25

ANALYSIS OF SPATIAL DATA: INTRODUCTION AND VISUALISATION

Chapter contributor: Javier Sanchez

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand the basic features of spatial data.
2. Differentiate the main types of spatial data used in veterinary epidemiology.
3. Describe the main approaches for visualising spatial data.
4. Generate visual presentations of different types of spatial data.

25.1 INTRODUCTION

The concept of space in epidemiology can be traced back to the 4th century BC; Hippocrates was one of the first to relate disease occurrence to the local environment. His book, *Air, Water and Places,* emphasises the importance of individual lifestyle, the role of wind, water sources and location of cities, and the influence these have on the presence and spread of disease.

It seems natural to think in terms of space when dealing with epidemiologic data. One of the main objectives in any epidemiologic investigation is the identification of patterns. While a disease pattern is typically the main interest, the distribution of other population characteristics might also be of interest when carrying out epidemiologic studies. Space is a basic component of the epidemiologic triad and it represents the physical location where hosts, disease agents and the environment interact.

A classical example of how geographical space might be related to the occurrence of disease is John Snow's investigation of cholera outbreaks in London in 1854 (see Chapter 1). That *Vibrio cholera* was the etiological agent of cholera was unknown at that time; however, Snow's previous investigations indicated that this microorganism might be transmitted by contaminated water. During the London outbreak, he observed that most of the deaths were in close proximity to a water pump on Broad Street—the closer to the pump people were, the higher the number of deaths were. He then produced a dot map showing the distribution of cases and the locations of water pumps in this area (for an overview of John Snow's disease mapping, see http://www.ph.ucla.edu/epi/snow/mapsbroadstreet.html).

Spatial epidemiology is a subdiscipline of epidemiology with the main focus being to describe and explain the spatial pattern of disease. It is also called environmental, landscape and geographical epidemiology. This branch of epidemiology introduces specialised tools to capture and analyse this type of data. Table 25.1 shows the main components of spatial analysis in epidemiology.

25.2 SPATIAL DATA

Specialised analytical software has been developed for managing and storing spatial data. A **geographic information system** (GIS) can be considered as a computer-based decision support system integrating specifically trained people, hardware, software and data where the main purpose is to store, analyse, present and disseminate geo-referenced data in an integrated environment. Despite the improved capacity of available hardware and software, there are still significant challenges involved in working with spatial data, including time spent collecting the information, the handling of the location of subjects such as animals that are frequently moving and finally, issues related to presentation of this information.

Spatial data usually consist of 2 components: **features** and **attributes** (Waller LA, 2004). Features include, for example, natural (*eg* rivers, mountains) and human-made entities (*eg* farms, roads). They have a geographical location (*eg* latitude/longitude, address), and are of a certain size, shape and have a specific orientation. The attribute (or tabular data) represents the measured values associated with each feature (*eg* herd size associated a farm). These data are usually stored in a database and are linked to the spatial data so they can be queried, joined and symbolised when producing maps in a GIS.

Table 25.1 Main components of spatial analysis in epidemiology

Spatial analysis	Object of analysis	Application
Disease mapping	Show variation in relative risk of a disease across a geographical area (map). Visual presentation of spatial variation in risk	Descriptive data analysis, predictive risk-mapping for informing risk-based surveillance, resource allocation, disease atlas construction
Ecological analyses	Geographical distribution of disease in relation to explanatory covariates, usually at an aggregated spatial level. Examine geographical variation in exposure to environmental variables in relation to health outcomes measured on a geographical scale	Focused on etiological questions (see Chapter 29)
Cluster detection	Assess whether disease occurrence is spatially clustered and where the clusters are located	Surveillance

25.2.1 Types of spatial data

Spatial information is generally defined as **points**, **lines**, **polygons** or **grids**. The data can be represented in **vector** or **raster** format.

Vector format
The vector format stores geographical features such as points, lines, and polygons as a series of X-Y coordinates and a feature identifier. Points, also called 'nodes', are represented by a pair of X-Y geographical coordinates. Lines, also called 'arcs', are represented by a segment between 2 points. Lines are commonly connected to form **networks** where the sequence and connections are as important as the geographical coordinates. Examples of lines are rivers, power lines, roads. Polygons are represented by a closed loop of coordinates (*eg* the starting node and ending node share the same X-Y coordinates) and they are described by a sequential list of nodes or **vertices**. Examples of polygons are census tracts, lakes, farm boundaries, land parcels, or watersheds. Fig. 25.1 depicts the 3 main features represented in vector format

Raster format
The raster format of spatial data consists of evenly distributed cells forming a grid organised into rows and columns like a spreadsheet. The space covered by each cell defines a spatial feature and information about attributes is stored for each grid cell. The relative position of each cell within the grid provides information about its geographical location. Grid data can be derived from scanned paper maps or obtained from arial photographs or satellite images. Fig.

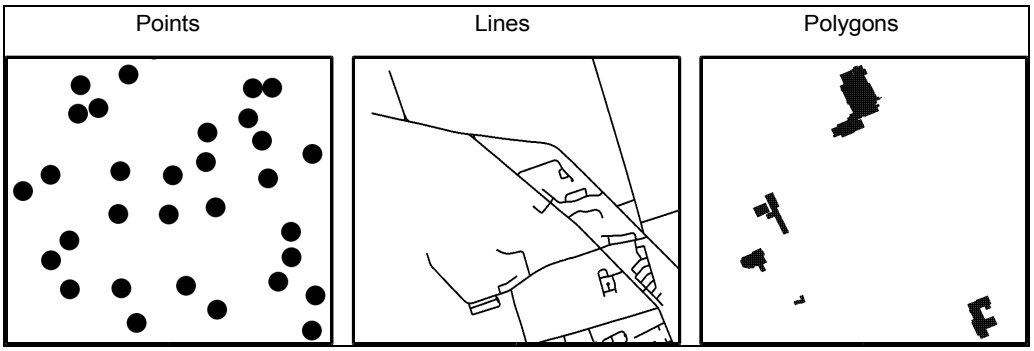


Fig. 25.1 Types of spatial data

25.2 shows a comparison of how a river can be represented using vector and raster format. When using a vector format, the river is represented as a line; an identifier (ID) is assigned to that line and each pair of geographical coordinates from that line will share the same ID. On the other hand, the raster format is represented by cells in a grid. Each cell has a unique geographical location and a value associated with it describing the relevant characteristics. In this example, a value of 1 is used to represent the location of the river, and 0 otherwise.

The vector format is suitable for describing discrete features such as rivers, roads, and buildings. It is less useful for spatial characteristics that vary continuously—soil type, vegetation index, or land use. The vector format allows a more accurate description of the location of a geographical feature than the raster format. On the other hand, raster data can be manipulated more quickly by the computer, but they are not as flexible as vector data for storing multiple attributes. The degree of detail that can be represented with the raster format is dependent on the size of the cell. The smaller each cell, the more accurately geographical features will be represented, but larger cells require data storage and processing. Fig. 25.3 compares 3 raster representations showing different cell sizes overlayed with a vector format of the same feature.

Veterinary epidemiologic data used for spatial analysis are usually represented as **discrete** or

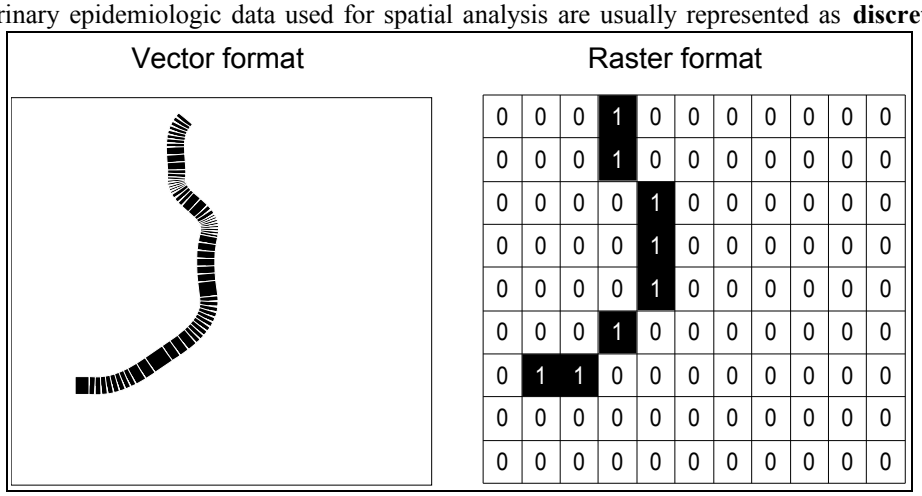


Fig. 25.2 Comparison of representation of spatial data using vector and raster formats for a line feature (eg river)

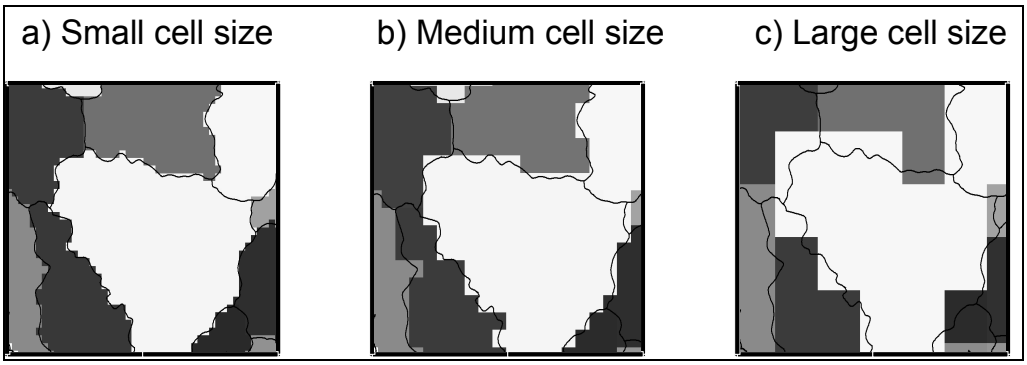


Fig. 25.3 Comparison of level of detail represented by different cell sizes using grid data

continuous geographical features. Discrete features represent the space filled with objects and each object is associated with potentially several measurable attributes. The main types of discrete feature are points and polygons. On the other hand, for continuous features, the (usually) single attribute associated with each geographical location varies continuously throughout space. For instance, the locations of diseased animals or infected farms represented as points or polygons are discrete features; whereas soil mineral levels or rainfall values are examples of continuous features because they vary continuously over space.

In the continuous field view, the main interest is in the representation of the continuity of the spatial phenomena. Usually, a finite number of locations are sampled and then a mathematical function is applied to predict the value of the attribute of interest (*eg* soil mineral) for all the non-sampled locations belonging to the study area.

The main characteristic of spatial data analysis is that the geographical location associated with each attribute is taken into account in order to understand the spatial relationships between these locations. Table 25.2 summarises the main spatial features used in veterinary epidemiology.

25.3 Spatial data analysis

Epidemiologic phenomena can be represented spatially using the discrete or continuous views as described above. The remainder of this, and the following, chapter is based on the approach to spatial data analysis suggested by Bailey and Gatrell (1995)—visualisation, exploration and modelling. Visualisation refers to the representation of the data being analysed as maps; exploration seeks to summarise and identify spatial patterns and relationships; and modelling relies on the specification of a statistical model and estimation of the parameters. It is acknowledged that there is some degree of overlap; for instance, some of the methods used in exploration require the visualisation of the data as maps and results of statistical methods for estimation of parameters also can be presented as maps. The following sections will discuss the visualisation methods used for discrete (point and area) and continuous spatial data. Chapter 26 will deal with exploration and modelling.

Table 25.2 Spatial data in veterinary epidemiology

Data type	Location	Attribute	Storage format*
Point	Locations themselves are the variable of interest and a dataset consists of a finite number of locations within a defined study region	Recorded at each location. Point locations with associated attribute values are known as marked point processes	Set of X-Y coordinates in vector format
Polygon (area or lattice)	Discrete areas, regular or irregular in shape. Areas may be spatially referenced by an adjacency matrix defining each area's proximity relationship with other areas	Counts of events within each area, or the value of some discrete or continuous variable defined in the area as a unit	Set of X-Y coordinates in vector format Grid cells with attribute value associated representing a polygon class in raster format
Continuous	All possible locations throughout a study area	Vary continuously across space. A finite number of locations are sampled to estimate the value of the attribute at any position within the study area (for example rainfall recorded at monitoring stations)	Grid cells with associated value representing the attribute of interest in raster format

Adapted from Stevenson (2003)

*Most common storage format in GIS

25.3.1 Visualisation

Once the spatial data have been entered and processed, standard tables and graphs can be produced to examine the results. In addition, 'maps' can be produced as a special type of output for this kind of information. The science and art of map-making is called **cartography** derived from the Greek *chartis (*map) and *graphein* (write). A map is a simplified depiction of a space which highlights relationships between components of that space. Specific details about map scale, projections, and symbolisation are beyond the scope of this chapter. Interested readers can refer to Brewer (2006); Kennedy & Kopp (2000); Monmonier (1996);and Waller (2004).

Data visualisation is now a commonly used spatial analysis method for supporting surveillance and monitoring activities of animal health. This involves the creation of maps showing the spatial and temporal pattern of diseases which are often used purely for describing spatial patterns or they may be used for generating casual hypotheses of potential risk factors of the disease.

This chapter, and Chapter 26, will use a dataset that represents the locations of avian influenza (AI) outbreaks between 2004 and 2006 in a northern region of Vietnam. It was extracted from a larger dataset for the whole country. Further information about these data are presented in Chapter 31 and a more detailed description of the dataset can be found in Pfeiffer *et al* (2007). The data consists of communes' centroid point locations and a dichotomous variable for each commune indicating whether at least one AI outbreak had occurred during the above time period. The objective of the analysis was to provide a more easily interpretable map of the spatial heterogeneity of AI outbreaks in this part of Vietnam. Fig. 25.4 shows a map of Vietnam highlighting the study area.

Fig. 25.4 Map depicting study area in Vietnam

Visualising point patterns

Point patterns refer to discrete data consisting of a series of point locations, $s_1, s_2 ... s_n$, in some study region R, at which events have occurred. It is important to mention that if the process we are working on relates to the occurrence of events, the pattern we are interested in is the 'event location'. This should be distinguished from where the interest is in the patterns of 'attribute values' between point locations across the study region (*eg* number of diseased animals or rainfall values). The objective of visualising spatial point patterns is to examine whether the spatial pattern of the observed events exhibits any systematic structure, as opposed to being distributed at random. Fig. 25.5 shows the 3 possible distributions of point patterns (random, clustered and regular).

Dot maps are the most commonly used type of visual representation for geographically referenced point data. They are used for showing any pattern related to the distribution of events (*eg* case and control farms). Events in a dot map are often labelled by using different colours and symbols for displaying different types of event or occurrence of the same event during different periods of time. Example 25.1 shows a dot map of the infection status of the communes in the study area during the study period. In general, this method will only be useful if the number of points is small and the points are not too densely clustered.

Visualising aggregated spatial data

Spatial data are called aggregated when geo-referenced epidemiological information is summarised at some geographical entity which can sensibly represent other entities associated with it, such as farms within a province to reflect, for instance, the total number of farms in a county, prevalence of disease per community *etc*. This information can be represented using both discrete point and polygon data.

Point data

Aggregated spatial point data can be shown as a 'proportional dot map'. The size of each point location on the map is proportional to the values of the attribute of a variable (*eg* number of AI outbreaks).

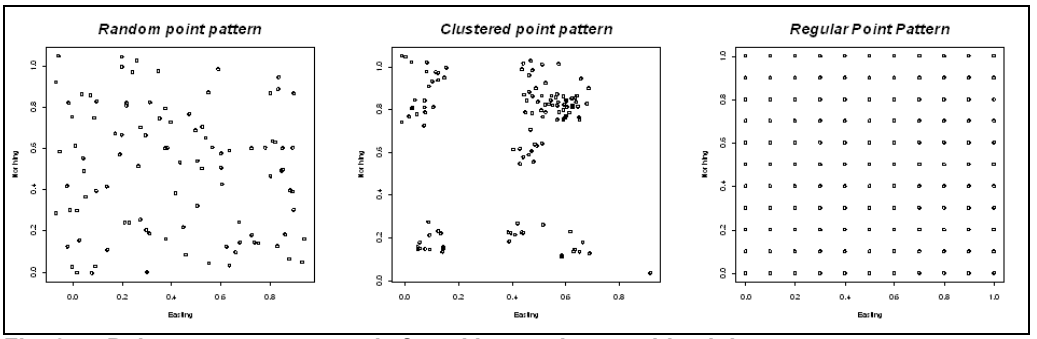


Fig. 25.5 Point patterns commonly found in veterinary epidemiology

Area or polygon data

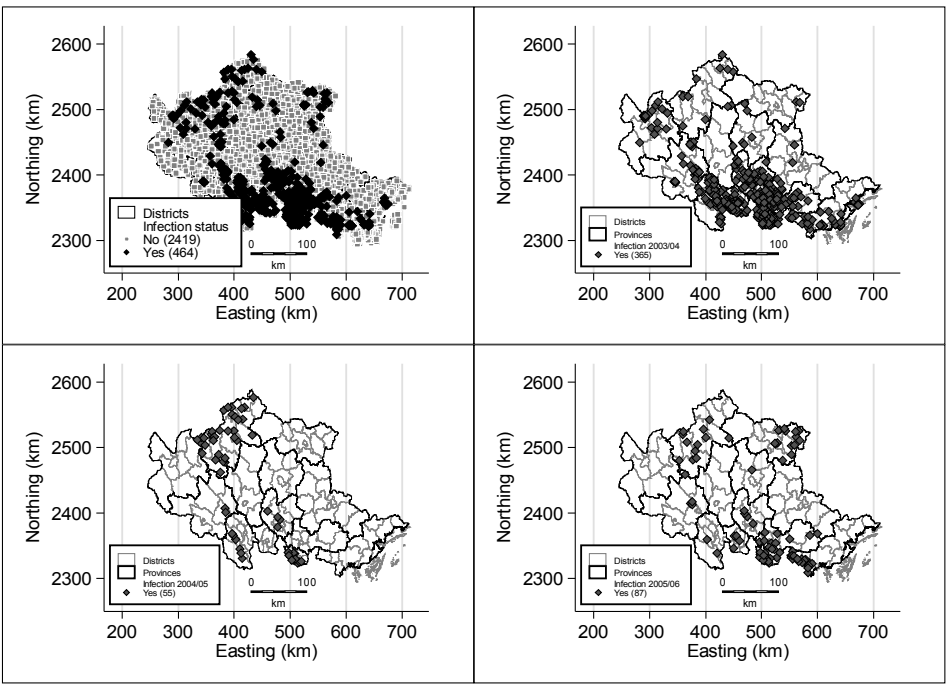
Plotting aggregated data is one of the most commonly used methods for visualising epidemiologic events. These maps are called **choropleth maps** where the space within a given polygon is shaded according to a colour scale representing the value of the attribute of interest usually at some level of aggregation. Two important issues need to be considered when producing this kind of map: 1) the number and width of class intervals and choice of colour palette, and 2) the **modifiable area unit problem** (MAUP).

The number and width of the class intervals will strongly affect the visual impression of the map and different approaches haven been proposed to determine the number of intervals or classes used to represent the attribute values. It has been suggested that class intervals should be selected according to the overall shape of the distribution and not the statistical frequency distribution. Evans (1977) and Pfeiffer, Robinson, Stevenson, Stevens, Rogers, Clements (2008) discussed in detail different approaches for defining class intervals that can be used for mapping. The most commonly used methods for selecting class intervals, which are implemented in many mapping and GIS softwares, will be mentioned in this chapter.

- **Manual breaks** Values are grouped according to cut-offs that the analyst considers to be meaningful.
- **Natural breaks** (Jenkins' method) Natural groupings in the data are automatically identified by a software algorithm that aims at maximising similarity within and differences between groups (*ie* it looks for big jumps in data values).
- **Quantile breaks** Classes are determined using the relevant values from the percentiles generated from the underlying data, so that each interval has approximately the same number of observations. This method can be misleading, as the breaks will result in a visually exaggerated difference between values that are close to either side of the break value.
- **Equal interval breaks** This method divides the difference between the largest and smallest attribute value into equal intervals. This method will not take account of the distributional characteristics of the underlying data, and should therefore only be used when the data are well understood.
- **Cluster breaks** Class intervals are defined using an iterative process that partitions the data into 'k' groups or clusters using standard cluster analysis algorithms.
- **Standard deviation breaks** The intervals are based on the number of standard deviation units around the overall mean, so the scale is in standard deviation units. Also, class intervals are defined as a proportion of the standard deviation, where the width of class

Example 25.1 Spatial visualisation of point data on AI outbreaks in northern Vietnam
data = Vietnam (District-d.dta, District-c.dta, Commune-d.dta)

AI infection status of 2,883 communes from 2004–2006. a) Overall distribution of infected and non-infected communes for all years (upper left). b) Infected communes in 2003/04 (upper right) c) Infected communes in 2004/05 (lower left) and d) Infected communes in 2005/06 (lower right).



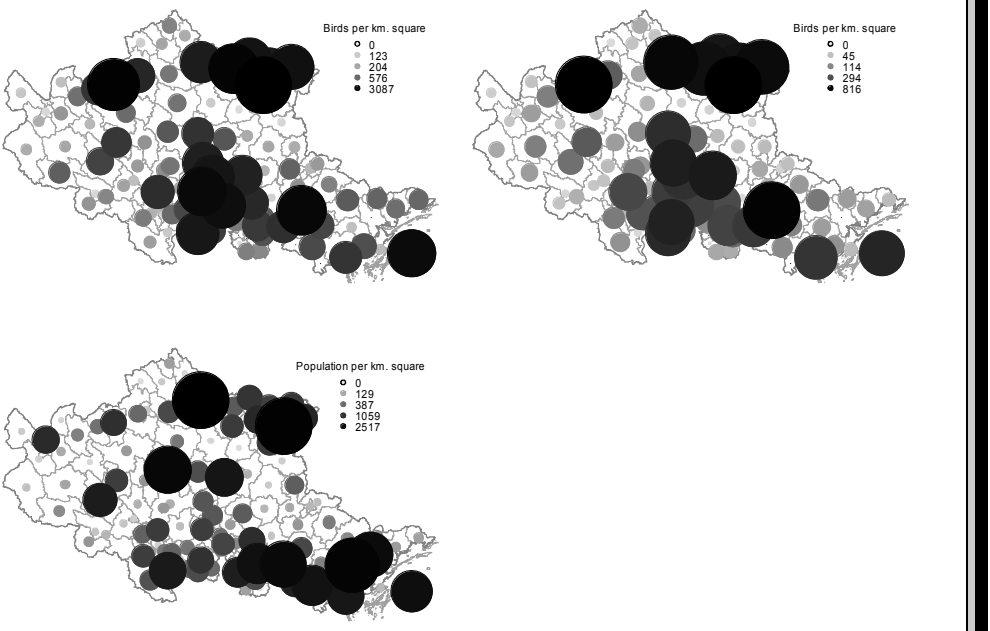
intervals is a fraction of the standard deviation, and this proportion depends on the number of intervals. This method, therefore, will indicate how far a local value is away from the overall mean value.

Regarding colour schemes, it is recommended to use continuous shading or gradation of colour to represent continuous values. For instance, light to dark colours can be used to represent low to high prevalence values. On the other hand, different hues such as yellow-green-blue, can be used to plot adjacent categories (*eg* land use) and contrasting colours for mapping diverging results (*eg* positive and negative regression residuals). A very useful resource for colour mapping for representing area data is the Colorbrewer website (http://www.colorbrewer.org). (**Note** Many of the figures in this chapter were generated, and are more informative when viewed, in colour—however, this book is printed in black and white!) Example 25.3 presents choropleth maps of the Vietnam outbreak data.

The MAUP results from the fact that the results of the analysis depend on the level of resolution used. This is mainly due to the fact that geographical areas usually represent arbitrary limits

Example 25.2 Chicken, duck and human densities in northern Vietnam
data = District-d-d.dta, District-c.dta

The proportional dot maps below show chicken (upper left), duck and human population densities (number of heads/km^2) for each commune in the study region. It indicates that the densities of chickens and ducks follow a similar pattern, with higher densities in the northwest and south-central communes. However, the human population appears more concentrated in the southwest (near Hanoi).



(political boundaries) and do not usually follow the spatial phenomena to be represented and larger areas tend to dominate the display and these areas tend to be sparsely populated giving a false impression of the magnitude of the event. Bailey and Gatrell (1995) recommend that spatial data be analysed at the smallest area unit available rather than aggregating the data using larger units.

An approach used infrequently to deal with the MAUP is a **cartogram**. With this type of map, the shape and size of the geographic feature (*eg* provinces) are distorted, typically proportionally to some attribute value (*eg* population size). Various mathematical algorithms have been developed to produce these maps. The method proposed by Gastner and Newman (2004) is called **diffusion cartogram.** This is commonly used and has been implemented in freely available tools (http://chorogram.choros.ch/scapetoad). Briefly, the method involves first overlaying a regular grid over the study region, then it computes the density of the attribute and rescales the size of each region proportional to this value but constrained by the boundaries of the study size. The overall shape of the cartogram will depend on the grid size used to produce

the equalised-density map. A very fine grid will produce a map with substantial local distortion in areas with high values of the attribute. A coarse grid will produce a map where areas are easier to recognise, but it will give a less accurate impression of the distribution of the attribute of interest. Once the map with the modified shapes has been obtained, then a choropleth map can be produced as described above. This kind of map helps to visualise the relationship among two variables plotted at the same time over the study region. Because it might be difficult to recognise the areas due to their distorted shape and size, it is useful to accompany the map by the undistorted version, and when showing it in computer software ideally should be dynamically linked with the cartogram map. Thus, careful attention should be given to these issues, given that different criteria to choose the class intervals, colour schemes and level of resolution can produce very different interpretations. Example 25.4 uses a cartogram to present the Vietnam AI outbreak data, adjusted for chicken density.

Visualisation of spatially continuous data

With this particular type of spatial data, the attribute value is a reflection of a spatially continuous process. In other words, it is possible to obtain a value for the attribute of interest at every geographic location across the study region. Generally, these values are obtained from a finite number of sampling locations and then interpolation methods (such as trend surface or spline regression, or kriging) are applied to generate a smooth surface for the data. This type of data includes soil mineral values, precipitation, temperature, elevation, air pollution *etc*. When visualising this type of data using only the values obtained at sample locations, the approach described above for showing variation in attribute values of point data can be used, where the size of the point represents the value of the measurement at each sampling location. Alternatively, interpolated values can be presented in 2 dimensional (2D) or 3 dimensional (3D) maps. Example 25.4 presents the distribution of elevation in Prince Edward Island as an example of cartograms.

25.4 ADDITIONAL TOPICS

25.4.1 Spatio-temporal data

Although not mentioned in this chapter, spatial data often has time as one of the attribute values, and this information is usually used to generate a series of maps using the methodologies described in this chapter. For instance, Example 25.1 shows point pattern maps of the number of communes with AI outbreaks between 2004 and 2006. Similar maps can be produced to represent spatio-temporal aggregated data.

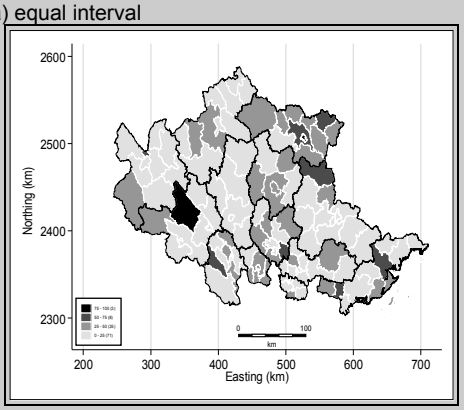
25.4.2 Dynamic visualisation of spatial data

Aggregated spatial data can be also visualised using a ‘dynamic’ link between the geographic data and the attributes of interest using standard graphical outputs (eg histograms). As described by Pfeiffer *et al* (2008), a method called ‘brushing’ can be used to interactively explore the spatial and statistical distribution of an attribute variable. This technique is available in the freely available GeoDA software (https://www.geoda.uiuc.edu/); Example 25.6 contains one screen shot from this program.

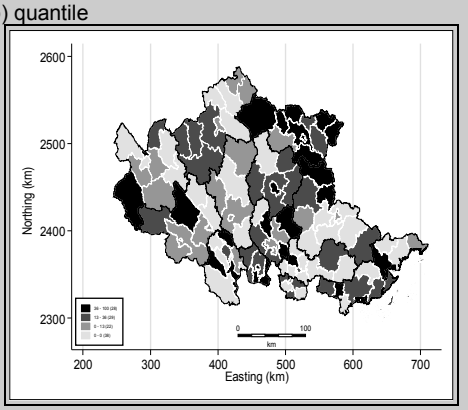
Example 25.3 Choropleth maps of AI outbreaks in northern Vietnam
data= District-d.dta, District-c.dta, Prov-c.dta

The following maps depict the prevalence of the AI outbreaks for each district. Three methods for defining class boundaries were used to generate these maps: a) equal-interval, b) quantile, and c) cluster method. In each map, darker colours represent higher prevalence values. The equal interval method (map a) highlights 3 districts with high prevalence values (*eg* greater than 75%), however many districts have prevalence values ≥ 25%, and from this map the main impression is that there are few districts with high prevalence. This is mainly due to the colour shading and scale used. In contrast, the quantile method (map b) provides a better representation of the distribution of the AI prevalence in the study region and is very similar to the class intervals obtained by the cluster method (map c). It appears that the cluster method produces an intermediate pattern between the equal interval and quantile maps. Maps a and c seem to indicate that a group of districts from the provinces in the central region (on the east-west axis) of the study region have higher prevalence than the remaining ones; however, there are individual districts across the study region with high values as well.

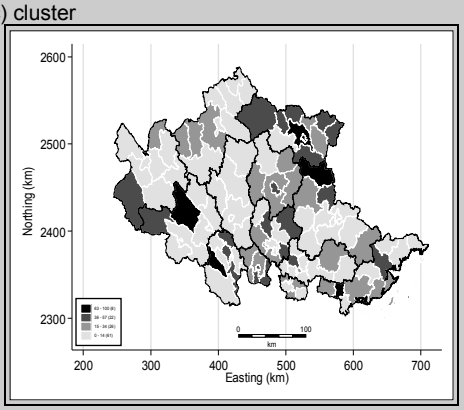
a) equal interval



b) quantile

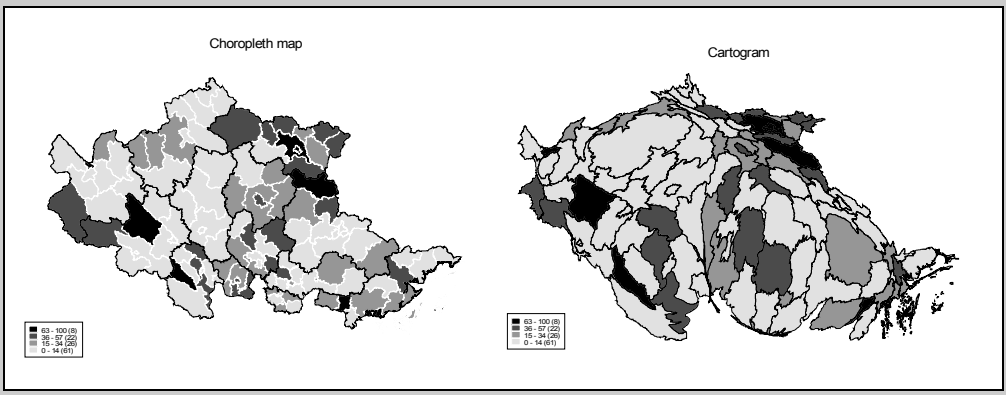


c) cluster



Example 25.4 Choropleth and cartograms for visualising spatially continuous data
data= District-d.dta, District-c.dta, Prov-c.dta, District_cart-d.dta, District_cart-c.dta

The following maps show the proportion of communes for each district with at least one AI outbreak during 2004-6, first as a standard choropleth map (a) and then as a cartogram (b) where the resized and shaped district boundaries are proportional to the chicken density in the district. It can be concluded from these maps that the districts with a large density of chickens do not tend to have the highest prevalence of communes with at least one AI outbreak. Although the shape of the study region is somehow distorted, the cartogram helps in visualising the spatial distribution of the attribute of interest and of its relationship with another variable, in this case the spatial distribution of the chicken density.



Example 25.5 Spatially continuous data on land elevation in Prince Edward Island

Two methods of presenting maps reflecting the land elevation (above sea level) for Prince Edward Island. Data were obtained from the website http://www.geobase.ca and the 3D map created using the 3D Analyst extension in ArcView 3.3. (ESRI, Redlands, CA, USA) .

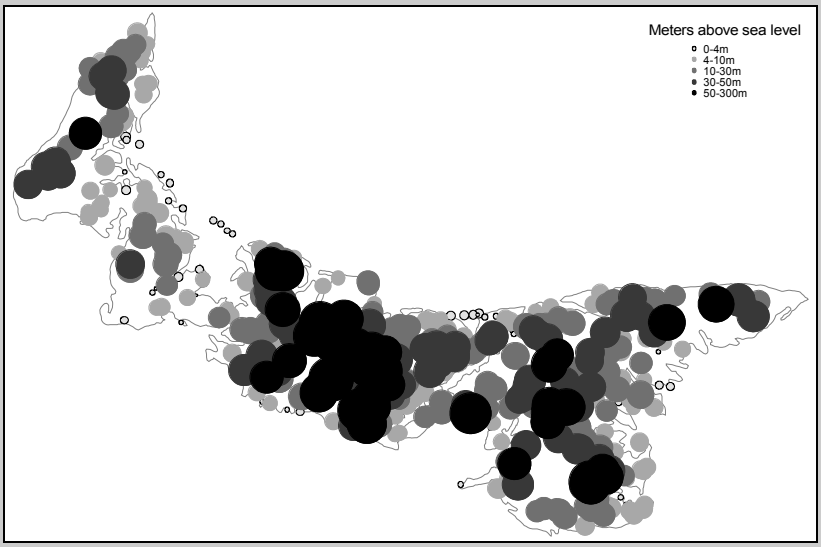


Fig. 25.6 Proportional dot map of PEI elevation

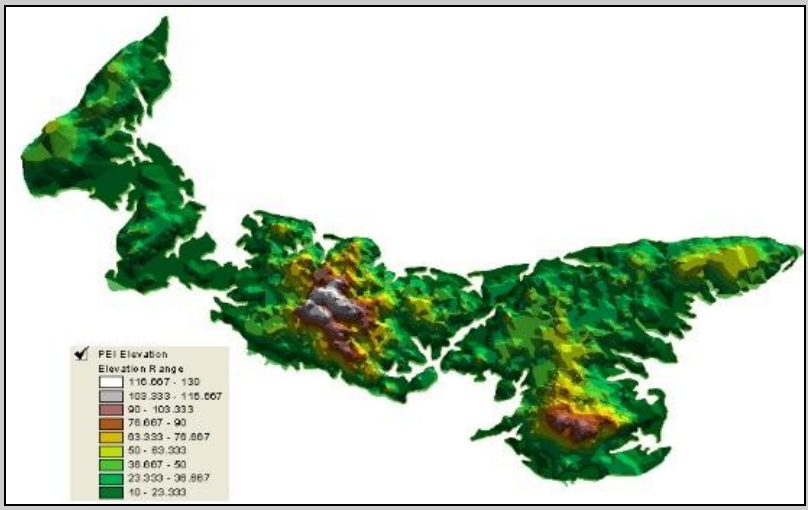


Fig. 25.7 Three-dimensional map of PEI elevation

Example 25.6 Dynamic visualisation of the upper 65th percentile distribution of AI in communes in northern Vietnam

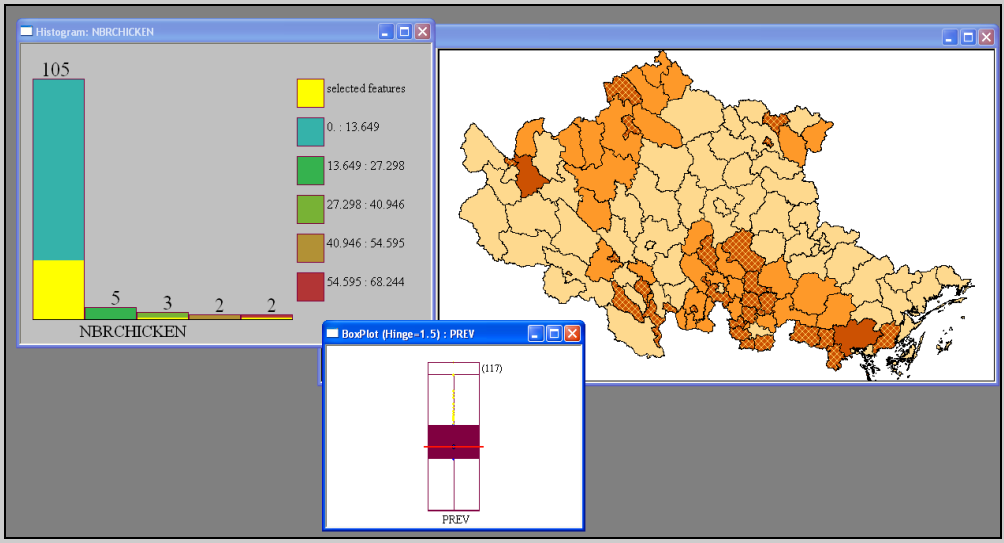


Fig. 25.8 Dynamic display of Vietnam data

It is impossible to adequately present a dynamic display within a static figure; this image is provided as an example only. It presents the location of districts where the number of communes with at least one outbreak of AI is in the 75 percentile (interactively selected in the boxplot). The histogram presents the frequency distribution of the number of chickens per district and marked in yellow the distribution of these same values for a subset of selected districts. The choropleth map depicts the distribution of the prevalence of AI by district (gridded districts represent the selected districts). The output was generated by the GeoDA software (https://www.geoda.uiuc.edu/).

References

Armstrong MP, Xiao N, Bennett DA. Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps Annals of the Assoc Am Geographers. 2004; 93: 595-623.

Bailey TC, Gatrell AC. Interactive Spatial Data Analysis. Longman; Harlow, UK. 1995.

Brewer CA. Basic mapping principles for visualizing cancer data using Geographic Information Systems (GIS) Am J Prev Med. 2006; 30: S25-36.

Evans I. The selection of class intervals Transactions of the Institute of British Geographers. 1977; 2: 98-124.

Gastner MT, Newman MEJ. Diffusion-based method for producing density-equalizing maps Proc Natl Acad Sci U S A. 2004; 101: 7499-504.

Kennedy M, Kopp S. Understanding map projections. Environmental Systems Research Institute, Inc; Redlands, CA, USA. 2000.

Monmonier M. How to lie with maps. The University of Chicago Press; Chicago, USA. 1996.

Pfeiffer D, Robinson T, Stevenson M, Stevens K, Rogers D, Clements A. Spatial Analysis in Epidemiology. Oxford University Press; Oxford. 2008.

Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ. An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data Vet J. 2007; 174: 302-9.

Stevenson M. The spatio-temporal epidemiology of BSE and FMD in Great Britain. [PhD Thesis]. Massey Univ.; Palmerston North, New Zealand. 2003.

Waller LA. Applied spatial statistics for public health data. John Wiley & Sons, Incorporated; Hoboken, NJ, USA. 2004.