A new spatial aggregation algorithm that improves performance of spatial cluster detection

Christiansen, Lasse Engbo; Van Meter, Karla

Publication date:
2008

A new spatial aggregation algorithm that improves performance of spatial cluster detection

Lasse Engbo Christiansen[1] and Karla Van Meter[2]

[1] Department of Informatics and Mathematical Modelling, Technical University of Denmark, DK

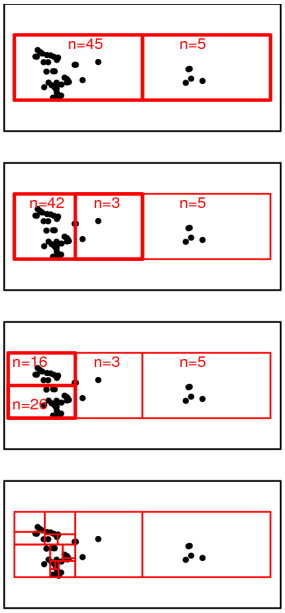
[2] Public Health Sciences, School of Medicine, University of California Davis, CA, US

Introduction

As more detailed spatial data becomes available, e.g. geocoded individual addresses rather than county wide counts, the computational burden in tests for spatial clustering increases; most tests have practical upper limits on the number of locations that can be used. Therefore, some spatial aggregation is needed. Typically political boundaries or a regular square grid are used. Political boundaries are subjective and may not have a useful resolution. The resolution of a square grid can easily be adjusted but when the density of the underlaying point process changes by orders of magnitude within the region of interest, the range of the population at risk per grid cell is very large which affects the performance of the test.

Methods

We present the concept of epiunits as areal units recursively subdivided to obtain a maximum population per epiunit. We propose four different approaches with different advantages: Pure spatial, pure density, spatial then grow, and spatial then density.



Here is the algorithm for generating a set of epiunits using the spatial then grow approach:

a) Select maximum population in each epiunit, n_{max}. Then define the boundary box of the area of interest.

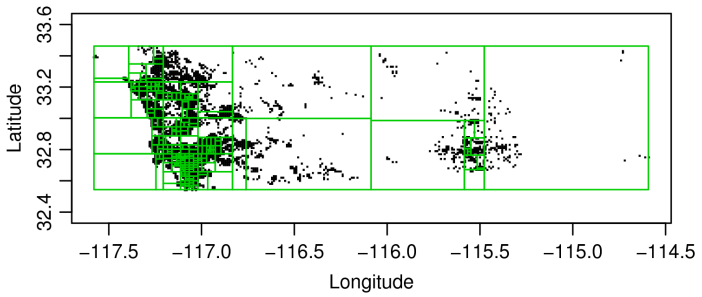
b) Divide the box along the longest edge.

c) Count the population in each of the two halves.

d) If any of the two counts is less than $n_{max}/2$ then remake the division by enlarging the half that has too few points until it has $n_{max}/2$. And re-count in each half.

e) Repeat b-e recursively for each of the halves with a population above n_{max}

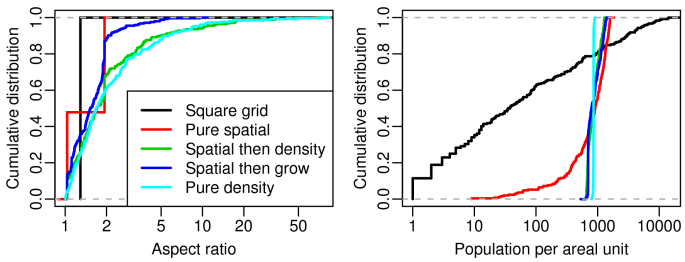
The different approaches are compared and contrasted using a simulation study where the point data is birth population data from the 1996 through 2000 State of California Department of Health Services Center for Health Statistics Confidential Birth Files. The mother's address at birth for each record was geocoded with a geocoding success rate of 93 percent. All 219,417 geocoded birth locations within San Diego and Imperial counties, with mixed rural and urban areas, are used as the study population. See Van Meter et al.[1] for further details.



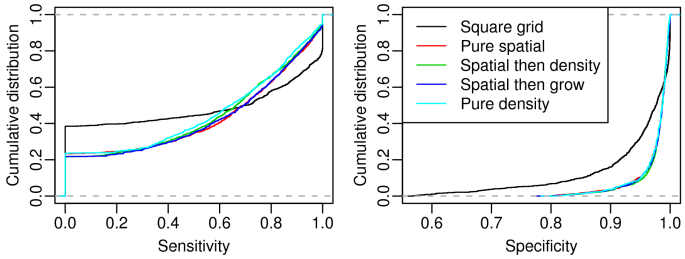
The simulations were done by selecting a random point and defining the nearest 2% of the points as the true underlying cluster. Then cases of a rare disorder were generated with a background risk of 0.004 outside the cluster and a pre-specified relative risk factor inside the cluster. Further, Episcan[2] was used to find the most likely cluster.

Results

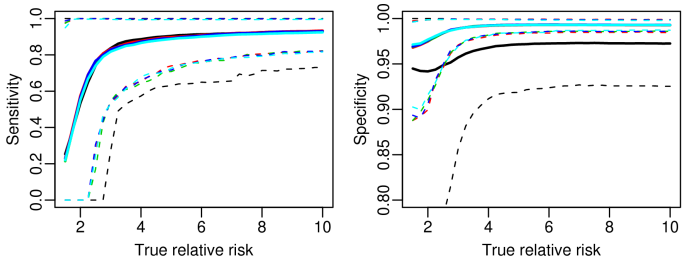
First looking at the cumulative distributions of the aspect ratios (length of longest edge divided by length of shortest edge) of the areal units created with the different approaches and the population per areal unit.



Simulation study with 1000 replicates showing the ability to identify an underlying cluster with a relative risk of two.



Both sensitivity and specificity depend on the true relative risk. Below the median is a solid line and the 5% and 95% quantiles are dashed lines.



Conclusions

- Based on the aspect ratio and count per areal unit it seems that the spatial then grow and pure spatial approaches are the best.
- The four epiunit approaches have similar performance in the simulation study.
- The traditional square grid approach has larger variance in sensitivities and generally lower specificities than the epiunit approaches.
- Given that the specificity is important in the search for explanatory environmental factors for rare diseases the pure spatial or spatial then grow sets of epiunits should be preferred over a square grid.

References

[1] Van Meter KC, Christiansen LE, Hertz-Picciotto I, Azari R, Carpenter TE.: A procedure to characterize geographic distributions of rare disorders in cohorts. Int J Health Geogr. 2008 May 28;7:26

[2] Christiansen L, Andersen J, Wegener H, Madsen H: Spatial Scan Statistics Using Elliptic Windows. Journal of Agricultural, Biological and Environmental Statistics 2006, 11:411-424