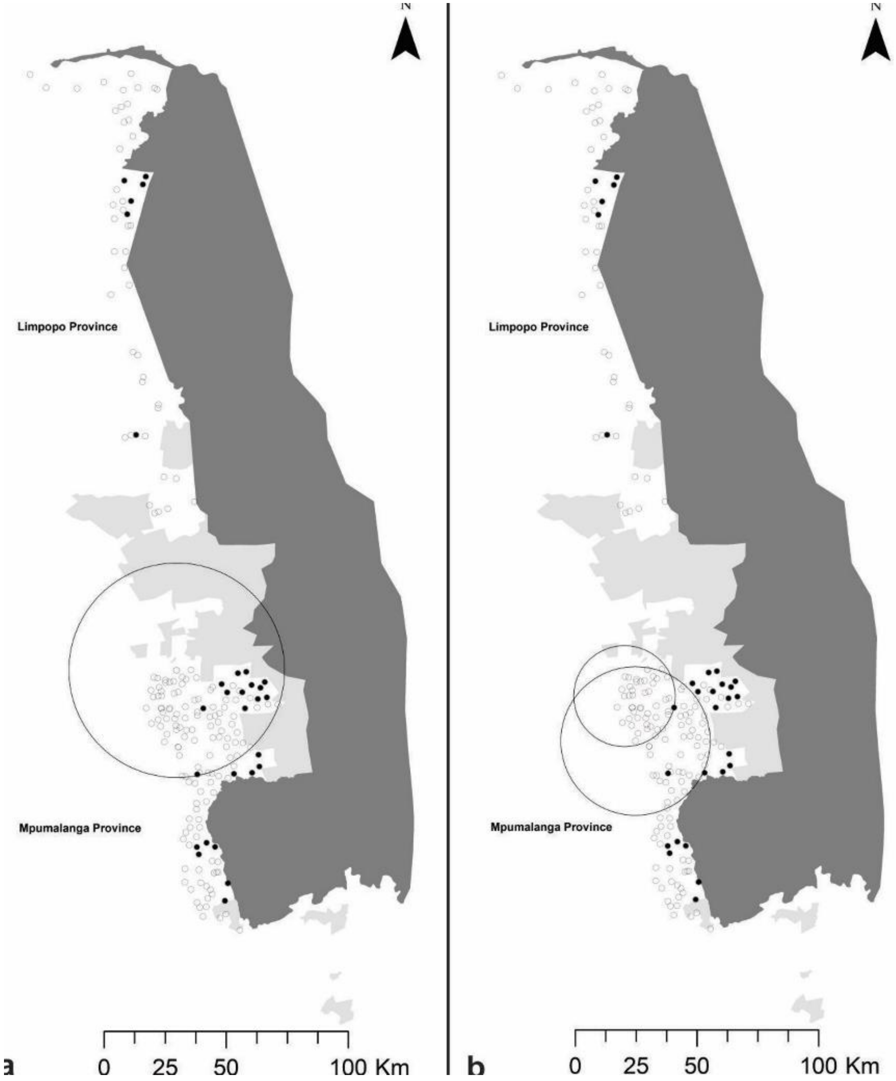
World Organisation for Animal Health
Founded as OIE

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37: 17-23.

100
1908 - 2008
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science

Local spatial clustering

- Spatial scan statistics
  - The approach can use multiple data types
  - Bernoulli or case/control data
  - Poisson for incidence rate data
- Local indicators of spatial autocorrelation (LISA)
  - Uses quantitative data
  - Modification of Moran’s I
  - Identifies the locations responsible for autocorrelation (clustering or dispersion)



Sirdar MM, Fosgate GT, Blignaut B, Mampane RL, Rikhotso O, Du Plessis B, Gummow B. Spatial distribution of foot-and-mouth disease (FMD) outbreaks in South Africa (2005-2016). *Trop Anim Health Prod* 2021;53:376.

Spatial scan

- Why use a scan statistic
  - We do not know where diseases will occur
  - We do not know their geographical extent

1. Obtain data for a set of spatial locations s_i
2. Choose a set of spatial regions S to search
3. Choose models of the data under null hypothesis H0 (no clusters) and alternative hypotheses H1(S) (cluster in region S).
4. Derive a score function F(S) based on H1(S) and H0
5. Find the most anomalous regions (i.e. those regions S with highest F(S))
6. Determine whether each of these potential clusters is actually an anomalous cluster

Spatial scan

- Create a regular or irregular grid of centroids for the study area
- Create an infinite number of circles around each centroid, with the radius anywhere from zero up to a maximum of50 percent of the population
- For each circle:
  - Obtain actual and expected number of cases inside and outside the circle
  - Calculate likelihood function
- Compare circles:
  - Pick circle with highest likelihood function as Most Likely Cluster
- Inference:
  - Generate random replicas of the data set under the null-hypothesis of no clusters (Monte Carlo sampling)
  - Compare most likely clusters in real and random data sets (Likelihood ratio test)

Spatial scan

- Adjusts for inhomogeneous population density
- Simultaneously tests for clusters of any size and any location using circular windows with variable radius
- The approach accounts for multiple testing
- It is possible to include covariates that might be a source of confounding
- The approach can be used for point or aggregated data
- Can analyse data based on multiple distributions
  - Bernoulli – case/control or cross-sectional data
  - Poisson – incidence rate
  - Normal – similar to a LISA analysis for continuous data (next slide)
  - Exponential – survival analysis
  - Ordinal – uncommon for veterinary medicine but could be cancer staging
  - Space-time permutation – when only case data available

LISA

- Local indicators of spatial autocorrelation
- Modification of the global Moran’s I
- Cannot distinguish between H-H and L-L clusters
- Conventional clustering fails to preserve contiguity

$$\mathrm{MC} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \frac{\sum_{i=1}^{n} z_{Y,i} \sum_{j=1}^{n} c_{ij} z_{Y,j}}{(n-1)/n}$$

Time-space cluster detection

- Ederer, Myers, and Mantel (EMM) test
  - One-sided test for clustering
  - Sensitive to changes in population at risk and therefore not recommended for many (>5) time periods
  - Can be calculated relatively easily by hand (spreadsheet)
- Scan statistics
  - Can be used to identify high-rate and low-rate clusters (not one-sided)
  - Can be used for many time periods
  - Can be computationally expensive and not possible to calculate by hand
  - Commonly employed in veterinary epidemiology

EMM space-time test

- **Clustering yes/no – not able to distinguish between random and uniform (dispersed) distributions if no clustering detected**
- **A test for time clustering in several time series simultaneously**
- **Not used as commonly since the availability of spatial scan tests**

$$\chi^2 = \frac{\left[\left|\sum m_i - E\left(\sum m_i\right)\right| - 0.5\right]^2}{\sum V(m_i)}$$

m_i = maximum number of cases in single time period for location i

$\sum m_i$ = sum of m_i over all areas; $E(\sum m_i)$ = expected value for sum

$V(m_i)$ = variance of areal maxima; 0.5 is the χ^2 continuity correction factor

Ederer, F., Myers, M.H., Mantel, N., 1964. A statistical problem in space and time: Do leukemia cases come in clusters? Biometrics 20: 626-638.

World Organisation for Animal Health
Founded as OIE

100
1908 - 2008
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science

Space-time scan

- Use a cylindrical window, with the circular base representing space and the height representing time
- For each cylinder
  - Obtain actual and expected number of cases inside and outside the cylinder.
  - Calculate likelihood function
- Compare cylinders
  - Pick cylinder with highest likelihood function as Most Likely Cluster
- Inference:
  - Generate random replicas of the data set under the null-hypothesis of no clusters (Monte Carlo sampling)
  - Compare most likely clusters in real and random data sets (Likelihood ratio test)

Space-time permutation

- **Case-only data**
- **For each cylinder, calculate the expected number of cases conditioning on the marginal totals**

$$\mu_{st} = \Sigma_s c_{st} \text{ x } \Sigma_t c_{st} \ / \ C$$

- **where c_{st} = number cases at time *t* in location *s***
- **and C = total number of cases**
- **Then calculate the test statistic**

$$T_{st} = [c_{st} / \mu_{st}]^{cst} \text{ x } [(C-c_{st})/(C- \mu_{st})]^{C-cst}$$

if $c_{st} > \mu_{st}$ = 1, otherwise

Test statistic $T = \max_{st} T_{st}$

World Organisation for Animal Health
Founded as OIE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science

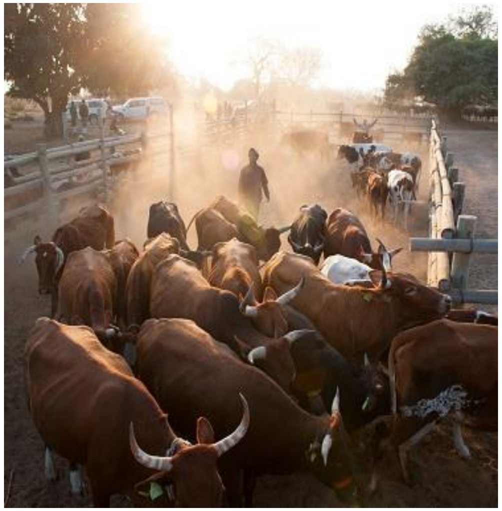
Space-time permutation

- Generate random replicas of the data set conditioned on the marginal totals by permuting the pairs of spatial locations and times
- Compare test statistic in real and random data sets using Monte Carlo hypothesis testing:
- p = rank(T_{real}) / (1 + number of replicates)

- Adjusts for purely geographical and purely temporal clusters
- Simultaneously tests for outbreaks of any size at any location using a cylindrical windows with variable radius (space) and height (time)
- Accounts for multiple testing
- Aggregated or non-aggregated data

Summary

- Spatial epidemiology concerns describing disease occurrence in terms of geographical location
- There are many procedures available to create risk maps and evaluate for the presence of clustering
- Some methods can be performed “by hand” while others require specialized software
- Risk maps (interpolation) are typically performed in GIS software
- SaTScan is free software (www.satscan.org) that is commonly used to investigate spatial clustering



Thank you

World Organisation
for Animal Health
Founded as OIE

100
1908 - 2008

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Veterinary Science