Advanced GIS Virtual Training- Investigation of spatial clustering of disease

Chris Compton, EpiCentre, School of Veterinary Science, Massey University

August 2021

Table of Contents

1 Background

Epidemiologists frequently investigate possible clusters of disease to provide information on its possible causes and for its control and prevention. An 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. These factors should be considered 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 of contaminants or toxins on public health, for example, aerial deposition of particles downwind 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.

2 Reasons for studying disease clustering:

- In an epidemiological study of the 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 of the disease in the study. This is because the distribution of cases alone may just reflect the underlying distribution of the population at risk, which needs to be taken into account to calculate the actual incidence or risk of disease.

2.1 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 for by the known variables.

2.1.1 Population surveillance

The detection of clusters of disease with known risk factors may prompt a government-led health response, where those factors can be controlled, for example from water-borne diseases or environmental pollution, or an effective vaccination programme in the face of an infectious disease. 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.

2.2 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" Lawson and Kuldorff (1999)
  - or less formally- an area within the study region of significant elevated risk of disease
  - also known as a 'hot-spot'

2.3 Spatial 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 pattern of occurrence of a variable across the study region
- A *second-order* effects describes small-scale variation due to interactions between neighbours

3 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) as:

- General, global or non-specific clustering detection methods
  - Investigate the overall clustering tendency of a disease in a study region, regardless of where that clustering is located
  - Are analogous to the assessment of autocorrelation
  - Assess the overall or global aspects of clustering
  - Do not estimate the spatial locations of clusters

  - Assume 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 pre-determined 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 modelled through the use of random effect regression models.

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.

A range of statistical methods are used to investigate and describe spatial clustering of disease. However, in this course we will only consider visual exploratory spatial data analysis (ESDA) methods and one statistical method to investigate localised non-focused spatial clusters of disease

3.1 Exploratory spatial data analysis methods for clustering of disease

The first step in an investigation of spatial patterns of disease occurrence is exploratory data analysis (ESDA) of disease occurrence, initially by simply creating a map of disease case locations. Such a map provides readily-understood information of the overall distribution of disease in the study area, and in particular, any variation in the number of cases between regions within the study area to locate where the main disease burden lies. When there are a large number of cases of disease and point locations may be plotted over one another, an additional GIS step to aid interpretation of the map is to smooth the distribution of cases by kernel-smoothing methods

A second ESDA step is to plot the distribution of both case and non-case locations to gain an impression of whether the distribution of cases merely reflects the distribution of the

underlying population at risk rather than an actual local increase in risk compared to other regions in the study area. However, the visual appearance of maps drawn in this way may be difficult to interpret and is subject to user-defined parameters such as colour and size of points on the map, and particularly the diameter of areas over which the kernel-smoothing function is set to operate on.

3.1.1 Exploratory analysis for spatial clusters of FMD in Myanamar with QGIS

This section continues that from the earlier teaching material on SRA for incursion and spread of FMD in Myanmar. Some of the spatial files created in that teaching will be reused in this section.

The authors acknowledge the following data sets provided by the Livestock Breeding and Veterinary Department (LBVD), Myanmar from the NZ MFAT-OIE SEACFMD project for the use of the two data sets used in this section-

1. The locations and administrative regions of FMD outbreak villages between March 2015 and Feb 2016, provided in the file "all_outbreaks_rev.csv"
2. The locations and administrative regions of all villages in Myanmar in the file "mmr_crty-level_pplp2_250k_mimu_Jan_2018_V0.csv"

3.1.1.1 Set up a new project with required data files

1. Create a new QGIS project as for MMR FMD Spatial Risk Assessment and save the QGIS project file in the same directory as previously
2. All but one of the raw data files for this analysis were already used in the previous SRA exercise and are located in subfolders with the "….." folder, and the new results files for this project can be added to subfolders within the existing "…" folder
3. Copy file of village locations for Myanmar (mmr_crty-level_pplp2_250k_mimu_Jan_2018_V0) to "RawData-Features" folder and map these point locations
    1. Click "Open Data Source Manager" on Toolbar Menu -> Dialog box … (Figure 3.1)
        1. Select "Delimited Text"
        2. In file name select the above file
        3. Check Geometry Definition is correct
        4. Click "Add"

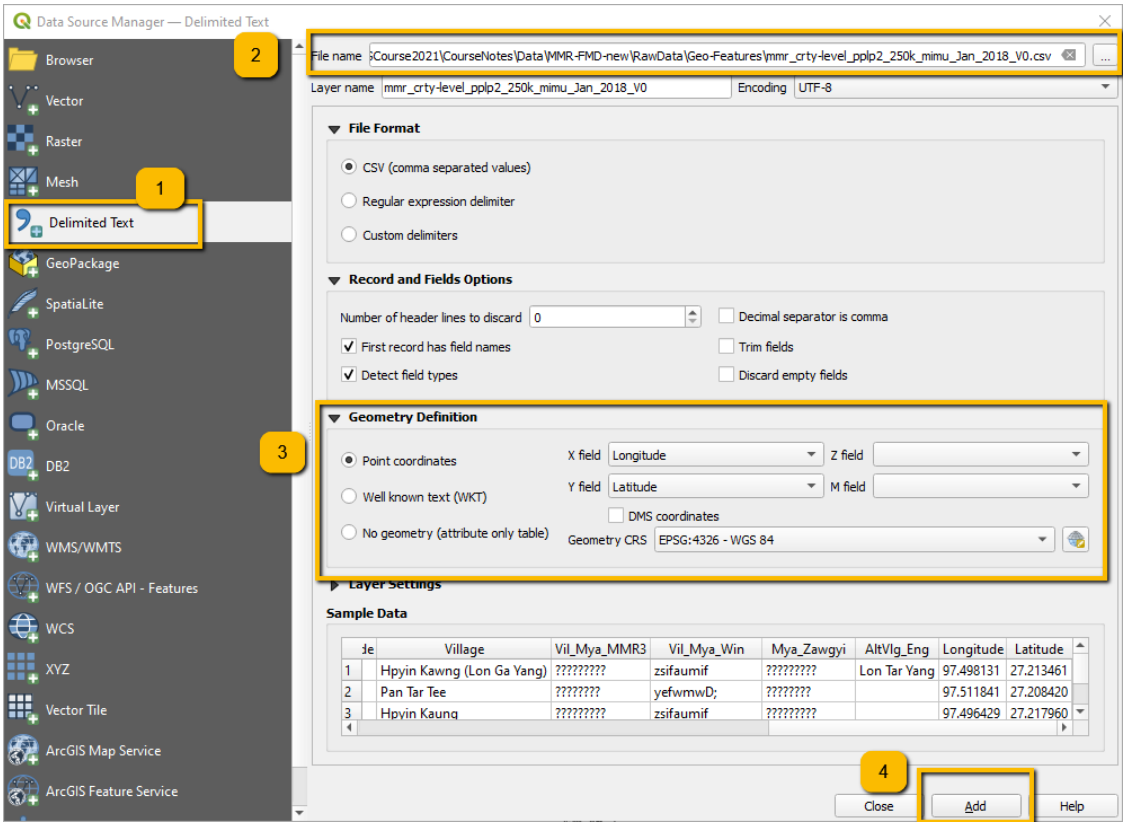

Figure 3.1: Create spatial file of Myanmar village locations

4. Reproject and save the file to the Project CRS
    1. Right click mmr_crty-level_pplp2_250k_mimu_Jan_2018_V0 in Layers Pane -> Export -> Save Features As … -> Dialog box (Figure 3.2)
        1. Click folder button to far right of File name and Select “ResData-Features” and name file “MMRVillages”
        2. In CRS combo box select “Project CRS:32646 - WGS 84 / UTM zone 46N”
        3. Click “OK”

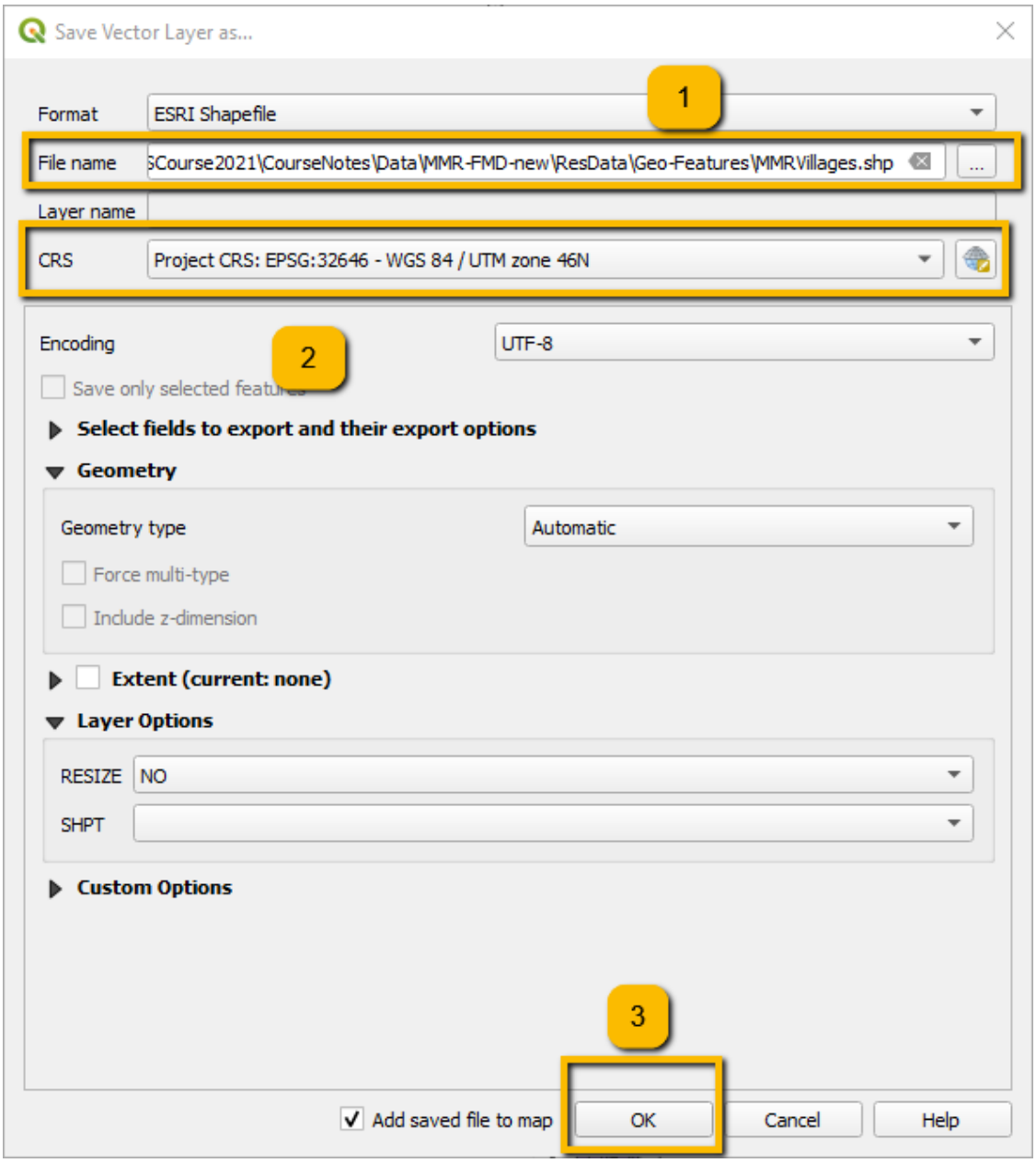

Figure 3.2: Save projected spatial file of Myanmar village locations

3.1.1.2 Map FMD outbreak villages only

1. Create a map of the study area
    1. From the Browser pane, drag the projected Myanmar country boundary map file (MMR_0) from the previous SRA project onto the map palette
2. You may also add a map of further administrative level regions to inform of the regions with the greatest (and least) number of case villages
    1. From the Browser pane, drag the projected Myanmar level 2 boundary map file (MMR_2) onto the map palette
3. Add map layer of FMD case village locations

1. Drag file “all_outbreaks_rev.shp” from “ResData-Features” folder created in the previous SRA project onto the Map palette

3.1.1.3 Create a kernel-smoothed density map (heatmap) of FMD outbreak villages only

1. Highlight “all_outbreaks_rev” in the Layers Pane
2. In Processing Toolbox search bar type “kernel density” and select “Interpolation -> double click”Heatmap (Kernel Density Estimation) -> Dialog box (Figure 3.3)
    1. Set radius to 100000 (100 km) to begin with (this setting is subjective only)
    2. Set number of rows in output to 200 (the other boxes self-fill)
    3. Click “Run”
    4. Click “Close”

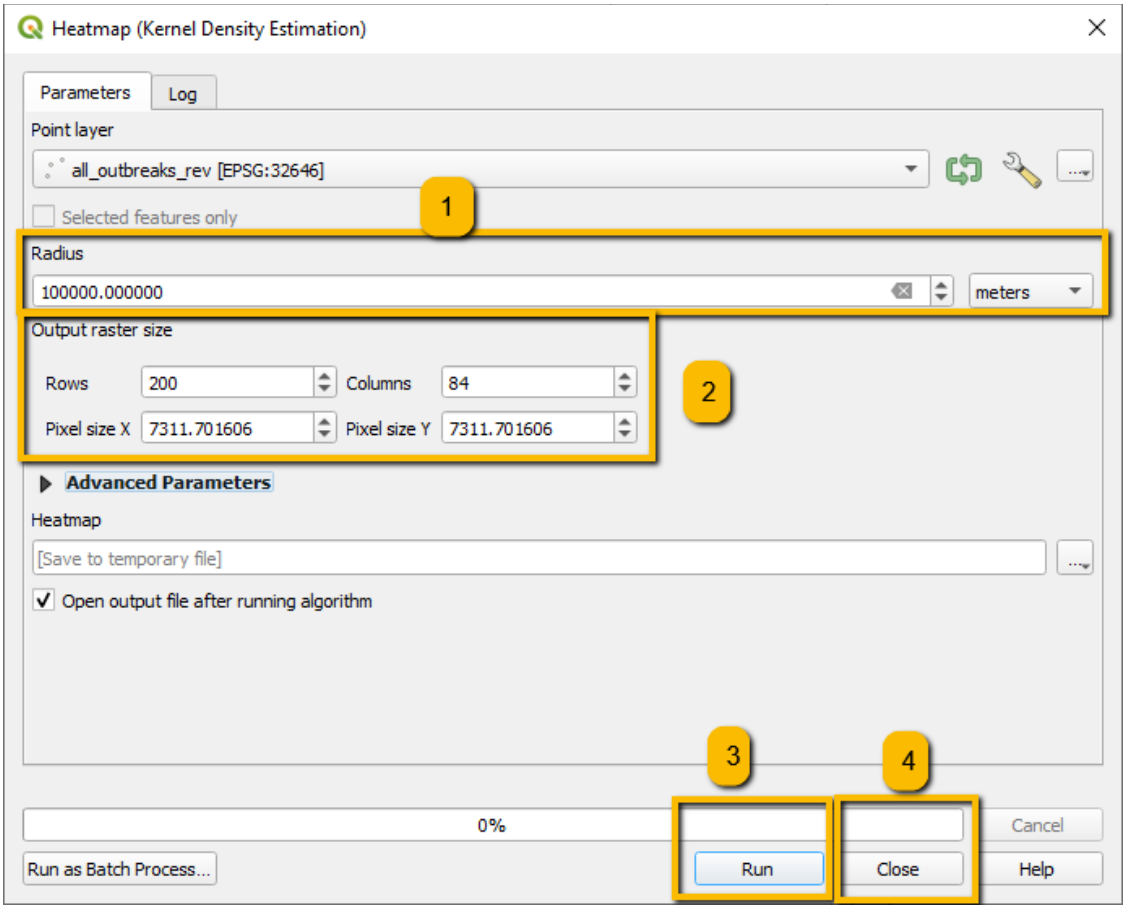

Figure 3.3: Create heat map of FMD outbreak villages in Myanmar

3. Save your heatmap as raster layer
    1. Right-click “Heatmap” in the Layers Pane -> Export -> Save As … -> Dialog box (Figure 3.4)
        1. Select folder “-Features” and create a new name for the file “Outbreaks_Heatmap” in the GeoTIFF format
        2. Click OK

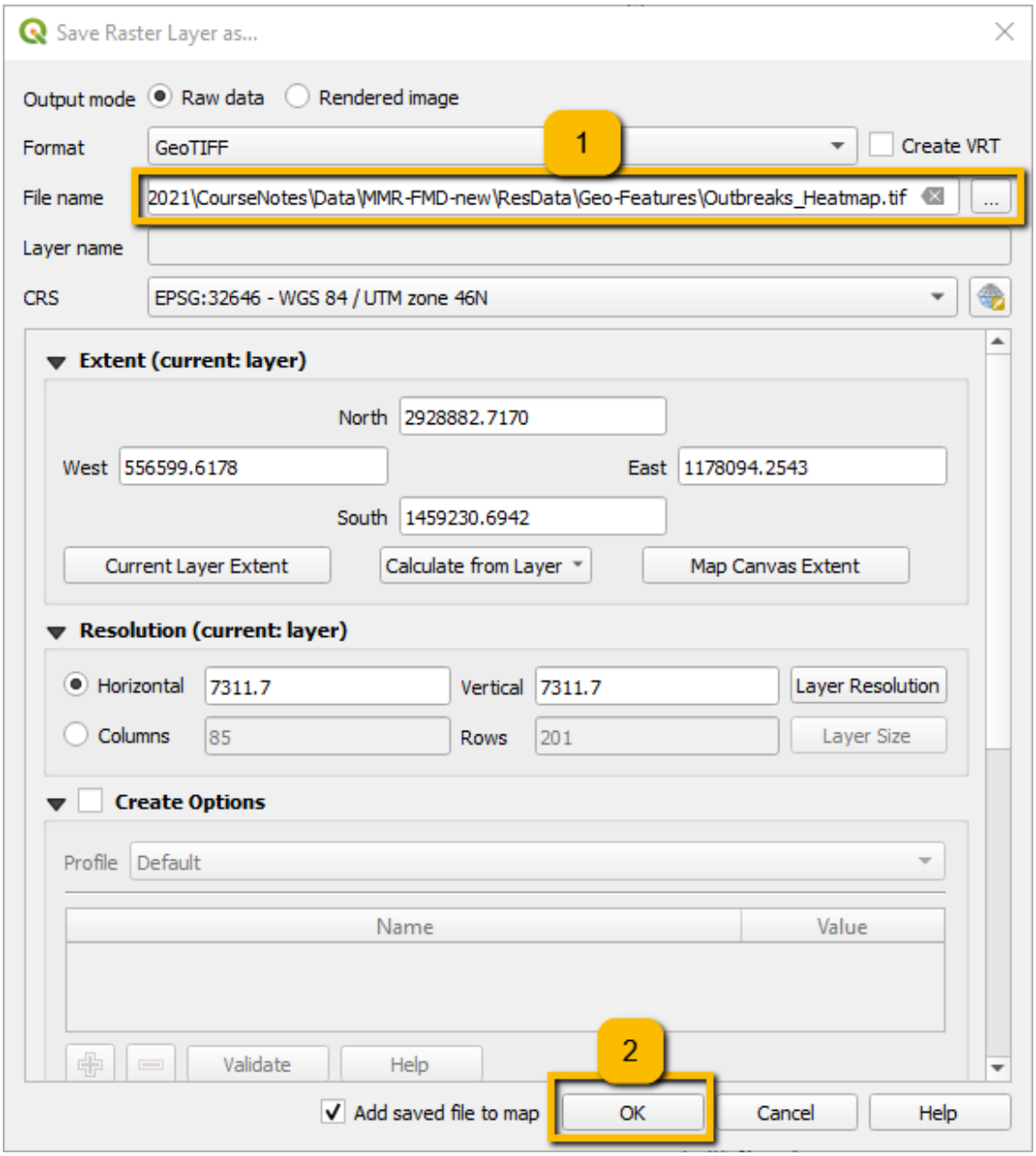

Figure 3.4: Save heat map of FMD outbreak villages in Myanmar as a raster file

3.
    1. (continued)
        4. Remove “Heatmap” from Layers Pane (it is a temporary file only and is not saved when a session is closed) -> Right click -> Remove layer
4. Edit the default view to better visualise the density of cases
    1. Highlight the newly-created file in the Layers Pane “Heatmap”
    2. Right click -> Properties -> Dialog box … (Figure 3.5)
        1. Select “Symbology” tab
        2. In “Render type” select “Singleband pseudocolor”

3. Select a “Color ramp” option (orange to red is a reasonable choice)
4. Click “OK”
5. Experiment with different radius dimensions

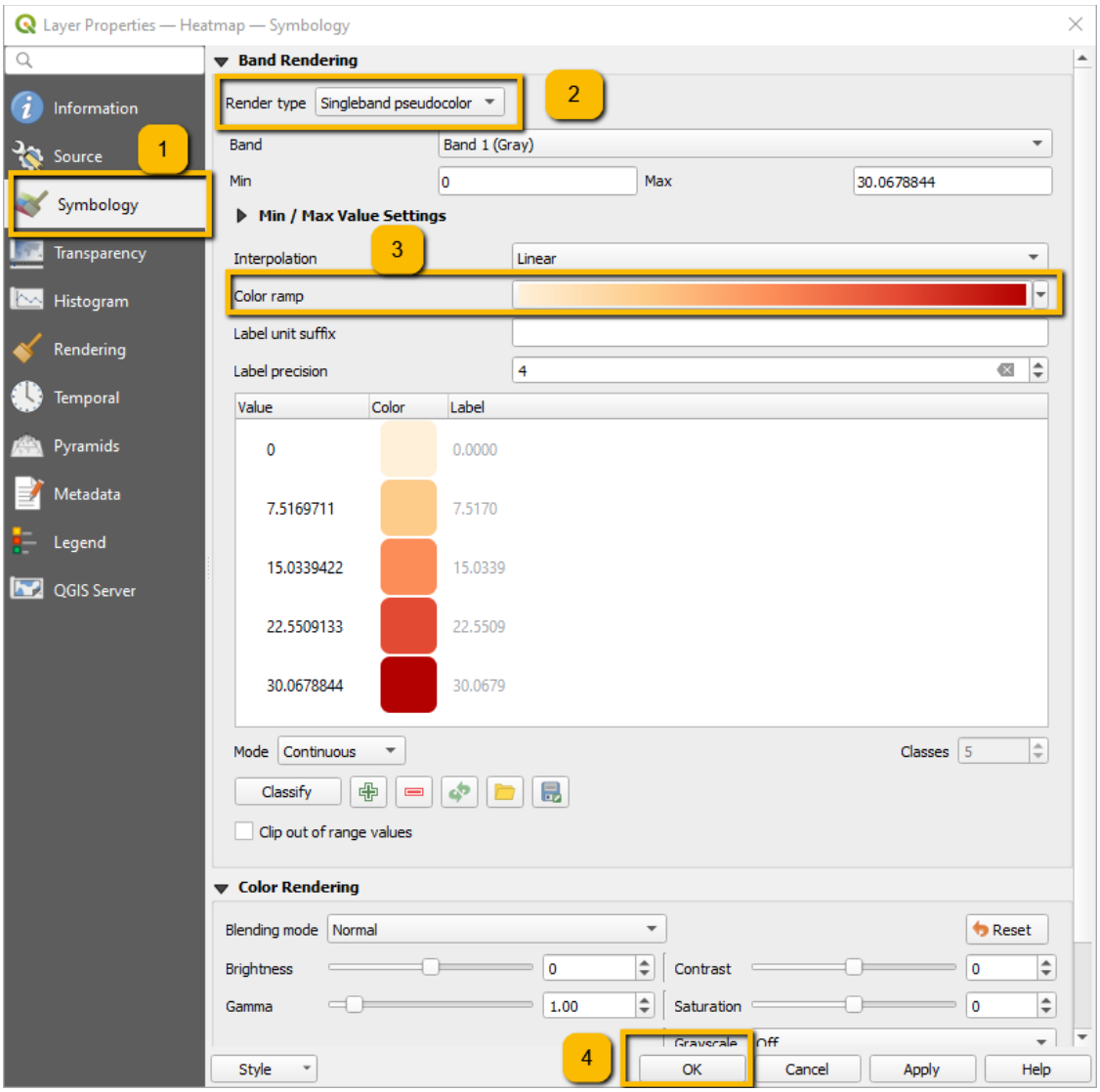

Figure 3.5: Edit symbology of heat map of FMD outbreak villages in Myanmar

5. Your heat map might look something like Figure 3.6)

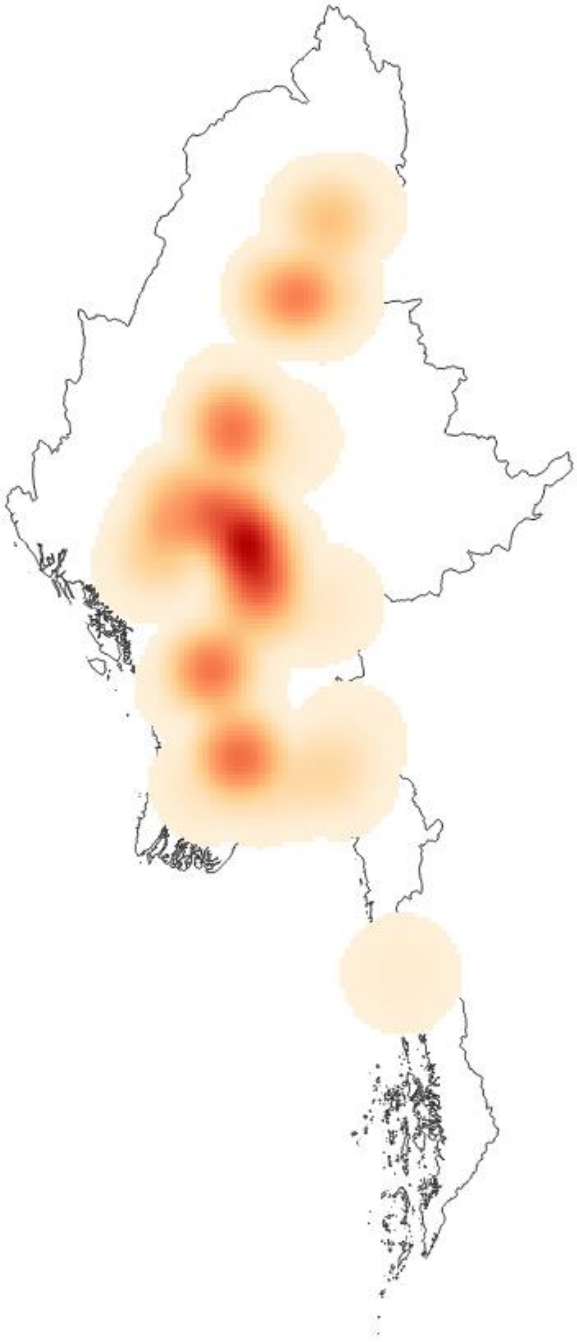

Figure 3.6: Heat map of FMD outbreak villages in Myanmar

Exercise 3.1 (Exploratory spatial data analysis of villages with FMD outbreaks in Myanmar)

Questions:

1. Describe in words the density of village outbreaks within the study region
2. What additional information is needed to adequately explain this pattern

3.1.1.4 Map both disease cases and non-cases together

1. From the Layers Pane drag both “MMRVillages” and “all_outbreaks_rev” onto the Map Palette and arrange the outbreak locations above the village locations so that the former are visible (Figure 3.7)

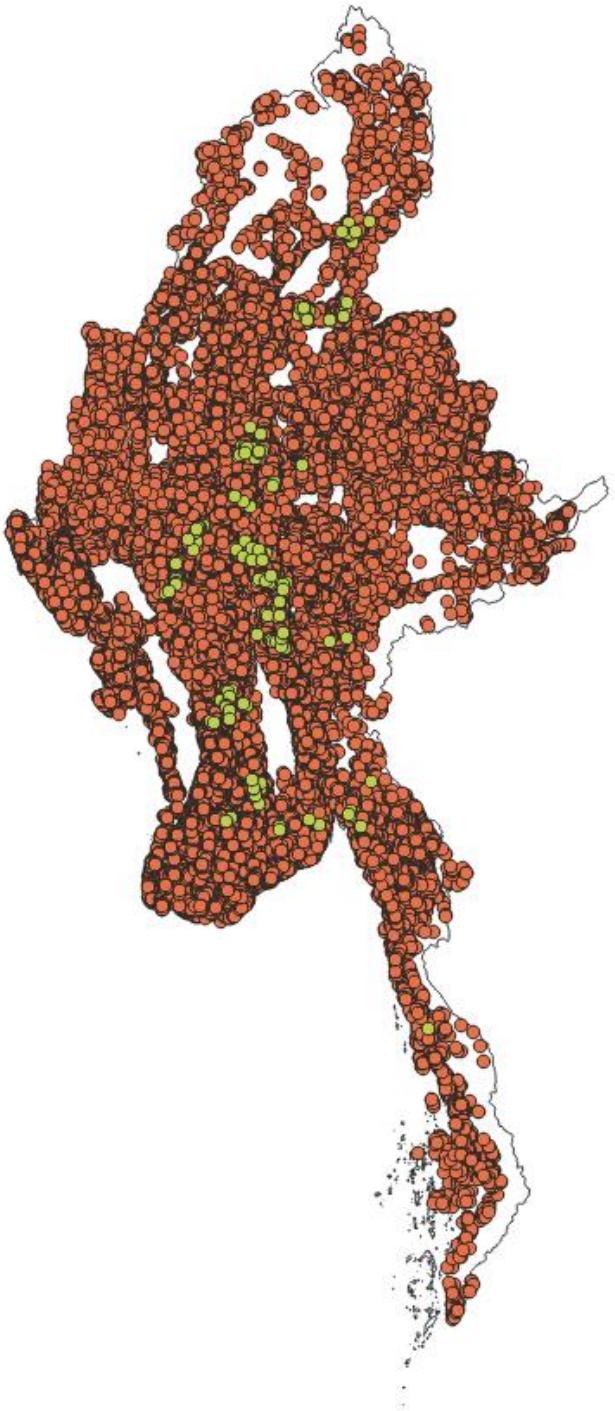

Figure 3.7: Locations of villages with FMD outbreaks (green) and villages without outbreaks (orange) in Myanmar

Exercise 3.2 (Exploratory spatial data analysis of villages with FMD outbreaks in Myanmar)

Questions:

1. What information can you gather from the map in Figure 3.7?
2. Construct a heat map for the density of "MMRVillages" using the same heat map parameters as for the outbreaks, and compare with the density of outbreaks .

- Does it appear that the density of outbreak villages is approximately the same as that for all villages combined? - Could the pattern of density of FMD outbreaks just be reflecting the underlying density of the population of villages at risk of an outbreak, or could other factors be responsible?

3.2 Localised non-focused cluster detection

A formal statistical test is a useful additional step to detect spatial clusters of disease because it removes some of the issues about subjective assessment.

Spatial clustering may be investigated in three different dimensions (linear, point and area) and from a range of different study and data types, for example case-control or surveillance studies with case and population count data, and dichotomous, categorical, rank or continuous data types. Individual statistical tests were 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.

3.2.1 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 of study units e.g. farms or villages
- How the test works:
  - A theoretical circular window is placed on a map of all study locations, for example the centroids of administrative regions
  - A scanning window of increasing radius is placed around one of many possible centroids by sequentially aggregating the nearest neighbour areas to create zones
  - The window radius may vary to a defined upper limit (up to 50% of study population is recommended)
  - For each window the risk of disease is compared with that of the study area outside the window
  - If using case-control data, controls should be selected from same source population as the cases
  - Significance testing is estimated by Monte Carlo sampling

  - The disease data may be either Bernoulli (zero for cases and one for controls) or Poisson (the number of cases and the population at risk)
  - The test adjusts for the heterogeneity of the population at risk by indirect adjustment to calculate the expected number of cases for each location
  - This test may be used as complement to a global clustering test
  - The test can be used to detect clusters with increased, decreased or both increased and decreased incidence of disease
  - The test reports the most significant primary and secondary clusters
- References: Kulldorff and Nagarwalla (1995), Kulldorff (1997)

3.2.2 Temporal scan test

- Data needed: Count of cases by time
- How the 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 of that interval in the study
  - The test can be generalised to account for temporal trends in the population size and incorporated in the SaTScan software
  - The test is most sensitive when the the scanning window is a similar interval as the duration of the clusters
  - It is recommended to set the scanning window on basis of known disease patterns, but the subjectivity of this setting can affect test results
- References: Kulldorff (2018)

References

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/{10.2307/2982708}.

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.

Kulldorff, M. 1997. "A Spatial Scan Statistic." *Communications in Statistics - Theory and Methods* 26 (6): 1481–96. https://doi.org/10.1080/03610929708831995.

———. 2018. "SaTScan- Software for Spatial, Temporal and Space-Time Scan Statistics." https://www.satscan.org/.

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.

Lawson, Andrew B., and M. Kuldorff. 1999. “A Review of Cluster Detection Methods.” 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, 99–110. John Wiley & Sons, Ltd, Chichester, United Analyst (Cambridge, United Kingdom). https://www.ebook.de/de/product/4253210/lawson_disease_mapping_risk_assessment.html.

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.

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.