Scale, resolution and missing data in a long term spatiotemporal epidemiological study

Robert Corner*, Grit Schuster and Bernhard Klingeisen
Department of Spatial Sciences
Curtin University
Perth Western Australia
*R.C.Corner@curtin.edu.au

Abstract—As part of Australia’s biosecurity effort we are working to understand the relationship between the spatiotemporal variability of climatic and environmental conditions and the occurrence of two vector-borne viruses; Murray Valley Encephalitis Virus (MVEV) which causes disease in humans and Bluetongue Virus (BTV) which causes disease in animals. This involves bringing together remotely sensed and, other environmental data with long term serological test data for disease occurrence in sentinel animals. In working with these datasets we are faced with a number of challenges involving scale, resolution and data completeness.

Keywords: Epidemiology, uncertainty, sparse data sets, Remote Sensing

1. INTRODUCTION

In working towards a goal of explanatory models for the occurrence of MVEV and BTV we have been analysing data collected for an area covering the north of Western Australia (for BTV and MVEV) and the Northern Territory (for BTV only). The overall study area covers about one million km² and lies in Köppen classification zones Tropical Savannah, Arid Steppe and Arid Desert (Peel, Finlayson and McMahon, 2007) but is subject to intense summer cyclone activity. Based on spatiotemporal signals that are detected in the environmental data and that can be related to the ecology of the viruses, models are being developed to enable the prediction of disease occurrence risk.

The data employed in this study come from a variety of sources and has a similarly varied set of quality parameters. The data used are as follows: Serological data from sentinel chickens and sentinel cattle, the MODIS MOD11A2 Land Surface Temperature and Emissivity product (LST), the MODIS MCD43A4 Nadir Bi-directional Reflectance Adjusted Reflectance data (NBAR), and rainfall data from the Tropical Rainfall Measurement Mission (TRMM). The main spatial and temporal characteristics of the data are listed in Table 1. Spatial support and resolution are given here as area measures, not pixel side measures.

The MODIS data sets alone comprises multiband datasets covering the study area for the full nine year operational period of that system. The total of nearly 500 data acquisition epochs results in a very large data cube. The other datasets cover the same time frame.

This research was supported by the Australian Bioscience Cooperative Research Centre for Emerging Infectious Diseases

<table>
<thead>
<tr>
<th>Data set</th>
<th>Spatial support</th>
<th>Spatial resolution</th>
<th>Temporal support</th>
<th>Temporal resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicken</td>
<td>unknown</td>
<td>Point data</td>
<td>≤ Temp res</td>
<td>≥2 weeks</td>
</tr>
<tr>
<td>Cattle</td>
<td>unknown</td>
<td>1850 km²(a)</td>
<td>≤ Temp res</td>
<td>≤1 year</td>
</tr>
<tr>
<td>LST</td>
<td>1 km²</td>
<td>1 km²</td>
<td>8 days</td>
<td>8 days</td>
</tr>
<tr>
<td>NBAR</td>
<td>0.25km²</td>
<td>0.25km²</td>
<td>16 days</td>
<td>8 days</td>
</tr>
<tr>
<td>TRMM</td>
<td>625km²</td>
<td>625km²</td>
<td>3 hour</td>
<td>1 day</td>
</tr>
</tbody>
</table>

* Average value, range is 6.0 to 6.00

Figure 1. Location of continuously monitored sentinel chicken flocks in the study area

- 189 -
II. DATA QUALITY ISSUES

The serological data and the remotely sensed data both exhibit different data quality problems with respect to support, resolution, data completeness and attribute accuracy.

A. Support and Resolution.

The support of the data may be regarded as the effective size of each sample (Isaaks and Srivastava, 1989) whereas the resolution describes how the data are presented for use.

The chicken data are nominal sampled twice a month, but are characterised by frequent missed observations. Their spatial location is known to high precision but the “support” for the data is problematic since an incidence of infection requires that a chicken be bitten by a mosquito which carries the disease. This is influenced by the complex ecology of the virus which is further detailed in Section IVA. The spatial distribution of the irregularly spaced flocks is such that each is “responsible” for an average of 60,000 km².

The cattle data are sampled at intervals of no longer than one year. An animal testing positive may have acquired the disease at any time (via a midge bite) during that year effectively rendering the support variable between samples but nowhere exceeding one year. Spatially the cattle may have been anywhere on a cattle station during that year. The average size of a Western Australian cattle station is 1850 km².

The satellite remote sensing data have variable resolution and support. The LST data are fairly straightforward - pixel size is 1km and the support will be fractionally larger than this due to the circular Instantaneous Field of View (IFOV) of the sensor. In the temporal domain the data are delivered every eight days as composites derived over an eight day period, support and resolution again being the same. The NBAR product has a pixel size of 500 m but the support will be a little larger due to circular IFOV. Temporally the data have a resolution of eight days (delivery cycle) but are a rolling composite over 16 days. The support is therefore 16 days.

B. Data Completeness

For the serological blood sampling data, if an animal tested positive to the virus of interest it certainly has become infected since the last testing session. Unfortunately the testing is not as regular or frequent as could be hoped and missing data are a serious problem. During the nine year study period, new chicken flocks were established at some sites, while others were discontinued, sampling sequences were also be interrupted by predation of flocks. Moreover, sampling of poultry is generally done on a voluntary basis and not all flocks were sampled regularly e.g. due to absence or illness of the collaborator.

Data completeness issues for the cattle serological data chiefly relate to the temporal support problem described above.

Data loss for the remotely sensed products is mainly due to cloud in the tropical areas of Australia. Since the TRMM data is based on cloud data this is not a problem but loss occurs in the LST and NBAR data streams. At particular times of the year any one pixel could have missing data for up to ten data delivery epochs.

C. Attribute Accuracy

The serological data essentially comprises virus-positive tests of chicken and cattle drawn from a variable sized cohort. For example the chicken sample size ranges from one to 27. With a large number of positive serological tests from a sample size of 27, we can be reasonably certain that disease is present in the area. Often, however, flock sizes and hence the number of birds sampled was relatively small. Therefore there is a chance of the virus being present in an area even though all birds at a test site returned negative tests.

The accuracy of the remotely sensed data attributes is again variable. The TRMM rainfall estimates have been compared to gridded rain gauge data and have been found to generally agree well (Ebert, Janowiak and Kidd, 2007). Although TRMM data do not deliver high precision rainfall rates, they provide the most accurate and spatially regular available information for the entire study area. Their spatial regularity renders them highly capable of capturing the spatial variability of rainfall events remote from meteorological stations.

The NBAR and LST data are also the products of sophisticated processes that are designed to produce the highest possible quality outputs (Salomon, Scharf, Strahler, Gao and Jin, 2006) (Wan, 2008).

III. FITNESS FOR USE

A key concept in considering data quality is that of fitness for use. There are a number of ways of assessing this, chiefly comparison to a set of accepted standards or an assessment of the risk of using the data (Agumya and Hunter, 2002). In this situation we have no option but to use the data, they are all that are available, particularly the serological data. Whilst we might recommend future protocols that address the data quality problems, for the historical record from which we will build our models we do not have that luxury.

In order to make a “use-based” risk assessment we must first examine the intended use of the data, which was to build explanatory models of disease occurrences (as indicated by serological data) for MVEV (chicken data) and for BTV (cattle data). The methods used so far are logistic regression, ideally suited to presence/absence data, and Generalised Additive Models (GAM). A range of independent environmental variables was employed to model the dependant, sero-positive or negative condition of the sentinel animals. In consideration of the risk of ignoring biological facts by using a mathematical model, selection of models and explanatory variables is not based solely on statistics, but also on the distinct ecologies of the two viruses which is described below.

IV. VIRUS ECOLOGY

A. Murray Valley Encephalitis Virus

This virus is transmitted by female mosquitoes, primarily Culex annulirostris species, which breed in shallow fresh water, particularly in receding floodwaters. The virus is
maintained in a cycle between the mosquitoes and water birds such as Herons and Egrets which act as reservoirs and amplifiers. Domestic fowl are also capable of infection, as are humans, with potentially fatal consequences. However humans do not pass the disease on even if bitten by a mosquito whilst suffering from the disease.

Both the vectors and hosts are mobile. Female Culex mosquitoes can travel at least 12km per day with the rate of dispersal of the population estimated at 22 km per day (Bryan, O'Donnell, Berry and Carvan, 1992). The time taken for the virus to develop and be able to be passed on by a mosquito is about eight to 24 days. The water bird hosts are migratory and typically exploit the resources of the floodwaters in northern Australia's wet season. However little is known about the movements of birds.

A model of MVEV should therefore include indicators of proximity to water, particularly floodwater where birds and mosquitoes will be prevalent, as well as water amounts as derived from rainfall. Temperature is also important for mosquito populations to establish and for virus replication.

B. Bluetongue Virus

BTV affects both sheep and cattle and, whilst not transmissible to humans, it has an effect on agricultural trade, particularly the export of live animals. It is transmitted by female Culicoides midges and can be maintained in a population of ruminants by cycling between animals and midges. The midge vector particular to Australia breeds in fresh wet cattle dung. The midges also have a preference for grassland. Hence moisture, vegetation information and the presence of cattle are necessary BTV model components. In addition temperature determines the vectors level of activity and length of the development cycle. Sheep, although highly susceptible to BTV, are not present in our study area.

V. DEVELOPED ENVIRONMENTAL VARIABLES

Whilst the set of developed environmental variables used to model each of the virus ecologies differ, they are based on the same fundamental remotely sensed parameters. The simplest of these developed variables are vegetation indices. Both the Normalised Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) (Huete et al., 2002) have been used.

More complex developed variables are those that combine both space and time. An example of these is a suite of seasonally accumulated rainfall variables covering selected periods throughout the rainy season. These included early wet season (Dec- Feb), early plus high wet season (Dec-Mar), entire wet season (Dec-May) and high wet season (Jan-Mar). Similar variables were developed for temperature, based on the Growing Degree Day concept, which accumulates the time for which temperature exceeds a particular limit. Further complexity is introduced by variables that determine periodic (monthly or seasonal) anomalies from long term trends.

In addition variables were developed to reflect the temporal lags in the development of MVEV in susceptible populations.

The calculation of all these developed variables will propagate the uncertainty in the original geophysical parameters by either amplifying it or smoothing it depending on the algorithm used.

VI. RESPONDING TO DATA QUALITY ISSUES

The vast amount of data demands pragmatic approaches to the data quality issues. Our responses have been designed to minimize the chance of false disease risk predictions. All data, except TRMM which were supplied to us already processed, have required both spatial and temporal treatment. These treatments have been different for the serological and remotely sensed datasets.

A. Serological Data

1) Cattle data.

Spatial support and resolution for the cattle serological data are both low. In order to relate the serological data for any one herd to environmental conditions, those conditions must be summarized for the entire cattle station. However it is reasoned that in any one year parts of the station will be more frequented by cattle than others. Whilst vegetation indices can provide information on biomass production they do not address the palatability of that biomass to stock. In order to improve this information ancillary data were used. The highest resolution data available on the soils and vegetation of the northern rangeland areas in Australia comes from a series of maps and reports which provide information at a Land System level and which include estimated carrying capacity of each Land System (Tille, 2006).

Although we still have to aggregate environmental variables over the entire area of a station we can apply a weighting scheme based on carrying capacity since areas of higher capacity are more likely to have been frequented by cattle during the year.

The poor temporal support and resolution are also dealt with by applying ancillary data. In this case it is the knowledge that the disease is particularly active in the wet season, and the operations are performed on the environmental data as described in section VIB below.

2) Chicken data

Whilst sampled more frequently than the sentinel herds, the sentinel flock data still required some work in order to create a data set appropriate for statistical analysis. It was judged safe to interpolate across sampling gaps if the following criteria were met:

- No seroconversion (from negative to positive) was detected in a flock when testing recommenced after a testing free period.
- No changes occurred to the flock composition during the testing free period.

Where a seroconversion had occurred after a testing free period that gap could not be filled by interpolation and a further strategy was applied in those cases. If the gap was longer than three months the flock was excluded, whilst for gaps less than three months, the seroconversion was assumed to have occurred during the middle of the sample free period.
Test sites where interpolation was not possible according to either of these rules were excluded from the analysis.

**B. Environmental Data**

Operations to ameliorate data quality problems with the environmental variables were mostly carried out at the level of the fundamental remotely sensed parameters.

1) *Land surface temperature data*

This data set contains many data gaps due to cloud cover and other atmospheric influences. The LST data include a "quality band" which was used to remove pixels with low quality (e.g., at the edge of clouds). These were additional to the pixels in the cloud mask from the MODIS LST algorithm. The remaining data were processed according to the following rules:

- Spatial gaps no wider than five pixels were filled with the spatial mean for the surrounding pixels.
- For spatial gaps of greater than five pixels, but lasting no longer that three data collection epochs, missing values were interpolated from temporally adjacent images.

In spite of this approach, especially during the wet season, a continuous time series cannot be generated and it is not possible to derive seasonal environmental variables. Environmental variables, such as maximum day temperature and minimum night temperature were generated only when more than 50% of the data were available.

2) *NBAR surface reflectance data*

The BRDF data were used for a number of purposes, these included the determination vegetation phenology (timing and intensity of growing season) from vegetation indices and the mapping of surface water. These required different gap filling strategies.

The development of phenology requires a knowledge of the maxima in the vegetation indices, so the unchanged data were used since any interpolated values would by definition not be extremes. Time series analysis of the data was then carried out using the Timesat software (Jonsson and Eklundh, 2004) which offers several curve fitting methods to smooth noisy data.

At the time of writing the surface water mapping algorithms are still being developed and require the differentiation, from water, of cloud, cloud shadow and salt affected pixels. The quality data for the NBAR product form part of a separate data set which increases the overall size of the data cube. Work is in hand to automate this process. In the meantime spatial aggregation of the data around the chicken flock sites suggest that the best results are obtained when the area within a 250 Km radius is considered. This is consistent with the little that is known about bird movements.

**VII. Results**

The development of explanatory models continues at the time of writing but preliminary results show that our strategies have achieved success.

Logistic regression models have been developed for MVEV using a range of rainfall variables. The explanatory power of the developed models, using the Receiver Operating Characteristic method, ranges from 0.7 for seasonal models to 0.81 for time lagged monthly models.

Generalised additive models for BTV in the Northern Territory have an explanatory power in excess of 0.81

**VIII. Conclusion**

This work has outlined a number of pragmatic solutions to issues of data quality in a diverse and disparate set of data.

**Acknowledgements**

We wish to thank all contributors to the Western Australian Arbovirus Surveillance and Research Program and the National Arbovirus Monitoring Program, for providing serological data. We also thank Luigi Renzullo from CSIRO Land and Water, Canberra for providing us with the daily compilations of the TRMM 3B42 data.

**References**


