THE ANALYSIS OF PUTATIVE SOURCES OF HEALTH HAZARD

Andrew B. Lawson
Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, USA

Keywords: hazard, risk, putative source, pollution, statistical modeling, health, epidemiology

Contents

Summary

This contribution discusses the use of statistical methods for the analysis of the locations of potential health hazard. Emphasis is placed on the use of statistical modeling in this analysis. Focused clustering is the main object of the analysis.

1. Introduction

The assessment of the impact of sources of pollution on the health status of communities is of considerable academic and public concern. The incidence of many respiratory, skin and genetic diseases is thought to be related to environmental pollution, and hence any localized source of such pollution could give rise to changes in the incidence of such diseases in the adjoining community.

In recent years, there has been growing interest in the development of statistical methods useful in the detection of patterns of health events associated with pollution

sources. In this review we consider the statistical methodology for the assessment of putative health effects of sources of air pollution or ionizing radiation. We consider study design issues, inference and modeling problems. We concentrate primarily on the data analysis of observed point patterns of events rather than specific features of a particular disease or outcome. Our purpose is to review statistical methods, so some published case studies of pollution sources of hazard may not appear.

A number of studies utilize data based on the spatial distribution of such diseases to assess the strength of association with exposure to a pollution source. Raised incidence near the source, or directional preference relate to a dominant wind direction may provide evidence of such a link. Hence, the aim of the analysis of such data is usually to assess specific spatial variables rather than general spatial modeling. That is, the analyst is interested in detecting patterns of events near (or exposed to) the focus and less concerned about aggregation of events in other locations. The former type of analysis has been named '*focused clustering*'. To date, most pollution source studies concentrate on incidence of a single disease (e.g. childhood leukemia around nuclear power stations or respiratory cancers around waste product incinerators).

The types of data observed can vary from disease event locations (usually residence addresses of cases) to counts of disease (mortality or morbidity) within census tracts or other arbitrary spatial regions. The two different data types lead to different modeling approaches. Spatial point process models are appropriate for event location data. In the case of count data, one may use properties of regionalized point processes. That is, an independent Poisson model for regional counts is often assumed and one typically uses log-linear models and related tests.

The effects of pollution sources often are measured over large geographic areas containing heterogeneous population densities (usually both urban and rural areas). As a result, the underlying intensity of the point process model is heterogeneous.

2. Study Design

In what follows, we consider a delimited geographical study area or window within which data concerning disease occurrence and exposure to the pollution source are collected. Issues concerning the strategic aims of the study must be considered prior to detailed consideration of the appropriate study region and data collection requirements.

2.1. Retrospective and Prospective Studies

During the 1980's, a number of studies of disease occurrence in geographical regions around putative sources of risk were carried out. Most of these were 'reactive', in that suspicion of a health risk, due to the past operation of a pollution source, instigated review of the historical evidence for a link between disease incidence and exposure to the source. In essence, *a retrospective* study of disease occurrence was carried out. In some cases, continued monitoring of the source was also recommended or initiated. However, solely, *prospective* studies of sources are seldom encountered. These two approaches and their respective strengths and weaknesses are well-known in the epidemiological literature.

Such studies of effects of pollution have a number of limitations, however. First, typically the emission characteristics of a source are not recorded for a suitable time period. Retrospective data on emissions may not be available and prospective monitoring data is expensive to collect over a long time period for a wide range of substances of interest. Often, no direct information is available on correlation between emission and disease occurrence. Furthermore, exposure and disease data are often collected by separate groups at different levels of resolution (even in prospective studies). Also, the nature of available data may be limited for particular diseases or health status indicators, or for particular time periods. Often, nationally-collected data rather than data from a specially designed study must be utilised. In some cases, the level of resolution in available data constrains the analysis considerably. For example, some diseases are reported only as counts from postal zones or census enumeration districts and not as exact addresses due to confidentiality. In that case, methods based on analysis of counts rather than point events are appropriate. Inevitably, such regionalization leads to some loss of information. For example, very small clusters cannot be detected if they occur within a large census tract as the aggregate disease rate for the tract as a whole may not differ from the background disease rate. Only if the spatial pattern of events occurs at a larger scale than the measurement unit will it be detectable in regionalised data. Finally, for chronic outcomes like cancers, the temporal lag between exposure and an event of interest may be on the order of years or decades. Mobility of individuals over such a time period can confound exposure-outcome relationships and cause prohibitive costs in prospective studies over large areas.

2.2. Study Region Design

The design of a study region or window is of great practical importance. Usually, a study will concern the distribution of events (e.g. incident disease cases) within a fixed map area of given size and shape. The choice of size and shape can have considerable impact on study results and, while often it is not possible to choose the most appropriate region, some consideration should be given to these issues.

2.2.1. Region Size

A study region should be defined which is of sufficient size that any effects of a putative source can be measured adequately. As it is often not possible to assess, a priori, the spatial scale of pollution effects, it is therefore important that a large region including the pollution source should be used. In many published studies a region is defined and the total incidence in the region is analyzed (compared to external 'control' regions). If a region is specified which is larger than the true pollution range then a localised effect within some part of the region may be diluted. On the other hand, a small region may truncate the evidence and not represent the complete effects in the population. In addition, the use of multiple region sizes may still induce problems in data analysis if a pollution effect occurs at a spatial scale different than those considered.

In previous studies, sizes of region, in radial units from a source, vary from less than 1 kilometer to 10 kilometers. Most study windows have areas between 10 and 100 square kilometers. Often, the size of region is defined by a natural break in the underlying

population. For example, the boundary of a town or physical barriers such as rivers, mountains, or coastlines may affect the region size (and shape). Practical data acquisition problems may limit the region size. Furthermore, exposure and outcome data may be available for different regions.

2.2.2. Region Shape

When one assesses exposure to a single pollution source, and one assumes that distance is a surrogate for exposure, then a circular region centered on the source yields the least sampling bias for detecting directional trends, in that sampling is equal in all directions. Square, rectangular, or other polygonal regions do not provide such unbiasedness. Of course, if the putative source is not central to the region then a circular window has no advantage. If population structure dictates the region shape and size then a polygonal region may have to be adopted, although some advanced statistical techniques can be used to allow for population sparse regions in regular windows.

When one examines multiple pollution sources, a rectangular or polygonal region should suffice. However, one should make some effort to provide 'similar' sampling detail in all directions from the sources in case directional anisotropy is present.

2.3. Replication and Control

Few studies examine replicated realizations of disease events around pollution sources. The main use of replication in such studies should be to provide estimates of variability not available from single realizations. An alternative use of replication is to study other areas where potential pollution sources exist but where no evidence has been demonstrated for adverse health links to the source or sources.

If substantial hypotheses concerning an individual source are to be examined then control areas may be of some use. However, the use of replication to provide increased sample size by pooling, without examination of variability, only provides evidence for hypotheses concerning the sources in general, and not as individual sites. Local effects, which may be 'unusually' marked at an individual site may be swamped in such a pooled sample.

In any study of disease incidence within a population, one must take some account of population structure. A standard epidemiological case-control design can be used where individuals are selected as controls and matched to cases with respect to confounding factors (e.g. age and occupation). Another standard approach in the conventional analysis of small area count data involves the use of strata-specific standardized rates to represent the 'background' population effect. The ratio of observed count to expected count, based on such standardization, can be used as a crude estimate of region-specific relative risk.

An alternative approach is to utilize a disease or group of diseases which is thought to represent the 'at risk' population in the area but is usually unaffected by the type of pollution being considered. This approach is designed for point event data where a 'background' point event map of a 'control' disease is available. This method could also

be used with count data, where counts of 'case' and 'control' diseases are available.

The goal is to find a 'control' disease which affects the same population with respect to possible confounding variables (e.g. age, occupation, smoking, etc.) yet is unrelated to the exposure of interest. While the existence of such a 'control disease' is subject to epidemiological debate, if such data are available, the statistical foundation of the methods is sound.

In many non-geographical studies in epidemiology, it is common to assign *individual* controls to cases i.e. each case has an individual control that is matched to the case on a selection of variables such as age, gender or exposure history. Such matched case-control studies can be implemented within a geographical setting, and details of the statistical issues relating to these studies and putative source examples are available.

3. Problems of Inference

The primary inferential problems arising in putative source studies are (a) post hoc analyses, and (b) multiple comparisons.

The well-known problem of post hoc analysis arises when prior knowledge of reported disease incidence near a putative source leads an investigator to carry out statistical tests or fit models to data to 'confirm' the evidence. Essentially, this problem concerns bias in data collection and prior knowledge of an apparent effect. Both hypothesis tests and study region definition can be biased by this problem. However, if a study *region* is noted a priori to be of interest because it includes a pollution source, one does not suffer from post hoc analysis problems if the internal spatial structure of disease incidence did not influence the choice of region.

Although much recent work examines the statistical methodology appropriate for analysis for single disease types, there is little consideration of how to accommodate multiple 'health markers' in the investigation of putative sources.

The multiple comparison problem has been addressed in several ways. Bonferonni's inequality may be used to adjust critical regions for multiple comparisons but the conservative nature of such an adjustment is well-known. Multiple comparison problems have also been addressed by the use of cumulative p-value plotting to assess the number of diseases yielding evidence of association with a particular source.

-
-
-

Bibliography

Bithell, J. and R. Stone (1989). On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *Journal of Epidemiology and Community Health 43*, 79-85.

Cressie, N. (1993). *Statistics for spatial data* (revised ed.). New York: John Wiley and Sons.

Diggle, P., S. Morris, P. Elliott, and G. Shaddick (1997). Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, A 160*, 491-505.

Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Jour. Royal Statist. Soc. A 153*, 349-362.

Diggle, P.J., S. Morris, and J. Wakefield (2000). Point-source modelling using matched case-control data. *Biostatistics 1*, 1-17.

Esman, N.A. and G.M. Marsh (1996). Applications and limitations of air dispersion modeling in environmental epidemiology. *Journal of Exposure Analysis and Environmental Epidemiology 6*, 339-353.

Lawson, A., A. Biggeri, and F.L.R. Williams (1999). A review of modelling approaches in health risk assessment around putative sources. In A.B. Lawson, D. Bohning , E. Lesaffre, A. Biggeri, J.-F. Viel and R. Bertollini (Eds.), *Disease Mapping and Risk assessment for Public Health*, Chapter 17, pp. 231-245. Wiley.

Lawson, A.B. (2001).*Statistical Methods in Spatial Epidemiology*. New York: Wiley.

Panopsky, M.A. and J. A. Dutton (1984). *Atmospheric Turbulence*. New York: Wiley.

Thomas, D.C. (1985). The problem of multiple inference in identifying point-source environmental hazards. *Environmental Health Perspectives 62*, 407-414.

Biographical Sketch

Andrew Lawson is Professor of Biostatistics in the Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina. He is the author of some 40 journal papers and many invited book chapters. He has also authored 4 books in the area of spatial statistical methods in epidemiology. He is an advisor to the World Health Organisation (WHO) on Disease Mapping and Risk Assessment. His research interests range across statistical methods in environmental and spatial epidemiology, directional statistics and environmetrics.