Feature Article

A New Protocol for Evaluating Putative Causes for Multiple Variables in a Spatial Setting, Illustrated by Its Application to European Cancer Rates

ROBERT R. SOKAL,[1]* NEAL L. ODEN,[2] MICHAEL S. ROSENBERG,[3] AND BARBARA A. THOMSON[4]
[1]*Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794-5245*
[2]*EMMES Corp., Potomac, Maryland 20854*
[3]*Department of Biology, Arizona State University, Tempe, Arizona 85287-1501*
[4]*Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3*

Dedicated to Professor F. James Rohlf in recognition of his numerous contributions to statistical approaches in biology.

ABSTRACT We introduce a statistical protocol for analyzing spatially varying data, including putative explanatory variables. The procedures comprise preliminary spatial autocorrelation analysis (from an earlier study), path analysis, clustering of the resulting set of path diagrams, ordination of these diagrams, and confirmatory tests against extrinsic information. To illustrate the application of these methods, we present incidence and mortality rates of 31 organ- and sex-specific cancers in Europe; these rates vary markedly with geography and type of cancer. Additionally, we investigated three factors (ethnohistory, genetics, and geography) putatively affecting these rates. The five variables were correlated separately for the 31 cancers over European reporting stations. We analyzed the correlations by path analysis, k-means clustering, and nonmetric multidimensional scaling; coefficients of the 31 path diagrams modeling the correlations vary substantially. To simplify interpretation, we grouped the diagrams into five clusters, for which we describe the differential effects of the three putative causes on incidence and mortality. When scaled, the path coefficients intergrade without marked gaps between clusters. Ethnic differences make for differences in cancer rates, even when the populations tested are ancient and complex mixtures. Path analysis usefully decomposes a structural model involving effects and putative causes, and estimates the magnitude of the model's components. Smooth intergradation of the path coefficients suggests the putative causes are the results of multiple forces. Despite this continuity of the path diagrams of the 31 cancers, clustering offers a useful segmentation of the continuum. Etiological and other extrinsic information on the cancers map significantly into the five clusters, demonstrating their epidemiological relevance. Am. J. Hum. Biol. 16:1–16, 2004. © 2003 Wiley-Liss, Inc.

The analysis of spatially varying population data presents researchers with a number of problems. Since neighboring localities frequently affect each other, the resulting data are spatially autocorrelated, i.e., not independent, as conventional statistical tests require. Positive spatial autocorrelation yields liberal test results, i.e., the null hypothesis of no difference among population samples will be rejected more frequently than indicated by the nominal significance level. This means that more observed differences among localities will be considered significant than are expected by chance sampling. The consequences of disregarding spatial autocorrelation during the analysis of biological data are discussed in another article in this volume (Sokal, 2004). Some methods for overcoming this problem are discussed in Sokal et al. (1993). When the analysis involves more than one variable, it may be extended to a study of the correlation structure of the variables over the set of correlated localities. The statistical significance of such correlations is also affected by the spatial autocorrelation of

Contract grant sponsor: National Science Foundation; Contract grant number: SBR 9419349 (to R.R.S.).

*Correspondence to: Dr. Robert R. Sokal, Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245. E-mail: sokal@life.bio.sunysb.edu

Received 2 June 2003; Accepted 8 August 2003

Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/ajhb.10231

the constituent variables and can yield excessively liberal test outcomes (Clifford et al., 1989). When, as is common in human genetic or epidemiological studies, different variables are extracted from diverse databases resulting from different studies, researchers are faced with the problem of missing values. This occurs because different studies may not correspond with respect to the set of variables analyzed or they may differ in the set of localities at which samples were taken. Frequently they will differ with respect to both criteria. Such problems further complicate any analysis of their covariation. Neglect of such problems has led to questionable interpretations of spatial genetic datasets (Sokal et al., 1999a,b).

In some studies several putative explanatory variables are measured at each locality and coefficients of correlation between these variables are calculated over the set of localities. Path analysis (Li, 1975; Sokal and Rohlf, 1995) can be used to help interpret intercorrelations between the observed variables and the putative causes and to determine the magnitude of the effects of the several possible explanatory variables. In the study presented below, we examine the effects of three potential explanatory variables (ethnohistory, genetics, and geography) on two kinds of cancer rates (incidences and mortalities). However, these rates are estimated for 31 different cancers, resulting in 31 separate datasets of five variables sampled for a varying number of localities. In such a situation, two contrasting approaches can be employed. In a top-down approach, the researcher computes average correlations between the explanatory variables and variables of interest based on all loci studied or on rates for all the recorded diseases. These average correlations can then be analyzed by means of a single, overall path diagram. The bottom-up approach investigates the diversity of the variation of the biological variables by studying the differences among path diagrams for each member of the set of kindred variables (each type of cancer). If there are many of these variables, the examination and comparison of their path diagrams with all others become tedious and path diagrams are clustered or ordinated to group them with other path diagrams showing similar patterns.

Although none of these methods is new when considered singly, their application in combination as illustrated in this article is novel and could be extended to numerous other examples in human biology. We encountered these problems during an analysis of data from human cancer rates and we believe that the methodology of their solution should be of interest not only to cancer epidemiologists but also to other researchers working with quite different variables.

The two response variables (cancer rates) and three putative explanatory variables were taken from available databases with estimates of each variable for each European locality. Coefficients of correlation between these variables were calculated separately for the 31 cancers over the European reporting stations, as well as over all cancers by averaging the correlations. All resulting correlation matrices of the five variables were analyzed by path analysis. To simplify interpretation, the 31 separate sets of path coefficients were summarized by k-means clustering and nonmetric multidimensional scaling. We interpret the resulting structure in terms of the differential effects of the three putative causes on incidence and mortality. To demonstrate their epidemiological relevance, we mapped etiological and other extrinsic information onto the resulting five clusters. References to the sources for these data and details on the methodology are furnished below.

MATERIALS AND METHODS

Data

During the 1980s and 1990s, the International Agency for Research on Cancer (IARC) compiled data on incidences and mortalities for numerous cancers (Waterhouse et al., 1982; Parkin et al., 1983, 1992; Muir et al., 1987; Smans et al., 1992; Zatonski et al., 1996). Cancer incidences are the numbers of new cases per year of a specified cancer per 100,000 persons in a population from a defined area, adjusted for the age-structure of the population. Cancer mortalities are the numbers of deaths per year due to a specific cancer per 100,000 persons in a population for a given area. We correlated interlocality differences in the two rates (symbolized as INCID and MORT, respectively) with ethnohistoric distances (ETH), genetic distances (GEN), and geographic distances (GEO) for the areas studied. All five variables in this study (INCID, MORT, ETH, GEN, and GEO) were computed as interlocality distances for the indicated variable. This was done for three reasons: 1) Relevant theory in population genetics (Nei, 1987), genetic epidemiology (Morton, 1982), and anthropology (Smouse and Long, 1992) is frequently formulated in terms of distances. 2) Cancer mortalities and incidences are spatially autocorrelated (Rosenberg et al., 1999; Upton and Fingleton, 1985). Significance tests for such data can be conveniently adjusted (by Mantel tests and their extensions, see below) when they are expressed as distances. 3) Ethnohistoric and genetic distances permit considerable data compression over the original values.

The cancer incidence distances (INCID) are based on absolute differences in incidences between pairs of areas. These rates, age-adjusted to the world standard (Higginson et al., 1992), come from four volumes (Waterhouse et al., 1982; Parkin et al., 1983, 1992; Muir et al., 1987) that report European incidence rates for four periods between 1968 and 1988. The number of organ-specific sites varies for different reporting stations, the maximal number being 45 for males and 47 for females. The maximal number of European localities per cancer site is 75, the maxima per country being: the former Czechoslovakia, 2; Denmark, 1; England, 8; Finland, 1; France, 6; Germany, 3; Hungary, 3; Iceland, 1; Ireland, 1; Italy, 9; Netherlands, 2; Norway, 1; Poland, 7; Portugal, 2; Romania, 1; Scotland, 5; Spain, 7; Sweden, 1; Switzerland, 5; former USSR, 8; and former Yugoslavia, 1. We investigated by means of nonparametric ordering tests whether there are time trends in the incidences. Finding none, we calculated average incidences over time based on from 1 to 4 of these rates to obtain the final incidence rates used to compute the distances INCID.

Cancer mortality distances (MORT) were computed from absolute differences in mortalities between pairs of areas. Such rates for Europe are available (in Smans et al., 1992) for 40 organ- and sex-specific cancers at 355 registration areas in the quondam European Economic Community (EEC) and in Zatonsky et al. (1996) for 36 cancers at 153 areas (but only 32 cancers at 194 areas) in Central Europe (CE). The mortality rates in these sources are stated as age-standardized deaths per 100,000 population size per annum (Higginson et al., 1992). At the time of reporting the mortalities (1970s), the EEC comprised Belgium, Denmark, Eire, France, Italy, Luxembourg, the Netherlands, the United Kingdom, and West Germany. The CE data are for 1983 to 1987 and include Austria, Bulgaria (1986/7), the former Czechoslovakia, East and West Germany (the West German data are for a later time span than that of their EEC counterparts), Hungary, Poland, Romania, and the former Yugoslavia. The two regions, EEC and CE, overlap in West Germany and share 34 organ- and sex-specific rates.

Since the incidences and mortalities were recorded from different sampling stations, finding matching localities became a problem. We permitted matches between localities up to 100 km apart (a lower bound for patch sizes for cancer mortalities in the EEC having been determined to be 342 km by Rosenberg et al., 1999). The matched localities (which ranged from 20 to 41 pairs) differed from cancer to cancer.

We omitted cancer sites with fewer than 20 localities for any one cancer rate or fewer than 20 matched localities in the combined database because correlations based on fewer than 20 paired observations would have been unreliable. This is especially true in view of the fact that all variables in the study are known to be spatially autocorrelated, hence are supported by fewer degrees of freedom than n-2 (Clifford et al., 1989). In order for a specific cancer to be included in our study, we had to be able to match 20 or more points for mortality, incidence, and genetics. Of these, the potentially limiting variables were incidence and genetics, there being an overabundance of mortality readings. Ethnohistory and geographic coordinates were never limiting since they were available for any locality in Europe. Regrettably, this screen eliminated some common cancers, such as colon/rectal cancer. We were left with the 31 organ- and sex-specific cancers, listed in Table 1 (15 for males, 16 for females).

The ethnohistorical distances (ETH) are computed from an ethnohistorical database (Sokal et al., 1996) for Europe from 2200 BC to 1970 AD, assembled in our laboratory and consisting of 1,750 “active” records describing nine types of ethnic movements and 1,710 “passive” records describing locations, assimilations, and wanderings within the same area. Each of the 3,460 records lists the name of a population unit (e.g., tribe, people) and their language family, when known; it reports the dates and defines the areas of movement and location. The

TABLE 1. Path coefficients[a] obtained for the 31 cancers and their cluster averages

Groups	Cancer sites	Cluster number	ETH→ MORT	INCID→ MORT	GEN→ MORT	GEO→ MORT	ETH→ INCID	GEN→ INCID	GEO→ INCID
Males	Bladder	IV	0.1223	0.1844	0.0521	0.0166	−0.0780	0.0841	0.1081
	Brain	IV	0.1005	0.2559	0.0850	−0.0321	0.0632	−0.0220	0.1627
	Gall Bladder	V	−0.0507	0.3645	−0.0119	0.0351	−0.1219	−0.0372	0.1185
	Hodgkins	IV	0.0783	0.4040	0.0763	0.0561	0.1175	0.0612	−0.0447
	Larynx	V	−0.1095	0.5231	0.0005	0.0635	0.1417	0.0821	0.0902
	Lung	V	−0.0321	0.4972	−0.0110	0.0823	0.1279	0.0155	0.0641
	Lymphoma	I	0.2044	−0.0147	−0.0586	0.1428	−0.1997	0.0038	0.1456
	Melanoma	IV	0.1627	−0.0088	0.0455	0.0513	0.0430	−0.0278	−0.1165
	Myeloma	I	0.0619	−0.0435	−0.0319	0.2390	−0.1211	0.0605	0.2288
	Oesophagus	V	0.0154	0.4444	−0.0449	−0.0616	−0.2067	0.1335	0.1960
	Pancreas	II	0.0180	0.1332	0.0461	0.3985	0.1641	−0.0448	0.1962
	Prostate	II	0.1099	0.3409	0.0446	0.2064	0.0319	0.0088	0.1155
	Stomach	III	0.2049	1.0001	−0.0089	−0.4395	−0.1987	0.0168	0.4742
	Testis	IV	0.1965	0.1098	0.1022	0.0442	0.0279	−0.0037	0.0406
	Thyroid	IV	0.0601	0.2984	0.0688	0.0326	−0.0338	0.0504	−0.0860
Females	Bladder	I	0.1629	0.2082	−0.0372	0.1751	−0.1780	0.0152	0.2786
	Brain	IV	0.1655	0.0780	0.0050	0.0318	0.1779	−0.0039	0.1703
	Breast	I	0.0145	−0.0746	−0.0731	0.4897	−0.1808	0.0211	0.3212
	Cervix	II	0.2265	0.2003	−0.0212	0.3532	−0.0175	−0.0288	0.0663
	Gall Bladder	V	0.0593	0.5692	0.0065	−0.1011	−0.2190	−0.0224	0.2361
	Hodgkins	IV	0.1834	0.1075	0.0761	0.0213	0.0827	0.0468	−0.0896
	Larynx	IV	0.0274	−0.0122	0.0478	0.0515	−0.0116	0.0393	0.0565
	Lung	V	0.1131	0.7399	−0.0358	0.1140	−0.0716	0.0148	0.2609
	Lymphoma	I	0.1554	−0.1779	−0.0439	0.2027	−0.1567	0.0494	0.2042
	Melanoma	IV	0.0877	0.1894	0.0139	0.1758	0.0553	−0.0407	0.0325
	Myeloma	I	0.0580	−0.0300	−0.0065	0.1831	−0.1083	−0.0231	0.2001
	Oesophagus	III	0.2039	0.8891	0.0394	0.0337	−0.1902	0.0058	0.3972
	Ovary	II	0.0897	0.2120	−0.0254	0.4711	0.2537	−0.0075	0.0590
	Pancreas	II	−0.0273	0.3096	0.0048	0.4312	0.1500	−0.0123	0.1330
	Stomach	III	0.1881	0.8133	0.0039	−0.3194	−0.2589	0.0289	0.4351
	Thyroid	IV	−0.0012	0.1974	0.0343	0.0240	0.1515	0.0367	0.1776
AVERAGE[b]			0.0919	0.2809	0.0110	0.1024	−0.0247	0.0161	0.1494
MEDIAN			0.0897	0.2082	0.0048	0.0561	−0.0175	0.0148	0.1456
Clusters[c]	I		0.1095	−0.0221	−0.0419	0.2387	−0.1574	0.0211	0.2297
	II		0.0834	0.2392	0.0098	0.3721	0.1164	−0.0169	0.1140
	III		0.1990	0.9008	0.0115	−0.2417	−0.2159	0.0172	0.4355
	IV		0.1076	0.1640	0.0552	0.0430	0.0542	0.0200	0.0374
	V		−0.0007	0.5230	−0.0161	0.0220	−0.0583	0.0310	0.1610

[a] We omit showing path coefficient ETH→GEN, assumed constant (0.2381) for all 31 cancers.

[b] The AVERAGE path coefficient values differ inconsequentially from the OVERALL values reported in figure 2 of Sokal et al. (2000), except for INCID→MORT, which should have been 0.2788, but was reported as 0.4834, owing to a recently discovered computational error. The corrected figure corresponds well with the average path coefficient for INCID→MORT, which is 0.2809 in the above table. The principal conclusions in Sokal et al. (2000) are not affected by the change.

[c] The cluster values are mean coefficients computed for the cluster members from the upper portion of the table.

ethnohistorical database can be found on the World Wide Web at http://life.bio.sunysb.edu/ee/msr/ethno.html. The program ETHNO (by N.L. Oden; available at the same Web address) estimates the admixture of populations from specific language families following an updating algorithm that uses optimal weights for each type of movement (Sokal et al., 1996). At the completion of the program there are vectors of estimated proportions of contribution by 17 language families and two unknown groups to the population mix at each of 2,216 land-based 1° × 1° quadrats in Europe. Most quadrats receive input from numerous other quadrats (26 on average). From these vectors we computed arc distances (Cavalli-Sforza and Edwards, 1967) between all pairs of quadrats. Sensitivity experiments showed that ethnohistoric-genetic correlations were robust against reasonable perturbations in time of movement, location, ethnic (language-family) designation, and completeness of the database (Sokal et al., 1996). To assemble ethnohistorical distance matrices we chose the set of quadrats that matched the locations for genetic, mortality, and incidence data.

The genetic distances (GEN) were computed from our genetic database of 3,481 samples for Europe (Sokal et al., 1989), which comprises 26 genetic systems (blood group antigens, proteins, enzymes, HLA, and immunoglobulins) with 93 allele or haplotype frequencies. With some exceptions, each system corresponds to a genetic locus. The distances were computed separately for each genetic system. The smallest number of genetic systems for any one cancer was 15 for males and 14 for females. For each cancer rate locality, a computer program found the closest genetic sampling point to form a matching pair of gene-frequency and mortality or incidence values. If the closest genetic point was more than 100 km from the cancer rate locality, the point was omitted from the study. This cutoff point was chosen as a conservative estimate based on earlier work by Sokal et al. (1989) demonstrating that the patch size for these gene frequencies ranges from 900–1500 km. We computed Prevosti distances (Prevosti et al., 1975) between gene-frequency samples and assembled them into matrices of genetic distances (GEN) of the same size as the matching mortality and incidence matrices. Prevosti distances between localities are average absolute differences in gene frequencies between these localities, averaged over all available genetic systems. As noted above, the minimal matrix size (number of locality samples) for which we kept results was 20. The correlations for the separate genetic systems were then averaged to yield the correlation coefficients used to compute the path coefficients in Table 1 explained below.

Geographic distances (GEO) were calculated as great-circle distances (in km) between all pairs of locations that matched those of the other four variables.

Analysis

We treated the data as point estimates throughout. Rates furnished for an entire country were treated as though they had been collected at the capital city. The initial computation was of all 10 pairwise correlations of the five distance variables of our study (INCID, MORT, ETH, GEN, and GEO) for each of the 31 organ- and sex-specific cancers. Earlier work with these data (Sokal et al., 1997, 2000; Rosenberg et al., 1999) demonstrated that the first four variables were spatially autocorrelated. For this reason, the distances were correlated and tested for significance by means of Mantel tests (Mantel, 1967; Sokal and Rohlf, 1995), with the matrix elements scaled to yield a correlation coefficient as the Mantel product. The number of pairs of observations being correlated, n, varied with the cancer and with the number of matches that could be obtained between gene frequency localities and cancer rates.

Because human gene frequency data are largely unbalanced, the GEN distances and their correlations with the other variables were computed separately for each genetic system. These correlations were averaged over the systems and their overall significance was computed by Fisher's method for combining probabilities (Sokal and Rohlf, 1995). Correlations r(ETH,GEN) and r(GEN,GEO) were held constant for all cancers at 0.2381 and 0.2297, respectively, based on the far more extensive European datasets in table 2 of Sokal et al. (1996) and table 1 of Sokal et al. (1997). The remaining eight correlations (of the 10 between all pairs of the five variables in this study) were independently computed for the 31 organ- and sex-specific cancers.

To represent the interrelations of the five variables, we turned to path analysis (Li, 1975; Sokal and Rohlf, 1995), a method for studying the direct and indirect effects of one set of variables—the causes (predictor variables)—on another set—the effects (criterion variables). Cause-and-effect relationships are depicted by single-headed arrows, correlated causes by double-headed arrows. The strength of single-headed arrows is estimated by *path coefficients,* which some readers may know as standard partial regression (beta) coefficients. Path coefficients are scale-independent measures of the effects of the predictor on the criterion variables, with the effects of other predictors held constant.

In this study we use path analysis to examine the relative strengths of the separate effects of the variables upon each other when constrained by our structural model, the path diagram in Figure 1. In this diagram, the putative causes ETH and GEO, as well as GEN and GEO, are connected by double-headed arrows to indicate correlations by remote factors that we cannot investigate further with the present data. The correlation r(ETH,GEN) is shown as a single-headed arrow ETH→GEN (a path coefficient), because ethnohistoric similarity will lead to genetic similarity, whereas the converse is unlikely. Ethnohistoric distances may affect

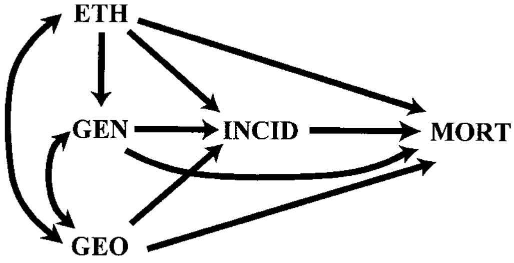

Fig. 1. Path diagram for pairwise correlations between distances or differences of the indicated variables. ETH, ethnohistory; GEN, genetics; GEO, geography; INCID, incidence; MORT, mortality. Double-headed arrows indicate correlations; single-headed arrows are path coefficients.

cancer rates not only through genetic distances (via the path ETH→GEN), but also directly expressed as cultural differences, as shown by paths ETH→MORT and ETH→INCID. Even though some of the ethnic admixtures in our model are quite ancient, it is possible that some cultural traits that affect cancer incidences will persist in the modern admixed populations. The cultural component of ETH that directly affects MORT may represent cultural factors affecting 1) the treatment and care of cancer patients, 2) the correctness of the diagnosis in the death certificate, and 3) varying death registration practices in different regions or countries. Similarly, we assume separate direct paths from GEN and GEO to INCID and MORT. The single-headed arrow INCID→MORT represents the direct effect of incidence on mortality, since morbidity precedes mortality. A reverse arrow is implausible.

The eight correlations among our five variables not indicated by double-headed arrows can be expressed in terms of the path coefficients and the two remote correlations (double-headed arrows) of Figure 1. These eight equations yield the magnitudes of the eight unknown path coefficients by standard methods for solving simultaneous equations. The resulting path coefficients for each organ- and sex-specific cancer are furnished in Table 1. We omit showing path coefficient ETH→GEN, assumed constant (0.2381) for all 31 cancers.

It is not customary to furnish statistical significance of path coefficients, the relative magnitudes of the coefficients being the main feature of interest. However, we can approximate the significance by examining the significance of the partial correlations between MORT and the three putative causes ETH, GEN, and GEO (in table 2 of Sokal et al., 1997) and of the partial correlations of INCID with ETH and GEN (in table 2 of Sokal et al., 2000). Because the individual elements of distance matrices violate the independence assumptions of ordinary significance tests for correlations, we employed the Mantel test (Mantel, 1967; Sokal and Rohlf, 1995), which evaluates the significance by a permutational approach. For partial correlations of distance matrices we used a multiple matrix extension of the Mantel test (Smouse et al., 1986). Combined probabilities (Sokal and Rohlf, 1995) over all cancer sites indicate significance at $P \leq 0.000{,}005$ for ETH→MORT, GEN→MORT, GEO→MORT, and for GEN→INCID. The significant spatial autocorrelation of cancer mortalities in Europe (Rosenberg et al., 1998) suggests the significance of the GEO→INCID path over all cancer sites. Only ETH→INCID is not significant.

We shall refer to each row of seven path coefficients in Table 1 as a *path coefficient profile* or *cancer profile.* To simplify the interpretation of the 31 profiles, we clustered them by the k-means method (MacQueen, 1967; for a thorough recent description, see Legendre and Legendre, 1998, p 349–355). This algorithm randomly assigns each profile to one of k clusters, then shuffles the profiles between clusters and, by a stepwise procedure, tries to minimize the sum of squared distances from each cluster member to the cluster centroid. Seven hundred separate attempts were made to optimize the k clusters for $k = 2$ to $k = 8$. We chose a value of $k = 5$, since the sum of squares within clusters for $k > 5$ did not decrease appreciably. We confirmed the membership of the five clusters (listed in column 3 of Table 1) by 5,000 additional random partitions into $k = 5$ clusters, which yielded identical cluster sizes and compositions each time. The numbers of cancers in clusters I to V are 6, 5, 3, 11, and 6.

For another perspective on this cluster arrangement, we employed ordinations of the 31 cancer profiles. Such techniques project the 6-dimensional profiles into low-dimensional spaces for ease of inspection and representation. We carried out principal components analysis (PCA; Krzanowski, 1988) and nonmetric multidimensional scaling (MDSCAL; Krzanowski, 1988) of the

31 × 31-covariance and distance matrices of the cancer profiles, respectively. The cancer profiles were projected into two- and three-dimensional spaces; their distributions were examined and compared to their membership in the *k*-means clusters. To check the adequacy of the ordinations, *minimum spanning trees* (Krzanowski, 1988) for the full-dimensional distance matrices were superimposed on the projected points.

RESULTS

For convenience, we group the path and correlation coefficients by magnitude into four classes. This classification is clearly arbitrary, but it can be justified by our extensive experience with these and similar data. The numerical thresholds defining the classes (given below) are low by conventional criteria. This is characteristic of correlations between distance matrices, which are usually far lower than those of the variables on which they are based (Sokal et al., 2000; Dutilleul et al., 2000). For this reason we call effects $> |0.45|$ *strong*, those ranging from $> |0.15|$ to $|0.45|$ *moderate*, those ranging from $> |0.03|$ to $|0.15|$ *weak*, and those falling between ± 0.03 *negligible*. In Figure 3 we represent these four classes by bold, thin solid, dashed, or dotted arrows, respectively.

The path coefficients for each cancer are shown in Table 1. In the row labeled AVERAGE, we also feature their average values over all cancer sites. These average values are quite close to the values shown in figure 2 of Sokal et al. (2000): mean absolute deviation between the two sets of coefficients equals 0.0172, after the correction given in footnote b in Table 1 is made. This similarity in results obtains despite the two sets of values resulting from different algorithmic procedures. The current results are the means of the 31 path coefficients, whereas the earlier results represent the unique path coefficient solution of the 10 average correlations.

We note considerable differences in magnitude among the seven average path coefficients shown in Table 1, ranging from –0.025 for ETH→INCID to 0.281 for INCID→MORT. These average path coefficients give us an estimate, for both sexes and over all cancer sites, of the relative strengths of the seven paths. However, these average values can be misleading because of the considerable variation among the coefficients for any one vector (with the exception of ETH→GEN, assumed constant and not shown in Table 1). Therefore, we examined the data further by applying two statistical tests to each vector. First, we tested each mean for a significant positive or negative deviation from zero by means of a *t*-test of the null hypothesis that the mean path coefficient of the vector equals zero. Next, we tested the equality of the frequencies of positive and negative path coefficients. An indication of this inequality is given by the median in the row so labeled in Table 1. The tests concurred in labeling the means of four of the vectors (ETH→MORT, INCID→MORT, GEO→MORT, and GEO→INCID) significantly above zero (at $P < 0.01$) and with more positive than negative path coefficients. Two vectors (GEN→MORT and ETH→ INCID) are clearly not significant by either test, and GEN→INCID yields an ambiguous result. While its mean is significantly above zero ($P < 0.05$) and there are 19 positive to 12 negative coefficients, this inequality is not significant. These tests establish that some of the path coefficients are not zero.

We can summarize our findings as follows. Ethnohistoric distances directly affect genetic distances moderately and mortality differences weakly, the direct effect on incidence differences being negligible. Thus, we have no evidence for cultural carcinogenic effects, but some for cultural influences on mortalities (Sokal et al., 1997, 2000; Berrino et al., 1995, 1999). Different political entities may practice different mortality registration procedures; may exhibit differential biases in certifying deaths due to specific cancers; may code identical death certificate information variously in different national vital statistics offices; or may vary in their practices of handling imprecise or illegible death certificates. Cultural differences also include national and regional variations in level of medical care (screening, health behaviors, and health advice) and the reliability of census estimates. The common factors underlying *r*(ETH, GEO)—which can be summarized as: geographic proximity is reflected by similar ethnohistories—act directly on mortality and indirectly via incidences. However mediated, ethnohistoric affinities contribute to differences in cancer mortalities. The direct effects of GEN on MORT and on INCID are both negligible. Although the effects of GEN are slight, we should stress that they are highly significant on both INCID (table 2 of Sokal et al., 2000) and MORT (table 2 of Sokal et al., 1997) for various cancers and overall as well. In each table the overall effect of GEN on the cancer rate analyzed is significant at $P < 10^{-5}$. The common factors producing *r*(GEN,GEO) suggest that geographic proximity is reflected by similar genetics, but the common factors act mostly via GEO. Geographic distances influence both incidences and mortalities weakly, both directly (possibly reflecting environmental similarities) and indirectly through their common factors (discussed above) with ethnohistoric and genetic distances.

The PCA accounted for 78.5% and 91.6% of the covariance in two and three dimensions, respectively. The corresponding MDSCAL final stress values were 0.189 and 0.092, accomplished in both cases after five iterations. Ordinations by either method are similar and useful representations of the multi-dimensional spatial structure of the 31 cancer profiles. The superimposed minimum spanning trees confirmed the adequacy of the ordinations, even in two dimensions. Since the minimum spanning tree connects nearest neighbors in the full-dimensional space, two- or three-dimensional ordination may introduce distortions in the true relationships of the projected objects. Such distortions are indicated when points that are far apart in the full-dimensional space appear close to each other in the reduced dimensional space (Krzanowski, 1988). In Figure 2 we illustrate the two-dimensional MDSCAL ordination, because such plots generally preserve both near and far distances better than a PCA of the same dimensionality (Rohlf, 1972). An example of a distortion in Figure 2 is the distance between female brain cancer and testicular cancer, which appears close in the two-dimensional ordination, but in actuality is much greater, as shown by the circuitous route by which the minimum spanning tree connects the two. However, when the entire assemblage of points (cancer sites) is considered, the two-dimensional ordination seems to represent the full-dimensional space quite well.

In Figure 2 we have indicated membership in the five *k*-means clusters. Three of the clusters, I, III, and IV, are well supported by the ordination. They are internally connected, i.e., one can travel along the minimum spanning tree from any member of one cluster to any other member of the same cluster without traversing any cancers in the other clusters. The two exceptions include cluster II, in which one member, cervical cancer, attaches to female melanoma in cluster IV rather than to other members in its cluster. Cluster V is the other exception, with two of its members, male lung and male larynx cancers, separated from other members of cluster V by male Hodgkins and male thyroid cancer of cluster IV.

Figure 2 reveals that, whereas clusters I, II, and III are located near the boundaries of the space defined by the axes of the graph, clusters IV and V occupy intermediate

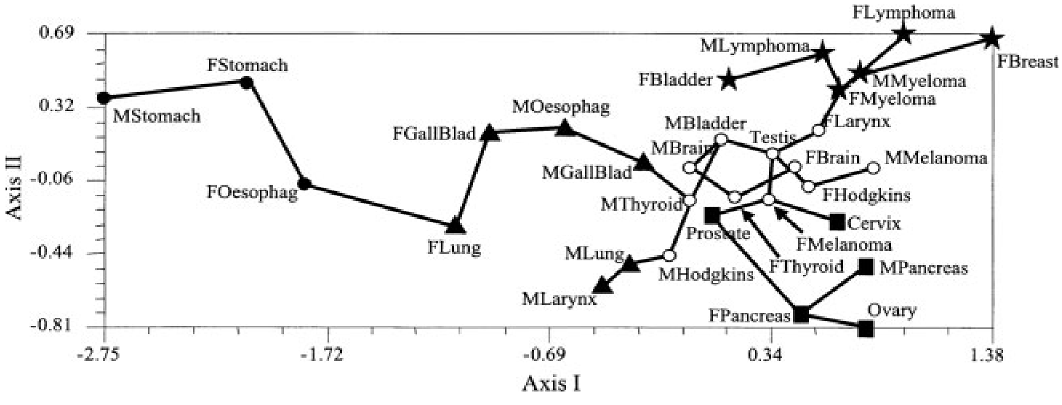

Fig. 2. Ordination in two dimensions by nonmetric multidimensional scaling of the 31 cancer path coefficient profiles. The lines connecting the cancers are edges of a minimum spanning tree in the full dimensional space. The names of cancers occurring in both sexes are preceded by F for female and M for male cancers. The cluster membership of each cancer is indicated by the following symbols: stars for cluster I, squares for cluster II, solid circles for cluster III, open circles for cluster IV, and triangles for cluster V.

regions of the space and form a gradual transition between the extremes. Thus, the cluster results, even though stable for 5,000 random partitions, do not necessarily describe discrete clusters with wide gaps between them. When the 30 edges of the minimum spanning tree are ordered from shortest to longest, the six edges in Figure 2 that are transitions from a member of one cluster to a member of another cluster are in positions 7, 14, 20, 22, 25, and 26. (There are six such edges for five clusters, rather than the four that might be expected, because of the above-mentioned two exceptions to internal connectedness within clusters.) Although there is the expected tendency toward long edges between clusters, the sum of the six ranks is not significantly higher than expected by chance sampling ($P = 0.148$ by an exact Wilcoxon rank sums test). This implies that the k-means clusters are not as well separated from each other as the robustness of their minimal sums of squares solution might suggest. As a corollary of this finding, we note that among the 10 longest edges in the minimum spanning tree, seven are within rather than between clusters. The cancer profiles are packed into the clusters at different densities and show differing dispersions. Clusters V and IV differ the most with respect to dispersion. The variance of the edge lengths of the former is $\sim$51 times that of the latter. Also differing are the diameters of the subgraphs of the clusters (the maximal distance along the edges between any pair of profiles—a measure of the "volume" of the hyperspace that the cluster occupies). Cluster V is the most strung out, cluster III the least. This is not immediately apparent from Figure 2, but it will be recalled that this figure is in the two major dimensions only, masking the extent of the spread of profiles into the other five dimensions. It appears that the effects of the three putative causes ETH, GEN, and GEO, as well as INCID→MORT, range widely and intergrade among the cancer profiles. Actually, Figure 2 is somewhat misleading, because it is plotted in two dimensions only. Clusters I, II, and III are not the mutually most distant. When distances between clusters are evaluated, the three most distant pairs of clusters in the full-dimensional space ordered by length are I–III, II–III, and III–IV.

In Figure 3 we show path diagrams for all five clusters based on their mean profiles furnished in the last five rows of Table 1. The arrows are coded to indicate the magnitudes of the correlations or path coefficients, employing the conventions described earlier. Even a casual inspection reveals substantial differences among the five diagrams. Cluster I (lymphoma and myeloma in both sexes, as well as female bladder and breast cancers) is characterized by moderate effects of GEO on INCID and MORT, and negligible effects of INCID on MORT. Cluster II (male and female pancreas plus cervix, ovary, and prostate cancers) exhibits the highest mean path coefficient GEO→MORT and a moderate mean path INCID→MORT. The strongest average path INCID→MORT (0.9008), a moderate effect ETH→MORT, and moderate negative effects GEO→MORT and ETH→INCID mark cluster III (male and female stomach and female oesophagus). In cluster IV (brain, Hodgkins, melanoma, and thyroid in both sexes, as well as male bladder, female larynx, and testis), we find a moderate INCID→MORT effect, with all other effects positive but weak or negligible. Cluster V (male and female gall bladder and lung plus male larynx and oesophagus) shows a strong path INCID→MORT, a moderate path GEO→INCID, and negligible or negative effects for the other paths.

DISCUSSION

We commence our discussion of these findings with 1) an interpretation of the three putative causes in this study—ETH, GEN, and GEO. This is followed by 2) a brief account of previous work with these data, 3) a detailed discussion of the epidemiological findings of the present study, and 4) the data-analytic techniques of this study with the implications of their results for the model of the intercorrelation of the variables. We conclude by 5) testing the validity of the data structure in hyperspace against extrinsic information from the epidemiological literature.

Interpretation of putative causal variables

ETH measures the differences in admixture proportions of the language-family affiliations of the populations that have given rise to the modern inhabitants of the areas concerned (Sokal et al., 1996, 1997, 2000). Although language-family affiliation of a given ethnic group is most unlikely to

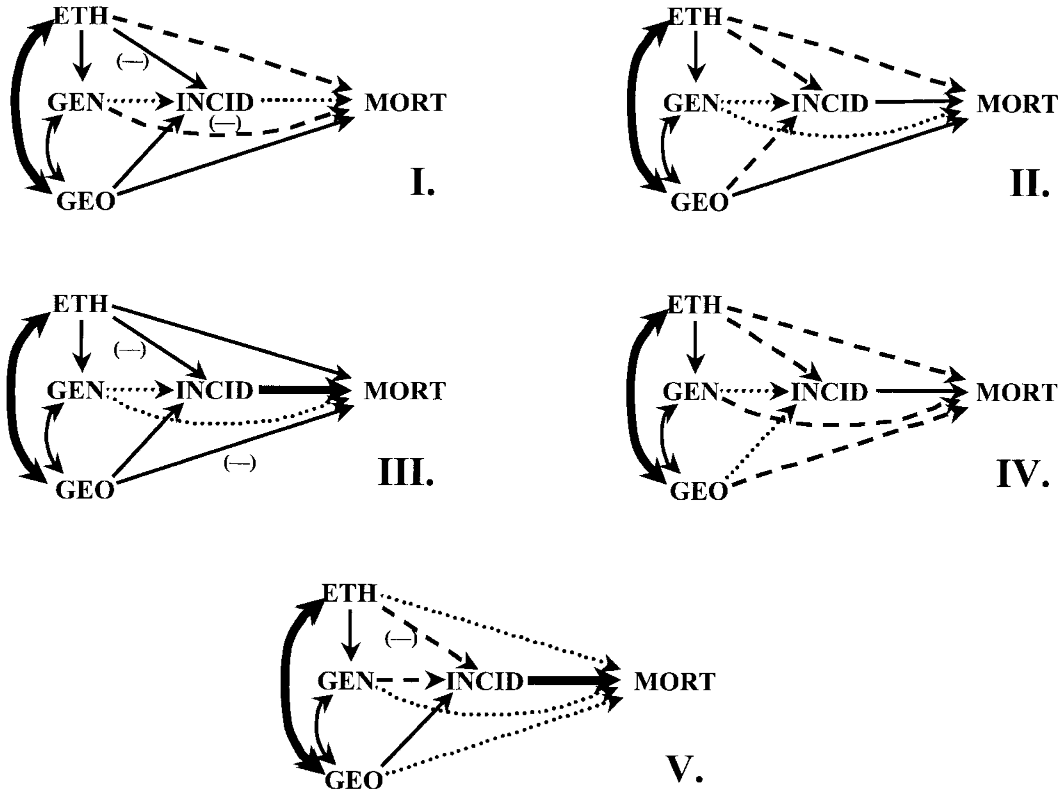

Fig. 3. Path diagrams based on average path coefficients of the five clusters of cancer path coefficient profiles. Cluster numbers are indicated next to the diagrams. For abbreviations of variables and meaning of arrows, see Figure 1 legend. The arrows are coded to indicate the magnitudes of the correlations or path coefficients. Strong effects > |0.45| are shown as bold arrows, moderate effects ranging from > |0.15| to |0.45| as thin solid arrows, weak effects ranging from > |0.03| to |0.15| as dashed arrows, and negligible effects falling between ± 0.03 as dotted arrows. The coding of the arrows is based on the numerical values of the average path coefficients in the last five rows of Table 1. Negative signs in parentheses (–) indicate negative path coefficients other than those of negligible magnitude. Numerical values for the correlation *r* (ETH,GEO) could be separately computed for each cancer. Their cluster means range from 0.4682 for cluster I to 0.4756 for cluster II.

have any direct effect on cancer rates, speakers of any one language family in most cases have a shared demographic history, leading to genetic and cultural similarities. Thus, ETH distances are a combination of genetic and cultural components, some of which are bound to affect cancer rates.

Our GEN estimates are conventional measures of genetic differentiation among populations, such as have been used for both classical loci and molecular genetic information. In this study the genetic distances are based on classical loci, since these are the only extensively sampled sources of genetic variation available for most countries of Europe. In theory, such distances express the genetic differentiation of a random sample of genes. In fact, these human genes are not chosen truly randomly. Numerous investigators over the years have sampled them for diverse reasons only rarely related to cancer. Although some of these loci have been associated with various cancers as early as the 1970s (Mourant et al., 1978), none of the loci in our database have alleles known to have major effects on the occurrence or severity of any cancer in this study. If cancer—like height—is driven by many small roughly additive contributions from different loci, we might expect GEN to be an important actor in our data. But if cancer is driven chiefly by a few important loci, it should not surprise us to find only weak effects of genetic distances on incidence and mortality.

There probably are differences among ethnic groups with respect to frequencies of alleles that strongly affect cancer rates but are not in our database. Such differences would be confounded with the cultural effects of ETH and would not be expressed through GEN, since the latter refers only to distances computed from the assemblage of loci in our database. Factors common to geographic and genetic distances represent the spatial autocorrelation of the gene pool engendered by isolation by distance as well as by the cohesiveness of ethnic units during migration and settlement. Isolation by distance (Wright, 1943, 1969) occurs when individuals are limited in their dispersal and tend to mate with partners within their range of dispersion. This will result in genetically similar population samples at nearby locations. This process gives rise to correlation between GEO and GEN, but it dies off exponentially with distance, gene flow between remote locations being limited.

Distances GEO correlate with both ETH and GEN. Geographically closer samples are more likely to have a similar ethnohistory and similar genetic structure than samples that are farther apart. For GEN the reason has already been discussed. For ETH it is the cohesion of ethnic units. Besides the covariates ETH and GEN, the remaining component of geographic distance (the direct paths from GEO to INCID and MORT) will include environmental differences between the areas being compared. In view of the important environmental determinants of carcinogenesis, we would expect the path coefficients representing these effects to be substantial for some cancers, as indeed they are (see Table 1).

Previous findings

We have mentioned that in earlier work with these data (Sokal et al., 1997, 2000; Rosenberg et al., 1999) we demonstrated significant spatial structure for both cancer rates in Europe. Average correlations of differences in cancer mortalities (MORT) with the three putative causes were found to be positive and ordered as follows by magnitude: GEO > ETH > GEN, whereas differences in cancer incidences (INCID) were correlated at a generally lower level, with GEO > GEN, but not at all with ETH. To estimate the relative strength of the various factors acting on the system, Sokal et al. (2000) also examined the interrelations of these five variables (INCID, MORT, ETH, GEN, and GEO) by means of path analysis of correlation coefficients averaged over all cancers. The analysis demonstrated the lack of cultural carcinogenic effects, but suggested cultural influences on mortalities through differences in recording procedures, such as the registration of deaths in different political entities. We found the relatively large correlation between ETH and MORT to be due to common factors behind the correlation of ETH and GEO. Geographic proximity results in similar ethnohistories. The direct effects of GEN were negligible and only their common effects with GEO played a role, accounting for the weak influence of GEN on INCID and MORT.

There are minor numerical inconsistencies between the results published earlier (Sokal et al., 1997, 2000) and those in this article because the present data are limited to 31 organ- and sex-specific cancers, fewer than in the earlier analyses. Furthermore, the number of matching localities for mortalities and incidences is frequently lower than the number of samples employed in the earlier separate analyses of the two variables.

We are unaware of work by others, in different disciplines, using the entire sequence of operations we have applied here. However, at least two studies (Nantel and Neumann, 1992; Leduc et al., 1992) carried out a part of the sequence by combining Mantel correlations with path analysis in ecological studies.

Epidemiological findings

Four of the mean path coefficients are unequivocally significant and positive. The most marked average path is INCID→ MORT. One might expect the magnitude of this effect to be even greater than observed, since there is a direct cause-and-effect relationship between incidence and mortality for a given cancer. However, the effect will differ among cancers because of varying treatment success. Additionally, we expect the effect to be modulated by geographical differences in treatment methods and quality of care. The effects of INCID on MORT in this study are possibly biased, but it is difficult to discern the direction of such a bias. In the same population, patients scoring positive for INCID in a time interval will be spread out to encounter

MORT among several subsequent intervals, leading to a weakening of the apparent link between INCID and MORT. In fact, some of our MORT records precede corresponding INCID records, weakening the link further. However, our variables are averages or summaries of individual behavior over time and space. Correlations between such averages may exceed the correlations between the individual observations themselves. Because both of these effects are operating, possibly in opposite directions, it is difficult to say whether INCID-MORT correlations are biased upward or downward.

We now turn to the other three significant paths. We have already seen that ETH→MORT can be explained as the effect of cultural differences on screening for the disease, health behaviors, health advice, and processing of death certificates. However, it may also reflect genetic differences between ethnic units that are not included among the loci that constitute our genetic distance GEN. The GEO→INCID path illustrates that geographic propinquity will result in similarity of environmental carcinogens, hence in similarity of incidence rates. The fourth path, GEO→MORT, is appreciable because both quality of care and death certification are spatially autocorrelated due to homogenizing influences on the health care system within political entities.

The remaining three mean path coefficients (GEN→INCID, GEN→MORT, and ETH→INCID) are negligible in magnitude. We have seen that only GEN→INCID differs significantly on the positive side from a mean vector of zero ($P < 0.05$). None of the three vectors has significantly more positive than negative paths. Thus, the direct effects of GEN on these cancer rates seem in doubt. It seems unlikely a priori for genetic differences to cause mortality differences, unless we are studying very specific genes that affect the outcome of a given course of treatment. We might have expected some differences in cancer incidences from genetic differences of populations, yet GEN→INCID is negligible as well. This supports our belief that the genetic distances we computed did not include major carcinogenic loci.

The mean path coefficient for ETH→INCID is negative, although it is not significantly different from zero. How are we to interpret the negative path coefficients in Table 1? There are 63 minus signs (29.0%) among the 217 path coefficients in this table. It is likely that most negative coefficients represent random variation around a negligible effect. The 12 most negative coefficients are all of moderate strength, i.e., between –0.15 and –0.45. Nine of the 12 occur in ETH→INCID and, in view of the lack of significance for the mean vector, may not warrant interpretation. However, the two most negative coefficients are for male and female stomach cancers in the GEO→MORT vector. These two cancers are in the outlier cluster III. It is possible to arrive at a structural explanation for negative paths a posteriori. However, we hesitate to ascribe epidemiological meaning to negative paths for the following reasons: 1) Our results are constrained by the model chosen for the path diagram. The model may not be correct or complete, forcing the negative contribution onto a path where it does not belong. The path-coefficient model we employed was based on linear functions estimated by ordinary least squares and these assumptions and procedures may not represent the true state of affairs. Since the solutions of the simultaneous equations leading to the path coefficients for each cancer are predetermined by the signs and magnitudes of correlation coefficients, errors in the estimates of the correlation coefficients can lead to errors in the signs and magnitudes of the path coefficients. 2) Although an incident case necessarily precedes a death, our incidence and mortality rates were sometimes sampled in reverse chronological order. 3) Measurement error may also play a role in our results.

Data-analytic methodology

Faced with 31 spatial datasets, one for each cancer, it is useful to carry out both approaches outlined in the Introduction—the top-down approach and the bottom-up approach. The former yields a summary of the interrelations of the five variables. The utility of such a summary will depend on the variance of the 31 values for each vector of path coefficients. In the cancer dataset this variance is substantial for four paths: ETH→MORT, INCID→MORT, GEO→MORT, and GEO→INCID. It is therefore only roughly indicative of the relations in the 31 datasets.

The bottom-up approach starts with separate path analyses of the 31 datasets.

Interesting differences can be seen, but when as many as 31 profiles have to be compared it is easy to lose sight of whatever structure there is among the profiles. One way to impose structure on such data is to cluster them. However, we must keep in mind that most clustering techniques will force the data into clusters regardless of whether or not the clusters are real (i.e., implying a common underlying generating mechanism). For this reason it is useful to examine the data by an ordination method as well. We saw that in the case of the cancer data, the profiles seem to be distributed in a continuum with few gaps. We would have missed this finding entirely had we not proceeded with the ordination of the data. From the ordination in Figure 2 we must conclude that the clusters, while stable, merely segment the data for convenient description of the hyperspace. They do not reveal a conventional clustering model with marked gaps between the clusters.

The smoothly intergrading effects of ETH, GEN, and GEO could imply at least two different models. Either the putative causes are single forces exerting a continuous range of effects, or they are the resultants of multiple forces, whose joint effects produce the continuous range of effects. The second model is the more likely one in this study, since ETH, GEN, and GEO all are undoubtedly multifactorial. ETH will have different cultural components as well as some genetic ones. GEN is affected by 26 different genetic systems in our data and GEO represents geographic space that isolates populations, as well as numerous spatially patterned environmental variables, such as temperature, rainfall, and insolation.

Confirmation of results

We deemed it desirable to obtain confirmation that the hyperspace segments represented by our clusters have epidemiological reality and utility. Can they be associated with relevant, independently obtained information?

Using Mantel tests (Mantel, 1967; Sokal and Rohlf, 1995), we carried out a series of assessments of design matrices representing the cluster membership indicated in column 3 of Table 1 against a series of design matrices depicting available extrinsic criteria. The results are summarized in Table 2. There are 13 male–female pairs of identical (organ-specific) cancers in our study. If our clusters are epidemiologically meaningful, we expect paired male and female cancers to share the same clusters. From Table 1, we can see that 10 of the 13 pairs actually do so. This result, featured in the first row of Table 2, is significant at $P = 0.0001$.

Investigation of the three pairs that do not share the same cluster (bladder, larynx, and oesophagus) reveals great sex differences in both incidence and mortality in these cancers. The male:female ratios of both statistics range from 3.41 to 23.48. The only other cancer with ratios in this range is lung cancer,

TABLE 2. Comparisons of k-means clusters of 31 path-coefficient profiles with available extrinsic criteria (pairwise and first-order partial correlations)

Type of comparison	Extrinsic criterion	Matrix correlation, r	P-value[a]	First-order partial r[b]	P-value[a]
Male–female cancer pairs		0.2322	0.0001		
Etiological categories[c]	Genetics	0.1704	0.0148	0.1490	0.0225
	Tobacco	−0.0265	0.4487	−0.0676	0.0505
	Other environmental factors	0.1163	0.0380	0.0991	0.0594
	Endogenous factors	0.0449	0.2271	0.0279	0.2533
	Infectious agents	0.0242	0.5385	−0.0109	0.4916
Predicted paths	ETH	−0.0557	0.2012	−0.0707	0.0968
	GEO	0.3533	0.0008	0.3237	0.0005
Mortality correlogram clusters	1-dimensional correlograms	0.4257	0.0001	0.4081	0.0001
	2-dimensional correlograms	0.0447	0.1928	0.0059	0.3849

[a] All P-values are permutational and based on 10,000 permutations, yielding $P = 0.0001$ as the lowest possible probability we were able to obtain.

[b] In the partial correlations, it is the design matrix showing the male–female pairs of cancer profiles that is kept constant.

[c] The numbers of cancers associated with the five etiological groups follow. The numbers in parentheses furnish a breakdown of the total number into the five cancer-profile clusters I through V: Genetics 14 (5, 0, 0, 7, 2); Tobacco 10 (1, 2, 1, 2, 4); Other environmental factors 24 (4, 1, 3, 10, 6); Endogenous factors 10 (1, 2, 0, 5, 2); Infectious agents 3 (0, 1, 2, 0, 0). The total numbers of cancers in clusters I to V are 6, 5, 3, 11, and 6.

for which male and female profiles both are grouped into cluster V, but are separated along the minimum spanning tree by two members of cluster IV (Fig. 2). In consequence, some of the path coefficients of the two sexes in the three cancers differ substantially as well, as can be seen in Table 1. This is responsible for placing male and female profiles into different k-means clusters.

When interpreting the remaining tests of matrix correlation reported in Table 2, we have to allow for the correlation between the path coefficient profiles of the paired male and female cancers. To do so, we calculated first-order partial correlations between the design matrices indicating cluster membership of cancer profiles and the design matrices representing the various extrinsic factors, holding constant a third matrix, the design matrix indicating the male–female cancer pairs. The significance of each partial matrix correlation was tested by the Smouse-Long-Sokal test (Smouse et al., 1986). We note that despite significant correlations (not shown) for most design matrices with the male-female cancer pair matrix, the effect of holding this matrix constant decreases the partial correlations only slightly and generally does not change the P-values very much. An attractive alternative to computing partial correlations in this manner would be to carry out the matrix comparison tests separately for male and female cancers. However, simulations (not shown) indicate very low power for tests with matrices of dimension 15 by 15 or 16 by 16. Therefore, we did not pursue this approach.

We coded etiological inferences for the 31 organ- and sex-specific cancers furnished in Higginson et al. (1992). We grouped the stated etiologies into the following five categories: genetics, tobacco, other environmental factors (e.g., dietary habits, occupational factors, industrial pollution, UV radiation), endogenous factors (e.g., metabolism, multiparity, hormones, age), and infectious agents (e.g., human papilloma virus). Details on the breakdown by cluster for each etiology are furnished in footnote c to Table 2. Following Higginson et al. (1992), male and female cancers of a pair were always coded the same. We compared membership of cancers in classes based on their etiologies to cancer clusters based on similarity of path diagrams. Specifically, we are *not* testing whether any of the etiological categories are carcinogenic. That had been established in previous work (see Higginson et al., 1992). We are testing whether given etiological categories are preferentially associated with one or more of our cancer-site clusters. We would consider such positive association to be supportive evidence for the epidemiological utility of the k-means clusters based on the path coefficient profiles. The results are shown in rows 2–6 of Table 2. The matrix correlation coefficients, r, measure whether the cancer sites with a common etiology are randomly distributed over cancer-site clusters I–V (the null hypothesis) versus clumping within one or more of the clusters (the alternative hypothesis). The first-order partial correlations measure the same hypotheses, except that the effect of male and female cancers being similar (located in the same cluster) is eliminated by computing the partial correlations in the manner described above. Of the etiological factors, genetics and the catchall category “other environmental factors” are significantly associated with the k-means clusters. The latter is only marginally significant when the first-order partial correlation is considered. Our results for tobacco deserve special mention. There is marginal evidence for overdispersion, i.e., the 10 tobacco-associated cancer sites among the 31 sites studied appear to be distributed more evenly among our five clusters than expected by random allocation. This is seen in the marginal significance ($P = 0.0505$) of the negative partial correlation (–0.0676) between cluster structure and tobacco-mediated cancers. This suggests that tobacco, the only specific etiological agent tested (the other categories are all more inclusive), is not associated with specific regional or ethnohistorical patterns in these European populations.

Next, we reexamined the etiologies of the cancers described in Higginson et al. (1992) and attempted to predict the causal path or paths (ETH, GEN, or GEO) along which they would act in the path diagrams. Differences in incidence and mortality of specific cancers in various ethnic groups were assigned to predicted ETH which, as can be seen in Figure 1, affects INCID and MORT directly, as well as affecting them via GEN. We reserved assignment to predicted GEN to those cancers associated with specific genetic factors (e.g., female breast cancer). (However, predicted GEN turned out to be identical to the design matrix for genetics in the etiological categories of Table 2. It is therefore not repeated in the table.) Predicted GEO was invoked only for climatic factors (e.g., UV radiation) or some geographically varying feature (e.g., iodine content of soils). Occupational factors, such as work in asbestos, dye, or rubber factories, were not assigned to a predicted category, since there was no clear association with ethnic background or geography. Correlation of predicted ETH and GEO with the matrix of cluster membership derived from k-means analysis of the path coefficient profiles are presented in rows 7 and 8 of Table 2. We cannot demonstrate significant association with our predictions for ETH, but we can for GEO (and would have for GEN, had we repeated the results of row 2 in Table 2).

Rosenberg et al. (1999) analyzed the spatial autocorrelation of the Western European cancer mortality data (Smans et al., 1992) that are part of our present dataset. In their tables 2 and 3, respectively, Rosenberg et al. (1999) present the results of k-means clustering of the one- and two-dimensional correlograms for these data. These authors obtained four clusters of the former and five of the latter, members of a cluster sharing similar spatial patterns of cancer mortality. We cannot expect too much agreement between the classifications for 31 cancers common to the study by Rosenberg et al. (1999) and the present analysis, as those from the previous study are based entirely on the spatial properties of mortality data and are limited to Western Europe. Yet we find considerable and highly significant agreement between the classes of path coefficient profiles and those of one-dimensional mortality correlograms, but no agreement with the two-dimensional mortality correlograms (see rows 9 and 10 of Table 2).

Overall, there is clear significant association between the k-means clusters and the second, third, and fourth type of comparison in Table 2, yielding $P < 0.002$ by a Bonferroni procedure (Sokal and Rohlf, 1995). We conclude, therefore, that etiological and other extrinsic information on the cancers is associated significantly with the five clusters of cancer profiles, demonstrating that the profiles carry epidemiologically meaningful information.

In an indirect manner, the results reported in this section also confirm the reality of the clusters of cancer profiles obtained by the k-means method and the validity of the underlying (dis)similarity structure of the profiles as revealed in the ordination. If the 31 path diagrams differed only randomly in the magnitudes of their path coefficients, it would be difficult to explain the significant association observed between the positions of the profiles in the variable-space (as witnessed by their membership in the five clusters) and the various extrinsic criteria.

In this study we have demonstrated the importance of ethnohistory in the determination of some cancer incidence and mortality rates. If even differences in ancient ethnic mixtures between pairs of sampling points in Europe show effects on cancer rates, one can expect that distinct ethnic minorities in modern populations will exhibit such effects *a fortiori*. Our results emphasize the importance of studying ethnic differences in epidemiological investigations.

ACKNOWLEDGMENTS

This article is contribution No. 1117 in Ecology and Evolution from the State University of New York at Stony Brook. We thank F. James Rohlf for valuable advice and assistance on computation and interpretation of the data.

LITERATURE CITED

Berrino F, Sant M, Verdecchia A, Capocaccia R, Hakulinen T, Estève J (eds.). 1995. Survival of cancer patients in Europe: the EUROCARE Study. Lyon: IARC Sci Pub No 132.

Berrino F, Capocaccia R, Estève J, et al. (eds.). 1999. Survival of cancer patients in Europe: the EUROCARE-2 study. Lyon: IARC Sci Pub No 151.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570.

Clifford P, Richardson S, Hémon D. 1989. Assessing the significance of the correlation between two spatial processes. Biometrics 45:123–134.

Dutilleul P, Stockwell JD, Frigon D, Legendre P. 2000. The Mantel-Pearson paradox: statistical considerations and ecological implications. J Agric Biol Environ Stat 5:131–150.

Higginson J, Muir CS, Muñoz N. 1992. Human cancer: epidemiology and environmental causes. Cambridge, UK: Cambridge University Press.

Krzanowski WJ. 1988. Principles of multivariate analysis. Oxford: Clarendon Press.

Leduc A, Drapeau P, Bergeron Y, Legendre P. 1992. Study of spatial components of forest cover using partial Mantel tests and path analysis. J Veg Sci 3:69–78.

Legendre P, Legendre L. 1998. Numerical ecology, 2nd English ed. Amsterdam: Elsevier.

Li CC. 1975. Path analysis—a primer. Pacific Grove, CA: Boxwood Press.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1. Berkeley: University of California Press. p 281–297.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209–220.

Morton NE. 1982. Outline of genetic epidemiology. Basel: Karger.

Mourant AE, Kopeć AC, Domaniewska-Sobczak K. 1978. Blood groups and diseases. Oxford: Oxford University Press.

Muir CS, Waterhouse J, Mack T, Powell J, Whelan S (eds.). 1987. Cancer incidence in five continents, vol. 5. Lyon: IARC.

Nantel P, Neumann P. 1992. Ecology of ectomycorrhizal-basidiomycete communities on a local vegetation gradient. Ecology 73:99–117.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Parkin DM, Smans M, Muir CS (eds.). 1983. Cancer incidence in the USSR. Cancer incidence in five continents, vol. 3. Lyon: IARC.

Parkin DM, Muir CS, Whelan SL, Gao YT, Ferlay J, Powell J (eds.). 1992. Cancer incidence in five continents, vol. 6. Lyon: IARC.

Prevosti A, Ocana J, Alonso G. 1975. Distances between populations of *Drosophila subobscura* based on chromosome arrangement frequencies. Theor Appl Genet 45:231–241.

Rohlf FJ. 1972. An empirical comparison of three ordination techniques in numerical taxonomy. Syst Zool 21:271–280.

Rosenberg MS, Sokal RR, Oden NL, DiGiovanni D. 1999. Spatial autocorrelation of cancer in Western Europe. Eur J Epidemiol 15:15–22.

Smans M, Muir CS, Boyle P (eds.). 1992. Atlas of cancer mortality in the European Economic Community. Lyon: IARC Sci Pub No 107.

Smouse PE, Long JC. 1992. Matrix correlation analysis in anthropology and genetics. Yrbk Phys Anthropol 35:187–213.

Smouse PE, Long JC, Sokal RR. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool 35:627–632.

Sokal RR. 2004. Raymond Pearl's legacy: the proper measure of Man. Am J Hum Biol 16: (in press).

Sokal RR, Rohlf FJ. 1995. Biometry, 3rd ed. New York: W.H. Freeman & Co.

Sokal RR, Harding RM, Oden NL. 1989. Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 80:267–294.

Sokal RR, Oden NL, Thomson BA, Kim J. 1993. Testing for regional differences in means: distinguishing inherent from spurious spatial autocorrelation by restricted randomization. Geogr Anal 25:199–210.

Sokal RR, Oden NL, Walker J, DiGiovanni D, Thomson BA. 1996. Historical population movements in Europe influence genetic relationships in modern samples. Hum Biol 8:873–898.

Sokal RR, Oden NL, Rosenberg MS, DiGiovanni D. 1997. Ethnohistory, genetics, and cancer mortality in Europeans. Proc Natl Acad Sci USA 94:12728–12731.

Sokal RR, Oden NL, Thomson BA. 1999a. A problem with synthetic maps. Hum Biol 71:1–13.

Sokal RR, Oden NL, Thomson BA. 1999b. Problems with synthetic maps remain: a reply to Rendine et al. Hum Biol 71:447–453.

Sokal RR, Oden NL, Rosenberg MS, Thomson BA. 2000. Cancer incidences in Europe related to mortalities, and ethnohistoric, genetic, and geographic distances. Proc Natl Acad Sci USA 97:6067–6072.

Upton GJG, Fingleton B. 1985. Spatial data analysis by example, vol I. Point pattern and quantitative data. Chichester, UK: John Wiley & Sons.

Waterhouse J, Muir CS, Shanmugaratnam K, Powell J (eds.). 1982. Cancer incidence in five continents, vol. 4. Lyon: IARC.

Wright S. 1943. Isolation by distance. Genetics 28: 114–138.

Wright S. 1969. Evolution and the genetics of populations, vol. 2. The theory of gene frequencies. Chicago: University of Chicago Press.

Zatonski W, Smans M, Tyczynski J, Boyle P (eds.). 1996. Atlas of cancer mortality in Central Europe. Lyon: IARC Sci Pub No 134.