A method for spatially moving correlation analysis

Bjørn Bølviken

Geological Survey of Norway, NO-7491 Trondheim, Norway

E-mail: bjorn.bolviken@ngu.no

ABSTRACT

A statistical method for spatially moving pair-wise correlation analysis is described. Examples are given of demonstrated associations for rates for multiple sclerosis versus environmental data for indoor radon and fallout of atmospheric magnesium. It is concluded that the method may disclose spatial geochemical associations that are not easily detected by other statistical techniques.

INTRODUCTION

Geomedicine, also called medical geology, is the science dealing with the influence of natural environmental factors on human and animal health (Bølviken 1998). A number of endemic diseases are connected with features of the environment, of which the relationships between caries and fluorine deficiency and between goitre and iodine deficiency are the most well known in Norway. Statistical analysis of the correlation between epidemiological and spatially distributed environmental data is often applied in order to cast light upon the pathogenesis of diseases with incompletely known aetiology. For large survey areas such correlations may vary from place to place, in other words be different within different sub-areas. However, the delineation of such sub-areas is normally not known beforehand. This calls for a statistical method of spatially moving correlation analysis. Such a method has been developed at the Geological Survey of Norway (Nilsen 1992, Ukkelberg et al. 1994, Lomheim 1996, Bølviken et al. 1997). The present paper gives a short description of this method together with some examples of application results.

MOVING PAIR-WISE CORRELATION ANALYSIS

Description of the method

The method can be applied to any set of digitised data consisting of pairs of (1) disease incidence rates and (2) an environmental parameter, both obtained at the same locations within a survey area. A circular window is defined around an observation station (S_1, coordinates X_1 and Y_1) on a map of the survey area in such a way that it encompasses a given number (n-1) of the nearest neighbouring stations. The total number of stations within the window is thus n. The figure n is chosen by the user, and may be any number greater than 2 and less than the total number N of stations within the survey area, see Fig. 1. The correlation coefficient is calculated for the n stations of the sub-area within the window (n-*2* degrees of freedom). The centre of the circle is then moved to a neighbouring station (S_2, coordinates X_2, Y_2), and the correlation coefficient is calculated for the n locations of the new position of the window. This procedure is repeated until a value for the correlation coefficient is obtained for every station (S) within the survey area.

Spatial distribution of correlation coefficients

The results of moving correlation analysis can be plotted on a map by symbols at the (X, Y) coordinates of the central station (S) of each position of the window. Such a map illustrates how the correlation coefficient varies between sub-areas. Two examples of results obtained with the method are given in Fig. 2. These results are based on the raw data in Fig. 3.

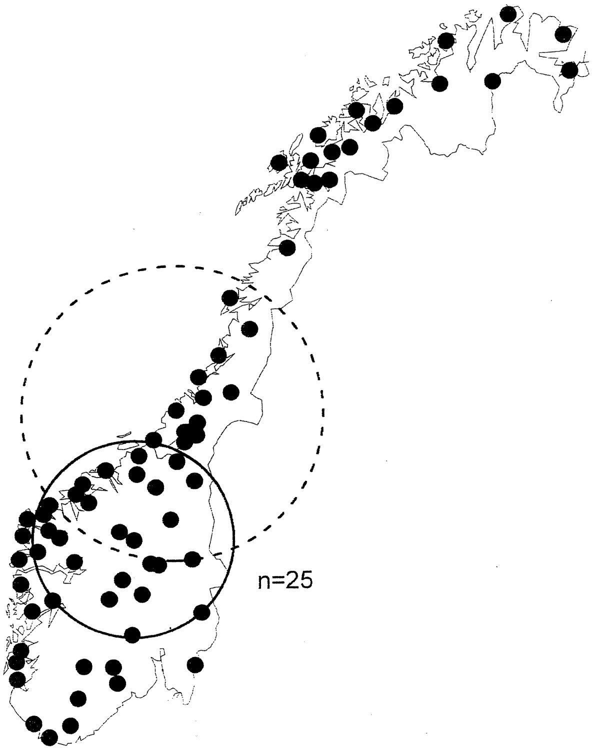

Figure 1. Principles of moving correlation analysis.

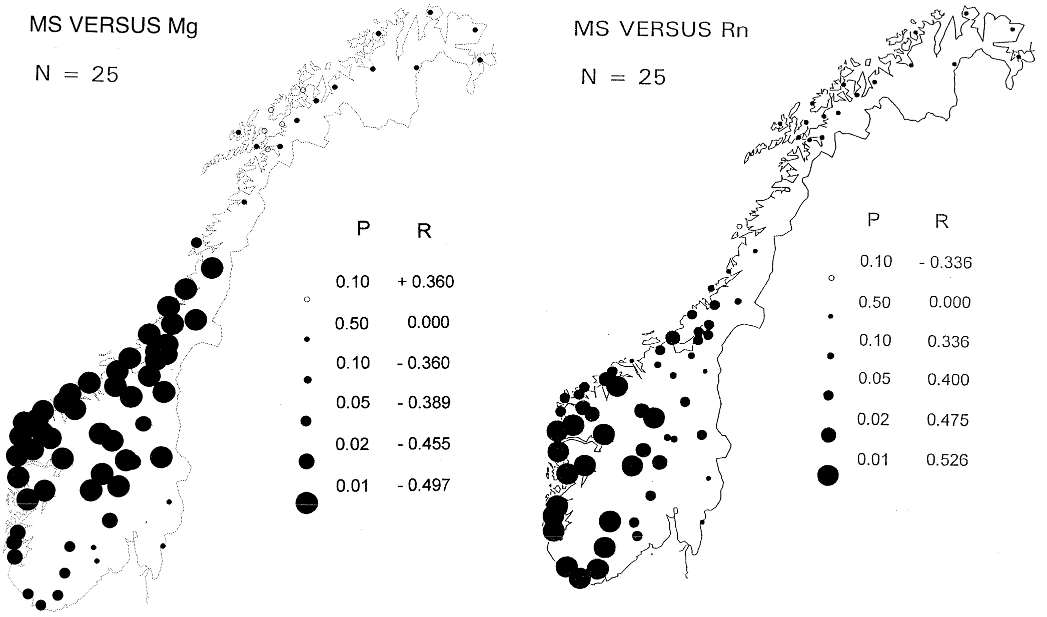

Figure 2. Moving Spearman Rank correlation coefficients for Norwegian rates of multiple sclerosis versus (left) atmospheric fallout of magnesium and (right) contents of radon in indoor air. After Bølviken et al. (2003). Copyright: Karger AG.

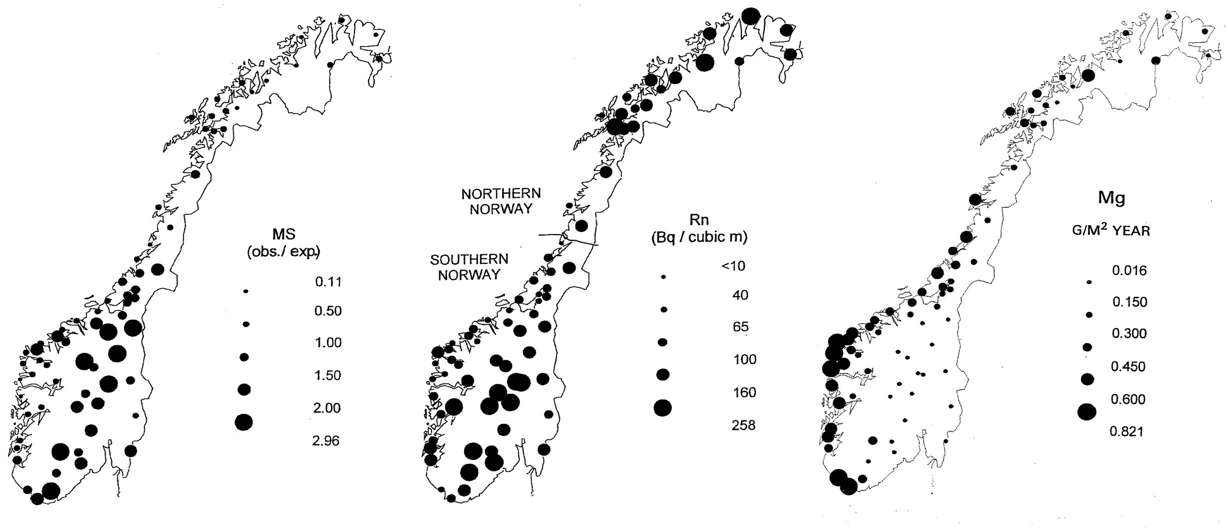

Figure 3. Distribution in Norway of (left) rates of multiple sclerosis, (middle) contents of radon in indoor air of dwellings and (right) fallout of athmospheric magnesium. Data from 73 rural municipality aggregares, each with a population of 10,000+. After Bølviken et al. (2003). Copyright: Karger AG.

Significance of obtained results

The following procedure can be used in order to test the significance of an obtained empirical distribution of correlation coefficients.

The environmental parameters in a given set of empirical data are randomly permutated 1000 times (or more) without repeating the same combination of data. Moving correlation coefficients are then calculated for each set of permutated environmental parameters versus the real set of epidemiological data using the coordinates of the epidemiological data for locations. For each simulated distribution, the correlation coefficients are put in order of increasing values, named V_1, V_2, V_3, …, V_N. Frequency distributions, F_1, F_2, F_3, …, $F_{1000,}$ are now calculated for each set of V, F_1 having the lowest, and F_{1000} the highest median. An adequate selection of these frequency distributions, for example F_{10}, F_{50} and F_{950}, F_{990}, can be plotted on probability paper, see Fig. 4, in order to illustrate the (in this case, respectively, 1%, 5% and 95%, 99%) significance levels of the various values obtained for the empirical correlation coefficients for the following zero hypothesis:

$H_0 1$: The empirical correlation coefficients are not significantly different from zero and randomly distributed.

Anomalous clusters of high correlation coefficients can be defined subjectively and tested by a zero hypothesis:

$H_0 2$: The average correlation coefficient for a selected anomalous area is not significantly different from the average correlation coefficient of permuted data from the same area.

Number of locations within the moving window

The sub-areas overlap; the degree of overlapping depends on the number *n* of locations within the moving window, see Fig. 1. In order to avoid problems with heterogeneous parameter distributions estimating significance levels at various locations, the number of stations within each position of the moving window (*n*) should be kept constant throughout the entire survey area. For inconsistent sampling density over the survey area, this means that the size of the sub-areas will vary. For a given set of data an optimal value of *n* should be tested by trial and error. A low *n* produces relatively high resolution and high variability in the results, while a high *n* gives low resolution, but more regular results, as is clearly indicated by the examples in Fig. 5.

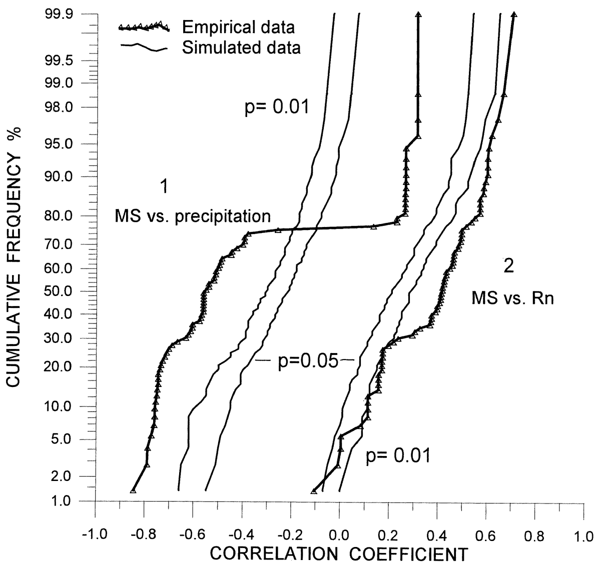

Figure 4. Cumulative frequency distribution plots of moving correlation coefficients for rates of multiple sclerosis versus (left) amounts of precipitation which correlates with fallout of Mg, and (right) Rn in indoor air. Data from stepwise moving windows, see Figs. 1 and 2. The simulated curves are obtained from permutations of the empirical data and indicate significance levels of p=0.05 and p=0.01.

SOME COMMENTS TO RESULTS OBTAINED BY THE METHOD

1. Significant correlations disclosed by this method may lead to formulation of new aetiological hypotheses. However, the user should always consider that the results may be caused by confounders and should, therefore, not by themselves be taken as a proof of causal relationships. Furthermore, similar to the outcome of other types of ecological studies, the results are valid only for population groups, and are not necessarily applicable for individuals (Morgenstern 1982).

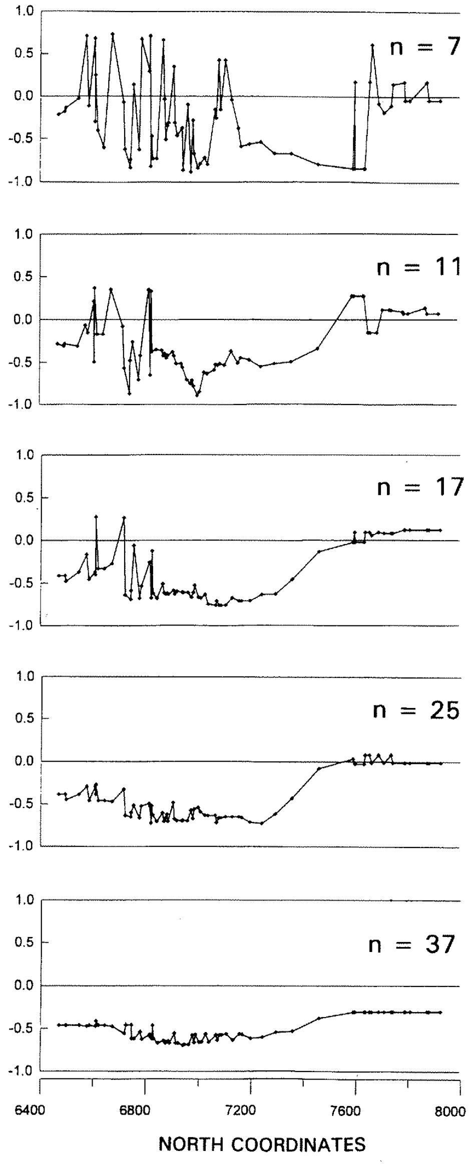

Figure 5. Moving correlation coefficients of rates of multiple sclerosis versus yearly fallout of atmospheric magnesium. Data from 73 rural Norwegian municipality aggregates. *n* = 7, 11, 17, 25, and 37 indicate number of observation points within the moving window. The values of the coefficients are interconnected with lines in order to illustrate their trends. After Bølviken et al. (1997). Copyright: Chapman and Hall.

2. The results of the testing of the method show that the correlation coefficient between a disease and an environmental parameter may vary between regions, being significantly high in some areas while insignificant in others. In fact the correlation in different subareas may even be of opposite sign (Lomheim 1996, Bølviken 1998). These features indicate that misleading results may easily be obtained if a geomedical correlation analysis is performed for a too large survey area as a whole. Possible significant correlations in some regions may then be drowned out. It is also indicated that a significant correlation found in a given area should not *a priori* be expected for other areas.

3. The most significant geomedical correlations may sometimes be found where an apparently beneficial parameter have relatively high values and an apparently harmful parameter have relatively low values, but where the disease rates simultaneously show relatively low levels. Such connections between diseases and environmental agents as well as other suggestive results warrant follow-up work with traditional epidemiological methods in order to see if they perhaps indicate causal relationships.

Conclusion

The described method for spatially moving pair-wise geomedical correlation analysis may disclose covariations that are not easily found by other statistical techniques. Intervention should not be realized until follow-up by more traditional epidemiological methods, such as case-control studies, has given interesting results. In the future, the spatially moving strategy could possibly be further developed applying multivariate, instead of bivariate statistical analysis.

Referanser

Bølviken B, 1998. Geomedicine (In Norwegian, English summary). *Nor J Epidemiol* **8**: 7-17.

Bølviken B, Nilsen R, Ukkelberg Å, 1997. A new method for spatially moving correlation analysis in geomedicine. *Environ Geochem Health* **18**: 143-53.

Lomheim L, 1996. Undersøkelse av samvariasjon mellom forekomst av multippel sklerose og noen miljøparametre i Norge (An investigation of co-variations between multiple sclerosis and some environmental parameters in Norway). In Norwegian. Institute of Geology and Mineral Resources Engineering, Technical and Scientific University of Trondheim. 55 pp, 6 appendices.

Morgenstern H, 1982. Uses of ecological analysis in epidemiological research. *Am J Public Health* **72**, 1336-44.

Nilsen R, 1992. Regional kartlegging av samvariasjon mellom geoparametre (Regional mapping of covariation between geoparameters). In Norwegian, English summary. Geological Survey of Norway Open File Report 92.263, 13 pp, 3 appendices, 11 maps, 6 figures and 7 tables.

Strand T, Green BMR, Lomås PR, Magnus K, Stranden E, 1991. Radon i norske boliger (Radon in Norwegian dwellings). In Norwegian, English abstract. National Institute of Radiation Hygiene, Østerås, Norway, 24 pp.

Ukkelberg Å, Bølviken B, Steinnes E, 1994. Multippel sklerose og geokjemiske miljøparametre (Multiple sclerosis and geochemical parameters of the environment). In Norwegian. Geological Survey of Norway Open File Report 94.099, 14 pp, 18 appendices.