The Effect of Uncertain Locations on Disease Cluster Statistics

Geoffrey M. Jacquez1 and Lance A. Waller2

Abstract. Disease clusters are space-time aggregations of adverse health events such as birth defects and cancer. Because of incomplete information the exact locations of health events are often unknown, but to date the effect of uncertain locations on disease cluster statistics has not been assessed. This research demonstrates that uncertain locations, if not accounted for, lead to test statistics, null distributions and p-values that differ substantially from those based on actual locations, and increases the probability of type II error (false negatives). Further research is needed to develop statistical methods and software to account for uncertain locations in disease cluster investigations.

INTRODUCTION

Public health agencies use cluster statistics to assess whether an alleged aggregation of disease cases is statistically unusual. The health agency may choose to stop a cluster investigation when the alleged cluster appears to be due to chance, otherwise a more detailed and expensive investigation may be warranted (Centers for Disease Control 1990). Point-based statistics assume exact geographic locations and are often used when health events are rare and the number of observations is small. They are coming into increasing use as spatially referenced health data become commonly available (Openshaw 1991), and models of spatial point processes (Diggle 1993, Lawson and Waller 1996), become increasingly sophisticated. However, the assumption of exact locations upon which point-based statistics are founded is often invalid. Does the violation of this assumption make a difference?

Sources Of Location Uncertainty

Point based statistics assume data of the form \((x, y, t)\) where \(x, y\) is the geographic coordinate (such as place of residence) and \(t\) is the date of onset, diagnosis or death. More often than not this location model is inappropriate because locations of the health events are uncertain. This uncertainty has several sources. It emerges when centers of areas such as zip-code zones or census tracts are used instead of exact place of residence. Uncertainty also arises when data are gridded (as, for example, in raster-based GIS) and the coordinates of the nearest grid node are used instead of exact locations. Conceptually, locations are almost always uncertain because humans are mobile rather than sessile and because

---

1President, BioMedware, Ann Arbor, MI.
2Assistant Professor, University of Minnesota, Minneapolis
health events and their causative exposures may occur anywhere within a persons activity space. Tobler et al. (1995) observed that in modern society a persons daily activity space is approximately 15 km and varies widely. Exact locations do not represent this location uncertainty.

There are many examples of uncertain locations in the literature. Cuzick and Edwards (1990) used the centers of postal code zones to represent place of residence in their study of childhood leukemia in North Humberside. Waller et al. (1995) mapped locations of cases of childhood leukemia in Sweden at parish centroids, and evaluated proximity to nuclear power plants using several methods. Location uncertainty arises in these two examples because area centroids are used instead of exact locations. In a study notable for its spatial resolution, Lawson and Williams (1994) used place of residence to assess a possible cluster of deaths from respiratory cancer near a smelter in Armadale, Scotland. Indoor air quality is known to differ substantially from that outdoors, and an unknown proportion of each individuals exposure to smelter fumes presumably occurred outside of the home. In this example location uncertainty arises because place of residence is used to represent exposures that occurred throughout each persons activity space.

We have not evaluated whether or not uncertain locations had an effect, if any, on these studies. Rather, our objective in citing these examples is to give readers a feeling for the ubiquity of the location uncertainty problem. In practice geographic locations are, to a greater or lesser extent, inherently uncertain because they are used as proxy measures of exposure. If we had more information we could explore dose-response relationships using controlled studies. When such detailed knowledge is lacking we instead use spatial relationships among cases, and their proximity to putative sources of hazard, as an uncertain measure of exposure. However, almost all point-based cluster statistics assume the spatial coordinates of health events are precise. What is the impact, if any, of using statistics that assume exact locations when the locations are actually uncertain?

METHODS

Perhaps the most frequently used paradigm for dealing with uncertain locations is the centroid model. This model assigns all cases occurring within an area the location of that area's centroid. It was used explicitly in the studies in Humberside and Sweden and is implied in the Armadale study. It obtains whenever data are gridded and is common in environmental epidemiology where insufficient knowledge of exposures and confidentiality concerns preclude exact knowledge of health event locations. Using a simulated AIDS epidemic in Michigan, we explored the performance of three disease cluster statistics at three levels of spatial resolution. We compared the test statistics, null distributions, p-values and statistical power under the centroid model to those obtained when actual locations were used.
Cluster Statistics

Three statistics for space-time clustering were compared, Mantel (1967), Knox (1964) and k-NN (Jacquez 1996a). These commonly used tests were chosen because they are based on different spatial proximity measures: geographic distance, adjacencies and nearest neighbor relationships, respectively (Jacquez 1996b). We choose tests based on several different proximity measures so that we could generalize the results to a wider range of other cluster tests.

Mantel's test is based on regression of the space distance between pairs of cases on the time distance between case pairs. This study used the reciprocal transformation suggested by Mantel. The Knox test employs space and time critical distances to classify pairs of cases as either near or far in space and time. For example, two cases are considered near to one another when the distance between them is less than the space critical distance. In this study we selected critical values to maximize the Knox test statistic, and used these critical values for all of the simulations. While choosing critical values in this fashion precludes statistical hypothesis testing, it was deemed appropriate for the purpose of the exploratory analyses in this study. The k-NN statistic is the intersection of the space and time nearest neighbor relationships and is described by Jacquez (1996a). All cluster statistics were calculated using the Stat! software (Jacquez 1994).

Figure 1. Boundaries corresponding to four levels of geographic uncertainty.

Geography

Four levels of geographic uncertainty were evaluated: none, county, bicounty and region (Figure 1). Places of residence of the at-risk population are shown in 1a by the ' +' symbol and were sampled based on population density distributions of the global demography project (Tobler, Deichmann et al. 1995). Centroids and areas are shown for counties (1b), bicounties (1c) and regions (1d). Area centroids are shown as squares.

HIV Simulation

A simulated epidemic of AIDS in Michigan was used to generate space-time distributions of cases, and is based on the model of Jacquez et al (1994). This HIV
model was chosen because it describes the dynamics of the epidemic and because the transmission parameters have been estimated. The model was simplified for this research and, while the simplified version cannot be said to represent HIV in human populations, it is a contagious process and should result in space-time clusters. If the cluster statistics perform poorly for this model we have little reason to expect them to perform well for diseases such as environmentally caused cancer, for which space-time clustering should be substantially weaker.

The modeled disease has 4 stages: healthy, primary infection, seropositive, and deceased. The vector of transmission probabilities was \((0.0, 0.3, 0.001, 0.0)\). Each element of this vector describes the probability, for a healthy individual, of infection per contact with an individual in disease stage \(i\). The underlying population consists of 500 individuals with a heterogeneous geographic distribution (Figure 1a). Two of these were selected as seeds to start the epidemic. At each time step each individual contacts 1 of any of 4 nearest neighbors, and a transmission event can occur when the contacted individual is infectious. Transition from healthy to primary infection does not occur in the absence of a transmission event. Probabilities describe transition from one disease stage to another. In each round there is a 5% probability that individuals in the primary infection stage will become seropositive, while, on average, 1% of seropositives enter the terminal stage.

The simulation is a three-step process; first transmission events (which cause healthy individuals to transit to primary infection) are evaluated. Next, stage transitions are determined using transition probabilities. Finally, after each iteration is complete the space-time location of each new seropositive is recorded and the simulation halts when 50 individuals become seropositive. The seropositive stage thus was used for the reporting of disease clusters. The model is stochastic and uses a uniform random number generator to evaluate the discrete transitions for each individual in the population. It is realistic because seropositives are monitored, and because stochasticity is introduced through probabilistic processes of contact, transmission and state-to-state transition.

Simulation Protocol

The HIV simulation model was run 252 times using the actual locations shown in Figure 1a, and new locations of the 2 initial infectives were chosen for each run. Each of these runs is a distinct realization of the AIDS epidemic and is called an 'AIDS simulation'. Each AIDS simulation produced a list of coordinate pairs describing the actual locations of 50 seropositives. Centroid locations for the county, bicounty and regional aggregation levels were assigned as appropriate (Figure 1). For example, at the county level a case occurring in county \(i\) was assigned the coordinates of that county's centroid. The actual test statistic \(T_A\) and reference (null) distribution \(g_A\) were calculated using the actual locations, and usually are not observable in field situations when centroids are used. The
centroid test statistic ($\Gamma_C$) and its null distribution ($g_C$) were then calculated using the centroid locations, and correspond to statistics calculated in studies where centroid locations are used. The $g_A$ and $g_C$ distributions used 249 randomization runs to generate reference distributions expected under the null hypothesis of no association between a case's location and the time at which the case became seropositive. This was accomplished by reallocating the times the cases became seropositive at random across the case locations. These statistics and distributions were recorded for each AIDS simulation and for each cluster statistic.

In all, 252 AIDS simulations were run. 249 Monte Carlo randomizations were used to generate each statistical null distribution. 188,244 Monte Carlo randomizations were needed to generate the $g_A$ distributions (252 AIDS simulation x 3 cluster tests x 249 randomizations) and 564,732 Monte Carlo randomizations were needed to generate the $g_C$ distributions (252 AIDS simulation x 3 cluster tests x 3 uncertainty levels x 249 randomizations). A grand total of 752,976 Monte Carlo randomizations were conducted using spatial uncertainty modeling software being developed by BioMedware.

RESULTS

Inquiry focused on four questions. How does uncertainty regarding the spatial locations of health events affect (1) the test statistic, (2) its null distribution, (3) p-values, and (4) statistical power of point-based cluster statistics?

Figure 2 shows the results at the county level. For Mantel's test, test statistics calculated from the centroid locations were consistently lower than those calculated using actual locations, and the upper 95% critical value of the $g_C$ reference distribution was not highly correlated with the 95% critical value of the $g_A$ reference distribution calculated from the actual locations. As a result, p-values using the centroid locations were essentially independent of p-values based on the actual locations (Figure 2a). This pattern of results holds for both the Knox test (Figure 2b) and the k-NN test (Figure 2c). Results, not shown due to space constraints, at the bicounty and regional aggregation levels are as bad or worse and demonstrate, for the HIV model and at the spatial scales considered, that p-values obtained under the centroid location model differ substantially from p-values based on actual locations. This is consistent with the results of Waller and Turnbull (1993), who reported changes in p-values at the census tract and block group aggregation levels for focused cluster tests.

For the 3 cluster statistics considered there is a substantial decrease in statistical power at the county, bicounty and regional aggregation levels. Statistical power was quantified as the proportion of simulation runs for which the p-value was less than or equal to 0.05. Using actual locations the statistical power was 0.592, 0.480 and 0.792 for the Mantel, Knox and k-NN tests, respectively. At the county level the drop in statistical power ranged from 3.67% (Knox) to 36.62% (k-NN). At the bicounty level the drop ranged from 6.58% (Knox) to 44.1% (k-NN).
Figure 2. Results at the county level, centroid p-values vs. actual p-values for Mantel (a) Knox (b) and k-NN (c) tests.

the regional level the drop in statistical power ranged from 22.60% (Knox) to 59.83% (k-NN).

These results illustrate that p-values from centroid data bear little resemblance to p-values based on actual locations. This causes a loss in statistical power when using centroid locations, and an increase in type II error (false negatives).

CONCLUSION

As mentioned in the introduction, cluster statistics are one of several criteria used by public health agencies to determine the future course of a cluster investigation. In such situations, p-values from disease cluster tests are used to quantify how unusual an observed aggregation of cases is. A very small p-value suggests the aggregation was unlikely to have been a chance event, and additional effort may be needed to identify environmental exposures that might explain the cluster. While generalization from a single simulation study is difficult, our results suggest that p-values based on uncertain locations are unreliable and should not be used as a quantitative basis for determining the future course of cluster investigations.

Several authors have commented on the apparent lack of statistical power in cluster investigations. This study suggests that actual clusters may be missed because of the loss of statistical power due to location uncertainty.

Even when place of residence is known to the level of the street address, geographic location is often uncertain because health events occur throughout a person's daily activity space, and because exposures (e.g. transmission events for infectious diseases, exposure to radiation and carcinogens for cancer, etc.) underlying health events occur outside as well as inside the home. This indicates a substantial need for ways of accounting for uncertain locations in general, and for point-based statistics in particular.

There are broader implications because the cluster tests considered are similar to statistics used in other fields. For example, nearest neighbor statistics are used extensively in ecology, and Mantel's test can be expressed as the Pearson product moment correlation which is used extensively in the environmental sciences. All GIS's employ data models that, depending on the GIS operation, may abstract locations to grid nodes or polygon centers, thereby introducing location
uncertainty. This study therefore may have implications for statistical analyses of data managed by Geographic Information Systems.

In conclusion, the software and methods developed in this research allow one to assess, given a specific geography and hypothesized space-time process, the impact of uncertain locations on test statistics, reference distributions, p-values and statistical power. Generalization from the single simulated disease process and geographies considered is problematic. Will the pattern of results hold at finer geographic resolutions? Because of their use in small area studies, it seems particularly important to assess the impact of uncertain locations for census tract and zip-code zone geography. Now that a substantial effect at the county level has been demonstrated, the methods developed in this preliminary study can be used to address this question and others like it.

ACKNOWLEDGMENTS

This research was funded by Small Business Innovation Research grant R43 CA65366-01A1 from the National Cancer Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCI.

REFERENCES


**BIOGRAPHICAL SKETCH**

Geoffrey M. Jacquez is President of BioMedware, Inc. of Ann Arbor, Michigan. He holds an M.S. in resource policy and law from the University of Michigan, and a Ph.D. (1989) in ecology and evolution from SUNY Stony Brook.

Lance Waller is Assistant Professor of Biostatistics in the School of Public Health at the University of Minnesota. He received his Ph. D. in operations research from Cornell University in 1991.

BioMedware was founded in 1990, develops software and statistical methods for the environmental and health sciences, and is the recipient of SBIR and STTR grants from the NIH.