Chapter 6 *

LINKING ENVIRONMENTAL AND HEALTH DATA: STATISTICAL AND EPIDEMIOLOGICAL ISSUES

6.1 Introduction
Exploration of associations between environment and health is an integral part of environmental epidemiology, either in the search for previously unknown dose-response relations, or to test hypotheses about such relations. The HEADLAMP methodology is an extension of this approach (WHO, 1995). Lying at the interface between epidemiology and public policy, it involves applying known dose-response relations, established in previous investigations and documented in the literature, to new empirical data as a basis for improved decision-making and policy support.

In general, the data used for environment and health linkage as part of HEADLAMP studies are derived from routine monitoring sources, although where necessary additional data may be collected from purpose-designed rapid surveys. In either case, the data often comprise series of data accrued over a long period of time, and are gathered in an aggregated form (e.g. at the small-area or regional level). The need to conduct aggregate data studies arises from the difficulty of acquiring individual level data, especially on environmental exposures and other covariates (Rothman, 1993). As such, the linkage of a health effect variable (e.g. excess mortality) to exposure and other characteristics of populations does not involve the direct use of individual records. Instead, the HEADLAMP methodology relies on analysing grouped data (Nurminen and Nurminen, 1999).

In the HEADLAMP approach, the aim of the environment and health linkage is not to discover new associations, nor to confirm suspected ones. Rather it involves using established scientific knowledge to assess the risks that exist, to identify need for action, to compare the choices available, and to monitor and evaluate the effects of such actions. As part of this process, the associations previously recognised in environment and health data are extrapolated to new data.

* This chapter was prepared by M. Nurminen
The basic method for this purpose is ecological analysis. In addition to the ecological method, there are a wide range of more specialized approaches, techniques and procedures which may involve, or be relevant to, environment and health linkage. Examples include the analysis of disease clusters (Rothman, 1990), studies of point-source exposures (Elliott et al., 1992), and quantitative risk assessment (Nurminen et al., 1999). Each of these may have value in particular circumstances, but each also involves problems and pitfalls of which the investigator needs to be aware. The use of one of the most important tools for exposure and disease mapping, geographical information systems (GIS) (Briggs and Elliott, 1995), is discussed in the next chapter. Together, these methods and tools give an investigator with ingenuity countless opportunities to analyse and exploit existing data at greatly reduced cost. In the process, considerable value is likely to be added to the data, to knowledge about environment-health relations in the area under study, and to the quality of decision-making.

Whatever method is used, if it is to be suitable for linking aggregated environment and health data, two important criteria must be met. First, the method must be simple, inexpensive to implement and operable with the available data, thus allowing rapid assessment. Second, it must produce statistically valid and scientifically credible results if they are to be used as a basis for action. This means that the method should be unbiased and sensitive to the variations in the data at hand. Ideally, it should yield results that agree with those that would be obtained from more comprehensive ad hoc studies (conducted on an individual level) and should provide some estimate of their accuracy and precision.

Section 6.2 below outlines the ecological method in general terms. Section 6.3 reviews time series analysis (TSA), which represents a special type of aggregate data method. Section 6.4 discusses the elements of quantitative risk assessment (QRA) and section 6.5 concludes that the linkage of environment and health data using ecological analysis is useful if used with care.

6.2 Ecological analysis

6.2.1 Background
The basic method for analysing aggregate-level data as part of HEADLAMP studies is ecological analysis. This involves the investigation of group level relations between environment and health, by analysing spatial or temporal variations in exposure and health outcome. First used in sociology (Robinson, 1950), it has often been criticised for producing fallacious results. Particular concern has focused on the potential bias which may be introduced by aggregation of data; a problem which Selvin
Statistical and epidemiological issues

(1958) termed the "ecological bias" or "ecological fallacy". Despite such theoretical shortcomings, however, ecological analysis has been widely used in environmental epidemiology, not least because it is relatively simple to perform, especially with the large, aggregated databases which are now available. For reasons of logistics and cost it may also be the only approach feasible where large population studies are required. Nevertheless, there has been a growing recognition that ecological or group level associations are not necessarily consistent with those measured at the individual level (Greenland, 1992). Thus, much of the subsequent discussion of ecological methods has focused on how to identify, deal with or avoid the various biases involved, and how to quantify their effects compared to individual level analyses. For the future, more extensive use of the method may be anticipated, stimulated in part by the development of new statistical techniques and GIS.

The ecological approach is a research technique used in observational studies to detect and recognize patterns of disease occurrence across space and time. It is also used to relate the rates of disease frequency to environmental, behavioural, and constitutional factors. The ecological design in epidemiology is also useful for the evaluation of intervention on risk factors for various diseases, for example, the effect of low-cholesterol diet on the future rate of ischaemic heart disease. Some environmental health problems are more readily approached by ecological studies than by general epidemiological studies. For example, the prevalence of asthma symptoms in relation to climate is applicable to ecological measurement (e.g. Hales et al., 1998), but the occurrence of respiratory symptoms associated with occupational exposure to airborne cobalt is less so. The reason for this is that the level of cobalt exposure of individuals in a worker population is also affected by personal hygiene, as cobalt can absorb through skin, whereas ubiquitous climatical pollution is an affliction of a population in the aggregate, rather than of its individuals. The ecological method thus derives epidemiological knowledge from the study of disease of human populations rather than the study of disease in human populations.

The grouping variate in ecological analyses is often a geographical region, although other factors such as time period, ethnicity, socio-economic class, etc., could also be used for grouping. The situations in which ecological studies are the appropriate design have been summarised in a series of methodological papers summarised in Poole (1994).

Given the availability of suitable exposure and health information, ecological analyses can be conducted in a number of different ways. The types of ecological study can be either explorative (disease mapping) studies, multigroup (disease-exposure correlation or regression) studies, or time-trend studies. Disease mapping can detect geographical disease
clusters without any direct incorporation of exposure information. The available exposure data allow an epidemiologist to study its association with disease outcome in a single population. Alternatively, one can compare the correlations or, preferably, the coefficients of regression models in two or more populations. In the multigroup comparison design, data on exposure to a risk agent and the health outcome are collected on a group basis for several regions. In either design, the data accrue in a relatively short span of time, but there are typically no multiple measurements over an extended time period. In time-trend ecological studies, a single population may be followed up in regard to its changes in exposure over time and the respective changes in the rates of disease over the same period of time.

Ecological analyses of dose-response relation can be potentially biased by several problems (e.g. model misspecification, confounding, nonadditivity of exposure and covariate effects, and noncomparable standardisation). Ecological correlations and rate estimates can be more sensitive to these sources of bias than individual level estimates, because ecological estimates are based on extrapolations to unobserved individual level data. In this chapter, emphasis is placed on the biases in group level estimates rather than on their counterparts in individual level designs. Thus the concern here is not so much with the utility of the ecological approach in its own right, as with its use as a proxy for individual based studies.

6.2.2 Advantages and disadvantages of ecological studies

Ecological studies prevail to be popular because they are often relatively easy to conduct using existing databases in a relatively short period of time. Thus, a judiciously implemented ecological approach can serve as a cost-effective alternative for screening or monitoring of many disease entities and environmental conditions across geographical areas. In practice, however, the true costs of this type of study are often hidden. Establishing and maintaining a monitoring system is expensive, and the apparent cost-effectiveness of this approach only comes from the ability to use uncosted, or subsidised, data from a pre-existing monitoring system.

Sometimes an advantage of the ecological approach is that it permits the study of very large populations (e.g. populations of entire countries). Nevertheless, the usefulness of an ecological analysis depends on the purpose of the study, whether it is basic science, public health, public policy, etc. For scientific issues of disease mechanisms, large populations are not necessarily needed. Moreover population probability samples also offer a better opportunity to study large populations without the limitations of ecological studies. When it is feasible to study large populations ecologically, relatively small increases in risk can be detected. The power of ecological studies, however, is not related to the size of the population...
Statistical and epidemiological issues

studied, but to the number of data points, the accuracy of the data and the power of the statistical analysis methods used. Even so, it is necessary to be cautious about concluding that very small risks are meaningful in practice, simply because they are statistically significantly elevated (see Nurminen, 1997b).

Plummer and Clayton (1996) have studied these important design issues, namely: (a) how large should sample surveys of population exposure be and (b) how should they be targeted on different sections of the study population? They summarised their results as follows: "The number of study populations has little relevance beyond a certain point, the power and precision being limited by the total number of disease events and by the size of the sample surveys used to estimate the distributions of determinants within populations." The determination of the optimal size of an ecological study requires also consideration of measurement error, the nature and effects of which are discussed below.

Ecological studies sometimes cover populations more markedly divergent in their exposures than those that can be readily obtained in studies of individuals. Limited within-population variability in exposure may also call for the study of multiple populations in a hybrid epidemiological investigation. For example, when cancer is studied there is the possibility of designed ecological studies in which population exposure is assessed by sample survey methods and compared with reliable cancer statistics. For a discussion of the statistical analysis of such multilevel studies, see Navidi et al. (1994) and Sheppard et al. (1996).

As noted previously, ecological studies are subject to unique biases not present in individual level studies. Therefore, the demand for methodological rigour is great. The various sources of bias in ecological data derive from linkage failures, i.e., an ecological study does not link individual disease events to individual exposure or covariate data (see, e.g. Nurminen, 1995a).

The ecological design provides no information at all on the joint distribution of the exposure and disease variate at the individual level. Thus, there is no way of knowing from the ecological data that individuals experiencing the health outcome have actually been exposed to the environmental risk factor. Inferences on individual level dose-response relations from ecological data are justified only under exceptional, rarely met conditions. Therefore, deriving individual level relations from ecological data should be viewed as a particularly tentative and exploratory process that may yield very tenuous and misleading results.

Problems may also exist with the available data. Routinely registered health event data (e.g. hospital discharges) may not suit the purposes of the ecological research in question, because of an unusable classification
system of diseases. It may also be difficult to define the population denominators (e.g. the catchment populations of hospitals) corresponding to the health event numerators. For a less severe health event, such as a mild asthmatic symptom, there may not be any records available at all.

Differences in the geographical basis of the available data may cause additional difficulties. Health data are usually available for administrative units such as municipal health care districts, municipalities or provinces, whereas environmental pollutants and other exposures are often available only for individual monitoring sites or for "natural" areas. Extra effort may be needed to create health and environmental data sets with comparable population subgroups. This may be done either by reallocating individuals to "pollution zones" based on their place of residence or, more commonly, by estimating pollution scores for each administrative area using mathematical models or spatial interpolation techniques. Some of the GIS techniques outlined in Chapter 7 are useful for this purpose.

6.2.3 Biases and problems in grouped level versus individual level studies
In this section, the various sources of bias in ecological data due to invalid study design or incorrect data analysis are considered. These include:
- Aggregation bias.
- Sampling error.
- Measurement error.
- Model misspecification
- Confounding.
- Nonadditivity of effects (effect modification).
- Noncomparable standardization.
- Temporal and spatial problems.
- Incorrect statistical or scientific inference.

Aggregation bias or cross-level bias refers to the incorrect estimates of exposure effect that result from the analysis of data aggregated across study groups (Robinson, 1950). Because the groups are typically exposed heterogeneously, aggregation bias is a more complex issue than a simple confounding by group (specification bias). The most recent attempt to solve this problem has been presented by King (1997). Because the geographical units on which the ecological sampling is based are divisible, ecological analyses are often done at several levels that may not give identical results. Unfortunately, aggregation bias cannot be identified by the examination of results using different aggregations. Ecological results that are similar at all levels of aggregation (e.g. county, economic area, state, region) can still be plagued by aggregation bias.
In addition to the sources of bias ingrained in individual level studies, ecological estimates of effect can be biased from effect modification by the group variate or confounding by the group variate. Covariates responsible for aggregation bias need not even be effect modifiers or confounders at the individual level (Greenland and Morgenstern, 1989).

The design of the sampling of the population in an ecological study has to account for sampling error (Cochran, 1977). This is a problem that has not been sufficiently touched upon in the literature on ecological studies. When the ecological estimates of exposure are based on sample surveys, they are subject to sampling error. If the studies are not based on routine data sources in which sampling errors are negligible (e.g. census data), then exposure variates have standard errors which will bias the regression coefficients. If estimates of the standard errors are available from surveys, these may be incorporated to correct for the bias. However, the sampled areas included in the analysis may differ in size and population density. To allow for the different amount of information contained in each group, a weighted regression should be used with weights proportional to the inverse of the variance of the observation unit.

The susceptibility of ecological estimates to measurement error can be a far more important source of uncertainty than the sampling error. Apart from basic demographic variates (such as sex, age, and vital status), most variates used in ecological analyses are measured with error. The design of the information on the samples in an ecological study is more complex than that of classical epidemiological (e.g. cohort) studies. This is so because the samples used to estimate the distributions of the disease, exposure and covariate distributions for an ecological study, are often independent. Therefore, the measurement errors that arise from this structure of an ecological study have to be considered separately for the exposure, disease outcome and covariates.

Measurement error has different effects on ecological and individual level studies. Independent non-differential misclassification of an exposure indicator will usually result in a biased estimate of the exposure effect that is directed away from the "no-effect" or null hypothesis in ecological studies. Intuitively, this happens because the exposure misclassification reduces the variation in the exposure prevalence across groups; although the group disease rates are unchanged, this effectively magnifies the exposure-effect relation (Sheppard et al., 1996). In individual level studies, random measurement error biases estimates of effect parameters toward their null value and overstates the precision of such estimates (Fuller, 1987).

It is necessary to exercise prudence before drawing firm conclusions about the role of any exposure component since measurement of the exposure level obtained by ecological means may be affected by large
random errors. Although it is recognised that such errors tend to bias the observed risk, insufficient attention has been given to the fact that the wrong variate can be identified as the main risk factor (when several variates are correlated with each other and measured with different random errors) simply because one variate may be measured with less error than the others. An analysis of the association of the airborne sulphuric pollutants SO$_2$ and SO$_4$ with daily hospital admissions in Ontario furnishes an example of this transfer of causality effect (Zidek et al., 1996). It is suggested that measurement errors be reduced whenever possible by repeating exposure measurements on the same units with different instruments having independent error factors.

Ecological studies in epidemiology typically deal with cause-specific mortality (and morbidity) rates rather than with total mortality. Therefore, misclassification of disease outcome can be a source of severe bias. This bias can be far greater than the sampling variability of the disease outcome. The following results are given by Greenland and Brenner (1993).

Imperfect disease specificity (i.e. false positive rate) induces no bias in the risk difference estimate, but this estimate is biased toward the null value by imperfect sensitivity. With disease misclassification, due to either imperfect sensitivity or to specificity, the linear regression estimate of risk ratio will be biased toward the null hypothesis (i.e. equal risks in the compared populations).

Another important issue in the study of ecological data concerns the sensitivity of these analyses to model mis-specification. The mathematical form of a model depends on many issues. The ecological relation in a particular group embodies the group means for the disease rate, exposure variate and covariate level across exposure-covariate strata. In general, this relation does not assume the functional form used in individual level studies (Richardson et al., 1987). This discrepancy may cause only little bias when the expected effects are small. Nevertheless, ecological summary rate ratios can be very sensitive to the chosen model form (Greenland, 1992). In contrast, the individual-level effect summaries of rate ratios appear insensitive to the choice of model structure (Maldonado and Greenland, 1993).

The choice of the regression model form also has implications on the epidemiological inferences. In linear-additive models, the estimated values of the disease rate parameters must be restricted so that they predict positive rates. To overcome the possible problem of extrapolation to negative values for rates, a logarithmic transformation of the rate parameter can be used. This transformation implies, however, that all continuous terms in the multiplicative model assume an exponential relation to one another. It is also difficult to find foolproof programs to fit these models. Fortunately, the
advances in the generalised linear (McCullagh and Nelder, 1983) and additive (Hastie and Tibshirani, 1990) models have opened new possibilities for more versa-tile modelling of ecological data.

Failure to identify, measure, or control important covariates of the dose-response relation is known as confounding. This problem is shared by cohort and case-referent study designs as well as all types of ecological design, but the problem is more perplexing in ecological studies than in individual level studies. This is true because ecological bias can be produced by other factors, such as effect modifiers acting independently of the confounders or tangling with their effects. Thus, the conditions for no confounding in ecological studies are logically independent of the conditions that guarantee no confounding in individual level studies. If the latter conditions are mistakenly applied in ecological studies, it can lead to omission of important covariates from the analysis. As Greenland (1992) argues, a covariate may be ignored at the individual level but not on the ecological level, or vice versa. There will be no ecological association of exposure distributions with disease outcome rates across groups if there are no exposure effects on either disease risk or the distributions of other risk factors by group (Greenland and Robins, 1994). This condition occurs even if, within each group, exposure levels are associated with the other risk factors.

If existing databases are used, the extent of those information sources imposes certain limitations. The use of routinely collected health and environmental data, by necessity, restricts confounder control possibilities to those covariates that have been measured. These variates usually do not include all the relevant covariates for the studied exposure-effect relation. Moreover, most covariates used in ecological regressions are either surrogates or crude measures of the true confounder. The problem is compounded by the need to measure the multivariate (joint) distribution of the confounders; univariate distributions of the covariates or a confounder score may not suffice to achieve full control of confounding.

The effect of exposure on disease outcome can vary according to the level of a covariate; this is termed "effectmodification". These covariates introduce statistical interaction with the exposure variate. Any individual level study that records exposure and covariate can analyse their interaction by including their product in the regression model. In an individual level cohort study, omission of the product term from the regression model induces no bias (Greenland, 1992). In stark contrast, omission of the group mean of the product term from the ecological regression model can lead to severe bias when this model is correct: The summary estimates of rate differences and rate ratios may even lie outside the range of their true covariate-specific values (Greenland and Morgenstern, 1989; Greenland
and Robins, 1994). Ecological bias caused by effect modification across areas can occur even when the number of areas is very large and there are no across-area or within-area associations of exposure with the covariate, and thus there is no confounding (Greenland and Morgenstern, 1991). In contrast, effect modification cannot by itself produce bias in individual level analyses (Miettinen, 1985).

There is need for standardization in ecological studies in epidemiology for variates whose distributions are not constant across population groups. This is important because published disease rates are invariably age-standardized, whereas published exposure rates are seldom standardized. In regression analysis of ecological data, the covariates need to be mutually standardized using the same standard distribution as used for the disease outcome. If a different standard (or no standardization) is applied to the covariates, the inclusion of the confounding variates in the regression model may even aggravate the bias. Similarly, if only the outcome rate is standardized, the bias may consequently increase. In addition, the exposure variate must be standardized using the same standard distribution; otherwise standardisation bias may result. (Greenland and Morgenstern, 1991)

In the study of chronic diseases using ecological studies it is important to consider the induction period. Ecological estimates of exposure may not correspond to the aetiologically relevant period for diseases with long induction times. If the exposure coincides more with that of the health outcome than with the relevant etiologic time, the relation is cross-sectional. In a longitudinal relation, exposures at some previous time(s) are considered. If the exposure level is stable over time, the distinction between the two relations is debatable. If it is very unstable, the distinction is critically important. Thus, there is need to focus on the most relevant time span of the aetiologic period.

Dynamic populations are characterised by in-migration and out-migration which may render the areal exposure estimates at earlier times poor proxies for the actual levels experienced by the study populations providing the disease rates (Greenland, 1992). One way of tackling the problem of migration would appear be to restrict ecological studies to more homogeneous populations. However, this restriction may lead to a vicious circle, because it generally means that the studied groups are smaller and, therefore, more easily subject to differential representation of the exposed or unexposed domains in the study base. Geographically static populations may also be unrepresentative of the wider population.

Failure to distinguish between the scientific object of an epidemiological research and the actual object of an empirical study may lead to incorrect inferences. In aetiological research, the object of inference is the same for both ecological and individual level studies. Although in
individual level studies the target effects are at the same level as the units of statistical analysis, the ecological analysis is coarser. Consequently, an ecological study can yield biased results for individual level effects and still be unbiased for ecological effects (Greenland and Morgenstern, 1989).

6.2.4 Strategies for minimisation of bias

In the preceding section, the sources and directions of various biases were considered. The primary strategy for the prevention of bias in ecological studies, as in epidemiological studies in general, must be the design of a valid study. If bias persists, there are limited statistical methods available for reducing bias in the analysis phase, although they will rarely eliminate bias entirely. They include:

- Multilevel modelling the exposure-effect relation.
- Coping with model assumptions.
- Influence analysis.
- Sensitivity analysis.
- Use of robust procedures
- Empirical Bayes methods.
- Correcting for nondifferential misclassification.
- Controlling for confounding.
- Modelling for nonlinearity and nonadditivity of effects.
- Adjusting for variates using comparable standardization.

Because of the developments in epidemiological study design and the progress of analytical multilevel modelling techniques, also called hierarchical regression (Greenland, 1998), it may be possible to alleviate the problems with inference based on grouped data by obtaining individual level data in samples of selected groups. Multilevel modelling allows for the simultaneous analysis of individuals and their ecologies. This approach examines the circumstances of individuals at one level and, simultaneously, the contexts or ecologies in which they are located at another level. The result is a strategy for the analysis and adjustment of aggregation effects in a regression analysis. In addition, the modelling provides a way of removing the bias due to the grouping variates if additional information about the individual-level covariances between the grouping variates is available. Pearce (1999), however, has argued that “epidemiologists need to learn to think in a multilevel way rather than just adding a multilevel modelling into their analytical toolkit.”

Most ecological studies are descriptive, and as such can be classified as exploratory data analysis. Influence analysis extends data description to exploration of the sensitivity of data summaries or exposure effects. It can, for example, be especially informative in ecological studies with small
groups, to visualise the impact of excluding some observations which stand out as statistical outliers.

Because ecological estimates are sensitive to biases, one should fit multiple regression models in an ecological analysis, especially when the true model form is unknown. Unfortunately, the parametric model for risk of disease is almost always mis-specified in practice. Greenland (1979) points out in the context of discussing the limitations of the logistic analysis of epidemiological data, that "... as with all statistical models, there is a danger that the ease of application of the model will lead to the inadvertent exclusion from consideration of other, possibly more appropriate models for disease risk." The choice of the model form has consequences on the effect estimates because of the frequent need in risk assessment for extrapolation to zero. In a sensitivity analysis, the model forms are varied to check the invariance of the results under new distributional assumptions.

An answer to the uncertainty about model specification is to employ robust methods (i.e. methods that function better than usual methods when the assumptions underlying the usual methods are violated). Random effects models can be used when the distribution of observations under the usual probability model show overdispersion, i.e. the mean exposure levels can exhibit extra Normal variation, or the disease occurrences can exhibit extra Poisson variation. Another proposed remedy is to use empirical Bayes methods which can successfully deal with several epidemiological problems, such as disease mapping, smoothing of unstable rates, and screening of multiple associations (Greenland, 1994).

Greenland and Brenner (1993) have provided a general method of adjustment for nondifferential misclassification of a binary exposure variate (e.g. smoker or nonsmoker) and disease outcome in ecological regression analysis. They derived simple correction formulae for the ecological regression estimates of risk difference and risk ratio. This method uses the concepts of sensitivity and specificity of exposure and disease measurement. The method can be applied specifically for exposure (Brenner et al., 1992a), disease outcome (Greenland and Brenner, 1993) and confounders (Brenner et al., 1992b).

The degree of confounding, and hence the strength of an effect, cannot be directly measured from the observed data; it should be evaluated against background disease risk, knowledge of subject-matter, logical argument, evidence from previous studies, and the particulars of the empirical setting in which the study is being conducted (Nurminen, 1997a). In general, the control of confounding in an ecological study is more demanding than in an individual level study because the measurement process for confounders is much more complicated. As in an individual level study, the ecological approach entails the problem that the crude measurement or approximate
measure for a confounder may be inadequate for achieving full control. Moreover, for potential confounders, such as diet, smoking and other lifestyle factors, multiple summaries of the joint distributions are needed for effective control. Unfortunately, the within-group joint distribution of the covariates is rarely available in ecological studies. In particular, the marginal summaries that may be available may prove to be too crude to provide effective control. If, however, confounder information is available in the disease registration system, then the analysis can be improved by stratification by the covariate (Brenner et al., 1992b).

The problem of confounding is aggravated if nonlinearity or nonadditivity of effects by the covariate is present. Ecological covariate summaries can be inadequate to detect and control confounding by a covariate with non-linear effects, and, also, when the effects are not additive (i.e. in the presence of effect modification). In addition, the covariate terms in the regression function must be adjusted to the same distribution as the disease and exposure using comparable standardization.

6.2.5 Designing, analysing and evaluating ecological studies
Ecological studies are based on a distinct methodological approach which sets them apart from individual level epidemiological studies. A number of specific factors thus have to be considered in either designing or analysing ecological studies, or in critically evaluating the end-results of such studies, (see Morgenstern, 1982; and Greenland, 1992):

- Ecological studies are much more sensitive to bias from model misspecification than are results from individual level studies. For example, deviations from linearity in the underlying individual level regressions can lead to inability to control confounding in ecological studies, even if no misclassification is present.
- Conditions for a covariate not to be a confounder differ in individual level and ecological analyses. Thus, a covariate may be ignorable at the individual level but not at the ecological level, or vice versa.
- In contrast to individual level studies, independent and nondifferential misclassification of a dichotomous exposure variate usually leads to bias away from the null hypothesis in ecological studies.
- Failure to mutually standardize disease, exposure, and covariate data for other confounders (not included in the regression model) can lead to bias.
- There is no ecological method available to identify or measure ecological bias.
- In the design of an ecological study, it is important:
to select areas are with populations that as homogeneously exposed as possible (i.e. minimize within-area exposure variation) by sampling smaller units for the analysis;

- to select populations which represent different extremes of exposure distribution (i.e. maximize between-area exposure variances);

- to select populations which are comparable with respect to covariate distributions; and

- to supplement, whenever feasible, approximately aggregated data with accurate data at the individual level in a hybrid epidemiological analysis.

In the analysis of ecological data it is important:

- to use weighted regression, instead of correlation, with weights proportional to the amount of information contained in each group;

- to include in the regression model all variates that are thought to be related to the grouping process;

- to examine multiple regression models with different and flexible structural forms beyond the standard linear form, such as exponential and product-term models, and nonparametric, smoothed curves;

- to test the basic assumptions in the model (i.e. the robustness of estimation techniques);

- to consider the ecological implication of different individual level model form specifications;

- to conduct an influence analysis by examining the effect of deleting from the analysis various areas with unusual outcome, exposure, or covariate combinations;

- to conduct a sensitivity analysis of ecological estimates to misclassification

- to take into account latency and induction periods separating causes and effects (e.g. consider the relevant exposures);

- to consider the effect of migration on areal exposure estimates; and

- to accompany ecological analysis by thorough consideration of biases unique to such an analysis, and to biases common to all epidemiological studies.

6.3 Time series analysis

Time series analysis (TSA) is a well-established technique in statistics. It was developed to a large extent for econometric applications, but has since been adopted in a wide range of disciplines. TSA is typically used to investigate patterns in series of observations, as a basis for identifying and quantifying causal relations. In environmental epidemiology, it is often applied to long sequential observations, such as mortality statistics, data
from morbidity registers (e.g. cancer registers, hospital discharge registers), or results from repeated health surveys. Eventhough there does not seem to be anything unique about time that is fundamentally distinct from any other covariate (e.g space) that defines the sampling units in an ecological study, the importance of time series warrants their analysis to be considered as a separate issue. In the case of HEADLAMP studies, TSA offers a valuable means of examining and comparing trends in environmental conditions and health effects in a study area, or to assess the effects of policy actions on environmental conditions and health outcome.

With simple data sets, it is a relatively straightforward method, and is supported by most well-equipped statistical packages. Where temporal patterns are complex, however (and thus where relatively complex models need to be used to describe the time series), however, it can be computationally and statistically demanding, and can pose severe problems for both implementation and interpretation. In recent years, it has been extensively applied in studies of air pollution and health, and thus efforts have been made to formalize and standardize the techniques used (e.g. Katsouyanni et al., 1995). Moreover, as temporal data series are extended and improved, the opportunities for time-series analysis will inevitably increase. Continued interest in the use of TSA may thus be expected.

Time series analysis looks at the relation between observations recorded at consecutive, usually equally spaced, discrete time points. While TSA is also a regression method, it predicts the health outcome not from independent covariates, but from values of the outcome at previous points in time. The minimal requirements are the abilities to plot the temporal series; derive new series (e.g. differenced series or smoothed series) and to plot these; to examine scatter plots of time-lagged values; to compute serial correlations periodograms; and to display these graphically. Current developments in graphical computing techniques for studying multi-dimensional relations will be valuable for TSA. Statistical computing aspects are especially important when the data sets used are large.

- Three basic approaches to time series analysis exist, namely:
  - Poisson autoregression analysis using generalized estimating equations (GEE);
  - Markov regression models using quasi-likelihood estimation (QLE); and
  - Poisson risk function model for time-stratified data using maximum-likelihood estimation (MLE).

It is beyond the scope and depth of this chapter to present the details of the statistics involved; it suffices here to outline in general terms how these models are applied in TSA.
**6.3.1 Regression models for time series analysis**

The time-series model can be understood as a subclass of the generalized linear model (McCullagh and Nelder, 1983) in which the exposure effects are multiplicative, the distribution of the errors is Poisson, and the link function is the (natural) logarithm. Thus, the model can be expressed as:

\[
\log[E(y_t)] = x_t' \beta
\]

where: \( y_t \) is the count of observed outcomes at time \( t \),
\( E(y_t) \) denotes the expected count,
\( x_t' \) is the (column) vector of covariates, and
\( \beta \), the vector of regression parameters, represents the effects of the covariates on the outcomes.

To account for the possibility of overdispersion and autocorrelation, the covariance matrix for the health outcomes on the units of observation is assumed to have a special form; the regression parameters are then estimated by the GEE (Liang and Zeger, 1986). GEE is used because the form of the joint distribution of the time-dependent measurements is so complex as to be intractable; that is, it cannot provide useful and interpretable information.

Overdispersion in Poisson counts can arise for at least two reasons. First, the risk of an adverse outcome occurring to an individual may not be equal for all individuals, but depends on previous events that happened to that individual; that is, it varies over time. The second reason is that the risk may remain constant over time but not necessarily be equal for all individuals.

Markov models can also be applied for regression analysis of time series data (Zeger and Qaqish, 1988). As serial observations are unlikely to be independent, in the Markov models the expected response at a given time depends not only on the associated exposure variates and covariates but explicitly also on health outcomes at previous times. The regression coefficients can be estimated using the QLE approach (McCullagh and Nelder, 1983). QLE allows one to estimate the regression relation without full knowledge of the error distribution of the response variate.

There is a fundamental distinction between the GEE approach, which is a "pure" regression model with autocorrelated errors, and the QLE approach, which is a mixed regression-autoregression model. Although these models may be considered as alternatives, in general the regression coefficients in the two models are different, because, as in any regression equation, the interpretation of a parameter depends on what other variates are included in the model. It is also normally inappropriate to assume that the error in the exposure variates is negligible. When data on measurement errors is lacking, estimates should be obtained from independent survey.
samples. An advantage of the QLE approach over the GEE approach is that competing models can be compared directly with each other using a deviance statistic.

A particularly problematic aspect of studying temporal relations arises when there are sharp peaks present of similar frequency in both response and exposure series. For example, in an epidemiological study of daily death rate and meteorological variates, seasonal fluctuations are likely to be present in all data sets. In other instances, long troughs or long-term trends may be present. TSA deals with this by studying the regressions separately in different seasons or periods. More commonly, though, TSA considers differences in adjacent time points, and the exposure-effect relation between them is modelled; it is then assessed whether any model so fitted accounts for all the relation present.

Previous use of Poisson autoregression analysis models has generally been based on the assumption that the series is time-dependent. However, it is not clear either that the GEE approach or the QLE approach has advantages over simpler model building procedures sufficient to compensate for their greater statistical complexity. All the autoregressive methods involve complex and computer-intensive estimation procedures. A much simpler way of dealing with temporal data may be to adopt the working assumption that repeated observations from a unit are time-independent of one another. One can then proceed by dividing the study data into subgroups (strata) and fitting a Poisson risk function model to the time-stratified series. In this approach, the assumption of constant risk (or rate) ratio can be alleviated by including time-dependent covariates in the linear predictor. Computational demands can be reduced by using the MLE methods available with existing software. Kuhn et al. (1994), for example, used this method for TSA and found that it compared favourably with the GEE approach.

### 6.3.2 Application of Poisson regression for time series

Given the potential use of Poisson regression to quantify time trends, it is worthwhile to consider some assumptions of the MLE method which may appear to be violated in the case of time-series data. Simple Poisson modelling requires that outcomes are independent. On first thought, it would seem untenable to assume that health outcomes occurring over time meet this requirement. For instance, the effects of social and environmental conditions are likely to persist at least in the short run. Poisson regression also assumes that the population subgroups are homogeneous with respect to the risk of adverse health outcome. This is another questionable assumption since the occurrence of, say, asthmatic or cardiac attacks do not occur at random but have predictable precursors and known patterns of risk.
Two points need to be emphasized here. The first is that the ordinary Poisson regression model requires that the study population meets the criteria of no overdispersion and heteroscedasticity conditional on the covariates. One effective way of removing overdispersion is to transform the data to a square root scale; this stabilizes the variance of the observed counts. The inclusion of time-dependent covariates may well result in conditional independence and help to define strata of homogeneous risk. Secondly, the Poisson regression allows the analysis of aggregate data to be comparable with the analytical methods used in cohort and case-cohort (or case-base) studies (Nurminen, 1992). Thus, although autoregressive time series analysis has been promoted as the preferred method in analyses of sequential observations over long periods of time, it may equally be argued that Poisson regression provides a simple and viable alternative for time-stratified data.

The minimum requirements for the epidemiological studies using time-trend ecological studies to be informative are basically the same as those required of a valid and precise epidemiological study in general. Inadequacies in the database and the sheer complexity of interactions among relevant variates both add to the problem of inferring the exposure-effect relation between pollution and health. Recall that the HEADLAMP approach was based on the idea of relying on an established relation, and that it should be applied on a local or national level to infer about excess risks. Thus it is not necessary to speculate on the kind of biological mechanism behind findings of such an associative nature. But even within this limited framework, at least part of the difficulties stem from methodology: an appropriate application of statistical methods for regression must necessarily cope adequately with the time series characteristics of pollution and health variates.

In general, the representation of confounders in the regression model should be guided by the concern for thoroughness of control, with the reservation that the efficiency of the study not jeopardised by the inclusion of covariates of the exposures which are not risk factors of the disease outcome. A major drawback of a time series design is the possible presence of unmeasured confounders. However, the time-trend study of short-term effects that uses long series of small units (days) often downplays such errors. An important feature of such studies is that the population followed up serves as its own control over time, and thus possible confounders can only be factors varying according to small time units (from day to day). Such factors can conceivably be meteorological and chrono-logical factors, which usually are accurately measured and easily recorded.

6.4 Quantitative risk assessment
Increasingly, authorities at the local, national and international level are faced with difficult decisions which involve weighing the social and economic benefits of technology against the health and environmental costs involved (McMichael, 1989). If these decisions are to be made on an informed basis, they require that health effects can be quantified. As a result, some form of quantitative risk assessment is necessary for regulatory purposes. Moreover, because the results of such assessments are often presented as a single number (for example, excess number of exposed disease cases), they give the appearance of scientific certainty and simplicity, both of which make the methods appealing to decision-makers.

In practice, however, the ability to quantify the health effects of development is often limited and valid methods of QRA are both uncertain and complex. For example, the methods are highly dependent on a series of assumptions and subjective choices which can have critical effects on the resulting risk estimates. Considerable care is therefore necessary in both using and interpreting results of QRA. For a review of methodological issues in epidemiologic risk assessment, see Nurminen et al. (1999).

Quantitative risk assessment can be defined as the application of a statistical relation between exposure and the associated health outcome to assess either the health risk to a population or the exposure level associated with a given risk. Thus, two main types of quantitative risk estimate can be distinguished:

- Risk analysis, which involves computation of the risk corresponding to a given level of exposure or dose $\delta$, for example, expressed in terms of excess risk or the number of extra disease cases; and
- Hazard analysis, which involves calculation of the exposure or dose corresponding to a given level of risk $\delta$, for example, the exposures estimated to cause adverse health outcomes in a certain percentage of exposed subjects.

Risk analysis may be also applied at two different scales. Individual risk refers to the probability that an individual will develop a disease as a result of exposure in a specified time period. Population risk refers to the expected number of cases of disease attributable to exposure in the population under study in a specified time period. These two measures may have different regulatory implications: the regulatory authorities may wish to evaluate either the risk to individuals who are exceptionally highly exposed or that to a large population whose average exposure could be much lower.
6.4.1 Uses and uncertainties of quantitative risk assessment

Risk analysis is not a true linkage method in the sense that local health data are not utilized. Instead, it uses a predefined association between exposure and health outcome to determine the risk to an exposed population. The relation between exposure and health is usually derived from independent studies, either within the study area or, more commonly, elsewhere. The particular advantage of risk analysis is thus that it can be applied in areas where insufficient health outcome data are collected to allow the relation between exposure and health to be determined locally. By the same token, QRA methods are the least resource-intensive, the easiest and the fastest to use of all the methods considered here. The success of the risk assessment process, however, depends on a number of issues such as the choice of the risk prediction models and the adequacy of exposure assessment. All of these are subject to large uncertainties, though the exact form and magnitude of these problems vary depending on the particular context and purpose of the analysis.

One of the most important difficulties in QRA lies in obtaining reliable estimates of the exposure-response relation. Results from epidemiological studies of one population cannot always be directly applied to others, due to differences in the range of exposures involved, in the methods of exposure estimation used, in the socio-economic contexts in which exposure occurs and in the baseline status of the populations concerned. A relation for exposure to air pollution derived from a developed country or city, for example, is likely to underestimate the risks in developing countries, where the baseline health status is poorer (Ostro, 1994). Similarly, differences in the way in which exposure or health outcome are defined or measured in different areas (e.g. in the design of the pollution monitoring network, the specific definition of the pollutants measured, or in diagnosis) may make it difficult to transfer relations from one area to another.

Particular care is also needed where the health outcome of concern is potentially related to more than one exposure. Both particulate matter and \( \text{SO}_2 \), for example, are known to contribute to respiratory diseases. In many areas, levels of the two pollutants are also highly correlated. When modelling the contribution of both, only one variable will remain statistically significant – the effect of the second will be subsumed within the first. When modelled separately, on the other hand, they may both show significant associations with health outcome. Summing these separate estimates of the effects will clearly exaggerate the estimated effect (e.g. the likely number of cases). Ideally, therefore, some measure of the combined effect should be obtained, by adjusting for the effect of the second exposure. In practice, this is often difficult, and in these cases a more
A further source of uncertainty in QRA is the presence of population heterogeneity. In environmental health linkage, risk factor data are usually collected and presented at high levels of aggregation. Aggregated risk estimates of this type can only be extrapolated back to the individual level if the population concerned is homogeneous. In reality, homogeneity within any population rarely if ever exists. Unrecognized risk factors may be expected to subject different people to different background disease risks. As a result, individual risks may differ substantially from those implied by the aggregated data. Usually, variance estimators tend to be upwardly biased when risks are heterogeneous, rather than low. In undertaking a risk analysis, therefore, one should always check for hidden heterogeneity before presenting aggregate population statistics. If heterogeneity is discovered, then population risk estimates based on the aggregate data may be misleading. The populations should either be subdivided into more homogeneous subpopulations, or the statistics should be presented with due cautions for interpretation.

6.4.2 Presentation and interpretation of results of risk assessments
The results from any risk assessment clearly need to be communicated to the decision-maker in an appropriate form. This implies that the results are both clearly presented, yet also suitably qualified with regard to their reliability. The interpretation of the results, both by the risk assessor and the risk manager and later by the governmental and non-governmental organizations as well as the general public, may be critically dependent of the methods used to present the results. This is especially crucial in linking environment and health data since the decision-makers may not be well versed in the specialized statistical methods used. Moreover, there is the need to present the linkage results in such terms that they can be easily transformed to inputs for a societal or an individual cost-benefit analysis, or disseminated to other stakeholders (e.g. the public).

At present, there are no standardized procedures for analysing and presenting results from environmental and health linkage. To a large extent, this reflects the many different methods used to analyse the data, and the inherent differences in the data themselves. As a result, a standardized approach for the linkage of environmental health data is often neither feasible nor necessary. It may not be feasible because of unresolvable differences in the data or methods available; it may be unnecessary because the study concerned does not involve comparisons across different areas or periods.
Standardization of methods is nevertheless beneficial insofar as it facilitates comparability. The diversity of analytical techniques so far applied in time-trend studies of air pollution and health, for example, has tended to hinder direct comparisons of the results, and made it difficult to derive general estimates of exposure-effect relations (e.g. from a meta-analysis). Lack of standardization also makes it difficult to verify the results of individual studies (e.g. by comparison with studies elsewhere) and reduces the opportunities to re-use the data at a later date. Standardization thus offers the possibility of obtaining added value from the data, and thus of improving the cost-effectiveness of data collection. One of the rare attempts to establish standardized procedures for time series analysis was the EU-funded APHEA project. This developed a standardized methodology to analyse data from 15 cities, representing a range of social, environmental and air pollution conditions across ten countries (Katsouyanni, 1995).

It is beyond the scope of this chapter to discuss how best to present the results of statistical analyses because they are well covered in many textbooks (e.g. Gore and Altman, 1982). Similarly, the technicalities in quantifying human health risks are not considered here because they have been described in books on risk assessment (e.g. Cox and Ricci, 1989). It is, however, useful to examine some of the general issues involved in the presentation of the results of linkage studies, as a basis for better informing the decision-maker.

The result of most interest to the health agency or risk manager in arriving at a decision is usually the quantitative estimate of exposure effect on health risk. It is this effect estimate which provides the platform for subsequent policy action. Two quantitative measures of effect are widely used:

- **Change in individual risk**, i.e. the increased or reduced likelihood of an individual experiencing a specified health effect due to a change in exposure level; and
- **Disease burden**, i.e. the number of excess cases of the specified health effect ("body count").

Table 6.1, for example, shows the average worklife risk of lung cancer for an individual exposed to silica, whereas Table 6.2 shows the excess numbers of lung cancer in the exposed population for both the currently prevailing exposure levels and for the lower control limits. To provide some perspective, the results of risk assessment are often expressed as hypothetical changes in risks. Thus, a risk-analyst might interpret the results of Table 6.2 as follows. Introduction of, and adherence to, an exposure standard of 0.2 mg m$^{-3}$ would produce a 22% reduction in the excess
Table 6.1. Average lifetime risk of lung cancer for silica-exposed men employed from age 20 to 60 years.

<table>
<thead>
<tr>
<th>Exposure level</th>
<th>Estimated risk (%)</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>1.87</td>
<td>0.21-31.3</td>
</tr>
<tr>
<td>( \leq 0.2 \text{ g m}^{-3} )</td>
<td>1.34</td>
<td>0.20-3.98</td>
</tr>
<tr>
<td>( \leq 0.1 \text{ g m}^{-3} )</td>
<td>0.83</td>
<td>0.15-1.92</td>
</tr>
</tbody>
</table>

Source: Leigh et al. (1997)

number of lung cancer cases. Alternatively, if the exposure standard was set at 0.1 mg m\(^{-3}\), a 46% reduction would be predicted.

The methods used for risk estimation inevitably give only approximate projections of risk, for they usually involve a myriad of assumptions, which cannot easily be verified. The presentation of simple point estimates of the expected risks and excess numbers thus tends to give a misleading impression of precision. Instead, it is important to provide clear information on both the assumptions and limitations involved. Cox and Ricci (1989), for example, suggest the following guidelines for the presentation of risk estimates:

- Risks should be presented in a sufficiently disaggregated form (showing risks for different subgroups) so that key uncertainties and heterogeneities are not lost in the aggregation.
- Confidence bands around the predictions of statistical models are useful, but uncertainties about the assumptions of the model itself should also be presented.
- Both individual risks and population risks should be presented, so that the equity of the distribution of individual risks in the population can be taken into account.
- Any uncertainties, heterogeneities, or correlations across individual risks should be identified.
- Sensitivity analyses should be used to assess the effects on estimates of the key assumptions involved.
Table 6.2. Excess lung cancer cases in a dynamic working population of 136,400 men exposed to silica in a 40-year follow-up period.

<table>
<thead>
<tr>
<th>Exposure level</th>
<th>Estimated number</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>1960</td>
<td>300-7760</td>
</tr>
<tr>
<td>≤ 0.2 g m⁻³</td>
<td>1524</td>
<td>270-3530</td>
</tr>
<tr>
<td>≤ 0.1 g m⁻³</td>
<td>1058</td>
<td>210-2200</td>
</tr>
</tbody>
</table>

Source: Leigh et al. (1997)

Linking environmental exposures to health outcomes is frequently achieved through the use of a regression model, for example a multiple logistic regression. Whatever method is used, presentation of results after allowance for covariates should be in a form similar to that which would be used if no covariates were included in the risk function. Merely quoting the coefficients from the logistic model does not achieve this and is in any case artificial, since the logit transformation would not be necessary if there had been only the one risk factor of interest, and no covariates. This does not mean that the risk-odds ratio would not be useful as an auxiliary parameter in risk modelling. The analyst should, however, also provide more informative measures of exposure effect, such as the absolute excess risk (risk difference) or the relative excess risk (risk ratio minus one)(see Nurminen, 1995b).

A minor, yet more than cosmetic, point in presentation of results from QRA is the number of significant figures. In this context, the inherent precision of the results needs to be acknowledged. It is not sensible, for example, to give a result as "49.35 expected disease cases per year" when the probable range is from 10 to 200. It might even be better not to give a single point estimate, but only to indicate the approximate confidence bounds. In presenting the results of a meta-analysis, the overall mean value can be shown along with the ranges for the lower and upper confidence limits.

QRA frequently present information in terms of probability measures. Probability distributions can be difficult for a nonspecialist to interpret. Although a plot of cumulative incidence rate (estimates of risk) allows one to read the median (and the percentiles of the distribution), the mean value cannot be determined from the plot. To avoid misinterpretations, therefore, it is important to present a plot of the cumulative distribution together with a graph of the incidence density curve, using the same horizontal scale, and to show also the mean risk on both curves (Ibrekk and Morgan, 1987).
To be of use for health policy making, epidemiological data often need to be interpreted. Traditional epidemiology is mostly concerned with the increased incidence associated with exposure to a risk factor, whereas policy-makers are more interested in the reduction of risk after the cessation of exposure. The importance of a risk factor for the incidence of a disease in a population is usually expressed as the aetiologic fraction, i.e. the proportion of the total incidence of the disease that can be attributed to that risk factor in the population (Miettinen, 1985). This indicates the proportion of incidence that could be prevented by the total elimination of that risk factor within the population.

In practice, prevention measures are rarely able to eliminate completely the prevalence of an environmental risk factor. As a result, a more useful measure is the potential impact fraction (Morgenstern and Bursic, 1982). This indicates the incidence that is avoided by a preventive intervention as a proportion of the incidence that would have occurred in that population without intervention. The potential impact fraction can be calculated when the prevalences of exposure to a risk factor in the population and the corresponding incidence density ratios or risk ratios are known.

In the traditional epidemiologic literature, the term potential impact fraction is often used to imply an immediate elimination of excess risk after termination of exposure. In reality, this risk reduction may take many years to achieve, due to the lag effects involved. Ideally, therefore, estimates of effect should incorporate a time dimension. For this purpose, a methodology based on the preventive impact fraction has been developed (Gunning-Schepers, 1989). This comprises a computer simulation model, PREVENT (Gunning-Schepers et al., 1993), that can estimate the health benefits for a population of changes in risk-factor prevalence. Results are presented in graphical or tabular form and include: the intermediate output variates (aetiologic fraction, trend-impact fraction, and potential-impact fraction) and the final output variates (disease-specific mortality, total mortality, disease-specific mortality difference, potential years of life gained, actual years of life gained, survival curves, and life expectancy at birth).

A preventive intervention programme is often difficult to sell politically since its effects take so long to become apparent. Indeed, in many cases, the effects are not expressed as real reductions in risk because of the demographic changes in the target population over time. This does not mean that prevention will have no beneficial effect. It does mean, however, that in order to see the effects it is important to show what would happen without the preventive intervention, and not merely to compare predicted effects with the current level of mortality. The potential utility of simulation models such as PREVENT in this respect lies in their ability to
provide more precise quantification of effect estimates over time, and to take account of multiple risk factors and possible effects of demographic changes on the effects of intervention (Gunning-Schepers et al., 1993).

Although risk estimates produced by risk analysis have traditionally been used as the justifiable basis for regulating risks, the public's perception of risk is much broader than the "body counts" on which the quantitative risk assessments have focused. The public frequently misperceives risks because of the biases in the information to which they are exposed (e.g. the news media, government reports and industry reports). The public also perceives risk in a much wider context than that used in environmental epidemiology, i.e. perceptions reflect dread of the unknown, social and political impact, outrage and stigma. This difference in risk perception calls for two-way communication between the risk-analysts, risk-managers and other policy-makers, on the one hand, and the general public on the other (Morris, 1990). Useful guidelines and suggestions on how to communicate results of QRA to the public have been published by the US Environment Protection Agency (Covello and Allen, 1988). These list "cardinal rules" for effective risk communication. In addition, a useful guide designed for industrial plant managers is available, which describes the technical information to be presented and provides guidelines for explaining risk-related numbers and risk comparisons (Covello et al., 1988).

6.5 Conclusions
The linkage of environmental and health data (or either of these with covariate data such as socio-economic or demographic information) is a vital part of the HEADLAMP approach. Unlike in traditional epidemiology, its aim is not to seek new environmental-health relations or confirm hypotheses; rather it is to use existing knowledge on such relations to help inform management and policy decisions, and raise awareness about the associations between environment and health. The methods are thus used essentially as a means of describing and monitoring the relations between environment and health, and to help assess and demonstrate the existing risks to the population concerned.

Any such data linkage must nevertheless be undertaken with care, for the relations between environment and health (whether expressed geographically or in terms of time trends) are often complex and fraught by uncertainties. Without an understanding of these complexities, it is all too easy to misinterpret the data. On the one hand, this may lead to complacency and lack of action, if risks are not correctly identified; on the other it may cause unnecessary anxiety and fear, if non-existent risks are inferred. It is important to recognize that these dangers may be created simply by presenting environment and health data together, for it is human
nature to search for associations. Since most observers will be unaware of
the complexities and subtleties of the data, misinterpretation is almost
inevitable. Data linkage thus needs to be recognized as a powerful but
treacherous tool. Applied carefully and correctly, it can greatly strengthen
decision-making; used carelessly, it will mislead. It is incumbent on the
analyst, therefore, to ensure not only that environment and health linkage is
conducted rigorously, but also that the results are presented and explained
clearly and unambiguously.

6.6 References
Brenner, H., Greenland, S. andfSavitz, D.A. 1992a The effects of
nondifferential confounder misclassification in ecologic studies.
*Epidemiology* 3, 456-9.
nondifferential exposure misclassification in ecologic studies.
*American Journal of Epidemiology* 135, 85-95.
Briggs, D.J, and Elliott, P. 1995 The use of geographical information
systems in studies on environment and health. *World Health
Wiley.
Covello, V. and Allen, F. 1988 *Seven Cardinal Rules for Risk
Communication*. USEPA, Washington, D.C.
Covello, V.T., Sandman, P.M. and Slovic, P. 1988 *Risk Communication,
Risk Statistics, and Risk Comparisons: A Manual for Plant
Cox Jr., L.A and Ricci, P.F. 1989 Risk, uncertainty, and causation:
Assessment of Environmental and Human Health Hazards: A
and Environmental Epidemiology: Methods for Small-Area Studies.*
Oxford University Press, Oxford.
Gore, S.M. and Altman, D.G. 1982 *Statistics in Practice*. British Medical
Association, London.
Greenland, S. 1979 Limitations of the logistic analysis of epidemiological


Navidi, W., Thomas, D., Stram, D. and Peters, J. 1994 Design and analysis of multilevel analytic studies with application to a study of air pollution. *Environmental Health Perspectives* 102 (Suppl. 8), 25-32.


