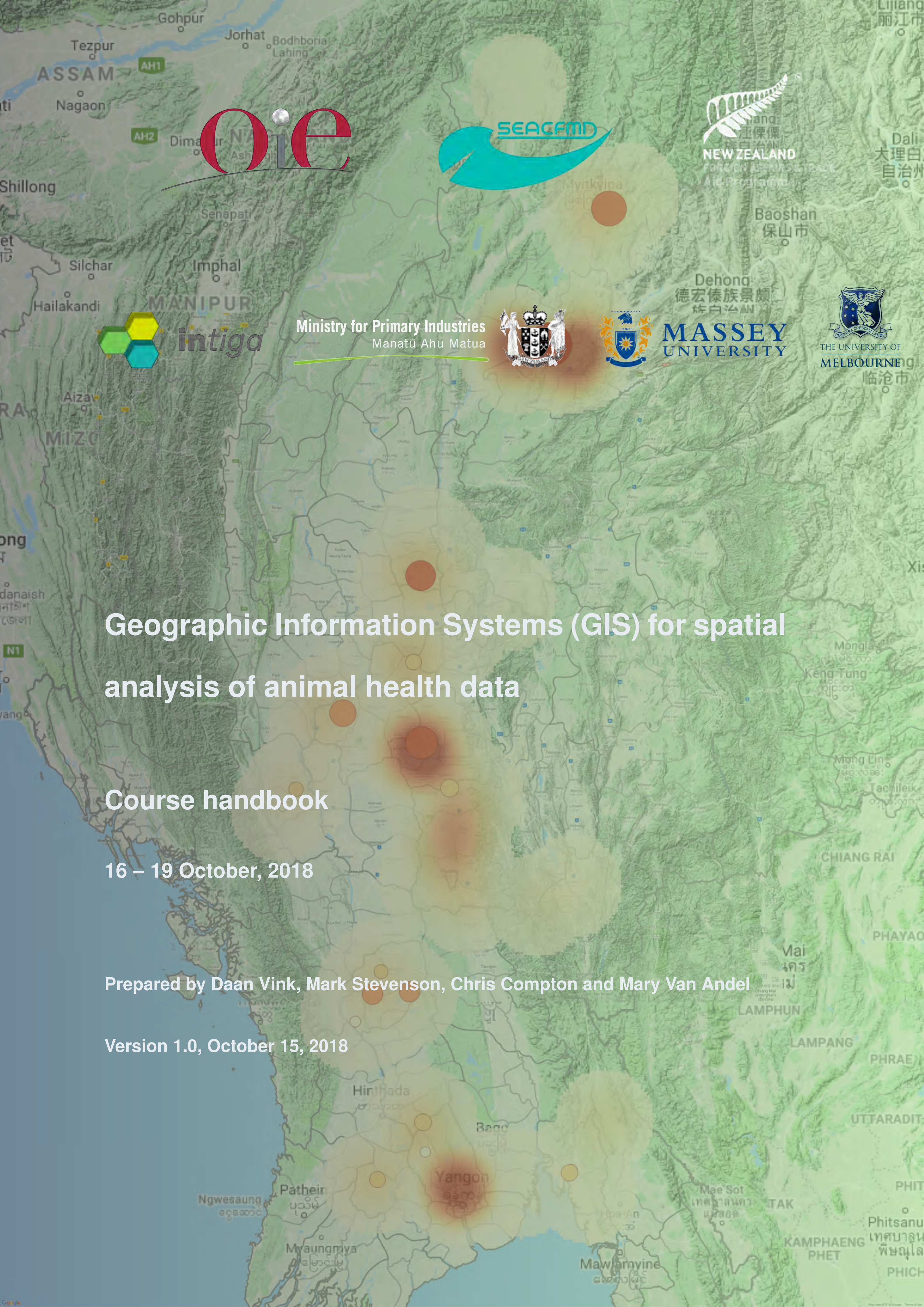


OIE

SEACFMD

NEW ZEALAND
Foreign Affairs & Trade
Aid Programme

intiga

Ministry for Primary Industries
Manatū Ahu Matua

MASSEY
UNIVERSITY

THE UNIVERSITY OF
MELBOURNE

Geographic Information Systems (GIS) for spatial analysis of animal health data

Course handbook

16 – 19 October, 2018

Prepared by Daan Vink, Mark Stevenson, Chris Compton and Mary Van Andel

Version 1.0, October 15, 2018

Contents

1 Course introduction

1.1 Background

Geographic Information Systems (GIS) are an essential tool to visualise an array of different data sets spatially through the creation of detailed maps. In recent years, applications of spatial epidemiology using GIS software have increased in their capacity to analyse and interpret data for enhanced understanding of relationships, disease patterns and trends. Open-source GIS software, such as `QGIS`, is easily accessible for use by Veterinary Services (VS) for animal disease surveillance, research and control purposes. GIS tools can also be used in a research capacity to better understand animal disease in relation to risk factor determination, spatial disease modelling, and distribution and prevalence studies.

In addition, GIS tools can be used practically to support VS in a range of activities relevant to disease control. For example, when an outbreak of an infectious disease occurs, veterinary officers can use GIS to map the locations of outbreaks and analyse this data to answer important questions of who, what, when and where in relation to the outbreak. Assessment of spatial trends in disease morbidity and mortality over time can support surveillance programmes and can be applied to detect aberrations and disease clustering. In organised disease response operations, spatial analyses can inform VS on how to prioritise the operationalisation and allocation of resources. In analogy with applications in surveillance and disease control, risk-based approaches using GIS tools are increasingly being used by technical staff to manage and focus the location of disease control and prevention efforts.

In recognizing the application of GIS software in animal disease surveillance and control, the World Organisation for Animal Health (OIE) Sub-Regional Representation (SRR) for South East Asia (SEA) is organising a four-day training course. This is the second such course; the first was organised in October 2017. Although foundation principles of spatial analysis and GIS will be covered, this is not designed as an introductory course, and the participants are expected to have some proficiency in spatial analysis and skills in using GIS techniques. The course will consolidate the material and learnings from the previous course and extend it further to increase the capability of participants to perform such work effectively.

1.2 Course details

1.2.1 Course overview

This course is aimed at personnel who have a role in epidemiological analysis of disease data at the national and regional level, such as WAHIS/ARAHIS and/or SEACFMD EpiNet Focal Points.

It focuses on spatial analysis of disease incidence, clustering and density estimation, and will also include a component on spatial risk assessment.

1.2.2 Course objectives and outcomes

The overall objective is for participants to strengthen their knowledge and skills of the analysis of disease data. By the end of this course, participants should:

1. Have improved their understanding of spatial epidemiology.
2. Be able to detect clusters of disease in space and time, identify potential hotspots and model the space-time cluster patterns.
3. Be proficient in undertaking a spatial disease risk assessment.
4. Have increased their proficiency in using tools such as `QGIS` and `R` to perform such spatial analysis.

1.2.3 Course structure and delivery

The following topics are included in the programme:

- **Topic 1: Exploratory spatial data analysis.** Using a combination of disease incidence data and census (population) data, participants will appropriately display the spatial distribution of disease events and quantify measures of disease (attack rate, mortality rate and case fatality rate) at the level of the epidemiological unit (usually the village). Where possible, an assessment of temporal trends will be made.
- **Topic 2: Analysing spatial heterogeneity of disease.** Global and localised disease cluster estimation techniques will be summarised. Relevant spatial scan statistics (purely spatial and spatiotemporal) will be computed to test the significance of disease clusters, using `SaTScan` and `QGIS`.
- **Topic 3: Assessment of spatial disease risk.** Utilising the outputs of the Multicriteria Decision Analysis (MCDA) process performed during the 2017 course, a spatial risk assessment for FMD will be performed.
- **Topic 4: An introduction to cluster analysis and spatial modelling.** This more advanced topic introduces participants to more formal data-driven analysis to develop space-time statistical models and quantify differences and associations.

Delivery is by technical experts proficient in GIS, spatial analysis of animal health and veterinary epidemiology from Intiga Consulting, Massey University EpiCentre, Melbourne University and the New Zealand Ministry for Primary Industries (MPI). There is an emphasis on the practical application of the material covered, including interpretation of the outputs, making use of the support and advice given by the course instructors. The lectures will be brief and serve to support the practical work. The practicals will be preceded by whole-group demonstration where relevant.

Participants may access all course materials, data, practical exercises, software and other resources in an online Learning Management System (LMS). Additional and updated materials can be added on the fly, and course participants may continue to use this after the course's conclusion. There will be communication with the participants prior to the course to ensure they can access this site. Participants may be requested to download and install the current versions of the software packages utilised.

1.3 Course programme

The course programme consists of three and a half days of classroom teaching followed by a session to discuss and consolidate the course learnings.

Part of the course will be delivered in parallel to two groups, implementing comparable material. One group will utilise `QGIS` for this while the other group will use `R`. The work in `R` will be more advanced; on the final day, and introduction to spatial modelling will be included.

Tuesday 16 October: GIS refresher and analysis of disease count data		
08:00 - 08:30	Registration	
08:30 - 08:45	Opening of the course	RA[1]
08:45 - 09:00	Participants introductions	
09:00 - 09:45	Course objectives, structure, logistics and practicalities	DV[2]
09:45 - 10:30	Lecture 1: Basics of GIS	MS[3]
10:30 - 11:00	*Morning tea / coffee*	
11:00 - 11:45	Lecture 2: GIS and spatial epidemiology	MS
11:45 - 12:30	Practical 1: Getting set up and creating a project: the 'base map'	DV
12:30 - 13:30	*Lunch*	
13:30 - 14:15	Presentation: Disease outbreak reporting in WAHIS	YQ[4]
14:15 - 15:00	Practical 2: Exploratory analysis of the disease data	DV
15:00 - 15:30	*Afternoon tea / coffee*	
15:30 - 17:00	Practical 2: Descriptive spatial analysis of the disease data	DV

Wednesday 17 October: Analysing spatial heterogeneity of disease		
08:30 - 09:15	Lecture 3: Analysis of disease count data /Disease distribution and clustering in space and time	MS
09:15 - 10:30	Practical 3: Analysis of disease count data plus visualisation of disease distribution and clustering	DV / MS
10:30 - 11:00	*Morning tea / coffee*	
11:00 - 12:30	Practical 3: Analysis of disease count data plus visualisation of disease distribution and clustering	DV
12:30 - 13:30	*Lunch*	
13:30 - 14:15	Lecture 4: Cluster detection using the spatial scan statistic	DV
14:15 - 15:00	Practical 4: SaTScan	DV
5:00 - 15:30	*Afternoon tea / coffee*	
15:30 - 17:00	Practical 4: SaTScan	DV

Thursday 18 October: Spatial modelling		
08:30 - 09:30	Lecture 5: MCDA, the REMBRANDT technique and WLC: spatial disease risk assessment	MS / DV

Group 1: Analysis using QGIS			***Group 2: Modelling using R***		
09:30 - 10:30	Practical 5: Spatial disease risk assessment using QGIS	DV / MvA[5]	09:30 - 10:30	Practical 5: Spatial disease risk assessment using R	MS / CC[6]
10:30 - 11:00	*Morning tea / coffee*				
11:00 - 12:30	Practical 5: Spatial disease risk assessment using QGIS	DV / MvA	11:00 - 12:30	Practical 5: Spatial disease risk assessment using R	MS / CC
12:30 - 13:30	*Lunch*				
13:30 - 14:15	Practical 5: Spatial disease risk assessment using QGIS	DV / MvA	13:00 - 14:15	Practical 5: Spatial disease risk assessment using R	MS / CC
14:15 - 15:00	Practical 5: Spatial disease risk assessment using QGIS	DV / MvA	14:15 - 15:00	Practical 5: Spatial disease risk assessment using R	MS / CC
15:00 - 15:30	*Afternoon tea / coffee*				
15:30 - 17:00	Practical 5: Spatial disease risk assessment using QGIS	DV / MvA	15:30 - 16:45	Practical 5: Spatial disease risk assessment using R	MS / CC

Friday 19 October: Spatial modelling, implementation and inference, application

Group 1: Analysis using QGIS			***Group 2: Modelling using R***		
08:30 - 09:15	Carry-over topics and work on own projects	DV / MvA	08:30 - 09:15	Lecture 6: Spatial disease modelling using R	CC / MS
09:15 - 10:30	Carry-over topics and work on own projects	DV / MvA	09:15 - 10:30	Practical 6: Spatial disease modelling using R	CC / MS
10:30 - 11:00	*Morning tea / coffee*				
11:00 - 12:30	Carry-over topics and work on own projects	DV / MvA	11:00 - 12:30	Lecture 6 / Practical 7: Spatial disease modelling using R	CC / MS
12:30 - 13:30	*Lunch*				
13:30 - 14:30	Group discussion: Applying cluster detection and analysis in real-life situations				ALL
14:30 - 15:00	What next? Discussion and perspectives; course conclusion				RA / DV

[1] Ronello Abila
[2] Daan Vink
[3] Mark Stevenson
[4] Yu Qiu
[5] Mary Van Andel
[6] Chris Compton

Part I

Reference material

2 Geography and epidemiology

2.1 Definition of a geographic information system

A geographic information system (GIS), as defined by University of Edinburgh's Dictionary of GIS terms, is 'a computer system for capturing, storing, checking, integrating, manipulating, analysing and displaying data related to positions on the Earth's surface.' Typically, a GIS is used for handling maps of one kind or another. These might be represented as several different layers where each layer holds data about a particular kind of feature (e.g. roads). Each feature is linked to a position on the graphical image of a map. The primary value of a GIS is that it defines precisely the location of objects and provides users with the ability to visualise the spatial arrangement of those objects. Knowledge of location allows complex calculations to be performed (such as working out the shortest route and the shortest travel time between two locations). For epidemiologists, the ability to visualise spatial data is a powerful method of describing the patterns of disease, and is a useful technique for identifying factors that potentially influence patterns of disease.

2.2 The elements of geographic data

Geographic data are built up from single elements, or facts, about the real world. In its crudest form, an element of geographic data (termed a datum) links: (1) place, (2) time, and (3) a descriptive property about place and time. For example, the statement: 'The temperature at 12 noon on 10 June 2003 at latitude 45° and longitude 60° was 25° Celsius' ties place and time to the property (or attribute) of atmospheric temperature. In many cases geographical data are slow to change and for this reason time is often omitted from geographic descriptions. On the other hand, atmospheric temperature changes constantly, so time is an important component of this type of representation.

The range of attribute information in geography is vast. Attribute information may be classified as follows.

- Nominal: attributes are nominal if they are given names or titles in order to distinguish one entity from another. Place names are a good example of nominal attributes (see the GeoNames website for a complete list of place names in different languages).
- Ordinal: attributes are ordinal if their values take on a natural order. For example, agricultural land may be classed in terms of soil quality with class 1 representing the best, class 2 second-best and so on.
- Numeric: examples include temperature, height above sea level, counts of numbers of cases of disease. Values vary on a discrete (for example, integer) or continuous scale.

Just as attributes can be classed into different types, so too can spatial objects. The various types of spatial object include:

- Points: spatial objects that have neither length nor breadth and therefore a dimension of zero; points may be used to indicate spatial occurrences or events; point pattern analysis is used to identify whether occurrences or events are interrelated;
- Lines: spatial objects that have length but no breadth, and hence a dimension of one; used to represent linear entities such as roads, pipelines and cables which are frequently assembled to form networks;
- Areas: spatial objects of two dimensions of length and breadth; may be used to represent natural objects such as countries, state boundaries or agricultural fields; areas may bound linear features and enclose points;
- Surfaces or volumes: spatial objects of length, breadth and depth; used to represent natural objects such as river basins, canyons, and mountains; surfaces are frequently derived by interpolating between lower dimension measurements such as point measurements of height;
- Time: often considered to be the fourth dimension of spatial objects.

2.3 Spatial resolution

The classification of geographic data into object types is dependent on scale. For example, at a low level of resolution a farm may be represented as a single point. At a higher level of resolution the same farm may be better represented as an area, where the exact farm boundaries are explicitly defined. At an even higher level of resolution the farm may be represented as a surface where, in addition to boundary information, details of height, aspect and slope are provided.

Figure 2.1 illustrates how, as spatial resolution increases, greater detail can be appreciated and the shape of spatial objects will change. Objects that appear as lines at low resolution (for example, in the left-hand and centre diagrams), are best represented by polygons when viewed at a high level of resolution (as in the right-hand diagram).

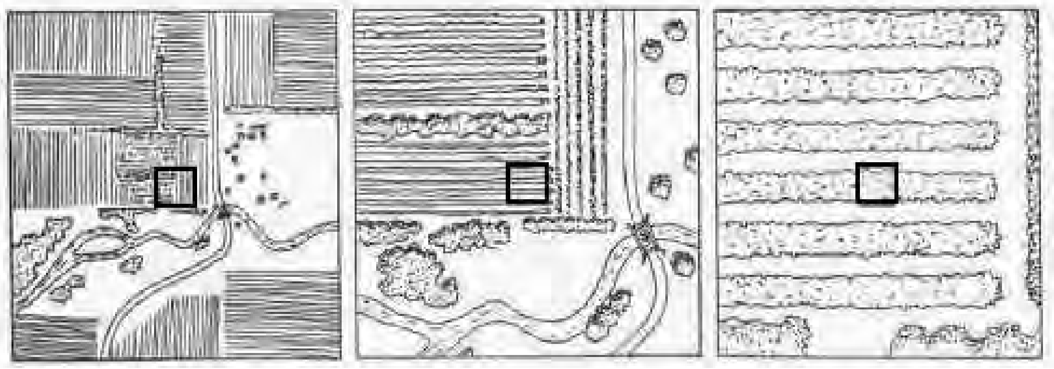


Figure 2.1: Different levels of spatial resolution.

In principle, if we collected enough items of geographic information we would be able to build a complete representation of the world. In practice any representation that is made is partial in that it must limit the level of detail provided or ignore change that may occur through time. One common way of limiting detail is by ignoring information that applies to small areas — that is, to reduce the spatial resolution. A second way is to regard many attributes as remaining constant over large areas.

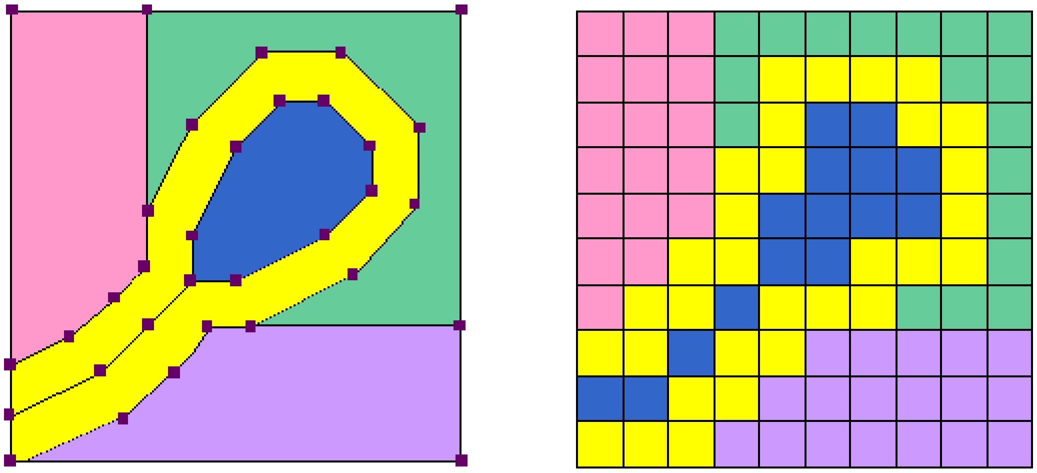
2.4 Data formats

Spatial data may be stored in either raster or vector formats with a GIS. Vector data are composed of points, lines, and polygons. Raster data sets are composed of rectangular arrays of regularly spaced square grid cells. Each cell has a value, representing a property or attribute of interest. While any type of geographic data can be stored in raster format, raster data sets are especially suited to the representation of continuous, rather than discrete, data.

2.4.1 Vector data

Vector data are composed of points, lines, and polygons. This spatial data model is known as 'arc-node topology.' Arcs are composed of nodes and vertices. Arcs begin and end at nodes, and may have 0 or more vertices between nodes. The vertices define the shape of the arc along its length. Arcs which connect to each other will share a common node.

Points represent discrete locations within a study area. These are either true points, such as the highest point on a mountain, or virtual points, based on the scale of representation. For example, a city's location on a driving map is represented by a point even though in reality a city occupies a defined area. Lines represent linear features such as rivers and roads. Each line is composed of a number of different coordinates, which make up the shape of the line, as well as the tabular record for the line vector feature. Polygons form bounded areas (Figure 2.2 (a)). Polygons are formed by bounding arcs, which keep track of the location of each polygon.



(a) Vector

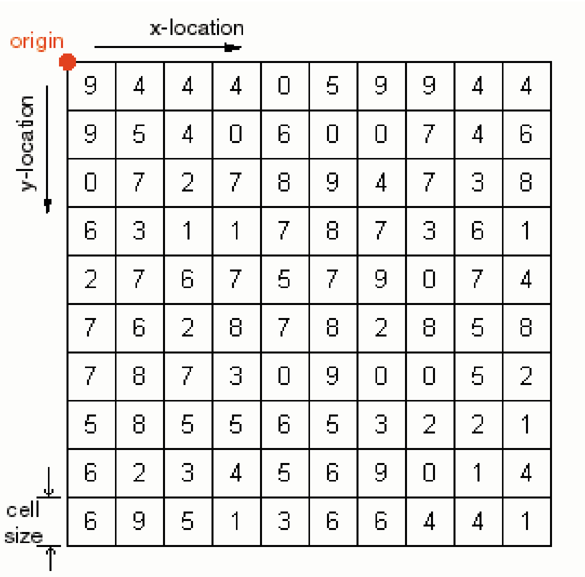
(b) Raster

Figure 2.2: A simple representation of spatial data formats.

2.4.2 Raster data

In a raster representation, geographic space is divided into an array of cells (a matrix), as shown in Figure 2.2 (b). These cells are sometimes called pixels.

Generally, cells are assigned a single numeric value, but with grid themes (a proprietary ArcInfo data format), cell values can also contain additional text and numeric attributes. All raster data sets are spatially referenced by a very simple method: only one corner of the raster theme is georeferenced (Figure 2.3). Because cell size is constant in both the horizontal and vertical directions, cell locations are referenced by row-column designations, rather than with explicit coordinates for the location of each cell's center. Different raster file formats may have an origin located at the lower left rather than at the upper left. Each cell or pixel contains a value representing some numerical phenomenon, or a code use for referencing to a non-numerical value.



9	4	4	4	0	5	9	9	4	4
9	5	4	0	6	0	0	7	4	6
0	7	2	7	8	9	4	7	3	8
6	3	1	1	7	8	7	3	6	1
2	7	6	7	5	7	9	0	7	4
7	6	2	8	7	8	2	8	5	8
7	8	7	3	0	9	0	0	5	2
5	8	5	5	6	5	3	2	2	1
6	2	3	4	5	6	9	0	1	4
6	9	5	1	3	6	6	4	4	1

Figure 2.3: Spatial referencing a raster surface.

With vector data, each point, node, and vertex has an explicit and absolute coordinate location. Raster cells, in contrast, are georeferenced relative to the theme's coordinate origin. This enhances processing time immensely in comparison to certain types of vector data processing. However, the file sizes of raster data sets can be very large in comparison to vector data sets representing the same phenomenon for the same spatial area. Also, there is a geometric relationship between raster resolution and file size. A raster data set with cells half as large (e.g. 10 m on a side instead of 20 m on a side) may take up 4 times as much storage space, because it takes four 10 m cells to fit in the space of a single 20 m cell.

Cells may either have a value ($0 - \infty$) or no value (null, or no data). The difference between these is important. Null values mean that data either fall outside the study area boundary, or that data were either not collected or not available for those cells. In general, when null cells are used in analysis, the output value at the same cell location is also a null value. Grid data sets can store either integer or floating-point (decimal) data values, though some other data formats can only store integer values. Typical simple image data will have limits on the number of unique cell values (typically 0 – 255).

When information is represented in raster format each component cell is assigned a single attribute value and all detail about variation within the component cell is lost. The size chosen for cells within a raster surface depend on the resolution of the data used to create the surface. The cell must be small enough to capture the detail required, but large enough so the data can be stored and analysed efficiently. As the homogeneity of the data increases so too can the designated cell size.

When creating raster data several rules may be applied to specify how a cell will be coded: in most situations the attribute with the largest share of the cell's area gets the cell attribute value. In other circumstances the rule is based on the central point of the cell and the attribute values at the central point are assigned to the cell. Although the largest share rule is almost always preferred, the central point rule is commonly used because it is quick to calculate.

2.4.3 Raster calculations

There are three basic categories of functions for the creation of raster surfaces: global, focal, and zonal. Global functions perform their algorithm on every cell in the data set (see the example in Figure 2.4). You can think of the global function calculation engine as starting at once cell location, performing a calculation once on the inputs at that location, and then moving on to the next cell location, and so on. Focal functions consider neighbourhoods, so that the output cell is the result of a calculation performed on either a group of cells determined by a window of cells (known as a kernel or focus) around each cell of interest. For example, a smoothing (low-pass filter) algorithm will take the mean value of a 3 × 3 cell kernel, and place the output value in the location of the central cell. If the kernel contains locations that are outside of the grid, these locations are not used in the calculation. Zonal functions perform analyses on a group of cells with a common value (a zone) in one of the inputs.

For most raster functions, other than those which simply identify a selected group of cells, operations take the conceptual format of an algebraic expression using arithmetic, relational, Boolean, logarithmic, trigonometric, and power operators. Table 2.1 summarises the features of raster and vector representations in a GIS.

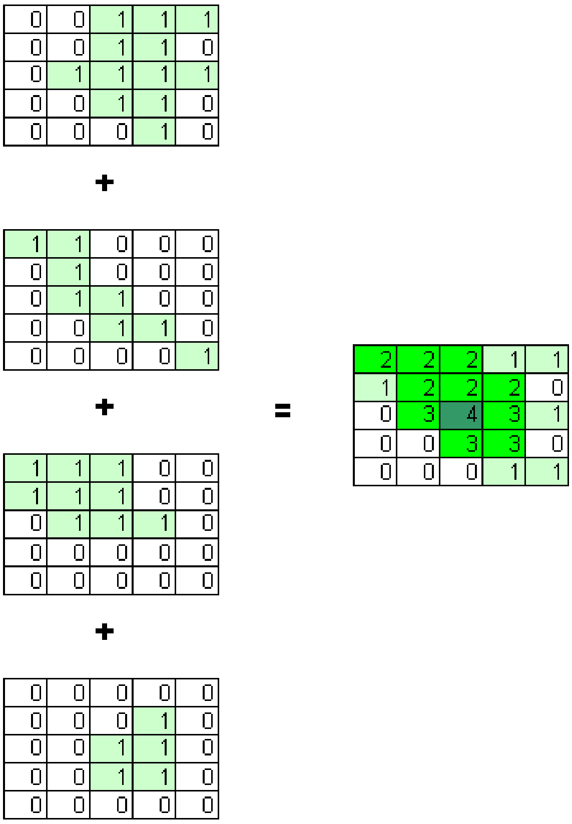
Table 2.1: Features of raster and vector representations in a GIS.

Factor	Raster	Vector
Source	Remote sensing	Printed maps, digitised maps
Applications	Environmental applications	Social, economic, administrative
Resolution	Fixed	Variable
Volume of data	Depends on size of cells used	Depends on level of detail
Efficiency	Efficient	Relatively inefficient

2.5 Georeferencing

Georeferencing refers to the process of assigning location information to geographic data. The primary requirements of a georeference is that it is unique so that there can be no confusion about the location that is referenced; and that its meaning is shared among all working with the information. In addition, a georeference must be persistent in time. In all cases georeferences are based on some type of coordinate system to define the location of points in two-dimensional or three-dimensional space.

René Descartes (1596 – 1650) introduced systems of coordinates based on orthogonal (right angle) axes. These two and three-dimensional systems used in analytic geometry are often



0	0	1	1	1
0	0	1	1	0
0	1	1	1	1
0	0	1	1	0
0	0	0	1	0

+

1	1	0	0	0
0	1	0	0	0
0	1	1	0	0
0	0	1	1	0
0	0	0	0	1

+

1	1	1	0	0
1	1	1	0	0
0	1	1	1	0
0	0	0	0	0
0	0	0	0	0

+

0	0	0	0	0
0	0	0	1	0
0	0	1	1	0
0	0	1	1	0
0	0	0	0	0

=

2	2	2	1	1
1	2	2	2	0
0	3	4	3	1
0	0	3	3	0
0	0	0	1	1

Figure 2.4: Simple addition of four raster surfaces (an example of a global raster calculation).

referred to as Cartesian systems. Similar systems based on angles from baselines are often referred to as polar systems. Before discussing the various types of coordinate systems used in geography, some background information on ellipsoids and map projections is provided.

The best model of the earth would be a 3-dimensional sphere in the same shape as the earth. Spherical globes are often used for this purpose. However, globes have several drawbacks: (1) they are large and cumbersome, (2) they are generally of a scale unsuitable to the purposes for which most maps are used, and (3) we usually want to see more detail than is possible to be shown on a globe. In addition standard measurement equipment cannot be used to measure distance on a sphere, as these tools have been constructed for use on flat surfaces.

The latitude-longitude spherical coordinate system expresses positions on the earth's surface in terms of angles rather than easting and northing (Cartesian planar) coordinates. Using this system, 'horizontal' or east-west lines are lines of equal latitude or parallels. 'Vertical' or north-south lines are lines of equal longitude or meridians. These lines encompass the globe and form a gridded network called a graticule. The line of latitude midway between the poles is zero latitude and is called the Equator. The vertical axis, which defines the line of zero longitude, is called the Prime Meridian. For most geographic coordinate systems, the Prime Meridian is the longitude that passes through Greenwich in the United Kingdom. Other countries use as prime meridians longitudes that pass through Bern, Bogota, and Paris.

The point where the Equator and the Prime Meridian intersect defines the Origin (0,0). The Earth's sphere is then divided into four geographical quadrants based on compass bearings from the origin. Above and below the Equator are north and south, and to the left and right of the Prime Meridian are west and east. Latitude and longitude are traditionally measured in decimal degrees or in degrees, minutes and seconds (DMS).

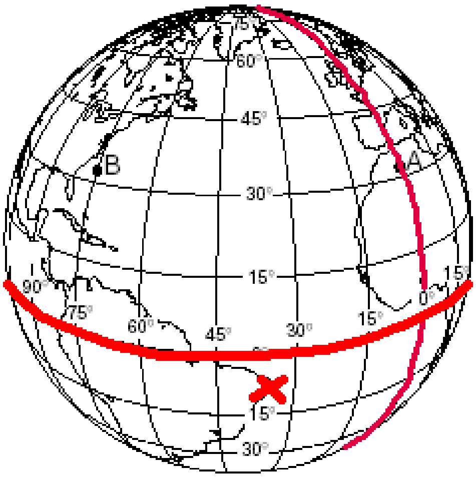


Figure 2.5: Latitude and longitude.

Latitudes are measured relative to the Equator and range from -90° at the South Pole to +90° at the North Pole. Longitude is measured relative to the Prime Meridian (shown as a red line running from north to south in Figure 2.5) positively, up to 180°, when travelling east and measured negatively, to -180°, travelling west. If the Prime Meridian is at Greenwich, then the most eastern point of South America (marked as a × in Figure 2.5) has a latitude of approximately -7° (negative, because it is south of the Equator) and a longitude of approximately -37° (negative, because it is west of the Prime Meridian at Greenwich). The latitude and longitude of this point would be written S 7° W 37°.

When the earth's shape is based on the WGS 84 Ellipsoid one degree of latitude at the Equator equals 110.57 kilometres. At the poles, one degree of latitude equals 111.69 kilometres. One degree of longitude at the Equator equals 111.05 kilometres. At 60° N one degree of longitude equals 55.52 kilometres. Note:

- Quote latitude values first, then longitude (you can remember this because 'latitude' comes before 'longitude' alphabetically).
- Use decimal degrees in preference to degrees, minutes and seconds.

Map projections are sets of mathematical models which transform spherical coordinates (such as latitude and longitude) to planar coordinates. Cartesian coordinate systems assign two coordinates to every point on a flat surface by measuring distances from an origin parallel to two axes drawn at right angles. These axes are often termed x and y, and the associated coordinates are termed x and y coordinates, respectively. Because it is common to align the y axis with the North in geographic applications, the coordinates of a projection on a flat sheet are often termed easting and northing. The basic types of projections are conic, cylindrical, and planar.

2.6 Datums

The actual shape of the Earth is spheroid (a slightly 'flattened' sphere) rather than perfectly spherical. To account for this, a large numbers of surveys have been conducted over the years, resulting in a large number of ellipsoid definitions (examples are given in Table 2.2). The generalised earth-centered coordinate system (WGS84) provides a good overall mean solution for all places on the earth. However, for specific local measurements, WGS84 does not account well for local conditions. In this situation, local datums are useful. The local North American Datum of 1927 (NAD27) more closely fits the earth's surface in the upper-left quadrant of the earth's cross-section. NAD27 only fits this quadrant, so to use it in another part of the earth will result in serious measurement errors. For mapping North America, in order to obtain the most accurate locations and measurements, NAD27 or NAD83/91 are used.

Table 2.2: Common ellipsoid definitions used in geography.

Title	Length of major axis (metres)	Flattening
Airy 1830	6377563	299.325
Australian National	6378160	298.250
Bessel 1841	6377483	299.153
Clarke 1866	6378206	294.979
Clarke 1880	6378249	293.465
Helmert 1906	6378200	298.300
GRS 80	6378137	298.257
South American 1969	6378160	298.250
WGS 72	6378135	298.260
WGS 84	6378137	298.257

Datum is a term you might also come across when dealing with map projections. A datum defines the position of the spheroid relative to the centre of the Earth. Local datums are often used to align a given spheroid to more closely to the Earth's surface in a particular area. A local datum is not suited for use outside the area for which it was designed.

2.7 Coordinate systems

Once geographic data are projected onto a planar surface, features must be referenced by a planar coordinate system. The geographic system (latitude-longitude) which is based on angles measured on a sphere, is not valid for measurements on a plane. Therefore, a Cartesian coordinate system is used, where the origin (0, 0) is toward the lower left of the planar section. The true origin point (0, 0) may or may not be in the proximity of the map data you are using. Coordinates are then measured from the origin point. However, false eastings and false northings are frequently used, which effectively offset the origin to a different place on the plane. This is done in order to minimise the possibility of using negative coordinate values (to make calculations of distance and area easier) and to lowwer the absolute value of the coordinates (to make the values easier to read, transcribe, and calculate). Systems for georeferencing can be divided into two groups:

1. Global systems, which are used to define position at all locations across the Earth's surface.
2. Regional systems, which are defined for specific areas, often covering countries, states, or provinces.

Whatever coordinate system is used, it must have the following features:

- It must be unique, so that there can be no confusion about the location that is referenced;
- Its meaning must be shared among all working with the information; and
- It must be persistent in time.

Table 2.3 lists and describes some of the commonly used systems of georeferencing.

Table 2.3: Commonly used systems of georeferencing.

System	**Domain**	**Resolution**	**Example**
Place name	Will vary	Varies by feature type	Sydney, Canada; Sydney, Australia
Postal address	Global	Size of location that has one postal address - typically a house or building, but may mean a large farm or station	105 Woodham Lane Addlestone, Surrey
Post code	Country	Area occupied by a defined number of mailboxes	The post code of Addlestone, Surrey, UK, is KT15 3NB
Telephone calling area	Country	Varies from country to country	If you are phoning a residence in New Zealand with a phone number that starts with 06, you know that the place you are calling is located somewhere in the south of the North Island.
Cadastral system	Local land authority	Area occupied by a single parcel of land	A map of land ownership maintained for the purpose of taxing land or creating a public record of land ownership.
Latitude and longitude	Global	Infinite	E113○59’53.0”, N22○22’36.6”
State plane coordinates	Unique to country or state	Infinite	UK national grid

2.7.1 The Transverse Mercator projection

The simplest of all projections is the Transverse Mercator projection where longitude is plotted as x and latitude as y. The result is a heavily distorted image of the Earth, with the poles extending across the entire top and bottom edges of the map (Figure 2.6). The Transverse Mercator projection has straight meridians and parallels that intersect at right angles. Scale is true at the equator or at two standard parallels equidistant from the equator.

2.7.2 The Universal Transverse Mercator system

The Universal Transverse Mercator (UTM) system is based on the Transverse Mercator projection and is often used in military applications and in datasets with global or national coverage. Under the UTM system the Earth is divided from east to west into 60 zones (called UTM zone numbers where numbers range from 1 to 60), with each zone corresponding to a width of 6° (Figure 2.7). Each zone is mapped by the Transverse Mercator projection with a central meridian in the centre of the zone. UTM zone number 1 applies to longitudes from W 180° to W 174° (a line drawn between the Soviet Union and Alaska, straight through the Pacific Ocean).

Each UTM zone is divided from north to south into designators. There are 20 latitudinal zones spanning the latitudes 80°S to 84°N and denoted by the letters C to X, omitting the letter O. Each of these is 8 degrees south-north, apart from zone X which is 12 degrees south-north. Areas are referenced by quoting the longitudinal zone number, followed by the latitudinal zone letter. For example, the southern end of South America is 19F. Locations within a UTM zone are measured in metres eastward from the central meridian and northward from the Equator. However, eastings increase eastward from the central meridian which is given a false easting of 500 km so that only positive eastings are measured anywhere in the zone. Northings increase northward from the equator with the equator’s value differing in each hemisphere. In the Northern Hemisphere the

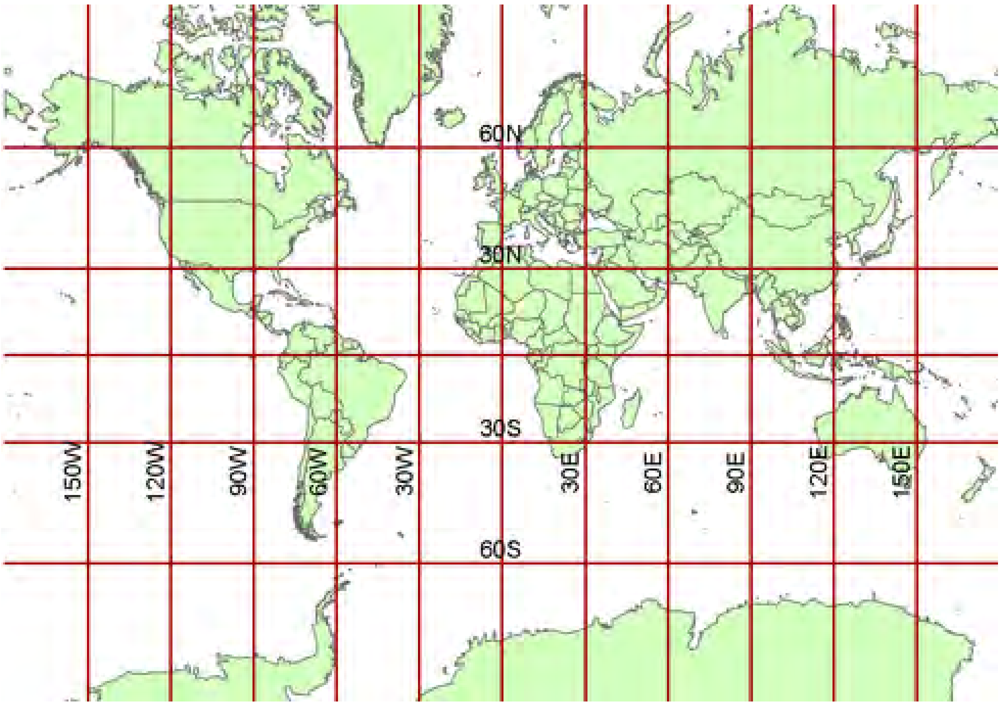


Figure 2.6: Map of the world plotted using the Transverse Mercator projection.

Equator has a northing of 0. For Southern Hemisphere locations the Equator is given a false northing of 10,000 km.

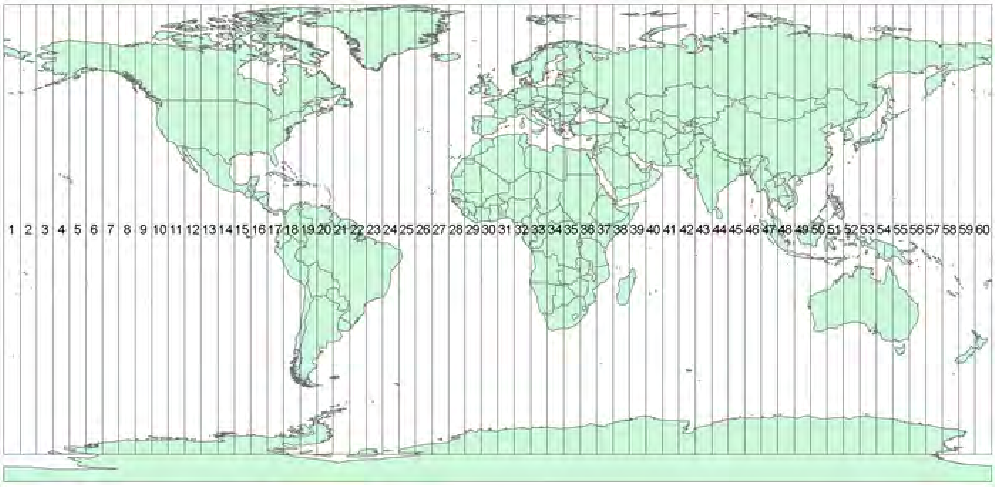


Figure 2.7: Map of the world plotted using the Universal Transverse Mercator (UTM) projection.

Because there are effectively 60 different projections in the UTM system, maps will not fit together across a zone boundary. Zones become such a problem at high latitudes that the UTM system is normally replaced with azimuthal projections centred on each pole above 80° latitude (these are known as UPS or Universal Polar Stereographic systems).

UTM coordinates are easily recognised because they commonly consist of a six-digit integer and letter (563257E, 4467843N for example). They are useful for spatial analyses conducted over large areas because distances can be calculated for points within the same zone with little error (typically no more than 0.04%). UTM grids are marked on many topographic maps and many countries project their topographic maps using UTM, so it is easy to obtain UTM coordinates from maps for input into digital datasets.

2.7.3 Transverse Mercator Grid Systems

Many countries have defined grid systems based on Transverse Mercator coordinates that cover their territory. The British National Grid (Figure 2.8) is an example of a national grid system based on the Transverse Mercator projection.

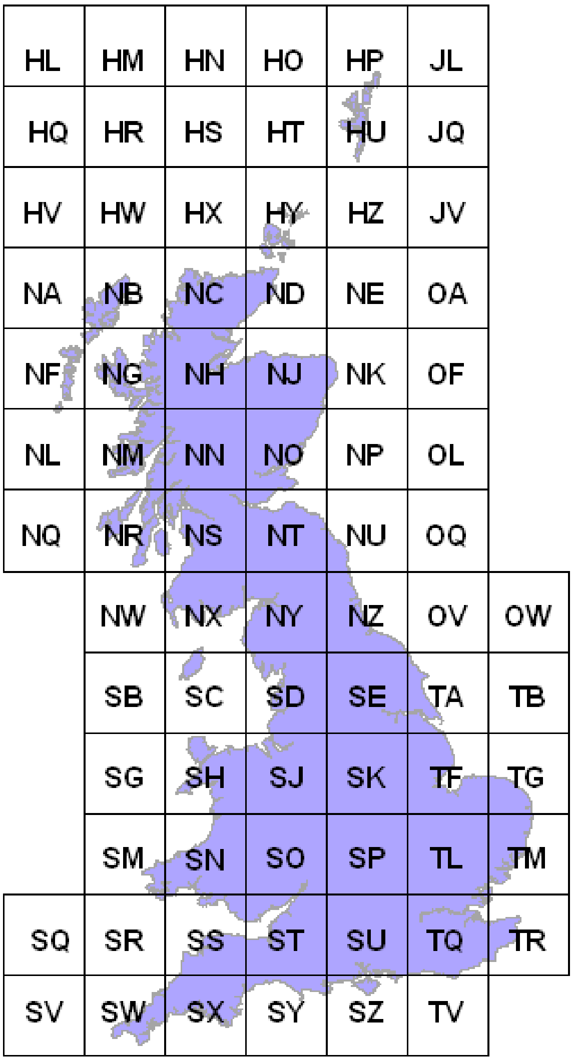


Figure 2.8: Map of the United Kingdom showing the boundaries of the British National Grid.

The British National Grid is administered by the Ordnance Survey of Great Britain and provides a unique georeference for every location in England, Scotland and Wales. The true origin of the system is at N 49° and W 2°. Grids are 100 kilometres from east to west and 100 kilometres from north to south. The first two letters of a British National Grid georeference define in which of the grids the location is situated (Figure 2.8). The next three digits represent multiples of 100 metres and define the easting coordinate (relative to the origin of the respective grid). The last three digits (again, in multiples of 100 metres) define the northing coordinate relative to the origin of the respective grid.

Take for example the British National Grid georeference SP254186. The origin of the SP grid (that is, its south-western corner) is 400 kilometres east and 200 kilometres north of the map origin (N 49° and W 2°, the most south-west corner of the grid labelled SV in Figure 2.8). The location is (254 × 100) / 1000 = 25.4 kilometres east and (186 × 100) / 1000 = 18.6 kilometres north of the origin of the SP grid. The grid coordinates for this location would be 4254000, 218600.

2.7.4 State plane coordinates

In the United States each state has its own State plane coordinate system. State plane systems were developed in order to provide local reference systems that were tied to a national datum. In the United States, the State Plane System 1927 was developed in the 1930s and was based on the North American Datum 1927 (NAD-27). NAD-27 coordinates are in Imperial units (feet).

2.8 Geography and spatial epidemiology

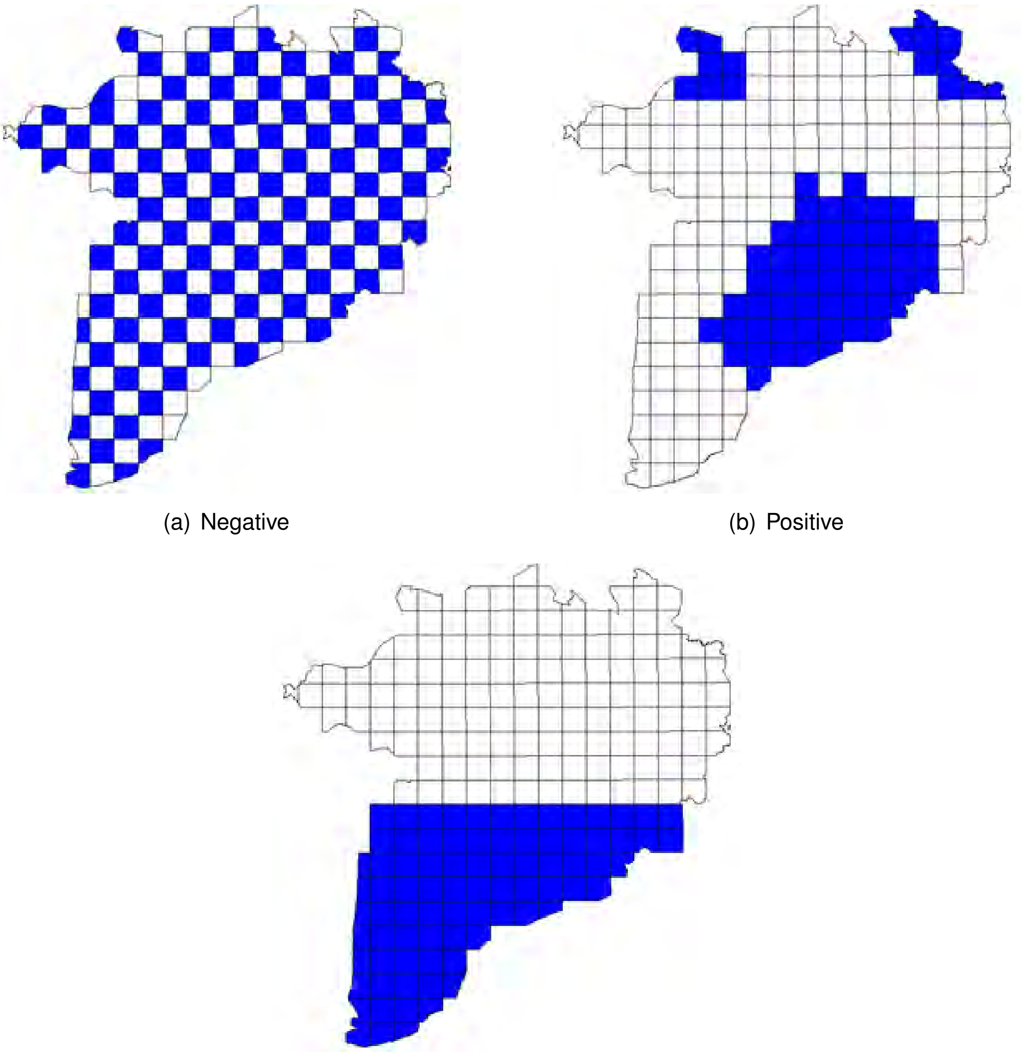
This section describes the characteristics that make spatial information different from other data types you may have dealt with, and explains why, when we analyse spatial data, these characteristics need to be accounted for.

2.8.1 Spatial autocorrelation

As a general rule spatial data tend to exhibit an increasing range of values (that is, they demonstrate increasing heterogeneity) with increasing distance. This may be stated in another way, in the form of Tobler's First Law of Geography (Tobler 1970): 'Everything is related to everything else, but near things are more related than distant things.'

Formally, this property is known as spatial autocorrelation and statistical tests are available to quantify the degree to which near and more distant things are interrelated. A similar concept, that of temporal autocorrelation, concerns the relationship between consecutive events in time. Spatial autocorrelation measures attempt to deal simultaneously with similarities in the location of spatial objects and their attributes. If features that are similar in location are also similar in attributes, the pattern as a whole is said to exhibit positive spatial autocorrelation. On the other hand, negative spatial autocorrelation exists when features which are close together in space tend to be more dissimilar in their attribute values than features that are further apart.

Figure 2.9 explains this concept more fully. Suppose that you have a country comprised of a series of grid cells. You travel around this country, taking soil samples from within each grid. The samples are then tested for the presence of fungal spores. Those grid cells where spores are found are represented by blue shading, those cells where fungus is absent are represented by white. In Figure 2.9(a), features that are adjacent are dissimilar: if we take any fungus-positive grid, there is a mix of positive and negative grids in its immediate vicinity. This pattern is typical of negative spatial autocorrelation. In Figure 2.9(b) features that are adjacent tend to be similar (clustered) — there is a large cluster of fungus-positive grids in the south east of the country (and two more towards the north). This pattern is typical of positive spatial autocorrelation. In 2.9(c) features that are adjacent are, on the whole, identical. This pattern is typical of extreme spatial autocorrelation.



(a) Negative

(b) Positive

(c) Extreme positive

Figure 2.9: An irregular area (a country, for example) divided into grids, with each cell of the grid coloured according to the presence or absence of a given attribute. Three types of spatial autocorrelation are shown.

In spatial data analysis an understanding of spatial autocorrelation has an important influence on the way in which we abstract and collect data, and how we draw inferences between events and occurrences. Spatial autocorrelation is important in the field of epidemiology for two reasons. Firstly, it helps us to generalise from a sample of observations in order to build a spatial data set. Secondly, the presence of spatial autocorrelation violates some of the key assumptions of many of the conventional statistical techniques used to quantify the relationship between two or more variables. Acknowledging the importance of geographic scale or level of detail is fundamental to understanding the likely strength and nature of spatial autocorrelation.

2.8.2 Spatial sampling

In aiming to represent the complexity of the real world, researchers have to employ abstraction and sampling of events and occurrences from a sample frame, defined as the group of eligible elements of interest. In reality, geographic representation is based on sampling in that elements of reality that are used are abstracted from the real world. In remote sensing each pixel is a spatially averaged reflectance value calculated at the spatial resolution characteristic of the sensor. In many situations we need to consciously select some observations and not others in order to create an abstraction of the area we wish to represent. This is because the resources available for any given project will never allow us to measure every single aspect of the elements that make up the region of interest.

In any application, where the events of interest are spatially heterogeneous we will require a large sample to capture the full variability of attribute values at all (or most) locations. Other parts of our study may be more homogeneous and a sparser sampling interval may be more appropriate. Both simple random and systematic random sampling designs may be adapted in order to allow a differential sampling interval over a given area — thus it may be possible to partition the sampling frame into sub-areas based on a knowledge of spatial structure, and specifically of the variability of the attributes that are being measured. Other, application-specific circumstances include the resources available for the project and accessibility of all parts of the study area for observation.

Example

You know that the concentration of a trace element in soil is closely related to soil type, and that there are two soil types within your area of study. Because of this you can restrict the number of samples that are taken within each soil type area, knowing that your samples taken at different locations throughout each area will yield a similar trace element concentration. In this case knowledge about the presence of spatial autocorrelation in soil trace element concentration has an important influence on your sampling strategy.

2.8.3 Spatial interpolation

In sampling part of reality it follows that you will need to exercise some level of judgement to ‘fill in the gaps’ — that is, to interpolate the sampled data so that your spatial representation will be in some way complete. To do this properly, you need to have a good understanding of issues related to autocorrelation. A literal interpretation of Tobler’s First Law of Geography implies a continuous, smoothly declining effect of distance upon the attribute values of adjacent spatial objects as you travel throughout a study region (that is, a decay function). The precise nature of the decay function used to represent the effect of distance is likely to vary depending on specific applications including linear distance, negative power distance and negative exponential distance.

In addition to appreciating the type of decay function, it is also important to be aware that the effects of distance may vary depending on direction. If the decay function is uniform in every direction it is said to be isotropic. If the decay function varies with direction it is said to be anisotropic.

Example

To continue our earlier example, having estimated trace element concentration at various sites you might then proceed to construct a contour surface to represent trace element concentration throughout the entire study area. The ability to interpolate the sampled data will depend on making assumptions about autocorrelation and a knowledge of how this autocorrelation may vary with direction. The notion of smooth, continuous variation underpins many of the techniques used to describe spatial data.

2.8.4 Aggregation

When reporting the spatial distribution of disease within a country it is common to aggregate information at an area level. This is done because the units of interest (for example, households or farms) occupy geographically unique locations and confidentiality restrictions usually dictate that uniquely attributable information is anonymised in some way.

Figure 2.10 provides an example where the standardised mortality ratio (SMR) of bovine spongiform encephalopathy (BSE) is plotted for arbitrarily defined areas of Great Britain. On the basis of Figure 2.10 we might legitimately conclude that there was a higher risk of BSE in the south of the country, compared with the north. While this might be the case at the area level we cannot necessarily draw the same conclusion at the individual farm holding level. In other words, if we picked a random holding from an area with a high SMR in the south of the country, we could not always guarantee that the selected holding would have experienced large numbers of BSE cases. Thus, while mapping disease at the area level is useful for identifying broad-scale spatial trends, it hides within-area heterogeneity of disease. This phenomenon (ecological fallacy) is something to beware of when drawing conclusions from disease data summarised at the area level.

2.8.5 The modifiable areal unit problem

When dealing with area data we need to consider how the zones of analysis affect results. If relationships between variables change with the selection of different areal units, the reliability of results is called into question. The effect of the selection of areal units on analysis, termed the modifiable areal unit problem (MAUP), is defined as: a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns (Openshaw 1984).

The MAUP has been most prominent in the analysis of socioeconomic and epidemiological data (see, for example Wong et al. 1999 and Nakaya 2000). Such areal data cannot be measured at a single point, but must be contained within a boundary to be meaningful. For example, it is not

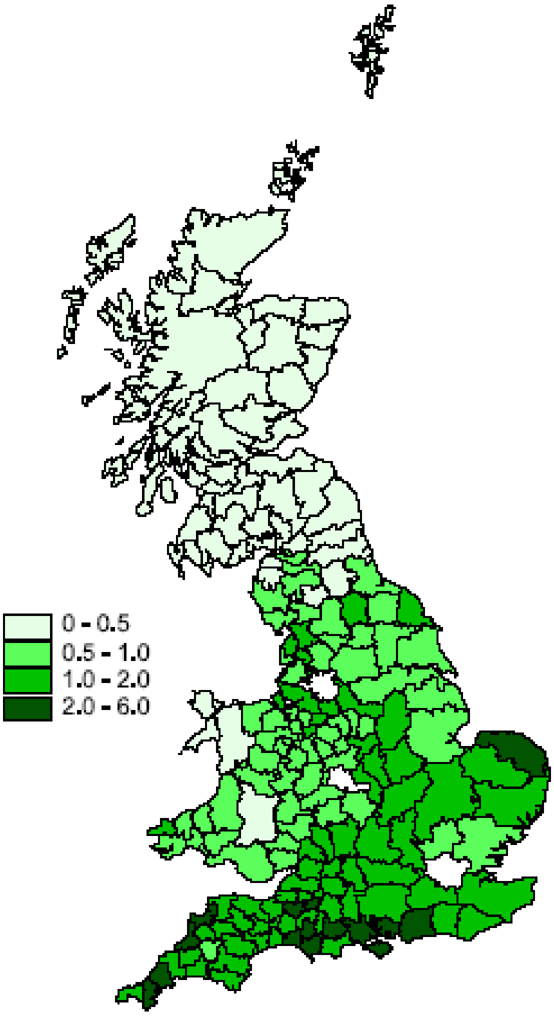


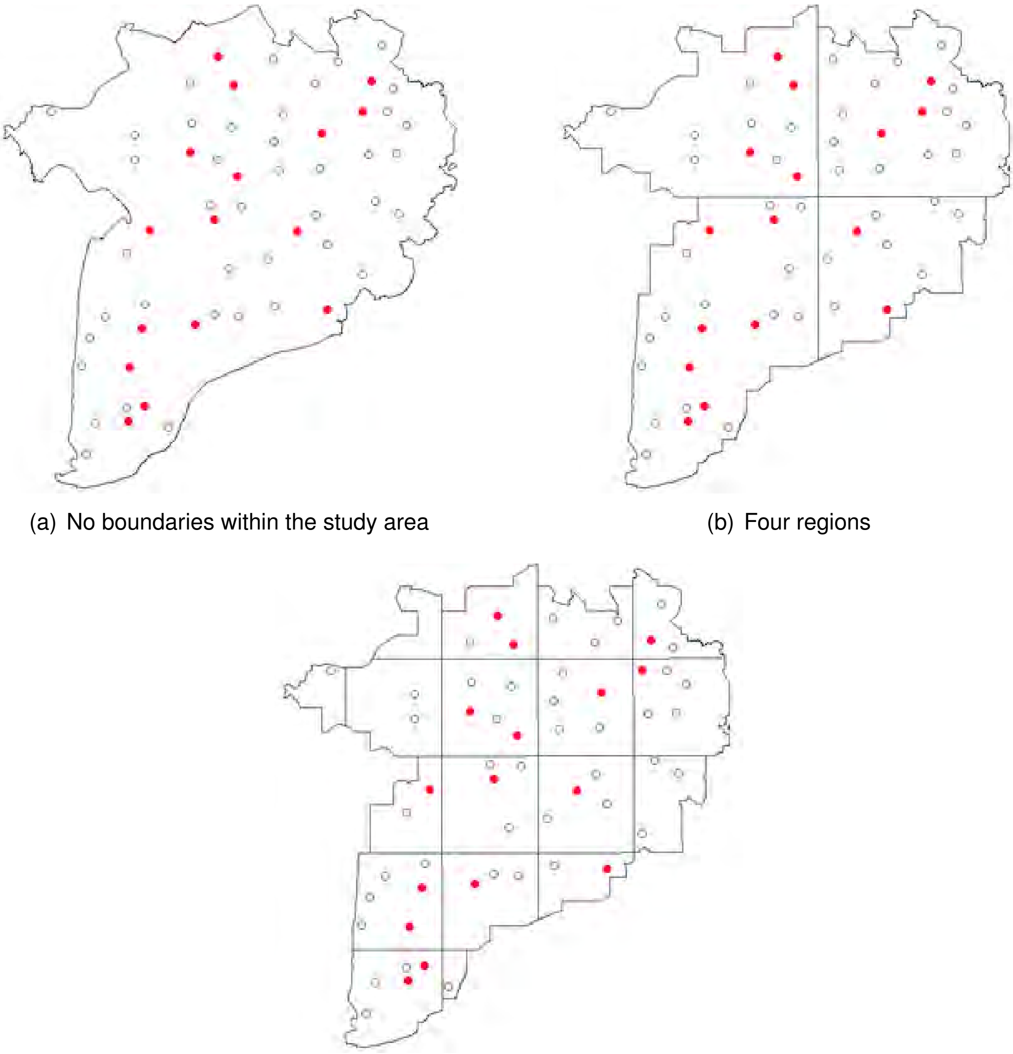
Figure 2.10: Choropleth map of district-level standardised mortality ratios (SMRs) for bovine spongiform encephalopathy in British cattle 1986 – 1997, for cattle born before the 18 July 1988 ban on feeding meat and bone meal to ruminants.

possible to measure the percentage of low-birthweight babies at a single point: this percentage must be calculated within a defined area. It is the selection of these artificial boundaries and their use in analysis that produces the MAUP.

The effects of the MAUP can be divided into two components: the scale effect and the zonation effect. The scale effect is the variation in numerical results that occurs due to the number of zones used in an analysis. As an example, the difference in numerical results between mortality rates by parish and by county in Great Britain is a scale effect.

The zonation effect is the variation in numerical results arising from boundary definitions that are used. For example, we might aggregate parishes within a county into larger areas made up of 30 parishes and report rates of disease for these arbitrarily defined areas. If we used groups of 20 parishes (rather than 30) the numerical differences in the rates reported would be a zonation effect.

It is necessary to understand the ways in which the MAUP affects the results of statistical analysis.



(a) No boundaries within the study area

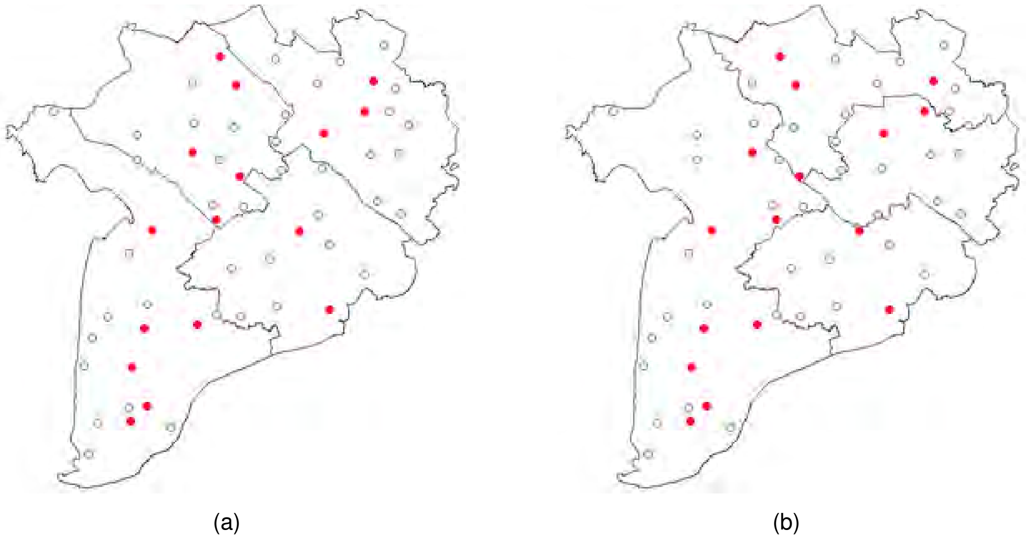
(b) Four regions

(c) 18 regions

Figure 2.11: Explanation of scale effects, for three areal units. A study area is comprised of farm holdings that are either disease positive (solid circles) or disease negative (open circles): (a) farm-level prevalence is 28% (16 of 57 farms are disease positive); (b) farm-level prevalence ranges from 18% to 40%; (c) farm-level prevalence ranges from 0% to 66%. As the size of each region gets smaller the variability of the prevalence estimate increases.

Caution is required, however, because there is a random aspect to the effects of the MAUP. It may be difficult to generalise about how different data sets with different spatial units are affected by the MAUP. This caution aside, the use of small areal units has a tendency to provide unreliable rates because the population used to calculate the rate is smaller. On the other hand, using larger areal units will provide more stable rates but may mask meaningful geographic variations evident with smaller areal units (Nakaya 2000). Choosing between the scale of zones depends upon the particular use and requirements of the data.

You should now have a general understanding of the MAUP and its effects on analysis. However, you will no doubt be wondering why no solutions to the problem have been advanced in this text. There are two reasons for this: (1) researchers have so far only begun to appreciate the effects of the MAUP on analysis; and (2) few generic and practical solutions exist.



(a) (b)

Figure 2.12: Explanation of zonation effects. The study area has been divided into four regions using two boundary definitions: (a) prevalence estimates range from 22% to 32%; (b) prevalence estimates range from 20% to 37%.

The weighting of areal units by population size, as well as complex statistical procedures are among the techniques currently being researched to address the MAUP. A simple strategy to deal with the problem is to undertake analysis at multiple scales or zones. So this is very much work in progress. Despite the lack of solutions, being cognisant of the fact that analysis results may be dependent on the zones used to aggregate data is an important step.

3 Sources of spatial data

3.1 Introduction

There are many sources of open-source and publicly-available spatial data repositories, and it can be time-consuming to identify specific data required for a project. It can also be frustrating, as often the descriptions (metadata) of downloadable spatial datasets is misleading, and the actual data cannot be accessed. Instead, only PDF maps can be downloaded, which are of little analytic value.

In addition, spatial datasets tend to be large. This results in long download times. Subsequently, some editing is usually required in the GIS software; the processing times for this, too, can be extensive. Don't underestimate the time required to prepare the spatial data for analysis.

In this chapter, we provide brief descriptions of some data sources that are likely to be useful for spatial analysis of livestock diseases. There are a number of supplementary data sources which provide statistical information which may be of use but are not explicitly spatial. These can have utility in spatial analyses. For instance, national livestock census data can be rendered useful by joining the lowest level of aggregation (e.g. commune, subdistrict or District) with administrative area shapefiles. The same can be done with socio-economic indicator data.

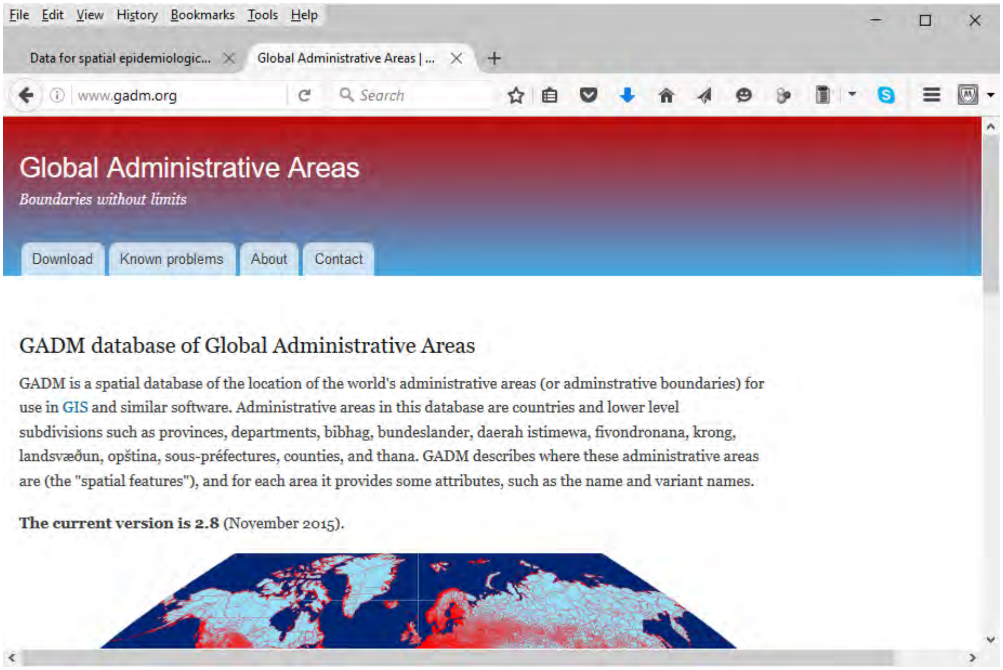
3.2 Geographic data

3.2.1 GADM database of Global Administrative Areas

The GADM database of Global Administrative Areas (http://www.gadm.org) is a database of digital maps of the world's adminstrative boundaries. Administrative boundaries in this database are countries and lower level subdivisions such as provinces, departments and communes. The GADM database is a resource of digital maps providing details of the boundaries of administrative areas (the 'spatial features') and, for each area, attribute information including area name and variant names. The current (September 2017) version of the GADM delimits a total of approximately 294,430 administrative areas. The data are available in ESRI shapefile, `.RData`, and Google Earth `.kmz` formats.

3.2.2 The OpenStreetMap project and data extracts

OpenStreetMapv (OSM) is 'a map of the world, created by people like you and free to use under an open license'. OSM is run by a non-profit foundation whose aim is to support and enable the development of freely-reusable geospatial data. It has a large and enthusiastic community base.


File Edit View History Bookmarks Tools Help
Data for spatial epidemiologic...
Global Administrative Areas | ...
www.gadm.org
Search
Download
Known problems
About
Contact

Global Administrative Areas

Boundaries without limits

GADM database of Global Administrative Areas

GADM is a spatial database of the location of the world's administrative areas (or adminstrative boundaries) for use in GIS and similar software. Administrative areas in this database are countries and lower level subdivisions such as provinces, departments, bibhag, bundeslander, daerah istimewa, fivondronana, krong, landsvæðun, opština, sous-préfectures, counties, and thana. GADM describes where these administrative areas are (the "spatial features"), and for each area it provides some attributes, such as the name and variant names.

The current version is 2.8 (November 2015).

Figure 3.1: The Global Administrative Areas website.

Unlike Google geographic data, these data can be freely downloaded and used; they can also be updated by users. OSM data are used in many websites, including navigational smartphone apps.

There are various ways to download and use OSM data in QGIS projects. A convenient way is to download an extract by region from the Geofabrik server (http://download.geofabrik.de). Data can be downloaded as either ESRI shapefiles or as OpenStreetMap zip files.

3.2.3 Populated places databases

There are a few public-facing databases of known populated places at country level, e.g.:

- Geonames geographical database: http://www.geonames.org
- National Geospatial-Intelligence Agency (NGA): http://geonames.nga.mil/gns/html/namefiles.html

The metadata are not very clear about the origin of these data. The definition of what constitutes a populated place, too, is unclear. Some of these data are very old and are likely to be inaccurate. Hence, cautious use is advised. However, they may be of some use, e.g. as a proxy for human population density (Figure 3.2).

3.3 Human geography, socio-economic and environmental data

There are many sources of data that are curated by a large range of institutions. These cover physical geography (e.g. land use and land cover, climate, elevation) as well as human geography (population, financial and economic indicators etc.).

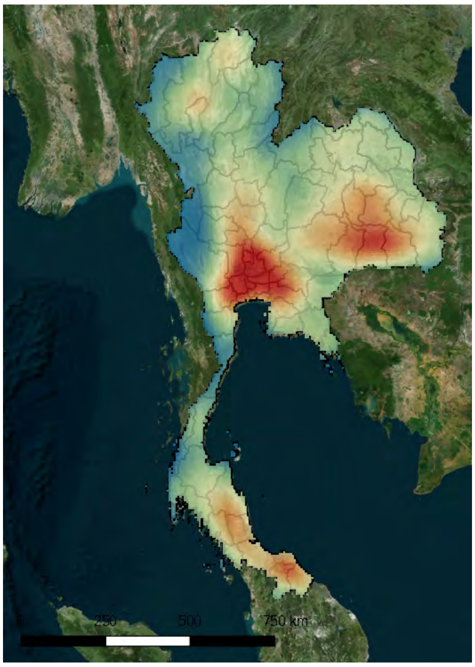


Figure 3.2: Heatmap showing density of populated places in Thailand (Source: NGA). This can be taken as a proxy for human population density.

Tip!

Unless you have a specific source of data, one way to identify potentially useful sources is to follow recommendations by others. Two examples we have found useful include

- Free GIS Data page: https://freegisdata.rtwilson.com
- GIS Geography page: http://gisgeography.com/best-free-gis-data-sources-raster-vector

Specific sources that may be useful include:

- UNEP Environmental Data Explorer: http://geodata.grid.unep.ch
- FAO GeoNetwork: http://www.fao.org/geonetwork/srv/en/main.home
- The World Bank DataBank (http://databank.worldbank.org/data/home) allows access and download of many subnational development indicators.

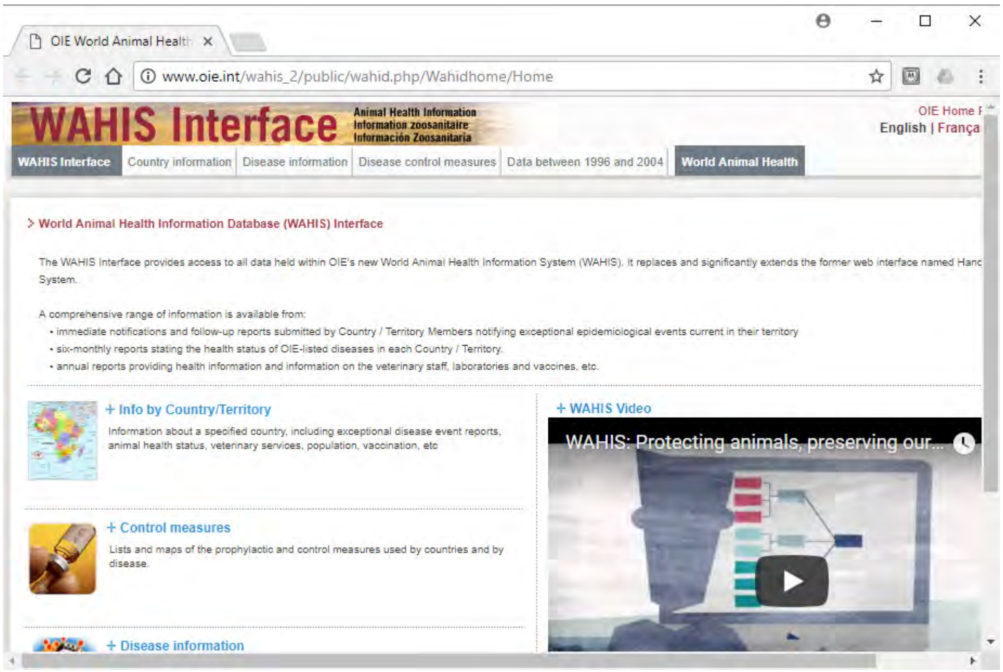
3.4 Livestock demographic and animal health data

3.4.1 The Gridded Livestock of the World

FAO's Gridded Livestock of the World raster maps provide estimates of livestock density (expressed as the number of animals per square kilometre) for grid cells of dimension 5 km by 5 km at the Equator. As of September 2017 raster maps of cattle, chickens (extensive and intensive), ducks, pigs (intensive and semi-intensive), sheep, goats are available. To download the data you'll need to create an account with a user name and password.

3.4.2 The World Animal Health Information System

The World Animal Health Information System (WAHIS) is a web-based application that has been designed to support veterinary services by facilitating the organisation and access to regional and global disease information. Timely and reliable access to disease information enhances early warning and response to transboundary and high impact animal diseases including emerging zoonoses. In addition it supports prevention, improved management and a progressive approach to disease control. To access WAHIS data, go to http://www.oie.int/wahis_2/public/wahid.php/Wahidhome/Home (Figure 3.3).



> World Animal Health Information Database (WAHIS) Interface

The WAHIS Interface provides access to all data held within OIE's new World Animal Health Information System (WAHIS). It replaces and significantly extends the former web interface named Hanc System.

A comprehensive range of information is available from:

- immediate notifications and follow-up reports submitted by Country / Territory Members notifying exceptional epidemiological events current in their territory
- six-monthly reports stating the health status of OIE-listed diseases in each Country / Territory.
- annual reports providing health information and information on the veterinary staff, laboratories and vaccines, etc.

+ Info by Country/Territory

Information about a specified country, including exceptional disease event reports, animal health status, veterinary services, population, vaccination, etc

+ Control measures

Lists and maps of the prophylactic and control measures used by countries and by disease.

+ Disease information

+ WAHIS Video

Figure 3.3: The World Animal Health Information System (WAHIS) home page.

The unique identifier, the observation date, the name of the locality in which the event occurred (and the longitude and latitude of the centroid of the locality in which the event occurred) are of most use for analytical purposes. As well as the number of affected animals, a strong feature is that the population at risk is also specified, which enables calculation of measures of outbreak severity.

3.4.3 Livestock census data

Livestock census data may be published on the web (e.g. via Ministry websites), or you may have access to these data directly through your job position. These data may be aggregated to District level or higher, but they may be available at village or commune level.

Such data may not be explicitly spatial, but it is apparent that by utilising georeferenced data at the same level as the census data, we can create a spatial layer which may be useful. This is best achieved by

1. Loading the spatial data layer in QGIS;
2. Preparing your census data in `*.csv` format, such that the relevant geographic indicator (e.g. District) matches exactly with the corresponding spatial layer;

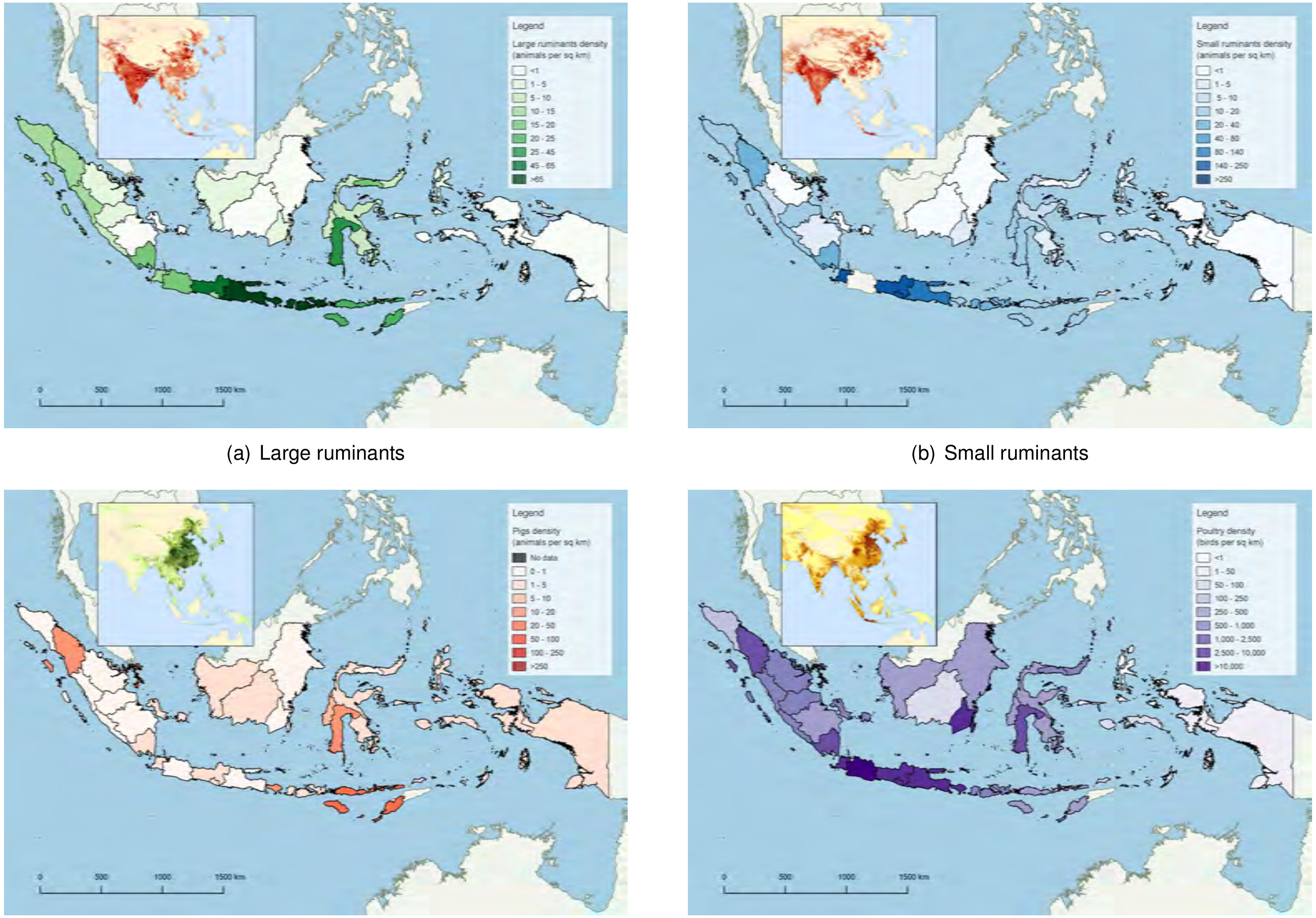
3. Importing this *.csv file into QGIS as an attribute-only table (i.e. one that has no geometry);
4. Creating a table join between the spatial layer and the data table.

Subsequently, choropleth or raster heatmaps can be made to visualise the livestock density. As an example, Figure 3.4 shows Province-level livestock density, comparing the outputs with the corresponding FAO GLiPHA map.

3.5 Geocoding locations

On occasions you will be provided with only the name of an outbreak location (without their corresponding longitude and latitude coordinates). In this situation, in order to define outbreak locations as a single point in space, it is necessary to geocode the outbreak location names. The resources page on the Veterinary Epidemiology at Melbourne University web site has tools for geocoding either a single or multiple addresses.

If you are using the multiple address facility your data will need to be saved as a Microsoft Excel (*.xlsx) file with two columns. The first column should be a numeric, unique identifier. The second column should list the address details as text. Because the geocoding engine uses Google Maps (with a limit on the number of addresses that can be geocoded on any given day without payment of a license) your spreadsheet should contain no more than 100 records.



(a) Large ruminants

(b) Small ruminants

(c) Pigs

(d) Poultry

Figure 3.4: Province-level livestock density in Indonesia, using public livestock census data.

4 Exploratory spatial data analysis (ESDA)

4.1 Introduction: exploratory data analysis (EDA)

If a dataset can be defined as “an organised collection of pieces of information about individuals”, then exploratory data analysis (henceforth abbreviated as EDA) can be considered as the initial interaction with a dataset which may identify important characteristics or trends that apply to the collection of individuals (generally referred to as the study population) as a whole.

The term EDA was coined by John Tukey, who published a book on it in 1977. He considered that EDA should be used by data analysts to examine their data sets. Rather than directly performing statistical hypothesis testing (confirmatory data analysis), the investigator should use EDA to formulate hypotheses and identify appropriate methods to subsequently test these.

EDA is not a specific technique so much as an approach that informs subsequent analysis. It incorporates different ways of summarising the data. A large number of terms may be used for this; these can be used interchangeably and it is not always clear how these terms relate to EDA. Such terms may include descriptive analysis or statistics; summary analysis or statistics; data visualisation or display; statistical graphics; etc. In recent years, the emergence of ‘big data’ and data mining have highlighted the application of EDA.

4.1.1 The importance of EDA

EDA is an important step in the analytic process, and it has several functions. It helps to enable us to understand the data.

EDA can be considered as an intermediary step between the collection of data and a formal analysis of these data (Figure 4.1). The optimal strategy for such formal analysis may not be clear from the outset; the specific structure and features of the dataset have a bearing on which analytic approach will be most effective. The outputs generated by EDA should be applied to assist with hypothesis generation and determining an appropriate strategy for this analysis. Indeed, in many cases this quantitative analysis will simply provide statistically robust evidence to support the trends detected by the EDA.

4.1.2 Aims and objectives of EDA

EDA is not a technique, but rather a variety of techniques that are employed as appropriate to

1. Generate insight into a data set (get a ‘feel’ for the data):
    - uncover underlying structure and identify important variables;

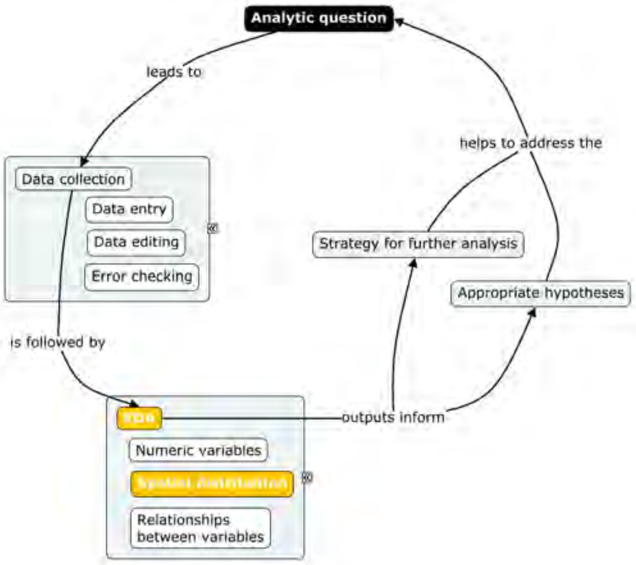


Figure 4.1: Concept map showing EDA as an element of statistical data analysis.

   - detect outliers and anomalies;
   - examine underlying assumptions (associations, confounding).
2. Identify potential discrepancies in the data (error checking).
3. Inform appropriate techniques for subsequent formal statistical modelling.
4. 'Prepare' the dataset for this analysis.

Many different data visualisation techniques can be applied to EDA (Figure 4.2). These are determined by the nature of the data (e.g. continuous or categorical); whether we are interested in univariate, bivariate or multivariate analysis (i.e. looking at one or comparing two or more variables with each other); and whether we are interested in temporal and/or spatial dimensions.

For numeric data, non-visual common EDA techniques include calculation of the 'five-figure summary': the two extremes (maximum and minimum, and hence the range), the median, and the quartiles. This summary adequately describes the distribution of these data. Note that boxplots are a good way to visualise and compare this summary!

4.2 Exploratory spatial data analysis (ESDA): EDA+

4.2.1 Introducing a spatial component into EDA

A fundamental principle of epidemiology is that for any given disease event, we have corresponding data related to:

- the *individual* (factors influencing why disease develops);
- the *time* (when?);
- the *place* (where?).

It is comparatively common to focus on the first (e.g. if we are interested in potential risk factors that are associated with the disease outcome). We often perform the second in a high-level way (e.g. stratification by year); more in-depth time-series analysis can follow if the data are suitably

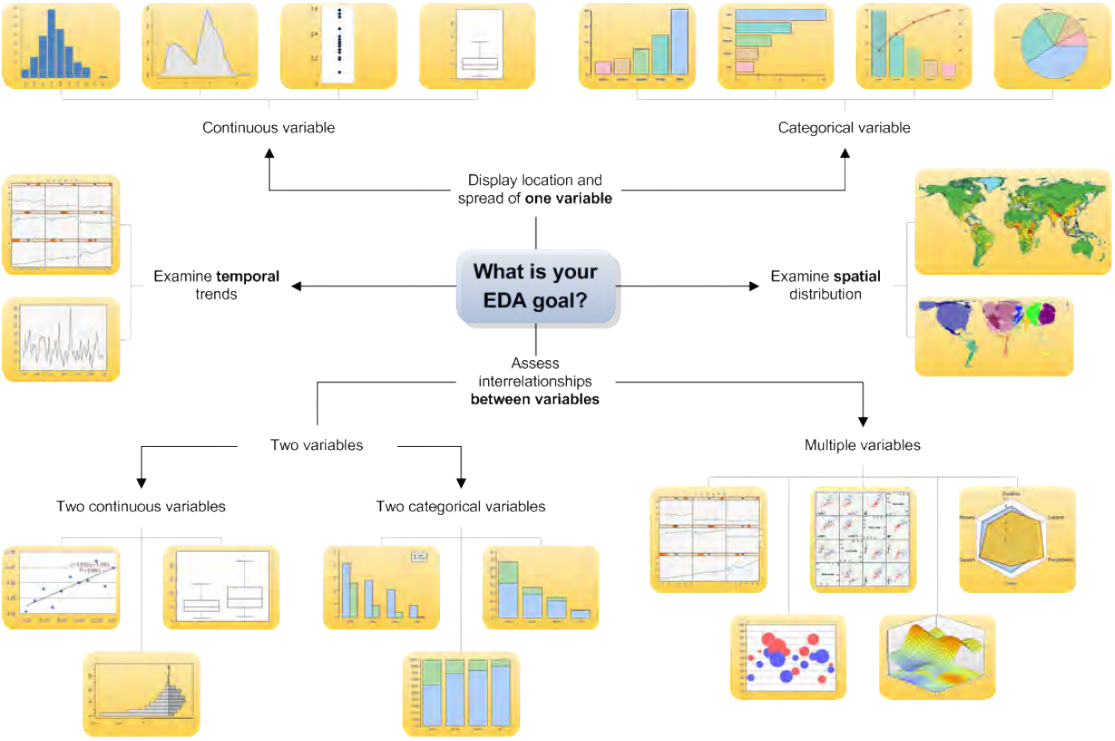

Continuous variable
Categorical variable
Display location and spread of **one variable**
Examine **temporal** trends
What is your EDA goal?
Examine **spatial** distribution
Assess interrelationships **between variables**
Two variables
Multiple variables
Two continuous variables
Two categorical variables

Figure 4.2: Decision chart to inform appropriate EDA data visualisation techniques.

detailed. However, the third is often overlooked if the dataset is not explicitly spatial (that is, contains coordinate data – see 4.2.3). Despite this, there is (albeit more limited) scope to do some spatial analysis. For instance, the data may contain an aggregated location variable (e.g. Distict or Province). Using open-source geographic data (see chapter 3), we are still able to develop useful visualisations.

In this section, we specifically consider the spatial component of EDA. This may be referred to as ‘exploratory spatial data analysis’ (ESDA). ESDA is no more or less than a special case of EDA, i.e. with a specific focus on spatial distribution of events. However, there are a few points of difference between EDA and ESDA.

4.2.2 What’s different about ESDA?

ESDA differs from EDA in a few aspects. The spatial distribution is not directly comparable to the distribution of a numeric data series. In both forms of exploratory analysis, outliers are of interest; in a numeric data series, these are unexpectedly high or low values, whereas in a spatial distribution, these are represented by ‘unusual’ locations. Conversely, in a numeric series, we are also interested in expressions of the central tendency of the data (i.e. the mean, median and mode). In a spatial context, the equivalent phenomenon is known as **clustering**.

A formal analysis of clustering is covered in detail in Chapter 5. In ESDA, we more or less limit the analysis to mapping of the data, visualising the data in different ways, and ‘eyeballing’ the distributions to assess whether there may be spatial trends. We may apply various types of stratification or choose visual display techniques that can accentuate the patterns that may be present in the data. Different drivers can influence the resulting distribution. For example, the spatial propagation of disease can be disaggregated into patterns of local spread (direct transmission), as opposed to spread due to trade and animal movement. These different dynamics are referred to as **spatial regimes**; although our data show a ‘mix’ of these dynamics, effective ESDA can uncover evidence of their existence, which can subsequently be investigated in more depth.

4.2.3 Types of spatial data and associated techniques for performing ESDA

Event data

In a spatial context, cases of disease are often referred to as **events**. These data refer to a specific location: they are point data. Therefore, each record includes coordinates.

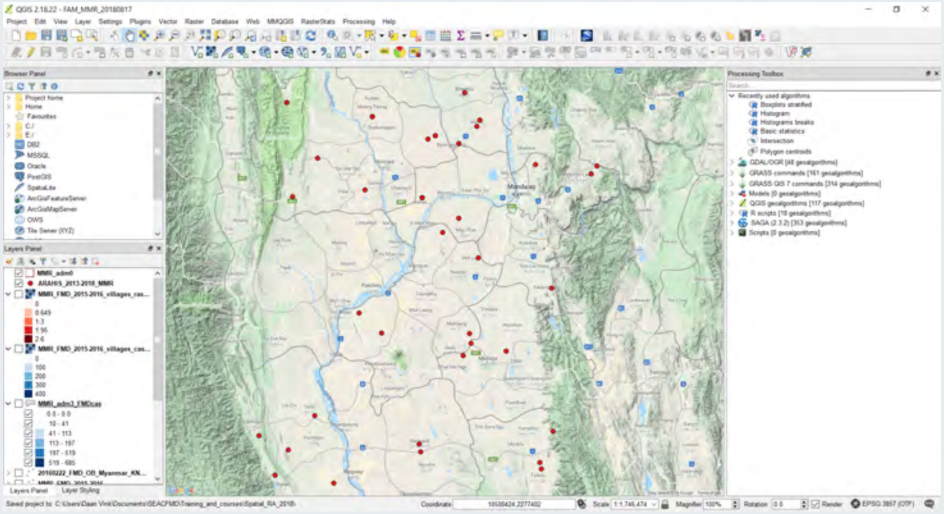
In the most atomic case, a record of event data will refer to an individual in a specified point location. In practice, one event record may contain information on multiple individuals (e.g. a herd of animals); the location may aggregated to a specified area, too. For instance, the level of observation may be at the farm or village level. In this case, the point location is expected to be representative of this area. Generally, the **centroid** is used.

The events may be of variable levels of accuracy, e.g. observed, confirmed (by laboratory diagnosis) or recorded / reported. As with other types of disease datasets, information on other variables may also be associated with the location of each record.

Example

The screenshot below shows a QGIS map view of ARAHIS data on reported FMD outbreaks, zoomed in on central Myanmar.

Each record represents a reported outbreak of FMD. The coordinates represents a village location in which the cases were observed. Each event contains data on the number of FMD cases, livestock species, and the population at risk (PAR) for each species; the date on which the first and last cases were reported; and, in some cases, the strains of FMD virus detected by sampling and laboratory analysis.



Saved project to: C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_RA_2018\

Coordinate 10500424,2277402 Scale 1:1,746,474 Magnifier 100% Rotation 0.0 ☑ Render EPSG:3857 (OTF)

Observing the distribution of points is a univariate assessment only. Bi- or even multivariate associations can be investigated using some or all of the following tricks:

- **Stratification by a second variable**. This can be a categorical variable (e.g. year of observation) or a continuous variable (e.g. attack rate of the outbreak). The latter is comparable to a histogram of continuous numeric data; the criteria for determining the break points becomes important.
- **Applying a colouring schema** according to another variable. This provides a visual cue if this second variable may also be clustered.
- **Sizing the points** according to another variable. This can be a highly effective method to visualise the variability in this variable.
- **Incorporating numeric techniques**. If the number of points is small, graphs such as histograms or pie charts can be included for each point.

Continuous data

Continuous spatial data are usually smoothed or constructed from event data. This represents a spatial distribution or 'surface' across the entire area under investigation. This surface is constructed by a computational process called **interpolation** (see section 2.8.3).

It follows that these are usually raster layers (see 2.4.2).

The resolution and accuracy of the interpolation depend on the number of observed or recorded events from which the surface is constructed. The degree of smoothing that is applied (referred to as the **radius**) depends on relevant characteristics of the event in question (e.g. dynamics of local spread) as well as the geographical scale on which the events are recorded.

Object data

Spatial object data contain information on specific **entities** which in themselves have no explicitly spatial meaning. They are discrete spatial features which may contain information associated with event data. They may be exhaustive (i.e. cover the entire geographical space) or can be associated with geographic coordinates.

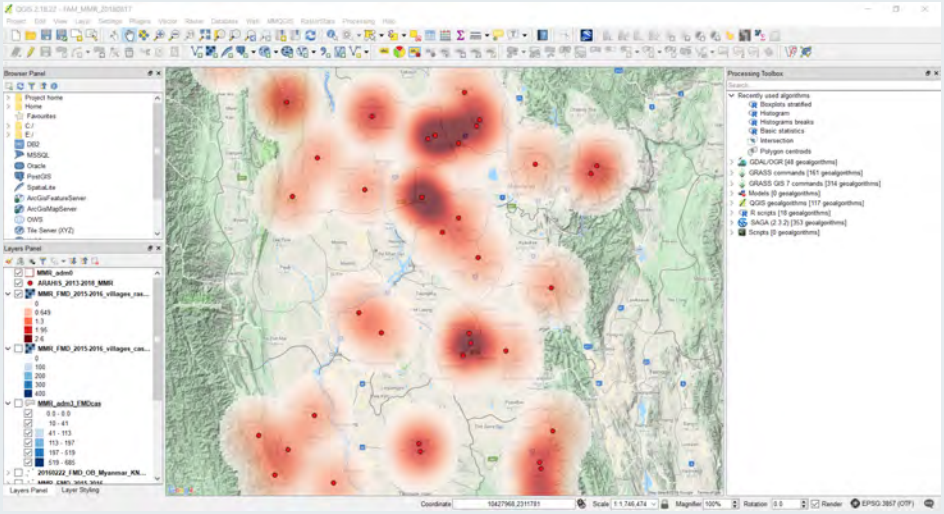
This form of data can be highly relevant because they are often associated with **aggregated** spatial data. The most common method of visualising an attribute in connection with such spatial object data is **choropleth maps**. In these maps, object features (which usually represent administrative areas) are shaded or patterned in proportion to the associated event data (for example, an attack rate or incidence rate of disease).

Another technique involves adjusting the area of each of each administrative division proportional to another variable; this is called a **cartogram**. While the resulting map may be distorted (sometimes grossly so), this provides another indication of variability across these objects.

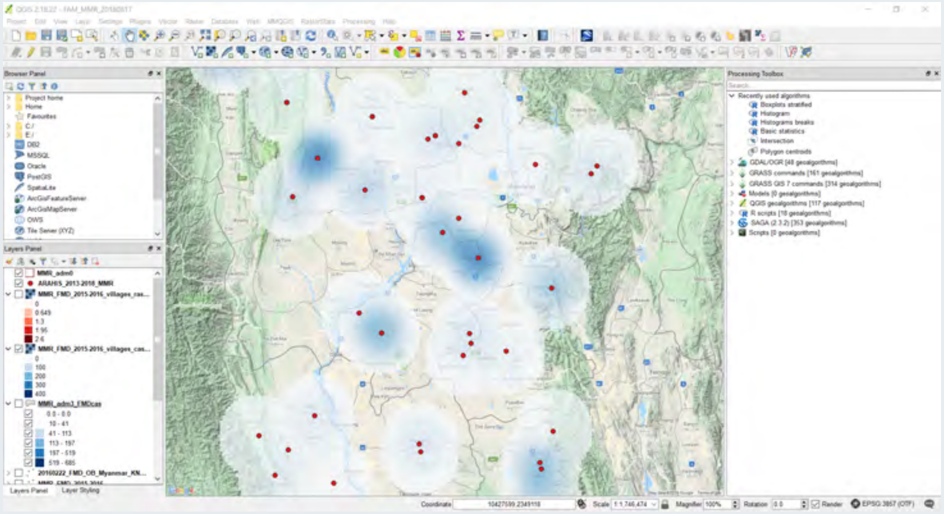
Note that aggregation of point event data to spatial object (polygon) data inevitably results in a loss of information. Once aggregation has been performed, the scope for disaggregation is limited. One method by which this can be done is by calculating a **centroid** of the polygons. While this is an approximation, it may sometimes be necessary to do so. This can be done on geographic grounds, but this can also be weighted by other criteria such as population density. If the polygon is not a convex shape, it is possible that the geographic centroid falls outside of the polygon; there are specific techniques which ensure that this is not the case.

Example

Continuing from the previous example, a raster heatmap was generated using the village point locations as the input layer. The radius chosen influences how concentrated or diffuse the surface is. Note that in this case, the heatmap reflects the density of the villages in which outbreaks were reported – it has no bearing on the numbers of cases or severity of the outbreaks!



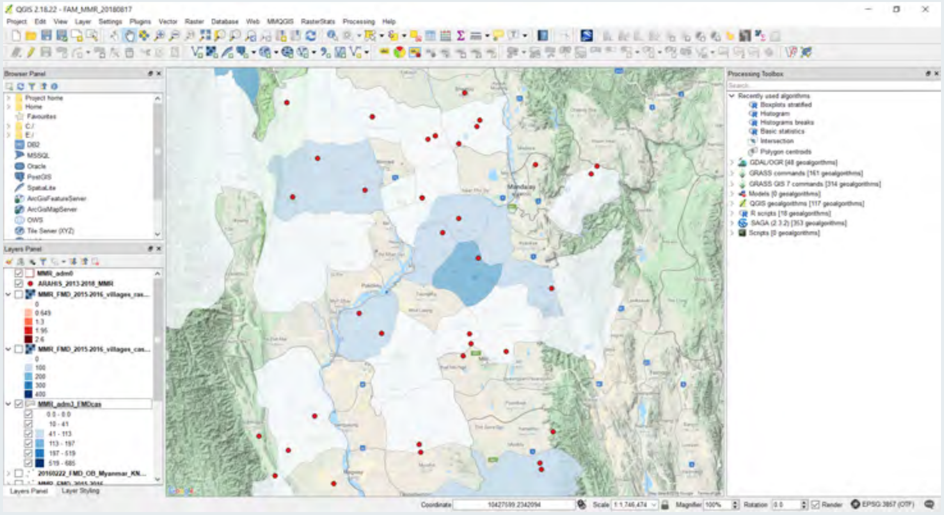
Such surfaces can be much more informative when they are weighted by another variable. The map view below shows the heatmap when each event was weighted by the number of FMD cases reported in large ruminants. This can be considered to be more representative of the distribution of disease.



Example

A very common type of spatial object data is administrative geopolitical boundaries. Frequently, the exact location of disease events could not be recorded, but information on the closest administrative level within which the event occurred (e.g. District of Province) can be retrieved.

Continuous data can also be extrapolated to spatial object data. Continuing from the previous example, a the median value of the raster heatmap of FMD density was computed for each District, and this was subsequently represented as a choropleth map. While less precise, such maps may be pragmatically useful as disease control decisions and prioritisation is often based on such political divisions.



4.3 A strategy for performing ESDA

As is clear from the previous sections, there is substantial overlap between a ‘standard’ exploratory data analysis and exploratory spatial analysis. From an epidemiological point of view, the differentiation between the two is artificial and arbitrary: measures of disease frequency, potential associations with risk factors, trends in time etc. are equally of relevance for any spatial analysis.

For this reason, it is strongly recommended to perform a ‘standard’ exploratory data analysis before commencing with the spatial analysis. This will give the investigator a good handle on the data, which can be subsequently augmented by the spatial assessment. The spatial analysis therefore deepens the understanding of the data in question: it should never be an end in itself.

A suggested workflow is as follows:

1. **Initial scan** (first pass):
   - inventarise what’s in the data (which variables are included);
   - get a crude idea of what state the data are in and how much work will be required to ready the dataset for analysis. For instance, an Excel workbook may consist of multiple sheets which are all formatted differently and may contain a lot of extraneous data.
2. **Prepare the data for analysis**. This includes reformatting, consolidation, error checking,

manipulation, creating new variables etc. It is easy to underestimate the time required to perform this! After completion, the data should be saved in a format that can be directly imported into GIS or statistical software. To avoid errors, only include the variables that will be analysed, and ensure they are complete.

3. Performing these steps may already have led to **questioning**: an elicitation of which variables could be analysed and what such information would help to clarify.
4. Subsequently, develop a **strategy** to answer these questions. Perform the EDA, initially in statistical software or Excel; then perform an ESDA using GIS software such as QGIS or ArcGIS.
5. Be systematic; record your findings and outputs; and reflect on their meaning. Don't be dogmatic – frequently, unexpected new questions will emerge as the EDA is being performed; these should be pursued. EDA provides a large scope to be innovative and creative, and a highly effective EDA will almost always result in a stronger subsequent analysis!

Example

Revisit the steps of investigating the reported cases of FMD outbreaks in Myanmar. At each step, consider whether the information was useful, and synthesize this to provide some information on FMD disease trends in the area.

Plotting the locations of FMD outbreaks, were there evident signs of spatial clustering?

Judging by the distribution of affected villages, on the basis of this very limited evidence, there does not appear to be much evidence of clustering. Two things are relevant here: firstly, we do not have information on the distribution of *all* villages – we really need to account for that before drawing this conclusion. Secondly, inclusion of the Google Physical map layer shows that outbreaks were limited to the plains. This is not surprising; again, we need to correct for the livestock density over the whole area before we can state this as a fact.

Did the surfaces provide more compelling information?

While there may be indications of a few potential clusters, the bandwidth (radius) and colour graduation used could influence this. The main conclusion here may be that outbreaks were reported throughout the whole area.

Was the choropleth map useful?

Not particularly; at best it may indicate that there was perhaps more FMD in the districts in the northern part. However, this is not convincing.

What else can we conclude?

The fact that FMD occurred throughout the area indicates that the disease is likely to be endemic here. However, additional data on the livestock populations, human demographics and potentially other risk factors is required to enable conclusions to be drawn with any degree of confidence.

5 Spatial clustering of disease

This text was summarised and modified from

- Chapters 4 and 5, Spatial Analysis in Epidemiology by Pfeiffer, Robinson, Stevenson, Stevens, Rogers and Clements, Oxford University Press, 2008.
- The SaTScan^{TM} User Guide for version 9.6 by Martin Kulldorf, March 2018.

5.1 Introduction: global versus local clustering

In general, spatial patterns of health data can be classified as regular, random, or clustered. The term ‘clustering’ is used to describe the spatial aggregation of disease events. However, to account for the distribution of the underlying population at risk or of various risk factors, a better definition is as follows:

Definitions

Spatial disease clustering is the spatial aggregation of health events after known influences have been accounted for.

More specifically, a disease cluster can be defined as ‘a geographically and/or temporally bounded group of occurrences of cases, of sufficient size and concentration to be unlikely to have occurred by chance’ (Knox, 1989).

The investigation of possible disease clustering is fundamental to epidemiology, with one of the aims being to determine whether the clustering is ‘real’ (statistically significant) and requiring further investigation or an intervention, or whether it is likely to be a chance occurrence, or whether it is just a reflection of the distribution of the population at risk. This is essential to ensure no disease clusters are missed, and conversely, to prevent wasting resources on spurious (false) clusters.

Investigating possible disease clusters requires a systematic approach and various cluster investigation protocols are available. Such protocols may include the collection of data on the disease frequency and potential risk factors; exploratory epidemiological analysis to scope disease trends in time and space; and structured application of global and local cluster detection techniques. Such techniques (and the output generated by them) can be complex. It can be challenging to ensure that the interpretation of the results (and the subsequent recommendations provided to the relevant decision-makers) are thoroughly considered, well-justified and clear, and underpinned by sound scientific evidence.

Epidemiologists differentiate between 'local' and 'global' clustering:

- **Global (non-specific) clustering methods** are used to assess whether clustering is apparent throughout the region under investigation but do not identify the location of clusters. They provide a single statistic that measures the degree of spatial clustering, the statistical significance of which can then be assessed. The null hypothesis for global clustering methods is simply that 'no clustering exists' (i.e. completer spatial randomness).
- **Local (specific) methods of cluster detection** define the locations and extent of clusters, and can be further divided into focused and non-focused tests:
    - *non-focused tests* identify the location of all likely clusters in the study region, while
    - *focused tests* investigate whether there is an increased risk of disease around a predetermined point, such as a livestock market.

As epidemiologists, of course we are also required to develop reasonable explanations and interpretations that explain why an identified disease clustering has developed. Clustering of a disease can occur for a variety of reasons including the infectious spread of disease, the occurrence of disease vectors in specific locations, the clustering of a risk factor or combination of risk factors, or the existence of potential health hazards such as localized pollution sources scattered throughout a region, each creating an increased risk of disease in its immediate vicinity.

Global measures of spatial clustering assume that the disease process is the same (stationary) for the whole area under investigation. Of course, this assumption is rarely met. Although we can investigate whether or not there is evidence for disease clustering across the area of interest as a whole, we are not able to understand local variability in disease patterns. Consequently, we may miss significant local clustering. Conversely, we are unable to exclude the parts of the area under investigation in which no significant clustering exists.

It is logical that the larger the region we're investigating, the higher the probability that there are regions with inherently different local relationships. Many factors contribute to this: livestock populations and densities, disease risk factors, trade and animal movements, geographic and climatological factors, disease control measures which may influence the propagation of infectious diseases, political factors et cetera. With very large spatial datasets (for instance, encompassing the entire south East Asia region), and particularly with large raster datasets such as remotely sensed images, global statistics such as Moran's I (5.2.2) run the risk of losing information on spatial autocorrelation since they summarize an enormous number of possibly dissimilar spatial relationships.

Local statistics overcome this problem by scanning the entire dataset, but only measuring dependence in limited portions of the study area, the bounds of which have to be specified. Thus, for clustering to be detected, it need not occur over the entire dataset, nor need it have the same characteristics throughout the study area. Covariate data can be included to increase the accuracy of the cluster estimation.

The aim is for such local statistics to estimate the characteristics of clusters, such as their location, size, and intensity. This can be difficult. Complications arise when dealing with variable population distributions across a study area, or lack of knowledge of the disease dynamics, or an incomplete understanding of relevant risk factors. This can affect the results, leading to a failure to detect clusters (accepting the null hypothesis of spatial randomness when it is false – a Type 2 error) or conversely, detection of clusters where in reality there are none (rejecting the null hypothesis of spatial randomness when it is true – a Type 1 error).

In this chapter, we first consider techniques to determine global spatial clustering, followed by local methods of cluster detection. We subsequently cover the most-commonly utilised tool to implement local clustering: SaTScan.

5.2 Global estimates of spatial clustering

5.2.1 Statistical concepts relevant to cluster analysis

Stationarity, isotropy, and first- and second-order effects

The concepts of stationarity, isotropy, and first-order (trend) and second-order (local) spatial effects are fundamental to cluster analysis. To summarise:

- A spatial process is termed *stationary* if the dependence between measurements of the same variable across space is the same for all locations in an area.
- If the dependence in a stationary process is affected by distance, but this is the same in all directions, the process is considered to be *isotropic*.
- *First-order effects* describe large-scale variations in the mean of the outcome of interest due to location or other explanatory variables.
- *Second-order effects* describe small-scale variation due to interactions between neighbours.

Monte Carlo simulation

Many tests for clustering use Monte Carlo simulation in order to determine the statistical significance of the cluster (i.e. does the observed spatial pattern differ significantly from the null hypothesis of complete spatial randomness?).

This involves:

- calculating the test statistic using the observed data, and then
- re-calculating it using a specified number (e.g. 99, 499, 999) of *simulated* data sets (or permutations);
- this simulated test statistic is subsequently used to generate the *expected distribution* of the test statistic under the null hypothesis;
- finally, the *likelihood* of obtaining the value for the test statistic derived from the observed data is then calculated, and expressed as the p-value.

As with other statistical tests, a Monte Carlo estimate of the p-value for a one-sided test is given by the proportion of test statistic values, obtained from the simulated datasets, that are greater than the value of the test statistic obtained when using the observed data.

As Monte Carlo methods rely on permutations of simulated datasets, slightly different p-values are obtained each time the test is run. However, using more simulations to estimate the distribution of the null hypothesis (e.g. 999 versus 99), means that smaller and more stable p-values can be calculated (e.g. p = 0.001 as opposed to p = 0.01). However, a problem with multiple testing is that the likelihood of wrongly rejecting the null hypothesis increases.

5.2.2 Methods for aggregated data

Autocorrelation statistics for aggregated data provide an estimate of the degree of spatial similarity observed among neighbouring values of an attribute over a study area. There are various autocorrelation statistics for aggregated data, four of which will are presented for your reference: Moran's *I*, Geary's *c*; Tango's excess events test and maximized excess events test.

Fundamental to all autocorrelation statistics is the weights matrix, used to define the spatial relationships of the regions so that those that are close in space are given greater weight in the calculation than those that are distant (Moran 1950). Neighbours can be defined based either on adjacency or distance. Methods based on adjacency (also known as contiguity) include rook contiguity (polygons are adjacent if they share a border), queen contiguity (polygons are adjacent if they share a border or corner), and higher-order contiguities (often called spatial lags) such as first-order (neighbours) or second-order adjacency (neighbours-of-neighbours). The concept of higher-order adjacencies is illustrated in Fig. 5.1. When using distance to define neighbours, polygons with their centroids located within a specified distance range are considered to be adjacent.

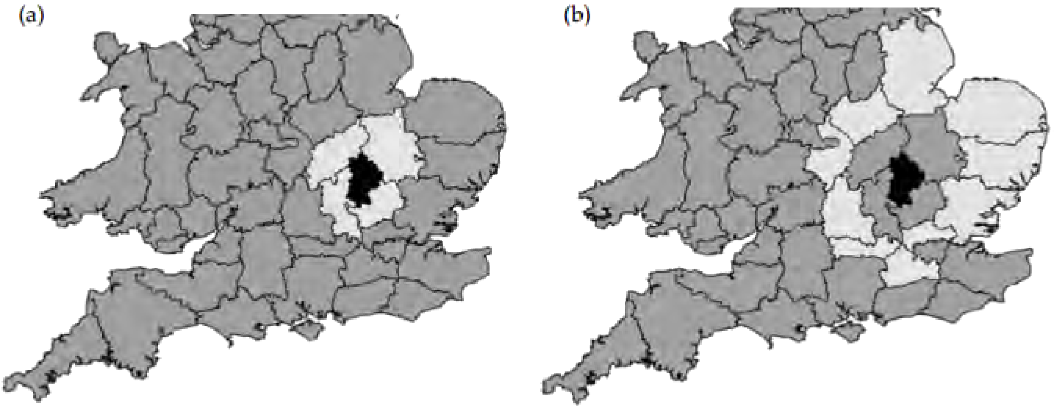


Figure 5.1: Maps illustrating (a) first-order adjacency (neighbours), and (b) second-order adjacency (neighbours-of-neighbours). In each instance the black county is the polygon of interest and the light-grey counties the defined neighbours.

Moran's *I*

Moran's *I* coefficient of autocorrelation quantifies the similarity of an outcome variable among areas that are defined as spatially related (Moran 1950). Moran's *I* statistic is given by:

$$I = \frac{n \sum_i \sum_j W_{ij}(Z_i - Zbar)(Z_j - \bar{Z})}{\sum_i \sum_j W_{ij} \sum_k (Z_k - \bar{Z})^2} \tag{5.1}$$

where Z_i could be the residuals $(O_i - E_i)$ or standardized mortality or morbidity ratio (SMR) of an area, and W_{ij} is a measure of the closeness of areas i and j. A weights matrix is used to define the spatial relationships so that regions close in space are given greater weight when calculating the statistic than those that are distant.

Moran's *I* is approximately normally distributed and has an expected value of $-1/(N-1)$ (where N equals the number of area units within a study region), when no correlation exists between neighbouring values. The expected value of the coefficient therefore approaches zero as N increases. Although Moran's *I* generally lies between +1 and -1, it is not bound by these limits. A

Moran's *I* of zero indicates the null hypothesis of no clustering, a positive Moran's *I* indicates positive spatial autocorrelation (i.e. clustering of areas of similar attribute values), while a negative coefficient indicates negative spatial autocorrelation (i.e. that neighbouring areas tend to have dissimilar attribute values). The significance of Moran's *I* can be assessed using Monte Carlo randomization.

Disadvantages of the autocorrelation test are the assumptions that the population at risk is evenly distributed within the study area, and that the correlation or covariance is the same in all directions (i.e. it is isotropic).

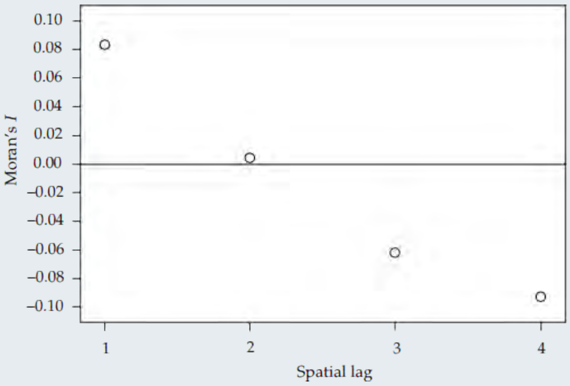
Although Moran's *I* is intended for use with continuous data it can also be used to analyse count data, even though, in such instances, any observed autocorrelation may simply be the result of variation in regional population sizes rather than any genuine spatial pattern in the disease counts (e.g. if regions with large populations are grouped together). Accounting for the population at risk by using regional incidence rates, instead of regional disease counts, increases the likelihood that any observed autocorrelation reflects a genuine spatial pattern rather than a heterogeneous population distribution.

A spatial correlogram is a series of estimates of Moran's *I* evaluated at increasing distances. Moran's *I* is plotted on the vertical axis and distance (spatial lag) is plotted on the horizontal axis. The correlogram can therefore be used to determine where, on average, spatial autocorrelation is maximized. Correlograms can be calculated based on distance or adjacency. There are various tests to determine whether a spatial autocorrelation value in a correlogram is statistically significant.

As Moran's *I* cannot adjust for a heterogeneous population density, Oden (1995) proposed the use of Oden's *Ipop*, a test statistic similar to Moran's *I* but in which rates are adjusted for population size.

Example

Moran's *I* was used to test whether there was clustering of counties in Great Britain in 1999 with respect to TB incidence rates. A weights matrix based on queen contiguity (i.e. polygons having a common border or corner) and a spatial lag of one (i.e. first-order adjacency) was defined and used in conjunction with Monte Carlo randomization (999 permutations), to calculate Moran's *I* for the data. These indicated the existence of non-significant, (p = 0.085) positive, spatial autocorrelation (I = 0.0832) in the TB incidence rates of neighbouring counties. In order to investigate spatial autocorrelation at higher-order spatial lags, weights matrices were defined and Moran's *I* statistics calculated for second- to fourth-order adjacencies, and used to plot a correlogram, which illustrates that for first- and second-order adjacencies spatial autocorrelation was positive, but negative for higher-order adjacencies. In other words, neighbouring or nearly-neighbouring counties (spatial lag of one or two) had similar TB incidence rates (either high or low), whereas counties further away from the one of interest (spatial lag of three or four) tended to have dissimilar incidence rates.



A correlogram showing Moran's *I* statistic computed for TB incidence rates in Great Britain in 1999 at first, second, third, and fourth-order spatial lags.

Geary's *c*

Geary's contiguity ratio, or Geary's *c*, is another weighted estimate of spatial autocorrelation (Geary 1954) but whereas Moran's *I* considers similarity between neighbouring regions, Geary's *c* considers similarity between pairs of regions. Geary's *c* ranges from zero to two, with zero indicating perfect positive spatial autocorrelation and two indicating perfect negative spatial autocorrelation, for any pair of regions. Geary's *c* is given by:

$$c = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{2\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)} \tag{5.2}$$

where n is the number of polygons in the study area, w_{ij} the values of the spatial proximity matrix, y_i the attribute under investigation, and the mean of the attribute under investigation.

Tango's excess events test (EET) and maximized excess events test (MEET)

Tango (1995) developed the excess events test (EET) to measure the 'closeness' among regions based on a distance matrix. Tango's EET is a weighted sum of excess events, as the statistic considers the difference between the observed rate of cases in each region and the expected rate, and then weights these differences by a measure of the distance between the regions, with a higher weighting given when the two locations are close.

Unfortunately Tango's EET requires the parameter λ (a measure of the spatial scale of clustering) to be specified, resulting in two problems. Firstly, λ is not generally known *a priori* and therefore several values of λ are tested creating issues of multiple testing. Secondly, choosing a large λ makes the test sensitive to geographically large-scale clustering while a small λ makes it more sensitive to small-scale clustering. To overcome these problems, Tango (2000) proposed the maximized excess events test (MEET). This test statistic searches for the value of λ which gives the smallest p-value of the observed value of the test statistic.

5.2.3 Methods for point data

Cuzick and Edwards' *k*-nearest neighbour test

Cuzick and Edwards (1990) developed a test for spatial clustering that takes into account the potentially heterogeneous distribution of the population at risk. It is based on the locations of cases and randomly selected controls from a specified region and includes a spatial scale parameter k, determined by the user. Scale in this instance refers to the number of nearest neighbours, and not geographic distance. For each case, the test counts how many of the k-nearest neighbours are also cases, such that if there are n_1 cases, and $m_i(k)$ represents the number of cases among the k nearest neighbours of case i so that $0 \leq m_i(k) \leq k$, for $i = 1, \ldots, n_1$, a test statistic T_k can be calculated as follows:

$$T_k = \sum_{i=1}^{n_1} m_i(k) \tag{5.3}$$

Thus, when cases are clustered, the nearest neighbour to a case tends to be another case and T_k will be large. However, when all cases have controls as their nearest neighbours T_k will be zero. The observed value of T_k can be compared with the distribution of values computed using Monte Carlo randomization of the dataset.

When data are available for the population at risk a modification of T_k is:

$$U_k = \sum_{j=1}^{n_1} (Y_j - E_j) \tag{5.4}$$

where circular regions are centred on each case and the radius of each circular region is chosen so that the expected number cases, E_j , is as close to the pre-defined value of k as possible, and Y_j is the number of cases within each region. Under the null hypothesis the expected value of U_k is equal to zero and the variance may be calculated.

Information on the exact locations of cases and controls is not always available, with locations instead being frequently assigned to the centre of administrative areas such as counties or parishes. As a result of assigning cases and controls to the same area 'ties' arise, precluding the calculation of Cuzick and Edwards' test statistic. Jacquez (1994) proposed an extension to Cuzick and Edwards' method that allows the test to be used in such situations.

A distinct advantage of Cuzick and Edwards' method is that it takes account of the heterogeneous distribution of the population at risk, as cases and controls are selected from the same population. In this way, the existence of any clustering of the population at risk, such as in built-up areas is accounted for. Furthermore, through the careful selection of controls, this method allows confounders to be accounted for. Disadvantages of the test include the fact that the user is required to select a value for the parameter k, and that interval data must be categorized as ‘case’ and ‘control’ locations resulting in a possible loss of information, although at the same time, this aspect of the test allows for flexibility in defining case and control locations. For instance, cases may be locations where disease has been detected or alternately, locations at which disease is above a specified threshold, such as the mean disease prevalence. Although Cuzick and Edwards’ test was originally developed for use with point data it can easily be adapted for aggregated data.

Ripley’s *K*-function

Second-order analysis describes the spatial dependence between events of the same type. The K-function is the most commonly-used method and identifies the distance at which clustering occurs. For an isotropic process with an intensity of λ points per unit area, the K-function at distance s may be defined as $K(s)$ such that $\lambda K(s)$ gives the expected number of events within a distance s of an arbitrarily-chosen event. Formally, $K(s)$ is defined as:

$$K(s) = \frac{1}{\lambda^2 R} \sum_{i+j} \sum I_s(d_{ij}) \tag{5.5}$$

where R equals the area of a region of interest, d_{ij} is the distance between the ith and jth events in R, and $I_s(d_{ij})$ is an indicator function which equals 1 if $d_{ij} - s$ and 0 otherwise. Where spatial autocorrelation is present, each event is likely to be in close proximity to other members of the same event type and, for small values of s, $K(s)$ will be large.

An important assumption of the K-function is that there are no first-order effects in the spatial pattern, as any evidence of spatial trend may influence the computed K-function. In addition, the variance of $K(s)$ increases with increasing distance, and therefore the K-function is suitable for estimating general tendencies toward clustering over distances that are small compared with the size of the region. As a rule of thumb it is recommended to restrict the range of s to no greater than 0.5 times the length of the shorter side of a rectangular study area.

Due to variations in the spatial distribution of the population at risk, a K-function computed only for cases may not be very informative. Instead, the K-function calculated for cases ($K_{case}(s)$) can be compared with one calculated for non-cases (or controls) ($K_{control}(s)$), with the difference between the two functions,

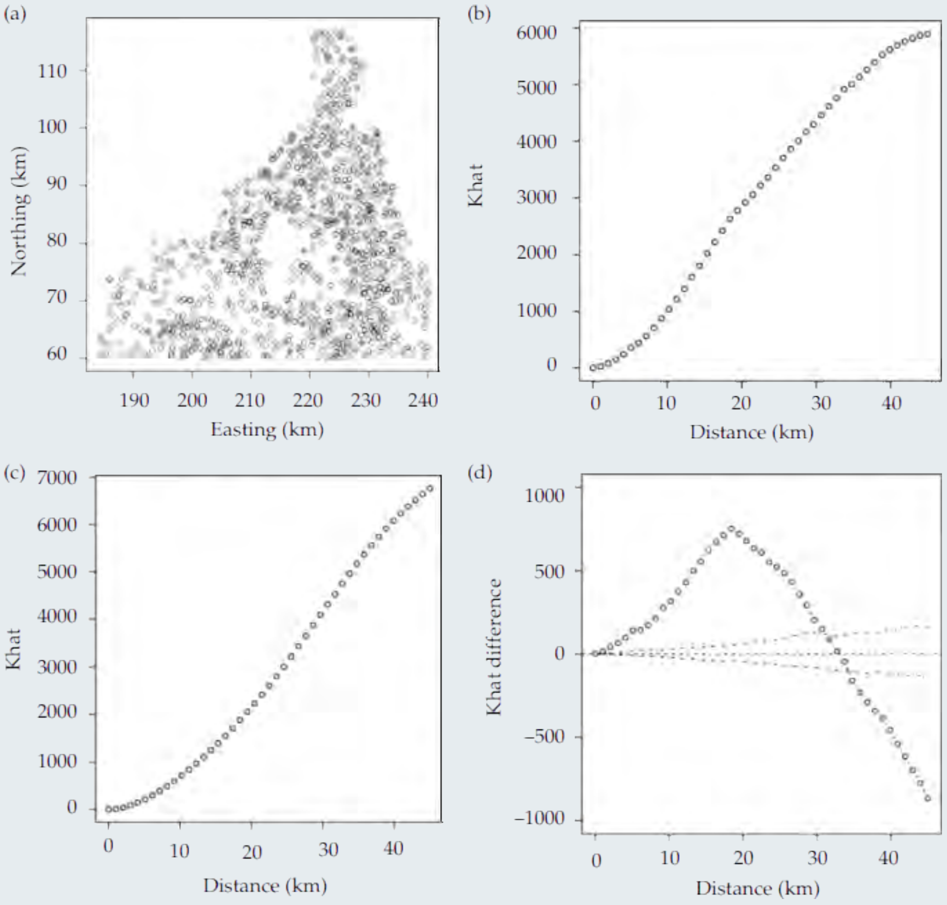
$$D(s) = K_{case}(s) - K_{control}(s) \tag{5.6}$$

representing a measure of the extra aggregation of cases over and above that observed for the non-cases. Monte Carlo randomization can then be used to randomly permute the locations of cases and non-cases/controls, and values of the difference function $D(s)$ computed for each permutation. The upper and lower bounds of these permutations are then plotted together with the observed difference function $D(s)$. Any deviation of $D(s)$ above the envelope formed by the upper and lower bounds indicates significant clustering of cases, relative to non-cases/controls.

Advantages of using the K-function to investigate clustering include the fact that it does not depend on the shape of the study region, and precise spatial locations of events are used in its estimation. It also takes into account the density of events in the region of interest, enabling spatial dependence to be compared among groups regardless of event prevalence. Furthermore, the test can accommodate an edge-correction weighting factor.

Example

The British cattle TB data were used to determine the distance over which clustering of TB-positive holdings was significant in a 60 60 km^2 area in the north-east of Cornwall in 1999. The TB-status of all holdings in this area is shown (a). K-functions were plotted for the TB-positive holdings (b) and for all the holdings (c) in the area of interest. Monte Carlo randomization was then used to randomly permute the locations of cases and non-cases, and values of the difference function ($D(s)$) computed for each permutation. The upper and lower bounds of these permutations were then plotted together with the observed difference function $D(s)$; this is shown in (d). Any deviation of $D(s)$ above the envelope formed by the upper and lower bounds indicates significant clustering of cases, relative to non-cases. In other words, compared with the spatial distribution of the holding population at risk, TB-positive holdings showed a greater tendency to be aggregated at distances of between 2 and 30 km, with maximum clustering occurring at a distance of 17 km.



(a) Easting and northing coordinates of holdings that tested positive (•) and negative (○) for TB in a 60 x 60 km^2 area of Cornwall, (b) K-function for TB-positive holdings, (c) K-function for all holdings, and (d) the difference between the two K-functions.

Rogerson's cumulative sum (CUSUM) method

Rogerson (1997) developed a cumulative sum (CUSUM) statistic for detecting changes in spatial pattern using a modified version of Tango's statistic. Whereas Tango's test is used retrospectively to identify clustering, Rogerson's modified version of the statistic aims to detect emerging clusters shortly after they occur, and can therefore be used for spatial surveillance. Owing to the problem of multiple testing, it would be inappropriate simply to re-calculate Tango's statistic after each new observation. Instead, once the test statistic has been determined for a particular set of observations, the expected value and variance of Tango's statistic after the next observation is estimated, based on the current value of the statistic. The expected value and variance is then used to convert the Tango's statistic that is observed after the next observation into a z-score, with all z-scores being incorporated into a CUSUM framework.

As Rogerson's CUSUM method is based on Tango's statistic, the test includes a measure of the spatial scale of clustering (λ) and thus, choosing a small value of λ makes the test more sensitive to small clusters and vice versa.

5.2.4 Investigating space-time clustering

While spatial patterns of disease are of great interest, space-time interactions are also important, particularly so when trying to determine whether a disease is infectious. In such instances it is necessary to evaluate whether cases that are close in space are also close in time and vice versa, adjusting for any purely spatial or temporal clustering. Various tests have been developed for this purpose; the tests for global space-time clustering are briefly reviewed here, while the local space-time cluster detection tests are dealt with in section 5.3.5.

It is worth noting that some tests look for clusters that relate to a fixed point in space, whilst others allow the spatial focus to move with time. In the first instance, the idea of a two-dimensional circular scan window is extended to that of a cylinder passing through time. In the second case, the geographical focus of a cluster may migrate with time, as long as it relates back to previous events.

There is also an important distinction between tests that require knowledge of the population at risk, and those that do not. Not requiring population or control data has obvious advantages in terms of ease of implementation, but the drawback is that it assumes that any change in the population at risk occurs evenly across the distribution under study. This would obviously have serious analytical implications, where interventions such as culling or vaccination may cause highly inhomogeneous changes in the population at risk.

The Knox test

Knox and Bartlett (1964) developed the first technique to identify spatio-temporal clustering of disease events and, although it has been the subject of much criticism, Knox's test has formed the platform from which subsequent tests have been developed. In this method, pairs of cases separated by less than a user-defined critical space-distance are considered to be near in space, and pairs of cases separated by less than a user-defined critical time-distance are said to be near in time. This classification allows pairs of points to be assigned to one of four cells in a 2 2 contingency table (near space — near time, near space — far time, far space — near time, far space — far time), and a test statistic, T_K is calculated as the number of pairs of cases that are near to one another in both space and time. The test statistic is compared against simulated results under a Poisson model, which Knox argues is the sampling distribution of the statistic in the absence of space-time clustering.

There are two principal limitations with Knox's test. Firstly, the choice of critical distances is subjective and secondly, that the critical distance in space does not vary with changing population density. This is unrealistic since the distance from case to case would decrease with increasing population density. Baker (1996) discusses the problems associated with specifying thresholds of proximity in space and time in the Knox test, and develops an adaptation that does not require unknown critical parameters to be specified, but instead allows for a range to be given for each. In its most flexible form the range can be specified from zero to the maximum space and time differences between any two pairs of cases. The test becomes more powerful as the ranges for thresholds can be specified with increasing accuracy, and reduces to the Knox test itself when the range for each parameter is reduced to zero.

Kulldorff and Hjalmars (1999) proposed a modification of the Knox test that overcomes these problems. Using the standard Knox test, significant space-time clustering is indicated at a range of critical distances whilst with their adaptation it is not. They show that changes in the distribution of the population at risk over time can have a strong influence on the standard test. These effects are particularly interesting with animal diseases where populations may be culled.

The space-time *K*-function

Diggle et al. (1995) extend existing second-order analysis methods for spatial data in order to investigate space-time interactions in point process data. They find second-order properties to be closely related to Knox's statistic. If $K_S(s)$ defines the K-function in space and $K_T(t)$ defines the K-function in time, the K-function difference $D(s,t)$ is:

$$D(s,t) = K(s,t) - K_S(s)K_T(t) \tag{5.7}$$

$D(s,t)$ estimates the cumulative number of cases expected within distance s and time-interval t of an arbitrarily-selected case attributable to the interaction between space and time. An alternative expression is:

$$D_0(s,t) = \frac{D(s,t)}{K_S(s) - K_T(t)} \tag{5.8}$$

which estimates, for given distance and time separations, the proportional increase in cases attributable to space-time interaction.

The Ederer-Myers-Mantel (EMM) test

Ederer et al. (1964) developed a cell occupancy approach for exploring space-time clustering whereby the study region is divided into a series of space-time sub-regions within which unusual distributions of cases are sought.

Mantel's test

Mantel (1967) reviews the methods of Knox and Bartlett (1964) and Ederer et al. (1964), and proposes a new test that compares inter-event distances in space and time against a null hypothesis that time and space distances are independent. The test statistic T_M is the sum, across all pairs of cases, of the spatial distances multiplied by the time distances. A transformation is used to reduce the effects of large space and time distances, which would not be expected to be correlated for contagious diseases. Significance levels are then tested using a standard Monte Carlo randomization process.

The EMM test and Mantel's test both have low statistical power for small numbers of study cases. True clinical disease excesses, as might result from proximity to a single pollution source, are more likely to be detected by the EMM method whilst the Mantel method is more likely to detect hotspots such as those due to a more general exposure to a putative source.

Barton's test

Barton et al. (1965) designed a test to detect changes in spatial patterns associated with the passage of time, based on analysis of variance and which tests the null hypothesis that these patterns do not change with time.

Jacquez's k nearest neighbours test

Jacquez (1996) developed a k nearest neighbours test for space-time interaction. The null hypothesis is of no association between time and space adjacencies (i.e. that the probability of two events being nearest neighbours in space is independent of the probability of their being nearest neighbours in time). This approach is based on the argument that geographic distance is not a good measure of spatial proximity in an epidemiological context.

5.3 Local estimates of spatial clustering

5.3.1 Introduction

Some of the more commonly-used methods for local cluster detection are reviewed below. They are divided into those primarily designed for aggregated data, and those for point data. Of course, point data can easily be aggregated and area data can be represented as points, for example by using the centroids of the areas.

The methods listed for point data are *'scanning statistics'*: they are based on variously defined circles that 'scan' the data for areas of elevated (or reduced) disease frequency. Such statistics are better suited to point, than to area data, since points fall clearly inside or outside a scan circle. Scan circles can be defined in terms of geographic distance (e.g. Openshaw's method), number of cases (e.g. Besag and Newell's method) or population size (e.g. Turnbull's method and Kulldorff's scan statistic).

Methods for investigating clusters around point sources are then reviewed, and finally methods that look for clusters in both space and time.

5.3.2 Methods for aggregated data

Getis and Ord's local G$i(d)$ statistic

Unlike Moran's I statistic (see Section 5.2.2), which measures the correlation between attribute values in adjacent areas, the G$i(d)$ local statistic (Getis and Ord 1992; 1996) is an indicator of local clustering that measures the 'concentration' of a spatially distributed attribute variable. The test statistic is calculated as:

$$\mathrm{G}i(d) = \frac{\sum_{j}^{n} w_{ij}(d)(x_j - \bar{x_i})}{S_i\sqrt{\frac{w_i(n-1-w_i)}{n-2}}}, j \neq i \tag{5.9}$$

where n is the number of areas within the region of interest and x_i is the observed value for area i

$$\bar{x_i} = \frac{1}{n-1} \sum_{\substack{j \\ j \neq i}}^{n} x_j \tag{5.10}$$

and w_{ij} is a symmetric binary spatial weights matrix:

$$w_i = \sum_{\substack{j \\ j \neq i}}^{n} w_{ij} \tag{5.11}$$

$$S_i^2 = \frac{1}{n-1} \sum_{\substack{j \\ j \neq i}}^{n} (x_j - x_i)^2 \tag{5.12}$$

It has been shown that E(G_i) = 0 and Var(G_i) = 1, and that the distribution of G_i under the null hypothesis of no spatial association among x_i is approximately normal. By comparing local estimates of spatial autocorrelation with global averages, the G$i(d)$ statistic identifies 'hotspots' in spatial data.

Moving towards exploring spatio-temporal clustering, Getis and Ord (1996) proposed the use of local statistics to quantify the pattern and intensity of spread of a disease away from the core of a hotspot by estimating a series of local statistics at different time periods. Local statistics can be used to estimate the intensity of a disease at various distances from a core location, and the time dimension can then be used to estimate the rate of spread.

Local Moran test

The local Moran test (Anselin 1995) detects local spatial autocorrelation in aggregated data by decomposing Moran's I statistic into contributions for each area within a study region. Termed Local Indicators of Spatial Association (LISA), the LISA statistic for each area is calculated as:

$$I_i = Z_i \sum_{\substack{j \\ j \neq i}}^{n} w_{ij} Z_j \tag{5.13}$$

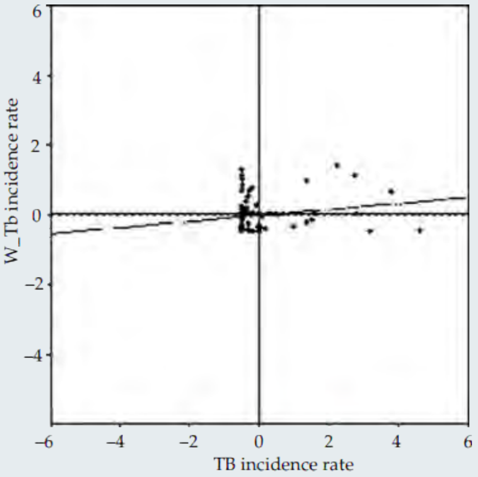
where Z_i and Z_j are the observed values in standardized form, and w_{ij} is a spatial weights matrix in row-standardized form.

These indicators detect clusters of either similar or dissimilar disease frequency values around a given observation. The sum of the LISAs for all observations is proportional to the global Moran's I statistic. There are two uses of LISA statistics: either as indicators of local autocorrelation or as tests for outliers in global spatial patterns in the form of a Moran scatterplot. In a Moran scatterplot the horizontal axis represents the vector of observed values and the vertical axis specifies the weighted average of neighbouring values. The extent of the 'mix' of pairs among the four types of association in the quadrants of the plot defined by the axes (low-high, high-high, high-low and low-low) provides an indication of the stability of the spatial association throughout the data. It may also suggest the existence of different types of association in different subsets of the data; for example, positive association in one area and negative association in another.

The following example continues on from the global Moran's I example presented in Section 5.2.2, which identified the existence of positive spatial autocorrelation among county-level TB-incidence rates in Britain in 1999 (I = 0.0832).

Example

Applying the local Moran test to the data allowed a visual representation of spatial autocorrelation to be obtained. The resulting Moran scatterplot showed that most of the points were in the lower left (low-low) and upper left (low-high) quadrants indicating the existence of both positive and negative spatial autocorrelation among county-level TB incidence rates in Britain in 1999. The negative spatial autocorrelation arose as a result of outlier counties with a low TB incidence rate being surrounded by neighbours with high TB incidence rates, while the positive spatial autocorrelation resulted from counties with a low TB incidence rate having neighbours with similar low incidence rates. This demonstrates that the LISA cluster map and LISA significance map are additional, useful methods of visualising the spatial autocorrelation.



Moran scatterplot of county-level TB incidence rates in Britain in 1999.

5.3.3 Methods for point data

Openshaw's Geographical Analysis Machine (GAM)

Openshaw's Geographical Analysis Machine (GAM, not to be confused with the Generalised Additive model statistical technique) was the first in a series of methods developed to explore disease data for evidence of spatial pattern (Openshaw et al. 1987). The GAM involves applying a fine grid across a study area and generating a series of circles of varying radii with their centres based at each intersection of the grid. The observed number of cases of disease within each circle is then compared with the expected number of cases, assuming the process under investigation follows a Poisson distribution. Circles that have a higher than expected occurrence of disease are retained, resulting in a large number of overlapping circles concentrated around 'disease centres'. Visual inspection is then relied upon to decide where clusters occur.

There are some severe computational limitations of the GAM methodology, as a consequence of which it is no longer the best option. The technique does not account for multiple testing, and therefore the change in radius and shifts in location of candidate clusters are not taken into account in the calculation of significance levels. Also, since distance is the method used to define the scan circles, circles of the same size can refer to different-sized populations and are therefore not directly comparable.

The primary purpose of mentioning the test here is that it was the predecessor of a series of tests for cluster detection based on scan circles, which addressed and overcame many of its deficiencies, adapting and improving upon its methodology.

Turnbull's Cluster Evaluation Permutation Procedure (CEPP)

Turnbull et al. (1990) developed the first test that was able to both locate and test the significance of disease clusters. The test, called the Cluster Evaluation Permutation Procedure (CEPP), creates a circular window for each area that contains a pre-determined number of individuals at risk, R. The number of individuals at risk in an area can be defined as p_i and the number of disease events as O_ii. If the number of individuals at risk (p_i), is less than R, then area i is included in the window and the area whose centroid is nearest to that of area i (say cell j) is included if $O_i + O_j < R$. If $O_i + O_j > R$ a fraction of the population of area j is added so that the total population at risk equals R. If $O_i > R$ then the window contains only a fraction of area i. For the fractional area included in a window, cases are allocated in the same proportion to the window as that of the population of the cell. A series of i overlapping windows is created with a population at risk of constant size, R. Turnbull's test statistic, as a function of R, equals the maximum number of cases in each of the windows. Monte Carlo simulation is used to evaluate the significance of the observed test statistic.

One of the conditions of this procedure is that cluster size (in terms of number of individuals at risk) must be defined *a priori* for the procedure to be valid.

Besag and Newell's method

Besag and Newell (1991) developed a method to overcome the problem in the GAM whereby circles of the same size can refer to different-sized populations and are therefore not directly comparable. Their test allowed the user to specify k, the expected cluster size. Typical values for k range between 2 and 10 for rare diseases. Each area with non-zero cases is considered in turn as the centre of a possible cluster. When evaluating an area it is labelled as 0 and the remaining areas are ordered according to their distance from area 0 and labelled 1, 2, . . . , $i - 1$. Using the notation defined in the previous section (O_i is the number of disease events) the statistic D_i is calculated such that $D_i = \sum_{j=0}^{i} O_j$ and $D_0 \leq D_1 \leq \ldots$ are the accumulated number of cases in cells 0, 1, . . . and $u_0 \leq u_1 \leq \ldots$ are the corresponding accumulated number of individuals at risk. M = min $i : D_i \geq k$ so that the nearest M areas contain the closest k cases. A small observed value of M indicates a cluster centred at cell 0. If m is the observed value of M, then the significance level of each potential cluster is:

$$\begin{gathered}\Pr\{M \leq m\} = 1 - \sum_{s=0}^{k-1} \exp(-u_m Q) \\ (u_m Q)^s / s!\end{gathered} \tag{5.14}$$

where Q equals the total number of cases in the study area divided by the total population at risk. The test statistic of clustering in the study area, T_{BN}, equals the total number of individually significant clusters, for example, at $p < 0.05$. The significance of the observed T_{BN} can be determined by Monte Carlo simulation.

Two limitations of this method, common to many other cluster detection tests, are firstly, the *a priori* choice of cluster size and secondly, the problem of multiple testing arising from the large number of potential clusters. Since the calculations to determine significance are based on the specified number of nearest-neighbour cases, selecting its value is an important issue. If it is too small then larger clusters cannot be detected, and if it is too large spurious clusters may be identified.

Alexander et al. (1991) adapted Besag and Newell's method into the nearest neighbour areas (NNA) test for local clustering.

Comparative performance assessment has showed that Besag and Newell's test and Kulldorff's scan statistic (5.3.3) give similar results but that the scan statistic is more likely to identify clusters in sparsely populated areas.

Kulldorff's spatial scan statistic

Building on Openshaw's GAM (5.3.3) and a generalization of Turnbull's CEPP (5.3.3), Kulldorff developed the spatial scan statistic (Kulldorff and Nagarwalla 1995), which brings together the advantages of each technique.

For each specified location, a series of circles of varying radii is constructed. Each circle absorbs the nearest neighbouring locations that fall inside it and the radius of each circle is set to increase continuously from zero until some fixed percentage of the total population is included. For each circle the alternative hypothesis is that there is an elevated risk of disease within the circle compared to that outside.

The test statistic T_{KN} is calculated as:

$$\begin{aligned} T_{KN} = \overset{\sup}{Z} \left(\frac{O(Z)}{p(Z)}\right)^{n(Z)} \left(\frac{O(Z^c)}{p(Z^c)}\right)^{n(Z^c)} \\ I\left(\frac{O(Z)}{p(Z)} > \frac{O(Z^c)}{p(Z^c)}\right) \end{aligned} \tag{5.15}$$

where Z^c indicates all circles except for Z, $O(\cdot)$ and $p(\cdot)$ are the observed number of cases and the population size in each area respectively, and $I(\cdot)$ is the indicator function. Monte Carlo simulation is conducted to compare T_{KN}, with the distribution of values generated under the null hypothesis.

Kulldorff implemented the spatial scan statistic in the SaTScan software, which searches for clusters in datasets using two different probabilistic models; a Bernoulli model where cases and controls are compared as Boolean variables, and a Poisson model where the number of cases is compared to the background population data and the expected number of cases in each unit is proportional to the size of the population at risk. Circle centres are defined either by the case and control/population data or by specifying an array of grid coordinates. Secondary clusters are computed, based on the degree of overlap allowed in the cluster circles, and includes the options no geographical overlap, and no cluster centres in other clusters.

The single most important subjective choice that has to be made when using the spatial scan statistic is specification of the maximum percentage of the population at risk (between 1 and 50%) that can be included in any one cluster. The SaTScan manual recommends specifying a high upper limit (i.e. 50%) of the population at risk, since SaTScan will then look for clusters of both small and large sizes without any pre-selection bias in terms of the cluster size. When looking for clusters of high rates a cluster of larger size indicates areas of exceptionally low rates outside the circle rather than an area of exceptionally high rates within the circle.

Until recently a major limitation of the spatial scan statistic was the use of a circular scanning window which decreased the likelihood of the statistic detecting non-circular clusters. However, the most recent version of SaTScan can implement an elliptic version of the spatial scan statistic, which uses a scanning window of variable location, shape, angle and size, thereby greatly increasing the ability of the statistic to detect non-circular clusters. Choosing the shape of the ellipsoid will depend on the nature of the data and the questions being asked, and the authors stress that the type of ellipsoid to be used must be chosen before running the model in order to prevent any pre-selection bias. For example, a long, narrow ellipsoid might be used to investigate potential clusters along a riverbank. The circular and elliptic scan statistics have similar power, with the circular scan statistic able to detect elliptic clusters and vice versa, although the elliptic scan statistic may provide a better estimate of the true area of the cluster. Although the elliptic scan statistic is more flexible than the circular scan statistic it still imposes a shape on potential clusters and therefore, for irregularly-shaped clusters (e.g. along a winding river) it may be more appropriate to use one of the non-parametric spatial scan statistics described in Section 5.3.3.

The spatial scan statistic has now been applied to an very wide range of health-related problems. This is due in part to its addressing many of the problems of earlier scan statistics but also because SaTScan, the freely available primary software for implementing the spatial scan statistic, makes it accessible to a wide, non-specialized audience.

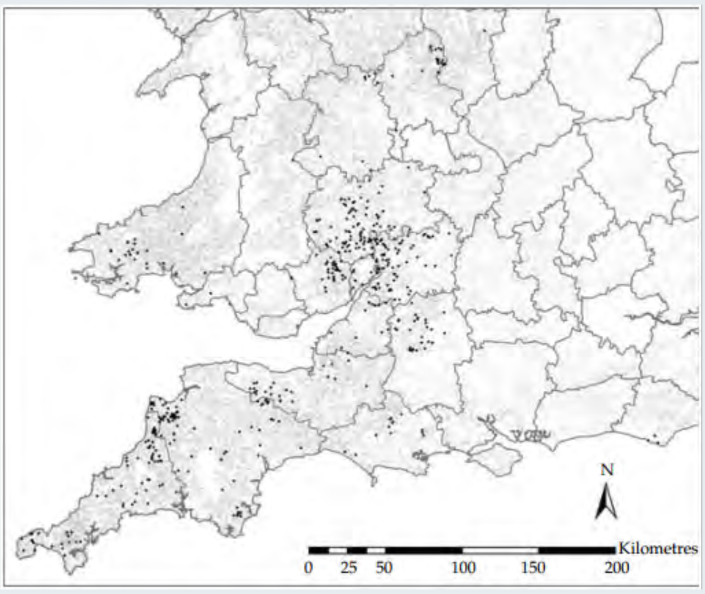
Non-parametric spatial scan statistics

A significant restriction imposed by all of the methods reviewed here is that they assume disease clusters are circular. Such scan statistics have a low power for detecting irregularly-shaped clusters, and may in fact identify an irregularly-shaped cluster as a series of small circular clusters. As these conditions seldom occur in reality, a number of investigators have tried to overcome this problem by developing a variety of tests that can locate irregularly-shaped disease clusters. Mention is only made here of the work of Tango and Takahashi (2005), who developed a flexibly shaped spatial scan statistic for detecting irregular-shaped clusters within small areas of a region, which can be implemented using the FlexScan software.

Example

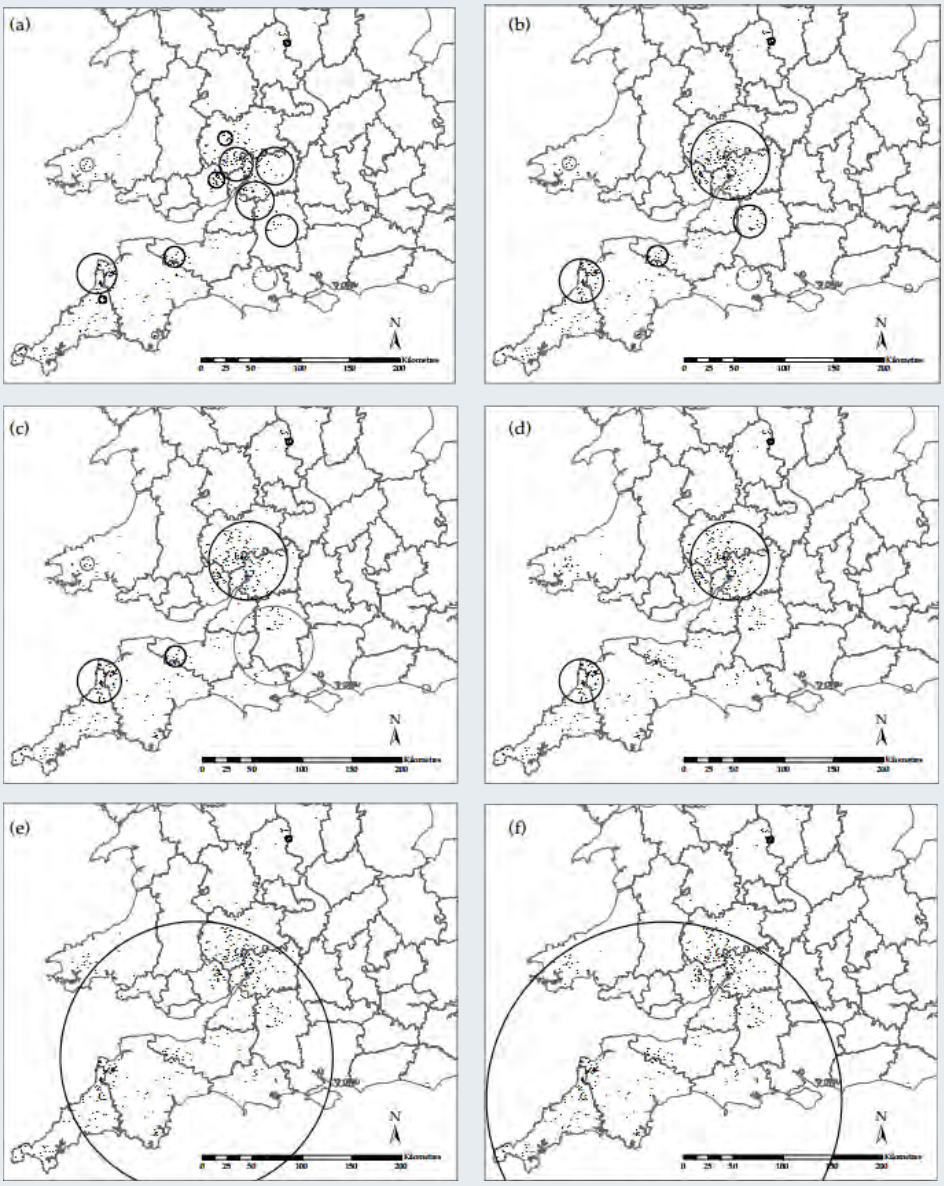
To demonstrate local cluster detection, Kulldorff's spatial scan statistic was applied to the British cattle TB breakdown herd data (1986-1997). Due to the nature of the TB data, incidence rates cannot be accurately derived, and therefore the analysis was restricted to the use of the Bernoulli model. All TB breakdown herds were included as cases, and compared against a sample of control herds and the combined data were used to define the circle centres for cluster detection (rather than specifying a special set of grid coordinates). Based on some preliminary analyses a 30% sample of control herds was chosen, above which no real change in the pattern of clustering occurred. The most restrictive option was selected for dealing with overlapping clusters, in which secondary clusters were reported only if they did not overlap with a previously reported cluster.

The map shows the distribution of herds (small grey dots) and of breakdown herds (larger black dots) in the south-west of Britain in 1997. In order to compare various options for implementing the spatial scan statistic, the 1997 dataset was used as this was the year with most recorded breakdowns (amongst those years within the dataset).



Distribution of herds (small grey spots) and tuberculosis breakdowns (large black spots) in the southwest of Britain in 1997.

The results of specifying different upper limits for cluster size when using the British cattle TB data for 1997 (with a 30% sample of control herds) are shown below, in which the selected upper limits were (a) 1%, (b) 5%, (c) 10%, (d) 20%, (e) 30% and (f) 50%.



Cluster analysis comparing different maximum percentages of the population at risk to be included in a cluster, using all TB breakdown herds as cases and a 30% sample of control herds, in the southwest of Britain, in 1977. Maximum percentage of the population at risk to be included in clusters was a) 1%, b) 5%, c) 10%, d) 20%, e) 30% and f) 50%. Bold black circles indicate P-values of <0.001; fine grey circles indicate P-values between 0.001 and 1.00.

The 1% upper limit produced ten highly significant ($p < 0.001$) and ten less significant ($1.0 > p > 0.001$) clusters. The 50% upper limit produced one enormous, highly significant cluster, one tiny highly significant cluster and one very small, less significant cluster. Moving from an upper limit of 1% towards one of 50% produced smaller numbers of increasingly large and increasingly significant clusters. Which of these patterns of clustering best reflects the epidemiology of TB in the southwest of Britain cannot be determined from this analysis.

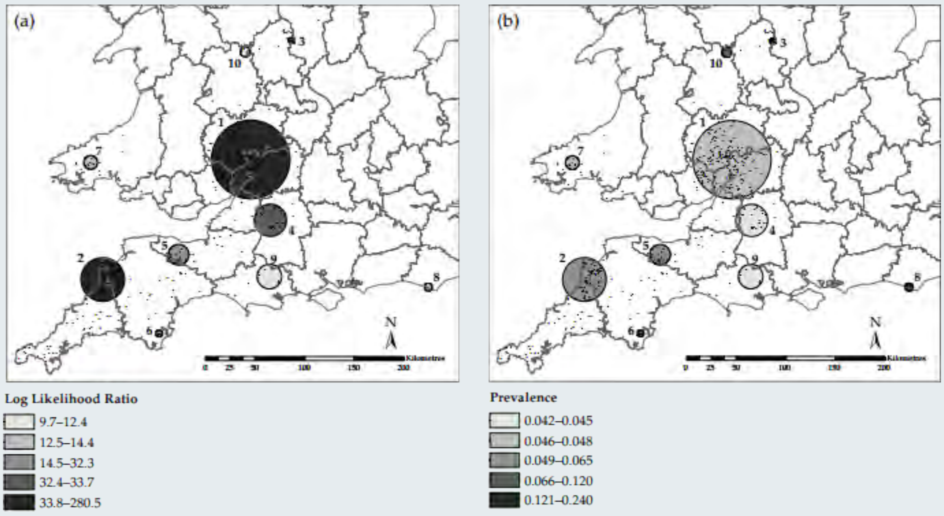
To explore the outcomes of the cluster analysis in greater detail consider the table below and the associated figure. The table shows some statistics for the (arbitrarily chosen) 5% maximum cluster size of the 1997 TB data.

Table. Details of the spatial clusters illustrated in map (b) above and the maps below. For each cluster the area is given, LLR is the log likelihood ratio, P (999) is the significance level based on 999 Monte Carlo replications (hence the plateau at 0.001), RR is the relative risk (observed/expected), which is directly proportional to Prev., (the prevalence corrected for sampling)

Cluster	Area (km^2)	LLR	P (999)	RR	Prev.
1	4,879	280.5	0.001	8.3	0.048
2	1,516	115.8	0.001	9.7	0.056
3	32	33.7	0.001	35.7	0.206
4	821	32.6	0.001	7.3	0.042
5	337	32.4	0.001	11.2	0.065
6	37	18.4	0.002	20.8	0.12
7	149	14.4	0.01	8.2	0.048
8	52	13.3	0.03	41.6	0.24
9	479	12.5	0.052	7.8	0.048
10	82	9.7	0.427	15.3	0.088

It is quite clear that some of the most significant clusters are of relatively low risk and vice versa. Cluster 8 for example, has by far the highest prevalence but is barely significant, while the similarly sized Cluster 3 also has a high prevalence but is highly significant. The very large, highly significant and so called primary cluster (Cluster 1), only has a relatively low prevalence.

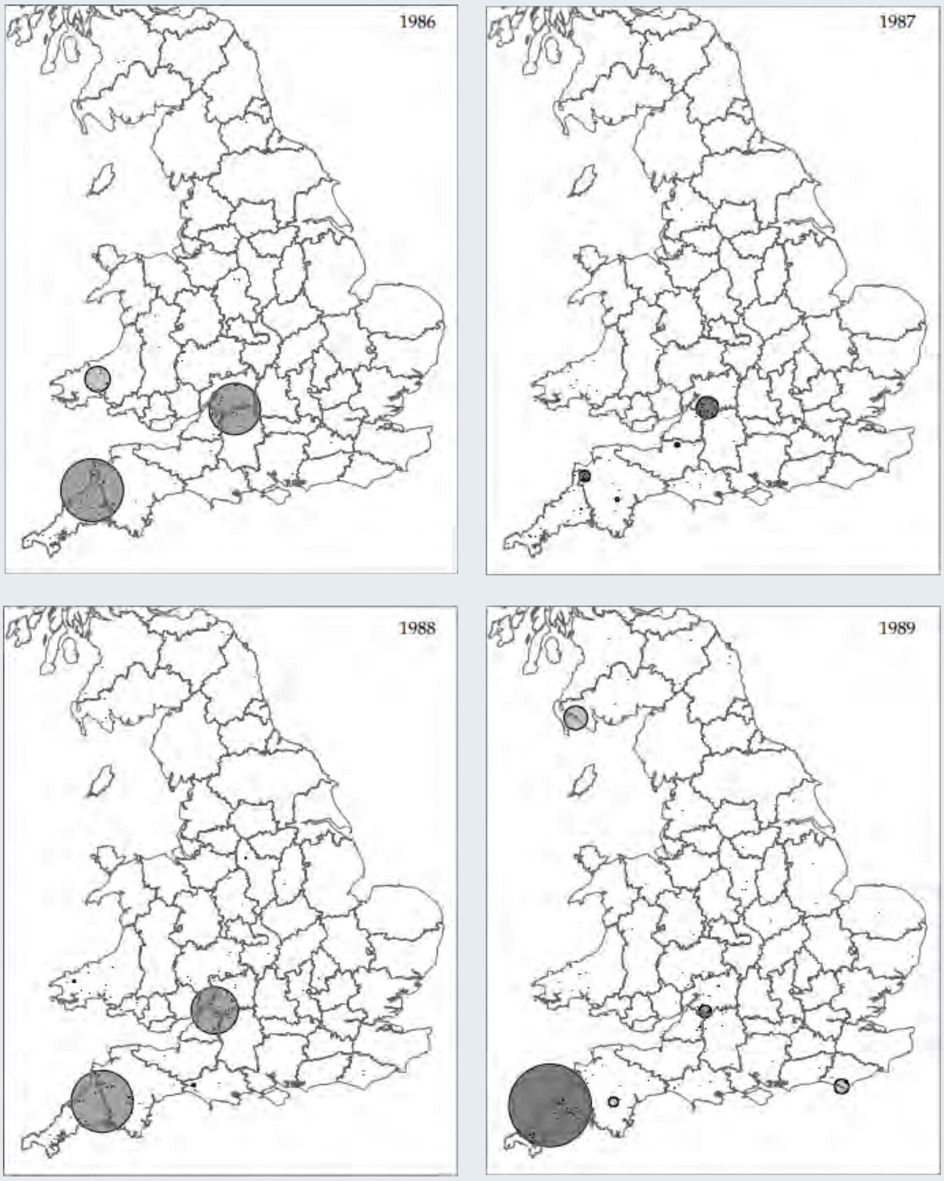
The figure illustrates these points through two contrasting representations of the analysis: the log likelihood ratio, which indicates the probability of a cluster being real (a), and the prevalence (b), which is directly proportional to the relative risk within a cluster but corrected for sampling of the population.



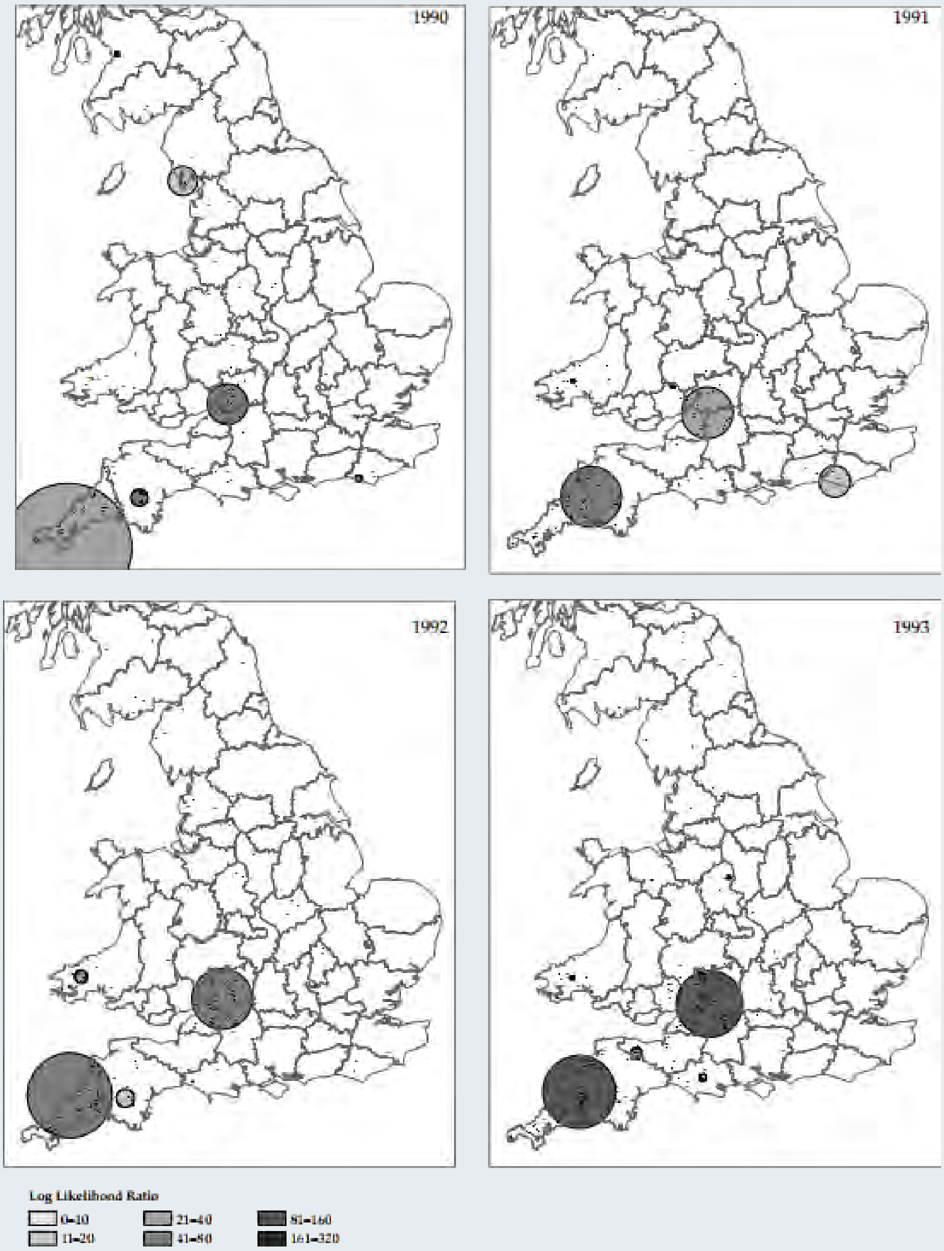
Detailed cluser analysis showing (a) log likelihood ratios (i.e. probability) and (b) prevalence (directly proportional to relative risk) of clusters of TB breakdown herds in the southwest of Britain, in 1977. Clusters were identified using the Bernoulli model using all TB breakdown herds as cases compared to 30% of control heads, with a maximum cluster size set to 5% of all points.

Cluster analysis was conducted for each year from 1986 to 1997. Non-overlapping clusters were identified using the Bernoulli model with all TB breakdown herds as cases compared against 30% of control herds and a maximum cluster size set to 5% of all points. The figures below show the evolution of TB clusters during this period. Annual clustering of TB breakdowns is represented by the log likelihood ratios, using the same scale for each year.

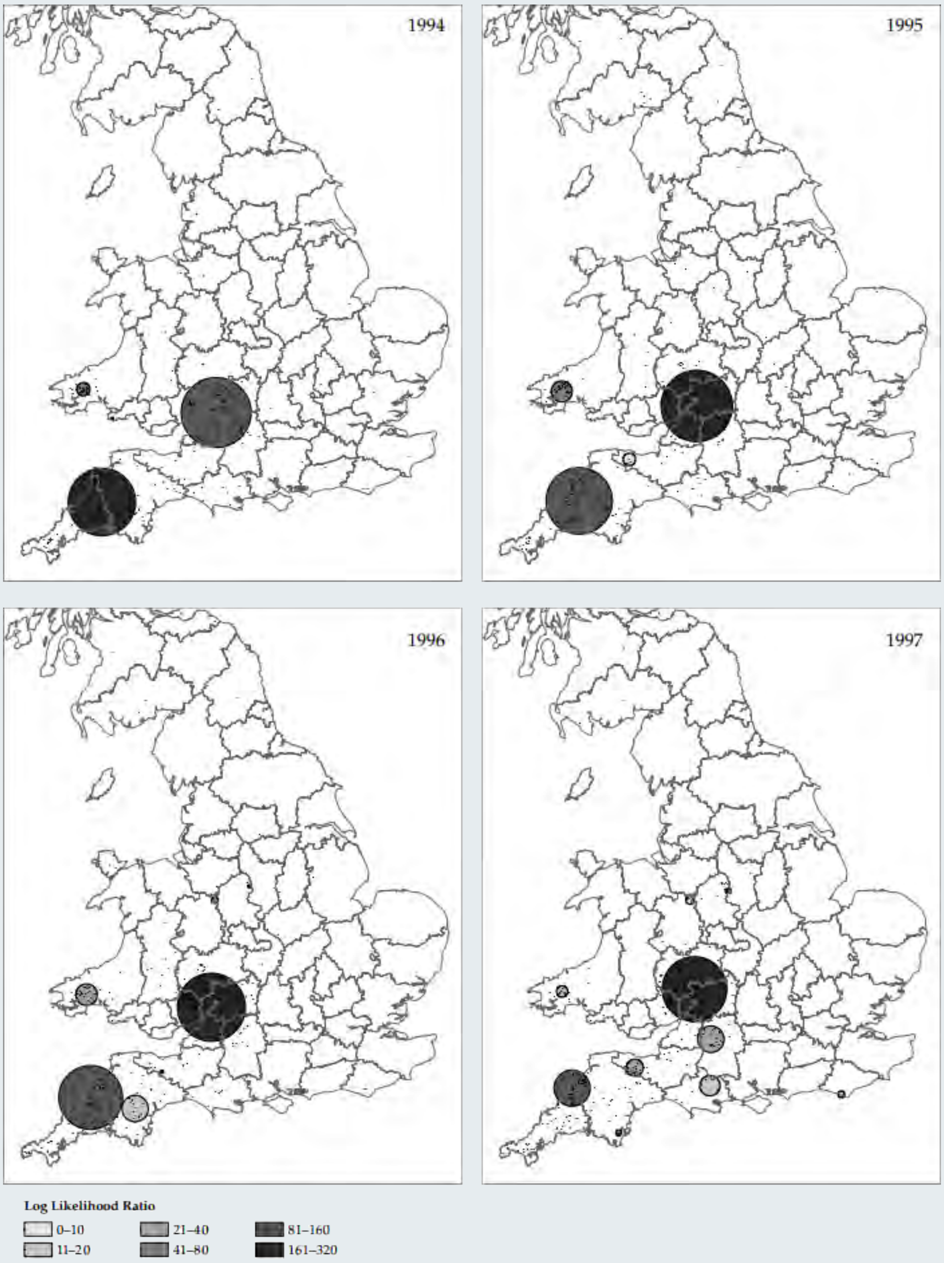
For most of the period two large and highly significant clusters were fairly constant. These were Cluster 1 (as labelled above) on the Gloucestershire, Hereford, and Worcester boundaries, and Cluster 2 on the Devon/Cornwall boundary. Some clusters tended to appear, disappear, and then re-appear in much the same place. Perhaps these are the main centres of endemicity that act as sources of infection via, for example, cattle movements into areas where prevailing environmental and epidemiological conditions allow new disease clusters to arise. Clearly there are clusters of bovine TB in the southwest of Britain, some persistent, some fleeting, and some recurring.



Cluster analysis of cattle TB data (1986-1989). Clusters are shaded by log likelihood ratio using the same scale for each year with darker clusters being the most likely. The distribution of breakdown herds in each year is superimposed on the cluster map.



Cluster analysis of cattle TB data (1990-1993). Clusters are shaded by log likelihood ratio using the same scale for each year with darker clusters being the most likely. The distribution of breakdown herds in each year is superimposed on the cluster map.



Cluster analysis of cattle TB data (1994-1997). Clusters are shaded by log likelihood ratio using the same scale for each year with darker clusters being the most likely. The distribution of breakdown herds in each year is superimposed on the cluster map.

5.3.4 Detecting clusters around a source (focused tests)

With focused tests, the location of a cluster centre is specified *a priori*, and the likelihood of that location truly being a cluster centre is then determined. Specifying this location can be a sensitive issue, which can be influenced by non-technical biases such as political factors and media influences.

Stone's test

Stone (1988) developed a class of tests for trend, the maximum likelihood ratio (MLR) and Poisson maximum (Pmax) tests, both of which use the first isotonic regression estimator, working on the assumption that there will be a monotonic decay of risk with increasing distance from any point source of a disease.

A number of adaptations of the original method have been developed. For example, Morton-Jones et al. (1999) propose an extension that allows for covariate adjustments via a log-linear model, which they illustrate using data on the incidence of stomach cancers near municipal incinerators. Diggle et al. (1999) adapted Stone's isotonic regression method to incorporate case-control data in addition to covariate information.

The Lawson-Waller score test

Lawson (1993) and Waller et al. (1992) developed the Lawson-Waller score test, sometimes referred to as the uniformly most powerful (UMP) test. The score test detects a decreasing trend in disease frequency associated with declining exposure to a point-focus. The test statistic T_{LW} is given by:

$$T_{LW} = \sum_{i=1}^{I} g_i(O_i - E_i) \tag{5.16}$$

$$Var(T_{LW}) = \sum_{i=1}^{I} g_i^2 E_i - O_i \left(\sum_{i=1}^{I} \frac{p_i g_i}{p_+} \right)^2 \tag{5.17}$$

where g_i denotes the exposure to the focus for an individual residing in area i, O_i is the observed number of disease cases in area i, E_i is the expected number of disease cases in area i, p_i is the size of the population at risk in area i, and p_+ is the total population size. Monte Carlo simulation is used to evaluate the significance of the test statistic T_{LW}.

Bithell's linear risk score tests

Bithell et al. (1994) and Bithell (1995) developed a set of tests known as 'linear risk score tests', where disease incidence is weighted by some distance function from a point source (e.g. $1/d_i$, $1/d_i^2$, or $1/rank_i$). Using the reciprocal of distance is appropriate for detecting an environmental hazard that declines with distance from a source and is relatively insensitive to the precise location of the assumed source. Using the reciprocal of rank is more appropriate when the relative proximity of residence is important, rather than actual distance, but it is more sensitive to the precise location of the putative source.

Diggle's test

Diggle (1990) developed a test that uses nonparametric kernel smoothing to describe natural variation in a disease (assuming a Poisson point process), and then a maximum likelihood test to evaluate the possibility of raised incidence around a pre-specified point source. Diggle and Rowlingson (1994) developed a conditional approach, which converts the original point process model into a non-linear binary regression model for the spatial variation in risk.

5.3.5 Space-time cluster detection

Kulldorff's space-time scan statistic

Kulldorff's spatial scan statistic can be 'focused' and used to search for clusters around a point source by making the point source the only coordinate pair in the special grid file.

The spatial scan statistic has also been adapted to look for clusters in space and time by extending the idea of a two-dimensional circular window to that of a cylinder passing through time. Prospective use of the space-time scan statistic, with repeated time periodic analysis, can be applied as part of a surveillance system to track active clusters of disease, for detecting the geographic location of emerging clusters and evaluating their significance. A later development of this prospective space-time scan statistic led to the development of the 'space-time permutation scan statistic', which does not require data on the background population at risk, but estimates expected disease occurrence based only on case data. This technique has been applied in a large number of scientific studies and publications.

5.3.6 Conclusion

The concepts and analytical methods discussed in this chapter extend beyond the mere visualization of spatial patterns. The use of statistical methods to assess whether observed patterns differ significantly from spatial randomness moves into the next stage of spatial analysis: that of exploration and from there, quantitative analysis.

A variety of methods have been presented by which spatial and spatio-temporal disease clusters can be identified statistically. New methods for detecting local clustering are continuously being developed. For example, methods using Bayesian (probabilistic) approaches or based on other statistical modelling techniques are increasingly applied.

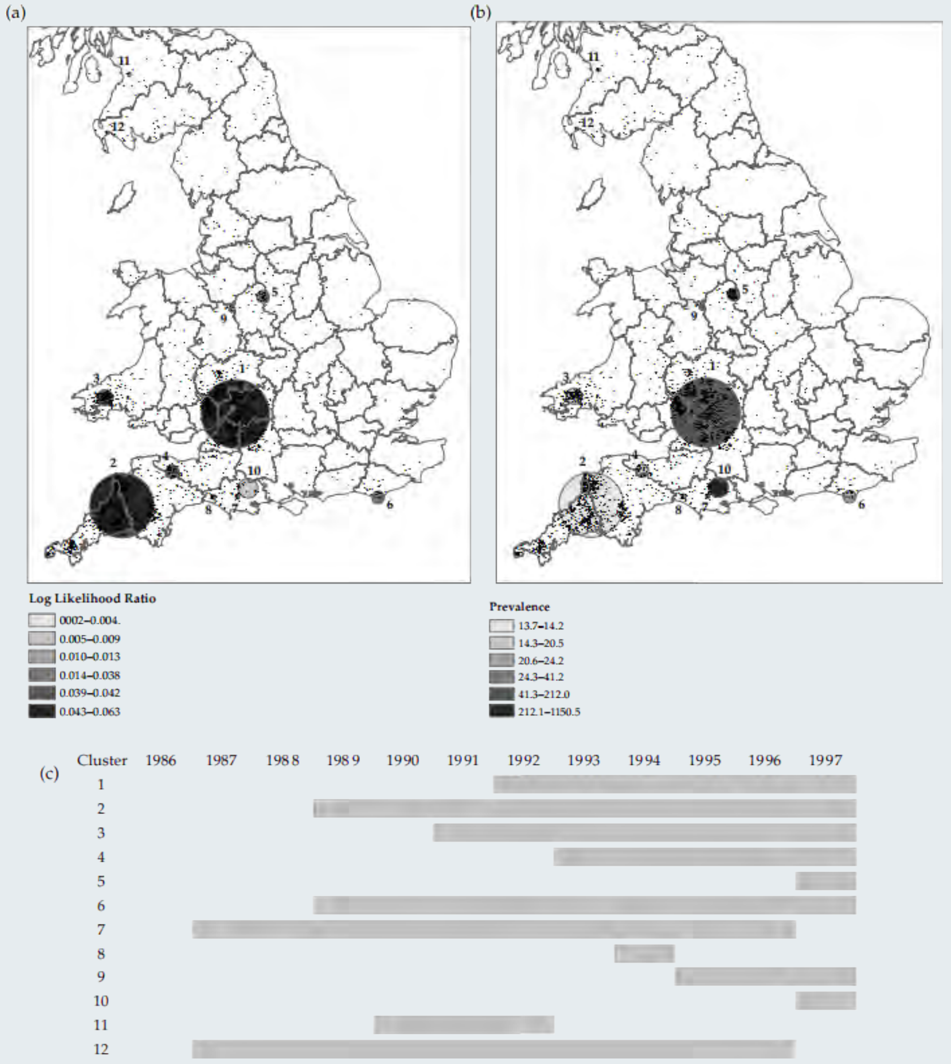
A frequently quoted limitation common to many cluster detection tests is the *a priori* choice of cluster size, as testing for a variety of cluster sizes results in problems of multiple inference. Analysis of the British cattle TB data, presented in this chapter as the motivating example, suggests this to be a central problem which can have profound effects on the results. There do not seem to be clear guidelines on how to deal with it. The attendant danger is that by exploring a range of maximum cluster sizes, an upper cluster size threshold can be chosen that presents a pattern of clustering best suited to support a particular argument, rather than that which best reflects reality. This may cast doubt on the validity of the numerous studies that have been reported using scan circles, and in particular those based on the spatial scan statistic.

Example

The space-time permutation of Kulldorff's spatial scan statistic was applied to the British cattle TB data from 1986 to 1997. The Bernoulli model was used and all TB-breakdown herds were included as cases compared to 30% of control herds. The maximum cluster size was set to 5% of all points. 12 space-time clusters of TB breakdowns were located.

Cluster	Area (km^2)	Start	End	LLR	P(999)	RR	Prev.
1	5,830	1992	1997	1,150.40	0.001	9.5	0.042
2	5,148	1989	1997	743.6	0.001	5.9	0.002
3	316	1991	1997	212.1	0.001	13.1	0.006
4	209	1993	1997	78.2	0.001	15.1	0.009
5	194	1997	1997	41.3	0.001	20.5	0.063
6	194	1989	1997	41.1	0.001	33.8	0.012
7	21	1987	1996	24.3	0.001	43.1	0.013
8	27	1987	1994	21.8	0.001	11.2	0.004
9	47	1995	1997	20.6	0.003	39.9	0.041
10	479	1997	1997	18.1	0.017	14.5	0.045
11	15	1990	1992	14.3	0.382	37.3	0.038
12	13	1987	1996	13.7	0.721	97	0.03

The results of the analysis are dominated by two large and highly significant clusters; the SW cluster had a much lower prevalence, illustrating the need to consider all characteristics of clusters when making comparisons.



Space-time cluster analysis showing a) log likelihood ratios, b) prevalence (directly proportional to relative risk), and c) duration of space-time clusters.

5.4 Cluster detection using SaTScan

SaTScan is a free software programme that analyses spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. It is designed for any of the following interrelated purposes:

1. Perform geographical surveillance of disease, to detect spatial or space-time disease clusters, and to see if they are statistically significant.
2. Test whether a disease is randomly distributed over space, over time or over space and time.
3. Evaluate the statistical significance of disease cluster alarms.
4. Perform prospective real-time or time-periodic disease surveillance for the early detection of disease outbreaks.

5.4.1 Statistical methodology, data types and methods

Scan statistics are used to detect and evaluate clusters of cases in either a purely temporal, purely spatial or space-time setting. This is done by gradually scanning a window across time and/or space, noting the number of observed and expected observations inside the window at each location. In the SaTScan software, the scanning window is an interval (in time), a circle or an ellipse (in space) or a cylinder with a circular or elliptic base (in space-time). Multiple different window sizes are used. The window with the maximum likelihood is the most likely cluster, that is, the cluster least likely to be due to chance. A p-value is assigned to this cluster.

Discrete and continuous probability models

Scan statistics use a range of different probability models, depending on the nature of the data and the purpose of the analysis. SaTScan can be used for discrete as well as continuous scan statistics.

- For **discrete scan statistics** the geographical locations where data are observed are non-random and fixed by the user. These locations may be
  - the *actual locations* of the observations, such as houses, schools or ant nests, or
  - a *central location representing a larger area*, such as the geographical or population weighted centroids of postal areas, counties or provinces.
- For **continuous scan statistics**, the locations of the observations are random and can occur anywhere within a predefined study area defined by the user. This could be an administrative area, or a bounded geographic area such as a rectangle.

For all discrete probability models, the scan statistic adjusts for the uneven geographical density of a background population. The following probability models are available:

- a *Poisson model*, where the number of events in a geographical location is Poisson-distributed (i.e. counts of disease events), according to a known underlying population at risk;
- a *Bernoulli model*, where individuals in the population are either cases or non-cases (i.e. 0/1 event data);
- a *space-time permutation model*, using only case data;
- a *multinomial model* for categorical data;
- an *ordinal model*, for ordered categorical data;

- an *exponential model* for survival time data with or without censored variables;
- a *normal model* for other types of continuous data; or
- a *spatial variation in temporal trends model*, looking for geographical areas with unusually high or low temporal trends.

A common feature of all these discrete scan statistics is that the geographical locations where data can be observed are non-random and fixed by the user. The data may be either aggregated at the census tract, zip code, county or other geographical level, or there may be unique coordinates for each observation. SaTScan adjusts for the underlying spatial inhomogeneity of a background population. It can also adjust for any number of categorical covariates provided by the user, as well as for temporal trends, known space-time clusters and missing data. It is possible to scan multiple data sets simultaneously to look for clusters that occur in one or more of them.

For continuous scan statistics, SaTScan uses a *continuous Poisson model*.

All analyses are conditioned on the total number of cases observed. For all discrete spatial and space-time analyses, the user must provide various identifiers in the data, the first two of which are required as a minimum, and the others dependent on the analysis to be performed:

- the *spatial coordinates* of a set of locations (coordinates file);
- for each location, information about the *number of cases* at that location (case file);
- for temporal and space-time analyses, the number of cases must be *stratified by time*, e.g. the time of diagnosis;
- depending on the type of analysis, *covariate data* about cases such as species, breed, age, gender, weight, length of survival etc. may also be provided;
- for the Bernoulli model, it is also necessary to specify the number of *non-cases* at each location;
- for the discrete Poisson model, the user must specify a *population size* for each location (population file). The population may vary over time.

Spatial, temporal and space-time scan statistics

Spatial scan statistic

The standard purely spatial scan statistic imposes a circular window on the map. The window is in turn centred on each of several possible grid points positioned throughout the study region. For each grid point, the radius of the window varies continuously in size from zero to some upper limit specified by the user. In this way, the circular window is flexible both in location and size. In total, the method creates an infinite number of distinct geographical circles with different sets of neighbouring data locations within them. Each circle is a possible candidate cluster.

The user defines the set of grid points used through a grid file. If no grid file is specified, the grid points are set to be identical to the coordinates of the location IDs defined in the coordinates file. The latter option ensures that each data location is a potential cluster in itself, and it is the recommended option for most types of analyses.

As an alternative to the circle, it is also possible to use an elliptic window shape, in which case a set of ellipses with different shapes and angles are used as the scanning window together with the circle. This provides slightly higher power for true clusters that are long and narrow in shape, and slightly lower power for circular and other very compact clusters.

It is also possible to define your own non-Euclidian distance metric using a special neighbours file.

Space-time scan statistic

This is defined by a cylindrical window with a circular (or elliptic) geographic base and with height corresponding to time. The base is defined exactly as for the purely spatial scan statistic, while the height reflects the time period of potential clusters. The cylindrical window is then moved in space and time, so that for each possible geographical location and size, it also visits each possible time period. In effect, we obtain an infinite number of overlapping cylinders of different size and shape, jointly covering the entire study region, where each cylinder reflects a possible cluster.

This may be used for either a single retrospective analysis, using historic data, or for time-periodic prospective surveillance, where the analysis is repeated for example every day, week, month or year.

Purely temporal scan statistic

The temporal scan statistic uses a window that moves in one dimension, time, defined in the same way as the height of the cylinder used by the space-time scan statistic. This means that it is flexible in both start and end date.

Seasonal scan statistic

The seasonal scan statistic is a purely temporal scan statistic where all the data is on a connecting loop, such as the year, where December 31 is followed by January 1. The key feature that distinguishes the seasonal scan statistic from the purely temporal scan statistic is that it ignores which year the observation was made and only cares about the day and month.

Spatial variation in temporal trends scan statistic

When the scan statistic is used to evaluate the spatial variation in temporal trends, the scanning window is purely spatial in nature. The temporal trend is then calculated inside as well as outside the scanning window, for each location and size of that window. The null hypothesis is that the trends are the same, while the alternative is that they are different. Based on these hypotheses, a likelihood is calculated, which is higher the more unlikely it is that the difference in trends is due to chance. The most likely cluster is the cluster for which the temporal trend inside the window is least likely to be the same as the temporal trend outside the cluster. This could be because of various reasons. For example, if the temporal trend inside the cluster is higher, it could be because all areas has the same incidence rate of a disease at the beginning of the time period, but the cluster area has a higher rate at the end of the time period. It could also be because the cluster area has a lower incidence rate at the beginning of the time period, after which it ‘catches up’ with the rest so that the rate is about the same at the end of the time period. Hence, a statistically significant cluster in the spatial variation in temporal trend analysis does not necessarily mean that the overall rate of disease is higher or lower in the cluster.

The spatial variation in temporal trends scan statistic can only be run with the discrete Poisson probability model. For it to work, it is important that the total study period length is evenly divisible by the length of the time interval aggregation, so that all time intervals have the same number of years, if it is specified in years, the same number of months if it is specified in months or the same number of days if it is specified in days.

5.4.2 A summary of the model types implemented in SaTScan

Bernoulli model

With the Bernoulli model there are cases and non-cases represented by a 0/1 variable. These variables may represent animals with or without a disease, or animals showing different degrees of clinical disease severity. They may reflect cases and controls from a larger population, or they may together constitute the population as a whole. Whatever the situation may be, these variables ar denoted as cases and controls, and their total will be denoted as the population. Bernoulli data can be analysed with the purely temporal, the purely spatial or the space-time scan statistics.

Example

For the Bernoulli model, cases may be newborns with a certain birth defect while controls are all newborns without that birth defect.

The Bernoulli model requires information about the location of a set of cases and controls, provided to SaTScan using the case, control and coordinates files. Separate locations may be specified for each case and each control, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc., with multiple cases and controls at each data location. To do a temporal or space-time analysis, it is necessary to have a time for each case and each control as well.

Discrete Poisson model

With the discrete Poisson model the number of cases in each location is Poisson-distributed. Under the null hypothesis, and when there are no covariates, the expected number of cases in each area is proportional to its population size, or to the person-years in that area. Poisson data can be analysed with the purely temporal, the purely spatial, the space-time and the spatial variation in temporal trends scan statistics.

Example

For the discrete Poisson model, cases may be stroke occurrences while the population is the combined number of person-years lived, calculated as 1 for someone living in the area for the whole time period and ½for someone dying or moving away in the middle of the time period.

The discrete Poisson model requires case and population counts for a set of data locations such as counties, parishes, census tracts or zip code areas, as well as the geographical coordinates for each of those locations. These need to be provided to SaTScan using the case, population and coordinates files.

The population data need not be specified continuously over time, but only at one or more specific ‘census times’. For times in between, SaTScan does a linear interpolation based on the population at the census times immediately preceding and immediately following. For times before the first census time, the population size is set equal to the population size at that first census time, and for times after the last census time, the population is set equal to the population size at that last census time. To get the population size for a given location and time period, the population size, as defined above, is integrated over the time period in question.

Space-time permutation model

The space-time permutation model requires only case data, with information about the spatial location and time for each case, with no information needed about controls or a background population at risk. The number of observed cases in a cluster is compared to what would have been expected if the spatial and temporal locations of all cases were independent of each other so that there is no space-time interaction. That is, there is a cluster in a geographical area if, during a specific time period, that area has a higher proportion of its cases in that time period compared to the remaining geographical areas. This means that if, during a specific week, all geographical areas have twice the number of cases than normal, none of these areas constitute a cluster. On the other hand, if during that week, one geographical area has twice the number of cases compared to normal while other areas have a normal amount of cases, then there will be a cluster in that first area. The space-time permutation model automatically adjusts for both purely spatial and purely temporal clusters. Hence there are no purely temporal or purely spatial versions of this model.

Caution

Space-time permutation clusters may be due either to an increased risk of disease, or to different geographical population distribution at different times, where for example the population in some areas grows faster than in others. This is typically not a problem if the total study period is less than a year. However, the user is advised to be very careful when using this method for data spanning several years. If the background population increases or decreases faster in some areas than in others, there is risk for population shift bias, which may produce biased p-values when the study period is longer than a few years.

For example, if a new large neighbourhood is developed, there will be an increase in cases there simply because the population increases, and using only case data, the space-time permutation model cannot distinguish an increase due to a local population increase versus an increase in the disease risk. As with all space-time interaction methods, this is mainly a concern when the study period is longer than a few years.

Multinomial model

With the multinomial model, each observation is a case, and each case belongs to one of several categories. The multinomial scan statistic evaluates whether there are any clusters where the distribution of cases is different from the rest of the study region. For example, there may be a higher proportion of cases of types 1 and 2 and a lower proportion of cases of type 3 while the proportion of cases of type 4 is about the same as outside the cluster. If there are only two categories, the ordinal model is identical to the Bernoulli model, where one category represents the cases and the other category represents the controls. The cases in the multinomial model may be a sample from a larger population or they may constitute a complete set of observations. Multinomial data can be analysed with the purely temporal, the purely spatial or the space-time scan statistics.

Example

For the multinomial model, the data may consist of everyone diagnosed with meningitis, with five different categories representing five different clonal complexes of the disease. The multinomial scan statistic will simultaneously look for high or low clusters of any of the clonal complexes, or a group of them, adjusting for the overall geographical distribution of the disease. The multiple comparisons inherent in the many categories used are accounted for when calculating the p-values.

The multinomial model requires information about the location of each case in each category. A unique location may be specified for each case, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc., with multiple cases in the same location. To do a temporal or space-time analysis, it is necessary to have a time for each case as well.

With the multinomial model it is not necessary to specify a search for high or low clusters, since there is no hierarchy among the categories, but in the output it is shown what categories are more prominent inside the cluster. The order or indexing of the categories does not affect the analysis in terms of the clusters found, but it may influence the randomization used to calculate the p-values.

Ordinal model

With the ordinal model, each observation is a case, and each case belongs to one of several ordinal categories. If there are only two categories, the ordinal model is identical to the Bernoulli model, where one category represents the cases and the other category represent the controls in the Bernoulli model. The cases in the ordinal model may be a sample from a larger population or they may constitute a complete set of observations. Ordinal data can be analysed with the purely temporal, the purely spatial or the space-time scan statistics.

Example

For the ordinal model, the data may consist of everyone diagnosed with breast cancer during a ten-year period, with three different categories representing early, medium and late stage cancer at the time of diagnosis.

The ordinal model requires information about the location of each case in each category. Separate locations may be specified for each case, or the data may be aggregated for states, provinces, counties, parishes, census tracts, postal code areas, school districts, households, etc., with multiple cases in the same or different categories at each data location. To do a temporal or space-time analysis, it is necessary to have a time for each case as well.

With the ordinal model it is possible to search for high clusters, with an excess of cases in the high-valued categories, for low clusters with an excess of cases in the low-valued categories, or simultaneously for both types of clusters. Reversing the order of the categories has the same effect as changing the analysis from high to low and vice versa.

Exponential model

The exponential model is designed for survival time data, although it could be used for other continuous type data as well. Each observation is a case, and each case has one continuous variable attribute as well as a 0/1 censoring designation. For survival data, the continuous variable is the time between diagnosis and death or depending on the application, between two other types of events. If some of the data is censored, due to loss of follow-up, the continuous variable is then instead the time between diagnosis and time of censoring. The 0/1 censoring variable is used to distinguish between censored and non-censored observations.

Example

For the exponential model, the data may consist of everyone diagnosed with prostate cancer during a ten-year period, with information about either the length of time from diagnosis until death or from diagnosis until a time of censoring after which survival is unknown.

When using the temporal or space-time exponential model for survival times, two very different time variables are involved. The first is the time the case was diagnosed, and that is the time that the temporal and space-time scanning window is scanning over. The second is the survival time, that is, time between diagnosis and death or for censored data the time between diagnosis and censoring. This is an attribute of each case, and there is no scanning done over this variable. Rather, we are interested in whether the scanning window includes exceptionally many cases with a small or large value of this attribute.

While the exponential model uses a likelihood function based on the exponential distribution, the true survival time distribution must not be exponential and the statistical inference (p-value) is valid for other survival time distributions as well. The reason for this is that the randomization is not done by generating observations from the exponential distribution, but rather, by permuting the space-time locations and the survival time/censoring attributes of the observations.

Normal model

The normal model is designed for continuous data. For each individual or for each observation, called a case, there is a single continuous attribute that may be either negative or positive. The model can also be used for ordinal data when there are many categories. That is, different cases are allowed to have the same attribute value.

Example

For the normal model, the data may consist of the birth weight and residential census tract for all newborns, with an interest in finding clusters with lower birth weight. One individual is then a ‘case’. Alternatively, the data may consist of the average birth weight in each census tract. It is then the census tract that is the ‘case’, and it is important to use the weighted normal model, since each average will have a different variance due to a different number of births in each tract.

While the normal model uses a likelihood function based on the normal distribution, the true distribution of the continuous attribute must not be normal. The statistical inference (p-value) is valid for any continuous distribution. The reason for this is that the randomization is not done by generating simulated data from the normal distribution, but rather, by permuting the space-time locations and the continuous attribute (e.g. birth weight) of the observations. While still being formally valid, the results can be greatly influenced by extreme outliers, so it may be wise to truncate such observations before doing the analysis.

In the standard normal model, it is assumed that each observation is measured with the same variance. That may not always be the case. For example, if an observation is based on a larger sample in one location and a smaller sample in another, then the variance of the uncertainty in the estimates will be larger for the smaller sample. If the reliability of the estimates differs, one should instead use the weighted normal scan statistic that takes these unequal variances into account. The weighted version is obtained in SaTScan by simply specifying a weight for each observation as an extra column in the input file. This weight may for example be proportional to the sample size used for each estimate or it may be the inverse of the variance of the observation.

If all values are multiplied with or added to the same constant, the statistical inference will not change, meaning that the same clusters with the same log likelihoods and p-values will be found. Only the estimated means and variances will differ. If the weight is the same for all observations, then the weighted normal scan statistic will produce the same results as the standard normal version. If all the weights are multiplied by the same constant, the results will not change.

Continuous Poisson model

All the models described above are based on data observed at discrete locations that are considered to be non-random, as defined by a regular or irregular lattice of location points. That is, the locations of the observations are considered to be fixed, and we evaluate the spatial randomness of the observation conditioning on the lattice. Hence, those are all versions of what are called discrete scan statistics. In a continuous scan statistics, observations may be located anywhere within a study area, such as a square or rectangle. The stochastic aspect of the data consists of these random spatial locations, and we are interested to see if there are any clusters that are unlikely to occur if the observations where independently and randomly distributed across the study area. Under the null hypothesis, the observations follow a homogeneous spatial Poisson process with constant intensity throughout the study area, with no observations falling outside the study area.

Example

The data may consist of the location of bird nests in a square kilometer area of a forest. The interest may be to see whether the bird nests are randomly distributed spatially, or in other words, whether there are clusters of bird nests or whether they are located independently of each other.

In SaTScan, the study area can be any collection of convex polygons, which are convex regions bounded by any number straight lines. Triangles, squares, rectangles, rhombuses, pentagons and hexagons are all examples of convex polygons. In the simplest case, there is only one polygon, but the study area can also be the union of multiple convex polygons. If the study area is not convex, divide it into multiple convex polygons and define each one separately. The study area does not need to be contiguous, and may for example consist of five different islands.

The analysis is conditioned on the total number of observations in the data set. Hence, the scan statistic simply evaluates the spatial distribution of the observation, but not the number of observations.

The likelihood function used as the test statistic is the same as for the Poisson model for the discrete scan statistic, where the expected number of cases is equal to the total number of observed observations, times the size of the scanning window, divided by the size of the total study area. When the scanning window extends outside the study area, the expected count is still based on the full size of the circle, ignoring the fact that some parts of the circle have zero expected counts. This is to avoid strange non-circular clusters at the border of the study area. Since the analysis is based on Monte Carlo randomizations, the p-values are automatically adjusted for these boundary effects. The reported expected counts are based on the full circle though, so the Obs/Exp ratios provided should be viewed as a lower bound on the true value whenever the circle extends outside the spatial study region.

The continuous Poisson model can only be used for purely spatial data. It uses a circular scanning window of continuously varying radius up to a maximum specified by the user. Only circles centred on one of the observations are considered, as specified in the coordinates file. If the optional grid file is provided, the circles are instead centred on the coordinates specified in that file.

5.4.3 Probability model choice and comparison

- All discrete probability models can be used for either individual locations or aggregated data.
- The discrete Poisson model is usually the fastest to run. The ordinal model is typically the slowest.
- With the discrete Poisson and space-time permutations models, an unlimited number of covariates can be adjusted for, by including them in the case and population files. With the normal model, it is also possible to adjust for covariates by including them in the case file, but only for purely spatial analyses. With the Bernoulli, ordinal, exponential and normal models, covariates can be adjusted for by using multiple data sets, which limits the number of covariate categories that can be defined, or through a pre-processing regression analysis done before running SaTScan.
- With the discrete Poisson model, population data is only needed at selected time points and the numbers are interpolated in between. A population time must be specified even for purely spatial analyses. Regardless of model used, the time of a case or control need only be specified for purely temporal and space-time analyses.
- The space-time permutation model automatically adjusts for purely spatial and purely temporal clusters. For the discrete Poisson model, purely temporal and purely spatial clusters can be adjusted for in a number of different ways. For the Bernoulli, ordinal, exponential and normal models, spatial and temporal adjustments can be done using multiple data sets, but it is limited by the number of different data sets allowed, and it is also much more computer intensive.
- Purely temporal and space-time analyses cannot be performed using the homogeneous Poisson model.
- Spatial variation in temporal trend analyses can only be performed using the discrete Poisson model.

Some additional considerations regarding model choice are as follows:

- **Few cases compared to controls.** In a purely spatial analysis where there are few cases compared to controls, say less than 10 percent, the discrete Poisson model is a very good approximation to the Bernoulli model. The former can then be used also for 0/1 Bernoulli type data, and may be preferable as it has more options for various types of adjustments, including the ability to adjust for covariates specified in the case and population files. As an approximation for Bernoulli type data, the discrete Poisson model produces slightly conservative p-values.

- **Bernoulli versus ordinal model.** The Bernoulli model is mathematically a special case of the ordinal model, when there are only two categories. The Bernoulli model runs faster, making it the preferred model to use when there are only two categories.
- **Normal versus exponential model.** Both the normal and exponential models are meant for continuous data. The exponential model is primarily designed for survival time data but can be used for any data where all observations are positive. It is especially suitable for data with a heavy right tail. The normal model can be used for continuous data that takes both positive and negative values. While still formally valid, results from the normal model are sensitive to extreme outliers.
- **Normal versus ordinal model.** The normal model can be used for categorical data when there are very many categories. As such, it is sometimes a computationally faster alternative to the ordinal model. There is an important difference though. With the ordinal model, only the order of the observed values matters. For example, the results are the same for ordered values '1 – 2 – 3 – 4' and '1 – 10 – 100 – 1000'. With the normal model, the results will be different, as they depend on the relative distance between the values used to define the categories.
- **Discrete versus homogeneous Poisson model.** Instead of using the homogeneous Poisson model, the data can be approximated by the discrete Poisson model by dividing the study area into many small pieces. For each piece, a single coordinates point is specified, the size of the piece is used to define the population at that location and the number of observations within that small piece of area is the number of cases in that location. As the number of pieces increases towards infinity, and hence, as their size decreases towards zero, the discrete Poisson model will be asymptotically equivalent to the homogeneous Poisson model.
- **Temporal data.** For temporal and space-time data, there is an additional difference among the probability models, in the way that the temporal data is handled. With the Poisson model, population data may be specified at one or several time points, such as census years. The population is then assumed to exist between such time points as well, estimated through linear interpolation between census years. With the Bernoulli, space-time permutation, ordinal, exponential and normal models, a time needs to be specified for each case and for the Bernoulli model, for each control as well.

5.4.4 Likelihood Ratio Test

For each location and size of the scanning window, the alternative hypothesis is that there is an elevated risk within the window as compared to outside. Under the Poisson assumption, the likelihood function for a specific window is proportional to:

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c} I() \tag{5.18}$$

where C is the total number of cases, c is the observed number of cases within the window and $E[c]$ is the covariate adjusted expected number of cases within the window under the null-hypothesis. Note that since the analysis is conditioned on the total number of cases observed, $C - E[c]$ is the expected number of cases outside the window. $I()$ is an indicator function. When

SaTScan is set to scan only for clusters with high rates, $I()$ is equal to 1 when the window has more cases than expected under the null-hypothesis, and 0 otherwise. The opposite is true when SaTScan is set to scan only for clusters with low rates. When the program scans for clusters with either high or low rates, then $I() = 1$ for all windows.

The space-time permutation model uses the same function as the Poisson model. Due to the conditioning on the marginals, the observed number of cases is only approximately Poisson distributed. Hence, it is no longer a formal likelihood ratio test, but it serves the same purpose as the test statistic.

For the Bernoulli model the likelihood function is:

$$\left(\frac{c}{n}\right)^{c}\left(\frac{n-c}{n}\right)^{n-c}\left(\frac{C-c}{N-n}\right)^{C-c}\left(\frac{(N-n)(C-c)}{N-n}\right)^{(N-n)(C-c)} I() \tag{5.19}$$

where c and C are defined as above, n is the total number of cases and controls within the window, while N is the combined total number of cases and controls in the data set.

The likelihood function for the multinomial, ordinal, exponential, and normal models are more complex, due to the more complex nature of the data. The likelihood function for the spatial variation in temporal trends scan statistic is also more complex, as it involves the maximum likelihood estimation of several different trend functions.

The likelihood function is maximized over all window locations and sizes, and the one with the maximum likelihood constitutes the most likely cluster. This is the cluster that is least likely to have occurred by chance. The likelihood ratio for this window constitutes the maximum likelihood ratio test statistic. Its distribution under the null-hypothesis is obtained by repeating the same analytic exercise on a large number of random replications of the data set generated under the null hypothesis.

The p-value is obtained through Monte Carlo hypothesis testing, by comparing the rank of the maximum likelihood from the real data set with the maximum likelihoods from the random data sets. If this rank is R, then $p = R/(1 + \#\text{simulation})$. In order for p to be a 'nice looking' number, the number of simulations is restricted to 999 or some other number ending in 999 such as 1999, 9999 or 99999. That way it is always clear whether to reject or not reject the null hypothesis for typical cut-off values such as 0.05, 0.01 and 0.001.

The SaTScan program scans for areas with high rates (clusters), for areas with low rates, or simultaneously for areas with either high or low rates. The most common analysis is to scan for areas with high rates, that is, for clusters.

5.4.5 Secondary clusters

For purely spatial and space-time analyses, SaTScan also identifies secondary clusters in the data set in addition to the most likely cluster, and orders them according to their likelihood ratio test statistic. There will almost always be a secondary cluster that is almost identical with the most likely cluster and that have almost as high likelihood value, since expanding or reducing the cluster size only marginally will not change the likelihood very much. The user can decide to what extent overlapping clusters are reported in the results files. The default is that geographically overlapping clusters are not reported.

There may also be secondary clusters that do not overlap spatially with the most likely cluster, and they may be of great interest. These are always reported. When there are multiple clusters in the data set, the secondary clusters are evaluated as if there were no other clusters in the data set. That is, they are statistically significant if and only if they are able to cause a rejection of the null hypothesis on their own strength, whether or not the other clusters are true clusters or not. That is often the desired type of inference. Sometime though, it is also of interest to evaluate secondary clusters after adjusting for other clusters in the data.

Note that the circle of a secondary cluster may overlap with the circle of a previously detected more likely cluster, and it may even completely encircle it so that the latter is a subset of the former. This does not mean that the more likely cluster is detected twice. Rather, the more likely cluster is treated as a 'lake' with no population and no cases, and the new secondary cluster consist of the areas around that 'lake'. This may for example happen if a city has a very high elevated risk, while the surrounding suburbs have a modest elevated risk. The same phenomena may occur when doing purely temporal or space-time analyses.

For purely temporal analyses, only the most likely cluster is reported.

5.4.6 Statistical adjustments

Covariate adjustments

A covariate should be adjusted for when all three of the following are true:

1. The covariate is related to the disease in question.
2. The covariate is not randomly distributed geographically.
3. You want to find clusters that cannot be explained by that covariate.

Example

Here are three examples:

- If you are studying cancer mortality, you should adjust for age since (i) older people are more likely to die from cancer (ii) some areas may have a higher density of older people, and (iii) you are presumably interested in finding areas where the risk of cancer is high as opposed to areas with an older population.
- If you are interested in the geographical distribution of birth defects, you do not need to adjust for gender. While birth defects are not equally likely in boys and girls, the geographical distribution of the two genders is geographically random at time of birth.
- If you are studying the geography of lung cancer incidence, you should adjust for smoking if you are interested in finding clusters due to non-smoking related risk factors, but you should not adjust for smoking if you are interested in finding clusters reflecting areas with especially urgent needs to launch an anti-smoking campaign.

When the disease rate varies, for example, with age, and the age distribution varies in different areas, then there is geographical clustering of the disease simply due to the age covariate. When adjusting for categorical covariates, the SaTScan program will search for clusters above and beyond that which is expected due to these covariates. When more than one covariate is specified, each one is adjusted for as well as all the interaction terms between them.

Covariate adjustment using the input files. With the Poisson and space-time permutation models, it is possible to adjust for multiple categorical covariates by specifying the covariates in the input files. To do so, simply enter the covariates as extra columns in the case file (both models) and the population file (Poisson model). There is no need to enter any information on any of the window tabs.

For the Poisson model, the expected number of cases in each area under the null-hypothesis is calculated using indirect standardization. Without covariate adjustment the expected number of cases in a location is (spatial analysis):

$$E[c] = p * C/P \tag{5.20}$$

where c is the observed number of cases and p the population in the location of interest, while C and P are the total number of cases and population respectively. Let c_i, p_i, C_i and P_i be defined in the same way, but for covariate category i. The indirectly standardized covariate adjusted expected number of cases (spatial analysis) is:

$$E[c] = \sum_i E[c_i] = \sum_i p_i * C_i/P_i \tag{5.21}$$

The same principle is used when calculating the covariate adjusted number of cases for the space-time scan statistic, although the formula is more complex due to the added time dimension.

Since the space-time permutation model automatically adjusts for purely spatial and purely temporal variation, there is no need to adjust for covariates in order to account for different spatial or temporal densities of these covariates. Rather, covariate adjustment is used if there is space-time interaction due to this covariate rather than to the underlying disease process.

Example

If children get sick mostly in the summer and adults mostly in the winter, then there will be age generated space-time interaction clusters in areas with many children in the summer and vice versa. When including child/adult as a covariate, these clusters are adjusted away.

Covariate adjustment using statistical regression software. SaTScan cannot in itself do an adjustment for continuous covariates. Such adjustments can still be done for the Poisson model but it is a little more complex. The first step is to calculate the covariate adjusted expected number of cases for each location ID and time using a standard statistical regression software package. These expected numbers should then replace the raw population numbers in the population file, while not including the covariates themselves.

The use of external regression software is also an excellent way to adjust for covariates in the exponential model. The first step is to fit an exponential regression model without any spatial information, in order to obtain risk estimates for each of the covariates. The second step is to adjust the survival and censoring time up or down for each individual based on the risk estimates his or her covariates.

For the normal model, covariates can be adjusted for by first doing linear regression using standard statistical software, and then replacing the observed value with their residuals.

Covariate adjustment using multiple data sets. It is also possible to adjust for categorical covariates using multiple data sets. The cases and controls/population are then divided into categories, and a separate data set is used for each category. This type of covariate adjustment is computationally much slower than the one using the input files, and is not recommended for large data sets. One advantage is that it can be used to adjust for covariates when running the multinomial or ordinal models, for which other adjustment procedures are unavailable. A disadvantage is that since the maximum number of data sets allowed by SaTScan is twelve, the maximum number of covariate categories is also twelve.

The adjustment approach to multiple data sets is as follows (when searching for clusters with high rates):

1. For each window location and size, the log likelihood ratio is calculated for each data set.
2. The log likelihood ratio for all data sets with less than expected number of cases in the window is multiplied with negative one.
3. The log likelihood ratios are then summed up, and this sum is the combined log likelihood for that particular window.
4. The maximum of all the combined log likelihood ratios, taken over all the window locations and sizes, constitutes the most likely cluster, and this is evaluated in the same way as for a single data set.

When searching for clusters with low rates, the same procedure is performed, except that it is then the data sets with more than expected cases that we multiply by one. When searching for both high and low clusters, both sums are calculated, and the maximum of the two is used to represent the log likelihood ratio for that window.

Spatial and temporal adjustments

Adjusting for temporal trends. If there is an increasing temporal trend in the data, then the temporal and space-time scan statistics will pick up that trend by assigning a cluster during the end of the study period. If there is a decreasing trend, it will instead pick up a cluster at the beginning of the time period. Sometimes it is of interest to test whether there are temporal and/or space-time clusters after adjusting for a temporal trend.

For the space-time permutation model, the analysis is automatically adjusted for both temporal trends and temporal clusters, and no further adjustments are needed. For the discrete Poisson model, the user can specify whether a temporal adjustment should be made, and if so, whether to adjust with a percent change or non-parametrically.

Sometimes, the best way to adjust for a temporal trend is by specifying the percent yearly increase or decrease in the rate that is to be adjusted for. This is a log linear adjustment. Depending on the application, one may adjust either for a trend that SaTScan estimates from the data being analysed, or from the trend as estimated from national or other similar data. In the latter case, the percent increase or decrease must be calculated using standard statistical regression software such as R, and then inserted on the Risk Adjustments Tab.

For space-time analyses, it is also possible to adjust for a temporal trend non-parametrically. This adjusts the expected count separately for each aggregated time interval, removing all purely temporal clusters. The randomization is then stratified by time interval to ensure that each time interval has the same number of events in the real and random data sets.

The ability to adjust for temporal trends is much more limited for the Bernoulli, multinomial, ordinal, normal and exponential models, as none of the above features can be used. Instead, the time must be divided into discrete time periods, with the cases and controls in each period corresponding to a separate data set with separate case and control files. The analysis is then done using multiple data sets.

Adjusting for day-of-week effects. Some data sets have a weekly pattern. If not adjusted for, that could create clusters, for example on a Monday, or from one Monday to the next Monday, simply because Mondays in general have more events than other days of the week. One way to adjust for this is to aggregate daily data into weeks, but that will reduce the temporal resolution. Another option is to select the day-of-week adjustment feature on the Spatial and Temporal Adjustment Tab, which will non-parametrically adjust for any weekly adjustment in the data. This feature is only available with the discrete Poisson probability model.

The space-time permutation model automatically adjusts for any purely temporal variation in the data, including day-of-week effects. Hence, with this probability model, there is never a need to do any special day-of-week adjustment. If different spatial locations have different day-of-week effects, that may lead to spurious space-time interaction clusters.

Example

Your disease data come from different medical clinics, but only some of the clinics are open on the weekends. That may result in weekend clusters at those clinics that are simply an artefact of their opening hours. To adjust for this, it is possible to adjust for the space-by-day-of-week interaction by selecting that option on the Spatial and Temporal Adjustment Tab. Doing this has exactly the same effect as including a day-of-week variable in the input case file. This feature is only valid for the space-time permutation model.

Adjusting for purely spatial clusters. In a space-time analysis with the Poisson model, it is also possible to adjust for purely spatial clusters, in a non-parametric fashion. This adjusts the expected count separately for each location, removing all purely spatial clusters. The randomization is then stratified by location ID to ensure that each location has the same number of events in the real and random data sets.

This option is not available for the Bernoulli, multinomial, ordinal, exponential, normal or space-time permutation models, in the latter case because the method automatically adjusts for any purely spatial clusters.

Adjusting for known relative risks. Sometimes it is known *a priori* that a particular location and/or time has a higher or lower risk of known magnitude, and we want to detect clusters above and beyond this, or in other words, we want to adjust for this known excess/lower risk. One way to do this is to simply change the population at risk numbers in the population file. A simpler way is to use the adjustments file. In this file, a relative risk is specified for any location and time period combination. The expected counts are then multiplied by this relative risk for that location and time.

Example

If it is known from historical data that a particular location typically have 50 percent more cases during the summer months June to August, then for each year one would specify a relative risk of 1.5 for this location and these months. A summer cluster will then only appear in this location if the excess risk is more than 50 percent.

This feature is only available for the discrete Poisson model.

5.4.7 Missing data

If there are missing data for some locations and times, it is important to adjust for that in the analysis. If not, you may find statistically significant low rate clusters where there is missing data, or statistically significant high rate clusters in other locations, even though these are simply artefacts of the missing data.

Some guidance is given below on how to adjust the different probability model types for missing data.

Bernoulli model	If cases are missing for a particular location and time period, remove the controls for that same location and time. Likewise, if controls are missing for a particular location and time, remove the cases for that same location and time. This needs to be done before providing the data to SaTScan. If both cases and controls are missing for a location and time, you are fine, and there is no need for any modification of the input data.
Multinomial and ordinal models	If one or more categories are missing for a particular location and time period, remove all cases in the remaining categories from that same location and time. This needs to be done before providing the data to SaTScan. If all cases in all categories are missing for a location and time, you are fine, and there is no need for any modification of the input data.
Discrete Poisson model	Use the adjustments file to define the location and time combinations for which the data is missing, and assign a relative risk of zero to those location/time combinations.
Continuous Poisson model	Redefine the study area buy using a different set of polygons, so that areas with missing data are excluded from the study area.
Space-time permutation model	It is a little more complex to adjust for missing data in the space-time permutation model, but still possible. First add day-of-week as a covariate in the analysis file. When a particular location / time period is missing, then for that location, remove all data for the days of the week for which any data is missing. Note that, in addition to adjusting for the missing data, this approach will also adjust for any day-of-week by spatial interaction effects. The same approach can be used with other categorisation of the data, as long as the categorisations is in some time-periodic unit that occur several times and is evenly spread out over the study period. Two more crude approaches to deal with missing data in the space-time permutation model is to remove all data for a particular location if some data are missing for that location or to remove all data for a particular time period for dates on which there is missing data in any location. The latter is especially useful in prospective surveillance for missing data during the beginning of the study period, to avoid removing recent data that are the most important for the early detection of disease outbreak.

5.4.8 Multivariate scan with multiple data sets

Sometimes it is interesting to simultaneously search for and evaluate clusters in more than one data set.

Example

One may be interested in spatial clusters with excess incidence of leukaemia only, of lymphoma only or of both simultaneously.

As another example, one may be interested in detecting a gastrointestinal disease outbreak that affects children only, adults only or both simultaneously.

If SaTScan is used to analyse one single combined data set, one may miss a cluster that is only present in one of the subgroups. On the other hand, if two SaTScan analyses are performed, one for each data set, there is a loss of power if the true cluster is about equally strong in both data sets. A SaTScan analysis with multiple data sets and the multivariate scan option solves this problem.

The multivariate scan statistic with multiple data sets works as follows (when searching for clusters with high rates):

1. For each window location and size, the log likelihood ratio is calculated for each data set.
2. The log likelihood ratios for the data sets with more than expected number of cases is summed up, and this sum is the likelihood for that particular window.
3. The maximum of all the summed log likelihood ratios, taken over all the window locations and sizes, constitutes the most likely cluster, and this is evaluated in the same way as for a single data set.

When searching for clusters with low rates, the same procedure is performed, except that we instead sum up the log likelihood ratios of the data sets with fewer than expected number of cases within the window in question. When searching for both high and low clusters, both sums are calculated, and the maximum of the two is used to represent the log likelihood ratio for that window.

6 Spatial models of disease risk

6.1 Introduction

Spatial risk assessment facilitates an informed appraisal of disease risk and prioritisation of areas within which to concentrate control activities, and enable decision makers to develop risk management strategies. A number of conditions must be met to perform such assessments. Firstly, they require a level of understanding of the underlying causal processes; it is assumed that there is spatial heterogeneity of disease risk across the area considered (or such a risk assessment is pointless!), and that risk groups or risky behaviours contribute to disease spread. This should lead to a set of criteria defining what is meant by 'high risk'. Secondly, relevant data (usually from a range of sources) must be available, plus a methodology of combining data from different sources to inform the net risk status. The most basic data needed would be georeferenced, quantitative information about disease occurrence and the population at risk. These data can often be complemented by various types of risk factor data such as attributes of potentially at-risk individuals or groups, their contact networks, or environmental information.

Disease status information can be collected using targeted or scanning surveillance activities. Once a disease event has been identified a record of the event is made by the herd / flock owner or, in the case of notifiable diseases, members of the state veterinary service. For diseases that are of transboundary significance, international animal health authorities such as the OIE are notified. In this situation a record of the disease event is recorded in the World Animal Health Information System (WAHIS).

WAHIS data, when combined with details of the spatial distribution of the animal population at risk, allows one to visualise the spatial pattern of disease risk. The higher the spatial resolution of the data for a given area, the higher the statistical power for detecting events that occur in small regions. Analyses aimed at explaining the variation in risk require access to risk factor information, which may be collected as part of targeted surveillance activities or may be accessed by linking surveillance data to census information or other risk factors such as environmental information.

6.2 Data-driven and knowledge-driven models of disease risk

The spatial analysis tools suitable for risk assessment include the whole range of methods from visualisation of disease events through to modelling techniques. Modelling techniques can be categorised into data-driven and knowledge-driven methods.

The former is characterised by the use of statistical methods for defining relationships between risk factors and disease risk as the outcome variable (usually using a regression-based approach)

– it follows that they tend to be used when details are available documenting disease events and the size (and location) of the population at risk.

Knowledge-driven modelling approaches are based on existing knowledge about the causal relationships associated with the disease risk of interest, and tend to be used when no disease event records are available or the quality of the details of disease events or the population at risk is questionable.

6.2.1 Data-driven models

Data-driven models rely on statistical analysis using data collected through surveillance and other means. They generate quantitative estimates of risk and the relative weights of risk factors, including the corresponding levels of statistical confidence and uncertainty. In common with other quantitative models, there is a perception of higher validity than knowledge-based ones due to the apparently more objective method of defining the relationships. However, it needs to be emphasized that they are strongly dependent on the quality of the data and the validity of the model in the context of a particular decision problem.

The statistical uncertainty associated with the outputs from such models should be presented together with the predicted risk estimates. For example, this can be done by presenting maps of the risk estimates together with maps of some specified upper and lower confidence limits. The extent of bias associated with these outputs should be presented in a qualitative commentary. Decision makers need to be informed of these constraints, so that these are incorporated into the decision-making process.

These models incorporate both spatial and temporal elements, and they tend to be mathematically complex. Further treatment of them is out of scope for this course.

6.2.2 Knowledge-driven models

A strength of the knowledge-driven approach is that they can incorporate information on transmission dynamics and disease spread without requiring quantitative data. A typical use of knowledge-driven approaches would be when a country that has not experienced an outbreak of highly pathogenic avian influenza wants to identify areas of the country where an incursion of the disease is more likely to occur.

Qualitative or quantitative risk estimates can be produced based on existing or hypothesised understanding of the causal relationships leading to disease occurrence. This means that either expert opinion or the application of participatory techniques can be used to elicit information that is not available in any other format. Disadvantages can be that the models become rather theoretical; as qualitative models, they may have a strong subjective element, and are only loosely connected to real data. Validation of the resulting maps is not always possible due to this lack of data, and is frequently limited to visual comparisons with existing data sources.

Models generated using this approach may be static or dynamic. Static models are defined as sets of linked rules of attribute information combined to produce risk estimates. Dynamic models are similar, but reproduce patterns of change in space *and* time with respect to disease status in a population as a result of specified spatial and attribute factors.

6.3 Spatial risk assessment using a knowledge-driven approach

6.3.1 Multicriteria Decision Analysis (MCDA)

One methodology for implementing knowledge-driven models is by using multicriteria decision analysis (MCDA). MCDA establishes preferences between options by reference to an explicit set of objectives that the decision maker has identified. It offers a number of ways of aggregating data on individual criteria to provide indicators of the overall performance of options.

MCDA methods come from the field of operations research and are commonly used in environmental, industrial and business management. Various analogous terms exist, e.g. multicriteria decision-making (MCDM). A number of different MCDA algorithms have been developed to analyse different types of decision problems. Because MCDA is widely applied in spatial analyses of non-disease applications, the use of the word ‘criteria’ is somewhat at odds with our standard epidemiological terminology; we are more familiar with words such as ‘factors’, ‘determinants’ or ‘indicators’.

In our context, MCDA is an analytical technique used to estimate the relative importance of risk factors hypothesised to contribute to a given disease outcome. It is simply a logical framework that facilitates structuring and definition of the problem, followed by a decision analysis to identify which criteria are relevant. These criteria can be evaluated using either quantitative and / or qualitative indicators. MCDA is performed by ranking and comparing alternatives based on these criteria. Relative importance ‘weights’ are applied to each of these, allowing one to provide an overall summary estimate of spatial risk.

MCDA involves a sequence of analytical steps:

1. Defining the objective(s) of the analysis.
2. Defining the factors and constraints relevant to the problem.
3. Defining the relationship between each factor and their association with the objective.
4. Standardising risk factor layers so they can be compared.
5. Defining the relative importance of each factor in relation to the objective.
6. Combining all factors and constraints to produce a final weighted estimate of suitability for each location in the study area.
7. Sensitivity analysis and validation.

This logic can be illustrated in a process diagram. Figure 6.1 shows an MCDA chart developed for a spatial assessment of FMD risk in Laos.

A key feature of MCDA is its emphasis on the judgement of the decision making team in establishing objectives and criteria and estimating relative importance weights. The advantages of MCDA over informal judgement procedures are as follows:

- The choice of objectives and criteria that any decision making group may make are open to interpretation and to change if they are felt to be inappropriate.
- Scores and weights, when used, are also explicit and are developed according to established techniques. They can also be cross-referenced to other sources of information on relative values, and amended if necessary.
- Scoring of comparisons can be sub-contracted to experts, so MCDA need not necessarily be left in the hands of the decision making body itself.

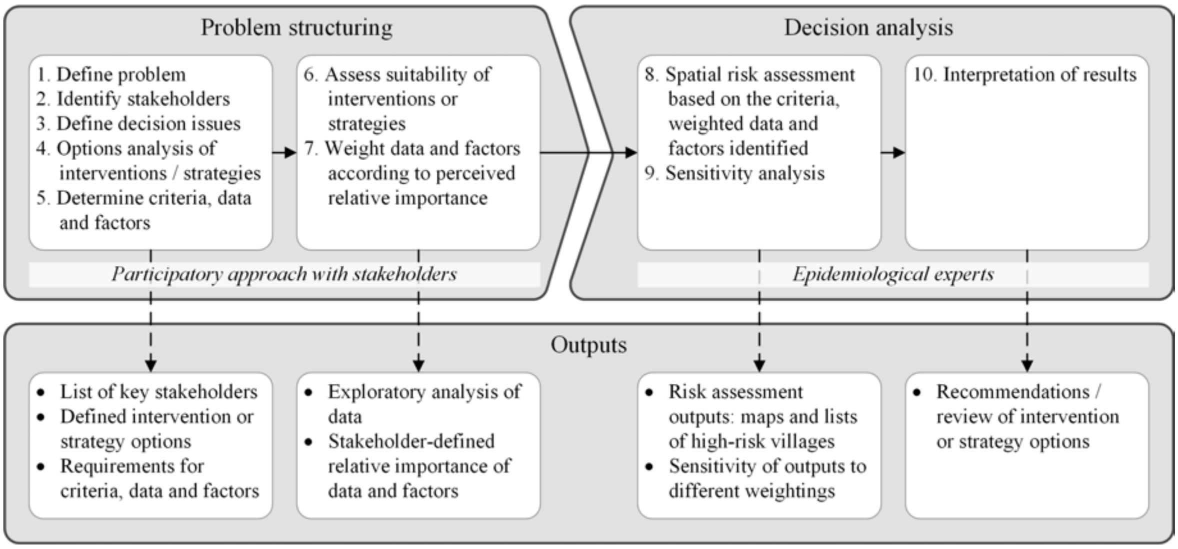

Problem structuring
1. Define problem
2. Identify stakeholders
3. Define decision issues
4. Options analysis of interventions / strategies
5. Determine criteria, data and factors
6. Assess suitability of interventions or strategies
7. Weight data and factors according to perceived relative importance
Participatory approach with stakeholders
Decision analysis
8. Spatial risk assessment based on the criteria, weighted data and factors identified
9. Sensitivity analysis
10. Interpretation of results
Epidemiological experts
Outputs
• List of key stakeholders
• Defined intervention or strategy options
• Requirements for criteria, data and factors
• Exploratory analysis of data
• Stakeholder-defined relative importance of data and factors
• Risk assessment outputs: maps and lists of high-risk villages
• Sensitivity of outputs to different weightings
• Recommendations / review of intervention or strategy options

Figure 6.1: Multicriteria decision analysis (MCDA) logic (adapted from Aenishaenslin et al. (2013)) applied to perform the FMD spatial risk assessment.

- MCDA can provide an important means of communication, within the decision making body and sometimes, later, between that body and the wider community.
- The scores and weights used in an MCDA provides an audit trail.

A strength of this technique is that in the absence of quantitative data for a criterion in a specific context (such as in knowledge-driven spatial risk assessment), MCDA methods allow for the incorporation of qualitative evaluations. Conversely, the subjectivity that pervades this can be a matter of concern.

We can differentiate two main approaches (although hybrids of these are possible):

1. **Expert consultation** following a structured or standardised approach (e.g. a two-stage Delphi process). The objective is to reach the greatest degree of consensus using such ‘expert opinion’.

2. A **participatory approach** which is inclusive of all stakeholders concerned by a particular issue, allowing them to actively engage in all stages of the decision analysis supported by MCDA. The techniques are more flexible, e.g. relying on focus group discussions and semi-structured interviews. The objective is to incorporate the fullest range of opinion and elicit a broad scala of knowledge.

The first approach strives for a degree of systematic rigour; it is more suited to problems which are discrete and well-defined, on which substantial literature has been published, and for which individual experts can be identified. This approach aims to be ‘semi-quantitative’. The second approach is qualitative. It can be highly effective when dealing with problems for which little published or structured data are available, or ones which involve a wide array of stakeholders. The advances in, and increased uptake of, participatory approaches over the previous decade or so attest to its effectiveness for handling complex, transdisciplinary and multi-sectoral decision-making problems.

6.3.2 Combining risk factor layers to derive an overall estimate of risk: the Weighted Linear Combination (WLC)

The basic principle of knowledge-driven models is to define a set of weighted rules for the factors or criteria which may influence disease risk at any given location. If multiple such criteria affect the disease outcome, a method needs to be adopted to combine them. One such method is the weighted linear combination (WLC) in which criteria are standardized for comparison on a common scale and then weights applied to each criterion so that more important criteria exert a greater influence on the outcome. There are a number of sources of information that may be utilized to determine the weights for WLC. Ideally, they would be derived from statistical data and published literature. However, in many cases such information does not exist; in this case, expert opinion or participatory elicitation of knowledge can be used, in extension to (or part of) the MCDA process described above. Finally, a weighted average across criteria is calculated.

There are a number of common pitfalls in the application of WLC are discussed. These are well reviewed by Malczewski (2000), and include the following.

1. **Data biases and incompleteness:** the attributes (i.e. factors and constraints) should be measurable and complete (i.e. cover all relevant aspects of the decision problem). Factor selection should not be based on data availability alone. However, for our purposes, comprehensive data on relevant disease factors are not always available and it is often necessary to select attributes from limited available data resources.

2. **Correlation between attributes:** this is referred to as a redundancy problem that gives rise to double-counting. In other words, if two factors are similar – that is, have the same direction or type of influence on the disease – and we include both, we are artificially amplifying their effect.

3. **Spatial scale and levels of aggregation (MAUP):** this is addressed in 2.8.4 above.

4. **Attribute linearity:** this occurs when transformation of the attribute for subsequent comparison does not take into account possible non-linear associations with the outcome.

5. **Incorrect weighting:** this may result from failing to consider the unit of measurement and range of the attribute.

There are different ways of calculating the weighted average across the criteria. The two most common methods are discussed below.

Linear additive approaches

If it can either be proved, or reasonably assumed, that the risk factors for disease independent of each other and if uncertainty is not formally built into the MCDA model, then a simple linear additive evaluation model is appropriate. The linear model allows one to combine each disease risk factor into a single combined overall value. This is done by multiplying the value assigned for each risk factor by the weight assigned to that factor and then adding all of the weighted scores together.

This simple arithmetic approach is only appropriate if the criteria are mutually preference independent. Most MCDA applications use this additive approach. Models of this type have a well-established record of providing robust and effective support to decision-makers working on a range of problems and in various circumstances.

The Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP) also develops a linear additive model, but, in its standard format, uses procedures for deriving the weights and the scores achieved by alternatives which are based, respectively, on pairwise comparisons between criteria and between options. Thus, for example, in assessing weights, the decision maker is asked a series of questions, each of which asks how important one particular criterion is relative to another for the decision being addressed.

The strengths and weaknesses of the AHP have been the subject of substantial debate among specialists in MCDA. It is clear that users generally find the pairwise comparison form of data input straightforward and convenient. On the other hand, doubts have been raised about the theoretical foundations of the AHP and about some of its properties. In particular, the rank reversal phenomenon has caused concern. This is the possibility that, simply by adding another option to the list of options being evaluated, the ranking of two other options, not related in any way to the new one, can be reversed. This is seen by many as inconsistent with rational evaluation of options and thus questions the underlying theoretical basis of the AHP.

At the core of the AHP lies a method for converting subjective assessments of relative importance to a set of overall scores or weights developed by Saaty (1980). The fundamental input to the AHP is the decision maker's answers to a series of questions of the general form, 'How important is criterion A relative to criterion B?'. These are termed pairwise comparisons. Responses are gathered in verbal form and subsequently codified on a nine-point scale, as shown in Table 6.1.

Table 6.1: Scores that might be assigned to the relative importance of each risk factor pair using the AHP.

How important is A relative to B?	Preference index
Overwhelmingly not more important	1/9
Very strongly not more important	1/7
Strongly not more important	1/5
Moderately not more important	1/3
Equally important	1
Moderately more important	3
Strongly more important	5
Very strongly more important	7
Overwhelmingly more important	9

A worked example

If decision makers indicate that they are of the opinion that risk factor A is very strongly more important than risk factor B, the A-B comparison is assigned a preference index 7 (Table 6.1) and the B-A comparison automatically is assigned a score of the reciprocal of 7, 1/7. Because decision makers are assumed to be consistent in making judgements about any one pair of criteria and since all criteria will always rank equally when compared to themselves, it is only ever necessary to make $0.5 \times n \times (n-1)$ comparisons to establish the full set of pairwise judgements for *n* criteria. A typical matrix for establishing the relative importance of three criteria is shown in the table below. The next step is to estimate the set of weights (three in the example shown in the table) that are most consistent with the relativities expressed in the matrix. A simple approach is to: (a) calculate the geometric mean of each row in the matrix; (b) total the geometric means, and (c) normalise each of the geometric means by dividing by the total just computed.

Say, for argument's sake we are conducting a MCDA to identify risk factors for highly pathogenic avian influenza into a country that has not previously had an outbreak. Our hypothesised risk factors are the presence or absence of backyard poultry (BYP), commercial poultry (CP), wetlands (WET), rivers (IRR), wild bird aggregation areas (WB), water bodies (WB), airports (AIR) and roads (ROA). A team of experts was assembled and provided the pairwise importance scores shown in the table.

	BYP	CP	WET	IRR	WB	WAT	AIR	ROA
BYP	1	-	-	-	-	-	-	-
CP	0.2	1	-	-	-	-	-	-
WET	3	5	1	-	-	-	-	-
IRR	3	5	1	1	-	-	-	-
WB	3	5	1	1	1	-	-	-
WAT	3	5	1	1	1	1	-	-
AIR	5	7	0.3	0.3	0.3	0.3	1	-
ROA	5	7	0.3	0.3	0.3	0.3	1	1

Once the pairwise scores have been assigned the matrix is completed, as shown in the following table.

	BYP	CP	WET	IRR	WB	WAT	AIR	ROA
BYP	1	5	0.3	0.3	0.3	0.3	0.2	0.2
CP	0.2	1	0.2	0.2	0.2	0.2	0.1	0.1
WET	3	5	1	1	1	1	3	3
IRR	3	5	1	1	1	1	3	3
WB	3	5	1	1	1	1	3	3
WAT	3	5	1	1	1	1	3	3
AIR	5	7	0.3	0.3	0.3	0.3	1	1
ROA	5	7	0.3	0.3	0.3	0.3	1	1

Each column of this table is then summed and each cell is divided by its respective column total. The final score for each risk factor is then the mean of each row, as shown in the following table.

	BYP	CP	WET	IRR	WB	WAT	AIR	ROA	Score
BYP	0.04	0.12	0.06	0.06	0.06	0.06	0.01	0.01	0.06
CP	0.01	0.03	0.04	0.04	0.04	0.04	0.01	0.01	0.03
WET	0.13	0.12	0.19	0.19	0.19	0.19	0.21	0.21	0.18
IRR	0.13	0.12	0.19	0.19	0.19	0.19	0.21	0.21	0.18
WB	0.13	0.12	0.19	0.19	0.19	0.19	0.21	0.21	0.18
WAT	0.13	0.12	0.19	0.19	0.19	0.19	0.21	0.21	0.18
AIR	0.22	0.17	0.06	0.06	0.06	0.06	0.07	0.07	0.10
ROA	0.22	0.17	0.06	0.06	0.06	0.06	0.07	0.07	0.10

We can now take these final risk scores and use them as a weight for each risk factor layer in a GIS.

The REMBRANDT technique

The REMBRANDT (Ratio Estimation in Magnitudes or deci-Bells to Rate Alternatives which are Non-Dominated) technique is an MCDA approach similar to the AHP. It uses systematic pairwise comparisons analysed on a multiplicative scale (as opposed to linear) scale.

REMBRANDT deals with three contended flaws in the AHP (and similar) methods:

1. Ratings are estimated on a logarithmic scale (as opposed to the linear 1 to 9 scale used in the AHP method).
2. Use of geometric means avoids the problem of rank reversal – the situation where relative rankings change after a factor is added or deleted.
3. Aggregation of scores by arithmetic means is replaced by the product of alternative relative scores weighted by the power of weights obtained from analysis of hierarchical elements above the alternatives.

Using the REMBRANDT technique, pairwise comparative judgements are made using a different numerical scale (Table 6.2).

Table 6.2: Scores that might be assigned to the relative importance of each risk factor pair using the REMBRANDT technique.

How important is A relative to B?	Preference index
Overwhelmingly not more important	-8
Very strongly not more important	-6
Strongly not more important	-4
Moderately not more important	-2
Equally important	0
Moderately more important	2
Strongly more important	4
Very strongly more important	6
Overwhelmingly more important	8

A worked example

Resuming our previous example, we populate a matrix of the form shown in the table below:

	BYP	CP	WET	IRR	WB	WAT	AIR	ROA
BYP	0	-	-	-	-	-	-	-
CP	-4	0	-	-	-	-	-	-
WET	2	4	0	-	-	-	-	-
IRR	2	4	0	0	-	-	-	-
WB	2	4	0	0	0	-	-	-
WAT	2	4	0	0	0	0	-	-
AIR	4	6	-2	-2	-2	-2	0	-
ROA	4	6	-2	-2	-2	-2	0	0

Once the pairwise scores have been assigned the matrix is completed, as shown in the following table.

	BYP	CP	WET	IRR	WB	WAT	AIR	ROA
BYP	0	4	-2	-2	-2	-2	-4	-4
CP	-4	0	-4	-4	-4	-4	-8	-8
WET	2	4	0	0	0	0	2	2
IRR	2	4	0	0	0	0	2	2
WB	2	4	0	0	0	0	2	2
WAT	2	4	0	0	0	0	2	2
AIR	4	6	-2	-2	-2	-2	0	0
ROA	4	6	-2	-2	-2	-2	0	0

With the REMBRANDT technique, when there are six to nine categories, the following approach is taken: (1) set γ to log(2) and transform each pairwise comparison score as $\exp^{\gamma^{i}}$; (2) calculate the geometric mean of each row of the matrix; and (3) sum the geometric means across rows and derive a weight for each risk factor by dividing each of the row geometric means by the row totals, as shown below.

	BYP	CP	WET	IRR	WB	WAT	AIR	ROA	gmeann	weight
BYP	1.00	16.0	0.25	0.25	0.25	0.25	0.06	0.06	0.35	0.03
CP	0.06	1.00	0.06	0.06	0.06	0.06	0.00	0.00	0.04	0.03
WET	4.00	16.0	1.00	1.00	1.00	1.00	4.00	4.00	2.38	0.19
IRR	4.00	16.0	1.00	1.00	1.00	1.00	4.00	4.00	2.38	0.19
WB	4.00	16.0	1.00	1.00	1.00	1.00	4.00	4.00	2.38	0.19
WAT	4.00	16.0	1.00	1.00	1.00	1.00	4.00	4.00	2.38	0.19
AIR	16.0	64.0	0.25	0.25	0.25	0.25	1.00	1.00	1.19	0.10
ROA	16.0	64.0	0.25	0.25	0.25	0.25	1.00	1.00	1.19	0.10

6.4 Spatial risk assessments of FMD

A number of FMD risk assessment studies have been undertaken:

- Gallego et al. (2007) used FMD outbreak data to investigate the impacts of different vaccination and intervention measures implemented in Colombia between 1982 and 2003; they identified long-term and cyclical trends in incidence, morbidity and measures of frequency, and concluded that "efficient and timely control and eradication of FMD is not likely to be forthcoming in countries where the basic elements of disease control are not in place and strongly enforced and where the disease has been endemic for years".
- Shiilegdamba et al. (2008) analysed outbreak data from epidemic FMD outbreaks in Mongolia between 2000 and 2002, and showed that there was significant localised spatial clus-

tering of FMD. They reported that while no causes for the outbreaks could be identified, illegal animal movements were likely to be implicated. They concluded that identification of spatial clusters are helpful to identify risk factors for FMD transmission, and can be applied to direct control and prevention measures in high-risk areas.

- A recent study by Amaral et al. (2016) applied a knowledge-based MCDA methodology, utilising the opinions of eight experts to determine a WLC of risk factors to develop risk maps for the introduction of FMD along the border between Brazil and Paraguay. The factors under consideration were very similar to our investigation, but their study did not incorporate any FMD-specific data.
- In contrast, a number of studies in Southern and East Africa applied more data-driven techniques to model spatial distribution and risk factors, using outbreak data in combination with other geographic data.
  - Hamoonga et al. (2014) determined associations between FMD outbreak intensity in Zambia and distance to the nearest major international border crossing, distance to the nearest major road, wetness index and elevation.
  - Sinkala et al. (2014) applied the spatial scan statistic to outbreak data, also from Zambia, to identify FMD clusters between 2004 and 2012.
  - Using Tanzanian data, Allepuz et al. (2015) found substantial spatial heterogeneity in high-risk areas from 2001 to 2006, and differentiated between endemic and epidemic phases. They determined that proximity to main roads was a consistent risk factor for FMD occurrence; other spatial factors (including cattle density) played a variable role in the risk of FMD.

Comparable spatial risk assessments have not yet been published in the South East Asia region. This is of high relevance, specifically considering the increasing emphasis on the application of risk-based approaches for the control and eradication of FMD, e.g. as applied by SEACFMD.

Part II

Course practicals

7 Practical 1. Getting set up

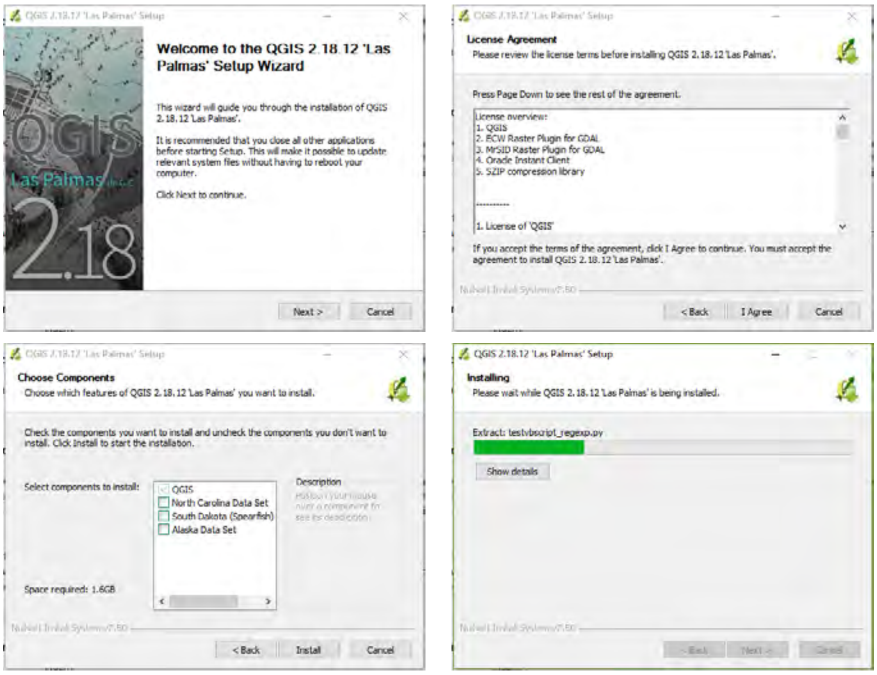
7.1 Objective

In this first practical, we will get set up with QGIS and the required functionality for this course, and identify and develop map layers containing spatial objects and data which will form the 'background' of our subsequent analysis – the 'base map'. We will also cover generating nice visual map output for inclusion in reports and other dissemination.

7.2 Preparing your QGIS setup

7.2.1 Installing QGIS

If you have not yet installed the current version of QGIS, 2.18.17 'Las Palmas', locate the installer and run it. The installation process is pretty standard (Figure 7.1).



QGIS 2.18.12 'Las Palmas' Setup

Welcome to the QGIS 2.18.12 'Las Palmas' Setup Wizard

This wizard will guide you through the installation of QGIS 2.18.12 'Las Palmas'.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

QGIS 2.18.12 'Las Palmas' Setup

License Agreement

Please review the license terms before installing QGIS 2.18.12 'Las Palmas'.

Press Page Down to see the rest of the agreement.

```
License overview:
1. QGIS
2. ECW Raster Plugin for GDAL
3. MrSID Raster Plugin for GDAL
4. Oracle Instant Client
5. SZIP compression library


----------

1. License of 'QGIS'
```

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install QGIS 2.18.12 'Las Palmas'.

QGIS 2.18.12 'Las Palmas' Setup

Choose Components

Choose which features of QGIS 2.18.12 'Las Palmas' you want to install.

Check the components you want to install and uncheck the components you don't want to install. Click Install to start the installation.

Select components to install:

- ☑ QGIS
- ☐ North Carolina Data Set
- ☐ South Dakota (Spearfish)
- ☐ Alaska Data Set

Description

Position your mouse over a component to see its description.

Space required: 1.6GB

QGIS 2.18.12 'Las Palmas' Setup

Installing

Please wait while QGIS 2.18.12 'Las Palmas' is being installed.

Extract: testvbscript_regexp.py

Figure 7.1: QGIS installation screens.

After installing, start QGIS Desktop. The main (or ‘Application’) window opens. You should see a graphical user interface similar to Figure 7.2.

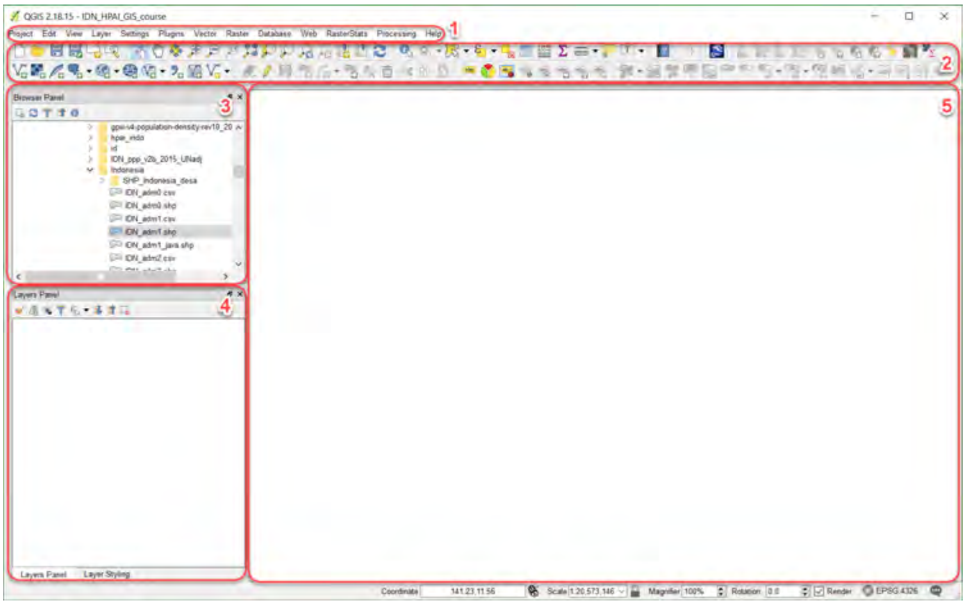
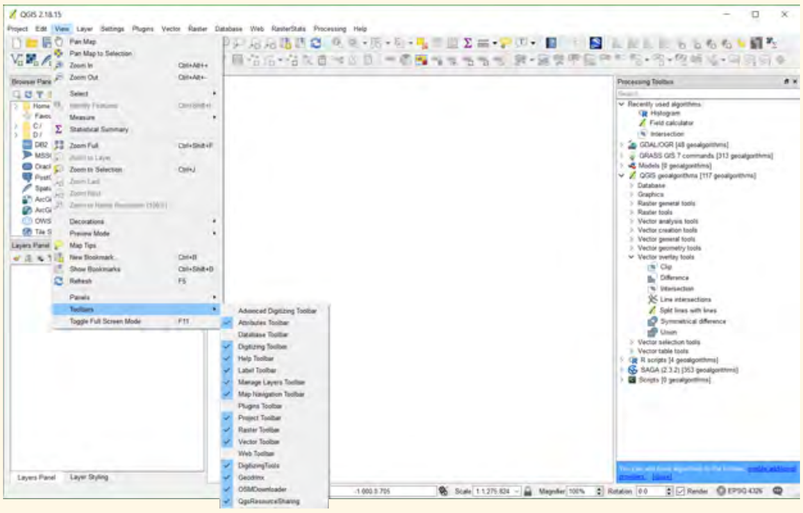

QGIS 2.18.15 - IDN_HPAI_GIS_course
Project Edit View Layer Settings Plugins Vector Raster Database Web RasterStats Processing Help
1
2
Browser Panel
3
Layers Panel
4
5
Layers Panel Layer Styling
Coordinate 141.23.11.56 Scale 1:20,573,146 Magnifier 100% Rotation 0.0 Render EPSG:4326

Figure 7.2: QGIS Application window, showing different panels. 1: Menu bar; 2: Toolbars; 3: Browser panel; 4: Layers panel; 5: Main map canvas.

Tip! Customise QGIS properties and settings

QGIS is highly customisable. You can personalise your QGIS setup in a few ways. This will make your work more efficient. Here are some tips:

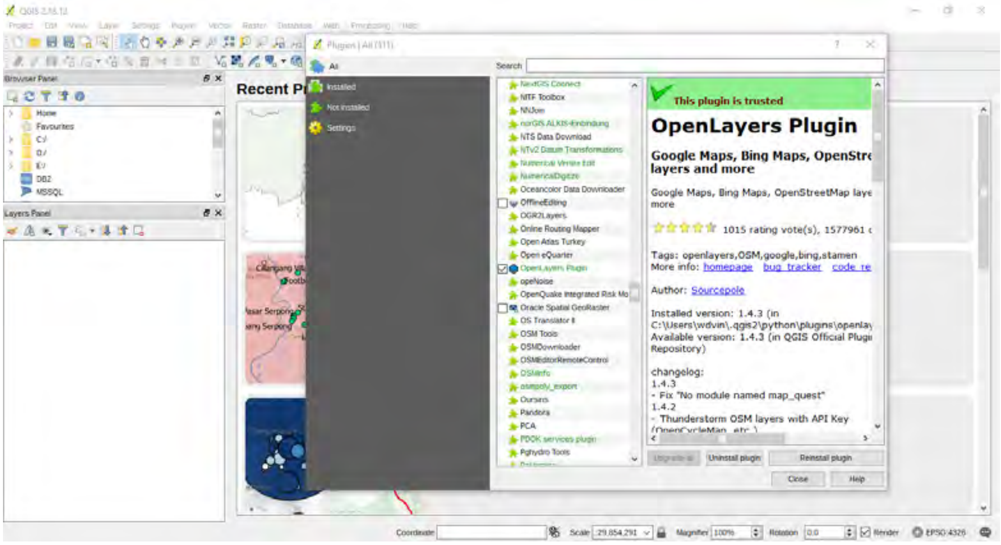
- Have a look at some of the settings under Settings – Options.
- You can specify your QGIS project settings under Project – Properties. The default setting of the CRS is WGS84 – this is the projection we will be using.
- You can determine which icon sets are displayed under View – Toolbars , and View – Panels. Experiment with moving these around. Eventually you will become used to your customised interface!



7.2.2 Downloading plugins

During this course, we will be using some functionality that is not included in the basic installation version; this functionality is provided by plugins. To install these, click on Plugins – Manage and Install Plugins. . . (Figure 7.3), and select and install the following plugins:

- OpenLayers Plugin
- Zonal Statistics plugin
- Spatial Query Plugin
- Cartogram



This plugin is trusted

OpenLayers Plugin

Google Maps, Bing Maps, OpenStre layers and more

Google Maps, Bing Maps, OpenStreetMap laye more

1015 rating vote(s), 1577961 c

Tags: openlayers,OSM,google,bing,stamen
More info: homepage bug tracker code re

Author: Sourcepole

Installed version: 1.4.3 (in C:\Users\wdvin\.qgis2\python\plugins\openlay
Available version: 1.4.3 (in QGIS Official Plugi Repository)

changelog:
1.4.3
- Fix "No module named map_quest"
1.4.2
- Thunderstorm OSM layers with API Key (OpenCycleMap, etc.)

Figure 7.3: Installing plugins in QGIS.

7.3 Preparing your project data directories

Of course, how you manage your data files and directories is entirely up to your own discretion. However, you will find out that when performing spatial analysis, more than other areas of analysis, the number and size of data files increases at a rapid rate! Good data management really helps to stay on top of all this information. Here we provide some generic tips for effective data management:

- **Work using a ‘project’ approach.** Take a few minutes to set up the directory structure you will be using in advance, for a specific piece of work. Don’t mix work done in different projects. It can be surprising how hard it is to locate a file when you need it!
- **Use a consistent directory structure.** Replicate this across different projects you may work on. This way, you will always know where to find, for example, your map image files.
- **Create sub-folders where required.** If you are generating outputs for different purposes, or if you are using spatial data from different sources, it’s a good idea to utilise different sub-folders for these.
- **Different folders for different formats.** For example, report documents, data, spatial files, graphical output, literature and information resources etc. It’s not a good idea to mix these different types of files and documents.
- **Use a naming convention.** This can be useful, if you’re really organised! For example, the prefix `IDN_Java_HPAI` for all files related to spatial analysis of HPAI on Java. Inserting a

date as a suffix can be useful as a timestamp (if you will work with different edited versions of a dataset).

New to GIS?

We have assumed the you have at least a basic knowledge and understanding of GIS principles and terms. If you have not used QGIS previously, a good reference for various actions is http://www.qgistutorials.com/en.

7.4 Loading your spatial data into QGIS

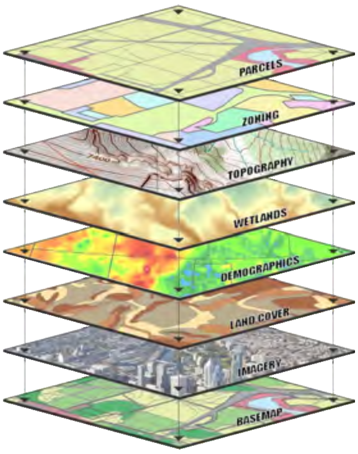
7.4.1 Creating a QGIS project

Click on Project – New or the blank page icon to create a new project. Before starting, save the project to your working folder (Project – Save As. . .) giving it an appropriate name.

Projects are used by QGIS to store details of the views which form part of a QGIS application. A project (`*.qgs`) file records the location of map files and data files associated with the project. Note that a QGIS project file doesn't contain actual spatial information, it simply records details of the locations where spatial data files are stored on your computer. If you move files that have previously been saved as part of an QGIS project the project file won't be able to load properly the next time you try to open it. Consequently, it's advisable to keep all your shapefiles in the same folder as the QGIS project file.

7.4.2 Loading spatial layers

The principle of map layers



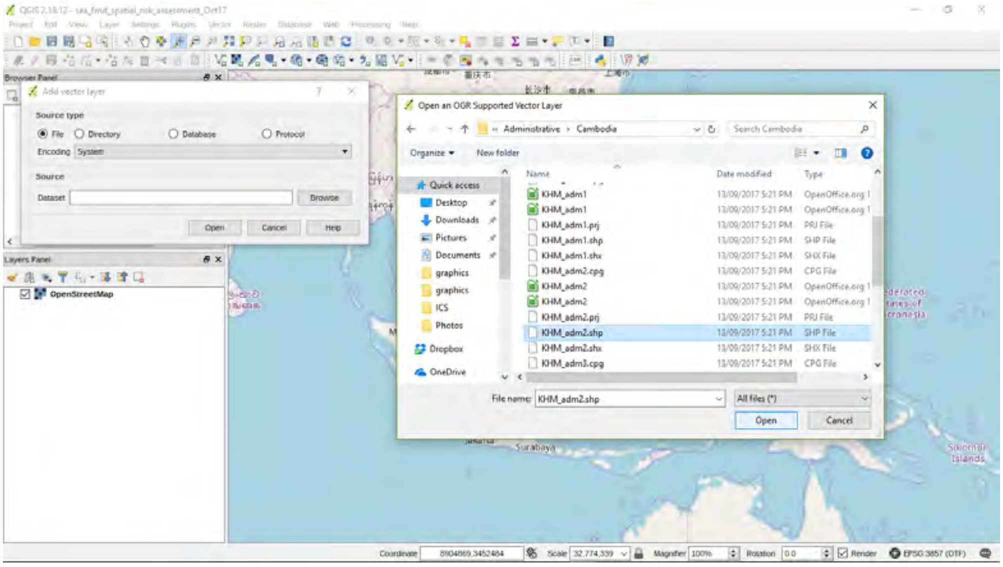
In QGIS, each factor or attribute which contains spatial data (vector or raster) is loaded as a layer. As we are normally interested in multiple spatial factors, we will work with multiple layers.

These layers can be individually displayed, edited or manipulated.

We can also combine the data in multiple of these layers into a single layer. (An analogy with numeric data is combining different variables to create a composite variable.)

Vector data

Vector data are composed of **points**, **lines**, and **polygons** (closed shapes). Points represent discrete locations. These are either true points (the actual location), or virtual points (e.g. the centroid) of a polygon. Lines represent linear features such as rivers and roads. Polygons form closed shapes.



Source type: ● File ○ Directory ○ Database ○ Protocol

Encoding: System

Dataset:

Name	Date modified	Type
KHM_adm1	13/09/2017 5:21 PM	OpenOffice.org 1
KHM_adm1	13/09/2017 5:21 PM	OpenOffice.org 1
KHM_adm1.prj	13/09/2017 5:21 PM	PRJ File
KHM_adm1.shp	13/09/2017 5:21 PM	SHP File
KHM_adm1.shx	13/09/2017 5:21 PM	SHX File
KHM_adm2.cpg	13/09/2017 5:21 PM	CPG File
KHM_adm2	13/09/2017 5:21 PM	OpenOffice.org 1
KHM_adm2	13/09/2017 5:21 PM	OpenOffice.org 1
KHM_adm2.prj	13/09/2017 5:21 PM	PRJ File
KHM_adm2.shp	13/09/2017 5:21 PM	SHP File
KHM_adm2.shx	13/09/2017 5:21 PM	SHX File
KHM_adm3.cpg	13/09/2017 5:21 PM	CPG File

File name: KHM_adm2.shp

All files (*)

Coordinate: 8904869,3452484 Scale: 32,774,339 Magnifier: 100% Rotation: 0.0

Figure 7.4: Opening a vector shapefile in QGIS.

As all points, nodes, and vertices have an explicit and absolute coordinate location, vector data processing can be slow (especially for large polygons such as country administrative boundaries).

You can load the vector shapefiles by clicking on the icon, navigating to the data folder, and then selecting and opening the relevant shapefile (Figure 7.4).

Tip! Use the Browser panel

Use the Browser Panel to navigate to the directories in which you've saved your spatial data files, for quick access.

If you've set a home directory for your project, this will be visible at the top.

If you navigate to the folders containing your spatial laters (vector or raster), you can open these directly by double-clicking on them.

Raster data

In a raster representation, geographic space is divided into a grid of cells (a matrix). Raster files are essentially the same as image files; in analogy with image files, the cells are sometimes called pixels. All cells have the same, defined size. The cell dimension therefore determines the total number of cells in the grid. Each cell has one (and only one) value within a certain range, representing a property or attribute of interest. Consequently, raster data sets are especially suited to the representation of a factor of interest which varies continuously in space, such as elevation, air pressure, temperature etc. Raster maps are often called surface maps, density maps, kernel maps or heatmaps. In animal health, two examples are

- representation of population density;
- estimation of a risk surface of disease.

If the cells are large, there will be fewer cells; as each cell has only one value, this results in less data being contained in the raster, and the raster file will be smaller. However, gradations and trends cannot be as well visualised: the resolution is less.

Because cell size is constant in both the horizontal and vertical directions, if we know the size of the matrix, one coordinate is sufficient to spatially reference the whole raster. By convention, this is called the 'origin' and it's normally the upper left corner. This makes raster data processing much faster. However, as each cell has a unique value, the file sizes of raster data sets can be very large in comparison to vector data sets.

You can load raster files by clicking on the icon, navigating to the data folder, and then selecting and opening the relevant raster file (this is usually a .TIF format).

7.5 Developing the 'base map'

7.5.1 What is a base map and why is it useful?

A base map is defined as "a background georeferenced image that gives a point of reference on a map. Basemaps are non-editable and provides aesthetic appeal such as aerial imagery, topography, terrain and street layers".

We will make a base map by loading administrative boundary data of the six countries in South East Asia reporting FMD. We will build on this basemap in a subsequent practical to investigate trends of disease incidence and reporting. In addition, we will load and prepare a number of other layers which will be used later during this course, so that they are available when we need them.

7.5.2 Creating a view

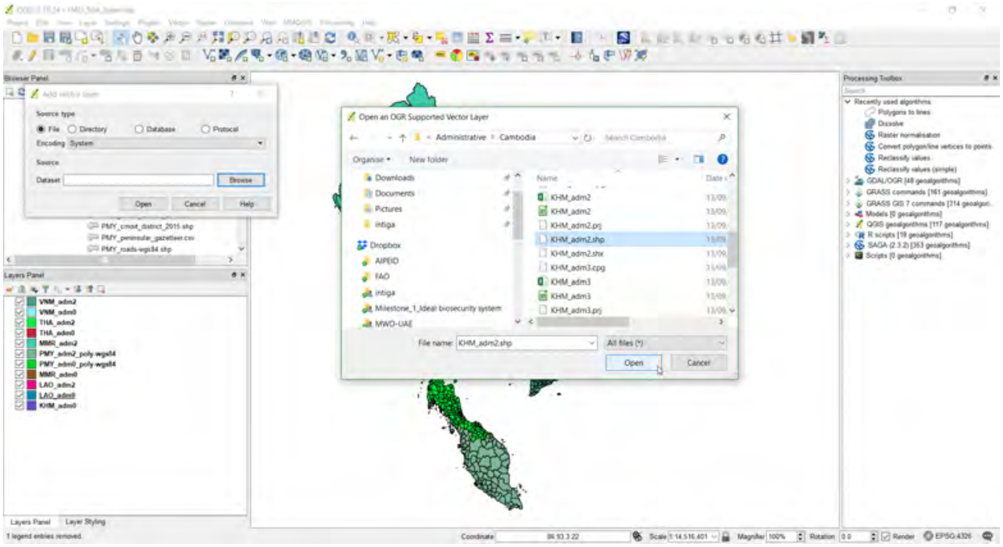
A View is a visual representation of spatial data and the attribute data associated with it. This is where the content and visual presentation of a map is defined. Within a View there can be multiple layers of spatial data.

Referring back to Figure 7.2, the View window consists of two sections:

- The section on the left is the Layers Panel. It lists the layers which are part of the map display by name and shows the symbols used to represent each layer. The check box next to each layer indicates if it is included in the map display. If multiple layers are selected, they will be rendered in order from top to bottom. You can reorder the layers by clicking and dragging.
- The second on the right is the main map canvas; this shows a preview of the map output.

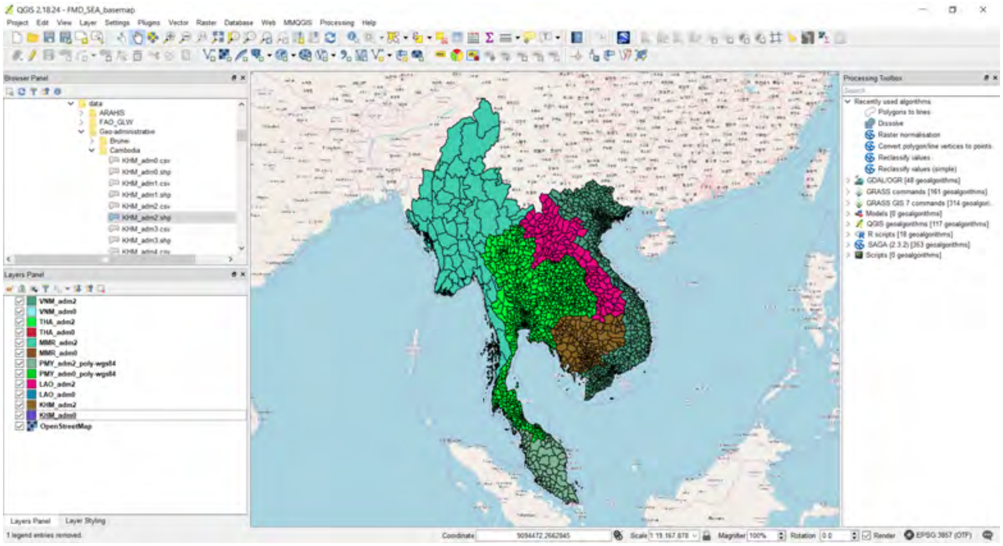
7.5.3 Country boundary data

You can access the country administrative boundaries data shapefiles from the course website. Download these and save them to an appropriate directory on your hard drive. For this basemap, load the country boundaries (adm0) and second-level administrative data (adm2) for Cambodia, Laos, peninsular Malaysia, Myanmar, Thailand and Vietnam.



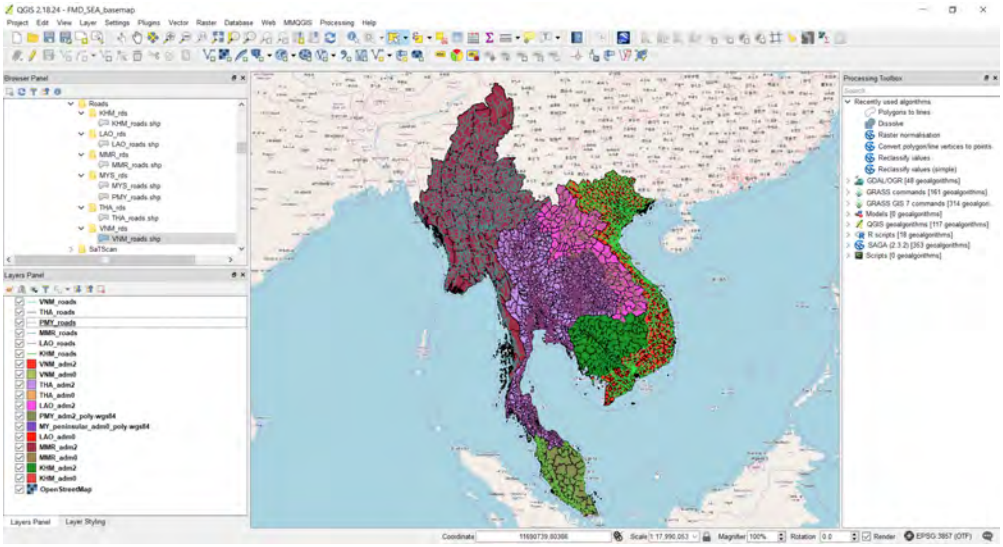
If you're using the Browser Panel, you can simply double-click on the relevant shapefile.

Finally, you should have something like the screenshot below. Note that we have included an OpenLayers background – this is not essential.



7.5.4 Road data

For each of the country, add the roads shapefiles.



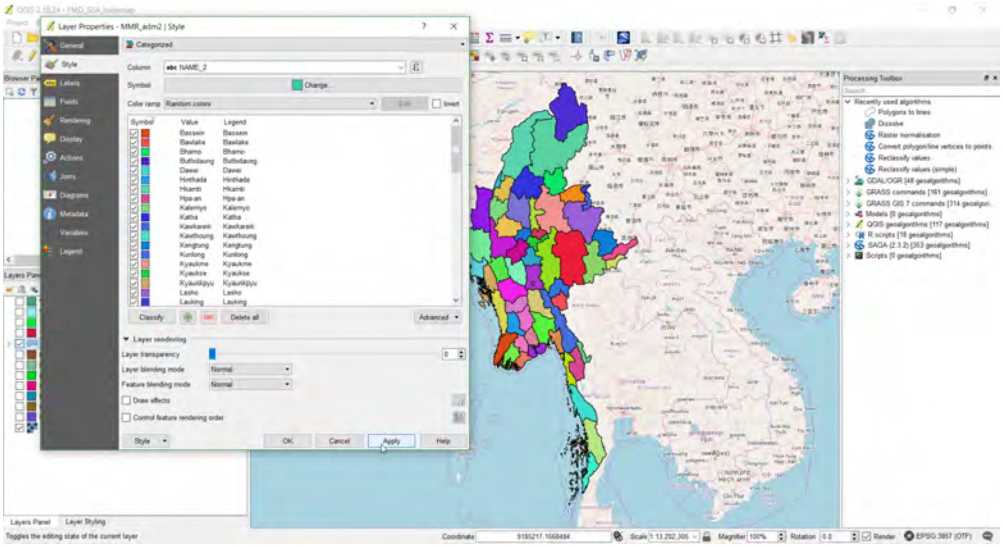
7.5.5 Changing the appearance of your layers

QGIS selects default colour and style attributes for our layers; we will usually want to edit the appearance of these layers so they look ‘useful’. This can be best done by clicking Layer – Properties, or by right-clicking on the layer in the Layers Panel and selecting Properties. The Layer Properties dialog opens. Spend some time re-familiarising yourself with the functionality; we will frequently be using this during the course.

Styling vector polygon layers

For a polygon vector layer:

- The appearance (fill colour and style) can be controlled under the Style tab:
  - We can choose Single symbol (one colour), Categorized (multiple colours for a categorical variable) or Graduated (multiple colours for a continuous variable). We won’t consider the other options currently.
  - Click on Simple fill and experiment with changing the fill colour, line colour, fill style, outline width and transparency. Change the symbol layer type to Outline: Simple line and see what happens.
  - For a given country (e.g. Myanmar), click on Categorized, select `NAME_2` as the column, and click Classify. This will shade the Districts of that country in different colours. You can specify or edit a colour ramp: this will be useful later.
  - Click Apply to view your changes.



Categorized

Column: abc NAME_2

Symbol: Change...

Color ramp: Random colors □ Invert

Symbol	Value	Legend
☑ 🟥	Bassein	Bassein
☑ 🟥	Bawlake	Bawlake
☑ 🟩	Bhamo	Bhamo
☑ 🟦	Buthidaung	Buthidaung
☑ 🟦	Dawei	Dawei
☑ 🟦	Hinthada	Hinthada
☑ 🟩	Hkamti	Hkamti
☑ 🟥	Hpa-an	Hpa-an
☑ 🟩	Kalemyo	Kalemyo
☑ 🟦	Katha	Katha
☑ 🟦	Kawkareik	Kawkareik
☑ 🟦	Kawthoung	Kawthoung
☑ 🟦	Kengtung	Kengtung
☑ 🟦	Kunlong	Kunlong
☑ 🟥	Kyaukme	Kyaukme
☑ 🟩	Kyaukse	Kyaukse
☑ 🟧	Kyaukpyu	Kyaukpyu
☑ 🟪	Lasho	Lasho
☑ 🟦	Lauking	Lauking

Layer rendering

Layer transparency: 0

Layer blending mode: Normal

Feature blending mode: Normal

□ Draw effects

□ Control feature rendering order

- Next, click on Labels. By default, the polygons are not labelled, but we can display these if we choose. Select Show labels for this layer and again choose NAME_2 as the column to label the polygons with. Experiment with changing the font, font size, colour, formatting, placement etc.
- At this stage, we won't consider the other options, but you can see that a lot is possible to fine-tune the display of your map!

Tip! Grouping layers

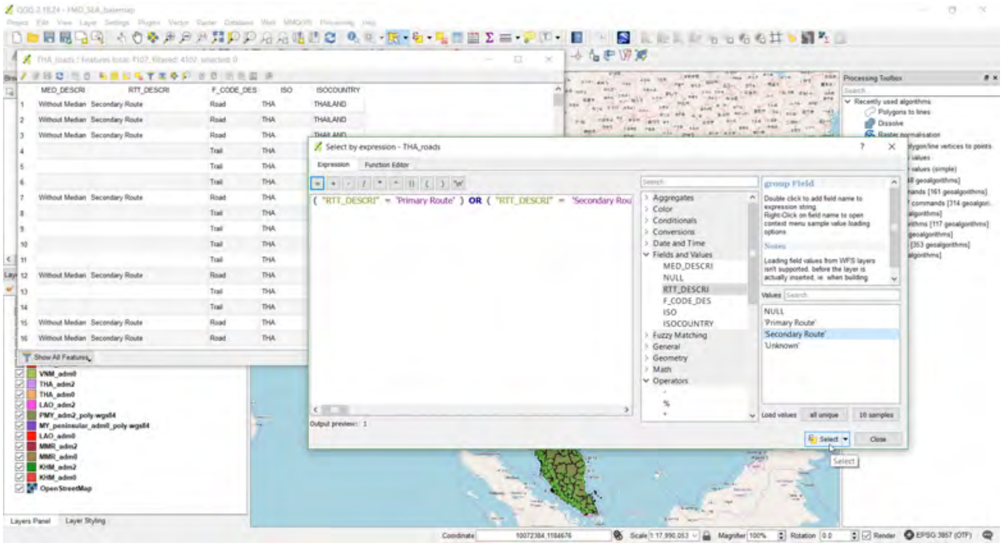
If you are actively working on a project, it's likely that you will end up with many layers. It's handy to group them by selecting the required layers, right-clicking – Group Selected. This group can be renamed (e.g. 'IDN adm layers') and collapsed.

Styling vector polyline layers

Now let's look at point or polyline data, such as the roads we added. Let's filter out all except primary and secondary roads.

- Let's have a look at the different types first. For any of the country roads layers, either click on Layer – Open Attribute Table, or right-click on the layer in the Layers Panel and then – Open Attribute Table.
- The Attribute Table displays the underlying spatial data. Each row represents one polyline. The column `RTT_DESCRI` shows the category of each. We want to filter this so that we only choose `Primary Route` and `Secondary Route`.
- Open the Expression editor:
  - We need to specify our two fields as well as an operator to select the records we're interested in.
  - Under Fields and Values, click `RTT_DESCRI`. To see all the values of this variable, click Load values – all unique.
  - In the Expression editor, open a bracket; double-click `RTT_DESCRI`; click the equals sign; double-click the primary route; and close the bracket.

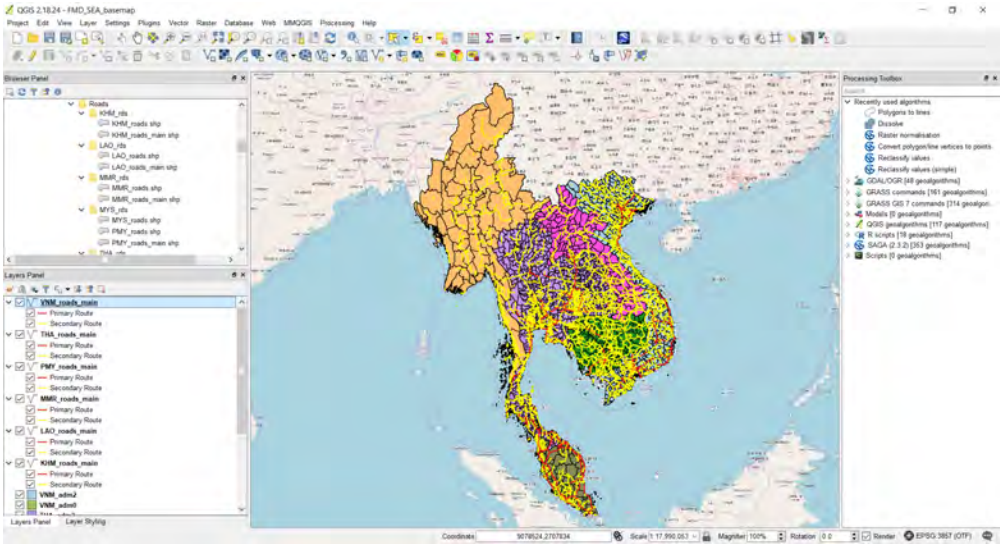
- Under Operators, double-click OR. Repeat the previous steps to specify secondary roads. You should have



Select by expression - THA_roads

```
( "RTT_DESCRI" = 'Primary Route' ) OR ( "RTT_DESCRI" = 'Secondary Rou
```

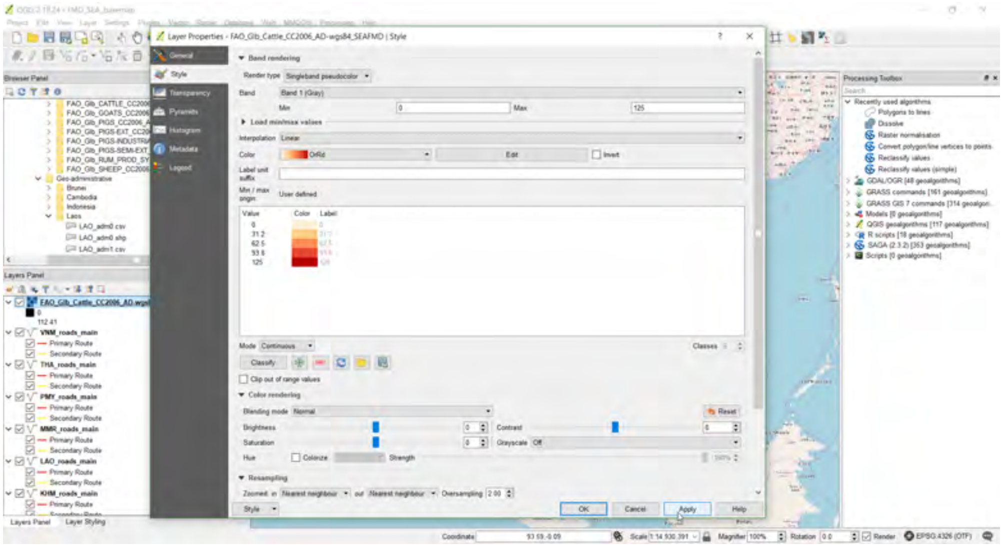
  - Finally, click on Select. You can now close the editor, and the attribute table.
- On the map, we can see our selected roads highlighted in yellow. We want to save this as a new shapefile:
  - Right-click on the layer in the Layers Panel, click on Save As. . . , browse to the folder containing the OSM data and save it with the same filename and the suffix `_main`. Make sure you check the box Save only selected features! Click OK, and the polyline layer will be added to your map view. Now, remove the full roads layer.
- Repeat this for the other five countries.
- We can now edit the appearance of the lines by opening the Layer Properties dialog as before:
  - To categorise our roads, change Single symbol to Categorized and select `RTT_DESCRI` as the category; click Classify.
  - Double-click on each of the road types to edit the appearance. Note that there are a number of options following default map symbology. However, let’s choose the colours red and yellow for primary and secondary roads, and set a line width of 0.5. You should see something like the screenshot below. Note that we can unselect layer categories (e.g. secondary roads) by unchecking the box.



Styling raster layers

Open the raster layer `FAO_Glb_Cattle_CC2006_AD-wgs84_SEAFMD`. The world map has been clipped to an extent covering the six countries. Subsequently:

- Open the Layer Properties dialog in the same way. Notice that the options are slightly different.
- You will have noticed that the raster image is black and white only. The minimum and maximum raster cell values are given. This is not very useful. Remember, the cell values represent continuous data. We want to display these in a series of coloured categories to be able to interpret this map. Therefore, we will need to change the Style settings:
  - Under the Style tab, change the Render type to Singleband pseudocolor. You will see that the minimum and maximum values (that is, the range) of the raster cell values is displayed. To create interval categories, click Classify.
  - By default, in Continuous mode, there are five classes and an equal interval is calculated to determine the categories. You can set the number of categories if you select Equal interval. Also, change the minimum and maximum values and re-classify to give 'neater' category intervals.
  - Also look at the different colour bands; you can edit these, or invert them if you want the colour scale to be reversed.
- The other property we will cover here is under the Transparency tab. Here, we can set the global transparency. This can be very useful when we are displaying our raster map as an 'overlay' of a base map layer; it allows us to simultaneously view underlying geographical map features.



Layer Properties - FAO_Glb_Cattle_CC2006_AD-wgs84_SEAFMD | Style

Band rendering

Render type: Singleband pseudocolor

Band: Band 1 (Gray)

Min: 0 Max: 125

Interpolation: Linear

Color: OrRd □ Invert

Min / max origin: User defined

Value	Color	Label
0	🟧	0
31.2	🟧	31.2
62.5	🟧	62.5
93.8	🟥	93.8
125	🟥	125

Mode: Continuous

□ Clip out of range values

Color rendering

Blending mode: Normal

Brightness: 0 Contrast: 0

Saturation: 0 Grayscale: Off

Hue □ Colorize Strength 100%

Resampling

Zoomed: in Nearest neighbour out Nearest neighbour Oversampling 2.00

Map backgrounds using the OpenLayers plugin

It is often attractive and interesting to have a map background. This is where the OpenLayers plugin is useful. Note that you must be connected to the internet to use it. You may use OpenStreetMap data or Google layers. Click on Web – OpenLayers plugin – OpenStreetMap – OpenStreetMap. Use the zoom icon 🔍 in the toolbar to zoom in on your region of interest.

Note that:

- Every time you change your view or zoom in or out, these spatial data need to be reloaded from the web. If your connection is slow or may drop out, or when you are actively working with different map layers, it is recommended to unselect this layer. You can then reselect it when required, e.g. when preparing the map for printing.
- As this is a base layer, it's recommended to move this layer right to the bottom in your Layers Panel.
- If you use this layer as a basemap, you will probably want to apply a level of transparency to the overlying layers.

7.6 Cartography and map design: generating visual output

Cartography is the study and practice of making maps. It combines science, aesthetics and technique. Maps represent the final outputs of data that have been processed to visualise essential geographical characteristics about the attributes or phenomena we are studying. Simply put, they are powerful graphical tools to convey information. Like all data visualisations, the maps we produce should 'tell a story' that can be interpreted by studying them alone.

There is undoubtedly a creative and aesthetic dimension to developing maps. They can be fun to make, and rewarding: people like maps because they can look very appealing. Also, unlike some of the specialised types of data visualisations we use for disease data, they are intuitive and do not need technical training to understand. Perhaps this explains why decision- and policy-makers, who may not have a technical background, tend to really appreciate maps. Consequently, it's worth investing time in ensuring your maps are as attractive as possible.

However, it's easy to forget that maps are one of our most sophisticated conceptual creations. Care needs to be taken to make sure our maps are clear and interpretable. Map design is an element of cartography which focuses on ensuring that the meaning is effectively conveyed. A map must be designed foremost with consideration to the audience and its needs. Maps are information rich, but displaying too much (or the wrong kind) information makes the message harder to see.

If you regularly produce maps, some of the practices and tips below should become the norm; you will increasingly develop and use templates and develop a personal 'style'.

7.6.1 Visual display of map layers

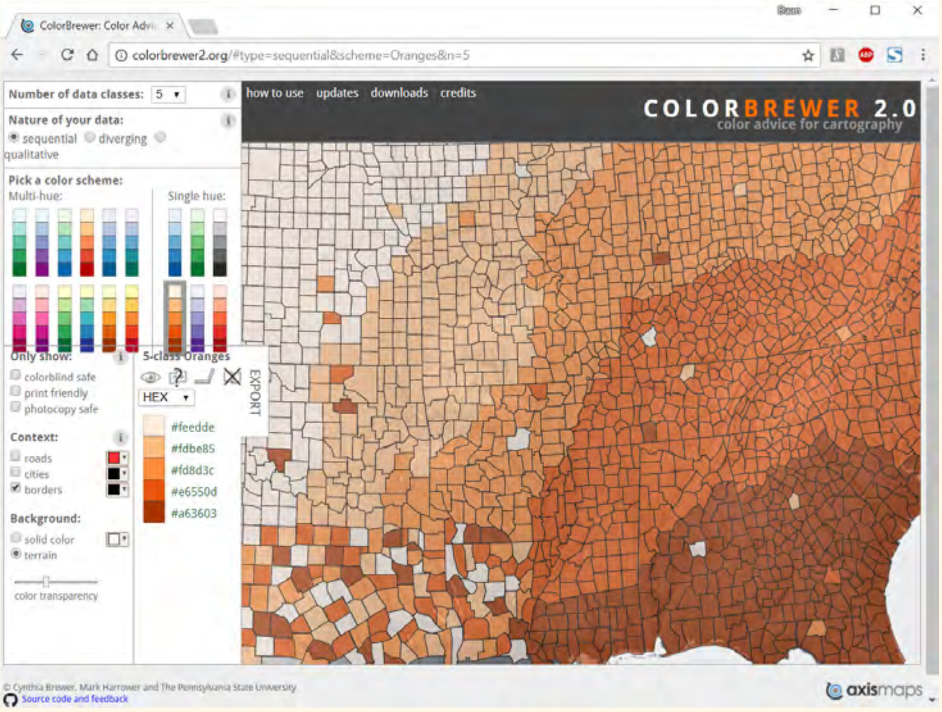
Our maps will show what we display in the main map canvas. Therefore, we edit and adjust the appearance of our map using the same techniques we implemented in 7.5.5 above; when we are satisfied we have finalised this, we can start designing our map, and supplementing it with the symbology discussed below.

Here are some tips and principles for you to consider to result in effective and beautiful maps:

- **Colour scheme.** The choice of colour scheme should relate to the logic in the data you are representing:
    - *Sequential:* used for quantitative ordinal data, i.e. higher values represent 'more' (e.g. poultry density). The convention is that the darker end represents the upper range and the lighter end represents the lower range.
    - *Diverging:* these emphasize the low and high ends of the data range. We use these when we're especially interested in the 'hot' and 'cold' areas, such as in disease (or disease risk) hotspots. A variant of this is the spectral theme used for thermal maps.
    - *Qualitative:* used when there is no specific order in the data (e.g. Provinces of a country).
    - **Colours and contrast.** Choose the number of colour categories wisely: generally, map readers can't tell the difference between more than six or seven levels. Ensure that the colour range (from lightest to darkest) is sufficient to be able to distinguish between the categories.
    - **Opacity and layer order.** When you wish to show the information from two or more layers at the same time, you will need to carefully consider that the overall combination of these layers adequately shows all of the required information in a way that can be differentiated and interpreted. The layer order becomes important. For example, if we want to relate livestock density to populated areas, our raster density map (the upper layer) needs to have a level of transparency so that we can simultaneously see the underlying physical layer. However, if it's too transparent, we are unable to see it properly.
    - **Clipping.** A final consideration relates to clipping (we will cover this in more detail later). Specifically, density maps should be clipped to the borders of the areas from which the data are derived; failure to do this is an artefact.

Tip!

Spend a few minutes playing with the ColorBrewer viewer.



Number of data classes: 5

Nature of your data: ● sequential ○ diverging ○ qualitative

Pick a color scheme:
Multi-hue:
Single hue:

Only show:
□ colorblind safe
□ print friendly
□ photocopy safe

Context:
□ roads
□ cities
☑ borders

Background:
○ solid color
● terrain

color transparency

5-class Oranges

HEX

#feedde
#fdbe85
#fd8d3c
#e6550d
#a63603

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University
Source code and feedback

7.6.2 Map elements: conventions and symbology

Some of the information displayed on maps should be standardised. It is good practice to get into the habit of ensuring these required components are always visible. Here is a checklist of elements you will need to consider all or some of the time.

- □ **Map title.** If the map is intended to stand alone (that is, not incorporated into a figure or a report), a map title is useful to indicate what is being displayed.
- □ **Additional text and graphics.** You may want to include the following, if appropriate:
  - ○ A textbox containing standardised disclaimer text and/or information.
  - ○ Organisational logos.
  - ○ The date of production.
- □ **Map elements.** All of the following should be displayed:
  - ○ A map legend: all the layers displayed on the map (with the exception of the basemap layers) should be explained in the legend. The symbology should be clear and appropriate. The layers in the legend should be ordered by priority, or clustered by theme.
  - ○ A scalebar: this should show sensible distance intervals (not too large, not too small).
  - ○ A north arrow or compass points: there are many types to choose from. This is a matter of personal preference. However, it is important to include this for orientation.
- □ **Other aesthetics and map elements**:
  - ○ If the map shows a relatively small geographic area, including an inset map is helpful for orientation and perspective.
  - ○ Visual hierarchy of layers. Your map layers should be arranged in such a way that

they are logically and sensibly displayed.

○ You may want to include frames around your map and map elements (legend etc.), adjust the opacity, and so forth.

○ All text on your map should be readable. This means that the font / typeface you use must be consistent and readable (Arial or Times New Roman are always safe choices); and the text size must be appropriate.

□ Use a replicable template. Using the Composer Manager (see below), you can develop templates which are pre-populated with these elements. This can save time and effort.

7.6.3 Generating map output using the Print Composer

The Print Composer is what we use to generate and export printable maps showing a specified region, including the additional and conventional map symbology and annotations mentioned above. It is not particularly intuitive or user-friendly, and many people struggle to master it effectively. However, it's worth some time and effort to do so, as it can save much time later on.

Let's continue with the basemap we developed previously, and imagine we want to display cattle density. To create a Print Composer, click on the New Print Composer icon, and enter the name of the composer you are creating. You can create and save multiple print composers for each QGIS project you are working on, or geographic sections within one project.

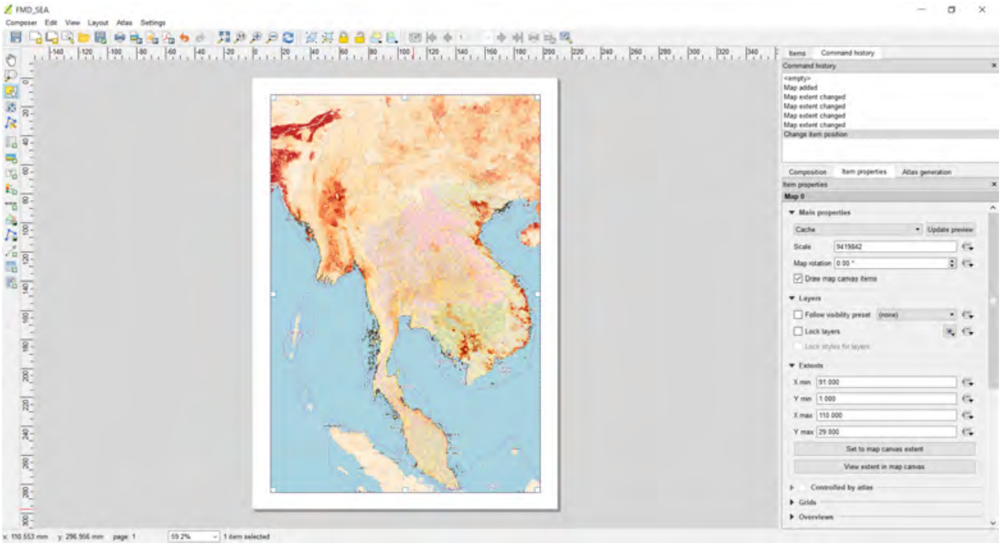
The Composer preview screen now opens. The following buttons provide functionality that you will find yourself using frequently:

Icon	Function
	Select / Move item
	Move item content
	Add new map
	Add image
	Add new label
	Add new legend
	Add new scalebar
	Export as image
	Zoom full
	Refresh view

Follow these steps to develop the Print Composer (don't be afraid to experiment if you are not familiar with this):

- Click on Add new map and drag out a map area. QGIS will try to fill this area with the map extent in you main screen.
- Click on Move item content to reposition the area you want to include in your map.
- Click on Select / move item to adjust your canvas size.
- Other fine-tuning and options for adjustment are given under the Composition and Item properties tabs.
- You can add a legend and a scalebar.

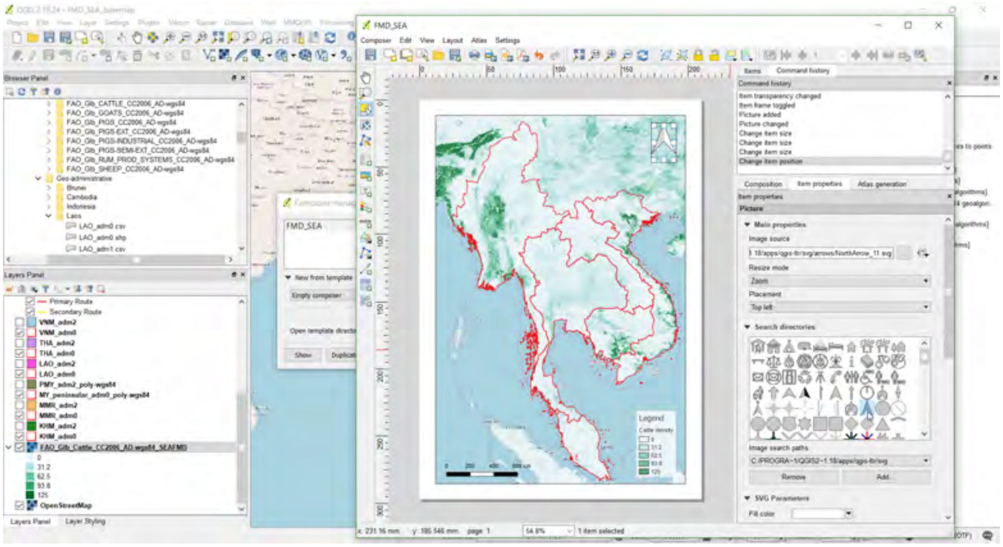
You should get something similar to the following screenshot, to set your map area (note that we've defined map extents which are the same as the raster layer).


FMD_SEA
Composer Edit View Layout Atlas Settings
Items Command history
Composition Item properties Atlas generation
Item properties
Map 0
▼ Main properties
▼ Layers
▼ Extents
Set to map canvas extent
View extent in map canvas
▸ Controlled by atlas
▸ Grids
▸ Overviews
x: 110.553 mm y: 296.956 mm page: 1 59.2% 1 item selected

Command history
<empty>
Map added
Map extent changed
Map extent changed
Map extent changed
Map extent changed
Change item position

Cache
Scale 9419842
Map rotation 0.00 °
☑ Draw map canvas items
☐ Follow visibility preset (none)
☐ Lock layers
Lock styles for layers
X min 91.000
Y min 1.000
X max 110.000
Y max 29.000

Note that when you want to change the presentation of the map (e.g. showing different layers or changing the rendering of the layers), go back to the main QGIS window and make these changes. Then go back to the Print Composer window and click Refresh: the map will now be updated.


QGIS 2.18.24 - FMD_SEA_basemap
Project Edit View Layer Settings Plugins Vector Raster Database Web
Browser Panel
Layers Panel
Layers Panel Layer Styling
FMD_SEA
Composer Edit View Layout Atlas Settings
Items Command history
Composition Item properties Atlas generation
Item properties
Picture
▼ Main properties
▼ Search directories
▼ SVG Parameters
Remove Add
Legend
Cattle density
Composer manager
New from template
Empty composer
Open template directory
Show Duplicate
x: 231.16 mm y: 185.546 mm page: 1 54.8% 1 item selected

FAO_Glb_CATTLE_CC2006_AD-wgs84
FAO_Glb_GOATS_CC2006_AD-wgs84
FAO_Glb_PIGS_CC2006_AD-wgs84
FAO_Glb_PIGS-EXT_CC2006_AD-wgs84
FAO_Glb_PIGS-INDUSTRIAL_CC2006_AD-wgs84
FAO_Glb_PIGS-SEMI-EXT_CC2006_AD-wgs84
FAO_Glb_RUM_PROD_SYSTEMS_CC2006_AD-wgs84
FAO_Glb_SHEEP_CC2006_AD-wgs84
Geo-administrative
Brunei
Cambodia
Indonesia
Laos
LAO_adm0.csv
LAO_adm0.shp
LAO_adm1.csv

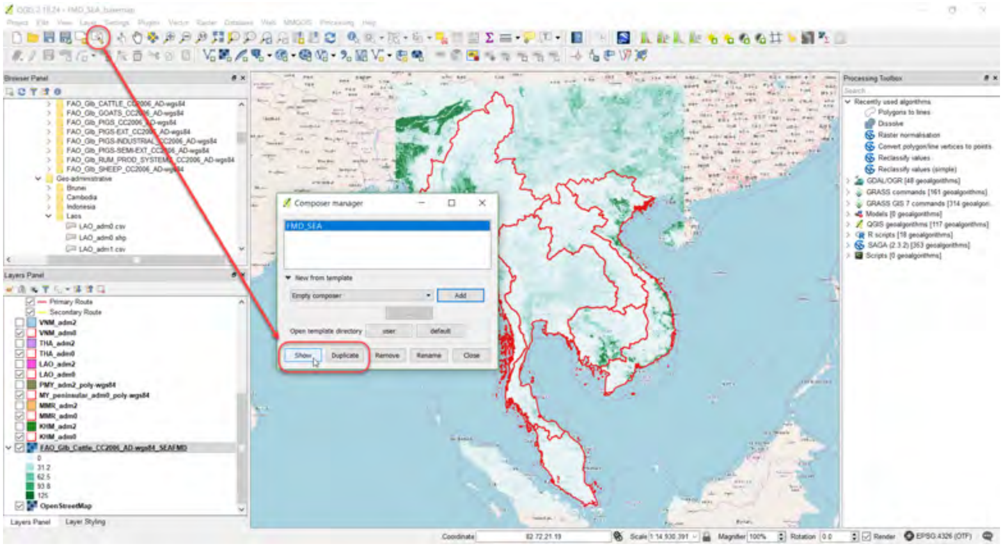
☑ Primary Route
☑ Secondary Route
☐ VNM_adm2
☑ VNM_adm0
☐ THA_adm2
☑ THA_adm0
☐ LAO_adm2
☑ LAO_adm0
☐ PMY_adm2_poly-wgs84
☑ MY_peninsular_adm0_poly-wgs84
☐ MMR_adm2
☑ MMR_adm0
☐ KHM_adm2
☑ KHM_adm0
☑ FAO_Glb_Cattle_CC2006_AD-wgs84_SEAFMD
0
31.2
62.5
93.8
125
☑ OpenStreetMap

Command history
Item transparency changed
Item frame toggled
Picture added
Picture changed
Change item size
Change item size
Change item size
Change item position

Image source
1.18/apps/qgis-ltr/svg/arrows/NorthArrow_11.svg
Resize mode
Zoom
Placement
Top left
Image search paths
C:/PROGRA~1/QGIS2~1.18/apps/qgis-ltr/svg
Fill color

Save your composer when you are happy with it. Finally, use the Export as image button to generate an image file of the map.

The next time you want to make a map of the same area, you can click on the Composer manager icon, and then use the option Show or Duplicate to reopen the map you developed previously.


Browser Panel
Layers Panel
Processing Toolbox
Composer manager
FMD_SEA
New from template
Empty composer
Add
Open template directory
user
default
Show
Duplicate
Remove
Rename
Close

8 Practical 2. Exploratory spatial data analysis (ESDA)

8.1 Objective

In this practical, we will perform some explanatory analysis of FMD outbreak data for the South East Asia region reported to ARAHIS for the period January 2013 – June 2018. This will be followed by spatial visualisation of these data. We will conclude the practical with a discussion on observations and findings.

We will use the whole dataset as the working example in this practical but if you wish to subset the data to analyse this for a specific country only, please do so. If you have additional data to supplement the FMD outbreak data, even better!

This work will lead into the subsequent work on global and local cluster detection.

8.2 Exploratory data analysis (EDA)

Before investigating aspects related to the spatial distribution, it is sensible to explore the data at a higher level. Things we may be interested in include the following:

1. A description of the dataset: what's in it and how complete is it?
2. About the outbreaks: investigation of the distribution of the number of cases reported; calculation of measures of disease frequency (attack rate, mortality rate, case fatality rate).
3. Stratification to assess the variability between the countries.
4. Trends in time: investigation of long term, period and seasonal trends. Some stratification to identify differences in these (e.g. seasonality) can also be done.

This practical can be performed in Excel or in R; we will provide some guidance on both. The practical utilises the file `ARAHIS_FMD_2013-2018.csv` in the data folder. Note that this dataset has already been 'sanitised' for convenience.

8.2.1 Description of the dataset

The data have been extracted from ARAHIS. They contain information on FMD outbreaks reported by six countries in South East Asia (Myanmar, Malaysia, Thailand, Cambodia, Laos and Vietnam) from January 2013 to June 2018. Among other things, the data include

- Unique identifier: `outb_ID`
- Country: `country_code` and `country`

- Outbreak location coordinates: `latitude` and `longitude`
- Administrative units: `province`, `district`, `subdistrict` and `epiunitname`
- Outbreak duration: `startdate`, `resolveddate` and `duration`; `month` and `year`
- Agent details: `diagnosis_type` and `serotype`
- Affected species: `species`
- Epidemiological units of interest: `atrisk`, `cases` and `died`

Partial data on other variables are present, but these are not likely to be useful.

8.2.2 EDA in Excel

Use pivot tables to create summary tables showing the numbers of reported outbreaks:

- per species
- per country
- per year and per month

What additional variables would you consider adding?

Use the pivot tables to make charts showing:

- the distribution of the numbers of reported cases (histograms)
- the numbers of cases reported per month for the entire period
- the numbers of cases reported per calendar month

You may wish to stratify these per country.

You will find that while pivot tables are a very useful feature of Excel, aggregating the data and formatting it in such a way that the charts we are interested in are not necessarily easy (or even possible) to perform. For this reason, it can be preferable to use specialised statistical software such as R to perform this data manipulation and chart generation.

8.2.3 EDA using R

You will find an `R` script file in the data folder which will perform some analyses on the data, and generate some graphics.

The boxplots, barplots and time series plots are shown in Figures 8.1 to 8.4.

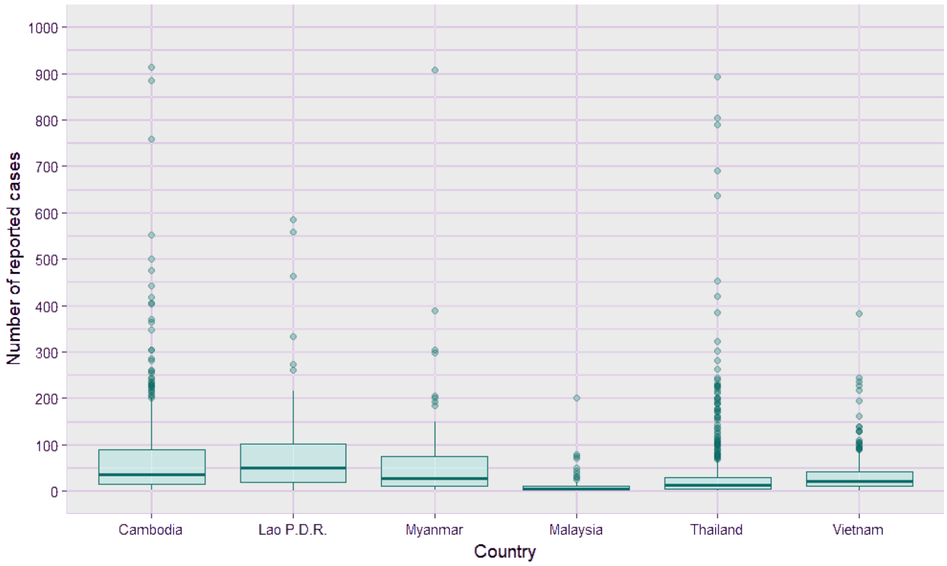


Figure 8.1: Boxplots showing distributions of numbers of cases reported in FMD outbreaks in SE Asia, 2013-2018

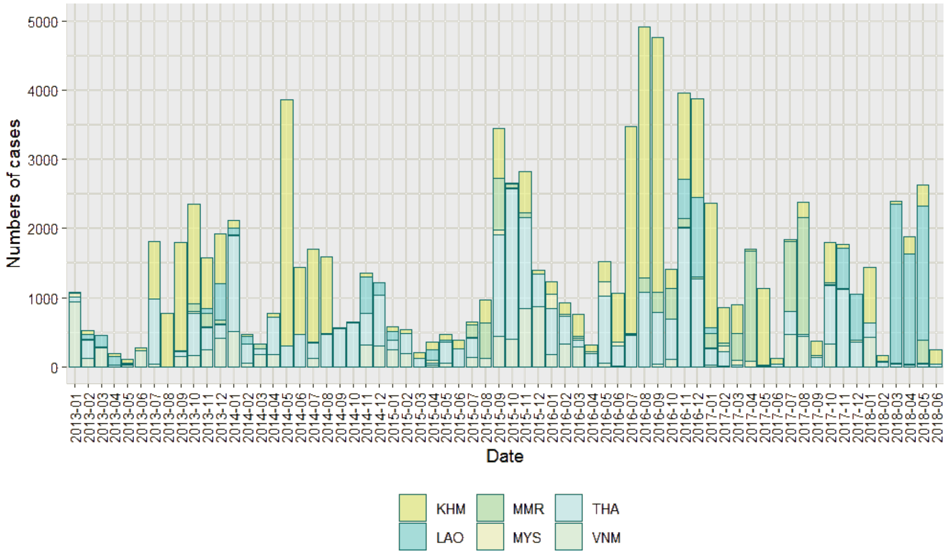


Figure 8.2: Stacked barplot showing monthly numbers of cases reported in FMD outbreaks in SE Asia, 2013-2018

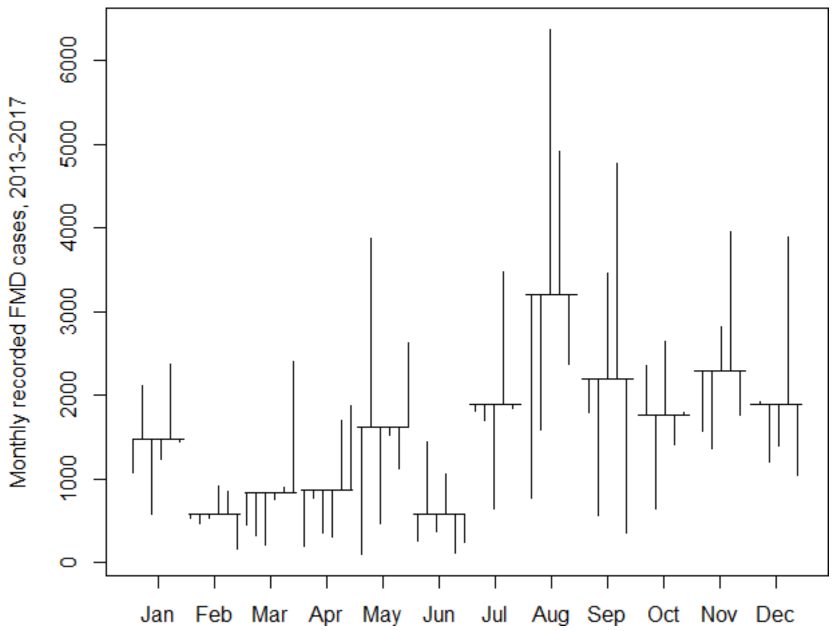


Figure 8.3: Monthplot showing median numbers of cases reported in FMD outbreaks in SE Asia as well as variance per year, 2013-2017

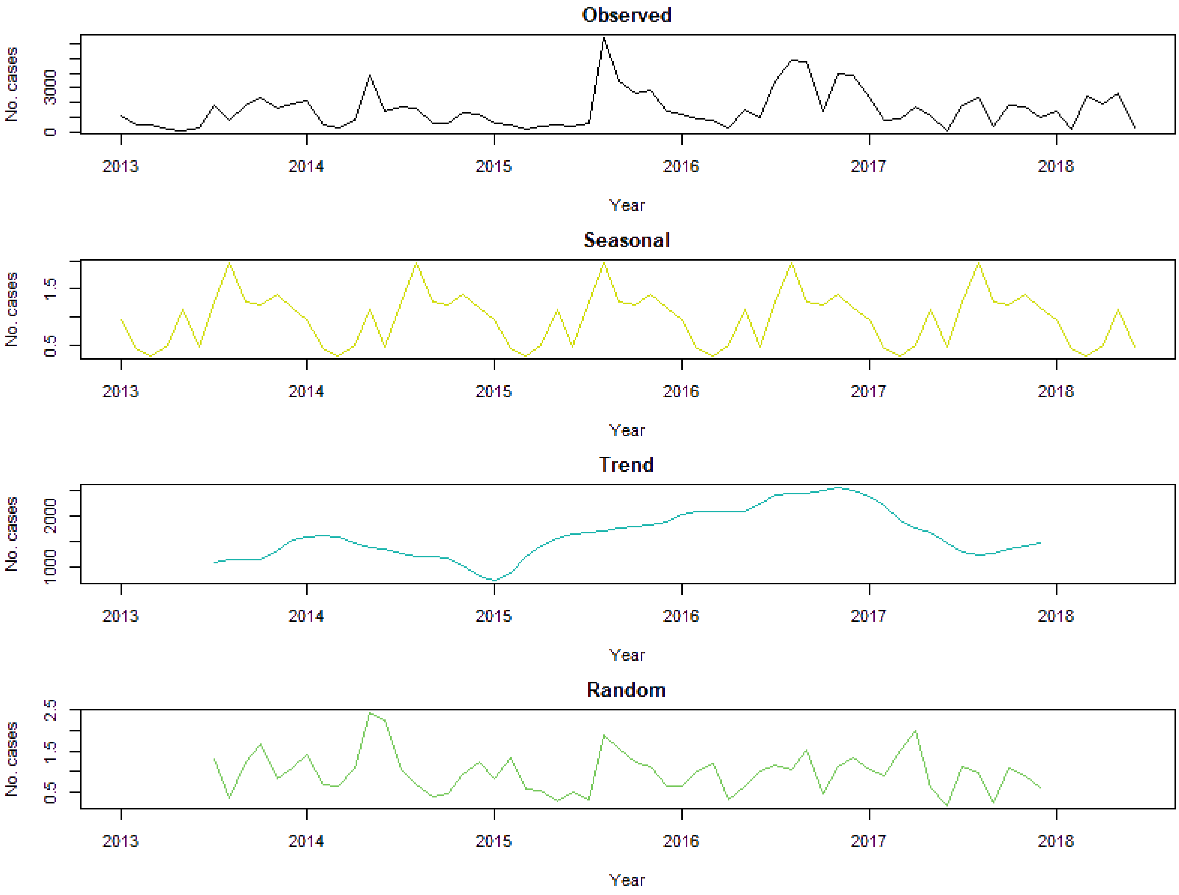


Figure 8.4: Decomposed time series of cases reported in FMD outbreaks in SE Asia, 2013-2017

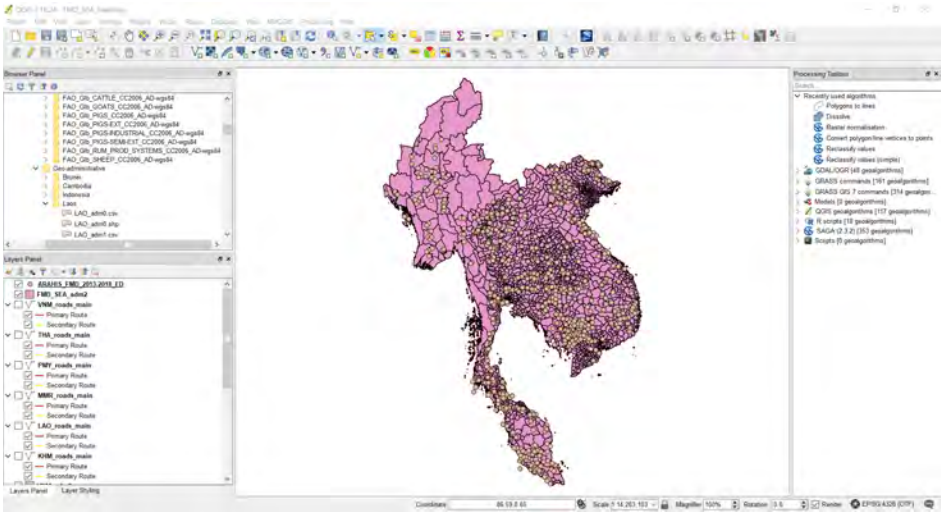
8.3 Exploratory spatial data analysis (ESDA)

In this section, we will investigate spatial aspects of the reported outbreak data to demonstrate how this enriches the EDA. Our starting point for this is the basemap we prepared earlier.

8.3.1 FMD outbreak data, 2011 – current

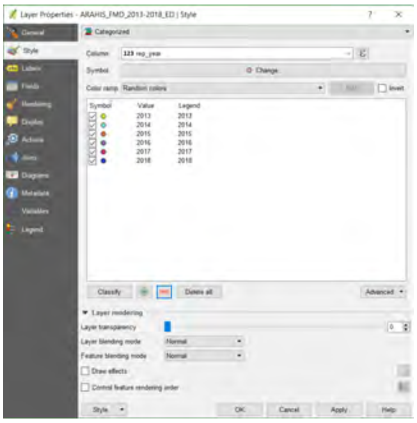
From our data manipulation in R, we exported the consolidated file of FMD outbreaks in large ruminants (cattle plus buffaloes) as `ARAHIS_FMD_2013-2018_ED.csv`. Load this into QGIS, and save it in a relevant directory as a shapefile.

Merge the `adm2` layers of each of the six countries (Vector – Data Management Tools – Merge Vector Layers), and save this as a new shapefile. We will use this layer to assess the numbers of cases at District level over the reporting period.



To provide some visual differentiation, do two things:

- Categorise the reported outbreaks by `rep_year` to stratify by year; this enable us to see if there is any obvious temporal effect.



Layer Properties - ARAHIS_FMD_2013-2018_ED | Style

Categorized

Column: 123 rep_year

Color ramp: Random colors

Symbol	Value	Legend
☑ ●	2013	2013
☑ ●	2014	2014
☑ ●	2015	2015
☑ ●	2016	2016
☑ ●	2017	2017
☑ ●	2018	2018

Layer rendering

Layer transparency: 0

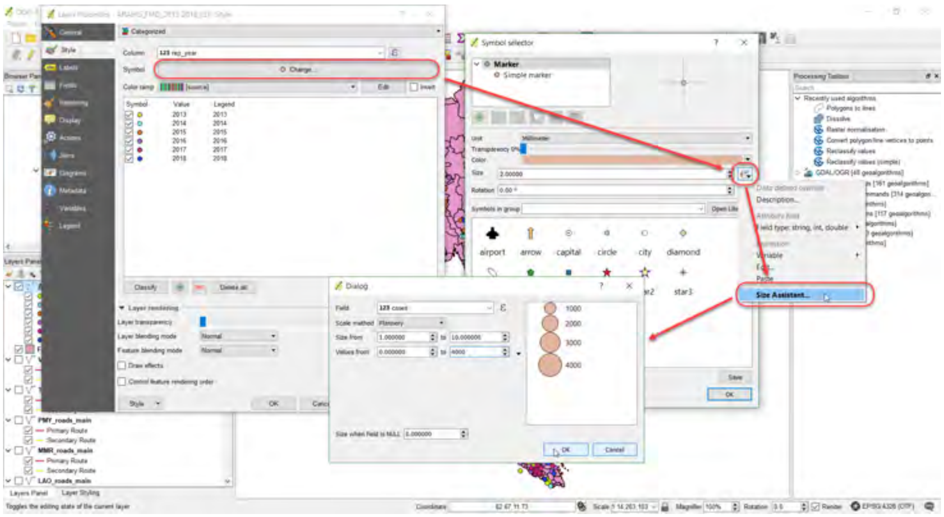
Layer blending mode: Normal

Feature blending mode: Normal

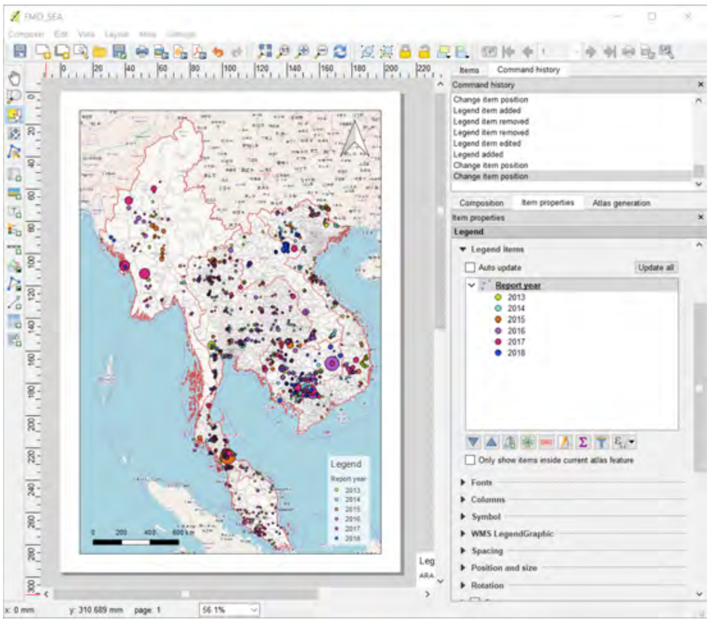
□ Draw effects

□ Control feature rendering order

- Scale the sizes of the points by the numbers of cases per report, by using the Size Assistant.



- With a bit of tweaking, we could generate something similar to the following.

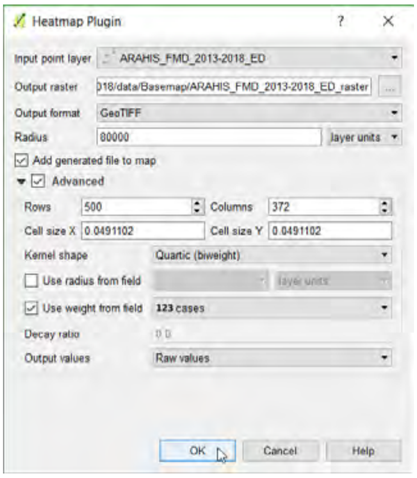


Question

Examine the output and ask yourself these questions:

- Can you see any clear evidence for clustering of outbreaks in space? Are there any marked differences between countries?
- Similarly, is there any evidence to suggest there was clustering in time?

Now click on Raster – Heatmap – Heatmap and select the outbreak points layer as the input. Specify and outputs file, set the radius to 80000 and specify the cases to weight the surface.



Heatmap Plugin

Input point layer: ARAHIS_FMD_2013-2018_ED

Output raster: 018/data/Basemap/ARAHIS_FMD_2013-2018_ED_raster

Output format: GeoTIFF

Radius: 80000 layer units

☑ Add generated file to map

☑ Advanced

Rows: 500 Columns: 372

Cell size X: 0.0491102 Cell size Y: 0.0491102

Kernel shape: Quartic (biweight)

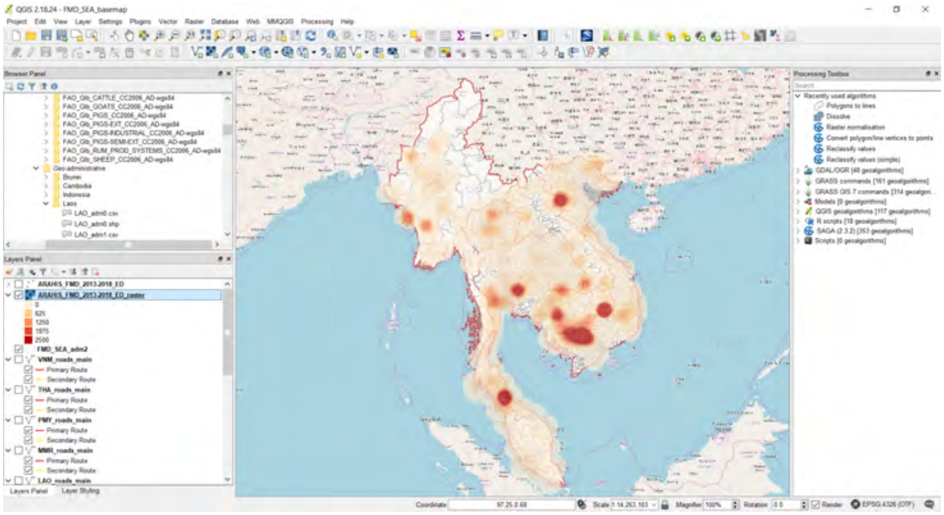
☐ Use radius from field

☑ Use weight from field: 123 cases

Decay ratio: 0.0

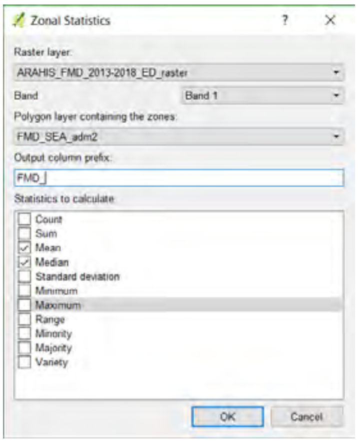
Output values: Raw values

After some manipulation you may get something like the following screenshot.



QGIS 2.18.24 - FMD_SEA_basemap
Project Edit View Layer Settings Plugins Vector Raster Database Web MMQGIS Processing Help
Browser Panel
Layers Panel
ARAHIS_FMD_2013-2018_ED
ARAHIS_FMD_2013-2018_ED_raster
625
1250
1875
2500
FMD_SEA_adm2
Processing Toolbox

Finally, we can look at the median numbers of cases per District. Click on Raster – Zonal statistics to calculate this from the raster layer; compute the mean and median (in the absence of knowledge of the distribution of the case numbers across all Districts, it is safer to use the median as our value of central tendency).



Zonal Statistics

Raster layer: ARAHIS_FMD_2013-2018_ED_raster

Band: Band 1

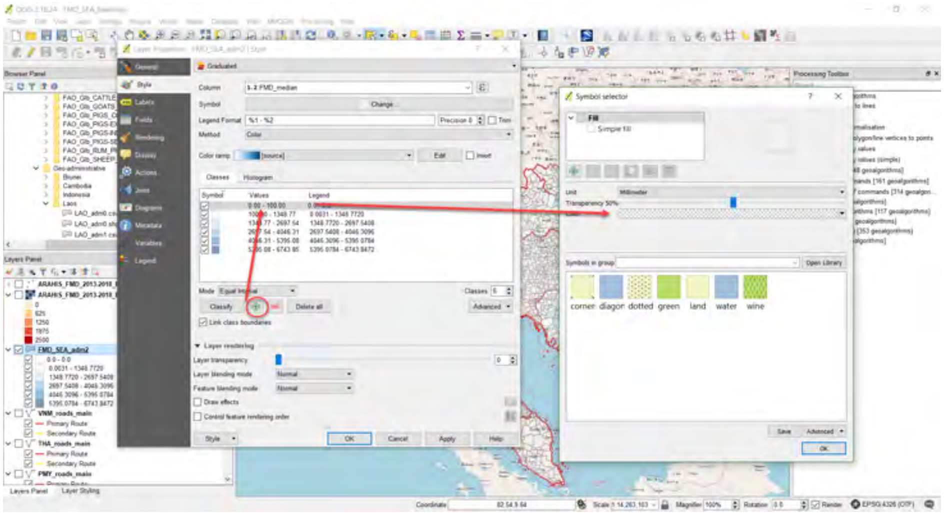
Polygon layer containing the zones: FMD_SEA_adm2

Output column prefix: FMD_

Statistics to calculate

☐ Count
☐ Sum
☑ Mean
☑ Median
☐ Standard deviation
☐ Minimum
☐ Maximum
☐ Range
☐ Minority
☐ Majority
☐ Variety

Now you can create a choropleth map by defining a graduated style for the median values. Add an extra category and set the values for this from zero to a small number (e.g. 100), and set a transparent colour for this category. This way, Districts which did not report any (or a negligibly small number of) cases are not shaded.



Once again, you can refresh the Print Composer to update your map output.

8.3.2 ESDA in preparation for the SaTScan practical

From what we have performed, it is clear that there are a number of reasons why a regional analysis is inappropriate:

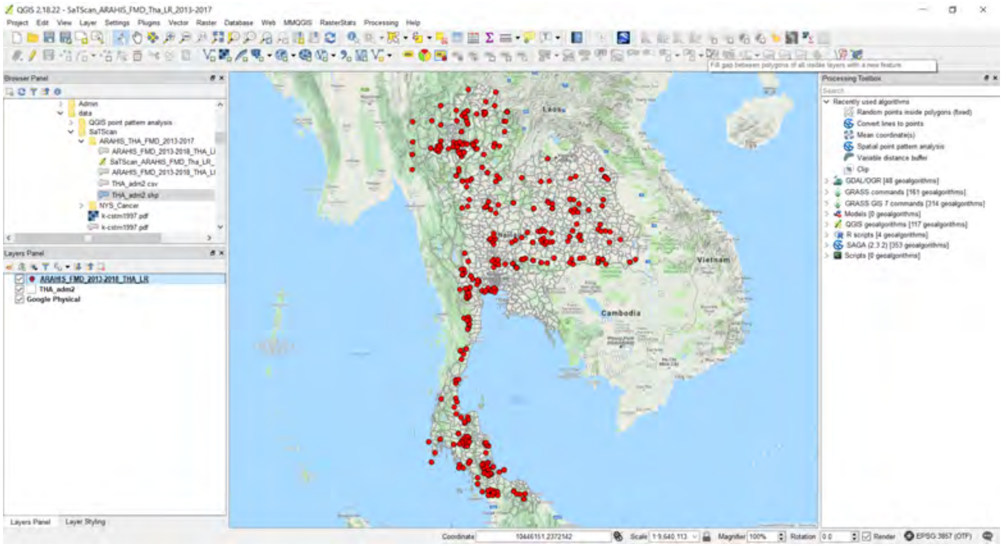
1. There are differences in the sizes of the Districts of the six countries, which represents an example of the modifiable areal unit problem (MAUP)(see chapter 2.8.5).
2. The quality and accuracy of the report data are likely to be variable.
3. Aggregation of data over this time period is likely to miss finer-scale temporal dynamics.
4. The reporting behaviours are likely to differ between countries.

Consequently, detection of clustering is not meaningful for the combined dataset of all six countries. For the next practical, in which we look for localised clusters of infection, we will therefore stratify by country. We will use the Thailand subset of the data.

Here, we can effectively repeat the ESDA we have just performed. This is very helpful because it will enable us to assess whether the outputs from `SaTScan` are consistent and make sense.

Follow this sequence of steps:

- Open the Attribute Table and use `Select Features Using an Expression` to remove all non-Thailand reports. Save this as a shapefile, e.g. (`ARAHIS_FMD_THA_2013-2018_LR.csv`).

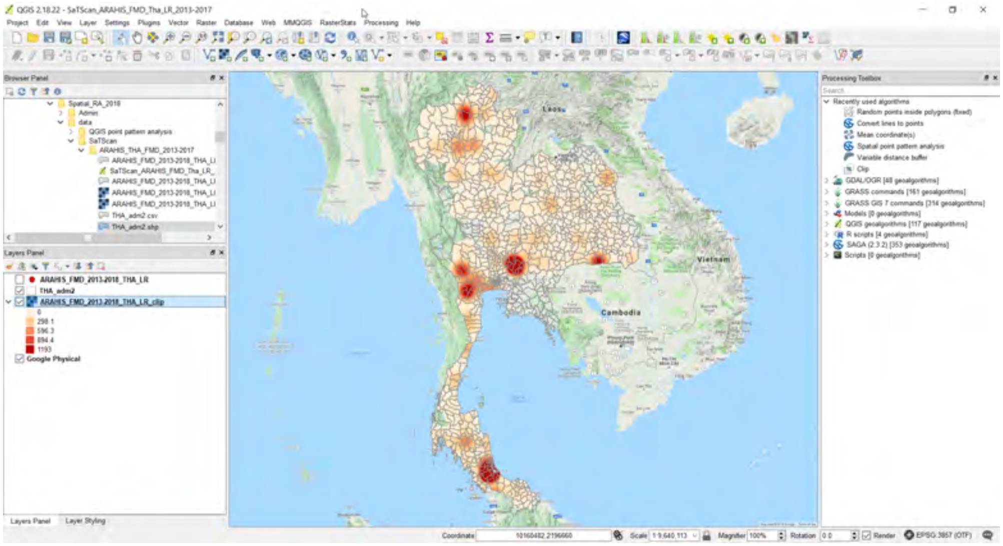


Question

Have a look at the distribution and ask yourself these questions:

- Do you have any comments on the geolocations of the reported outbreaks?
- Is there any evidence (based on the reports alone) of spatial clustering – if so, where do you expect these clusters to be?

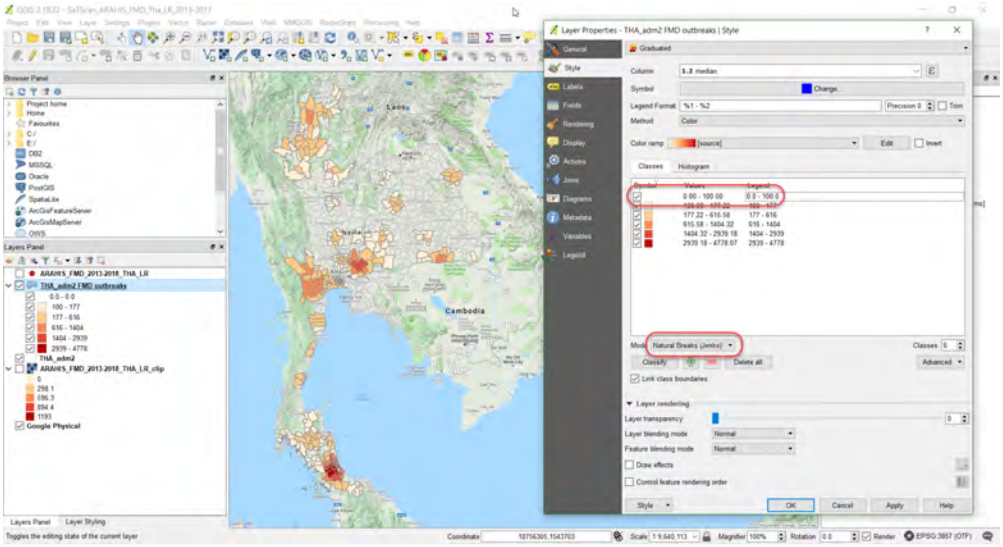
- Next, generate a heatmap of the point locations of the outbreaks, specifying a radius of 50,000. Apply a weighting by the number of cases. Clip the resulting raster layer to the Thailand country border, and display the heatmap something like this:


QGIS 2.18.22 - SaTScan_ARAHIS_FMD_Tha_LR_2013-2017
Browser Panel
Layers Panel
ARAHIS_FMD_2013-2018_THA_LR
THA_adm2
ARAHIS_FMD_2013-2018_THA_LR_clip
0
298.1
596.3
894.4
1193
Google Physical
Processing Toolbox
Laos
Vietnam
Cambodia

Does it look like the outbreaks may be clustered?

For the SaTScan practical, we are interested in the number (or count) of reported cases per outbreak. As the data contain information on the population at risk, they do allow us to calculate relevant measures of frequency such as the attack rate – we could consider using this instead.

- As `SaTScan` identifies clusters based on administrative polygon areas, let's convert the raster density to a district-level expression of FMD incidence over the observation period. Click on `Raster` → `Zonal statistics` → `Zonal statistics`, select the raster layer and extract the median to the `THA_adm2` polygon layer. Then display this as a choropleth map, as we have done previously. There are different ways of doing this but for the visual display, we recommend using Natural Breaks (Jenks) categorisation, adding one more category and setting this from 0 to 100, and making this category transparent (we are effectively 'filtering out' districts with very few or no outbreaks). You should end up with something similar to this:



Layer Properties - THA_adm2 FMD outbreaks | Style

Graduated

Column: 1_2 median

Symbol: Change...

Legend Format: %1 - %2 Precision 0 □ Trim

Method: Color

Color ramp: [source] Edit □ Invert

Symbol	Values	Legend
☑	0.00 - 100.00	0.0 - 100.0
☑	100.00 - 177.22	100 - 177
☑	177.22 - 615.58	177 - 616
☑	615.58 - 1404.32	616 - 1404
☑	1404.32 - 2939.18	1404 - 2939
☑	2939.18 - 4778.07	2939 - 4778

Mode: Natural Breaks (Jenks) Classes: 6

☑ Link class boundaries

Layer rendering

Layer transparency: 0

Layer blending mode: Normal

Feature blending mode: Normal

□ Draw effects

□ Control feature rendering order

Layers Panel

□ ● ARAHIS_FMD_2013-2018_THA_LR

☑ THA_adm2 FMD outbreaks

☑ 0.0 - 0.0

☑ 100 - 177

☑ 177 - 616

☑ 616 - 1404

☑ 1404 - 2939

☑ 2939 - 4778

☑ THA_adm2

□ ARAHIS_FMD_2013-2018_THA_LR_clip

0

298.1

596.3

894.4

1193

☑ Google Physical

Toggles the editing state of the current layer

Coordinate 10756305,1543703 Scale 1:9,640,113 Magnifier 100% Rotation 0.0 ☑ Render EPSG:3857 (OTF)

9 Practical 4. Space-time cluster scanning using SaTScan

9.1 Objective

SaTScan is free software that analyses spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. It is designed to detect spatial or space-time disease clusters, and to determine if these are statistically significant. For an overview of the computational methodology, refer to the resources provided in this course; we won't go into any details here.

In this practical, we will use the purely spatial scan statistic to analyse the geographical distribution of reported Foot-and-Mouth Disease (FMD) cases in Thailand over the period 2013 to 2018, and determine if there are any statistically significant geographical clusters of FMD. In other words, we will determine if there are any geographical areas with more FMD cases than would be expected if the risk of incidence of FMD was evenly distributed across the country. We will be using SaTScan for the computation but we will perform data editing and visualisation in QGIS.

We will conclude the practical by interpreting our results, the requirements in terms of data availability and accuracy, and assessing the usefulness of the technique.

As more advanced exercises, we will assess whether there is evidence of longer-term spatio-temporal clustering (stratification by calendar year) or short-term spatio-temporal clustering (stratification by season). You may tackle these later if you do not have sufficient time during the course.

9.2 Data required

You will find the data in the relevant folder. These include the following files:

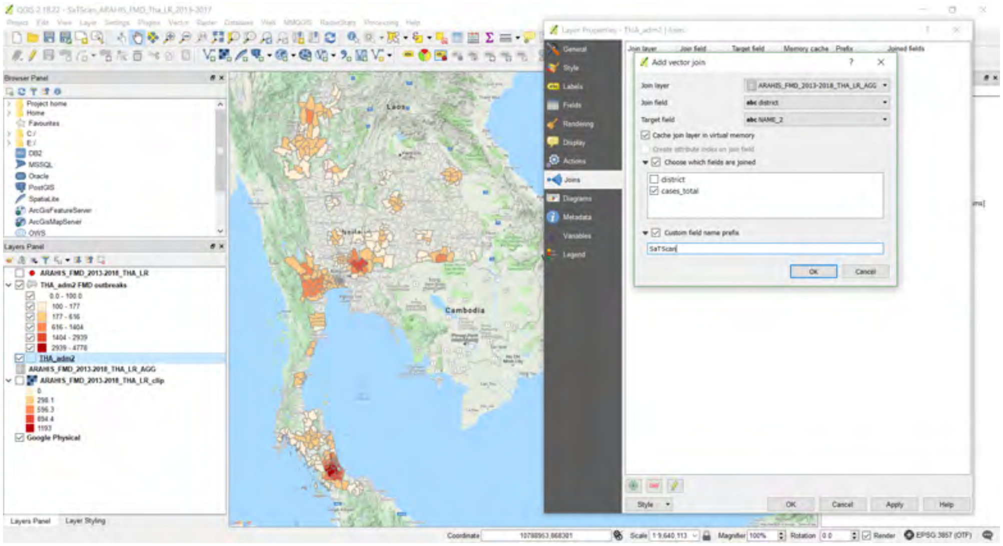
1. ARAHIS extract of FMD outbreaks for the period January 2013 to June 2018.
2. For the purely spatial analysis, these data will be subsetted to Thailand reports only, for the entire period, and aggregated to District.
3. For the spatiotemporal analysis, we will use the subset of Thailand reports for the calendar years 2013 to 2017.

We should have already performed some ESDA on these data – if not, refer to Chapter 8.3.2.

9.3 Preparing to run SaTScan

As with many standalone software packages, SaTScan is fiddly to run and quite a lot of data editing and preparation is required before it will successfully do this. We will do as much of this in QGIS as possible.

- SaTScan requires aggregated data, not point data. From our FMD outbreak data, we know that the District is also recorded (note that this field is separate from the point geolocation given and there may be discrepancies – we will come back to this later). So first, open the ARAHIS_FMD_THA_2013-2018_LR.csv file in Excel. Aggregate the data using pivot tables to find the total cases per District. Copy and paste this to a new workbook and save it as ARAHIS_FMD_THA_2013-2018_LR_AGG.csv or similar.
- Import this file into QGIS as a geometry-only table. Make a join with the THA_adm2 layer to add the total cases for each District; save this as THA_adm2_FMD_cases. If you use the prefix 'SaTScan', the new field name will be SaTScancas.



Add vector join

Join layer: ARAHIS_FMD_2013-2018_THA_LR_AGG

Join field: abc district

Target field: abc NAME_2

☑ Cache join layer in virtual memory

□ Create attribute index on join field

☑ Choose which fields are joined

□ district

☑ cases_total

☑ Custom field name prefix

SaTScan

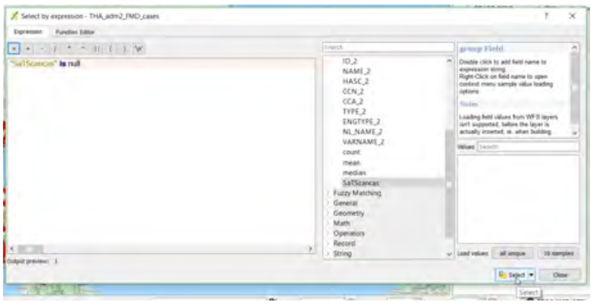
Working with joined data

We often make use of table joins when doing a spatial analysis. This is usually done when we're 'attaching' aggregated disease data to administrative polygons, like we're doing here.

We can, for example, generate choropleth maps this way. However, if you open the Attribute Table and toggle editing, you will notice that you can't make any calculations using the joined data. This is because these data are held in another table!

Overcoming this limitation is simple. If you save the layer which has the joined data as a new shapefile, the joined data become a part of this new layer. You can now perform any calculations.

- To make the join permanent, save the THA_adm2 layer as a new shapefile.
- Notice that many of the Districts are blank as no cases were reported from here. It is best to replace these values with zeroes. Open the attribute table and click the Select Features Using an Expression button. Select the SaTScan case number field (in the Fields and Values list) and type "SaTScancas" is null. Click Select.

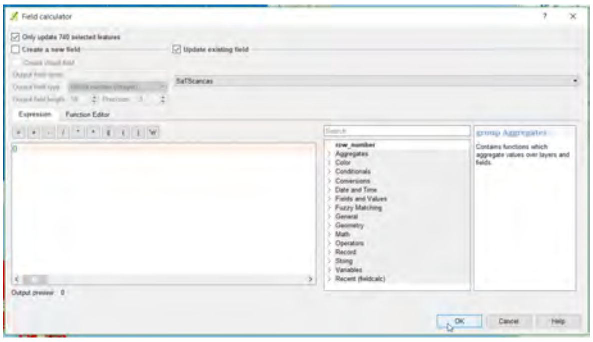


Select by expression - THA_adm2_FMD_cases

"SaTScancas" is null

Output preview: 1

Now go back to the attribute table and open the Field Calculator. Tick the `Update Existing Field` box, making sure you select the `SaTScancas` field. Enter 0 in the expression box and click OK.



Field calculator

☑ Only update 740 selected features

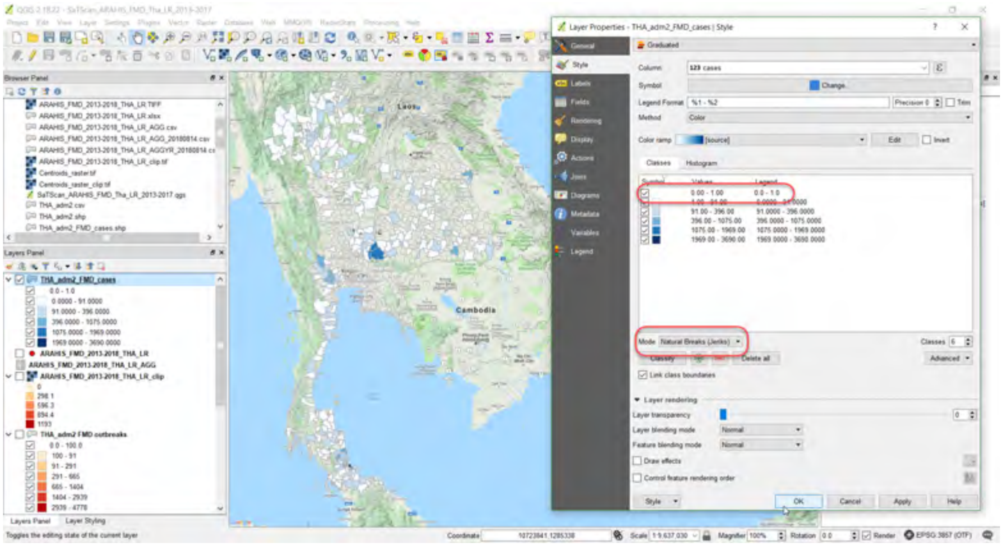
☐ Create a new field

☑ Update existing field: SaTScancas

Expression: 0

Output preview: 0

- Visualise this as a choropleth map, exactly as you did with the aggregated point data above. Comparing the two choropleth maps, do you have any comments? If you think there are clusters, do you think they are the same?



Layer Properties - THA_adm2_FMD_cases | Style

Graduated

Column: 123 cases

Legend Format: %1 - %2

Method: Color

Color ramp: [source]

Values	Legend
0.00 - 1.00	0.0 - 1.0
1.00 - 91.00	0.0000 - 91.0000
91.00 - 396.00	91.0000 - 396.0000
396.00 - 1075.00	396.0000 - 1075.0000
1075.00 - 1969.00	1075.0000 - 1969.0000
1969.00 - 3690.00	1969.0000 - 3690.0000

Mode: Natural Breaks (Jenks)

Classes: 6

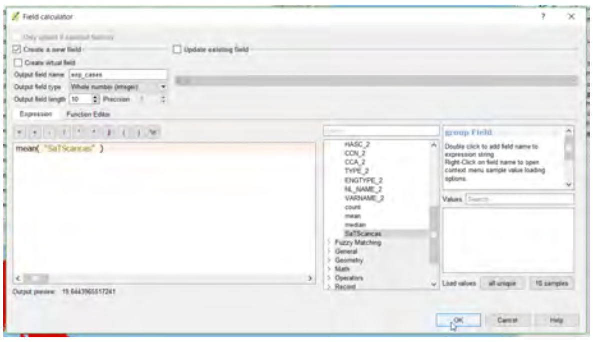
☑ Link class boundaries

Layer transparency: 0

Layer blending mode: Normal

Feature blending mode: Normal

- Before we can run `SaTScan`, two extra pieces of information are required:
  1. **Population data:** the algorithm compares the observed number of cases per areal unit with the expected number, so we need to specify the expected number of cases per District. This is not easy considering we are dealing with outbreak data. Pragmatically, the best approximation is to calculate the mean of reported cases, over the observation period, and add this as a column to our dataset – that is, we expect the average number of cases to occur in all Districts. To do this, in the Attribute Table, open the Field Calculator and add a new field `exp_cases` (or similar). Select `Aggregates` → `mean`, double-click, and then select `Fields and Values` → `SaTScancas`. Click OK to add this number (33).



Field calculator

☑ Create a new field ☐ Update existing field

☐ Create virtual field

Output field name: exp_cases

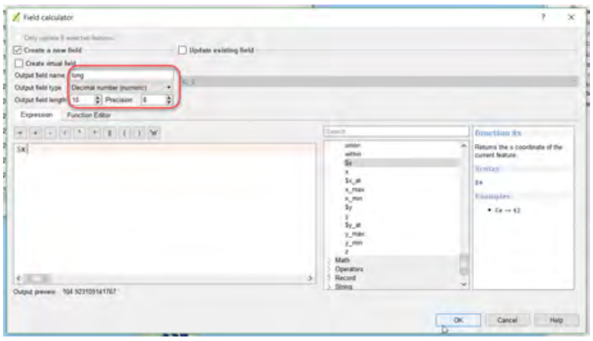
Output field type: Whole number (integer)

Output field length: 10 Precision

Expression: mean("SaTScancas")

Output preview: 19.6443965517241

2. **Location data:** SaTScan requires latitude and longitude data to execute its algorithm. The best approximation is to use the centroid of each THA_adm2 polygon. Use the Vector → Geometry Tools → Polygon centroids tool to generate a temporary Centroids layer. Open the Attribute Table: although this is a points layer, the coordinates are not displayed. To calculate these, make the layer editable (pencil tool), click on the Field Calculator and create a new field long (decimal field, precision 8). Under the Geometry menu, double-click $x and OK.



Field calculator

☑ Create a new field ☐ Update existing field

☐ Create virtual field

Output field name: long

Output field type: Decimal number (numeric)

Output field length: 10 Precision: 8

Expression: $x

Output preview: 104.923109141767

Repeat to calculate the latitude lat, $y. Finally, go to your layer that contains the other data (THA_adm2_FMD_cases) and make a join with the Centroids layer to incorporate the longitude and latitude fields. Once again save this layer as a new shapefile to make the join permanent.

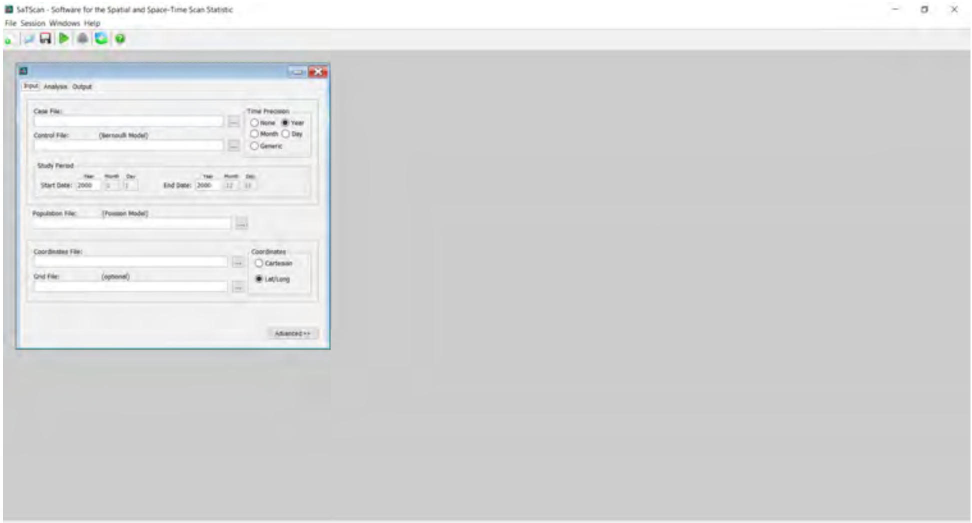
- The last step is to 'clean up' this shapefile by opening the Attribute Table and removing all the fields *except* ID_2, NAME_2, SaTScancas, exp_cases, long and lat. Note that it is the .dbf file in this shapefile that will be used for SaTScan.

9.4 Running SaTScan

9.4.1 SaTScan Software download, installation and launch

You can download the free SaTScan software from http://www.SaTScan.org. In order to obtain the download password, you need to register, providing your name, email address, affiliation and country. The SaTScan software can generate *.kml files which can be opened in Google Earth. However, we will be generate *.shp files which we will open in the QGIS project we created for this practical.

Launch the SaTScan software and select Create New Session from the start window. You should now see the input data tab. Save the session before continuing, e.g. as THA_FMD_2013-2018!



SaTScan - Software for the Spatial and Space-Time Scan Statistic
File Session Windows Help
Input Analysis Output
...
Advanced >>

Case File:

Time Precision: ○ None ● Year ○ Month ○ Day ○ Generic

Control File: (Bernoulli Model)

Study Period

Start Date: Year 2000 Month 1 Day 1 End Date: Year 2000 Month 12 Day 31

Population File: (Poisson Model)

Coordinates File:

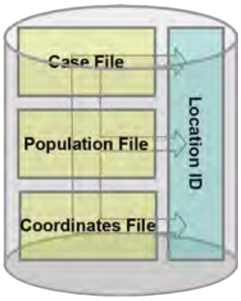
Coordinates: ○ Cartesian ● Lat/Long

Grid File: (optional)

The SaTScan software has three main tabs for specifying the input data, analysis parameters and output formats. Each of these tabs has the option to specify advanced settings.

9.4.2 Input Data tab

The first of the three main tabs is used to specify the input data. We will be using a Poisson (count) model. Three input data fields are required: the reported **FMD incident cases**, the **expected cases** (implemented in SaTScan as the Population) and the **geographical information**. Each of these fields need to have a common identifier: the **location ID**, which represents the name or code for a geographical unit such as province, district, township, village etc.



In our example, all these data are combined in one filed; the work we did above was required to acquire these fields.

Input tab

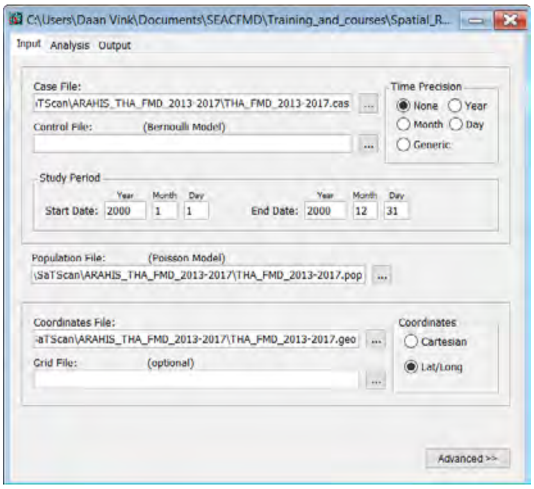
- SaTScan defines its own input file formats. The SaTScan Import Wizard can read several common file formats including *.csv, *.xlsx, *.xls, *.dbf, *.txt and *.shp. We will use the *.dbf file which we generated saving our shapefile in 9.3.
- Once the import file has been selected, you must assign columns to each of the required SaTScan variables. For the case file, we first need to specify the column for the location ID. Click on the word ‘unassigned’ that is to the right of the Location ID variable name, and then select ID_2. The next step is to repeat the same procedure for the number of cases (SaTScancas). We don’t need to assign any other variables.
- Click on Next. You are now asked to specify the file name and the directory in which you want to save the case file. Note that you need to specify the file extension as *.cas to assign it as the case file for the current analysis.
- Finally, click Import. The Time Precision radio button should revert to None.
- As we don’t have control data, we can leave the Control File field blank. Although our observation period is 2013 to 2018, we aren’t stratifying by year, so we also don’t need to specify the Study Period.

Population data

- Repeat the same process (loading the same *.dbf file) to specify the population data. Assign columns to the Location ID and Population variables. The numbers in the population file can be either expected counts, like we have (i.e. exp_cases), or raw population numbers.
- Click Next and save this as the population file, with the file extension *.pop.

Geographical data

- Thirdly, we cannot do the spatial analysis without information about the spatial location of the FMD cases and the expected cases. Using the same location IDs as for the cases and the population, we have to specify the geographical coordinates for each location ID. We have determined the centroids of the Districts as our location IDs.
- Repeat the same process (loading the same *.dbf file) to specify the coordinates file. Assign the location ID as well as the latitude and longitude columns. You need to save this file with the *.geo extension, then import. The geographical locations can either be specified as latitude and longitude (as in our case), or as Cartesian coordinates. When latitude / longitude is used, SaTScan draws concentric circles on the earth's surface, and no map projection is used.



C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_R...

Case File: iTScan\ARAHIS_THA_FMD_2013-2017\THA_FMD_2013-2017.cas

Control File: (Bernoulli Model)

Time Precision: ● None ○ Year ○ Month ○ Day ○ Generic

Study Period

Start Date: Year 2000 Month 1 Day 1

End Date: Year 2000 Month 12 Day 31

Population File: (Poisson Model) \SaTScan\ARAHIS_THA_FMD_2013-2017\THA_FMD_2013-2017.pop

Coordinates File: aTScan\ARAHIS_THA_FMD_2013-2017\THA_FMD_2013-2017.geo

Grid File: (optional)

Coordinates: ○ Cartesian ● Lat/Long

9.4.3 Analysis tab parameter settings

After providing SaTScan with the data needed to run an analysis, you need to tell SaTScan the type of analysis to perform. Click on the Analysis tab.

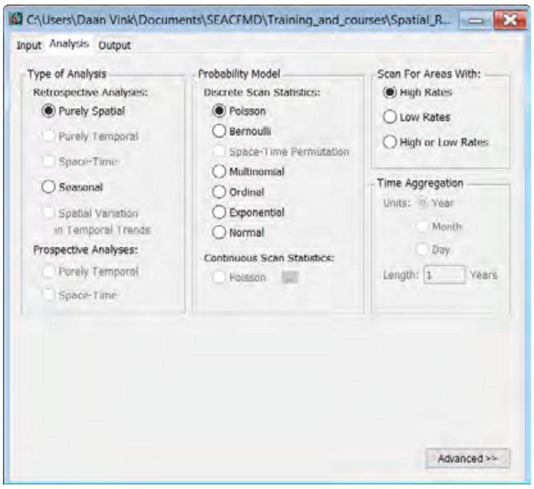
- The first choice is to decide whether to do the analysis using a purely spatial, a purely temporal or a space-time scan statistic. We are not interested in time, so we will do a purely spatial analysis.
- Next, it is necessary to choose the probability model. We have count data from each outbreak, with a background of expected cases. Under the null hypothesis, we assume that the FMD cases are independent of each other. For such data, it is suitable to use the Poisson probability model.
- Scan statistics are typically used to detect clusters of cases; that is, areas with a larger number of cases than would be expected by chance. This indicates areas where there may

be a higher risk for the disease. Sometimes it is also of interest to look for areas with fewer cases than expected, where the risk of the disease is lower; this is not the case here, so select the `High Rates` option.

- The last option on the Analysis tab is for Time Aggregation, but that is only relevant for purely temporal and space-time analyses. Since we are doing a purely spatial analysis, this option is greyed out, and we can ignore it.

Open the Advanced tab to look at some options there:

- In the 'Spatial Windows' tab, the default in `SaTScan` is to look for clusters covering up to half the total expected counts / half the population at risk. You can reduce this proportion if this covers a very large area, resulting in the detection of large clusters. We will keep the default.
- In the 'Inference' tab, look at the number of Monte Carlo replications. This should be at least 999 when running an actual analysis, although it can be set to 0 or 9 for a trial run. A larger number is always better on statistical grounds, as it will increase the statistical power of the analysis. Above 999, the increase in power is very marginal. The drawback of a larger number is that the analysis takes a longer time to run. A good rule of thumb is to use 999 for large data sets that are computer intensive to run, while using 9,999 or 99,999 for smaller data sets that are quick to run. We will retain the default of 999.



C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_R...

Input Analysis Output

Type of Analysis
Retrospective Analyses:
● Purely Spatial
○ Purely Temporal
○ Space-Time
○ Seasonal
○ Spatial Variation in Temporal Trends
Prospective Analyses:
○ Purely Temporal
○ Space-Time

Probability Model
Discrete Scan Statistics:
● Poisson
○ Bernoulli
○ Space-Time Permutation
○ Multinomial
○ Ordinal
○ Exponential
○ Normal
Continuous Scan Statistics:
○ Poisson

Scan For Areas With:
● High Rates
○ Low Rates
○ High or Low Rates

Time Aggregation
Units: ● Year ○ Month ○ Day
Length: 1 Years

Advanced >>

9.4.4 Output tab options

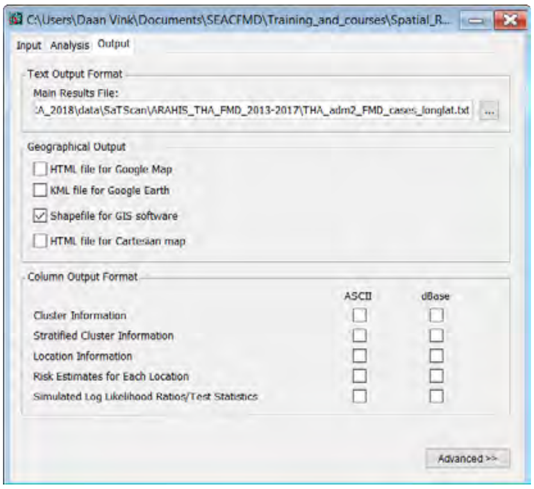
`SaTScan` gives several options to view and save the results of the scan statistic analysis. You will need to make these selections before you execute the `SaTScan` session.

Click on the Output tab:

- There are three sections on the main output tab for text, geographical and column output format.
- Click on the icon to determine the location of `SaTScan` output file, which will be saved as a text file (i.e. extension `*.txt`) and will have several important sections including a summary of the data, location IDs of each location included in each cluster, coordinates and radius of each cluster, population, number of cases, number of expected cases, relative risk and *p*-value for each cluster detected.
- Under `Geographical Output`, tick the option `Shapefile for GIS software`. You can also generate a `*.kml` file if you want to open the output in Google Earth.

- Click on the Advanced tab. `SaTScan` will typically find multiple overlapping clusters, some of which are almost identical to each other. In `SaTScan` there are options to report a different number of these overlapping clusters based on different criteria. For this practical, select `No Geographical Overlap`, which is the most restrictive choice. With this option, a secondary cluster will only be reported if it does not overlap with a more likely and previously reported cluster.

After you have made all the selections on the output tabs, you are ready to run the `SaTScan` analysis.


C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_R...
Input
Analysis
Output
...
Advanced >>

Text Output Format

Main Results File: :A_2018\data\SaTScan\ARAHIS_THA_FMD_2013-2017\THA_adm2_FMD_cases_longlat.txt

Geographical Output

☐ HTML file for Google Map
☐ KML file for Google Earth
☑ Shapefile for GIS software
☐ HTML file for Cartesian map

Column Output Format

	ASCII	dBase
Cluster Information	☐	☐
Stratified Cluster Information	☐	☐
Location Information	☐	☐
Risk Estimates for Each Location	☐	☐
Simulated Log Likelihood Ratios/Test Statistics	☐	☐

9.4.5 Running SaTScan and intepreting the results

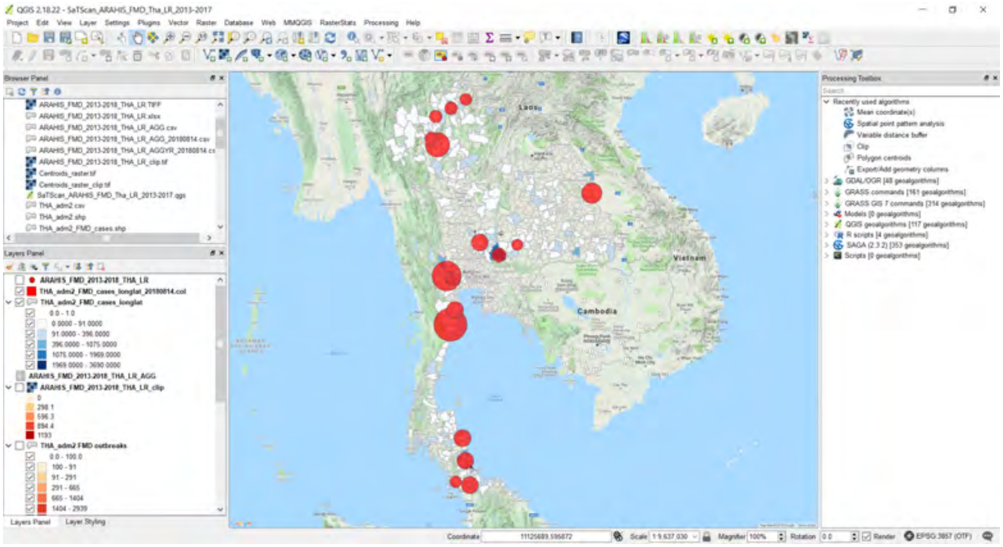
To run the program, from the menu, first select ‘Session’ and then ‘Execute’. Alternatively, and much faster, just click on the green triangle button.

During the `SaTScan` analysis, a dialog shows the progress, as well as warnings or errors. After completing the analysis, this window automatically opens the text-based results file (or you can open the specified `*.txt` file generated in any text editor).

- You can scroll through this window to see all the results, as well as the parameter settings you used. The list of parameter settings also includes the names of the optional results files that you requested, if any. These optional files contain the results in column and row format, for easy incorporation and further analysis using GIS and other software products.
- Check for any warnings or errors. These are designed to help the user find any data problems that may exist in the input files as well as unintended parameter settings. The most common errors are problems with the input files, such as a location ID that is present in the case file but missing in the geographical coordinates file.

After `SaTScan` has completed its computations (hopefully without errors!), check the file directory for the generated shapefile; it will be have an added `.col` suffix.

- Head back to `QGIS` and import this output. The clusters will be shown as concentric circles with variable radius.


QGIS 2.18.22 - SaTScan_ARAHIS_FMD_Tha_LR_2013-2017
Project Edit View Layer Settings Plugins Vector Raster Database Web MMQGIS RasterStats Processing Help
Browser Panel
ARAHIS_FMD_2013-2018_THA_LR.TIFF
ARAHIS_FMD_2013-2018_THA_LR.xlsx
ARAHIS_FMD_2013-2018_THA_LR_AGG.csv
ARAHIS_FMD_2013-2018_THA_LR_AGG_20180814.csv
ARAHIS_FMD_2013-2018_THA_LR_AGGYR_20180814.c
ARAHIS_FMD_2013-2018_THA_LR_clip.tif
Centroids_raster.tif
Centroids_raster_clip.tif
SaTScan_ARAHIS_FMD_Tha_LR_2013-2017.qgs
THA_adm2.csv
THA_adm2.shp
THA_adm2_FMD_cases.shp
Layers Panel
ARAHIS_FMD_2013-2018_THA_LR
THA_adm2_FMD_cases_longlat_20180814.col
THA_adm2_FMD_cases_longlat
0.0 - 1.0
0.0000 - 91.0000
91.0000 - 396.0000
396.0000 - 1075.0000
1075.0000 - 1969.0000
1969.0000 - 3690.0000
ARAHIS_FMD_2013-2018_THA_LR_AGG
ARAHIS_FMD_2013-2018_THA_LR_clip
0
298.1
596.3
894.4
1193
THA_adm2 FMD outbreaks
0.0 - 100.0
100 - 91
91 - 291
291 - 665
665 - 1404
1404 - 2939
Layers Panel
Layer Styling
Laos
Vietnam
Cambodia
Processing Toolbox
Recently used algorithms
Mean coordinate(s)
Spatial point pattern analysis
Variable distance buffer
Clip
Polygon centroids
Export/Add geometry columns
GDAL/OGR
GRASS commands
GRASS GIS 7 commands
Models
QGIS geoalgorithms
R scripts
SAGA (2.3.2)
Scripts
Coordinate
Scale
Magnifier 100%
Rotation 0.0
Render
EPSG:3857 (OTF)

- Open the Attribute Table to inspect the output. Also open the generated `*.txt` file; this contains more detailed information on the clusters.

Questions

- How many clusters were estimated?
- If you sort by the `RADIUS` field, you can see that many of these have a radius of 0 (i.e. they are detected as clusters, but are very small). How many non-zero clusters are there?
- Compare the clusters first with the choropleth map of aggregated numbers of cases we generated in 9.3 and the raster density map we generated in **??**. What are your comments?

9.5 Spatio-temporal analysis

In this practical, we have looked only at a purely spatial clustering over the entire observed time period. Incorporating the temporal trend may be more relevant and informative. We can investigate trends on different time scales:

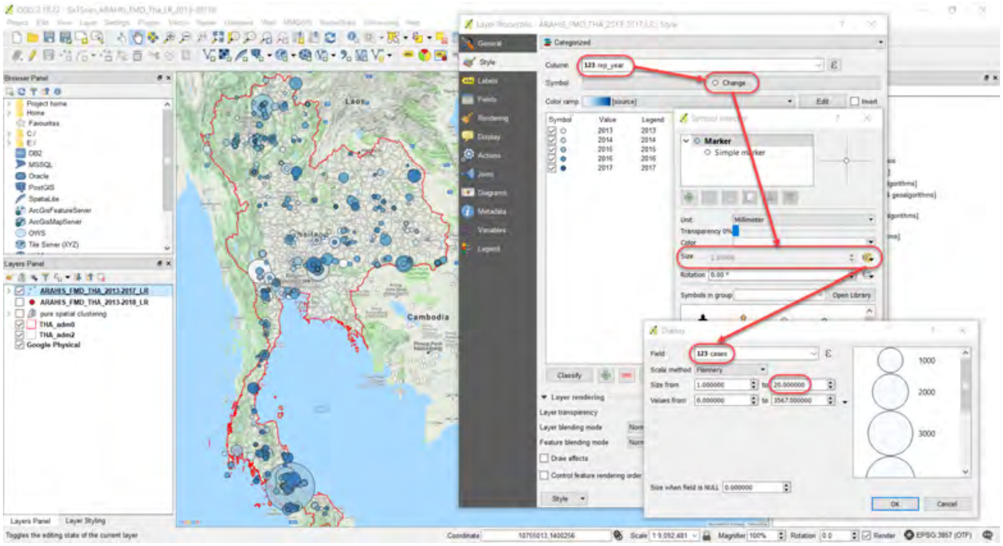
- The period temporal trend (for the period 2013 to 2017, e.g. on an annual basis).
- The seasonal trend, by stratifying per month or by quarter.

This practical will be quicker to carry out than the previous one; the data structure required for `SaTScan` is different, and requires less editing.

9.5.1 Preparing the data

- We will use the same source data (`ARAHIS_FMD_2013-2018_LR.csv`). In **??**, we created a shapefile for the Thai subset – `ARAHIS_FMD_THA_2013-2018_LR.shp`. This is our starting point.
- We can't use the data for 2018, as it is an incomplete year. So, save this layer as a new shapefile `ARAHIS_FMD_THA_2013-2017_LR.shp`, open the Attribute Table and remove all 2018 reports.

- Categorise the points by year and scale the size according to the number of cases, for visual examination. Is there clear evidence of clustering in time as well as space?



Categorized

Column: 123 rep_year

Symbol: ○ Change

Color ramp: [source]

Symbol	Value	Legend
☑ ○	2013	2013
☑ ○	2014	2014
☑ ●	2015	2015
☑ ●	2016	2016
☑ ●	2017	2017

Unit: Millimeter

Transparency 0%

Rotation: 0.00 °

Field: 123 cases

Scale method: Flannery

Size from: 1.000000 to 20.000000

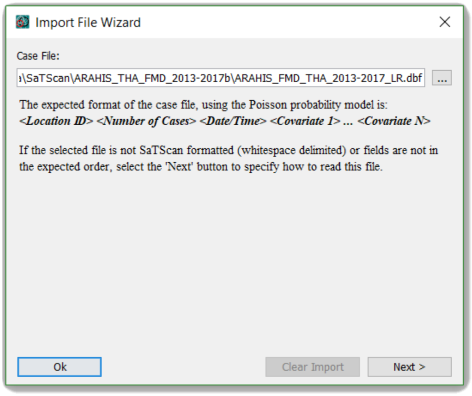
Values from: 0.000000 to 3567.000000

Size when field is NULL: 0.000000

- Note that we can't specify the location ID as District, as we did previously (why?). We will use the ID field in the data as the unique outbreak ID.

9.5.2 Running the model

- Open SaTScan and start a new session. We will load the ARAHIS_FMD_THA_2013-2017_LR.dbf file, as we did previously. For the case data, we now need to specify the time settings. For the time precision, set Day (as the rep_date variable is at this level), and define the reporting period. Screenshots of the settings for the *.cas file are as follows:

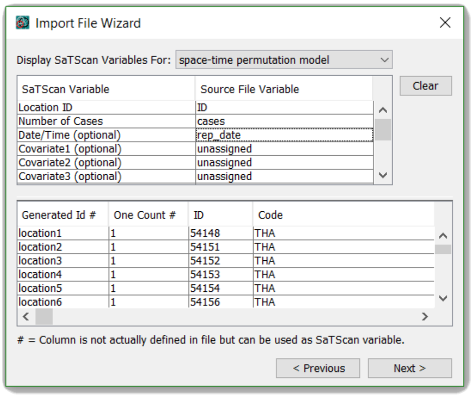


Import File Wizard

Case File: ı\SaTScan\ARAHIS_THA_FMD_2013-2017b\ARAHIS_FMD_THA_2013-2017_LR.dbf

The expected format of the case file, using the Poisson probability model is: *<Location ID> <Number of Cases> <Date/Time> <Covariate 1> ... <Covariate N>*

If the selected file is not SaTScan formatted (whitespace delimited) or fields are not in the expected order, select the 'Next' button to specify how to read this file.



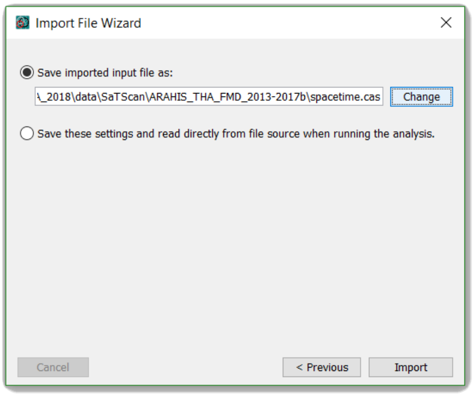
Import File Wizard

Display SaTScan Variables For: space-time permutation model

SaTScan Variable	Source File Variable
Location ID	ID
Number of Cases	cases
Date/Time (optional)	rep_date
Covariate1 (optional)	unassigned
Covariate2 (optional)	unassigned
Covariate3 (optional)	unassigned

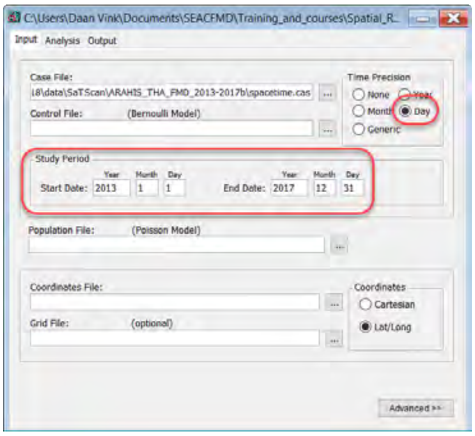
Generated Id #	One Count #	ID	Code
location1	1	54148	THA
location2	1	54151	THA
location3	1	54152	THA
location4	1	54153	THA
location5	1	54154	THA
location6	1	54156	THA

= Column is not actually defined in file but can be used as SaTScan variable.


Import File Wizard

● Save imported input file as: _2018\data\SaTScan\ARAHIS_THA_FMD_2013-2017b\spacetime.cas

○ Save these settings and read directly from file source when running the analysis.


C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_R...

Input Analysis Output

Case File: 18\data\SaTScan\ARAHIS_THA_FMD_2013-2017b\spacetime.cas

Control File: (Bernoulli Model)

Time Precision: ○ None ○ Year ○ Month ● Day ○ Generic

Study Period

	Year	Month	Day		Year	Month	Day
Start Date:	2013	1	1	End Date:	2017	12	31

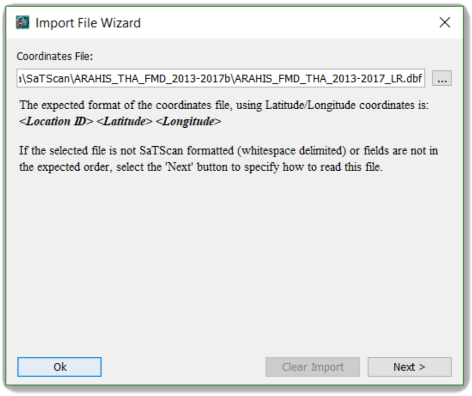
Population File: (Poisson Model)

Coordinates File:

Grid File: (optional)

Coordinates: ○ Cartesian ● Lat/Long

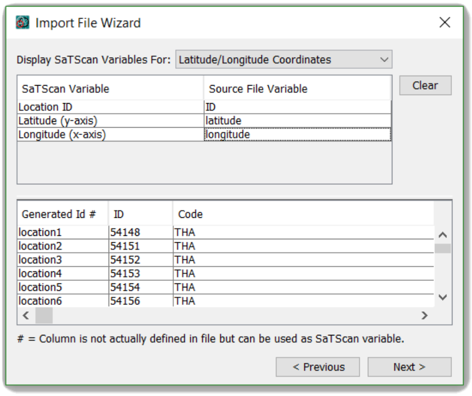
- As this isn't a Poisson model, we don't need to provide counts of the expected cases.
- Specify the coordinates file using the same `*.dbf` input, as before:


Import File Wizard

Coordinates File: \SaTScan\ARAHIS_THA_FMD_2013-2017b\ARAHIS_FMD_THA_2013-2017_LR.dbf

The expected format of the coordinates file, using Latitude/Longitude coordinates is: *<Location ID> <Latitude> <Longitude>*

If the selected file is not SaTScan formatted (whitespace delimited) or fields are not in the expected order, select the 'Next' button to specify how to read this file.

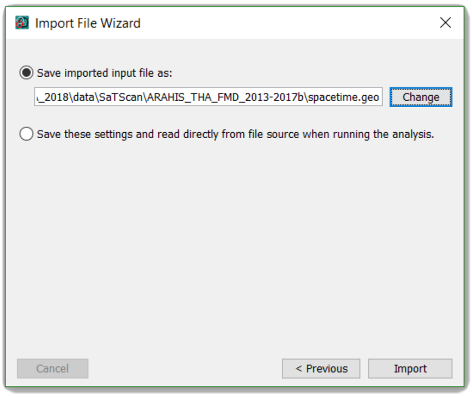

Import File Wizard

Display SaTScan Variables For: Latitude/Longitude Coordinates

SaTScan Variable	Source File Variable
Location ID	ID
Latitude (y-axis)	latitude
Longitude (x-axis)	longitude

Generated Id #	ID	Code
location1	54148	THA
location2	54151	THA
location3	54152	THA
location4	54153	THA
location5	54154	THA
location6	54156	THA

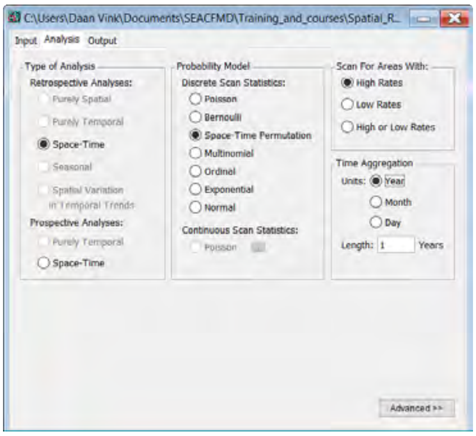
= Column is not actually defined in file but can be used as SaTScan variable.


Import File Wizard

● Save imported input file as: _2018\data\SaTScan\ARAHIS_THA_FMD_2013-2017b\spacetime.geo

○ Save these settings and read directly from file source when running the analysis.

- Now go to the Analysis tab. Select the `Space-Time` → `Space-Time Permutation` radio buttons. Start by specifying a time aggregation of one year. This means that the input data (which is specified as reporting date of the outbreaks) will be aggregated to the calendar year – `SaTScan` will try to identify outbreak clusters by space for each of the years. Finally, in the Output tab, specify the results `.txt` filed and tick the `Shapefile for GIS software` button.



C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_R...

Input | Analysis | Output

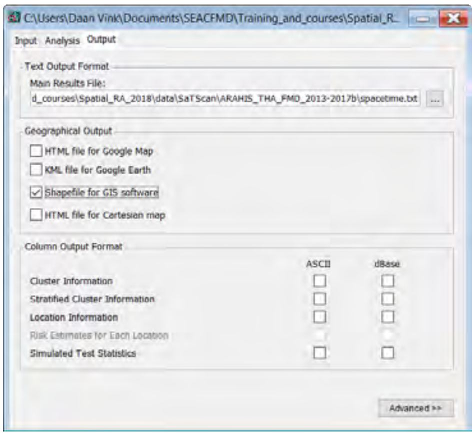
Type of Analysis
Retrospective Analyses:
○ Purely Spatial
○ Purely Temporal
● Space-Time
□ Seasonal
□ Spatial Variation in Temporal Trends
Prospective Analyses:
○ Purely Temporal
○ Space-Time

Probability Model
Discrete Scan Statistics:
○ Poisson
○ Bernoulli
● Space-Time Permutation
○ Multinomial
○ Ordinal
○ Exponential
○ Normal
Continuous Scan Statistics:
○ Poisson

Scan For Areas With:
● High Rates
○ Low Rates
○ High or Low Rates

Time Aggregation
Units: ● Year ○ Month ○ Day
Length: 1 Years

Advanced >>



C:\Users\Daan Vink\Documents\SEACFMD\Training_and_courses\Spatial_R...

Input | Analysis | Output

Text Output Format
Main Results File: d_courses\Spatial_RA_2018\data\SaTScan\ARAHIS_THA_FMD_2013-2017b\spacetime.txt ...

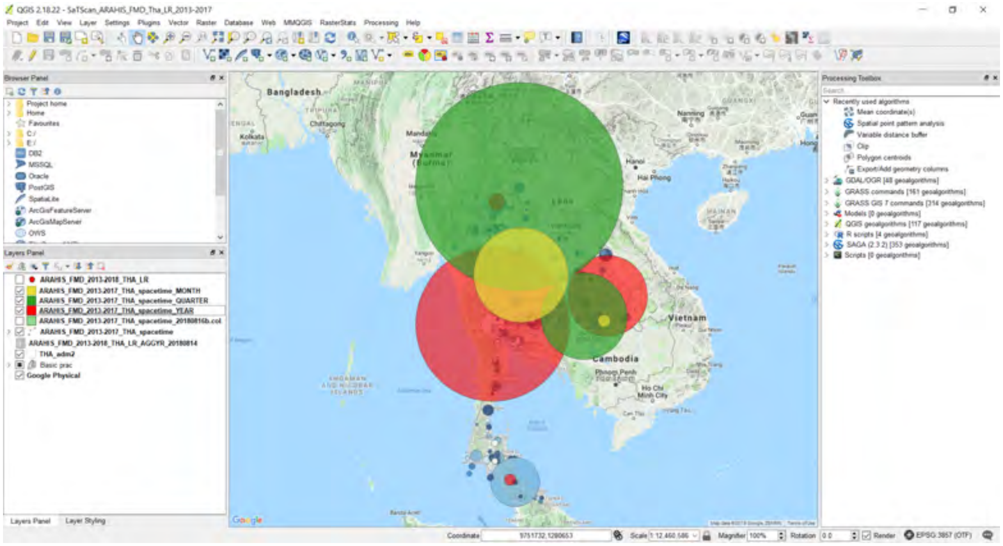
Geographical Output
□ HTML file for Google Map
□ KML file for Google Earth
☑ Shapefile for GIS software
□ HTML file for Cartesian map

Column Output Format

	ASCII	dBase
Cluster Information	□	□
Stratified Cluster Information	□	□
Location Information	□	□
Risk Estimates for Each Location		
Simulated Test Statistics	□	□

Advanced >>

- Now run the model; as it is aggregating to year, it should run very quickly. Go to QGIS and open the shapefile. Also open the attribute table. In which years were clusters detected? Is there any consistency?
- Let's look if there are spatiotemporal trends on more granular time scales:
  - Going back to SaTScan, reset the time aggregation to 3 months. This will scan for clusters per quarter.
  - Repeat this for an aggregation to month. Note that the shorter the aggregation period, the longer the model will take to run, as it has more statistics to compute!



QGIS 2.18.22 - SaTScan_ARAHIS_FMD_Tha_LR_2013-2017

Layers Panel
□ ● ARAHIS_FMD_2013-2018_THA_LR
☑ 🟨 ARAHIS_FMD_2013-2017_THA_spacetime_MONTH
☑ 🟩 ARAHIS_FMD_2013-2017_THA_spacetime_QUARTER
☑ 🟥 ARAHIS_FMD_2013-2017_THA_spacetime_YEAR
□ 🟩 ARAHIS_FMD_2013-2017_THA_spacetime_20180816b.col
☑ ARAHIS_FMD_2013-2017_THA_spacetime
ARAHIS_FMD_2013-2018_THA_LR_AGGYR_20180814
☑ THA_adm2
■ Basic prac
☑ Google Physical

Questions

- Interpret the results. Are they consistent – are the cluster months identified and the quarter and year results? Are they spatially consistent?
- Note that these results are from the most restrictive spatial setting (Output → Advanced → No Geographical Overlap); you may want to experiment with what happens when you select progressively less restrictive settings!

10 Practical 5a. Spatial risk assessment using QGIS

10.1 Objective

In this practical, we will demonstrate how to implement knowledge-driven spatial risk assessment to generate a risk surface for FMD in cattle in Malaysia. Refer to Chapter 6 for background information.

To save time, we will use the outputs of an MCDA exercise performed previously to inform the multipliers required for our Weighted Linear Combination (WLC). We will implement this WLC using different spatial layers. Ultimately, the objective is to identify potential high-risk areas for FMD – disease 'hot spots'.

The MCDA exercise was performed to elicit information on the relative influence of each of these factors on the occurrence of FMD outbreaks. This is knowledge-driven: it relies on expert opinion, both for the selection of the spatial risk factors as well as the estimation of their relative importance. If we had appropriate data and evidence from literature, a more quantitative (data-driven) approach would be preferable.

From the MCDA exercise, the following spatial risk factors were identified:

- Direct contacts of livestock during grazing.
- Livestock movements following legal trade.
- Illegal or unregulated livestock movements.
- Contact of ruminants with pigs.
- Contact of livestock with feral or wild-living pigs.

10.2 Spatial layers and data

A summary of the spatial data sets required for these analyses is provided in Table 10.1. We have made various assumptions to perform this analysis; this exercise is crude, but demonstrates a methodology which could be refined.

Workflow to generate FMD spatial risk maps

This exercise follows a sequence of steps which facilitates the analysis and avoids pitfalls:

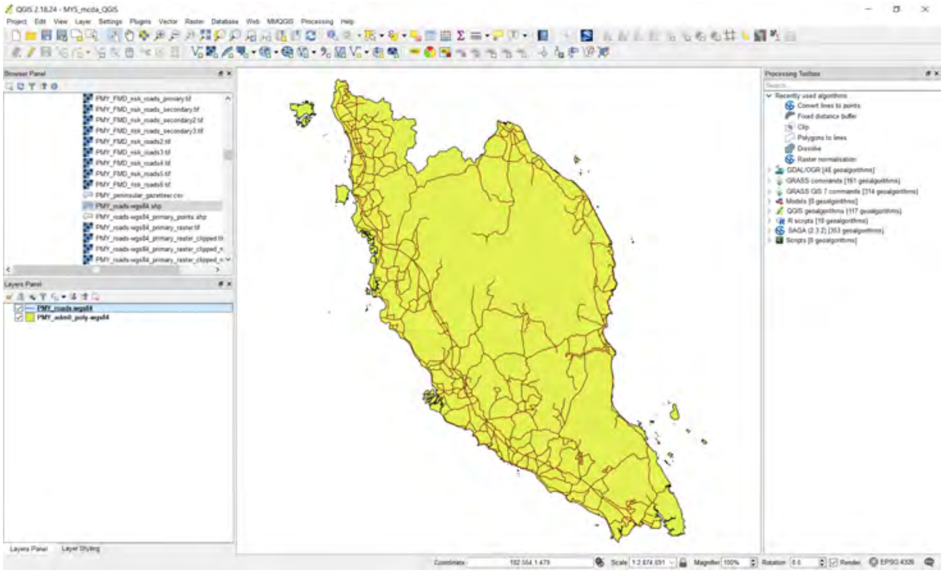
1. Identify the spatial layers (from the MCDA process).
2. Load and prepare these layers.
   - (a) Raster layers: clip to the country boundary to eliminate edge effects.
   - (b) Vector layers: convert to raster layers, clipping to the country boundary.
3. Deal with null values and zeroes in the rasters:
   - (a) Use the `Raster – Extraction – Clipper` to clip the layer to standardised extents that are outside of the country boundary;
   - (b) Reclassify null values with zeros (in the `Processing Toolbox`, `SAGA – Raster Tools – Reclassify values`).
   - (c) Finally, re-clip to the country boundary.
4. Normalise and align (i.e. specify same extents and numbers of rows and columns) the rasters.
5. Perform the WLC calculation, and generate map output.

Table 10.1: Names and file formats for each of the spatial data layers required for an FMD incursion spatial risk assessment, Peninsular Malaysia.

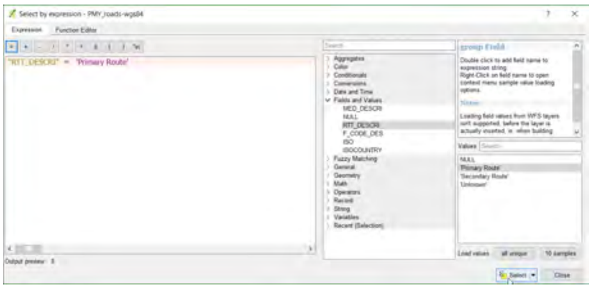
Layer	Format	Details
`PMY_adm0_poly-wgs84.shp`	SHP	Boundary map of Peninsular Malaysia.
`PMY_cmovt_district_2015.csv`	CSV	Number of cattle moved into each district, 2015.
`PMY_roads-wgs84.shp`	SHP	Peninsular Malaysia roads.
`PgIntDn_8k_201507081.tif`	TIF	Semi-intensive pigs – FAO Gridded Livestock of the World.
`PgExtDn_8k_201507081.tif`	TIF	Extensive pigs – FAO Gridded Livestock of the World.

10.2.1 Loading map layers and preparation

Import the boundary file `PMY_adm0_poly-wgs84` into QGIS. Also import the roads polyline layer, `PMY_roads-wgs84`.



As we assume that legal animal movements take place on main (primary) roads and illegal animal movements take place on minor (secondary) roads, use the expression editor to create two separate layers for primary and secondary roads, as we have performed previously.



Select by expression - PMY_roads-wgs84

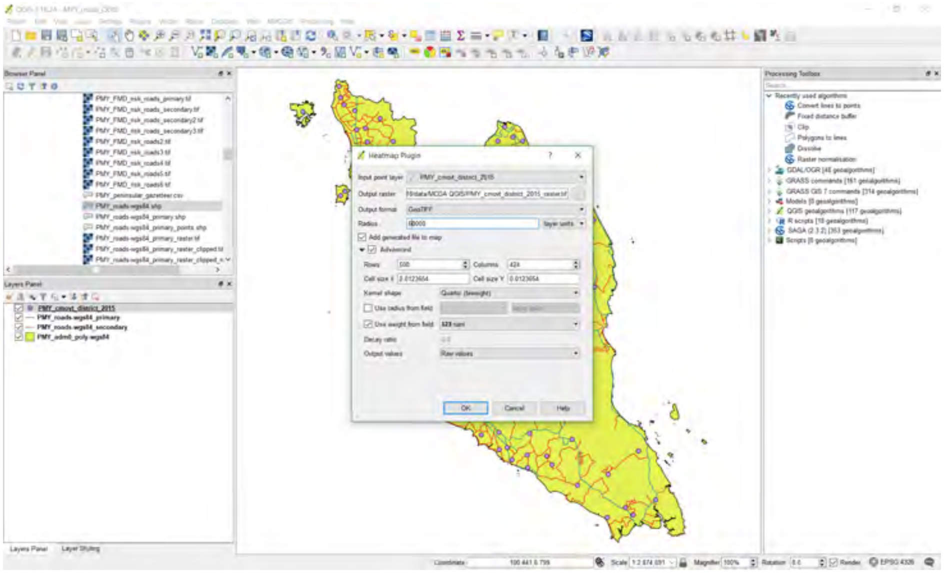
"RTT_DESCRI" = 'Primary Route'

Output preview: 8

Also, load the datafile `PMY_cmovt_district_2015.csv`. You will notice that the point location reflects the centroid of Districts (you can check this by loading the `adm2` file). If you look at the data, notice that it contains information on the numbers of consignments and the numbers of animals. We will use these data as a proxy for livestock contacts.

10.2.2 Animal contacts (grazing practices)

Generate a raster heatmap of the cattle movement data layer, specifying a radius of 60,000 and weighting the points (Districts) by the numbers of animals moved. Then clip the raster to the `adm0` boundary.



Heatmap Plugin

Input point layer: PMY_cmovt_district_2015

Output raster: ...\data\MCDA QGIS\PMY_cmovt_district_2015_raster.tif

Output format: GeoTIFF

Radius: 60000 layer units

☑ Add generated file to map

☑ Advanced

Rows: 508 Columns: 424

Cell size X: 0.0123654 Cell size Y: 0.0123654

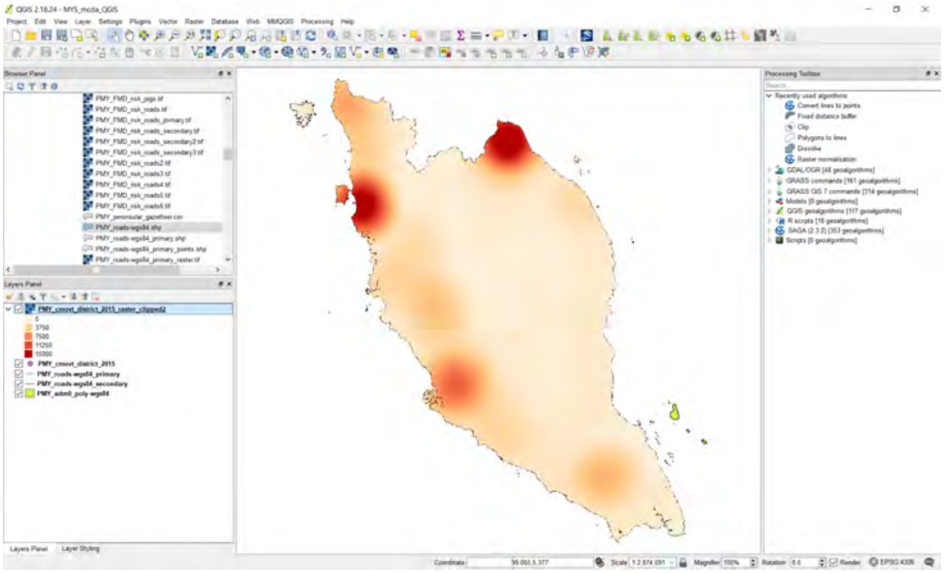
Kernel shape: Quartic (biweight)

☐ Use radius from field

☑ Use weight from field: 123 sum

Decay ratio: 0.0

Output values: Raw values


QGIS 2.18.24 - MYS_mcda_QGIS
Project Edit View Layer Settings Plugins Vector Raster Database Web MMQGIS Processing Help
Browser Panel
PMY_FMD_risk_pigs.tif
PMY_FMD_risk_roads.tif
PMY_FMD_risk_roads_primary.tif
PMY_FMD_risk_roads_secondary.tif
PMY_FMD_risk_roads_secondary2.tif
PMY_FMD_risk_roads_secondary3.tif
PMY_FMD_risk_roads2.tif
PMY_FMD_risk_roads3.tif
PMY_FMD_risk_roads4.tif
PMY_FMD_risk_roads5.tif
PMY_FMD_risk_roads6.tif
PMY_peninsular_gazetteer.csv
PMY_roads-wgs84.shp
PMY_roads-wgs84_primary.shp
PMY_roads-wgs84_primary_points.shp
PMY_roads-wgs84_primary_raster.tif
Layers Panel
PMY_cmovt_district_2015_raster_clipped2
0
3750
7500
11250
15000
PMY_cmovt_district_2015
PMY_roads-wgs84_primary
PMY_roads-wgs84_secondary
PMY_adm0_poly-wgs84
Layers Panel
Layer Styling
Processing Toolbox
Search...
Recently used algorithms
Convert lines to points
Fixed distance buffer
Clip
Polygons to lines
Dissolve
Raster normalisation
GDAL/OGR [48 geoalgorithms]
GRASS commands [161 geoalgorithms]
GRASS GIS 7 commands [314 geoalgorithms]
Models [0 geoalgorithms]
QGIS geoalgorithms [117 geoalgorithms]
R scripts [18 geoalgorithms]
SAGA (2.3.2) [353 geoalgorithms]
Scripts [0 geoalgorithms]
Coordinate
Scale
Magnifier 100%
Rotation
Render
EPSG:4326

10.2.3 Animal movements

We assume that animals are legally transported using the primary roads, and that animals are illegally transported using minor roads. We also assume that the risk of dissemination of infection decreases proportional to the distance from these roads, up to a maximum distance (we will be conservative and specify 10 km).

One way of doing this is by generating a buffer around the roads. A problem with this is that in this case, the infection risk is equal and constant for the entire distance (rather than decaying)(Figure 10.1 a)). An alternative approach is to first convert the polyline layer to a points layer; we can then generate a heatmap from this, which is reflective of a risk reduction proportional to distance (Figure 10.1 b)).

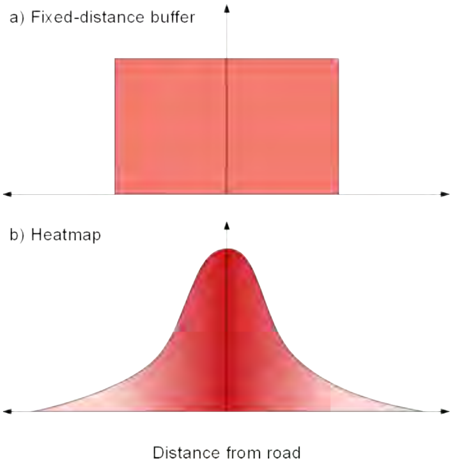
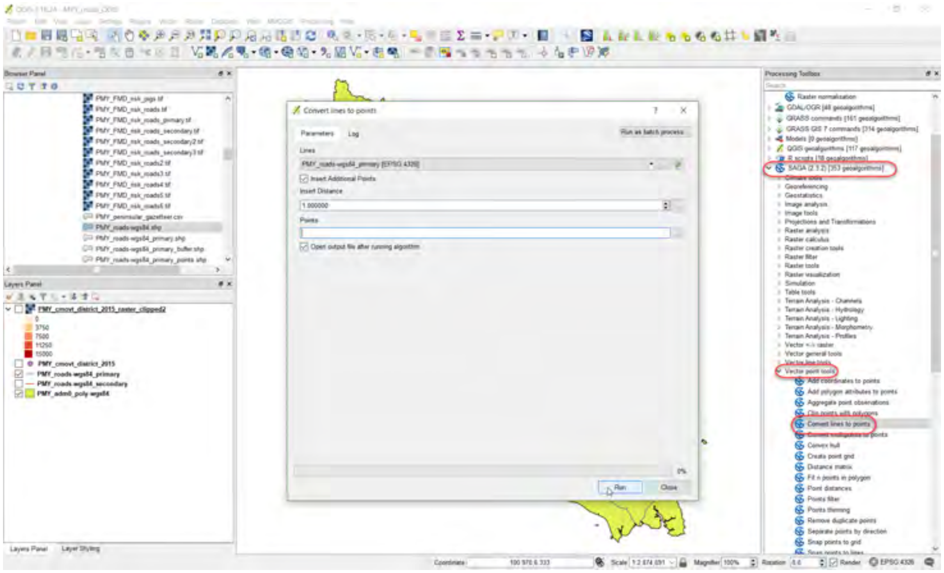


Figure 10.1: Methods for specifying FMD risk (y-axis) as a function of distance from roads.

- Open the Processing Toolbox and under SAGA, open Vector point tools followed by Convert lines to points. Specify the primary roads layer and run the algorithm to create a points layer. Repeat this for the secondary roads.



Convert lines to points

Parameters | Log | Run as batch process...

Lines: PMY_roads-wgs84_primary [EPSG:4326]

☑ Insert Additional Points

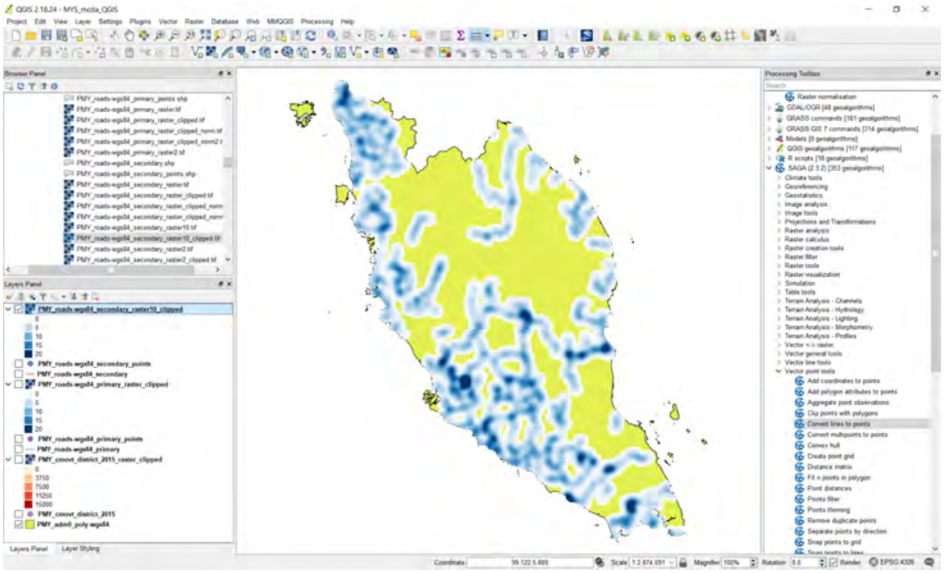
Insert Distance: 1.000000

Points:

☑ Open output file after running algorithm

Run | Close

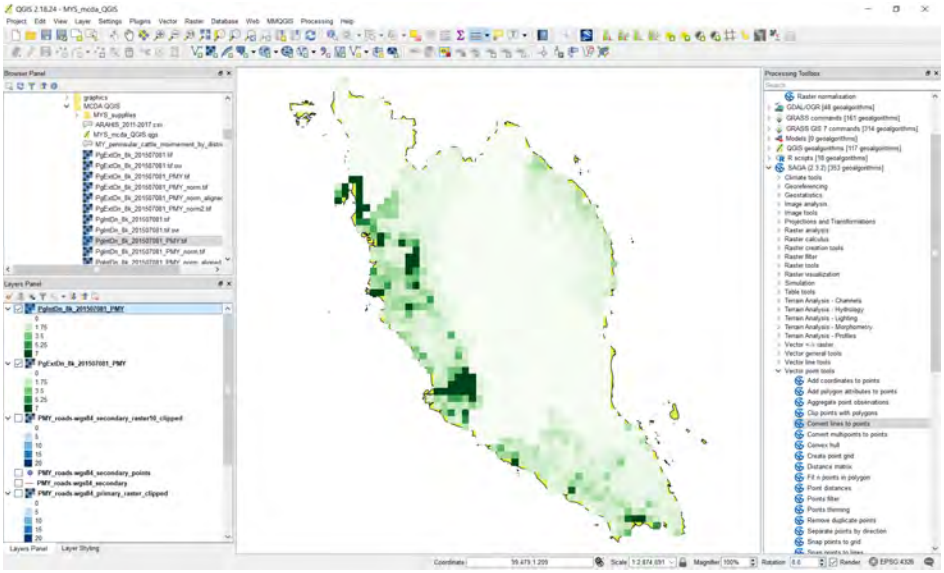
- Now, generate heatmaps of both these layers, specifying a radius of 10,000. Clip the layers to the boundary. You should end up with something similar to the screenshot below.



10.2.4 Contacts with pigs

Finally, we assume that pigs represent a reservoir of infection that can be transmitted to other FMD-susceptible livestock. We differentiate between contacts or proximity to farmed pigs and feral pigs. For the first, we use the Gridded Livestock of the World layer `PgIntDn_8k_201507081.tif`; for the second, we use the layer `PgExtDn_8k_201507081.tif`.

To prepare for these, load the raster layers and clip these to the `adm0` boundary.

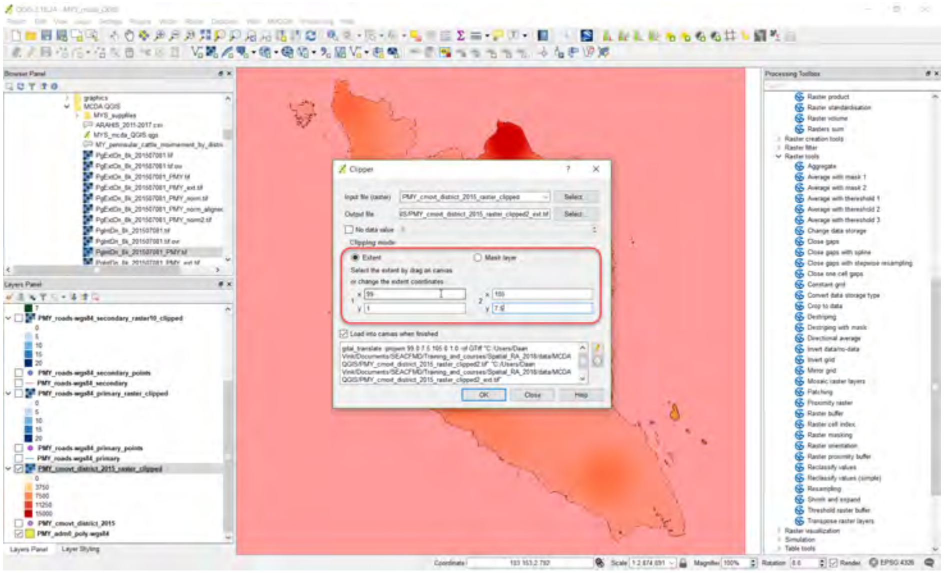

QGIS 2.18.24 - MYS_mcda_QGIS
Browser Panel
Layers Panel
PgIntDn_8k_201507081_PMY
PgExtDn_8k_201507081_PMY
PMY_roads-wgs84_secondary_raster10_clipped
PMY_roads-wgs84_secondary_points
PMY_roads-wgs84_secondary
PMY_roads-wgs84_primary_raster_clipped
Processing Toolbox
Layers Panel
Layer Styling

10.3 Standardising the layers

10.3.1 Clip to standardised extents

The extents of the raster layers should be the same for all the layers. So we should define these extents so that they are identical.

For each layer in turn, use the Raster – Extraction – Clipper to clip the layer to standardised extents that are outside of the country boundary. For our layers, use x (99, 105) and y (1, 7.5).



Input file (raster) PMY_cmovt_district_2015_raster_clipped

Output file IS/PMY_cmovt_district_2015_raster_clipped2_ext.tif

□ No data value

Clipping mode

● Extent ○ Mask layer

Select the extent by drag on canvas
or change the extent coordinates

1 x 99 y 1

2 x 105 y 7.5

☑ Load into canvas when finished

```
gdal_translate -projwin 99.0 7.5 105.0 1.0 -of GTiff "C:/Users/Daan Vink/Documents/SEACFMD/Training_and_courses/Spatial_RA_2018/data/MCDA QGIS/PMY_cmovt_district_2015_raster_clipped2.tif" "C:/Users/Daan Vink/Documents/SEACFMD/Training_and_courses/Spatial_RA_2018/data/MCDA QGIS/PMY_cmovt_district_2015_raster_clipped2_ext.tif"
```

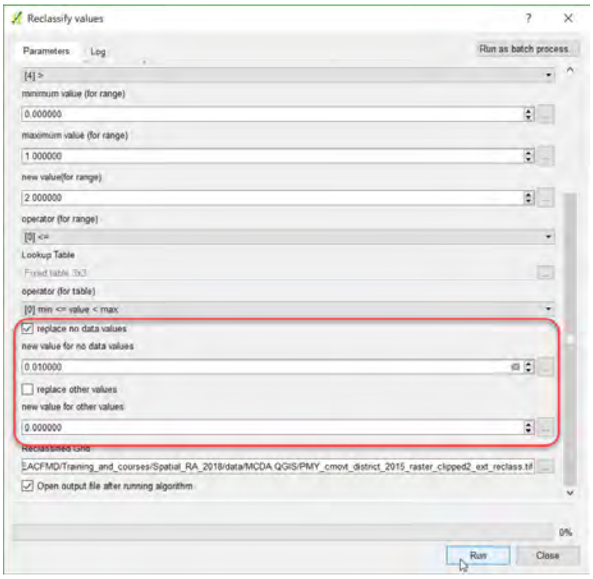
10.3.2 Dealing with null values and zeroes

A consequence of expanding the raster extents is that the raster layers now contain null values. In addition, the minimum value in the raster is zero, which can also be problematic.

Before normalising our layers, we should first convert all null values to zeroes, and subsequently add a very small offset of 0.01 to each value raster cell. We can do this in one action.

For each of the five raster layers:

- In the Processing Toolbox, open SAGA – Raster tool – Reclassify values. Select the raster layer.
- Scroll to the bottom and check that the box `replace no data values` is ticked; specify a value of 0.01. Untick the `replace other values` box. Specify and save the reclassified raster file.



[4] >

minimum value (for range)

0.000000

maximum value (for range)

1.000000

new value(for range)

2.000000

operator (for range)

[0] <=

Lookup Table

Fixed table 3x3

operator (for table)

[0] min <= value < max

☑ replace no data values

new value for no data values

0.010000

☐ replace other values

new value for other values

0.000000

Reclassified Grid

EACFMD/Training_and_courses/Spatial_RA_2018/data/MCDA QGIS/PMY_cmovt_district_2015_raster_clipped2_ext_reclass.tif

☑ Open output file after running algorithm

10.4 Normalising and aligning the raster layers

Before we can perform the WLC calculation, it is important to realise that all of these raster layers have different scales (range of lowest to highest value and distribution). Hence, we need to scale the data to make the layers comparable. The formula to scale any series of values is simple:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

In words, this means that a normalised value (let's call it z_i) takes the single raster cell value (x_i) minus the smallest value in the raster ($\min(x)$) divided by the difference between the largest value and the smallest value. As $\min(x)$ is often zero (or very small), it can often be ignored.

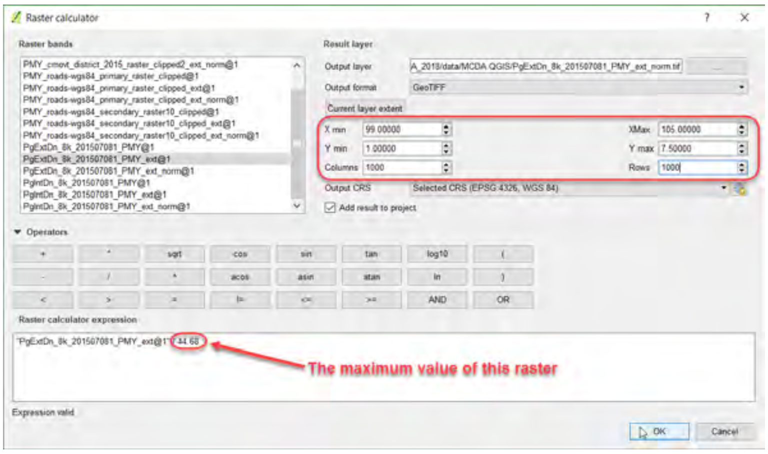
10.4.1 Performing the calculations

There is an algorithm to easily perform this in the SAGA toolbox (SAGA – Raster calculus – Raster normalisation). However, this may not be successful.

A manual way of performing this is to use the Raster calculator:

- For each layer, open the Properties – Metadata dialog and make note of the minimum and maximum values in the raster.

- Then open the Raster Calculator and write an equation to normalise the layer (as in the screenshot below). Note that we already applied a 0.01 offset. In most cases, the minimum value is disregarded as it is negligibly small.
- Confirm that all rasters have the same extents as we defined above.
- Also specify the same number of rows and columns for the output rasters (e.g. 1000 rows and 1000 columns).



Raster bands:

PMY_cmovt_district_2015_raster_clipped2_ext_norm@1
PMY_roads-wgs84_primary_raster_clipped@1
PMY_roads-wgs84_primary_raster_clipped_ext@1
PMY_roads-wgs84_primary_raster_clipped_ext_norm@1
PMY_roads-wgs84_secondary_raster10_clipped@1
PMY_roads-wgs84_secondary_raster10_clipped_ext@1
PMY_roads-wgs84_secondary_raster10_clipped_ext_norm@1
PgExtDn_8k_201507081_PMY@1
PgExtDn_8k_201507081_PMY_ext@1
PgExtDn_8k_201507081_PMY_ext_norm@1
PgIntDn_8k_201507081_PMY@1
PgIntDn_8k_201507081_PMY_ext@1
PgIntDn_8k_201507081_PMY_ext_norm@1

Output layer: A_2018/data/MCDA QGIS/PgExtDn_8k_201507081_PMY_ext_norm.tif
Output format: GeoTIFF
Current layer extent
X min: 99.00000 XMax: 105.00000
Y min: 1.00000 Y max: 7.50000
Columns: 1000 Rows: 1000
Output CRS: Selected CRS (EPSG:4326, WGS 84)
☑ Add result to project

```
"PgExtDn_8k_201507081_PMY_ext@1" / 44.68
```

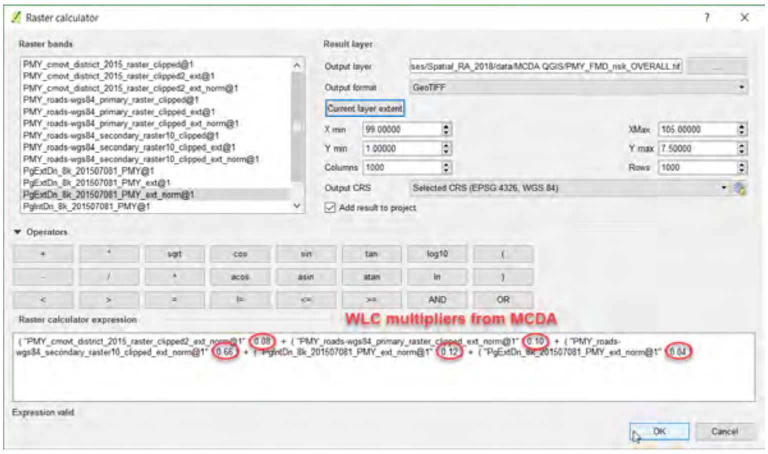
Expression valid

10.5 Applying the WLC using REMBRANDT weights

Finally, we apply the multipliers calculated using the MCDA process and REMBRANDT technique to each of the layers and then sum these to obtain our final risk surface.

From the MCDA exercise using the REMBRANDT technique, we determined the following weights for the five raster layers:

Spatial risk factor	Weight
Illegal movement of livestock	0.66
Swill feeding (i.e. semi-intensive pigs)	0.12
Legal movement of livestock	0.10
Grazing practices	0.08
Wild animal movement (free-roaming pigs)	0.04



Raster bands:

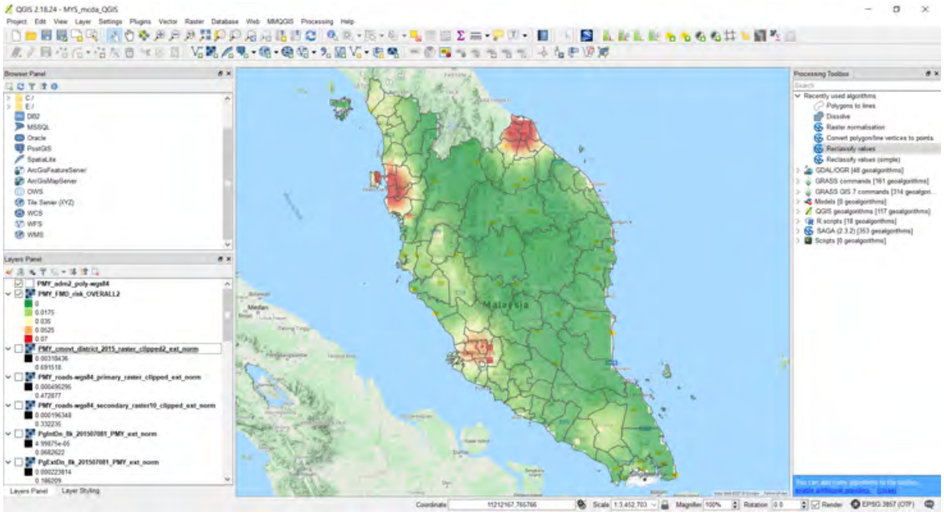
PMY_cmovt_district_2015_raster_clipped@1
PMY_cmovt_district_2015_raster_clipped2_ext@1
PMY_cmovt_district_2015_raster_clipped2_ext_norm@1
PMY_roads-wgs84_primary_raster_clipped@1
PMY_roads-wgs84_primary_raster_clipped_ext@1
PMY_roads-wgs84_primary_raster_clipped_ext_norm@1
PMY_roads-wgs84_secondary_raster10_clipped@1
PMY_roads-wgs84_secondary_raster10_clipped_ext@1
PMY_roads-wgs84_secondary_raster10_clipped_ext_norm@1
PgExtDn_8k_201507081_PMY@1
PgExtDn_8k_201507081_PMY_ext@1
PgExtDn_8k_201507081_PMY_ext_norm@1
PgIntDn_8k_201507081_PMY@1

Output layer: ses/Spatial_RA_2018/data/MCDA QGIS/PMY_FMD_risk_OVERALL.tif
Output format: GeoTIFF
Current layer extent
X min: 99.00000 XMax: 105.00000
Y min: 1.00000 Y max: 7.50000
Columns: 1000 Rows: 1000
Output CRS: Selected CRS (EPSG:4326, WGS 84)
☑ Add result to project

```
( "PMY_cmovt_district_2015_raster_clipped2_ext_norm@1" * 0.08 ) + ( "PMY_roads-wgs84_primary_raster_clipped_ext_norm@1" * 0.10 ) + ( "PMY_roads-
wgs84_secondary_raster10_clipped_ext_norm@1" * 0.66 ) + ( "PgIntDn_8k_201507081_PMY_ext_norm@1" * 0.12 ) + ( "PgExtDn_8k_201507081_PMY_ext_norm@1" * 0.04 )
```

Expression valid

You can now style the resulting risk surface map, and generate a nice printable map using the Print Composer.


QGIS 2.18.24 - MYS_mcda_QGIS
Project Edit View Layer Settings Plugins Vector Raster Database Web MMQGIS Processing Help
Browser Panel
C:/
E:/
DB2
MSSQL
Oracle
PostGIS
SpatiaLite
ArcGisFeatureServer
ArcGisMapServer
OWS
Tile Server (XYZ)
WCS
WFS
WMS
Layers Panel
PMY_adm2_poly-wgs84
PMY_FMD_risk_OVERALL2
0
0.0175
0.035
0.0525
0.07
PMY_cmovt_district_2015_raster_clipped2_ext_norm
0.00318436
0.691518
PMY_roads-wgs84_primary_raster_clipped_ext_norm
0.000495295
0.472877
PMY_roads-wgs84_secondary_raster10_clipped_ext_norm
0.000196348
0.332236
PgIntDn_8k_201507081_PMY_ext_norm
4.99875e-05
0.0682622
PgExtDn_8k_201507081_PMY_ext_norm
0.000223814
0.186209
Layers Panel
Layer Styling
Malaysia
Medan
Processing Toolbox
Search
Recently used algorithms
Polygons to lines
Dissolve
Raster normalisation
Convert polygon/line vertices to points
Reclassify values
Reclassify values (simple)
GDAL/OGR [48 geoalgorithms]
GRASS commands [161 geoalgorithms]
GRASS GIS 7 commands [314 geoalgori...
Models [0 geoalgorithms]
QGIS geoalgorithms [117 geoalgorithms]
R scripts [18 geoalgorithms]
SAGA (2.3.2) [353 geoalgorithms]
Scripts [0 geoalgorithms]
Coordinate 11212167,765766
Scale 1:3,452,703
Magnifier 100%
Rotation 0.0
Render
EPSG:3857 (OTF)

10.6 Further work and uses of the risk surface

There are several further options which you may wish to explore:

1. You can extrapolate the District-level FMD risk by using the Zonal statistics tool, as we have done previously. Such information can be useful for targeting of risk communication or disease control efforts.
2. If you have geolocation data for villages, you can overlay these and extract a village-level FMD risk estimate.
3. As a form of validation, you can overlay the locations of reported FMD outbreaks and assess the degree to which these occurred in higher risk areas.

Of course, there is scope for refinements of the analysis and performing more sophisticated options. This should be based on evidence (literature or field data) or by utilising techniques such as MCDA. Some examples might include:

- Inclusion of other ruminant livestock layers (e.g. buffaloes) – these can be assigned other weights.
- Instead of using all the roads, reduce this to the roads along which most livestock travel (effectively, movement corridors). These can be added as another layer, or it can replace the roads layer.
- Incorporation of additional layers, e.g. related to climate and vegetation, risk points for FMD transmission (e.g. markets and slaughterhouses), transboundary livestock movements, etc.

10.7 Discussion

Advantages of performing this work is that it enables us to generate a comprehensive risk surface. By utilising open-source spatial data and non-FMD data such as livestock census information, as well as a technique for utilising local knowledge and expert opinion, we are able to perform this analysis with limited disease data.

However, we need to also be aware of the limitations and dangers of doing this:

- Subjective nature of expert elicitation to perform MCDA and determine the WLC.
- Limitations in the data: old or inaccurate census figures (or interpolated spatial data such as in this case); underreporting (poor or biased disease data).
- It is possible to generate an authoritative-looking map which may be taken at too much value by decision-makers. It is very difficult to assess the accuracy of the outputs. It is possible to perform the analysis in such a way as to suggest hotspots exist, whereas in fact the true risk may be much more homogeneous (e.g. in an endemic setting).

11 Practical 5b. Spatial risk assessment using R

11.1 Objective

In this practical, we will demonstrate how to implement knowledge-driven spatial risk assessment to generate a risk surface for FMD in cattle in Myanmar. Refer to Chapter 6 for background information.

Outputs of an MCDA exercise performed previously with stakeholder groups identified: (1) grazing practices; (2) legal animal movement; (3) the presence of backyard (semi-intensively managed) pigs; (4) illegal animal movement; and (5) wild animal movement as contributors to the risk of FMD incursions in South East Asian countries.

Our task now is to retrieve appropriate spatial data for each of these risk factors and to manipulate that data into a format suitable for spatial risk analysis. Subsequently, we will use the multipliers required for our Weighted Linear Combination (WLC) tp develop an FMD incursion risk map for Peninsular Malaysia for the purpose of illustration.

We will be performing this exercise using `R`. You will be able to follow the sequence of steps and run the code to obtain the outputs. We will compare these outputs to those obtained using `QGIS`, and discuss the pros and cons of using the different software packages.

11.2 Spatial layers and data

A summary of the spatial data sets required for these analyses is provided in Table 11.1.

Table 11.1: Names and file formats for each of the spatial data layers required for an FMD incursion spatial risk assessment, Peninsular Malaysia.

Layer	Format	Details
`PMY_adm0_poly-wgs84.shp`	SHP	Boundary map of Peninsular Malaysia.
`PMY_cmovt_district_2015.csv`	CSV	Number of cattle moved into each district, 2015.
`PMY_roads-wgs84.shp`	SHP	Peninsular Malaysia primary roads.
`PgIntDn_8k_201507081.tif`	TIF	Semi-intensive pigs – FAO Gridded Livestock of the World.
`PMY_roads-wgs84.shp`	SHP	Peninsular Malaysia secondary roads.
`PgExtDn_8k_201507081.tif`	TIF	Extensive pigs – FAO Gridded Livestock of the World.

11.2.1 Preparing for the analysis

You will find an R script file in the data folder which will perform some analyses on the data, and generate some graphics.

Load the required packages for this exercise and set your working directory:

```
library(raster); library(rgdal); library(ggplot2); library(spatstat);
library(maptools); library(rgeos); library(RColorBrewer); library(RODBC);
library(tidyr); library(dplyr); library(psych)

setwd("C:\\Users\\User_name\\Documents\\...\\data\\MCDA_R\\")
```

11.2.2 Loading map layers

Our boundary map of Peninsular Malaysia is projected in latitude longitude format. Because parts of our analysis involve thinking in terms of distance we project the data to Cartesian coordinates so that distance is expressed in metres (or kilometres) as opposed to decimal degrees. Retrieve the EPSG code for the UTM 48N projection for Malaysia:

```
EPSG <- make_EPSG()
EPSG[grep("48N", EPSG$note), 1:2]
```

11.2.3 Boundary map

The EPSG code for UTM 48N is 32648. Load the Malaysian boundary map and reproject to UTM 48N:

```
mypol.ll <- readOGR(dsn = getwd(), layer = "PMY_adm0_poly-wgs84")
mypol.llext <- extent(mypol.ll)
mypol.utm <- spTransform(mypol.ll, CRS("+init=epsg:32648"))
```

Our analyses (later) will involve some of the functions used by the `spatstat` package. Create an observation window for `spatstat`:

```
mypol.w <- as(as(mypol.utm, "SpatialPolygons"), "owin")
```

Dilate the observation window by 1 km:

```
mypol.w <- dilation(mypol.w, 1000)
```

11.2.4 Animal movements (by road)

To quantify risk associated with legal and illegal animal movement we will use the Malaysian road shape file downloaded from the GADM database of Global Administrative Areas:

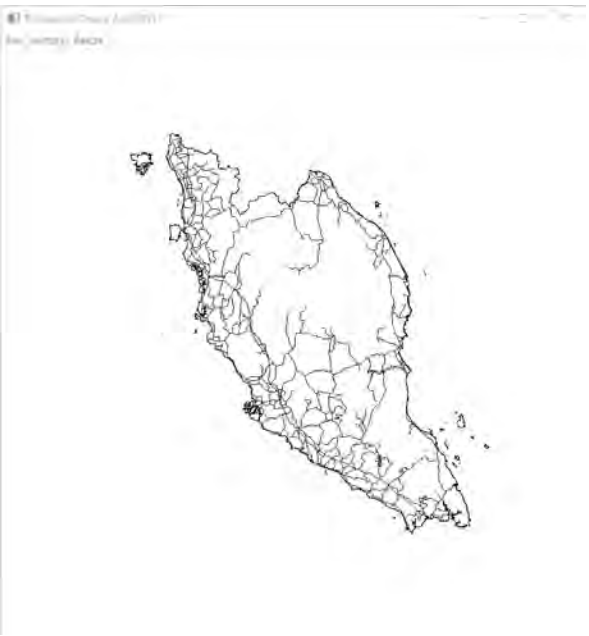
```
myroad.ll <- readOGR(dsn = getwd(), layer = "PMY_roads-wgs84")
myroad.utm <- spTransform(myroad.ll, CRS("+init=epsg:32648"))
```

Clip the road spatial data layer to ensure they're all within the boundaries of Peninsular Malaysia:

```
myroad.utm <- crop(x = myroad.utm, y = mypol.utm)
```

Plot to check:

```
windows(); plot(mypol.utm)
plot(myroad.utm, add = TRUE)
```



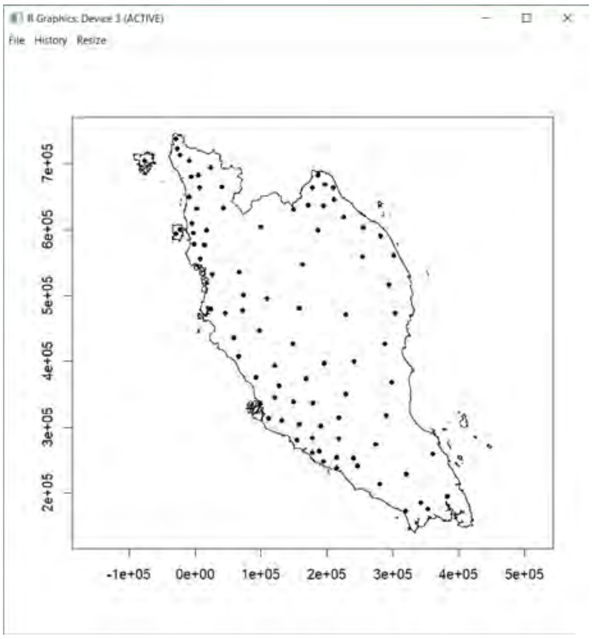
11.2.5 Grazing practices

Read in the district-level Pensinsular Malaysian cattle movement data for the 12-month period beginning January 2015:

```
mymov <- read.table("PMY_cmovt_district_2015.csv",
header = TRUE,
sep = ",")
```

Create a `spatialPointsDataframe` from the movement data and reproject to UTM 48N:

```
coords <- SpatialPoints(mymov[, c("lon", "lat")])
mymov.ll <- SpatialPointsDataFrame(coords, mymov)
proj4string(mymov.ll) <- CRS("+init=epsg:4326")
mymov.utm <- spTransform(mymov.ll, CRS("+init=epsg:32648"))

windows(); plot(mybord.shp, axes = TRUE)
points(mymov.utm, pch = 16, cex = 0.75)
```



Create a spatstat `ppp` object for district centroids:

```
mymov.ppp <- ppp(x = coordinates(mymov.utm)[,1], y = coordinates(mymov.utm)[,2],
marks = mymov.utm$nani, window = mypol.w)
```

Kernel smooth the movement data using a relatively wide bandwidth (20 km). Here we weight the kernel smoothing function by the number of cattle entering each district:

```
mymov.im <- density(mymov.ppp, weights = mymov.utm$nani, sigma = 20000, dimyx = c(200,200))
```

Express the kernel smoothed plot as the number of cattle per square kilometre (1 square metre = 0.000001 square kilometres):

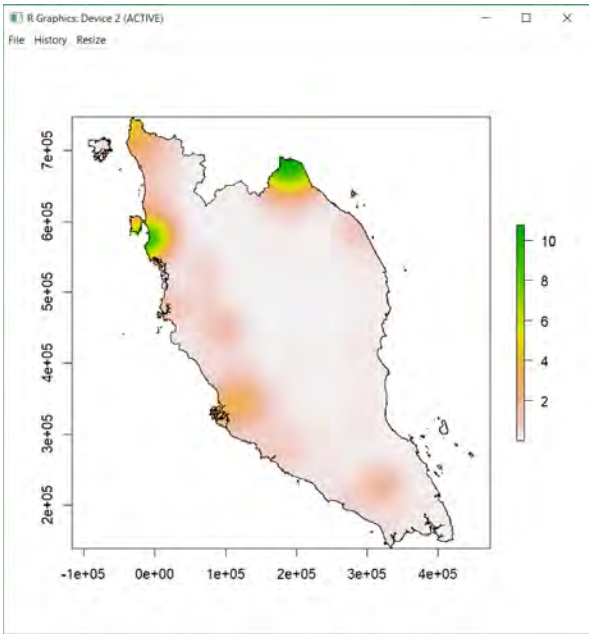
```
mymov.im$v <- mymov.im$v / 0.000001
```

Convert `mymov.im` to a raster object:

```
mymov.r <- raster(mymov.im)
```

Plot to check:

```
windows(); plot(mymov.r)
plot(mypol.utm, add = TRUE)
```



Main roads (legal animal movements)

First we create a subset of the primary roads:

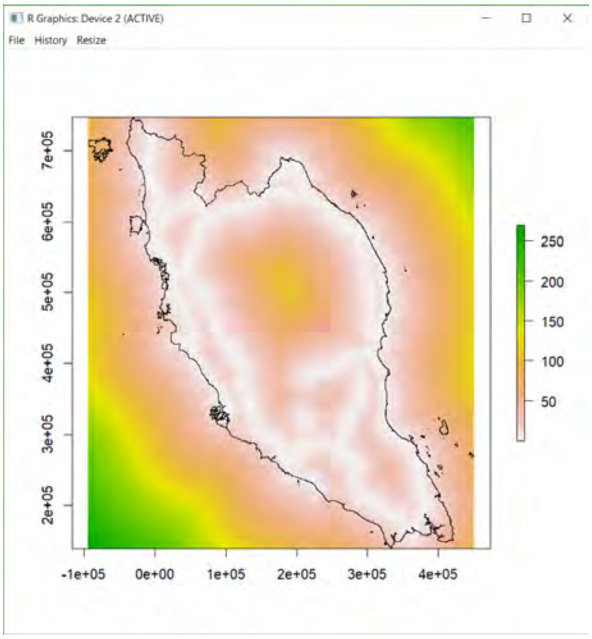
```
id <- myroad.utm$RTT_DESCRI == "Primary Route"
mymroad.shp <- myroad.utm[id ,]

windows(); plot(mypol.utm)
plot(mymroad.shp, add = TRUE)
```

Create a `spatstat im` object defining distance to the nearest main road, expressing this in kilometres:

```
mymroad.psp <- as.psp.SpatialLinesDataFrame(from = mymroad.shp, window = mypol.w)
mymroad.im <- distmap(mymroad.psp)

mymroad.im$v <- mymroad.im$v / 1000

windows(); plot(mymroad.im)
plot(mypol.utm, add = TRUE)
```

Finally, we rasterise the im object:

```
mymroad.r <- raster(mymroad.im)

windows(); plot(mymroad.r)
plot(mypol.utm, add = TRUE)
```



Feeder roads (illegal animal movements)

Develop a spatial data layer of secondary roads. Our assumption is that illegal movements of FMD susceptible species will be via secondary (as opposed to main) roads.

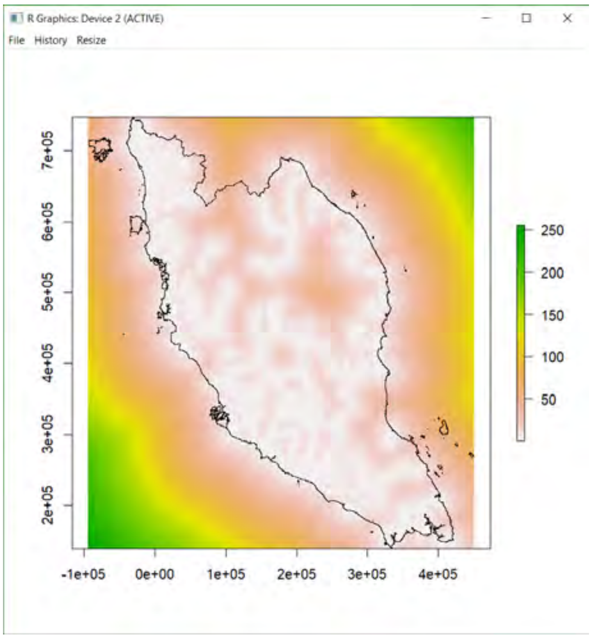
```
id <- myroad.utm$RTT_DESCRI == "Secondary␣Route"
myfroad.utm <- myroad.utm[id,]
```

Create a `spatstat im` object defining distance to the nearest feeder road, expressing this in kilometres:

```
myfroad.psp <- as.psp.SpatialLinesDataFrame(from = myfroad.utm, window = mypol.w)
myfroad.im <- distmap(myfroad.psp)

myfroad.im$v <- myfroad.im$v / 1000

windows(); plot(myfroad.im)
plot(mypol.utm, add = TRUE)
```

Rasterise the im object:

```
myfroad.r <- raster(myfroad.im)

windows(); plot(myfroad.r)
plot(mypol.utm, add = TRUE)
```



11.2.6 Pig density

We use the FAO Gridded Livestock of the World raster layer for this.

We previously identified as risk factors backyard pigs (semi-extensive) and wild pigs (extensive). These are available as separate layers.

Backyard (semi-intensive) pigs

Read in the semi-extensive pig raster layer from FAO's Gridded Livestock of the World:

```
pig.r <- raster("PgIntDn_8k_201507081.tif")
```

Crop the raster to the Malaysian boundary shape file spatial extent:

```
pig.crop <- crop(pig.r, mypol.llext, snap = "out")
```

Dummy the raster with a spatial extent equal to the cropped raster, but full of NA values:

```
crop <- setValues(pig.crop, NA)
```

Rasterise the catchment boundaries, with NA outside the catchment boundaries:

```
bndry.r <- rasterize(mypol.ll, crop)
```

Put NAs in all the raster cells outside of the shape file boundary:

```
mybyp.r <- mask(x = pig.crop, mask = bndry.r)
```

Set the projection of the pig raster:

```
crs(mybyp.r) <- CRS("+init=epsg:4326")
```

Re-project to UTM 48N:

```
mybyp.r <- projectRaster(mybyp.r, crs = CRS("+init=epsg:32648"))
```

Wild animal movements (pigs)

Read in the extensive pig raster layer from FAO's Gridded Livestock of the World:

```
pig.r <- raster("PgExtDn_8k_201507081.tif")
```

Crop the raster to the Malaysian boundary shape file spatial extent:

```
pig.crop <- crop(pig.r, mypol.llext, snap = "out")
```

Dummy the raster with a spatial extent equal to the cropped raster, but full of NA values:

```
crop <- setValues(pig.crop, NA)
```

Rasterise the catchment boundaries, with NA outside the catchment boundaries:

```
bndry.r <- rasterize(mypol.ll, crop)
```

Put NAs in all the raster cells outside of the shape file boundary:

```
myexp.r <- mask(x = pig.crop, mask = bndry.r)
```

Set the projection of the pig raster:

```
crs(myexp.r) <- CRS("+init=epsg:4326")
```

Re-project ot UTM 48N:

```
myexp.r <- projectRaster(myexp.r, crs = CRS("+init=epsg:32648"))
```

11.3 Normalising the layers

Because each of the numeric values represented in each of our different spatial data layers are expressed on different scales (e.g. density of backyard pigs, number of kilometres from the nearest main road) we need to standardise the data to a common scale. The maximum score procedure (Malczewski 1999) scales values to 0 to 1. For linearly increasing criteria the transformation is:

$$x'_{ij} = \frac{x_{ij}}{x_{\max}}$$

And for linearly decreasing criteria:

$$x'_{ij} = 1 - \frac{x_{ij}}{x_{\max}}$$

```
# Grazing practices:
tmymov.r <- (mymov.r / cellStats(mymov.r, stat = max))
windows(); par(mfrow = c(1,2), pty = "s"); hist(mymov.r); hist(tmymov.r)

# Legal animal movement:
tmymroad.r <- 1 - (mymroad.r / cellStats(mymroad.r, stat = max))
windows(); par(mfrow = c(1,2), pty = "s"); hist(mymroad.r); hist(tmymroad.r)

# Backyard (semi-intensive) pigs:
tmybyp.r <- (mybyp.r / cellStats(mybyp.r, stat = max))
windows(); par(mfrow = c(1,2), pty = "s"); hist(mybyp.r); hist(tmybyp.r)

# Illegal animal movement:
tmysroad.r <- 1 - (myfroad.r / cellStats(myfroad.r, stat = max))
windows(); par(mfrow = c(1,2), pty = "s"); hist(myfroad.r); hist(tmyfroad.r)
```

```
# Wild animal movement:
tmyexp.r <- (myexp.r / cellStats(myexp.r, stat = max))
windows(); par(mfrow = c(1,2), pty = "s"); hist(myexp.r); hist(tmyexp.r)
```

11.4 Resampling of processed rasters

A spatial raster data set can be thought of as a data matrix with a defined number of rows and columns. Because our intent is to combine each of the risk factor data layers, it is important that they uniform in terms of the number of rows and columns. The GIS technique ‘resampling’ is the process used to align one raster data set with another. For this example we will re-sample each of the input rasters to the dimensions of the `tmymroad.r` raster layer.

```
# Check the dimensions of each raster in turn and re-sample, if necessary. What is the
resolution of the cattle movement raster layer?
tmymov.r

# Answer: 200    200. Resample tmymov.r to match tmymroad.r:
tmymov.r
rtmymov.r <- resample(x = tmymov.r, y = tmymroad.r, method = "ngb")

# Legal animal movement. All OK.
tmymroad.r

# Backyard (semi-intensive) pigs:
tmybyp.r
rtmybyp.r <- resample(x = tmybyp.r, y = tmybcross.r, method = "ngb")

# Illegal animal movement. All OK.
tmyfroad.r

# Wild animal movement:
tmyexp.r
rtmyexp.r <- resample(x = tmyexp.r, y = tmybcross.r, method = "ngb")
```

11.5 Multicriteria decision analysis (MCDA)

Read in the pairwise comparison responses from each of the 30 participants:

```
dat <- read.table("SEACFMD_incursion_risks.csv", header = TRUE, sep = ",")
```

Create a matrix to capture the responses from each of the participants. Set row and column number for each factor:

```
dat$row <- 0; dat$col <- 0

dat$row[dat$factora == "Grazing␣practices"] <- 1
dat$row[dat$factora == "Legal␣animal␣movements"] <- 2
dat$row[dat$factora == "Swill␣feeding"] <- 3
dat$row[dat$factora == "Illegal␣animal␣movements"] <- 4
dat$row[dat$factora == "Wild␣animal␣movement"] <- 5

dat$col[dat$factorb == "Grazing␣practices"] <- 1
dat$col[dat$factorb == "Legal␣animal␣movements"] <- 2
dat$col[dat$factorb == "Swill␣feeding"] <- 3
dat$col[dat$factorb == "Illegal␣animal␣movements"] <- 4
```

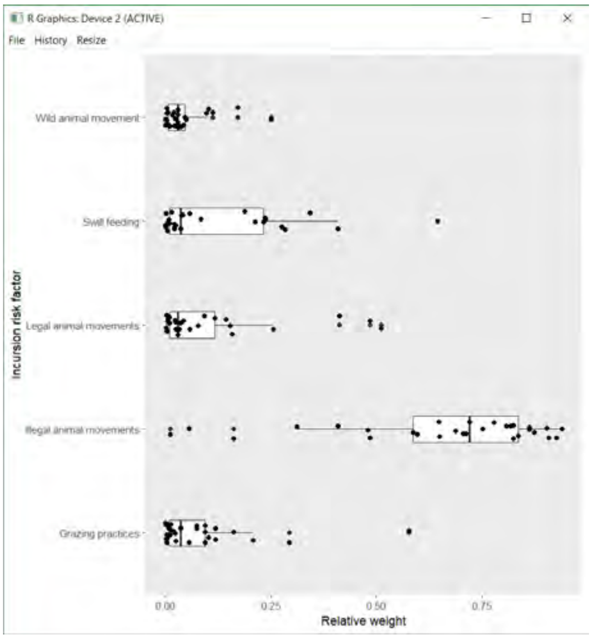
```
dat$col[dat$factorb == "Wild animal movement"] <- 5

id <- c(); country <- c(); factor <- c(); nmean <- c()
```

Loop through each of the responses provided by each of the 30 participants and, for each of the five FMD incursion risk factors, calculate the REMBRANDT weights:

```
who <- unique(dat$id)

for(i in 1:length(who)){
        tmp <- subset(dat, subset = id == who[i])

        # Create a 5 * 5 matrix:
        rval <- data.frame(matrix(rep(0, times = 25), nrow = 5))

        for(j in 1:nrow(tmp)){
                rval[tmp$row[j], tmp$col[j]] <- tmp$score[j]
                rval[tmp$col[j], tmp$row[j]] <- -tmp$score[j]
        }

        colnames(rval) <- c("graz", "lmov", "swill", "imov", "wild")
        rownames(rval) <- c("graz", "lmov", "swill", "imov", "wild")

        # Gamma transformation:
        rval <- exp(rval * log(2))

        # Take the geometric mean of each row:
        rval$gmean <- apply(rval, MARGIN = 1, FUN = function(x) geometric.mean(x,
        na.rm = TRUE))

        # Normalise the data:
        rval$nmean <- rval$gmean / sum(rval$gmean)

        id <- c(id, who[i])
        country <- c(country, as.character(tmp$country[1]))
        factor <- c(factor, rownames(rval))
        nmean <- c(nmean, rval$nmean)
}

rval <- data.frame(id, country, factor, nmean, rank)
rval$factor <- factor(rval$factor, labels = c("Grazing practices", "Illegal
animal movements", "Legal animal movements", "Swill feeding", "Wild animal
movement"))
```

Make a box and whisker plot showing the distribution of REMBRANDT weights for each FMD incursion risk factor:

```
windows(); ggplot(data = rval, aes(x = factor, y = nmean)) +
geom_boxplot(width = 0.25) +
xlab("Incursion risk factor") +
ylab("Relative weight") +
geom_jitter(width = 0.1) +
coord_flip()
```



11.6 Applying the WLC using REMBRANDT weights

Finally, we apply the multipliers calculated using the MCDA process and REMBRANDT technique to each of the layers and then sum these (the 'raster stack') to obtain our final risk surface.

```
# Illegal movement of livestock: 0.66
wtmyfroad.r <- calc(tmyfroad.r, function(x) x * 0.66)
windows(); plot(wtmyfroad.r)
plot(mypol.utm, add = TRUE)

# Swill feeding (i.e. semi-intensive pigs): 0.12
wrtmybyp.r <- calc(rtmybyp.r, function(x) x * 0.12)
windows(); plot(wrtmybyp.r)
plot(mypol.utm, add = TRUE)

# Legal movement of livestock: 0.10
wtmymroad.r <- calc(tmymroad.r, function(x) x * 0.10)
windows(); plot(wtmymroad.r)
plot(mypol.utm, add = TRUE)

# Grazing practices: 0.08
wrtmymov.r <- calc(rtmymov.r, function(x) x * 0.10)
windows(); plot(wrtmymov.r)
plot(mypol.utm, add = TRUE)

# Wild animal movement (free-roaming pigs): 0.04
wrtmyexp.r <- calc(rtmyexp.r, function(x) x * 0.04)
windows(); plot(wrtmyexp.r)
plot(mypol.utm, add = TRUE)
```

Create a raster stack comprised of each of the weighted raster layers, then sum these:

```
fmd.stack <- stack(wtmyfroad.r, wrtmybyp.r, wtmymroad.r, wrtmymov.r, wrtmyexp.r)
fmd.r <- stackApply(fmd.stack, indices = c(1,1,1,1,1), na.rm = TRUE, fun = sum)
```

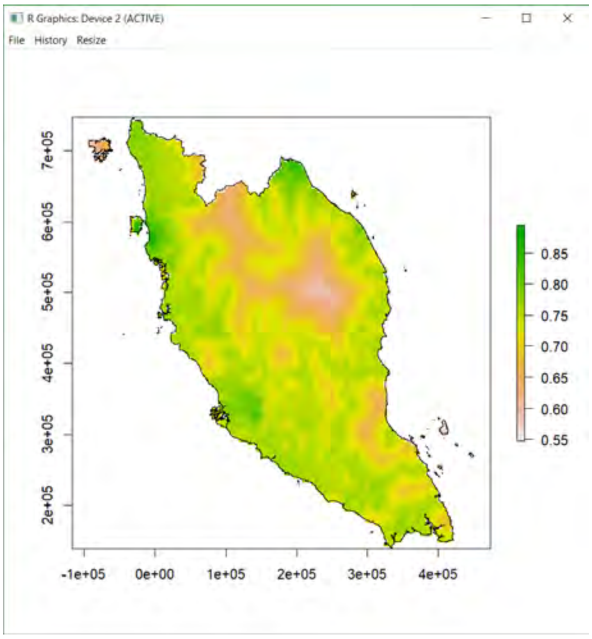
11.7 Plotting the output

Plot to check:

```
windows(); plot(fmd.r)
plot(mypol.utm, add = TRUE)
```

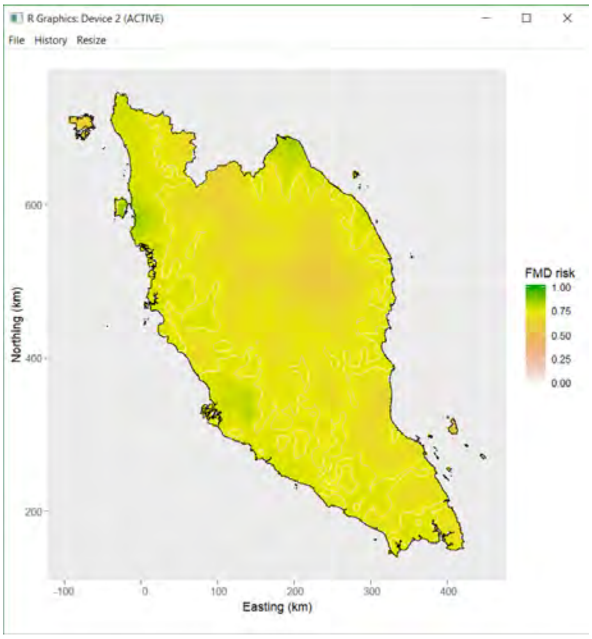
Prepare the map for publication. Crop fmd.r:

```
mypol.utm <- readOGR(dsn = getwd(), layer = "MY_peninsular_adm0_poly-wgs84")
mypol.utm <- spTransform(mypol.utm, CRS("+init=epsg:32648"))
mypol.ext <- extent(mypol.utm)

fmd.crop <- crop(fmd.r, mypol.ext, snap = "out")
crop <- setValues(fmd.crop, NA)
bndry.r <- rasterize(mypol.utm, crop)
fmd.r <- mask(x = fmd.crop, mask = bndry.r)
crs(fmd.r) <- CRS("+init=epsg:32648")

windows(); plot(fmd.r)
plot(mypol.utm, add = TRUE)
```



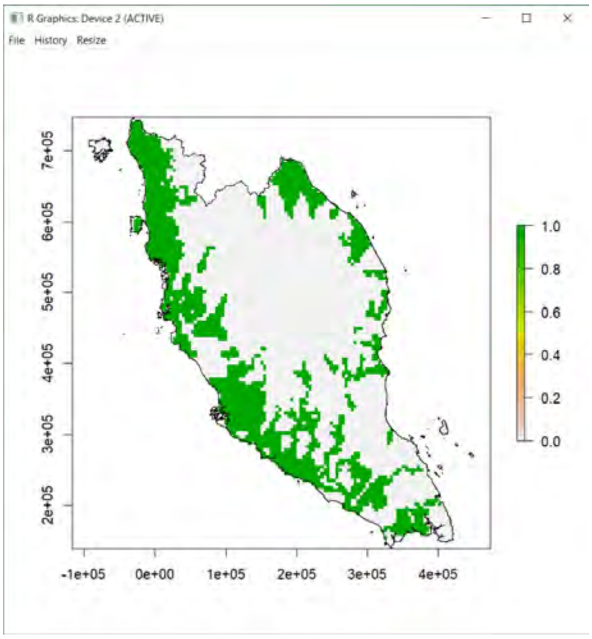
The same map using ggplot2:

```
fmd.df <- data.frame(rasterToPoints(fmd.r))
mypol.df <- fortify(mypol.utm, region = "ID_0")

mformat <- function(){
function(x) format(x / 1000, digits = 2)
}

breaks <- seq(from = 0, to = 1, length = 5)
cols <- rev(terrain.colors(n = 5, alpha = 1))

windows(); ggplot(data = fmd.df) +
geom_tile(aes(x = x, y = y, fill = index_1)) +
scale_fill_gradientn(colours = cols, breaks = breaks, limits = c(0,1)) +
geom_contour(data = fmd.df, aes(x = x, y = y, z = index_1), breaks = 0.75, colour = "white") +
geom_polygon(data = mypol.df, aes(x = long, y = lat, group = group), col = "black",
 fill = "transparent") +
coord_equal() +
scale_x_continuous(labels = mformat()) +
scale_y_continuous(labels = mformat()) +
labs(x = "Easting␣(km)", y = "Northing␣(km)", fill = "FMD␣risk")
```



A simpler map for policy makers:

```
tfmd.r <- fmd.r
tfmd.r[tfmd.r <  0.75] <- 0
tfmd.r[tfmd.r >= 0.75] <- 1

windows(); plot(tfmd.r)
plot(mypol.utm, add = TRUE)
```



11.8 Model verification

Read in the ARAHIS outbreak data for 2011 – 2017:

```
arahis <- read.csv("ARAHIS_2011-2017.csv", header = TRUE)
```

Select only those FMD outbreaks that occurred in Malaysia:

```
id <- arahis$country == "Malaysia"
arahis <- arahis[id,]
```

Create a `spatialPointsDataframe` and reproject to UTM 48N:

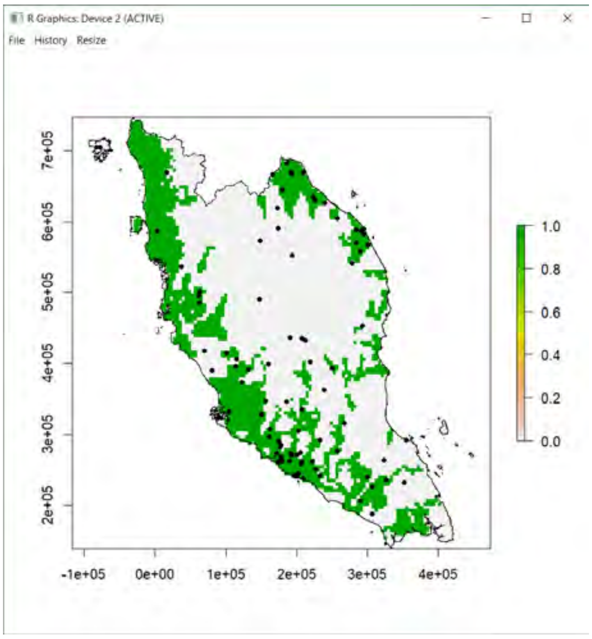
```
coords <- SpatialPoints(arahis[, c("longitude", "latitude")])
arahis.ll <- SpatialPointsDataFrame(coords, arahis)
proj4string(arahis.ll) <- CRS("+init=epsg:4326")
arahis.utm <- spTransform(arahis.ll, CRS("+init=epsg:32648"))
```

Drop those point locations outside of the `mypol.w`:

```
id <- inside.owin(x = coordinates(arahis.utm)[,1], y = coordinates(arahis.utm)[,2],
w = mypol.w)
arahis.utm <- arahis.utm[id,]
```

Plot to check:

```
windows(); plot(tfmd.r)
points(arahis.utm, pch = 16, cex = 0.75)
plot(mypol.utm, add = TRUE)
```



What proportion of the total number of FMD outbreaks were in the areas defined as high risk by our MCDA?

```
arahis.utm$fmd <- raster::extract(x = tfmd.r, y = arahis.utm)
sum(arahis.utm$fmd[arahis.utm$fmd == 1], na.rm = TRUE) / length(arahis.utm$fmd)
```

A total of 68 of the 101 FMD outbreaks that occurred in Peninsular Malaysia between January 2012 and April 2017 (68%) occurred in the high risk areas defined by our MCDA.

12 Practical 6. Investigation of spatial clustering in R

12.1 Preliminaries and acknowledgements

12.1.1 R software set-up

The setup for analysis in this document uses a “Project” in “RStudio” . RStudio is a software interface to R, that is, R is opened when you open R Studio and you interact with R through RStudio. RStudio provides tools for working with R which help the analyst in many ways, including providing “Projects” to aid file organisation.

File organisation and folder structure is a personal preference, but a basic structure with informative names which suits many purposes is:

- “MyProjectName” folder
  - “Code” sub folder
  - “Data” sub folder
  - “Docs” sub folder

This structure will work with the following code for accessing the data that are provided below. (Note that the sub folders are all on the same “level”).

Create a new project in the following steps:

1. Open R studio

2. Click in on “File- New Project. . . ”

3. Choose to create a “New Directory” if none already exists, or “Existing Directory” if one already exists (for example, the “Code” directory)

4. Browse to and create a new directory, or just browse to the directory you want, and then click “Create Directory”

Create a new .R file from which you run your code, and save and name it appropriately e.g. “MyRFile.R” (you must add the “.R” extension manually). R code is identified in the shaded rows, and can be copied from this PDF into the upper left window in R Studio, and run from there.

Note that code or text which is preceded on a line with a “#” character is ignored by R. These comments are used to organise the structure of the exercises and briefly explain what is happening to aid understanding. Some code is commented out for the sake of brevity, but you are encouraged to run these (and create your own in addition) to aid your understanding.

The data for the exercises first needs to be copied into the directory chosen to store it ("Data" in this case). You are now ready to go.

R packages and base R itself, are frequently updated. The authors do this to fix bugs and add new features, which is great. Unfortunately, sometimes the code in the updates "breaks" old code that previously worked. Therefore, if your versions of R or the packages used in the following code either predate those used to produce the accompanying exercises, or sometime in the future if you return to these code and they don't work, then one reason for the problem may be that the software versions don't match. If you can't fix the problem, you can install the versions that are defined in 2.4 at the end of the last content chapter.

12.1.2 Exercises and getting on-line help

Alongside the notes are exercises and questions to check the student's understanding of the principles and application of the learnt methods. Additional "extension" exercise are available to develop further skills and understanding.

You are encouraged to use the 'help' resources in R to understand the functions. If the function used are in packages loaded at the start of the exercise (using the 'library' function), then simply typing "?function-name" in the "Console" screen (usually lower left screen in R Studio) will bring up the help page for that function in the "Help" tab (usually lower right window in R Studio). When R functions are referred to in this document, the package in which they are located is indicated first, followed by a colon (:), and then the function name, for example 'graphics::plot'. Cited references and additional other helpful resources for learning spatial epidemiology at the introductory level are found the Bibliography at the end of this document.

12.1.3 Data sets used with these notes and exercises

Extensive use is made of different data sets in the exercises. These data are from three main sources- R packages, other public on-line sources, and from previous spatial epidemiology or GIS course. The data from R packages are simply imported into the work space by loading the relevant library with the 'library' function, then using the 'data'(data set name) function. An explanation of how the data were obtained and a description of the fields is available by using the 'help?' function as previously described, with the name of the data set in parentheses beside it. These data are from a variety of scientific fields, with a few human and animal health examples.

The second group of data sets was obtained from the website for the book "Veterinary Epidemiologic Research" (Dohoo, Martin, and Stryhn 2009) and relate to the relevant Chapters 25 and 26. These data represent the results of surveillance for avian influenza (AI) between 2004 and 2006 in a region in northern Vietnam, extracted from a larger data set of the whole country, and carry the prefix "viet" in their. Several spatial data sets are available for different administrative levels, including coordinates of communes (actually the centroids of their areas) and attribute data on whether or not the commune experienced one or more outbreak of AI in the surveillance period. Further information on these data are available in Pfeiffer et al. (2007)

Finally, we borrow from other datasets used in a previous course which is acknowledged below. The "cheshire" data set contains locations of farms and dates of infection with foot-and-mouth disease in the 2001 outbreak in the UK county of that name. The Scottish lip cancer data set

(“SC_lip_cancer”) consists of spatial polygon files of the districts in that country, and attributes of the number of cases of that cancer and the population in each district.

The exercises are designed to introduce methods and stimulate your thinking about them. They are not complete analyses in themselves, and may not even be the best or only approach to the analytical problem. A deeper understand than can be taught in a course can only come from practicing the methods and further reading on the subjects, and better still conducting your own research!

12.1.4 Acknowledgements

Many of the data exercises (and material for the accompanying presentation) originate from Mark Stevenson’s comprehensive notes “Investigation of Spatial Patterns of Animal Disease” provided for the Introductory Spatial Epidemiology course in “118.817 Advanced analysis of epidemiologic data” held at Massey University 13 - 17 Nov 2017, and I want to acknowledge his work that has contributed to these notes.

References

Dohoo, I., W. Martin, and H. Stryhn. 2009. Veterinary Epidemiologic Research. 2nd ed. Atlantic Veterinary College Inc., Charlottetown, Canada.

Pfeiffer, D.U., P.Q. Minh, V. Martin, M. Epprecht, and M.J. Otte. 2007. “An Analysis of the Spatial and Temporal Patterns of Highly Pathogenic Avian Influenza Occurrence in Vietnam Using National Surveillance Data.” Veterinary Journal 174 (2): 302–9. https://doi.org/10.1016/j.tvjl.2007.05.010.

12.2 Investigation of spatial clustering in R

12.2.1 Background and definitions

Investigation of possible clusters of disease is a core role of epidemiologists that aims to provide information on the causes of disease and means for it’s control and prevention. Analysis of the spatial and temporal patterns of disease occurrence provide a quantitative description of the apparent problem and possible new insights into the cause of the disease that might have otherwise been unnoticed with other methods.

Many factors may determine the spatial distribution of disease occurrence, such as the density and species of animal populations, movement and trade in these populations, geographic and climatic factors that affect the animals, their environment or disease vector species, or existing disease control programmes. Due consideration of these factors is important when applying methods to evaluate clustering of disease.

The history of methods to detect clusters of disease has grown since the 1980’s out of increased concerns about actual or potential adverse environmental effects on public health, for example from nuclear power stations. However, many diseases will show geographic (and possibly temporal) clustering for other reasons that are associated with the disease, and not the one postulated. Hence, in some situations diseases may appear to cluster, even when the known aetiology doesn’t suggest it should be observed.

Reasons for studying disease clustering:

- In epidemiological study of aetiology of disease
- In public/population health disease surveillance
- In response to disease cluster (outbreak) alarms to evaluate whether further investigations are warranted

In each setting, it is important to account for the population at risk in the study, as the distribution of cases may just reflect the underlying population distribution.

Study of disease aetiology

Differential disease rates in large scale or localised geographic areas have long been used to study disease aetiology. The same methods can be used to detect areas with high rates of disease in which to conduct further epidemiologic studies (especially cohort studies for which it is important to enrol subjects with a relatively increased risk of disease to achieve sufficient statistical power in a cost-efficient way).

A key feature of these methods for purely spatial or space-time data is that they are able to pinpoint the location of clusters. However, the interest may alternatively be in general nature of distribution of disease, for example, to understand whether the disease is likely to be infectious (may exhibit both space-time interaction and spatial clustering), or has risk factors that vary geographically (only spatial clustering). When assessing the association between geographic risk variables and disease risk in an ecological analysis, it is important to adjust for any spatial autocorrelation not explained by the known variables. When individual subject exposure is unavailable or unknown, focused cluster tests can be used to move from pure hypothesis to established risk factors and the conduct of further epidemiologic studies.

Population surveillance

The detection of clusters of disease with known risk factors may prompt a public health response where those factors can be reduced or removed, for example from water-borne diseases or environmental pollution. Additionally, geographic areas with high mortality rates may be searched for, and adjustment made for the distribution of incidence instead of the population at risk, to thereby identify areas of substandard treatment or screening. Alternatively, areas of reduced risk may indicate successful treatment or screening programmes from which others can learn, or they may alternately reflect areas of under-reporting of cases.

Response to disease cluster alarms

These are more commonplace in the human public health setting, for example for leukemia incidence around nuclear power stations, or clusters of cases of swine flu (in humans) in 2009 or outbreaks of meningitis.

Definitions of clustering

- When the exact extent or form of the clusters to be studied are unknown, a cluster is:
  - ” a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance” (Knox 1989 cited in @Lawson1999)

  - or less formally- an area within the study region of significant elevated risk
  - also known as a ‘hot-spot’
- When the extent or form of the cluster has been previously defined, a cluster occurs when
  - the study region displays a pre-specified cluster structure

Other relevant statistical concepts

- A spatial process is stationary when the dependence between measurements of the same (outcome) variable is the same for all locations in the study area
- A spatial process is isotropic when it is affected by distance from a location, but the effect is the same in all directions
- A first-order effect describes large-scale variations in the mean of the same variable across the study region
- A second-order effects describes small-scale variation due to interactions between neighbours

12.2.2 Methods to investigate clustering of disease

A range of methods exist to investigate spatial and temporal clustering of disease because of the various data types that might be obtained from epidemiologic studies, and because no single method is clearly superior to the alternatives. Cluster evaluation methods are clearly useful to assess whether an excess of cases have occurred to inform further responses. However, several problems are likely to be met when attempting to evaluate clustering:

- Cases of disease may be rare or distributed over an extended period of time, which means that they may not be detected
- Information on the population at risk may be unavailable or of poor quality
- The background occurrence of disease in the population may be unknown

The types of clustering methods have been defined by Besag and Newell (1991) and Tango (1999):

- General, global or non-specific clustering detection methods
  - An analysis of the overall clustering tendency of a disease in a study region, regardless of where that clustering is located
  - Analogous to assessment of autocorrelation
  - Assesses the overall or global aspects of clustering
  - These methods do not estimate the spatial locations of clusters
  - Assumes that the disease process is the same (stationary) throughout the study region and therefore give no indication of local variability of clustering
  - Most easily interpreted when applied to regions where the factors that determine disease occurrence are relatively uniform
- Localised non-focused spatial cluster detection methods
  - Aim to define the location and intensity of any clusters if they exist
- Localised focused spatial cluster detection methods
  - Examine clusters of pre-defined characteristics and their extent around a predetermined focus point
  - Examples of focused local clustering methods are mainly from putative sources of health hazard for humans from environmental contamination, but can be applied in

the veterinary setting, for example with infectious diseases to putative sources such as livestock markets or transport routes

It is important to consider the structure of hypotheses that could include clustering components in any analysis of geographic health data. If the disease of interest naturally clusters (beyond that explained by the background population), then this form of clustering should be investigated. This form of clustering may arise from unobserved covariates and should be considered as heterogeneity, and can often be modeled through the use of random effects. Additionally, close attention needs to be paid to the assumptions and methods of the models to verify that they are appropriate for the study data and research question. In particular, when optional parameters need to be specified for the method, these should be chosen a priori on the basis of the biology of the disease being studied and possibly also on what other authors have published, so as not to 'adjust' the method to obtain a pre-determined result. Unfortunately, guidance in the literature is not always available or clear for each situation the analyst encounters, so these methods should be used with caution.

Global methods to investigate clustering or autocorrelation

Ripley's K-function difference

- Spatial data: Point
- Data needed: Coordinates of point locations of cases and controls (or population at risk)
- How test works:
  - Calculates the number of events of the same type occurring within a certain distance of a randomly-chosen case
  - When spatial autocorrelation is present then events of the same type are likely to be in closer proximity to one another, and for small distances, K will be larger than expected
  - Assumes a stationary point process and that the direction to neighbour does not affect the spatial process (isotropy), but options exist is software for inhomogenous point process data
  - Reflects the second-order effects of a spatial process
  - Edge effects can be adjusted with a choice of methods
  - Because the variance of the K function increases with distance, it is recommended that the pre-defined distance scale for the function is no more than one third of the linear extent of the study area or half the length of the shorter side of the rectangular study area
  - Therefore this test is most suitable for testing clustering over a relatively small extent of the study region
  - Spatial heterogeneity of population at risk (or controls) is accounted for by use of a difference function (cases - controls)
  - Results are visualised in a plot of the K against distance, which should have a parabola shape if process CSR
    - Expresses the scaled expected number of excess cases occurring at a given distance from a random case
    - Significance of difference tested by Monte Carlo simulation which also allows statistical inference in presence of non-stationarity
    - Any deviation of the difference function above the envelope formed by the upper and lower bounds of the expected difference indicates significant clustering of cases

  - Difference function also estimates the spatial extent or distance over which clustering occurs, but the extent must be small in relation to the whole study region
- References: Ripley (1977), Diggle and Chetwynd (1991), Waller and Gotway (2004)

Ex 6.1

```
## K-function difference library(smacpod) data(grave) # load data levels(grave$marks) # check
levels of marks attribute to ensure that cases are correctly ordered plot(grave) kd1 = kdest(grave)
plot(kd1, iso ˜ r, ylab = “difference”, legend = FALSE, main = “”) kd2 = kdest(grave, nsim = 99,
level = 0.8) # Simulate to create envelope plot(kd2, ylim = c(-2e+06, 6e+06), legend = TRUE)
## The two bands are each for the minimum and maximum differences between the cases and
controls, ## respectively ## Statistical test for clustering using difference in K functions pval <-
kdplus.test(kd2) pval$pvalue detach(package:smacpod)
```

Q 6.1 How would you interpret the test result? Does it agree with your first visual impression of the data in the plot? Is there evidence in the data of clustering of cases, and over what range of distances?

A “by hand” calculation of the K-function difference from previous the previous K-function exercise using epi.incin data is shown below. The explicit steps help understand how the test works

```
##For bivariate data, estimate separate point patterns for cases and controls ## and calculate
the difference of the K-functions ## Peaks or troughs in the plot above or below the simulation
confidence envelope ## are evidence of clustering or regularity, respectively data(epi.incin);
dat.df <- epi.incin dat.df$status <- factor(dat.df$status, levels = c(0,1), labels = c(“Neg”, “Pos”))
names(dat.df)[3] <- “Status” dat.w <- convexhull.xy(x = dat.df[,1], y = dat.df[,2]) dat.w <- dilation(dat.w,
r = 1000) coords <- matrix(c(dat.w$bdry[[1]]$x, dat.w$bdry[[1]]$y), ncol = 2, byrow
= FALSE) pol <- Polygon(coords, hole = FALSE) pol <- Polygons(list(pol),1) pol <- SpatialPolygons(list(pol))
pol.spdf <- SpatialPolygonsDataFrame(Sr = pol, data = data.frame(id =
1), match.ID = TRUE) pol.map <- fortify(pol.spdf) dat.ppp <- ppp(x = dat.df[,1], y = dat.df[,2],
marks = factor(dat.df[,3]), window = dat.w) ## Estimate the area of the study region: lenx <-
max(dat.df$xcoord) - min(dat.df$xcoord); lenx leny <- max(dat.df$ycoord) - min(dat.df$ycoord);
leny ## The study area is 17960 x 16550 metres. Compute the K-function and plot it: r <-
seq(from = 0, to = 1000, by = 10) pos.kenv <- envelope(split(dat.ppp)$Pos, fun = Kest, r = r,
nrank = 2, nsim = 99, verbose = FALSE) neg.kenv <- envelope(split(dat.ppp)$Neg, fun = Kest,
r = r, nrank = 2, nsim = 99, verbose = FALSE) pos.ggplot <- ggplot(data = pos.kenv, aes(x =
r, y = obs)) + geom_line() + geom_ribbon(aes(ymin = lo, ymax = hi), colour = “#cccccc”, alpha
= 0.25) + geom_line(data = pos.kenv, aes(x = r, y = theo), col = “red”, linetype = “dashed”) +
ylab(“K(r)”) + theme(aspect.ratio = 1) neg.ggplot <- ggplot(data = neg.kenv, aes(x = r, y = obs))
+ geom_line() + geom_ribbon(aes(ymin = lo, ymax = hi), colour = “#cccccc”, alpha = 0.25) +
geom_line(data = neg.kenv, aes(x = r, y = theo), col = “red”, linetype = “dashed”) + ylab(“K(r)”) +
theme(aspect.ratio = 1) grid.arrange(pos.ggplot, neg.ggplot, ncol = 2) kdif.spat <- data.frame(r =
r, obs = pos.kenv$obs - neg.kenv$obs, theo = pos.kenv$theo - neg.kenv$theo, lo = pos.kenv$lo -
neg.kenv$lo, hi = pos.kenv$hi - neg.kenv$hi) ## Plot the results: ggplot(data = kdif.spat, aes(x =
r, y = obs)) + geom_line() + geom_ribbon(aes(ymin = lo, ymax = hi), colour = “#cccccc”, alpha =
0.25) + geom_hline(yintercept = 0, linetype = 2) + ylim(-3E06, 3E06) + labs(x = “Distance (m)”, y
= “K-function difference”)
```

Extension exercise 6.1

```
## Load Vietnam AI data library(foreign); library(sp); library(spatstat) library(ggplot2); library(ggforce);
library(smacpod) commune.d <- read.dta(“..//Data//commune-d.dta”) dat.c.all.df
<- data.frame(xcoord = commune.d$x_coord, ycoord = commune.d$y_coord, status = com-
```

mune.d$infected) dat.c.all.df$status <- factor(dat.c.all.df$status, levels = c("No", "Yes"), labels = c("Neg", "Pos")) summary(dat.c.all.df) ## Create observation window dat.c.all.w <- convexhull.xy(x = dat.c.all.df$xcoord, y = dat.c.all.df$ycoord) dat.c.all.w <- dilation(dat.c.all.w, r = 1) ## Create a ppp objects for positive and negative communes: pos.ppp <- ppp(x = dat.c.all.df$xcoord[dat.c.all.df$status == "Pos"], y = dat.c.all.df$ycoord[dat.c.all.df$status == "Pos"], window = dat.c.all.w) neg.ppp <- ppp(x = dat.c.all.df$xcoord[dat.c.all.df$status == "Neg"], y = dat.c.all.df$ycoord[dat.c.all.df$status == "Neg"], window = dat.c.all.w) ## Plot data coords <- matrix(c(dat.c.all.w$bdry[[1]]$x, dat.c.all.w$bdry[[1]]$y), ncol = 2, byrow = FALSE) pol <- Polygon(coords, hole = FALSE) pol <- Polygons(list(pol),1) pol <- SpatialPolygons(list(pol)) pol.spdf <- SpatialPolygonsDataFrame(Sr = pol, data = data.frame(id = 1), match.ID = TRUE) pol.map <- fortify(pol.spdf) g <- ggplot() + geom_point(data = dat.c.all.df, aes(x = xcoord, y = ycoord, colour = status, shape = status), alpha = 0.5) + geom_polygon(data = pol.map, aes(x = long, y = lat, group = group), col = "black", fill = "transparent") + scale_shape_manual(values = c(1, 16)) + scale_colour_manual(values = c("royalblue", "red")) + labs(x = "Easting (km)", y = "Northing (km)", fill = "status") + coord_equal() g ## Combine positive and negative communes com.ppp <- ppp(x = dat.c.all.df$xcoord, y = dat.c.all.df$ycoord, window = dat.c.all.w, marks = dat.c.all.df$status) levels(com.ppp$marks) # check levels of marks attribute to ensure that cases are correctly ordered # plot(com.ppp) Uninformative because of over-plotting ## Run K difference test kd1 = kdest(com.ppp) plot(kd1, iso ˜ r, ylab = "difference", legend = FALSE, main = "") # Simulate to create envelope- use only 19 sims to shorten processing time kd2 = kdest(com.ppp, nsim = 19, level = 0.8) plot(kd2, legend = TRUE) ## The two bands are each for the minimum and maximum differences between the cases and controls, ## respectively ## Statistical test for clustering using difference in K functions pval <-kdplus.test(kd2) pval$pvalue

Cuzick-and-Edwards' k-nearest neighbour test

- Spatial data: Point
- Data needed: Coordinates of point locations of cases and controls
- How test works:
  - Accounts for heterogeneous distribution of population at risk by comparing coordinates of cases and controls
  - Requires a pre-determined scale parameter k which defines the number of nearest-neighbours (may be > 1)
  - Null hypothesis is that nearest neighbour to a randomly-selected case is just as likely to be a control as another case
  - Test statistic is number of cases that are nearest neighbours to each individual case- when cases are clustered, the nearest neighbour to a case will tend to be near to another case
  - Assumes controls were randomly selected from same source population as cases
  - If controls are selected on basis of potential confounding variables (for example by matching) then their effect can be adjusted for
  - May also be used with aggregated areal data
- References: Cuzick and Edwards (1990)

Cuzick-and-Edward's k-nearest neighbour test—- qnn.test(com.ppp, q = 1:5) # Search up to 5 nearest neighbours detach(package:smacpod)

Extn. Q 6.1 Is their evidence of spatial clustering of cases in this data set, and over what distances? Do the test results agree?

Moran's I test for autocorrelation Spatial data: Areal data with continuous or ordinal attribute Data needed: Areal aggregated data, neighbour matrix (optionally with weights) How test works: - Commonly-used test for global clustering - Originally developed for event counts within sub-regions, but typically used now with standardised measures such as rates because of differences in sizes of populations at risk in different sub-regions - Can also be applied to Bernoulli (0/1) data and residuals from a regression model - Assesses the similarity between pairs in their deviation from the global mean - Neighbouring values that are similar result in a positive I, and conversely if different, I is negative - A positive value of I implies clustering, and a negative value dispersion or repulsion - The expected value of I is 1/(N1) 1 / (N 1) where N N = the number of points in the study, and approaches zero when N N increases - If used for proportion data, the population-size weighting is lost, and therefore, a population-adjusted statistic , Odens Ipop I p o p that accounts for the population, is preferred - Test results are sensitive to the form of neighbour graph and weights, therefore check against Monte Carlo test results - Assumptions: - If analysing only count data the population at risk should be evenly distributed over the study region - If disease proportions or rates are used and sub-region populations differ, then the assumption of constant variance of the measured attribute will be violated. One option to account for this includes Oden's method (Oden 1995, an adaptation of Moran's I) - First order or spatial trends are absent in the study region - The correlation between sub-regions are the same in all directions, that is, the data are isotropic - Weights matrix is symmetric (see spdep::moran.test for solution when it isn't) - A useful additional method is to create a 'spatial correlogram' to investigate the spatial lag or distance over which autocorrelation occurs - Moran's I (or Geary's C- see below) is calculated for different distances or spatial lags and plotted against those same lags or distances - The results should be interpreted with caution from a spatial correlogram if there are few remaining included observations when the spatial lag order increases - Also, the adjacent values in these plots are correlated, and therefore correlations at greater lags should be cautiously considered - References: Moran (1950), Cliff and Ord (1981), Assuno and Reis (1999), Ward and Carpenter (2000)

Ex 6.2

```
library(rgdal); library(ggmap); library(ggplot2); library(epiR); library(classInt); library(RColorBrewer); library(plyr) library(rgeos); library(sp); library(maptools); library(spdep) ## Import data file. Districts are numbered consecutively from 1 to 56. ## Retrieve disease event details: sclip.df <- read.table(“..//Data//SC_lip_cancer.csv”, header = TRUE, sep = “,”) #head(sclip.df) ## The disease event table uses the same district numeric codes as the shape file ## Calculate the standardised morbidity ratio, its standard error and 95% confidence interval: exp <- (sum(sclip.df$obs) / sum(sclip.df$pop)) * sclip.df$pop smr <- round(epi.conf(as.matrix(cbind(sclip.df$obs, exp)), ctype = “smr”), digits = 2) sclip.df <- cbind(sclip.df, exp, smr) names(sclip.df)[c(7:10)] <- c(“smr”, “smr.se”, “smr.lci”, “smr.uci”) ##Import polygon shapefile: scpol.shp <- readOGR(dsn = “..//Data”, layer = “SC_districts-LL”, verbose = FALSE) scpol.shp@data$DISTRICT ## Convert polygons to a data frame for plotting using fortify: scpol.df <- fortify(scpol.shp, region = “ID”) scpol.df$id <- as.numeric(scpol.df$id) ## Join columns from scpol.shp@data to scpol.df: scpol.df <- join(x = scpol.df, y = scpol.shp@data) ## Merge columns from sclip.df to scpol.df: scpol.df <- merge(x = scpol.df, y = sclip.df, by.x = “id”, by.y = “id”) ## Create fixed class intervals for plotting ppal <- brewer.pal(n = 5, name = “Reds”) q <- classIntervals(sclip.df$smr, n = 5, style = “fixed”, fixedBreaks = c(0, 1, 2, 4, 10, Inf)) qColours <- findColours(q, ppal) ## Gradient scale choropleth map using ggplot: ggplot(data = scpol.df) + geom_polygon(aes(x = long, y = lat, group = group, fill = smr)) + geom_path(aes(x = long, y = lat, group = group), colour = “grey”, size = 0.25) + scale_fill_gradientn(limits = c(0, 10), colours = brewer.pal(n = 5, “Reds”), guide = “colourbar”) + labs(x = “Longitude”, y = “Latitude”, fill = “Lip cancer SMR”) + theme(legend.position = c(0.2, 0.8), legend.background = element_rect(colour =
```

NA, fill = NA)) + coord_map() ## We begin by creating queen contiguity neighbours. ## Tracts sharing boundary points are defined as neighbours. ## The poly2nb function accepts a SpatialPolygonsDataFrame as its (first) argument: scpol.01nb <- poly2nb(scpol.shp, queen = TRUE) plot(scpol.shp, border = “#7f7f7f”); plot(scpol.01nb, coords, add = TRUE) ## Check neighbours of individual districts: id <- 11; neigh <- unlist(scpol.01nb[id]) col <- rep(“white”, times = 56) col[id] <- “red”; col[neigh] <- “yellow” plot(scpol.shp, col = col, border = “#7f7f7f”, axes = TRUE) ## Now define second order neighbours: sc.nblag <- nblag(scpol.01nb, maxlag = 2) scpol.02nb <- sc.nblag[[2]] class(scpol.02nb) <- NULL ## And again, check the neighbours of individual districts: id <- 11; neigh <- unlist(scpol.02nb[id]) col <- rep(“white”, times = 56) col[id] <- “red”; col[neigh] <- “yellow” plot(scpol.shp, col = col, border = “#7f7f7f”, axes = TRUE) ## Sometimes it might of interest to identify cumulative neighbours: sc.cnblag <- nblag_cumul(nblags = sc.nblag) ## And plot to check: id <- 11; neigh <- unlist(sc.cnblag[id]) col <- rep(“white”, times = 56) col[id] <- “red”; col[neigh] <- “yellow” plot(scpol.shp, col = col, border = “#7f7f7f”) ## Create spatial weights for defined neighbours ## Row-standardised- sum of weights of neighbours add to 1 scpol.02nb <- nb2listw(scpol.01nb, style = “W”) scpol.02nb ## Binary weights- 1 for neighbours, 0 for non-neighbours scpol.03nb <- nb2listw(scpol.01nb, style = “B”) scpol.03nb ## General binary weights use the glist argument to include extra a priori ## information about spatial relationships between ## areas. Here we allow the weight to be inversely proportional to distance. coords <- coordinates(scpol.shp) # Define coordinates of polygon centroids dsts <- nbdists(scpol.01nb, coords) idw <- lapply(dsts, function(x) 1/(x/10000)) scpol.04nb <- nb2listw(scpol.01nb, glist = idw, style = “B”) scpol.04nb ## Finally, weight matrices can be imported as a *.gal file (written from GeoDa). ## If you are using a GeoDa generated ## *.gal file make sure the row ordering of your data is the same as the order ## in which polygons are listed in the shape ## file of your study area. scpol.01nb <- read.gal(“..//Data//SC_districts_BNG.gal”) coords <- coordinates(scpol.shp) plot(scpol.shp, border = “#7f7f7f”, axes = FALSE) plot(scpol.01nb, coords, add = TRUE) title(main = “GAL order 1 links with first nearest neighbours in red”) col.knn <- knearneigh(coords, k = 1) plot(knn2nb(col.knn), coords, add = TRUE, col = “red”, length = 0.08) ## Run Moran’s I test with and without simulation and plot spatial correlogram moran.test(sclip.df$obs, listw = scpol.02nb) EBImoran.mc(sclip.df$obs, sclip.df$exp, listw = scpol.02nb, nsim = 999) plot(sp.correlogram(scpol.01nb, sclip.df$obs, order = 3, method = “I”))

Q. 6.2 Is there evidence of global spatial autocorrelation and over what number of neighbouring districts? Are these results consistent with the choropleth map of standardised morbidity ratios previously plotted in 6.1?

Extension exercise 6.2

Global spatial clustering of areal data- Morans I test for autocorrelation—- ## Load required packages library(rgdal); library(ggmap); library(ggplot2); library(epiR); library(classInt); library(RColorBrewer); library(plyr) library(rgeos); library(sp); library(maptools); library(spdep) library(DCluster) ## Import data on 117 districts with number of communes and number affected with AI in any year viet_dist.shp <- readOGR(dsn = “..//Data”, layer = “viet_district_poly”, verbose = FALSE) summary(viet_dist.shp) # summary of records in data frame str(viet_dist.shp@data) ## Convert No_commune from factor to integer viet_dist.shp@data$No_commune <- as.integer(viet_dist.shp@data$No_commune) ## Create id variable in viet_district.shp for districts id <- data.frame(id = 1:nrow(viet_dist.shp@data)) viet_dist.shp@data <- cbind(viet_dist.shp@data, id) summary(viet_dist.shp) ## Estimate SMR of each district for plotting ## Calculate the number of expected infected communes, the standardised morbidity ratio, ## its standard error and 95% confidence interval from the data attributes exp <- (sum(viet_dist.shp@data$All_AI) / sum(as.integer(viet_dist.shp@data$No_commune))) *

as.integer(viet_dist.shp@data$No_commune) smr <- round(epi.conf(as.matrix(cbind(viet_dist.shp@data$All_AI exp)), ctype = “smr”), digits = 2) viet_dist_ai.df <- data.frame(id = 1:nrow(viet_dist.shp@data), exp, smr) names(viet_dist_ai.df)[c(3:6)] <- c(“smr”, “smr.se”, “smr.lci”, “smr.uci”) obs <- viet_dist.shp@data$All_AI pop <- viet_dist.shp@data$No_commune viet_dist_ai.df <- cbind(viet_dist_ai.df, obs, pop) summary(viet_dist_ai.df) ## Plot SMR with sp.plot ## Create cut-points and transform SMR into categorical variable cat.smr <- cut(viet_dist_ai.df$smr, breaks = c(0, 0.0000001, 0.5, 1, 2, 5, Inf), include.lowest = TRUE, right = FALSE, dig.lab = 2) levels(cat.smr) <- c(“0”, “0 - 0.5”, “0.5 - 1”, “1 - 2”, “2 - 5”, “5+”) summary(cat.smr) #Create a dataframe with all the information needed for the map plotting and further analysis map.cat.smr <- data.frame(id = viet_dist.shp@data$id , exp = exp, cat.smr = cat.smr) summary(map.cat.smr) #Merge the categorised SMR for each sub-region with the spatial polygon into a new shape file viet_dist_smr.shp <- merge(viet_dist.shp , map.cat.smr, by.x = “id”, by.y = “id”) ## Plot with spplot ## Greyscale spplot(obj = viet_dist_smr.shp, zcol = “cat.smr”, col.regions = gray(seq(0.9,0.1, length=6)), asp=1) ## Colourscale gradient scale my.palette <- brewer.pal(n = 7, name = “OrRd”) spplot(viet_dist_smr.shp, zcol = “cat.smr”, col.regions = my.palette) rm(cat.smr, map.cat.smr, my.palette) ## Create neighbours and weights ## define district centroids coords <- coordinates(viet_dist.shp) ## Create queen contiguity neighbours. viet_dist.01nb <- poly2nb(viet_dist.shp, queen = TRUE) plot(viet_dist.shp, border = “#7f7f7f”, axes = TRUE); plot(viet_dist.01nb, coords, add = TRUE) summary(viet_dist.01nb) ## 1 district without link (an island) which(card(viet_dist.01nb) == 0) # id = 117 id <- 117; neigh <- unlist(viet_dist.01nb[id]) col <- rep(“white”, times = 117) col[id] <- “red”; col[neigh] <- “yellow” plot(viet_dist.shp, col = col, border = “#7f7f7f”, axes = TRUE) rm(id, neigh, col) ## Identify nearest neighbour (this may be artificial as don’t know boat connections. . .) id <- 82; neigh <- unlist(viet_dist.01nb[id]) col <- rep(“white”, times = 117) col[id] <- “red” plot(viet_dist.shp, col = col, border = “#7f7f7f”, axes = TRUE) # No neighbour link to island rm(id, neigh, col) ## Manually add a link to the islands disrict not automatically connected viet_dist.01nb[82] viet_dist.01nb[117] viet_dist.01nb[[82]]= as.integer(sort(c(viet_dist.01nb[[82]], 117))) viet_dist.01nb[[117]]= as.integer(82) ## Check edit summary(viet_dist.01nb) which(card(viet_dist.01nb) == 0) # id = 0 i.e. no districts without neighbours id <- 117; neigh <- unlist(viet_dist.01nb[id]) col <- rep(“white”, times = 117) col[id] <- “red”; col[neigh] <- “yellow” plot(viet_dist.shp, col = col, border = “#7f7f7f”, axes = TRUE) # neighbour link created rm(id, neigh, col) ## Weights ## Row standardised viet_dist.02nb <- nb2listw(viet_dist.01nb, style = “W”, zero.policy = TRUE) ## Binary weights viet_dist.03nb <- nb2listw(viet_dist.01nb, style = “B”, zero.policy = TRUE) ## General binary weights use the glist argument to include extra a priori ## information about spatial relationships between ## areas. Here we allow the weight to be inversely proportional to distance. dsts <- nbdists(viet_dist.01nb, coords) idw <- lapply(dsts, function(x) 1/(x/100)) viet_dist.04nb <- nb2listw(viet_dist.01nb, glist = idw, style = “B”, zero.policy = TRUE) ## Run Moran’s I tests, Moran scatterplot and spatial correlogram moran.test(viet_dist_ai.df$smr, listw = viet_dist.02nb, zero.policy = TRUE) moran.mc(viet_dist_ai.df$smr, listw = viet_dist.02nb, nsim = 999, zero.policy = TRUE) plot(sp.correlogram(viet_dist.01nb, viet_dist_ai.df$smr, order = 5, method = “I”))

Extn. Q 6.2 Do the test results support the SMR plots and what you know about the biology of the disease? Do the tests as applied violate any assumption?

Geary’s C autocorrelation (or contiguity ratio) statistic

- Spatial data type: Areal
- Data needed: Counts of observed or expected number of cases, or residuals from regression model, spatial weights matrix (for some software implementations)
- How test works:

  - A weighted estimate of spatial autocorrelation that considers similarities between pairs of sub-regions as defined by the neighbourhood matrix (unlike Moran's I which considers all the defined neighbours)
  - The test statistic varies between 0 (perfect positive autocorrelation) and 2 (perfect negative autocorrelation)
  - Note that the test is sensitive to the assumptions used in the form of neighbour relationships and spatial weights (notably for non-symmetric weights more so than for Moran's I), but that solutions exist for these problems (see spdep::geary.test)
- References: Geary (1954)

Getis and Ord's global G test for spatial autocorrelation

- Spatial data type: Areal data
- Data needed: Number of cases and population at risk in each sub-region of study area
- How test works:
  - Measures overall concentration or lack of concentration of all pairs of points within a distance of each other
  - G test may be used in conjunction with I test to reveal any patterns not revealed by the latter
  - Inappropriate for regression residuals, use cautiously when distance between sub-regions compartively small or large
- References: Cliff and Ord (1981)

Tango's statistic (excess events test, EET) for general clustering

- Spatial data type: Areal
- Data needed: Count of observed and expected number of cases in a sub-region, neighbours list with spatial weights
- How test works
  - Measures the difference between the observed and expected rate of cases in each sub-region
  - Weights the differences by the distance between the sub-regions so that smaller distances have greater weighting
  - Requires a predetermined value for the scale over which clustering occurs (which may not be known) which in turn makes the test more sensitive to clustering over that distance
  - This creates problems of both multiple testing when comparing different values for the scale parameter, and of 'picking winners'
  - These problems addresse with Tango's maximum excess events test (MEET)
- References: Tango (1995), Tango (2000)

Ex. 6.3

Global spatial clustering- Tango's statistic—- ## Create data in form required for test ## Create data frame of observed and expected number of infected communes + locations coords <- coordinates(viet_dist.shp) viet.df <- data.frame(Observed = viet_dist.shp@data$All_AI, Expected = exp, x = coords[, 1], y = coords[, 2]) summary(viet.df) ## Use previously-created neighbour lists tango.stat(viet.df, viet_dist.02nb) tango.test(Observed˜offset(log(Expected)), viet.df, model="poisson", R=99, list=viet_dist.02nb)

Extn. Q. 6.3 Do the results of Tango's test agree with those of the Moran's I test? If not, what reasons might explain these findings?

Localised non-focused spatial cluster detection

Spatial clustering may be investigated in three different dimensions (linear, point and area) and where for a range of different data types: case-control or case and population count data, dichotomous, categorical, rank or continuous). Each test was primarily developed to be used with one data type, but it its possible to aggregate point to areal data, and possibly areal to point data (by using areal centroids). However, the scanning methods for point data are less suitable for areal data as sub-regions may not neatly fall within the scanning circle.

Randomness of runs and sign test This test is appropriate when investigating linear clusters, for example when the event locations refer to a road or coastline, or row of pens or hutches:

- Appropriate for linearly distributed, case-control data
- Runs are adjacent similar points
- More runs equates to greater clustering

Besag and Newell's statistic for spatial clustering

- Spatial data type: Areal
- Data needed: Number of cases and population at risk in each sub-region of study area
- How test works:
  - Investigates whether clusters of cases in a region exist greater than a pre-determined number k k (usually 2 - 10 for rare diseases)
  - May use Poisson or negative binomial distributions
  - At every area where a case was detected, the number of neighbouring regions required to reach k k is counted
  - If number of regions required to reach k k is fewer than expected, then it is identified as a cluster
  - Drawbacks of this test
    - Need to define cluster size before testing- a value that is too small may not identify larger clusters, or a value too large may falsely identify clusters by chance; therefore care needs to be applied when setting k k
    - The potentially large number of clusters tested introduces the problem of multiple testing and Type I error
- References: (Besag and Newell 1991)

Getis-Ord localG spatial statistics

- Spatial data type: Areal
- Data needed: Counts of cases and population at risk aggregated by sub-region within study region, spatial neighbour matrix with spatial weights or distance-based measures
- How test works
  - Used for detection of localised clusters (in a diagnostic sense)
  - Measures the degree of association that result from the concentration of weighted points (centroids of areas) and all other weighted points within a predefined radius from the original weighted point

  - The local estimate of spatial autocorrelation is compared with a global average to identify 'hotspots'
- References: Getis and Ord (1992), Ord and Getis (1995)

Local Moran's I statistic

- Spatial data type: Areal data
- Data needed: Count of cases by sub-regions, a neigbourhood matrix with spatial weights
- How test works
  - The global Moran's I statistic is decomposed into local Moran's I values of spatial autocorrelation around each sub-region
  - A local indicator of spataital association (LISA) statistic is calculated for each zone based on the neighbour weights object
  - Indicates clusters of local autocorrelation (similar or dissimilar disease frequency)
  - May be used to investigate outliers in global spatial patterns by use of a 'Moran scatterplot'
    - X-axis represents the standardised local value for the sub-region
    - Y-axis represents the weighted average of the standardised neighbouring values
    - The location of the points in the 4 quadrats of the plot indicates the type of clustering mechanism for each point
      - Lower-left (low-low) and and upper-right (high-high) quadrats indicate clustering
      - Upper-left (low-high) and lower-right (high-low) are dissimilar to their neighbours
      - The slope of a linear regression line fitted to the data represents the Moran's I statistic
- References: Anselin (1995)

Extn. Ex. 6.4

```
## Localised non-focused spatial cluster detection- Local Moran's I statistic—- resI <- localmoran(x = smr$est, listw = viet_dist.02nb) hist(resI[,5]) # Distribution of p-values for significance of clustering around locations mean(resI[,1]) # Mean statistic moran.plot(x = smr$est, listw = viet_dist.02nb) rm(resI)
```

Extn. Q. 6.4 Interpret the findings from these tests in light of your other findings?

Kulldorff's spatial scan test

- Spatial data type: Areal or point
- Data needed: Polygons of areal units, counts of cases and either controls or population at risk, or dichotomous infection status
- How test works:
  - Theoretical circular window placed on map of all study locations (centroids of areal units)
  - A scanning window of increasing radius is placed around one of many possible centroids by sequentially aggregating nearest neighbour areas to create zones
  - Window radius may vary to a defined upper limit (up to 50% of study population recommended)
  - For each window the risk of disease is compared with that outside the window

  - If using case-control data, controls should be selected from same source population as cases
  - Significance measures estimated by Monte Carlo sampling
  - Data may be either Bernoulli (cases and controls) or Poisson (the number of cases and the population at risk)
  - Adjusts for heterogeneity of population at risk by indirect adjustment to calculate the expected number of cases for each location
  - Test may be used as complement to a global clustering method
  - The test can be used to detect clusters with increased, decreased or both increased and decreased incidence of disease
  - Reports most significant most likely and secondary clusters
- References: Kulldorff and Nagarwalla (1995), Kulldorff (1997)

Extn. Ex. 6.5

```
## Kulldorff’s spatial scan test (point data)—- library(smacpod) res<- spscan.test(com.ppp, case
= 2, nsim = 99, alpha = 0.05, maxd = 75, cl = NULL, longlat = FALSE) ## Collect results in
a dataframe for plotting xcoords <- c(res$clusters[[1]]$coords[1], res$clusters[[2]]$coords[1],
res$clusters[[3]]$coords[1], res$clusters[[4]]$coords[1], res$clusters[[5]]$coords[1]) ycoords
<- c(res$clusters[[1]]$coords[2], res$clusters[[2]]$coords[2], res$clusters[[3]]$coords[2],
res$clusters[[4]]$coords[2], res$clusters[[5]]$coords[2]) radius <- c(res$clusters[[1]]$r,
res$clusters[[2]]$r, res$clusters[[3]]$r, res$clusters[[4]]$r, res$clusters[[5]]$r) res.spscan.df
<- data.frame(xcoords = xcoords, ycoords = ycoords, radius = radius) g + geom_circle(aes(x0
= xcoords, y0 = ycoords, r = radius), data = res.spscan.df) rm(xcoords, ycoords, radius)
detach(package:smacpod) ## Kulldorff’s spatial scan test (areal data)—- library(smerc)
out.s = scan.test(coords = coords, cases = viet_dist_smr.shp@data$All_AI, pop =
viet_dist_smr.shp@data$No_commune, nsim = 49, alpha = 0.05, lonlat = FALSE) ## Number
of significant clusters identified length(out.s$clusters) ## 3 ## Plot clusters starting with
most significant id <- out.s$clusters[[1]]$locids col <- rep(“white”, times = 117) col[id] <- “red”
plot(viet_dist.shp, col = col, border = “#7f7f7f”, axes = FALSE) rm(id) ## Plot clusters starting
with next most significant id <- out.s$clusters[[2]]$locids col[id] <- “orange” plot(viet_dist.shp, col
= col, border = “#7f7f7f”, add = TRUE) ## Plot clusters starting with next most significant id <-
out.s$clusters[[3]]$locids col[id] <- “orange” plot(viet_dist.shp, col = col, border = “#7f7f7f”, add
= TRUE)
```

Extn. Q. 6.5 How are the results of the 2 tests similar or dissimilar? Give reasons for your answers.

Flexibly-shaped spatial scan test

- Spatial data type: Areal
- Data needed: Counts of cases and population at risk for each areal unit
- How test works
  - As for Kulldorff’s spatial scan statistic, except uses both circular and non-circular search windows by amalgamating areal units
  - Need to define maximum number of regions (< 30) to include in a cluster
  - The power to detect circular clusters is less than that of Kulldorff’s spatial scan statistic, but is able to detect non-circular clusters that cannot be captured by any circular window
  - Has a high positive predictive value for the detected significant most-likely cluster
- References: Tango and Takahashi (2005)

Extn. Ex. 6.6

Localised non-focused spatial cluster detection- Flexibly Shaped Spatial Scan Test—- ## Create binary spatial adjacency matrix from queen neighbours list object viet_dist.01nb.bm <- nb2mat(viet_dist.01nb, style = "B") out.f <- flex.test(coords = coords, cases = viet_dist_smr.shp@data$All_AI, pop = viet_dist_smr.shp@data$No_commune, w = viet_dist.01nb.bm, alpha = 0.05, k = 10) ## Number of significant clusters identified length(out.f$clusters) ## 6 plot(viet_dist.shp, col = color.clusters(out.f))

Extn. Q. 6.6 Is there agreement between the different tests on the same data? How would you interpret them?

Localised focused spatial cluster detection

Several methods are available in R packages for detecting localised and focused clusters of disease, but because these are less-frequently applied in animal health settings, only one is mentioned in detail. For further information, the reader is referred to Stone's test (Stone 1988 and implemented in DCluster::stone.test), Lawson-Waller test (Lawson 1993 and implemented in spatstat::berman.test),

Kulldorff and Nagarwalla's statistic for spatial clustering

- Spatial data type: Areal data, aggregated counts of cases and population for sub-regions, Bernoulli or Poisson data
- Data needed:
- How test works:
    - Pre-define a location for the point source of cluster and the maximum proportion of population in the study region to include in circular window of investigation
    - Data set created so that sub-region of interest is in the first row
    - Subsequent rows sorted in ascending order of distance to the centre of the sub-region of interest
    - Should not be used if point source is determined from the data itself
- References: Kulldorff and Nagarwalla (1995), Ward and Carpenter (2000)

Other methods for cluster detection

Temporal scan statistic

- Data needed: Count of cases by time
- How test works:
    - Originally proposed by Naus (1966) for use in stable population and analogous to spatial scan statistic
    - The test statistic is the maximum number of cases in a predefined "window" of time found by scanning all time series or that interval in the study
    - Generalised to account for temporal trends in the population size and incorporated in the SaTScan software
    - Test most sensitive when the scanning window of similar interval as duration of clusters
    - Recommended to set scanning window on basis of disease patterns, and subjectivity of this setting can affect test results

- References: Kulldorff (2018)

Dynamic minimum spanning tree scan test

- Spatial data type: Areal
- Data needed: Counts of cases and population at risk for each areal unit
- How test works
  - Similar to other scan methods
  - Identifies only non-overlapping clusters
  - Pre-define the proportion of the population and the maximum radius to consider in each zone
- References: Assuno et al. (2006)

12.2.3 Session information

The following code prints a summary of the operating system and versions of R and packages used in compiling this document. This information is useful for checking whether different package version numbers are responsible for any failure of the scripts to run or to debug other code errors.

sessionInfo() ## R version 3.5.1 (2018-07-02) ## Platform: x86_64-w64-mingw32/x64 (64-bit) ## Running under: Windows 7 x64 (build 7601) Service Pack 1 ## ## Matrix products: default ## ## locale: ## [1] LC_COLLATE=English_New Zealand.1252 ## [2] LC_CTYPE=English_New Zealand.1252 ## [3] LC_MONETARY=English_New Zealand.1252 ## [4] LC_NUMERIC=C ## [5] LC_TIME=English_New Zealand.1252 ## ## attached base packages: ## [1] stats graphics grDevices utils datasets methods base ## ## other attached packages: ## [1] kableExtra_0.9.0 readxl_1.1.0 ## ## loaded via a namespace (and not attached): ## [1] Rcpp_0.12.19 rstudioapi_0.8 xml2_1.2.0 ## [4] knitr_1.20 magrittr_1.5 hms_0.4.2 ## [7] munsell_0.5.0 rvest_0.3.2 viridisLite_0.3.0 ## [10] colorspace_1.3-2 R6_2.3.0 rlang_0.2.2 ## [13] httr_1.3.1 stringr_1.3.1 tools_3.5.1 ## [16] xfun_0.3 htmltools_0.3.6 yaml_2.2.0 ## [19] rprojroot_1.3-2 digest_0.6.17 tibble_1.4.2 ## [22] crayon_1.3.4 bookdown_0.7 readr_1.1.1 ## [25] evaluate_0.11 rmarkdown_1.10 stringi_1.1.7 ## [28] compiler_3.5.1 pillar_1.3.0 cellranger_1.1.0 ## [31] scales_1.0.0 backports_1.1.2 pkgconfig_2.0.2

References

Knox, E.G. 1989. "Detection of Clusters." In Methodology of Enquiries into Disease Clustering, edited by P. Elliot, 17–20. London School of Hygiene; Tropical Medicine, Biological Journal of the Linnean Society of London.

Besag, J., and J. Newell. 1991. "The Detection of Clusters in Rare Diseases." Journal of the Royal Statistical Society Series A-Statistics in Society 154 (1): 143–55. https://doi.org/%7B10.2307/2982708%7D.

Tango, T. 1999. "Comparison of General Tests for Spatial Clustering." In Disease Mapping and Risk Assessment for Public Health, edited by Andrew Lawson, Annibale Biggeri, Dankmar Bohning, Emmanuel Lesaffre, Jean-Francois Viel, and Roberto Bertollini. John Wiley & Sons, Ltd, Chichester, United Analyst (Cambridge, United Kingdom). https://www.ebook.de/de/product/4253210/lawson_disease_mapping_risk_assessment.html.

Ripley, B.D. 1977. "Spatial Patterns." Journal of the Royal Statistical Society Series B 39 (2): 172–212.

Diggle, P.J., and A.G. Chetwynd. 1991. "Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations." Biometrics 47 (3): 1155–63. https://doi.org/10.2307/2532668.

Waller, L.A., and C.A. Gotway. 2004. Applied Spatial Statistics. John Wiley & Sons. https://www.ebook.de/de/product/3606503/waller_gotway_applied_spatial_statistics.html.

Cuzick, J., and R. Edwards. 1990. "Spatial Clustering for Inhomogenous Populations (with Discussion)." Journal of the Royal Statistical Society Series B 52 (1): 73–104.

Oden, N. 1995. "Adjusting Moran's I for Population Density." Statistics in Medicine 14 (1): 17–26. https://doi.org/10.1002/sim.4780140104.

Moran, P.A.P. 1950. "Notes on Continuous Stochastic Phenomena." Biometrika 37: 17–23.

Cliff, A.D., and J.K. Ord. 1981. Spatial Processes: Models and Applications. Pion, London, UK.

Assuno, R.M., and E.A. Reis. 1999. "A New Proposal to Adjust Moran's I for Population Density." Statistics in Medicine 18 (16): 2147–62. https://doi.org/10.1002/(SICI)1097-0258(19990830)18:16<2147::AID-SIM179>3.0.CO;2-I.

Ward, M.P., and T.E. Carpenter. 2000. "Techniques for Analysis of Disease Clustering in Space and in Time in Veterinary Epidemiology." Preventive Veterinary Medicine 45 (3-4): 257–84. https://doi.org/10.1016/S0167-5877(00)00133-1.

Geary, R.C. 1954. "The Contiguity Ratio and Statistical Mapping." The Incoporated Statistician 5 (3): 115–45.

Tango, T. 1995. "A Class of Tests for Detecting 'General' and 'Focused' Clustering of Rare Diseases." Statistics in Medicine 14 (21-22): 2323–34. https://doi.org/10.1002/sim.4780142105.

Tango, T. 2000. "A Test for Spatial Disease Clustering Adjusted for Multiple Testing." Statistics in Medicine19 (2): 191–204. https://doi.org/10.1002/(SICI)1097-0258(20000130)19:2<191::AID-SIM281>3.0.CO;2-Q.

Getis, A., and J.K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." Geographical Analysis 24 (3): 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Ord, J.K., and A. Getis. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." Geographical Analysis 27 (4): 286–306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

Anselin, L. 1995. "Local Indicators of Spatial Association—Lisa." Geographical Analysis 27 (2): 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Kulldorff, M., and N. Nagarwalla. 1995. "Spatial Disease Clusters: Detection and Inference." Statistics in Medicine 14 (8): 799–810. https://doi.org/10.1002/sim.4780140809.

Kulldorff, M. 1997. "A Spatial Scan Statistic." Communications in Statistics - Theory and Methods 26 (6): 1481–96. https://doi.org/10.1080/03610929708831995.

Tango, T., and K. Takahashi. 2005. "A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters." International Journal of Health Geographics 4. https://doi.org/10.1186/1476-072X-4-11.

Stone, R.A. 1988. "Investigations of Excess Environmental Risks Around Putative Sources: Statistical Problems and a Proposed Test." Statistics in Medicine 7 (6): 649–60. https://doi.org/10.1002/sim.4780070604.

Lawson, A.B. 1993. "On the Analysis of Mortality Events Associated with a Prespecified Fixed Point." Journal of the Royal Statistical Society. Series A, (Statistics in Society) 156 (3): 363–77. https://doi.org/10.2307/2983063.

Naus, J.I. 1966. "Power Comparison of Two Tests of Non-Random Clustering." Technometrics 8 (3): 493–517. https://doi.org/10.1080/00401706.1966.10490382.

Kulldorff, M. 2018. "SaTScan- Software for Spatial, Temporal and Space-Time Scan Statistics." https://www.satscan.org/.

Assuno, R., M. Costa, A. Tavares, and S. Ferreira. 2006. "Fast Detection of Arbitrarily Shaped Disease Clusters." Statistics in Medicine 25 (5): 723–42. https://doi.org/10.1002/sim.2411.