Chapter 3

Space-time variations and group correlations

Geographical analysis

The objectives of cartography

Like all phenomena which vary across regions, spatial differences in cancer occurrence can be represented on a map. A remarkable degree of sophistication has been achieved in this area. Geographers are convinced that a map can provide, through the simple play of colours, both an overall impression of major differences between regions (such as the juxtaposition of plains and mountains on a geophysical map) as well as a partial or detailed view of the characteristics of a given region.

The design of a map is not only based on aesthetic concerns. In contrast to a table of regional results, a map provides supplementary information on the contiguity and the proximity of regions. The fact that neighbouring regions might be similar with regard to the phenomenon under study can be an essential element in interpretation.

The cartographic illustration of mortality by cause is not a new idea. It has for some time formed the basis for political discussions on inequalities between regions and been a tool for health planners, for example, in the regional planning of health services. There has been a revival of interest in this approach over the past few years mainly as a result of the development of specific computing techniques. In the field of cancer, the development of cartography is relatively recent, with some notable exceptions such as Figure 3.1, showing crude cancer mortality in Switzerland for the period 1911-1914 [1].

Over the past few years, a number of cancer atlases have been produced, generally from mortality data. Examination of these atlases reveals many differences in the methods used, suggesting that their objectives differed somewhat. Some are designed to show only broad spatial patterns (for example, through a limited number of regions or colours), others indicate a systematic attempt to show, by magnification, highly localized differences through the use of a rich array of colours or a fine division of geographical units. Despite these differences, it seems obvious that the main objective of cancer atlases is to provide basic information for etiological research. Their implicit goal is therefore to allow the image of geographical variation in the rate of a given cancer to be superposed on other maps, real or imaginary,

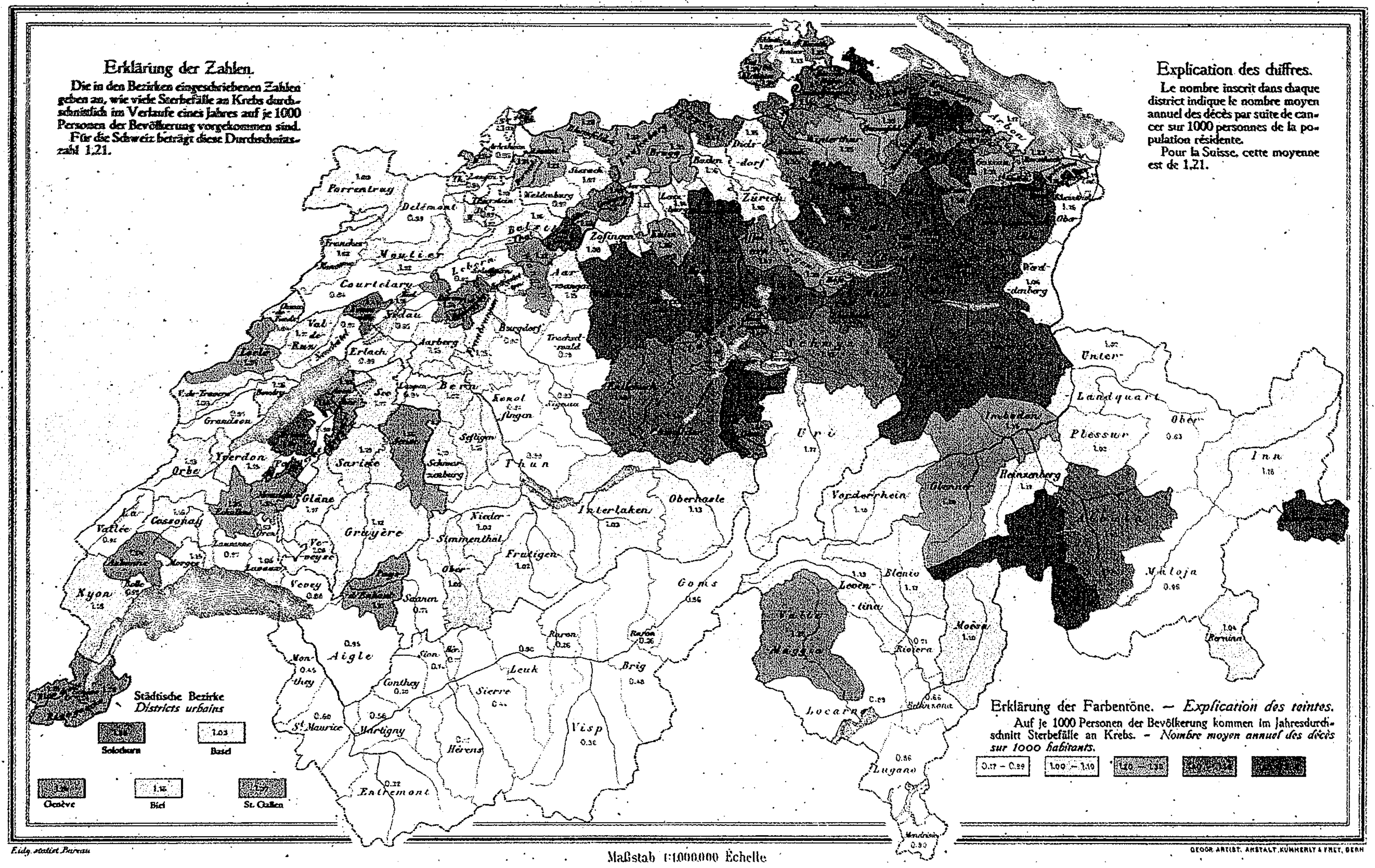


Figure 3.1 Cancer mortality (crude rates) in Switzerland in 1901-1910
Source : Swiss Bureau of Statistics [1]

of one or more environmental characteristics or individual behaviour potentially implicated in the variation of cancer risk. It is not certain that such superposition of these factors can be achieved with a single cartographical representation, since all the evidence suggests that exposure to diverse etiological factors can be distributed at different scales.

If interest is in factors which vary locally, the map would be expected to define zones in which the incriminated exposure can be found. In this case, a detailed subdivision is adopted to obtain relatively homogeneous zones with respect to the exposure under consideration. For example, mesothelioma is particularly frequent in Italy in coastal areas where naval construction, known as a source of exposure to asbestos, is concentrated [2] (Figure 3.2).

If, on the other hand, interest is in factors which are distributed more widely over the spatial map (such as cultural and regional behaviour, or climatic conditions), the objective will no longer be to show the level of risk in a particular area compared to adjacent areas but to provide a more homogeneous representation of broad patterns in the phenomenon. If the intensity of the phenomenon varies progressively from one region to another across all or part of the country under consideration, the differentiation of the areas should visually show this gradient. Such a progression

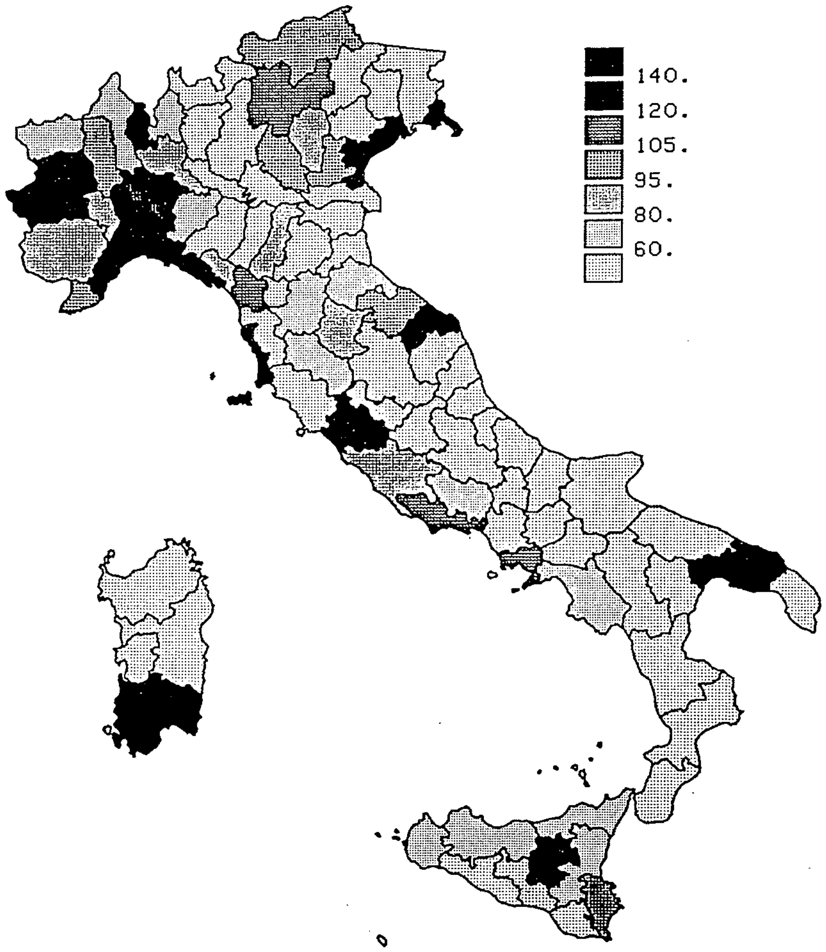


Figure 3.2 Mesothelioma mortality, men, 1975 – 1977
Source : Cislaghi et al. [2]

could in fact suggest a dose-response relationship with the level of exposure, whereas a finer subdivision could be influenced by local variations which are irrelevant to the phenomenon under consideration. An interesting example of geographical variation on a large scale is provided by mortality for malignant melanoma of the skin [3,4]. In the USA (Figure 3.3a), mortality due to this cancer increases as the latitude decreases, while in Europe (Figure 3.3b), the phenomenon is inverted. In the USA, increased exposure to ultraviolet light in the more southern regions results in a detectable increase in melanoma mortality. The way in which this country has been populated by migrants of different origins has led to an unplanned adjustment for ethnicity. Immigrants from different ethnic backgrounds are effectively distributed randomly throughout the country. In Europe, in contrast, factors linked to ethnicity are the most important determinant of melanoma risk and mask the effect of place of residence; individuals most susceptible to ultraviolet light have remained in the north, while recently adopting a life style involving significant exposure to the sun.

Beyond the objectives illustrated by these examples, cancer atlases which have appeared so far have been works of general scope destined for a wide readership. Thus their authors have often made compromises such that the atlases do not necessarily answer the needs of etiological researchers. Nevertheless, the techniques which they apply are fundamental tools which have been used for a long time in descriptive epidemiology to solve etiological problems. As early as 1848, John Snow identified the source of the epidemic which ravaged London by using a map by district of mortality rates due to the disease. Joint study of this map and that of the areas covered by different water suppliers revealed similarities which convinced Snow to follow his investigations at the level not only of the district but also of individual houses [5]. This more detailed approach was rendered necessary because the old part of London was served by two companies, the Lambeth Society and the Southwark and Vauxhall Society. Analysis of the water showed among other things differences between the companies not only in the content of organic material but also its acidity, which undoubtedly affected the conditions for bacteria growth. These geographical observations led Snow to identify the vehicle of the then unknown agent of the disease, *Vibrio cholerae*.

Since that time, the representation of risk or exposure by means of a geographical map and the tools for analysing geographical distributions have advanced considerably. The following sections describe both aspects.

Methods

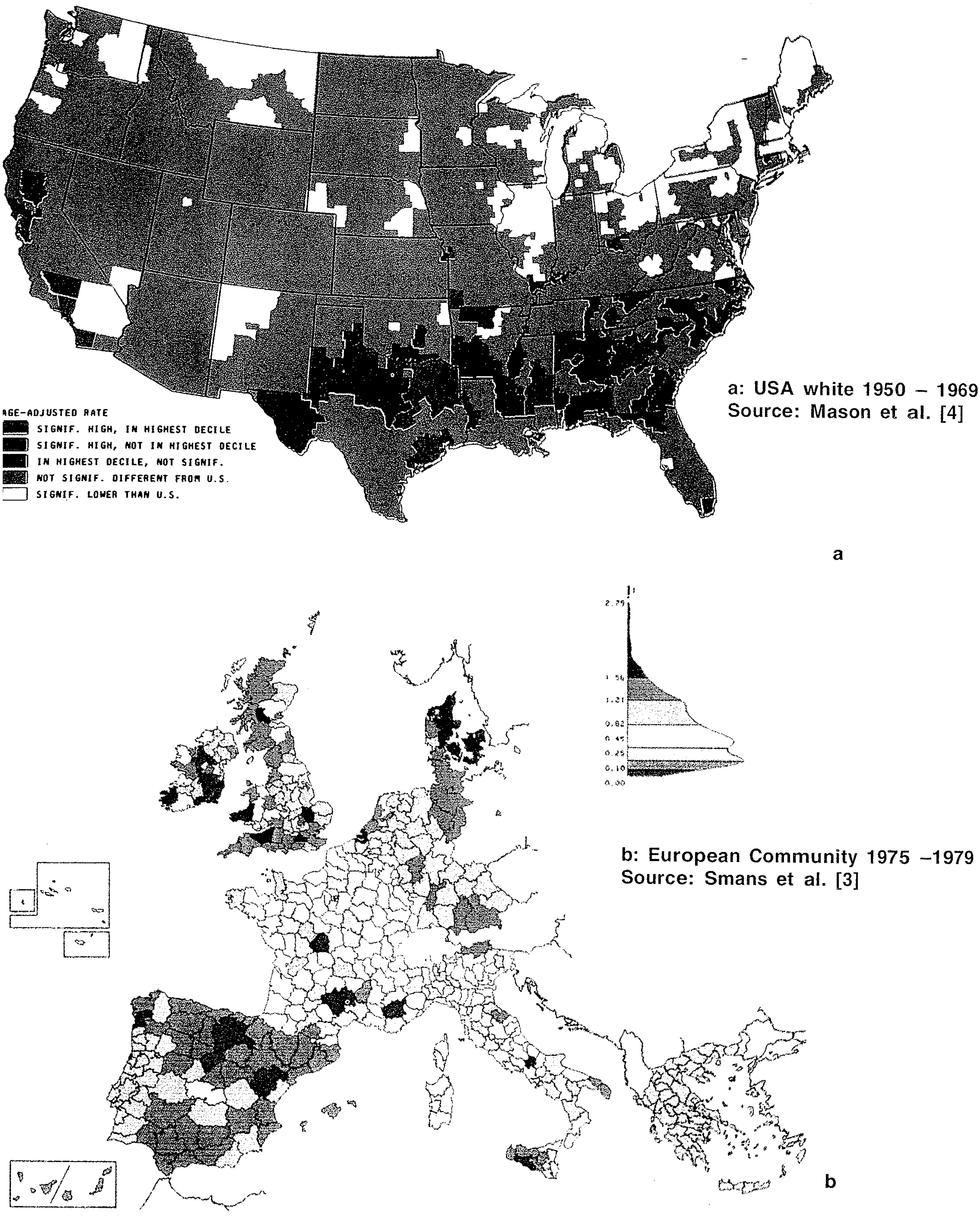
Geographical division

Geographical representation of cancer frequency is provided by the juxtaposition of areas of different colours or shades, each of which represents a level of frequency. The boundaries and especially the number of the areas determine the degree of detail of the map and thus its overall appearance. As has been indicated, the issues are different when the goal is to produce a series of maps fulfilling a purely descriptive need, such as an atlas of selected cancer sites, or to indicate regions corresponding to a risk or to a given exposure in the context of a specific etiological investigation. The geographical level at which data are available (numerator and denominator) is not always the most important constraint. In practice, difficulties are more likely to occur because of the need to ensure statistical stability for the risk estimates in each, or at least most, areas. It is important to maintain an appropriate ratio between the incidence or mortality from one region to another and the corresponding random variation. For example, it would be unreasonable to define areas which only include four expected cases on average, if the objective is to classify areas into categories representing relative differences of 25%. In this situation, the coefficient of variation of the rate is of the order of $1/\sqrt{4} = 50\%$ (see Chapter 2, page 53). Accordingly, geographical units which are sparsely populated are often grouped together.

If data are available, it is sometimes preferable not to work with administrative subdivisions. For example, in the Finnish study on the relationship between life style and cancer incidence, communities have been grouped together to form areas of 10 000 people, characterized by their geographical proximity as well as their similarity with respect to appropriately selected socioeconomic variables [6].

In some situations, the definition of areas is in response to a specific etiological problem. The goal of cartography is then to illustrate a specific hypothesis, for example, to evaluate the effect of radiation around a nuclear power station or of pollution on the frequency of respiratory cancer. The objective then is to form one or more areas in which the exposure being studied is homogeneous. Recording information from small geographical units becomes essential. Because of this requirement, many countries have introduced systems by which data from population censuses and periodic reports (such as death by cause) are available for geographical units defined by appropriate cartesian coordinates [7]. When the source of risk is at a specific point, the usual approach would be to define the area as all squares located within a circle around this point (Figure 3.4) or between concentric circles, in order to demonstrate a dose-response relationship.

In other examples, the whole region is divided into areas depending on the intensity of exposure, as determined by measurements made at specific points in the region (e.g., measurement of ultraviolet light at meteorological stations). The aim is to divide the region into homogeneous areas around points where measurements have been carried out. Dirichlet's mosaic provides a simple and elegant solution [8]: the region to be mapped is divided into areas such that each point in a specific area is closer to the measurement point situated in it than to any other measurement point. This tiled area is obtained by connecting the perpendicular bisectors of the sides of triangles formed by the measurement points. A more sophisticated solution is based on interpolation from the measurements using polynomial regression. Division into areas of homogenous exposure can be constructed from contour lines of the resulting surface. This method can also be used after having artificially localized a regional measurement (e.g., rate per resident) at the centre


AGE-ADJUSTED RATE
SIGNIF. HIGH, IN HIGHEST DECILE
SIGNIF. HIGH, NOT IN HIGHEST DECILE
IN HIGHEST DECILE, NOT SIGNIF.
NOT SIGNIF. DIFFERENT FROM U.S.
SIGNIF. LOWER THAN U.S.
a
2.79
1.56
1.31
0.82
0.45
0.25
0.10
0.00
b

a: USA white 1950 – 1969
Source: Mason et al. [4]

b: European Community 1975 –1979
Source: Smans et al. [3]

Figure 3.3 Melanoma mortality; women

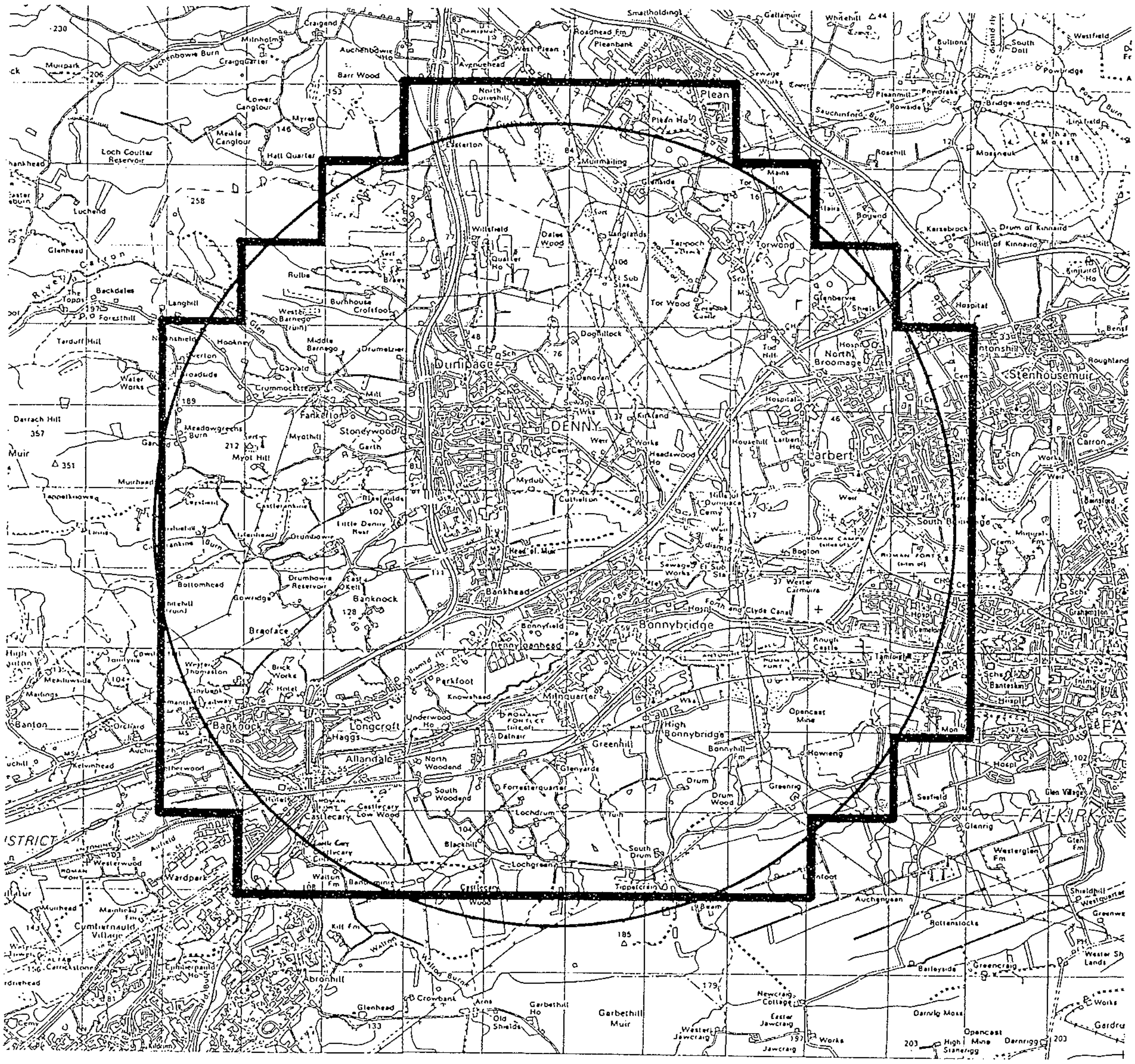


Figure 3.4 Construction of a circle with a radius of 5 km from a 1 km-grid square
Source: Carstairs et al. [7]

of gravity of the region, which is obtained by weighting according to population density.

When small adjacent geographical areas are grouped together to create homogeneous aggregates with respect to exposure, it is obviously important to check that the level of exposure can be considered equal in the areas which have been grouped together. One method of grouping based on the statistical significance of the differences in exposure between adjacent regions will be discussed below (see page 134).

The first objective of these diverse techniques is thus to create areas of more homogeneous risk by departing from the constraints of the politico-administrative subdivisions. Note that when the techniques involve grouping or interpolation, they have the additional advantage of smoothing the exposure data, eliminating the inconvenience of large random fluctuations which usually affect small area statistics.

This is even more evident when the methodology is used for the description of incidence or mortality; for example, a polynomial regression has been used to represent curves of stomach cancer mortality in Italy [9] in a purely descriptive context (Figure 3.5).

Choice of a risk indicator

When the objective is to show variation in risk as opposed to crude rate or number of cases, the graphical representation should use a risk indicator which is adjusted for age. Both direct and indirect standardization methods have been used for this purpose in published atlases.

For direct standardization, either the world or European population is most often used. This choice undoubtedly reflects the desire to expand the atlas's role to international comparisons. Nevertheless, the various atlases which have appeared are seldom comparable, because of the large variation in the choice of the risk categories and colours. None is based on the cumulative rate (Chapter 2, page 60) which would be the most readily interpretable index on a probability scale and make the various maps directly comparable.

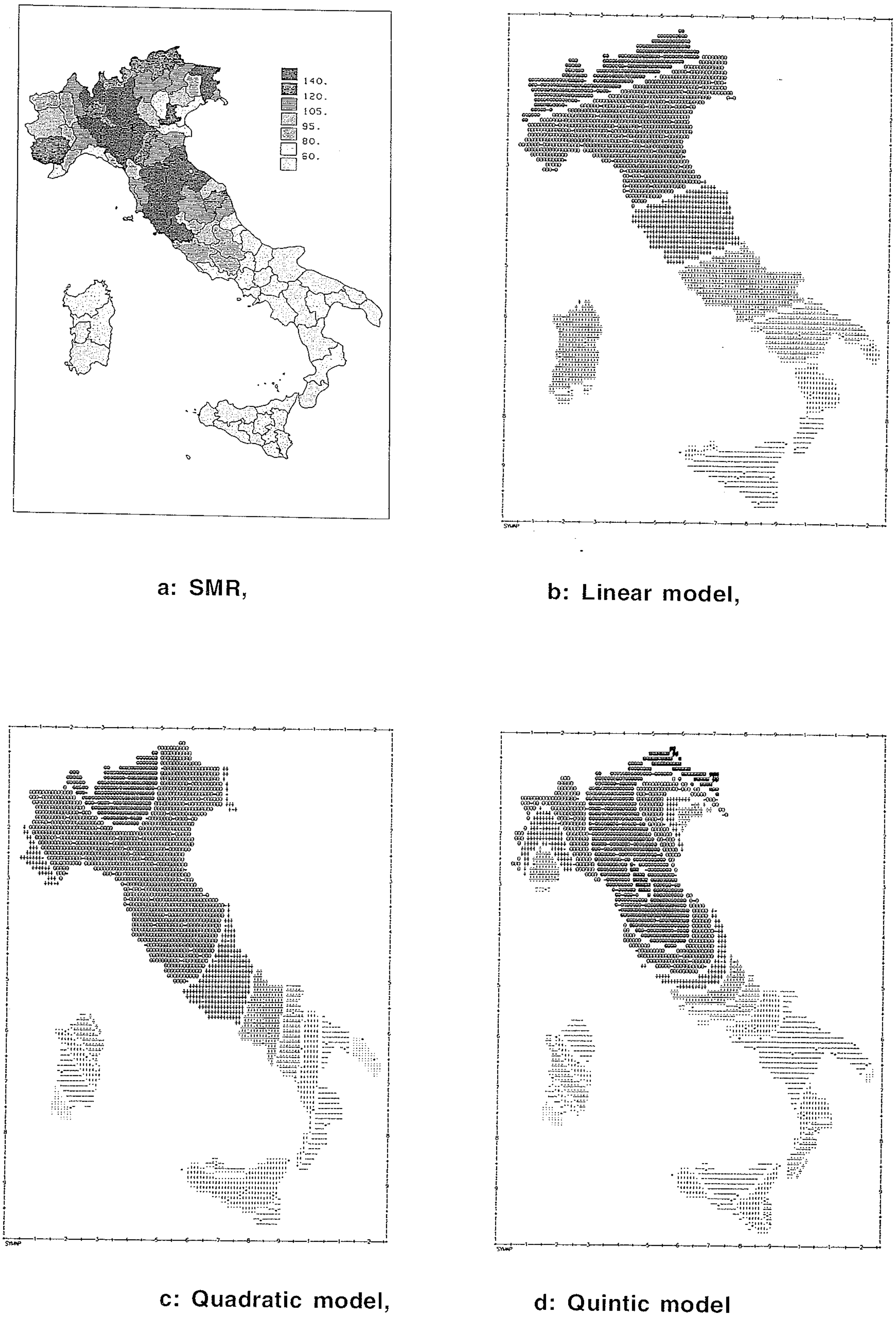
Many authors have chosen indirect standardization. This option is justified if the primary objective of a cancer atlas is to represent risk variations within a country. Geographical areas are then classified by their standardized mortality or morbidity ratio (SMR). This index generally has the advantage of providing more precise statistical estimates than the directly standardized rate (Chapter 2, page 100). The reference rate adopted for the calculation of the SMR is in general the incidence or mortality estimated in the region being mapped.

Definition of risk classes

We have already seen that the number of risk classes cannot be determined without taking into account the statistical precision of the risk indicator. Precision is equally relevant in the choice of scale and class limits, as we shall see below.

A priori, a larger number of classes should provide a more detailed picture of risk variation. However, dividing the area too finely diminishes the effect of the colour or shading contrasts required to distinguish the risk variation clearly. Moreover, as a general rule, the homogeneity of classes is proportional to their number: if there are few classes, differences between values in the same class could be much larger than those existing between the central values of two adjacent classes, which are nevertheless represented by different colours.

The colours chosen to represent the various levels of risk differ substantially from one atlas to the next. A principle generally applied is to make the zone representing average risk the least coloured. Zones of increasing (or respectively decreasing) risk are represented by colours which are arbitrarily chosen, but sufficiently contrasting visually. The chromatic intensity progressively decreases from extreme risk classes to intermediate classes.



a: SMR,

b: Linear model,

c: Quadratic model,

d: Quintic model

Figure 3.5 Stomach cancer mortality in Italy; men, 1975-1977
Source: Cislaghi et al. [2] and personal communication

It is not surprising that many authors choose red and green to characterize respectively an increase or a decrease compared to a standard risk. Culturally, red indicates danger while green represents ecology. Note also that the range of shades is not obligatorily centred on the average index value but can be distributed asymmetrically around the average index value such that those classes representing risk increase are broken down more finely, as has been done in the Chinese cancer atlas [10]. The subtlety of options used in the various atlases reveals an intention to use the physiology of visual perception, particularly in the choice of colours, to best communicate the desired message [11, 12].

The simplest procedure involves setting the limits of the classes based on an equal division of all risk index values after disregarding extreme values when they are outliers. Under this method, the scale depends on the data, and does not lend itself to comparison between maps of different cancer sites or the two sexes.

When the index is a relative measure (for example, an SMR), the same scale is adopted for all sites. Transition between colours is then immediately interpretable in terms of relative risk increases: for example, a relative risk scale increasing by steps of 25% from left to right open-ended categories. These categories at the extremities of the scale are defined by the maximum number of classes to be used in mapping. This approach has been frequently used in atlases, as it has the advantage of allowing comparisons to be made between sites and between sexes. In the French atlas [13], for example, it can be seen that stomach cancer mortality is one and a half times higher in Brittany than in the rest of the country for both men and women, and that the maps for both sexes are similar. However, this type of comparison is of little value when the standard levels used in the maps being compared (SMR = 1) are very different from each other. For example, for lung cancer in France, the comparison of zones characterized by values between 125 and 150 of the SMR for men and women is not directly informative, because of the difference in background risk between the two groups.

An examination of maps using fixed limits for risk categories shows that the geographical variation in risk is extremely variable between sites. Thus in the French atlas, maps representing oesophageal cancer are more variegated than those for colon cancer. This methodology may be better suited to a public health perspective than to etiological research, in which all real risk differences can be of interest.

The proportion of each colour on the map is directly dependent on whether or not a fixed scale is adopted. If distribution of risk is narrow, the map will be largely monochromatic. If the distribution tends to be bimodal, the map will be largely made up of colour zones representing high and low risk respectively. If the distribution is equally spread, all the selected colours will be almost equally used.

In order to describe all observed variability, the original scale has been replaced in some atlases by grouping together risk classes based on percentiles. For example, in the Scottish atlas directly standardized rates have been divided into seven classes with limits determined by the 5, 15, 35, 65, 85 and 95 percentiles [11]. The middle class therefore includes 30% of the values. By definition, this method leads to the use of a different scale for each site and for both sexes. Each of these scales is a function not only of the risk values but also of the shape of their distribution.

For a given number of classes, the use of percentiles makes the apparent variability of the risk index equal and maximal. It is thus impossible to judge the size of this variability visually, as each map makes the same use of the different colours and extreme values are no longer apparent. On the other hand, when variations in risk are small, any contrasts, gradients or autocorrelative phenomena can be clearly appreciated.

In order to reconcile the advantages of a relative measure with those of a measure expressed on an absolute scale, a division based on a logarithmic scale has been used in the Chinese atlas. All maps can then be built with one scale regardless of site or sex. In this system, the increase in risk for a class compared to the level of risk of the class immediately preceding it is represented on a multiplicative and not an additive scale, so that only pronounced variations are apparent; this is well illustrated by the map of oesophageal cancer in China (Figure 3.6) [10].

We have noted on several occasions that risk estimates are subject to statistical fluctuations that can be of different magnitude in different regions. Taking this variability into account will modify the interpretation of the map. For example, little significance will be attached to the high value of female mortality for cancer of the buccal cavity in France in the département of Cantal [13]: the value of the SMR is equal to 1.75 but its confidence interval (0.98; 2.88) does not exclude unity.

It is generally accepted that maps produced by the principles described above are usefully complemented by information on variability of the risk indices. In some situations, maps can be simply accompanied by an appended table providing the required data, such as the standard error or the confidence interval of the index. Others attempt to give a geographical view of variability by juxtaposing a map of risk with a map of degree of significance for the same areas [14]. Interpreting the two maps together is not always easy, but it can demonstrate that differences can be significant without being large, if the number of cases is high and/or the populations under study large. Thus the majority of European atlases show significant differences between regions for colon cancer, even though the variation in risk for this cancer is generally relatively small.

Some maps attempt to combine the size of the variation and its degree of significance on one single scale. The atlas of cancer mortality in England and Wales used the following four categories [15]: significantly increased risk; increased risk, but not significant; not increased risk; significantly decreased risk. Such a scale allows all rates significantly increased with respect to the reference rate to be placed at the top of the colour hierarchy even if the increase is in reality very small. Risks which are substantially increased, but not significantly so, will appear lower down in this hierarchy. In practice, the procedure is acceptable only if the geographical areas are divided equally (in terms of population), such that the statistical variability is of the same order for a given site.

The difficulties described above can be minimized or avoided in the interest of compromise. However, the study of spatial data, especially for specific problems, requires a more rational approach to account for random variability. The methods described below are more suitable in these situations.

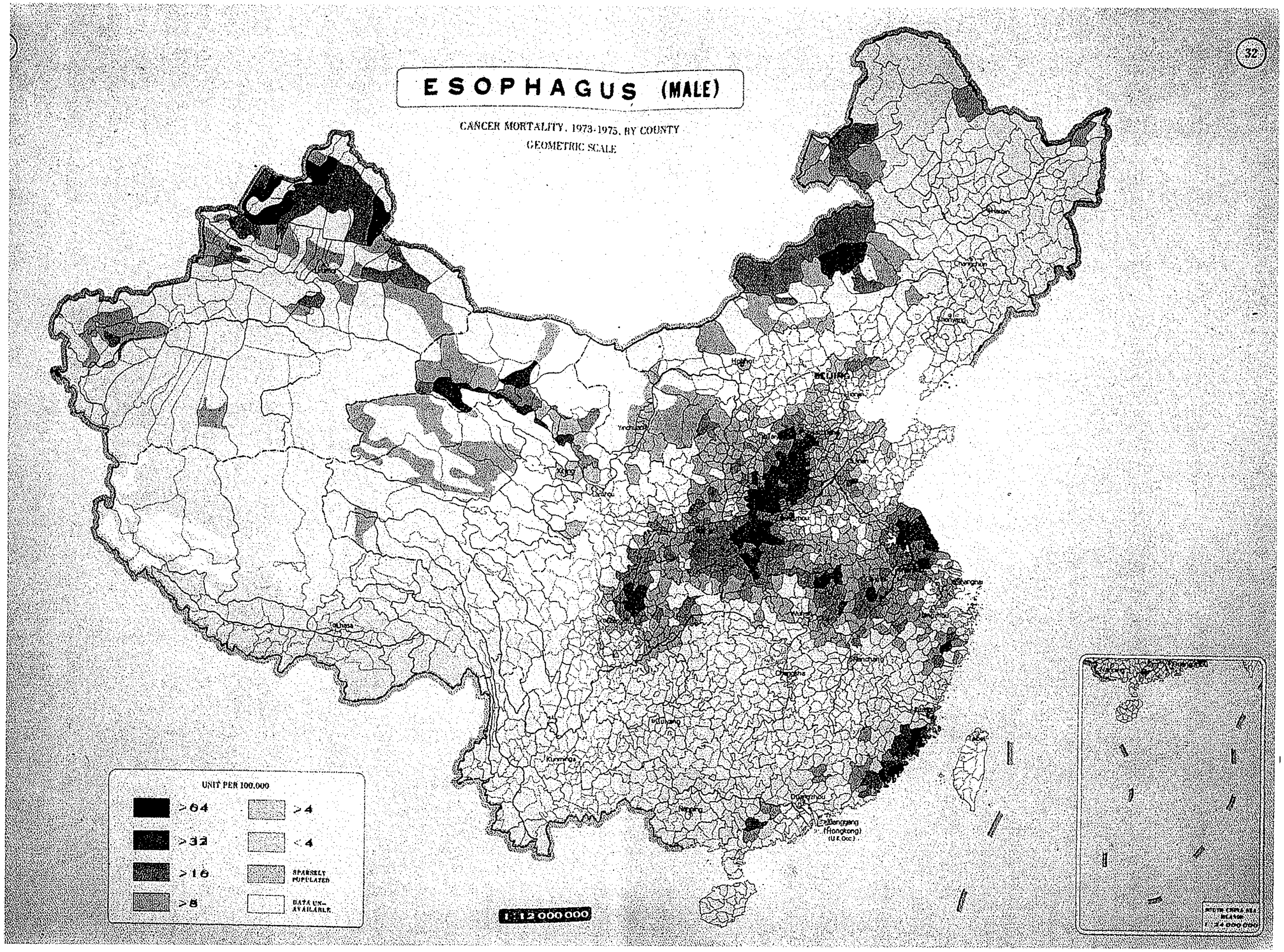


Figure 3.6 Oesophageal cancer mortality in China; men, 1973-1975
Source: China Map Press [10]

Tools to interpret geographical data

Autocorrelations

Graphical representation of disease frequency is not the only objective of the geographical study of the disease. Although this objective is important, quantitative answers to certain simple questions should also accompany the presentation of the data, to facilitate their interpretation.

The first of these questions concerns geographical variability: are rates different from one region to another? A homogeneity test such as that given in Chapter 2 (page 87) can obviously be carried out, but is of little value because it does not take into account the spatial structure of the geographical units being studied. As has been suggested previously, neighbouring regions are often subject to similar cancer risks: exposure to factors influencing the level of incidence or mortality is often more similar in neighbouring regions than in distant regions. Exposure can also vary continuously in a particular direction, resulting in a risk gradient such as those cited for melanoma mortality in the USA and Europe (Figures 3.3a and 3.3b) [4]. When the direction of the gradient is already known, as in this example, the significance of the variation in risk can be evaluated using a test on one degree of freedom (Chapter 2, page 90). However, in the majority of situations, no assumptions can be made about the direction of the gradient and the validity of the test can be questioned if the direction was suggested by observation of the map.

The spatial distribution of risk factors leading to local correlations in disease rates will generally be more complex than the risk factor distributions which determine larger-scale geographical patterns described above. For small areas it is of interest to measure and test the similarity of disease rates on a much finer scale. Local variations in processes which determine cancer incidence or mortality in the area under study are the focus of interest rather than overall trends. We therefore need to evaluate the correlation of risks in adjoining regions, also referred to as the spatial autocorrelation of the random process which gives rise to the observed geographical variations of incidence. A significant autocorrelation is frequently found. Taking this correlation into account using methods described below results in a more satisfactory description of the spatial distribution of risks and thus a better representation of incidence.

Even when the risks are the same over all regions studied, their estimation can result in a spatial correlation simply because the most accurate estimates, which are those in the most populated regions, are also found most often in neighbouring regions. The values observed in these regions will therefore be close simply because they estimate the common risk value better. This autocorrelation of the population sizes in the different geographical units is common and should be kept in mind, since, in this situation, the spatial autocorrelation observed is not in the risks but only in their estimates.

If there is no autocorrelation in risks, the test of geographical homogeneity reduces to the classical comparison of several groups. The presence of spatial correlation in risk establishes heterogeneity *de facto*, but its absence does not confirm homogeneity.

Finally, it is possible that a substantial variation on a large scale and spatial autocorrelation on a finer scale can be observed simultaneously. Methods described below for modelling spatial processes can be used in this situation. Here, we will simply show how to evaluate spatial autocorrelation from risk estimates based on the SMR. We first describe the different indices available, as if the risks were directly observable.

Suppose that the spatial structure of the geographical units is defined by a matrix of weights **W**, the elements w_{ij} of which measure the geographical proximity of the regions i and j. Most often, **W** will be an adjacency matrix whose elements w_{ij} are equal to 1 if i and j are adjacent and zero otherwise. Moreover, let X_i be the spatial process defined by the relative risks of disease ρ_i in the different regions (i = 1,...,n) (for example, $X_i = \log(\rho_i)$ or $X_i = \text{rank}\ (\rho_i)$]. Moran's coefficient [16] measures autocorrelation of the spatial process X_i using an index which is very close to the classical correlation coefficient :

$$I = \frac{n \sum_{i \neq j} w_{ij} (X_i - \overline{X})(X_j - \overline{X})}{S_0 \sum_i (X_i - \overline{X})^2} \tag{3.1}$$

where S_0 is the sum $\sum_{i \neq j} w_{ij}$ which, in the case of an adjacency matrix, is the number of pairs of areas with a common border.

Geary's coefficient [17] measures the average squared difference between risks observed in adjacent areas, and should be small in the case of spatial correlation:

$$C = \frac{n-1}{2\,S_0} \; \frac{n \sum_{i \neq j} w_{ij} (X_i - X_j)^2}{\sum_i (X_i - \overline{X})^2} \tag{3.2}$$

Two other indices have been used for investigating the geographical distribution of cancer risks. Ohno [18] suggested using the number of adjacent areas of the same colour on a map of incidence or mortality. Smans [19] recommended calculating the average difference in ranks of adjacent areas. As we shall see below, these statistics are in fact similar to the statistics used to evaluate time-space clustering. The first is similar to that introduced by Knox to analyse time-space clustering (see page 131) [20] and the second can be written:

$$D = \frac{1}{S_0} \sum_{i \neq j} w_{ij} \left| \text{rank}(\rho_i) - \text{rank}(\rho_j) \right| \tag{3.3}$$

Assuming that X_i have independent and identical normal distributions (under the null hypothesis of no autocorrelation), the means and variances of I and C are given by the formulae:

$$E(I) = -\frac{1}{n-1}$$

$$\mathrm{Var}(I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n-1)(n+1)S_0^2} - \left[-\frac{1}{n-1}\right]^2 \tag{3.4}$$

$$E(C) = 1$$

$$\mathrm{Var}(C) = \frac{(2S_1 + S_2)\,(n-1) - 4\,S_0^2}{2\,(n+1)\,S_0^2} \tag{3.5}$$

where S_1 and S_2 are functions of w_{ij} defined by

$$S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ij})^2$$

$$S_2 = \sum_i (w_{i.} + w_{.i})^2 \text{ with } w_{i.} = \sum_j w_{ij} \text{ and } w_{.i} = \sum_j w_{ji}$$

The means and variances of the statistics proposed by Ohno and Smans can be obtained directly from the formulae given by Mantel in the context of detecting time-space clustering (see page 133, formulae (3.22) and (3.23)).

Unfortunately, these formulae which only depend on the spatial structure **W**, are of little more than theoretical interest. As we saw above, spatial autocorrelation in risk estimates caused by heterogeneity in population sizes can be detected by these tests even if the risks are identical across areas. These theoretical values would only be valid if the population density was constant.

In practice, spatial autocorrelation in risks can only be tested by randomization procedure using the correct null hypothesis described below.

Let k_{xi}, m_{xi} be the number of cases and the person-years in the population of age x of area i. To test the existence of spatial autocorrelation against the null hypothesis of homogeneity. The total number of cases $k_{x.}$ in the different areas are distributed proportionally to the populations m_{xi} according to the multinomial model (Chapter 2, page 87). The estimates of ρ_i in each area are calculated for each simulation and, from these, the autocorrelation statistic and its distribution under the null hypothesis are calculated. Table 3.1 gives the mean and the standard error of the statistics I and D obtained by the above method for some cancer sites in the département of Isère in France [21].

Several patterns emerge from this analysis : in men, testicular cancer has a distribution with a significantly positive autocorrelation, while the homogeneity test detects no difference. This finding is noteworthy, given that this cancer is of such low incidence that the homogeneity test has in any case little power. Autocorrelation

Table 3.1 Autocorrelation of risks for selected cancer sites in the département of Isère in France

	K	Moran statistic: I ([a])				Smans statistic: D				Homogeneity ([c])
		Observed	Expected ([b])	Standard error	Z	Observed	Expected ([b])	Standard error	Z	
Males										
Testis	87	30.72	8.17	9.26	2.44	11.75	14.10	0.93	–2.53	42.1
Brain	164	12.34	1.37	9.58	1.15	13.44	14.88	0.96	–1.50	70.9
Kidney	217	–6.53	–1.08	9.33	–0.58	14.89	15.07	0.96	–0.18	43.4
Mouth	431	–6.76	–1.86	9.39	–0.52	13.54	14.89	0.93	–1.45	60.1
Colon-rectum	1 081	14.47	–1.81	9.28	1.75	13.63	14.98	0.93	–1.45	60.3
Females										
Brain	119	8.45	3.78	9.52	0.49	14.12	14.50	0.95	–0.40	56.1
Kidney	124	8.09	1.77	9.41	0.67	12.93	14.83	0.96	–1.98	49.9
Colon-rectum	985	0.30	–1.70	9.37	0.21	14.30	14.94	0.93	–0.69	97.6
Breast	2 208	39.44	–1.76	9.52	3.96	11.60	14.91	0.94	–3.52	100.1

([a]) I is the autocorrelation of the logarithm of the SMRs, multiplied by 100

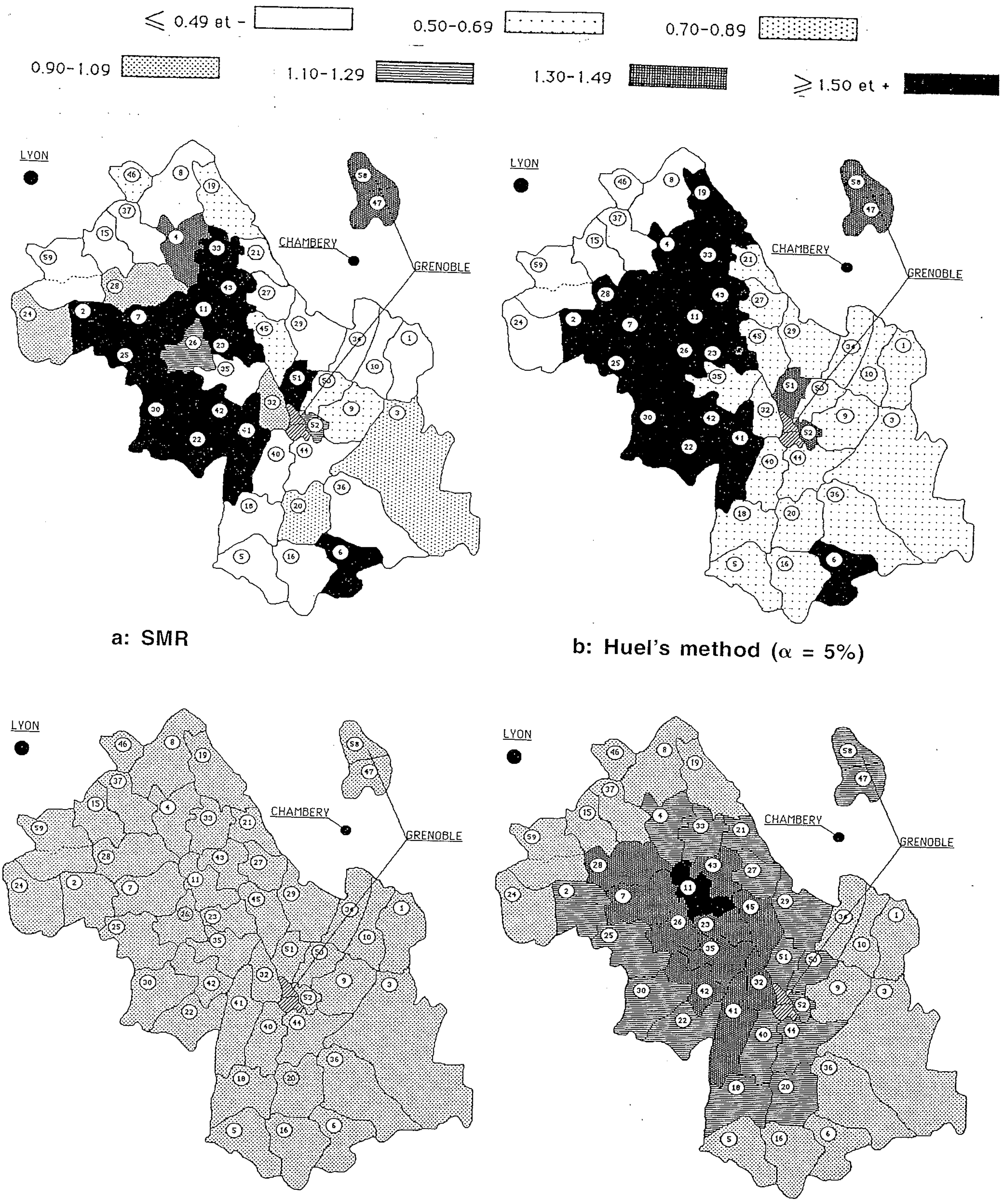
([b]) Under the assumption of a uniform risk in the département. Under the assumption of a normal distribution with uniform variance but no autocorrelation, the mean and standard error of I would be –2.27 and 9.12; those of D would be 15.33 and 0.93

([c]) This column gives the value of χ^2 for homogeneity (Chapter 2, page 89); the critical value at the 5% level is 60.5.

indicated by high values for I and D is illustrated in Figure 3.7 showing the geographical variation of testicular cancer incidence in the département of Isère. In males, oral and brain cancers have nonhomogeneous distributions without autocorrelation; the distribution of kidney cancer seems completely random. In females, brain cancer also has a random distribution. The statistic D detects a significant autocorrelation for kidney cancer, suggesting that in this case it is more powerful than I, which detects no autocorrelation. Colorectal cancer has geographical variation without significant autocorrelation while breast cancer shows both heterogeneity and autocorrelation. It is worth noting that the means of I and D can deviate from their theoretical values obtained by formulae (3.4) and (3.22) considerably when the number of cases is small but only slightly for more frequent cancers such as colorectal and breast. At the same time, the variances of I and D remain approximately constant and close to their theoretical values (see table 3.1, note ([b])).

Identifying risk clusters

The preceding sections have shown how to describe and interpret the spatial distribution of incidence or mortality using the basic geographical unit from which the data are usually collected. The aim of this section is to present methods for studying spatial distribution on a finer scale. These methods may require a knowledge of the place of incidence for each case.



a: SMR

b: Huel's method (α = 5%)

c: Empirical Bayes method (without autocorrelation)

d: Empirical Bayes method (with autocorrelation)

Figure 3.7 Testicular cancer incidence in Isère (France), 1979-1984
Source: Colonna [21]

Methods for studying the spatial distribution of biological or economical phenomena have been developed principally in the specific context of ecology and geography. In the medical area, this type of research has been carried out mainly for communicable diseases, with the goals of identifying clusters of infection and describing routes of transmission. The use of these methods in the epidemiology of noncommunicable diseases is relatively new. It is due mainly to the recognition of a geographical component among the risk determinants of these diseases.

The existence of apparently unusual clusters of cases in some regions and the concern caused by such aggregations among the resident populations are at the basis of this type of epidemiological research. Thus, the observation of a cluster of leukaemia cases around the nuclear installation at Sellafield in the UK [22] has led to much controversy and used as a further argument for the creation of a national system for collecting incidence and mortality data from small geographical areas [7]. Many epidemiologists were dissatisfied that the cluster had not been detected by the existing surveillance system and that it was ultimately revealed to the public by the lay press. Although the causes of this increased incidence remain to be established, the resulting research has led to new results, in particular concerning the spatial distribution of leukaemia.

Before describing the methods of analysis, the notion of case aggregation or clustering should be clearly defined. A number of clusters are nothing more than a misinterpretation of the observations, often as a result of confusing random phenomenon with regular or uniform phenomenon. This difficulty arises because of our frequently inaccurate picture of what is taken to be the normal reference situation, against which unusual rates of incidence are judged.

The problem of demonstrating the existence of a cluster often arises in the following circumstances:

- a geographical region exists in which disease incidence is *a priori* homogeneous over all areas within it.
- the disease is of unknown etiology and rare in each unit of the geographical region.
- the number of units in the region is sufficiently large to allow the geographical distribution of the disease to be studied.

A cluster is thus made up of one or more adjacent units in which the number of cases observed is inconsistent with the possibility of an homogeneous risk in the region under study, that is, of a random distribution of cases in all units of the region. Thus testicular cancer incidence in Isère [21], discussed in the preceding section and on page 140 clusters around canton 11, as demonstrated by bayesian methods given in this section (Figure 3.7 d).

It is necessary to distinguish the situation in which data are collected to test the possible excess of cases around the source of exposure that is, the hypothesis is proposed *before observing* the data, from that in which the hypothesis about the origin of the observed increase in risk is formulated *after making the observations*. In the latter situation, study of the distribution of cases in the whole geographical region can provide the basis for confirming or denying the unusual nature of the observation. This approach, which inevitably leads to statistical tests on a large number of degrees of freedom, is extremely conservative. On the other hand, tests on one degree of freedom, based on a more specific alternative, are not acceptable as they are designed for the situation in which the hypothesis precedes the observation (for example, variation of risk with distance from a source of exposure). Thus, specific methods are required.

The distinction between the two situations cited above is not always obvious. For example, the existence of a nuclear reactor or a toxic waste outlet in the vicinity of a leukaemia cluster might not always provide an *a priori* hypothesis. A systematic study of suspect environmental situations and the bias caused by the selective publication of significant results can lead to confusing interpretations. In the following section we provide a brief survey of the principal methods used to examine such clustering.

When the hypothesis precedes the observation, we would generally attempt to verify that risk increases with proximity to the source of the exposure. A trend test, in which the weights are the distances between the study areas and the source of exposure, can be used for this purpose [23]. The test's power nevertheless depends on the way in which risk decreases with distance and on the distribution of population density according to distance. Schulmann and coworkers [24] have suggested transforming the distances in such a way that the population density remains constant while still maintaining the topological structure of the area (sometimes known as isodemographical maps). There has been little research on the influence of the choice of proximity measurement on the power of corresponding tests.

Stone has proposed a method which is largely independent of the relationship between risk and distance [25]. Although it could be presented in a rather theoretical framework (estimation of risk under the constraint that it decreases with increasing distance), the method is based on a fairly intuitive principle; the essential idea is to construct a sequence of areas of increasing size around a source of exposure using available incidence or mortality data, then to choose the area for which the ratio between observed and expected cases or deaths is highest. In other words, the SMR is evaluated for that area for which the effect is maximum. The statistic thus accumulates the information available to test the assumption of homogeneity of the risk. This function of the observations no longer follows a Poisson distribution, given the way in which the area on which it is based was selected. Stone has shown how the level of significance of the test can be calculated exactly. In practice, it is often simpler to proceed by simulating the multinomial distribution of the number of cases observed in the constructed sequence of non-overlapping areas, conditional on the total number of cases observed in the region under study.

When several identical sources of exposure can be studied, the fact of living close to one of these sources can be considered a potential risk factor and the statistical significance of its effect can be evaluated in a geographical analysis. For example, the risk of leukaemia in small geographical areas as a function of the proportion of people living near a nuclear installation has been studied using a log-linear model [26] or more traditional approaches based on the SMR [27]. These methods can nevertheless suffer from methodological weaknesses inherent in ecological studies (see page 148). Kinlen [28,29] has shown that other conditions, such as living in a 'new town' created in the middle of a rural area, can also be linked to a high risk of childhood leukaemia. This factor could be confounded with the proximity of nuclear installations in the evaluation of leukaemia risk; its effect might be difficult to separate from the potential effect of radiation in an ecological study.

The problem of spatial aggregation of leukaemias and lymphomas has often been raised. In particular, an attempt has been made to evaluate the hypothesis that these diseases have a viral etiology, by assessing whether the spatial distribution of cases is random or is a cluster distribution. In this approach it is implicitly accepted that the viral hypothesis automatically leads to clustering; this latter inference can be questioned today in the light of recent findings on viral mechanisms and of the existence of a long latency period between infection and disease. Irrespective of any specific hypotheses, however, such studies are of value: beyond the test of randomness of the spatial distribution, it is of interest to identify clusters of disease, which can lead to further investigation in the geographical areas thus identified. A better understanding of the aggregative structure of the spatial distribution of a disease results in a more objective analysis of any supposed excess in risk.

The methods proposed rely on the study either of the distribution of cases in small geographical areas defined *a priori* or of the distribution of distances between cases observed over the whole geographical area under consideration. Generally speaking, the studies of homogeneity in risk are based on geographical areas with small populations and limited numbers of cases. Usually, about half the areas do not contain a single case. The test of homogeneity described in Chapter 2 (see page 87) is clearly inappropriate. An acceptable test should be able to detect deviations from randomness, which could either result from the preferential occurrence of excess cases in geographical units where there were already subjects with the disease, or be the consequence of small excess risk in several areas, the overall distribution of risk having however a small variance. In this second situation, few excess cases would be found in each unit, but cases in excess of the expected number would be found in the units where risks were higher. These alternatives to randomness are known as contagious distributions; the second differs however from the strict concept of contagion for which it is the presence of a subject with the disease which increases the probability of healthy subjects developing the disease. A powerful test against the alternative of heterogeneous risks with a small variance distributed around a common value has been proposed by Potthoff and Whittinghill [30,31] and used in the above context by Muirhead and Ball [32].

Recall that heterogeneity is demonstrated when the g multinomial distributions corresponding to g age groups (or more generally to g risk categories)

$$\left(k_{x.}, p_{xi} \, ; \, p_{xi} = \frac{\rho_i \, m_{xi}}{\sum\limits_i \rho_i \, m_{xi}} , 1 \leq i \leq n \right) \qquad 1 \leq x \leq g \qquad (3.6)$$

are not compatible with the hypothesis $\rho_i = 1,\ 1 \le i \le n$. Potthoff and Whittinghill have shown that for such a distribution a powerful test against the alternative of small variation of ρ around 1 is based on the statistic [30]

$$U_x = \left(\sum_{i=1}^{n} \frac{k_{xi}(k_{xi} - 1)}{p^*_{xi}} \right) \tag{3.7}$$

where p^*_{xi} is the value specified by the null hypothesis $\rho_i = 1,\ 1 \le i \le n$

$$p^*_{xi} = \frac{m_{xi}}{\sum_i m_{xi}}$$

Thus, U_x in this situation becomes:

$$U_x = \left(\sum_{i=1}^{n} m_{xi} \right) \left(\sum_{i=1}^{n} \frac{k_{xi}(k_{xi} - 1)}{m_{xi}} \right) \tag{3.8}$$

Note that U_x is based on the number of pairs of cases observed in different units, weighted by the inverse of the number of person-years accumulated by the corresponding population. This weighting has an intuitive explanation, the occurrence of a pair of cases being all the more indicative of clustering if the population is small. Note also that units with only 0 or 1 case make no contribution to this statistic. It can be shown that the mean and variance of U_x, under the null hypothesis, are:

$$E(U_x) = k_{x.}(k_{x.} - 1)$$

$$\mathrm{Var}(U_x) = 2(n - 1)E(U_x) \tag{3.9}$$

The test of homogeneity is thus constructed by summing the information from different age groups as has been done several times previously:

$$T_1 = \frac{\sum_{x=1}^{g} [U_x - E(U_x)]}{\sqrt{\sum_{x=1}^{g} [\mathrm{Var}(U_x)]}} \tag{3.10}$$

Table 3.2 shows brain cancer incidence in five cantons of the département of Isère and Potthoff and Whittinghill's test applied to these data. Numbers in parentheses have been observed while those which precede them correspond to a fictitious incidence, constructed to provide an example of a contagious distribution.

With T_1 equal to 3.488, the distribution of cases does not appear to be random, even though the classic test of homogeneity gives the value 1.63 for a χ^2 on four

Table 3.2 Potthoff and Wittinghill's test using data on brain cancer from five cantons of the département of Isère, France ([a])

Age group	Canton					Potthoff and Whittinghill's test				
	1	2	3	4	5	$k_{x.}$	U_x	$E(U_x)$	$Var(U_x)$	T_1 ([b])
1	0 (1) 4 766	0 (0) 4 139	2 (0) 7 876	1 (2) 26 102	0 (0) 11 473	3	13.803	6	48	1.1263
2	0 (0) 8 038	2 (1) 5 345	0 (0) 9 162	0 (1) 28 438	0 (0) 13 864	2	24.265	2	16	5.5661
3	1 (1) 9 934	0 (0) 5 959	0 (0) 10 666	1 (1) 29 424	0 (0) 15 131	2	0.000	2	16	−0.5000
4	0 (1) 9 860	0 (0) 5 908	0 (0) 13 312	2 (1) 34 772	0 (0) 14 004	2	4.478	2	16	0.6195
5	0 (1) 8 688	0 (0) 5 284	0 (0) 19 200	3 (4) 48 835	2 (0) 12 133	5	27.084	20	160	0.5601
6	2 (2) 6 923	0 (0) 5 070	0 (1) 14 641	3 (3) 49 304	2 (1) 12 529	7	50.445	42	336	0.4607
7	0 (0) 7 557	2 (1) 5 639	0 (1) 10 260	2 (2) 39 611	0 (0) 14 392	4	31.384	12	96	1.9783
8	1 (1) 8 245	0 (0) 5 335	0 (0) 8 377	4 (4) 30 233	0 (0) 12 557	5	25.699	20	160	0.4506
9	0 (2) 7 844	1 (1) 4 517	2 (0) 7 456	0 (0) 24 795	0 (0) 9 919	3	14.627	6	48	1.2453
10	0 (0) 7 452	0 (0) 4 285	3 (3) 7 427	2 (2) 24 400	1 (1) 9 590	6	47.298	30	240	1.1166
11	2 (2) 6 557	0 (0) 4 044	0 (0) 6 507	1 (1) 25 420	2 (2) 9 364	5	26.911	20	160	0.5464
12	2 (2) 5 746	0 (1) 3 445	0 (0) 5 287	3 (3) 24 103	2 (1) 8 746	7	39.077	42	336	−0.1595
13	0 (3) 4 250	0 (0) 2 352	0 (0) 3 510	5 (5) 18 189	3 (0) 6 272	8	71.089	56	448	0.7129
14	0 (1) 2 880	0 (0) 1 498	2 (0) 2 384	2 (3) 14 341	0 (0) 4 566	4	25.114	12	96	1.3385
15	2 (2) 2 626	0 (0) 1 304	0 (0) 1 931	1 (1) 13 502	0 (0) 4 523	3	18.192	6	48	1.7597
16	0 (1) 1 867	0 (0) 805	2 (0) 1 269	0 (1) 9 094	0 (0) 3 226	2	25.628	2	16	5.9070
17	0 (0) 1 153	0 (0) 442	0 (0) 743	0 (0) 5 319	0 (0) 1 892	0	0.000	0	0	0.0000
18	0 (0) 643	0 (0) 233	0 (0) 300	0 (0) 2 886	0 (0) 892	0	0.000	0	0	0.0000
Total	10 (20) 105 029	5 (4) 65 604	11 (5) 130 308	30 (34) 448 768	12 (5) 165 073	68	445.1	280	2 240	3.488

([a]) Fictitious incidence corresponding to a contagious distribution. The number of cases which were actually observed are in brackets; the second line gives the person-years of observation.

([b]) This column gives the T_1 test for each age group separately. The total value of T_1 is calculated from formula (3.10).

degrees of freedom (p = 0.80). On the other hand, this test statistic is equal to 25.1 when applied to observed data (p = 0.00005). The Potthoff and Whittinghill test on these same data gives a value of T_1 equal to 0.55. As the value is not significant it shows that this test is not powerful enough to detect certain types of heterogeneity. It is important to realize that T_1 is a powerful test only against the alternative of risk dispersion discussed at the beginning of this section. The statistic is constructed to detect a trend towards contagion and cannot detect even substantial heterogeneity in the absence of aggregation of this type. In particular, it can be shown that the application of this method to testicular cancer data does not provide a significant result ($T_1 = 1.405$; $p = 0.16$) despite clustering of cases (Table 3.1; Figure 3.7d).

The test does not take into consideration the spatial structure of the geographical units being analysed. It is therefore not constructed specifically for geographical analyses. Muirhead and Butland [33] have suggested that the test be applied to several different levels of geographical grouping in order to define the scale on which the phenomenon of aggregation occurs.

A related approach to that described above consists of regrouping the geographical units in such a way that the expected numbers based on a homogeneous distribution of risk are identical in the newly-formed groups. The problem of heterogeneity of populations is thus removed. The randomness of the distribution of the number of cases can then be tested simply by verifying that it follows a Poisson distribution. Such a test based on the same principle as above [31] is given by a statistic known in plant ecology as the dispersion index, defined as the ratio of the observed variance to the observed mean [34]. When ϕ is calculated over n units and n is large, $(n - 1)\,\phi$ is approximately distributed as a χ^2 on $(n - 1)$ degrees of freedom. Then :

$$T_2 = \sqrt{2\,(n-1)\,\phi} - \sqrt{2\,(n-1)-1} \qquad (3.11)$$

can be considered to be a standard normal random variable. When several risk groups are to be distinguished (for example, age groups), stratification can be used, as before.

Urquardt and coworkers [35] developed this approach further, including an algorithm to group units. This procedure takes into account the variations in population density, to construct study units which lead back to the simple case of the Poisson distribution. This idea has also been used in the dual approach, which involves working with distances between cases. If the population density is uniform in the geographical area under study, the distribution of the number of cases in each unit of area would be Poisson, with the mean given by the product of the surface area and the average number of cases per unit surface area. Thus, the number of cases in a circle with a radius r would follow a Poisson distribution with mean $\lambda\pi r^2$. The probability that the distance from a given point to the closest case was less than r would be equal to the probability that the corresponding circle only contained one case, that is $e^{-\lambda\pi r^2}$. In other words, the square of the distance from a given point to the nearest case has an exponential distribution with parameter $\pi\lambda$. More generally, when distances are ranked, if R_j is the distance from a given point to the jth nearest case (neighbour of order j), it can be shown by using the same principle (see Chapter 2, page the relationship between χ^2 and Poisson distributions) that $2\pi\lambda\, R_j^2$ has a χ^2 distribution on 2j degrees of freedom. Thus, study of the distribution of distances between neighbouring cases (from the first or jth order) provides a means of evaluating the randomness of a spatial distribution. Unless distances are transformed appropriately, population density cannot be considered constant and the distance to the jth case will not have the stated property. Nevertheless, the method still provides a useful statistic to define a test of randomness or-to characterize the geographical units which are at excess risk and which need further investigation.

Cuzick and Edwards [36] have proposed to comparing the proximity of n_0 cases to that of n_1 controls representing of the population residing in the region under study. For example, for a childhood disease, these controls could be births of the same sex preceding and following the case in the regional birth register. A test of spatial aggregation is constructed by determining the pairs of subjects (cases and controls) which are neighbours of order j, and counting among these pairs those in which both members are cases. An excess of such pairs compared to the expected number under the assumption of no aggregation (i.e., if the labels 'case' or 'control' are randomly distributed among the $n_0 + n_1$ subjects) will indicate spatial aggregation of cases. The statistic is then defined by:

$$T_k = \sum_{i \neq j} X_{ij} Y_{ij}$$

where $X_{ij} = 1$ if j is the label of a kth order neighbour of i, and 0 otherwise, and $Y_{ij} = 1$ if i and j are cases, and 0 otherwise.

Cuzick and Edwards also suggest other statistics to analyse the structure of distances in the group of cases and controls. They describe the distribution of these statistics under the null hypothesis of no spatial aggregation, and analyse their power to detect certain types of spatial aggregation. The controls in this approach are used to evaluate the density of people at risk in the area under consideration. A similar approach would be possible if this density was known from other sources : the expected number could then be calculated and it would not be necessary to resort to a sample of controls.

Besag and Newell [37] suggested defining areas of investigation around each case by circles with radius given by the distance to the nearest neighbour of order j. The possibility of a cluster around the case under consideration can then be identified from the evaluation of the population at risk in this circle, and hence the number of expected cases, under the hypothesis of homogeneity of risks. In fact, because of the nature of the available data, the region being examined around a given case is not exactly a circle : it is constructed by successive accumulation of small areas of known population. At each stage, the centre of gravity of the area being added is the closest one to the area added at the previous stage. The procedure stops when j cases are obtained in the resulting region (the initial case being excluded) and the expected number is calculated. A circle around the case under consideration is then drawn on the map each time that the probability of observing j cases in the region is less than a specified probability level (for example α = 5%). The number of expected cases in the region at the level α can obviously be calculated taking into account the presence of several risk classes (e.g., age, sex, urban or rural residence) if the population at risk can be characterized according to the values of these parameters. The method is well suited to detect potential clusters in a region for which the population is known on a small geographical scale. In particular, it can be used to identify clusters when the contagious nature of the distribution has already been demonstrated (this last condition is in fact necessary because if $\alpha = 5\%$, Besag and Newell's test will identify 5% of cases as defining a cluster in a purely random distribution).

In practice, these methods are limited by the imprecise information available on the location of cases, and the necessity of placing them at the centres of gravity of the geographical units being studied. The references cited in the bibliography provide more details in this regard.

Time-space clustering

If differences in demographic structure and the prevalence of risk factors across regions are sufficiently stable over time, the spatial distribution of incidence tends to remain constant. Further, time trends will tend to be identical between geographical units. This baseline situation corresponds to the absence of time-space interaction. One possible disruption to this state of equilibrium is the occurrence of change in risk at a given time in one area of the geographical region being studied. The resulting excess of cases defines *time-space clustering*. In investigation of cancer, for which the latency period between the start of exposure and the onset of disease is usually very long, it is uncertain that the identification of such clusters has led to meaningful epidemiological results. Nevertheless, the statistical methods suggested for this type of data merits a brief review.

One of the first studies in this area was by Knox [20] who examined the distribution in space and time of 96 cases of childhood leukaemia. He assumed that any two cases within a kilometre of each other were spatially close and that any two cases occurring within a month of each other were close in time. He then noted 152 pairs which were close in time and 25 pairs close in space. The observation of five pairs close in both time and space led him to the conclusion that there was time-space interaction. He based his conclusion on an analysis of the 2 × 2 table, classifying the 4560 pairs of subjects (96 × 95/2) into four categories according to their spatial and temporal proximity. Under the assumption of absence of interaction between these two variables, the expected number of subjects close in space and time was estimated as $25 \times 152/4560 = 0.83$. Furthermore, considering that the number of occurrences of such pairs follows a Poisson distribution, he calculated that the probability of observing a value greater than or equal to five was 0.0017, and thus highly improbable under the null hypothesis.

In fact, David and Barton [38] have shown that the mean and variance of the number of pairs belonging simultaneously to two distinct and independent relationships (for example, time and space) can be derived from the number of subjects N and the number of edges a_i and b_i $1 \leq i \leq N$ connecting related subjects in the respective graphs[1] of the two relations S and T which define proximity in space and

[1] A relationship can be represented graphically by a set of points (subjects, $1 \leq i \leq N$), and by a set of segments linking points which are in the relation. The subjects are the vertices of the graph and the segments are its edges. When the relationship is not symmetrical, the segments are replaced by vectors when (i,j) is in the relation and (j,i) is not.

time. Specifically, by characterizing the graph of a relationship by a matrix with elements equal to 1 if the pair (i,j) is in the relationship and 0 when it is not, U, the number of pairs which are in both relationships S and T, can be written in the form:

$$U = \sum_{i \neq j} X_{ij} Y_{ij} \tag{3.12}$$

where X_{ij} and Y_{ij} are the elements of the matrices of the graphs of S and T. Let P be the number of edges of a relationship and Q be the number of pairs of edges of this same relationship, then the number of edges connecting i, the total number of edges and the number of pairs of edges in the relation S can be written in the form:

$$a_i = \sum_{i \neq j} X_{ij} \, ; \quad P_s = \frac{1}{2} \sum_{i=1}^{N} a_i \, ; \quad Q_s = \sum_{i=1}^{N} \frac{a_i(a_i - 1)}{2} \tag{3.13}$$

with similar relationships holding for P_T and Q_T as functions of Y_{ij} through b_i, the number of edges connecting i in the realation T. David and Barton's result can then be written:

$$E(U) = \frac{2P_s P_T}{N(N-1)} \tag{3.14}$$

$$E(U^2) = \frac{2P_s P_T}{N^{(2)}} + \frac{4Q_s Q_T}{N^{(3)}} + \frac{4(P_s^{(2)} - 2Q_s)\,(P_T^{(2)} - 2Q_T)}{N^{(4)}} \tag{3.15}$$

where

$$N^{(k)} = N\,(N-1) \,.....\, (N-k+1)$$

In the example given by Knox, we have $P_s = 25$, and $P_T = 152$; Barton and David calculate Q_S and Q_T to obtain a variance of 0.802, showing that the hypothesis of Poisson variation is acceptable and that consequently Knox's conclusions are correct.

This approach can obviously be applied to situations other than the evaluation of time-space clustering. For example, to test the homogeneity of risk in a series of g families, each of size n_j and including k_j subjects with a genetic defect, we calculate the number of pairs of affected subjects in the same family

$$U = \sum_{j=1}^{g} \frac{k_j\,(k_{j-1})}{2} \tag{3.16}$$

if S denotes the relationship of belonging to the same family and T is the relationship of sharing a genetic defect, then $a_i = n_j - 1$ for all members of family j. If, in addition, K is the total number of cases, then $b_i = K - 1$ when i is a case and $b_i = 0$ for other subjects. Application of formulae (3.13) then gives:

$$P_S = \frac{1}{2}\sum_{j=1}^{g} n_j(n_j - 1) \qquad P_T = \frac{1}{2}K(K-1) \tag{3.17}$$

$$Q_S = \frac{1}{2}\sum_{j=1}^{g} n_j(n_j - 1)(n_j - 2) \qquad Q_T = \frac{1}{2}K(K-1)(K-2) \tag{3.18}$$

which immediately gives the mean and variance of U, from formulae (3.14) and (3.15) above.

The statistic U has been generalized by Mantel [39] by allowing X_{ij} and Y_{ij}, the indicators of proximity in space and time of the pair (i, j), to assume values other than 0 and 1. Furthermore, Mantel's method does not require the relationships S and T to be symmetric, so that it can account for very general situations such as the relationship of proximity discussed in the previous section. Cuzick and Edwards' method [36], presented earlier, is within the scope of this approach. Mantel's result is discussed by Cliff and Ord [40], whose work we will return to in more detail. Below, the method for calculating the moments of the statistic U are given.

First, the quantities S_0, S_1, S_2, defined by the following formulae, are calculated:

$$S_0 = \sum_{i \neq j} X_{ij} \tag{3.19}$$

$$S_1 = \frac{1}{2}\sum_{i \neq j} (X_{ij} + X_{ji})^2 \tag{3.20}$$

$$S_2 = \sum_{i} (X_{i.} + X_{.i})^2 \tag{3.21}$$

The quantities T_0, T_1 and T_2 are defined by similar formulae as functions of Y. Mantel has shown that under the hypothesis of no correlation between X_{ij} and Y_{ij}, the expected value and variance of U are given by:

$$E(U) = \frac{S_0 T_0}{N(N-1)} \tag{3.22}$$

$$E(U^2) = \frac{S_1 T_1}{2N^{(2)}} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4N^{(3)}} + \frac{(S_0^2 + S_1 - S_2)(T_0^2 + T_1 - T_2)}{N^{(4)}} \tag{3.23}$$

where $N^{(k)}$ is defined as in formula (3.15). Smans and Ohno's statistics given in the previous section are of this kind. In particular, the mean and the variance of D is derived from formulae (3.22) and (3.23) in the case of uniform population density.

Smoothing and the empirical Bayes method

Data from small geographical areas can be more informative in the analysis of disease occurrence than those for larger geographical areas, for example by allowing more homogeneous risk groups to be constructed. However, the size of the populations in these areas being studied then implies that most statistics, in particular measures of incidence and mortality, are subject to large random variability that makes the direct interpretation of data difficult. Use of a smoothing procedure then becomes necessary.

Although numerous smoothing methods have been proposed, their statistical properties have been relatively poorly investigated. The polynomial regression method discussed above (Figure 3.5) and the moving average method used in the Finnish cancer atlas [41] and elsewhere do not totally address the problem of inhomogeneity of populations. These methods can therefore produce extreme risk estimates for areas with small populations.

When sufficient information is available, groups of contiguous geographical zones can be formed by objectively defining similarity based on determinants of risk as geographical or socioeconomic variables. The SMR can then be calculated in the resulting areas to obtain more stable estimates, as was done for the atlas of cancer incidence in the Isère [42]. In a related approach, Huel [43] proposed grouping geographical zones based on similarity of incidence or mortality itself. This method assumes an extremely strong autocorrelation since it is based on the idea that contiguous geographical zones are *a priori* alike, in the absence of evidence to the contrary. Contiguous zones are grouped according to the following algorithm :

- Define a *coefficient of similarity* or *distance* between areas which measures their proximity with regard to the variable being considered (e.g., the Mantel-Haenszel statistic comparing incidence or mortality in two neighbouring areas; see page 77).
- Choose a cut-off point in the coefficient beyond which two areas cannot be grouped (e.g., significant difference at level α).
- Group two contiguous areas when their similarity is greater than that between each of the two areas with all other neighbours.

Iteration of step 3 leads to a unique solution if, at each step, all distances between neighbouring areas formed at the previous step are different. In this situation, one area can be grouped with only one of its neighbours.

This method has several advantages. It can eliminate spurious excesses of risk that a simple description using SMRs might produce. It can also reveal the minimal spatial structure compatible with the precision of the observations. On the other hand, the method suffers the inevitable arbitrariness of the choice of the cut-off point for similarity. The variability in the number of neighbours across regions raises another problem: a region with few neighbours probably has a greater chance of remaining isolated and thus attracting attention. This method has been systematically used in the atlas of cancer incidence in the Isère in France and the results appear to confirm this point. Figures 3.7a and 3.7b show the map of testicular cancer incidence based on SMRs and the smoothed map using Huel's method as applied by

Colonna [21]. The method confirms the existence of a spatial structure which had previously been detected by Moran and Smans' coefficients of autocorrelation. Furthermore, it shows a high-risk region. Note that the SMRs do not provide any clear indication of spatial structure because of the small observed numbers available in each geographical unit.

As noted previously, the description of risk in a group of geographical units raises the problem of simultaneous estimation of a series of values for which the available statistical information is of variable precision. Furthermore, a series of comparisons of geographical units taken two at a time does not necessarily lead to a ranking. It is likely that Huel's method provides the best solution that can be obtained using a series of tests of this kind.

The preceding discussion has also shown that the construction of a map implicitly or explicitly involves two steps: firstly, the establishment of a class of geographical units by risk level, then a grouping of these units into large risk categories from which the scale of the map is constructed. If this grouping is carried out on the basis of centiles, two methods to estimate incidence or mortality which rank the geographical units in the same order, or almost the same order, are equivalent and lead to the same graphical representation. As a consequence, the choice between various methods of standardization is not a major problem, since the rank correlation between the resulting measure of risk is usually high. Similarly, the fact that the random variability of the estimators is large compared to that of the underlying risks that they estimate only causes difficulty when the units contain populations of varying sizes: in this situation, estimates of risk in small population units based on small numbers are likely to be misclassified and have an unjustified weight in the final definition of risk categories. In this case, the classification of regions by incidence or mortality level should take into account not only the estimated value of the risk, but also the precision with which risk is estimated.

The empirical Bayes approach is probably the most satisfactory solution which has been proposed to date for this problem. Basically, this method [44] does not allow imprecise estimates to appear among the extreme values simply on the basis of their imprecision.

Suppose the map is defined by n geographical units in which O_i cases have been observed and E_i cases were expected under the hypothesis of equality of risk in different units. Then the relative risk ρ_i of each area compared to the standard risk is classically estimated by the SMR, O_i/E_i (see Chapter 2, page 100). We have seen that O_i can be considered to have a Poisson distribution with mean $\rho_i E_i$. Up to this point, in the classical approach, ρ_i was considered fixed and totally unknown. Now we suppose that the observations are the result of two successive, random mechanisms. The first, determined by the risk factors for the disease, generates the values ρ_i which then become the n realizations of the same underlying random variable determining the risk levels in different regions. The second mechanism leads to observations O_i from the Poisson distribution with mean $\rho_i E_i$. The geographical variability to be described obviously corresponds to the first of these mechanisms. In practice a model is chosen to describe the distribution of the relative risks ρ_i, which relies on available *a priori* information about them such as the prevalence of the risk factors in the regions, or a possible autocorrelation in risks detected by one of the methods discussed above. Several classes of distribution can appear to be reasonable in this context. If the aim is simply to impose some form of cohesion on the estimates and avoid extreme estimates from lightly populated regions, the gamma distribution is an appropriate choice, for reasons explained below. Its density is:

$$\gamma(\rho) = \frac{1}{s\,\Gamma(r/s)} \left(\frac{\rho}{s}\right)^{r/s-1} e^{-\rho/s} \qquad (3.24)$$

where the function Γ is classically defined by the integral

$$\Gamma(x) = \int_0^{\infty} t^{x-1}\, e^{-t}\, dt \qquad (3.25)$$

The mean and the variance of this distribution are r and rs, as can be verified by recalling that $\Gamma(x + 1) = x\Gamma(x)$. Thus r is the mean risk in the group of regions under study and s is a scale factor indicating the size of the geographical variability relative to this mean risk.

If the risks in different regions constitute a sample from this distribution, the probability that k deaths (or cases) are observed in region i is:

$$\Pr(O_i = k) = \int_0^{\infty} \gamma(\rho)\, e^{-\rho E_i} \frac{(\rho E_i)^k}{k!}\, d\rho \qquad (3.26)$$

that is

$$\Pr(O_i = k) = \frac{\Gamma(k + r/s)}{k!\,\Gamma(r/s)} \left(\frac{sE_i}{1 + sE_i}\right)^k \left(\frac{1}{1 + sE_i}\right)^{r/s} \qquad (3.27)$$

which can be written :

$$\Pr(O_i = k) = \frac{(r/s)\,(r/s + 1)\ldots(r/s + k - 1)}{k!} \left(\frac{sE_i}{1 + sE_i}\right)^k \left(\frac{1}{1 + sE_i}\right)^{r/s} \qquad (3.28)$$

Thus, the marginal distribution of O_i is a negative binomial distribution with parameters sE_i and r/s having mean rE_i and variance $rE_i(1 + sE_i)$. This distribution, which serves as a paradigm for cluster distributions, is particularly appropriate here. Effectively, if there is heterogeneity in risks, the distribution of cases in the different geographical units will differ from the random scatter represented by the Poisson distribution, and the cases will tend to group together in higher-risk regions.

Using the distribution of observations in the set of geographical units allows r and s to be estimated by the method of maximum likelihood, thus giving the mean risk and the variance of the distribution of ρ. This marginal distribution is however of limited interest. The main aim of disease mapping in this case is to obtain an estimate of risk in the area i which takes into account both *a priori* information about the distribution of ρ and *a posteriori* information provided by the value k taken by O_i. The *a posteriori* distribution of ρ in this region is used for this purpose, using the fact that the observed value of O_i is k. From Bayes' theorem, the probability density of ρ can be written:

$$\pi(\rho| k) = \frac{\gamma(\rho)\,\Pr(O_i = k| \rho)}{\Pr(O_i = k)} \tag{3.29}$$

$$\pi(\rho \mid k) = \frac{1}{u_i\,\Gamma(k + r/s)}\,(\rho/u_i)^{k+ r/s-1}\,e^{-\rho/u_i} \tag{3.30}$$

where

$$u_i = \frac{s}{1 + s\,E_i}$$

This is the density of a gamma distribution with parameters k + r/s and u_i. The *a posteriori* mean and the variance of ρ are therefore:

$$\hat{\rho}_i = \left(\frac{r}{s} + k\right)\left(\frac{s}{1 + s\,E_i}\right) = \frac{r + ks}{1 + s\,E_i} \tag{3.31}$$

$$\hat{v}_i = \hat{\rho}_i\left(\frac{s}{1 + s\,E_i}\right) \tag{3.32}$$

If r and s are known, these formulae will provide a Bayesian estimate of ρ_i, that is, both a value for ρ_i and the variance of the chosen estimator. In fact, r and s must be estimated from the marginal distribution as indicated above, explaining the use of the term 'empirical' in this method. Similarly, the variability of the estimator cannot be characterized by $\hat{v}_i$, since, the estimation of r and s introduces additional variation which is not taken into account in $\hat{v}_i$.

Replacing k by O_i in (3.31), the estimator $\hat{\rho}_i$ can be written

$$\hat{\rho}_i = \frac{r/s + (O_i/E_i)E_i}{(1/s) + E_i} \tag{3.33}$$

that is, as the weighted average of the mean risk r and of the ratio O_i/E_i, the SMR of the region i. Since s is the parameter characterizing the variance rs of the *a priori* risk distribution, the following observations can be made:

- For a given variance of the geographical distribution, the estimates will be closer to the SMR as E_i increases; however, on the other hand, less precise estimates are moved closer to the mean risk (r).
- If the variance of the geographical distribution (s) is very large, there is effectively no *a priori* information and the empirical Bayes estimates are close to the SMRs.
- When all the SMRs are equally precise, the only effect of their collective estimation will be to reduce the range of the estimates by bringing them all closer to the mean risk.

In spite of its attractive features, the gamma distribution can still be questioned as the only constraint it imposes on the estimates is in their variances. Generally, it would be of interest to incorporate in the *a priori* distribution of ρ additional data concerning the spatial structure under study or a series of covariables characterizing the geographical units being described. These objectives could be achieved by making the parameters r and s of the gamma distribution depend on the spatial structure and covariables. However, because of technical difficulties in incorporating the spatial structure, studies of this type have generally resorted to the arsenal of autoregressive Gaussian spatial processes used in other areas of investigation.

In this context, suppose that the variables X_i = Log ρ_i are Gaussian, with mean depending on a number of covariables (**z**) and correlation depending on the spatial structure (**W**). Besag [45] has shown that such a model can be specified using the conditional expectation of X_i in the form:

$$E\,(X_i \mid X_j\,,\, j \neq i) = \mu_i + a \sum_{j \neq i} w_{ij}\,(X_j - \mu_j) \tag{3.34}$$

$$\mathrm{Var}\,(X_i \mid X_j\,,\, j \neq i) = \sigma^2 \tag{3.35}$$

$$\mu_i = \boldsymbol{\beta}\,\mathbf{z}_i \tag{3.36}$$

where **W,** with elements w_{ij}, is most often the indicator matrix of proximity and $\boldsymbol{\beta}$ is a set of parameters to be estimated. More precisely, this specification is equivalent to a Gaussian model with mean $\boldsymbol{\mu}$ and variancecovariance matrix $\sigma^2(\mathbf{I} - a\mathbf{W})^{-1}$.

This model is especially appropriate for regular geographical subdivisions in which each unit has the same number of neighbours. In practice, this condition is rarely met, and it seems more satisfactory to suppose that the conditional variance of X_i increases as the number of neighbouring regions decreases. A model proposed by Besag and Kempton [47] and examined in detail by Mollie [48] fulfils this objective. This model (mixed model) assumes that the observations result from the sum of two processes : the first T_i is a normal random process with mean μ_i, constant variance σ^2 and without autocorrelation. The second, U_i, which has zero mean and maximal autocorrelation, is obtained by supposing that the conditional expectation of the U_i is the mean of observations U_j in the neighbouring units, and that the conditional variance of U_i is $\tau^2/w_{i.}$, where $w_{i.} = \sum_j w_{ij}$ is the number of neighbours of unit i. The conditional variance of X_i then depends on the number of neighbours, and the autocorrelation of the process $X_i = T_i + U_i$ depends on the relative size of the variances σ^2 and τ^2. The bigger the ratio σ^2/τ^2, the smaller is the spatial autocorrelation, while it is maximized for $\sigma^2 = 0$.

The use of such an *a priori* model for the distribution of risks requires numerical methods, because the marginal and *a posteriori* distributions are no longer expressed in a simple analytical form [44, 46].

In practice, the method gives estimates influenced not only by the mean risk of the region under study but also mean risks in areas neighbouring the unit where the risk is being estimated. This method is especially useful in preventing undue attention being focused on areas with small numbers and randomly raised SMR, when they are surrounded by areas of low risk.

Mollié [48] provides a particularly convincing example of the effectiveness of these methods, using gall bladder cancer mortality in French men. Figure 3.8 shows the SMR for 94 French départements, as well as smoothed estimates produced by the methods described above. The gamma distribution provides little insight into the spatial structure while this structure becomes apparent using models which take into

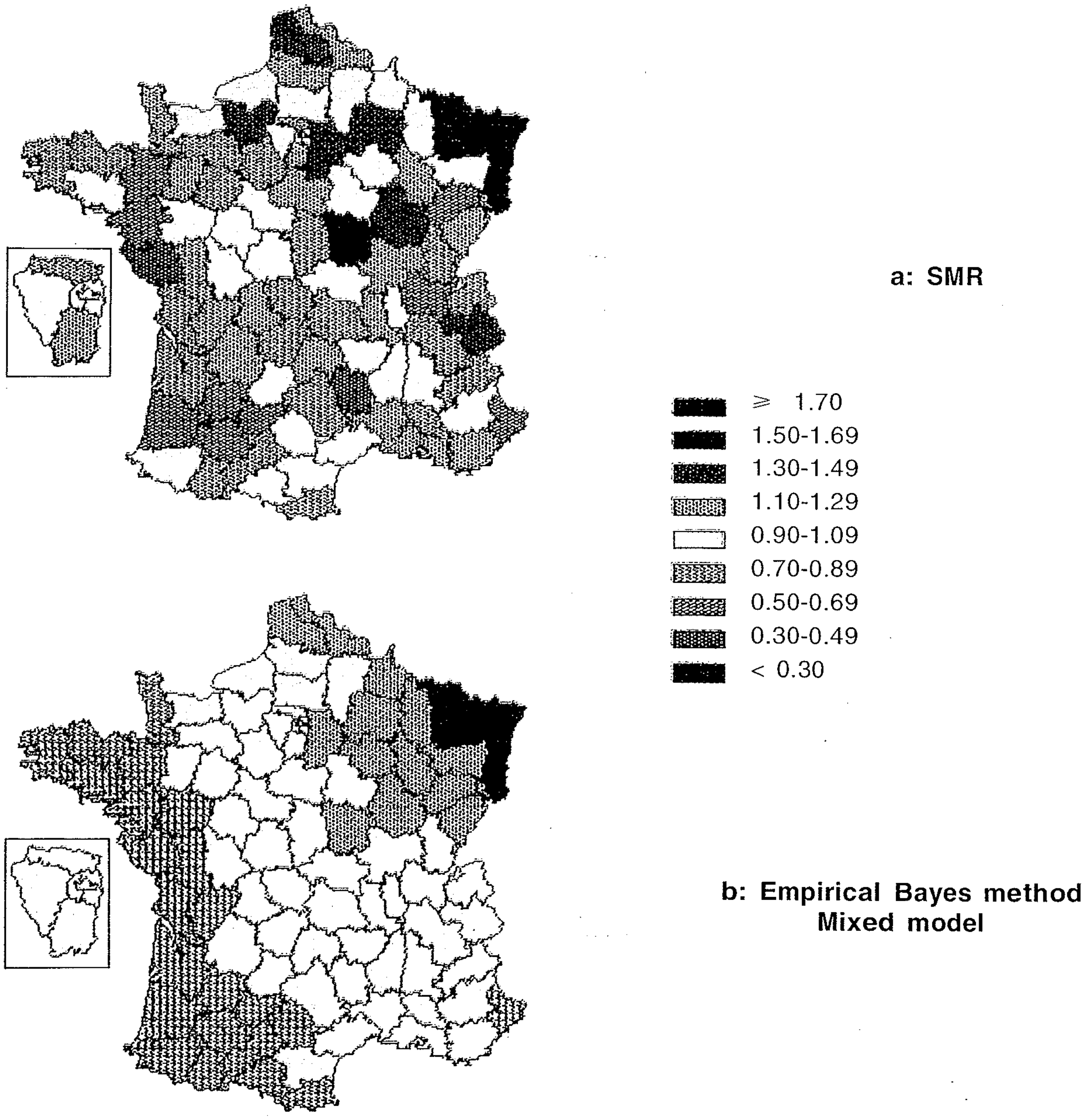


Figure 3.8 Gallbladder cancer mortality in France; men, 1971-1978
Source: Mollié [48]

account the strong autocorrelation of the spatial process. It is worth noting that the indicated gradient continues beyond the regional borders, the mortality rate for this cancer being particularly high in north east Europe.

The example of testicular cancer in the Isère used above [21,42] is also helpful in demonstrating how this method may incorporate the *a priori* information. The SMRs by canton vary from 0 to 628.7 but are considered not to differ from 1 by the homogeneity test (Table 3.1). The choice of the gamma distribution as the *a priori* description of risk is not justified because of the autocorrelation demonstrated previously. Its use gives risk estimates between 96.9 and 101.6 within the département; they are much more compatible with homogeneity of risk than the crude estimates but they completely ignore the local characteristics of the risk process. On the other hand, using the above model (formulae 3.34, 3.35, and 3.36) as *a priori* distribution provides estimates with a strong spatial autocorrelation and suggests the existence of higher-risk areas. This is reflected in the range of estimates (95.4; 162.0) which is larger than that obtained from the *a priori* gamma distribution. The second set of estimates should be preferred because the data are not compatible with an absence of autocorrelation. It is clearly more logical in this case to use an estimation method which takes into account the spatial organization of the geographical units, to allow a better appreciation of the geographical variation in risk.

Concluding remark

We conclude this section on geographical methods with a cautionary remark. The recent rapid development of these methods results more from a preoccupation with the environment than from new biological knowledge generating hypotheses to be examined. Although legitimate, these preoccupations have led to the introduction of some confusion and may well generate substantial report bias. The increase in the number of situations in which excess risk is investigated has tended to invalidate the statistical methods used in this context which are not designed to deal with this multiple test of randomness.

In these situations, epidemiologists can be caught between two extreme positions : either they may accept as having been stated *a priori* a hypothesis which was in reality suggested by the observations; from this point on, the hypothesis will be confirmed simply by a suitably chosen test. Alternatively, they can deny the existence of any excess risk in the particular case presented to them and look in the armoury of available tests for the most conservative one which will simply show that their own *a priori* ideas cannot be disproven by statistics. This ambiguity emphasizes the need to adopt an approach dictated by a biological hypothesis which integrates research from other disciplines. When there are no data of this kind, a good theoretical knowledge of the tools being used is the only support available. With this knowledge, wrong conclusions resulting from excessive confidence in statistical significance alone can be avoided. Thus, for many reasons, the contribution of geographical studies to etiological research is uneven, and depends on the context in which they are applied. Although they are useful, a number of these methods are at best tools of preliminary investigation.

Ecological studies

Aim and methodological principles

Correlation studies, also called *ecological studies*, have fundamentally the same objective as the methods of analytical epidemiology, that is, to detect associations between risk and exposure levels and then suggest, or preferably confirm, explanatory hypotheses. As with all methods in descriptive epidemiology, it is the group rather than the individual which constitutes the basic statistical unit.

Correlation studies are often seen as equivalent to geographical analyses of the determinants of risk. Indeed, the procedure is frequently used with groups that are geographically defined, whether by region or by country. It is a logical development of the studies described in the previous sections, and represents the most straightforward approach to try to explain geographical heterogeneity. Nevertheless, the methods have a much wider use, applying to all situations which involve investigating the relationship between the frequency of an event in several groups and a parameter characterizing the average exposure of individuals in the groups, no matter how the groups are defined.

Ecological studies also represent a natural extension of the pairwise comparisons often made in descriptive epidemiology, in that they provide a synthesis of the information obtained from these comparisons. Their advantage is especially obvious when many factors are presumed to act simultaneously and the average exposure of the group can be determined for each one of them. In this situation, it is not particularly informative to simply examine rates and levels of exposure to different factors. In theory, the specific effects of each factor could be assessed by simultaneously accounting for them in a multivariate analysis. In addition, correlations across groups should offer a further opportunity to confirm the existence of a relationship between exposure and risk if it is possible to demonstrate a dose-response relationship. In the following section, however, it will be seen that ecological studies are subject to a number of weaknesses which limit their value and make their interpretation difficult.

Correlation studies are often justified on the grounds that they use available data on groups which have been formed for other reasons, but nevertheless reflect different levels of the exposure being studied. As with other methods in descriptive epidemiology, ecological studies are based on the implicit assumption that the groups on which the study is based correspond to a categorization of exposure of acceptable specificity. It will be seen later that the homogeneity of exposure within groups is an important determinant of the method's success.

When groups are not defined *a priori*, the way in which they are formed using available data is obviously of crucial importance. In an ideal situation where these data are available at an individual level, groups could be formed by categorizing individuals with respect to increasing, if not homogeneous, exposure levels. The situation arising in this case is then strictly identical to that of an analytical study.

In general, study of the relationship between exposure and risk level is based on a graphical representation, in which each group under consideration appears as a point, situated on two axes characterizing respectively the two measures in question. For example, in Figure 3.9, data on tobacco consumption and the cumulative risk of lung cancer are shown for European countries. An important feature is the shape of the resulting scatter of points : concentration of points around a simple, especially linear, function tends to support the determining role of the exposure in the statistical explanation of risk level.

The effect of exposure can be quantified by fitting a regression line which predicts incidence or mortality as a function of the level of exposure. Later, we will see that this method is more appropriate than the calculation of the correlation coefficient, which is nevertheless the procedure most often used.

Technical aspects of the calculation and interpretation of regression and correlation are presented briefly below.

Figure 3.10a shows the linear function Y = 2X + 1 when X is between 0 to 1; the value of Y depends only on X and its variability is similarly defined by that of X:

$$\mathrm{Var}(Y) = 4\ \mathrm{Var}(X)$$

Figures 3.10b, c and d show how such a relationship is changed when a random component of increasing variance is added to the deterministic element 2X + 1 defining Y. Table 3.3 provides numerical values corresponding to these figures and details of the calculations for Figure 3.10c. In this example X is assumed to be

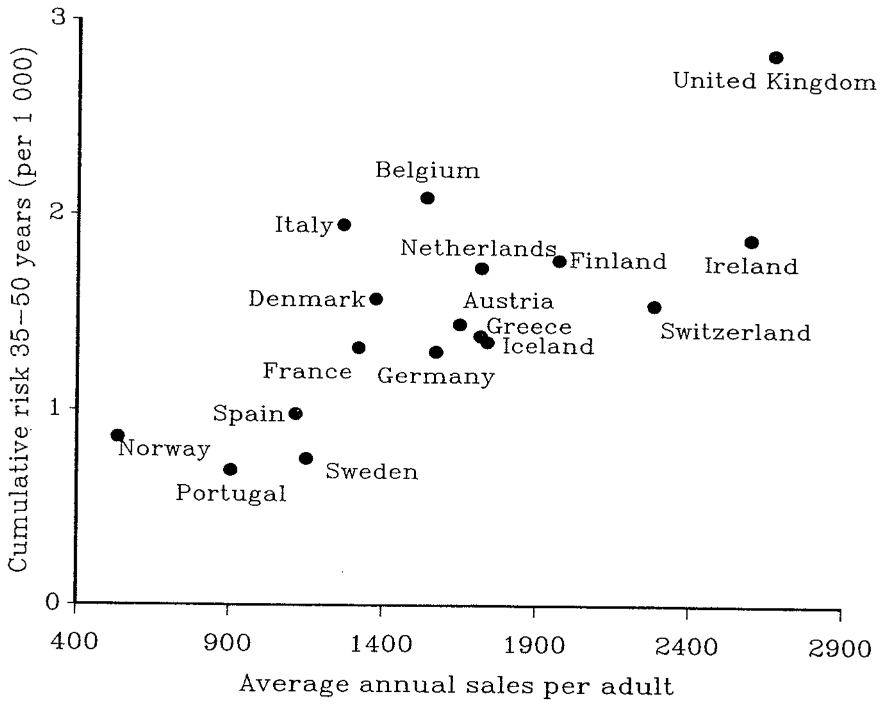
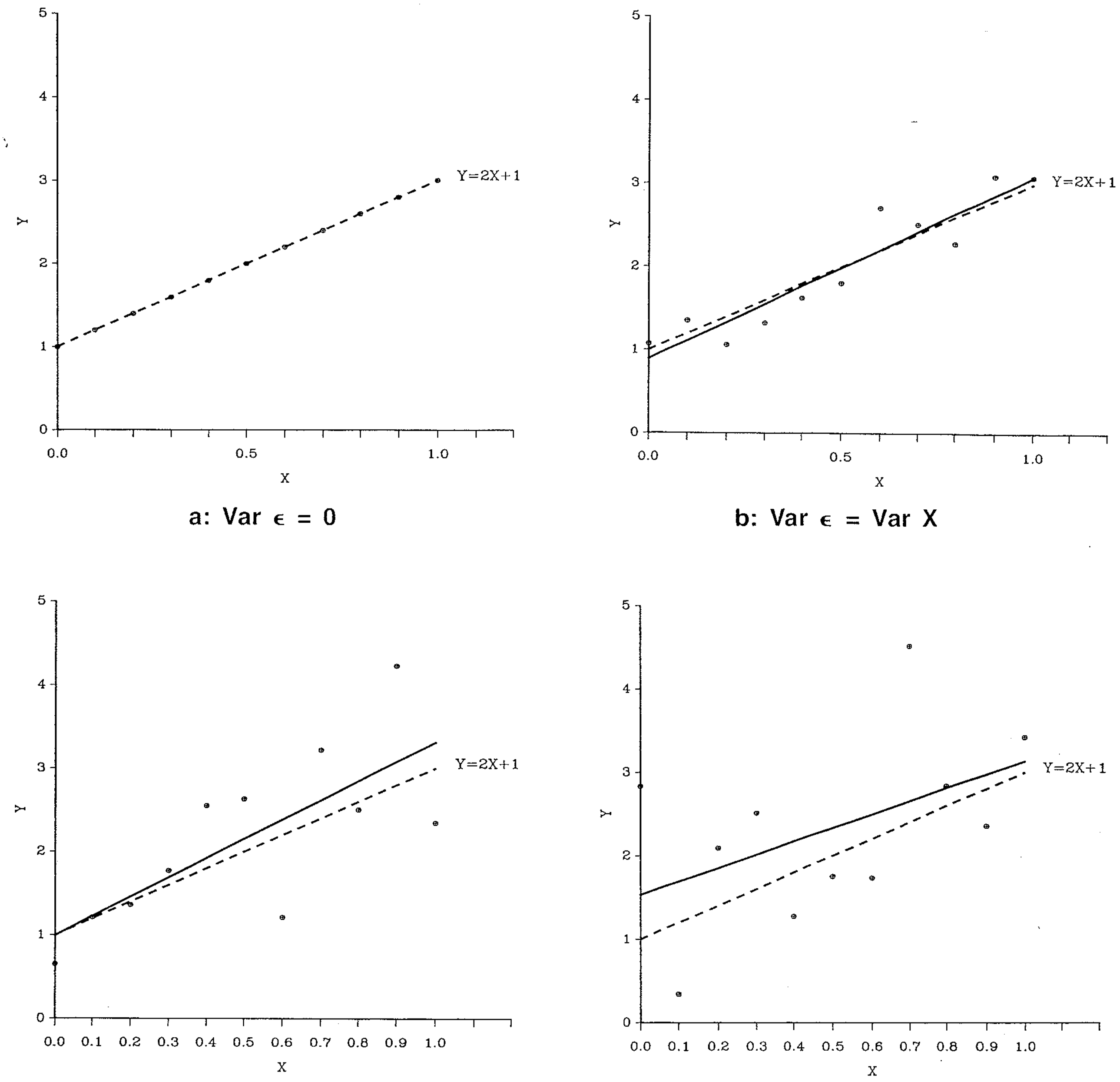


Figure 3.9 National sales of cigarettes (1955-1964) and risk of lung cancer in European countries (average risk in males and females born around 1925; see Table 1.1)



a: Var ε = 0

b: Var ε = Var X

c: Var ε = 4 Var X

d: Var ε = 16 Var X

Figure 3.10 Least squares estimate of the linear relationship $Y = 2X + 1 + \epsilon$, simulated data

controlled, that is, it takes the values x_i, i = 1, n defined *a priori* (here from 0 to 1 by steps of 0.1). The classical model used to represent this type of data is

$$Y = aX + b + \varepsilon$$

where the errors ε are assumed to be independent with the same normal distribution

$$\varepsilon \rightsquigarrow N(0, \sigma_\varepsilon^2) \tag{3.37}$$

It expresses a relationship in which the mean of the normal variable Y depends linearly on the variable X and the conditional variance σ_ε^2 is the same for all values of X. The variability of Y is thus the result of its structural variability due to the relationship with X and the random variability added by the error ε, which may be due to other determinants of Y not accounted for by the model:

$$\mathrm{Var}(Y) = a^2\,\mathrm{Var}(X) + \sigma_\varepsilon^2 \tag{3.38}$$

In Figure 3.10, σ_ε^2 respectively has the value Var(X) (Figure 3.10b), 4Var(X) (Figure 3.10c) and 16Var(X) (Figure 3.10d). In Figure 3.10c, only half of the variance of Y is due to the structural relationship linking X and Y.

The accuracy of the prediction of Y that can be made from knowing X is often measured by the percentage of the variance of Y which is due to its relationship with X. This relationship is written as $\rho^2 = a^2\mathrm{Var}(X)/\mathrm{Var}(Y)$. Its values are respectively 100%, 80%, 50% and 20% in the four diagrammes of figure 3.10 above. This figure show that the accuracy of the prediction, therefore ρ^2, depends on the random variability σ_ε^2. The above formula indicates that it is also a function of the structural variance. The less the slope, the smaller the value of ρ^2, for given random variability and variance of X; ρ^2 is obviously zero when the slope is horizontal, because X no longer provides any information on Y.

In practice, a and b are not known. They can be estimated by the maximum likelihood method which, for the model (3.37), is equivalent to the method of least squares: the estimates $\hat{a}$ and $\hat{b}$ of a and b are the values which minimize the deviance D(a,b), that is, the sum of squares of the differences in the model

$$D(a,b) = \sum_{i=1}^{n} (Y_i - aX_i - b)^2 \tag{3.39}$$

A simple rearrangement shows that

$$\hat{a} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\hat{\mathrm{Cov}}(X,Y)}{\mathrm{Var}(X)} \tag{3.40}$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

where $\overline{X}$ and $\overline{Y}$ are the observed averages of X and Y.

Calculation of the regression line from data shown in Figure 3.10 c is given in Table 3.3.

Table 3.3 Data from Figures 3.10(b,c,d) and calculation of the regression line for Figure 3.10c

Unit	Y^b (1)	Y^c (2)	Y^d (3)	X (4)	X^2 (5)	$(Y^c)^2$ (6)	$\bar{X}Y^c$ (7)	$\hat{Y}^c$ (8)
1	1.07	0.65	2.83	0.0	0.00	0.422	0.000	0.99
2	1.35	1.21	0.34	0.1	0.01	1.464	0.121	1.22
3	1.06	1.36	2.09	0.2	0.04	1.850	0.272	1.46
4	1.32	1.77	2.51	0.3	0.09	3.133	0.531	1.69
5	1.62	2.55	1.27	0.4	0.16	6.502	1.020	1.92
6	1.80	2.63	1.75	0.5	0.25	6.917	1.315	2.15
7	2.71	1.21	1.73	0.6	0.36	1.464	0.726	2.38
8	2.51	3.22	4.51	0.7	0.49	10.368	2.254	2.62
9	2.28	2.50	2.83	0.8	0.64	6.250	2.000	2.85
10	3.09	4.23	2.35	0.9	0.81	17.893	3.807	3.08
11	3.08	2.34	3.42	1.0	1.00	5.476	2.340	3.31
Total	21.89	23.67	25.63	5.5	3.85	61.739	14.386	23.67

The regression of Y^c on X is obtained from columns 5, 6 and 7 of Table 3.3 using the following calculation:

$$\bar{X} = 0.5 \qquad \sum (X_i - \bar{X})^2 = 1.1$$

$$\bar{Y} = 2.15 \qquad \sum (Y_i - \bar{Y})^2 = 10.81$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i Y_i - n\overline{XY}) = 2.55$$

$$\hat{a} = \frac{2.55}{1.10} = 2.32$$

$$\hat{b} = 2.15 - (2.32)(0.5) = 0.99$$

Then, column 8 gives estimated values of Y which define the observed regression line:

$$\hat{Y}_i = \hat{a}X_i + \hat{b}$$

By writing:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - \bar{Y} - \hat{a}(X_i - \bar{X})]^2$$

and by developing the second member of the equation, the relationship analagous to (3.38) is obtained:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \hat{a}^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (3.41)$$

which explains the fact that the observed variance of Y is made up of the variance due to the regression and the residual variance; in the present example, this is written:

$$10.81 = (2.32)^2 \times 1.1 + 4.89$$

The value 4.89 obtained for $D(\hat{a}, \hat{b})$ can be calculated in principle from the formula $\sum_i (Y_i - \hat{Y}_i)^2$. In practice, rounding errors prevent a precise result from being obtained in this way and the value is obtained by subtraction using formula (3.41). The percentage of variance explained by the regression is therefore:

$$\hat{\rho}^2 = \frac{\hat{a}^2 \mathrm{Var}(X)}{\hat{\mathrm{Var}}(Y)} = \frac{5.92}{10.81} = 0.55$$

this equation provides an estimate of the exact value ρ^2 which, in this example, was set *a priori* to 0.50.

The correlation coefficient, classically defined by the formula:

$$\rho = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{\mathrm{Var}\,(X) \cdot \mathrm{Var}\,(Y)}} \tag{3.42}$$

is estimated by:

$$\hat{\rho} = \frac{2.55}{\sqrt{10.81 \times 1.1}} = 0.74$$

which is the square root of the percentage of variance explained, and has the same sign as a.

Variations of $\hat{a}$, $\hat{b}$ around their respective expected values a = 2 and b = 1 are described by a bivariate normal distribution. In particular, the variance of the estimate of the slope can be shown to be

$$\mathrm{Var}\,(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_i (X_i - \overline{X})^2} \tag{3.43}$$

this result, which can be easily obtained from formula (3.40), shows that the estimate of a is more precise when the variance of X is large. In other words, a is estimated more accurately when the range of values of X is wide, as intuition would suggest.

Further, $D\,(\hat{a}, \hat{b})\,/\sigma_\varepsilon^2$ can be shown to follow a χ^2 distribution on n − 2 degrees of freedom, leading to an estimate of σ_ε^2 (which has a value in this example of 4Var(X) = 0.40) equal to

$$\hat{\sigma}_\varepsilon^2 = \frac{D(\hat{a}, \hat{b})}{n-2} = \frac{4.89}{9} = 0.54$$

A 1 − α level confidence interval around â can be constructed as

$$[\hat{a} - t_{\alpha/2}\,(n-2)\,S_{\hat{a}}\,;\ \hat{a} + t_{\alpha/2}\,(n-2)\,S_{\hat{a}}]$$

where

$$S_{\hat{a}} = \frac{\hat{\sigma}_\varepsilon}{\sqrt{\sum (X_i - \bar{X})^2}} = 0.70$$

is the standard error of $\hat{a}$ and $t_{\alpha/2}(n - 2)$ is the value exceeded with probability $\alpha/2$ by a Student t distribution on $n - 2$ degrees of freedom. From data in Figure 3.10c where $n = 11$ and $t_{0.025}(9) = 2.26$, the confidence interval of a is equal to [0.74 ; 3.90].

From data in Figures 3.10 b and d, a calculation not shown here leads to estimates of ρ^2 respectively equal to 0.88 and 0.23, as compared to the true values which are 0.80 and 0.20.

Now suppose that the data in Figure 3.10 c give an incidence or mortality rate Y calculated in n groups characterized by the proportion X of subjects exposed to a risk factor; such a relationship obviously expresses a positive association between risk and exposure. The statistical significance of the increase in this risk is evaluated by testing the hypothesis $a = 0$. This test is simply carried out by calculating:

$$t = \frac{\hat{a}}{S_{\hat{a}}} = \sqrt{(n - 2)}\ \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}} = 3.3 \tag{3.44}$$

which can be compared to the critical value of a Student t distribution on $n - 2 = 9$ degrees of freedom. In this example, the test leads to rejection of the hypothesis $a = 0$. On the other hand, in Figure 3.10 d, although the estimate of a is 1.61, the hypothesis $a = 0$ cannot be rejected as the formula above provides a value of t equal to 1.64 for $\rho^2 = 0.23$. The random component has blurred the structural relationship between X and Y. Here, the confidence interval of the slope [-0.60 ; 3.82] is probably more informative than the probability associated with Student's t test ($p = 0.14$), which reveals nothing about the power of the test carried out and *a fortiori* about the precision of the estimate obtained.

This model is nevertheless not really suitable for describing random fluctuations of incidence or mortality, which are a function of the number of expected cases. It may seem preferable to suppose that k_i, the number of cases (or deaths) observed in each group, follows a Poisson distribution with mean $M_i(aX_i + b)$, where M_i is the corresponding number of person-years, and to account for the heterogeneity in the variances implied by this distribution, if the sizes of the groups being studied are very different. This would be particularly relevant if the relationship were log-linear instead of linear; calculation of the regression line could then be modified by taking the predicted variable as $Y = \mathrm{Log}(k/M)$ and by supposing that the error variance is proportional to that predicted by the Poisson distribution. This leads to a *weighted regression* in which the function D(a,b) becomes:

$$D(a, b) = \sum_{i=1}^{n} w_i\,[Y_i - (aX_i + b)]^2 \tag{3.45}$$

where the weights w_i are proportional to the information provided by each observation, or, as a first approximation,proportional to the number of cases observed. In fact, it is unfortunately unlikely that the random component of the number of cases is limited to its Poisson part: other factors which have not been taken into account possibly play a more important role. Consequently the difference between $Y = \text{Log}(K/M)$ and $aX + b$ is the sum of a first component due to the random variation of K/M and a second attributable to geographical variation in risk associated with factors other than X. In practice, this second error, sometimes called *extra-Poisson variation*, renders the suggested weight insufficient and its advantages debatable (see page 182 for a discussion of this problem in the context of time trends). Furthermore, in practice, Y is often the logarithm of a directly standardized rate; its random variability can then no longer be of the Poisson type.

Examples of the use of this method will be given later. Firstly, we turn attention to a specific problem raised by the inherent nature of ecological studies.

Strengths and limitations of a measure of group exposure

Group versus individual exposure

The effect of errors in the measurement of exposure on the risk estimates has been largely studied in the context of analytical studies. It has been shown that these errors lead systematically to underestimation of risk when they are nondifferential, ie, independent of the status – case or non-case – of the individuals being studied. The problem is just as common, but rarely discussed in the context of ecological studies. In this situation, exposure is most often estimated from data collected for other reasons, which generally provide only an indirect measure of possible risk factors. For example, sales of a given product only partially reflect its consumption, because losses and unregistered imports are not taken into account. Furthermore, exposure is only characterized by a single value for the whole group, leading to more or less serious consequences depending on the type of exposure being considered.

When the exposure is collective by definition, it is often reasonable to assume that this single collective value is a good measure of individual exposure for all members of the group. Thus, in the study already cited of the association between water hardness and the incidence of cardiovascular disorders, there is little doubt that the quality of the local water is a good indicator of individual exposure for the residents of the district. A similar situation would apply in a study of the effects of sun exposure or natural radiation. A descriptive study in this case is conceptually the same as an analytical study. In the examples given, research carried out on individuals would rely on exactly the same data.

Most often, exposure is individual in nature and rarely homogeneous within a group, either because all members are exposed but at very different levels, or because exposure is either present or absent and the exposure of the group therefore amounts to the proportion of exposed individuals. In practice, distribution of exposure can often have both characteristics. Cigarette consumption represents an example of a heterogeneous distribution of individual exposure, but the heterogeneity may be even more marked as, for example, in the study of occupational risks. In fact, heterogeneity of exposure is the norm in ecological studies, generally as a result of the fact that groups are defined using available data, which usually characterize the exposure only indirectly.

Under these conditions, it is not appropriate to assume that all individuals classified as belonging to a given group have actually experienced the same extent of exposure, as is done in an analytical study: as a consequence, when it is stated that group A is defined as more exposed than group B, it is actually known that group B will include subjects more exposed than some subjects in group A and vice versa. Most often in an ecological study, hierarchical classification of groups based on the degree of exposure is thus only valid for the averages.

Intuitively, the quality of the information that can be derived from an ecological study based on group measurements depends on the relative magnitude of the variability of exposure within groups with respect to its variability between groups. For example, it is doubtful whether a correlation study of the relationship between meat consumption and colon cancer, conducted in districts of the same country, could provide an interpretable result because variations in average consumption between districts would probably be too small in comparison to individual differences within districts.

On the other hand, the more the groups formed for the study can provide a representative classification of individual exposure, the more one is tempted not only to establish the existence of a relationship between exposure and risk, but also to quantify the relationship.

Risk estimation in the context of an ecological study

Consider the situation in which individual exposure is characterized by a dichotomous variable (exposed/unexposed) and where therefore the exposure in each group is defined by the proportion of exposed subjects.

In contrast to a study based on individual follow-up (cohort study), a correlation study cannot use the distribution of events (whether deaths or incident cases) in exposed and unexposed subjects to calculate risk in the two subgroups and the relative risk of exposure. Nevertheless, it is still possible to estimate the relationship between risk and the factor under study when event data are available for a series of n groups. Table 3.4 presents data for the ith group.

Table 3.4 Distribution ([a]) of events (deaths or incident cases) and person-years in a cohort study and a correlation study

	Exposed	Unexposed	Total
Events	d_{1i}	d_{0i}	$\mathbf{D_i}$
Person-years	$\mathbf{m_{1i}}$	$\mathbf{m_{0i}}$	$\mathbf{M_i}$

[a] Data available from correlation study are in bold type.

If λ_{1i} and λ_{0i} are the unknown rates for the exposed and unexposed, the expected number of events in group i can be written:

$$E(D_i) = m_{0i}\,\lambda_{0i} + m_{1i}\,\lambda_{1i} \tag{3.46}$$

If $p_i = m_{1i}/M_i$ characterizes the proportion of exposed subjects in a group i, the rate in this group is:

$$\mu_i = \lambda_{0i} + p_i\,(\lambda_{1i} - \lambda_{0i}) \tag{3.47}$$

that is, the sum of the baseline risk and the additional risk attributable to exposure for a subset of the group. If risk in the different groups depends entirely on whether or not an individual is exposed, it is independent of other individual characteristics. λ_{0i} and λ_{1i} then do not depend on i and the incidence rate μ_i is a linear function of p_i, the proportion exposed in the group. In fact, if $\delta = \lambda_{1i} - \lambda_{0i}$, we have the model:

$$\mu_i = \mu_0 + \delta p_i \tag{3.48}$$

where δ is independent of i. In other words, if the baseline risk is constant ($\lambda_{0i} = \mu_0$), and if the relative risk ($R = \lambda_{1i} / \lambda_{0i}$) of exposed subjects does not depend on the group, the relationship (3.47) can be written:

$$\frac{\mu_i}{\mu_0} = (R - 1)p_i + 1 \tag{3.49}$$

and thus R can be estimated by:

$$\hat{R} = 1 + \frac{\hat{\delta}}{\hat{\mu}_0} \tag{3.50}$$

This estimate of relative risk is based on the assumption that the expected number of cases in each group depends only on the proportion of exposed and on absolutely no other characteristic of the group. Although this condition is often accepted implicitly, it is not routinely satisfied: hence the limited value in practice of this type of relationship (see the following section). On the other hand, these calculations have a theoretical value in showing that when the assumption is true, the relationship between risk and exposure is linear and the slope of the regression line is the important parameter.

The ecological fallacy

A number of authors have noted that the study of the association between exposure and risk based on grouped data can lead to false conclusions. An example frequently cited in this context [49] is Durkheim's study on suicide rates in four areas of western Europe in the nineteenth century [50]. Durkheim relied on the observation that the suicide rate increased with the proportion of Protestants in a given region to conclude that Protestants committed suicide more often than Catholics. Was this conclusion valid? It may have been that suicide was in fact more frequent among Catholics and increased the more they found themselves in the minority and experienced social pressures predisposing to suicide. This explanation is nevertheless unlikely, because it would require an extremely rapid increase in the suicide rate among Catholics as the proportion of Protestants increased. Indeed, Durkheim ultimately showed that this was not the case. Logically, it was not implausible and reveals one of the major contradictions in the ecological approach: the average level of an exposure factor can have a positive association with the incidence rate in the group, even when the same factor is associated negatively with individual risk within the group. This paradox has many causes. As an example, imagine that the increases in the average income of a group can lead to increased risk behaviour among the poorest of the group. The study of cervical cancer in Finland illustrates this situation (see below, page 157). This intrinsic weakness of correlation studies is known as *the ecological fallacy*.

Secondly, it should be emphasized that the ecological approach is particularly vulnerable to the effects of confounding variables; not only does the approach not allow for control as does a study carried out at an individual level, but it also tends to transform other risk factors into confounding variables, even when they are independent of the factor being studied at an individual level. For example, in an investigation of the relationship between the proportion of wood workers and lung cancer incidence using data from 25 Swiss cantons, smoking will induce confounding if consumption changes with the proportion of wood workers in each canton, even if the two factors in question are independent at the individual level.

To illustrate this point, consider the situation of two dichotomous factors given in Table 3.5. In an ecological study, only the data in bold type are known for each group in the study. If the two factors are independent and there is no interaction (on a multiplicative scale), it is clear that the relative risk for one of them can be estimated from the complete data without taking account of the other.

The marginal estimate of relative risk corresponding to the first factor $(d_{1.}/m_{1.})/(d_{0.}/m_{0.})$ is equal to the estimate obtained after stratifying by the second;

Table 3.5 Distribution ([a]) of events (d) and person-years (m) in the presence of two risk factors

Factor No 1	Factor No 2		Total
	Exposed	Unexposed	
Exposed	$d_{11}/\mathbf{m}_{11}$	$d_{10}/\mathbf{m}_{10}$	$d_{1.}/\mathbf{m}_{1.}$
Unexposed	$d_{01}/\mathbf{m}_{01}$	$d_{00}/\mathbf{m}_{00}$	$d_{0.}/\mathbf{m}_{0.}$
Total	$d_{.1}/\mathbf{m}_{.1}$	$d_{.0}/\mathbf{m}_{.0}$	$\mathbf{D}_i / \mathbf{M}_i$

([a]) Data available from correlation study are in bold type.

since the independence of the exposure factors implies an equal distribution of person-years, we can write:

$$\frac{m_{01}}{m_{11}} = \frac{m_{00}}{m_{10}} = \frac{m_{0.}}{m_{1.}}$$

furthermore, as an absence of interaction between the two factors implies that the relative rates are equal in the groups exposed and unexposed to the second factor, the above formula shows that the relative rate which would be obtained after stratification is also the marginal relative rate.

This property is important for the validity of analytical studies where there are several risk factors under study which are independent within groups, but which have a different distribution from one group to another. Membership in the group can then be taken as a categorical confounding variable. In such studies where data for a given factor are available at an individual level, it is possible to calculate an unbiased estimate of the overall relative risk after adjusting for the group as a factor, in the absence of precise information about any other factors. In an ecological study, where the group itself is the unit of analysis, it is by definition impossible to proceed in this way.

Using the example in Table 3.5, let the relative risks corresponding to the two factors be R_1 and R_2, and the proportions exposed to each factor in group i be p_{1i} and $p_{2i.}$ Then the relationship previously established between baseline risk and risk in the group becomes (see 3.49):

$$\mu_i = \mu_0 [1 + (R_1 - 1)p_{1i} + (R_2 - 1)p_{2i} + (R_1 - 1)(R_2 - 1)p_{1i}\, p_{2i}] \qquad (3.51)$$

This relationship shows not only the need to introduce p_{2i} in the regression equation despite the independence of the two factors at an individual level but also the inadequacy of linear adjustment[2].

Table 3.6 illustrates this situation from fictitious data. Five groups, each comprising 100 000 person-years, are divided according to level of exposure to a factor for which the relative risk is constant and equal to 2 in each group. The regression of the death rate against the proportion exposed leads to estimates:

$$\hat{\mu}_0 = -\, 0.1367 \quad \text{and} \quad \hat{\delta} = 7.56$$

these values are not meaningful, because they provide a negative value for the estimated relative risk $\hat{R}$ (3.50). If the baseline risk is taken to have the value $\mu_0 = 1$, which was used to generate the data for Table 3.6, the relative risk estimated from equation (3.49) is 4.43, a number much greater than its true value of 2.

In reality, the data have been generated assuming that two factors distributed independently in each group act multiplicatively on the risk of death. The proportions

[2] Formula (3.51) is only valid for two independent factors with a multiplicative effect. It can be checked that, in general, the last term of (3.51) is $p_{12i} [R_{12} - (R_1 + R_2 - 1)]$; it is equal to zero only when the effects are additive.

Table 3.6 Correlation study. Example of a possible relationship between mortality rate and percentage of exposed subjects

Group	Deaths		Person-years (thousands)		Relative risk	Rate (per 1000)
	Exposed	Un-exposed	Exposed	Un-exposed		
1	56	112	20	80	2	1.68
2	84	98	30	70	2	1.82
3	220	110	50	50	2	3.30
4	360	120	60	40	2	4.80
5	420	90	70	30	2	5.10
Total	1 140	530	230	270	1.96	3.34

of subjects exposed to the second factor in the five groups were respectively 10%, 10%, 30%, 50%, 50% and the relative risk corresponding to the second factor was 5.

If a linear model with two factors is fitted to the data, by an extension of the procedure used above (see page 158 and formula (3.56) for the method to estimate coefficients), the following relationship is obtained:

$$\mu_i = 0.67 + 1.67\,p_1 + 6.33\,p_2$$

which does not provide correct relative risks. Only fitting p_1, p_2 and p_1p_2 would in principle result in an exact estimation of the coefficients of the relationship (3.51), respectively 1, 1, 4, 4. In fact, models of this type are rarely fitted, either because the factors to be taken into consideration are not known or because the necessary data are not available.

In addition, factors associated with the group which act on the variable of interest are not necessarily dichotomous, but are often defined by a number of categories or are of a quantitative nature. Equation (3.46) can be generalized to account for these situations if the distribution of exposure is known in each group, through a model linking exposure and incidence. In the same way as before,

$$\begin{aligned} E(D_i) &= \int_e m_i(e)\,\lambda_i(e)\,de \\ &= M_i\,\lambda_i(0)\left[1 + \int_e (r_i(e) - 1)\,dp_i\right] \qquad (3.52) \end{aligned}$$

where $M_i = \int_e m_i(e)\,de$ is the total number of person-years of exposure and $dp_i = \frac{1}{M_i} m_i(e)\,de$ characterizes the distribution of exposure e in group i.

If the baseline risk $\lambda_i(0)$ and the relative risk $r_i(e)$ are not dependent on the group, we have, as before:

$$E(D_i) = M_i\,\mu_0\left[1 + \int_e (r(e) - 1)\,dp_i\right] \qquad (3.53)$$

Moreover, if risk is a simple function of exposure (for example, $r(e) = 1 + \alpha e$), the incidence rate in group i can be written as a function of the mean exposure in the group (in the preceding example: $\frac{\mu_i}{\mu_0} = 1 + \alpha\bar{e}_i$). Again, this relationship is only valid in the absence of confounding factors.

In conclusion, caution is required in the interpretation of correlation studies, as a number of risk factors which are known to be independent at the individual level can be associated at the group level. It is only under particular conditions of independence of the factors at a group level, such as when they are equally distributed throughout the groups, that this confounding effect is no longer present. For example, failure to account for sex would produce substantial bias in an analytical study of health in relation to an occupational exposure but would probably be without consequence in a geographical correlation study of the same exposure, because the sex ratio varies little from one population to another.

Despite these critical remarks, ecological studies can play an important role in epidemiological research. Some factors exhibit weak interindividual variation within populations, whereas the populations differ substantially in terms of mean levels of exposure. In this situation, the ecological approach can be very informative if carried out in conjunction with study on individuals. In addition to environmental factors, culturally determined behavioural factors, such as diet or sexual practice, can sometimes lend themselves to group studies with regard to exposure measurement. Ecological studies are not necessarily less accurate than studies of individuals. Some biases due to self-reporting, such as interviewer bias and recall bias, may even be avoided.

A review of the literature in this area shows the wide diversity in the applications of the basic principle. In most situations, the method is justified by the need to control for the effects of potential confounding factors. Some of the techniques used will be described in the following section.

Specific techniques and examples

Definition of groups

An example of the grouping of the subjects is provided by an ecological study of occupational risk of nasal cancer by Gardner and Winter [51]. The population census in England and Wales (carried out by sampling) provided the percentage of the male population employed in different occupations for each of 1366 local ad-

Table 3.7 Number of deaths from nasal cancer in the male population as a function of the percentage of workers employed in the furniture and upholstery industry [51]

Category	Percentage of workers	Number of workers in the industry (a)	Total number of workers (b)	Number of districts	Observed number of deaths	Observed expected radio (c)
1	0.00	0	73 488	75.5	76.5	0.98
2	0.00	0	65 245	130.6	63.5	0.82
3	0.00	0	58 166	323.2	71.0	0.91
4	0.06	47	73 268	51.4	64.1	0.82
5	0.12	85	71 684	44.3	67.9	0.87
6	0.16	111	70 414	43.0	82.0	1.05
7	0.19	134	70 781	34.8	90.8	1.17
8	0.21	147	68 675	48.9	75.9	0.98
9	0.26	177	68 690	48.2	57.9	0.74
10	0.30	191	64 456	39.4	70.6	0.91
11	0.34	228	67 382	64.6	84.5	1.09
12	0.39	265	68 832	54.3	67.0	0.86
13	0.44	305	69 480	47.9	75.6	0.97
14	0.50	336	67 943	50.6	67.0	0.86
15	0.59	392	66 500	51.8	75.3	0.97
16	0.72	474	66 176	28.8	91.4	1.17
17	0.82	557	68 219	54.9	94.6	1.22
18	0.98	680	69 571	45.7	100.2	1.29
19	1.40	963	68 940	63.1	68.2	0.88
20	3.12	2 153	68 958	65.0	111.9	1.44

(a) Furniture and upholstery.
(b) Based on the 1971 census of 10% of the male population aged between 15 and 64 years.
(c) Expected number of deaths in each group is 77.8.

ministrative districts. The authors grouped these geographical units into a small number of areas which would have had the same risk for the cancer under study if age had been the only determinant of the disease. This grouping was carried out using the following procedure for each occupational category for which the risk was to be investigated. First, the districts were ranked according to the percentage of the population employed in the category. The number of expected deaths was then calculated for each district based on national age-specific rates. Finally, the districts were grouped such that each of the newly formed units had the same number of expected deaths from nasal cancer. In order to get this result, the total expected cases in some districts could not be allocated to one unit and had to be divided between two successive units. The observed numbers in these districts were then allocated to the two units in proportion to the expected number of cases. The 20 new units thus formed were then considered to have the same *a priori* risk, with age no longer having a confounding effect in the correlation study.

Having formed the groups, the authors carried out a regression of the observed number of deaths on the percentage exposed in the 20 groups, and tested the significance of the slope. As a result, they showed an association between mortality due to nasal cancer and employment in the furniture and upholstery industry, and the leather industry, which is free of the confounding effect of age.

The classical regression of age-adjusted rates on the proportion of people employed in a given sector of activity may appear *a priori* equivalent and simpler than the above procedure.However,while adjusting for age takes account of the differing proportion of younger people across districts, it does not account for the fact that the proportion of the population employed in the relevant sector of activity is highest in groups with the largest proportion of younger people (see formula 3.51). This method of forming groups thus has specific advantages from the point of view of eliminating the effect of age. In addition, combining groups can in some circumstances eliminate other confounding factors, especially those which have geographical autocorrelation.

The detailed results given by the authors (Table 3.7) illustrate the calculation of relative risk by fitting a regression line of risk against the proportion exposed, as described above. For the furniture and upholstery industry, the mortality rate of group i is defined by the fitted line (using the notation in formula 3.48):

$$\hat{\mu}_i = 0.9133 + 0.1640 p_i$$

where p_i is the proportion of workers in this occupational category. The increase in risk with this proportion is highly significant (χ^2 = 20.02 on one degree of freedom). Note that the authors could have estimated the relative risk by:

$$\hat{R} = 1 + \frac{0.1640}{0.9133} = 1.18$$

This relatively small increase in risk is surprising, especially as it relates to an industry for which the association with nasal cancer has already been established. It is possible that the percentage of workers actually exposed to the carcinogens (such as wood dust and leather dust) represents only a small fraction of the workers employed in this sector; this dilution effect is the most likely explanation for the underestimation of true risk.

The authors of this study propose that the idea of combining groups into homogeneous units could be extended to the situation where control for confounding factors, such as socioeconomic status, is required. They recognize, however, that the combination is much more difficult to achieve, and that true homogeneity of groups cannot be attained. Generalizability of the approach is, in any case, limited by the requirement that data are available for small geographical units.

In some situations, exposure is so poorly characterized by the defined exposure variable that erroneous conclusions can result. The study of breast and cervical cancer incidence in Finnish municipalities as a function of a socioeconomic indicator illustrates this phenomenon. Teppo and coworkers grouped 500 Finnish communes into five categories by percentage of inhabitants in the upper social class. When they examined variations in breast cancer incidence, they found, as expected, an increase in risk with the proportion of women 'exposed' according to the above definition. It is known that women at higher risk of breast cancer are generally from the well-off classes (where risk factors such as lower parity and later marriage are more prevalent). On the other hand, the gradient observed for cervical cancer was in the same direction as that for breast cancer (Figure 3.11), and contrary to the relationship between high incidence and lower socioeconomic classes established in previous studies.

Discussing later the results of this ecological study in the light of an analytical study of the above association, the same authors [52] concluded that the risk factors for cervical cancer are more difficult to identify by the ecological approach than those of breast cancer. Under the assumption that cervical cancer is primarily associated with sexual history, it is possible that the diversity of individual exposure resulting from different sexual behaviour is greater than for breast cancer risk factors like parity and dietary factors even in small geographical units such as municipalities. In other words, the ratio between inter- and intra-municipality variation in exposure to breast cancer risk factors could be greater than the corresponding ratio for cervical cancer. This explanation is, however, only partially satisfactory, and raises questions about the characterization of exposure in the ecological study. In particular, the reduction to two social classes undoubtedly yields a measure of low specificity for exposure to risk factors for cervical cancer, and it is likely that in the group defined as exposed, there is in fact a heterogeneous exposure to the true risk factors for cervical cancer. In addition, this heterogeneity can differ from one municipality to another. Finally, it can be assumed that the population subgroups for which cervical cancer risk is particularly high (marginal groups, prostitutes) are generally more represented in urban municipalities. Given that these municipalities are defined as most exposed on the basis of having a large proportion of residents from the upper social class, an apparently positive relationship between cervical cancer risk and upper social class is the result. In fact, the number of subjects actually exposed to

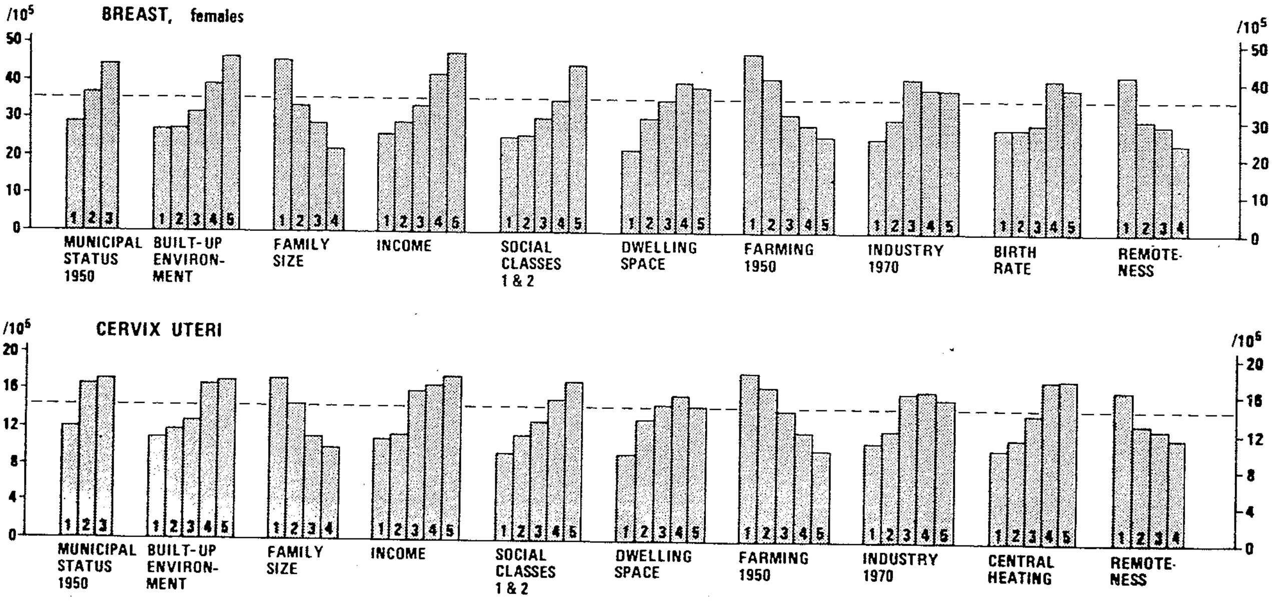


Figure 3.11 Standardized incidence rates of breast and cervical cancer by socioeconomic characteristics in Finnish municipalities, 1955-1974 (Finnish population as standard)
Source: Teppo et al. [6]

risk factors for cervical cancer increases with the proportion of persons in the upper social class, at the same time as the heterogeneity of the group increases.

The exposure indicator used is therefore doubly inadequate: not only does it not define the populations at risk, but it cannot characterize exposure to cervical cancer risk factors. We are faced here with the same type of problems as were discussed above in the context of Durkheim's study of the relationship between suicide and religion. Subjects actually exposed cannot be those identified by the defined exposure criteria. In this situation, a hasty interpretation of observed relative risk will inevitably lead to an ecological fallacy.

Multivariate analysis

When potential confounding factors cannot be controlled for by an appropriate grouping, the necessary adjustment must be carried out in the statistical analyses. The regression method described on page 142 for a single variable can be extended without difficulty to several variables, and appear, *a priori,* to be an appropriate tool for studying the relationship between cancer risk and multiple environmental factors. This method has been used often, mainly in exploratory epidemiological analyses. Its methodological principles will be explained using an example in which the method discussed on page 142 is extended to two variables. The only new concept required when going from one variable to two or more is that of *partial correlation,* which expresses the specific association between a single exposure variable and the risk measure, that is, the association which would be observed if all other factors were held constant.

Firstly, suppose that we wish to estimate the association of Y with two variables X_1 and X_2. As previously, the estimates of a_1 and a_2 in the relationship:

$$Y = a_1X_1 + a_2X_2 + b + \varepsilon \tag{3.54}$$

are obtained by minimizing the deviance $D(a_1,a_2,b)$ corresponding to the sum of the squares of the deviations in the model:

$$D(a_1, a_2, b) = \sum_{i=1}^{n} (Y_i - a_1X_{1i} - a_2X_{2i} - b)^2 \tag{3.55}$$

If Var and Cov are the estimates of variance and covariance, then:

$$D(a_1, a_2, b) = n[\mathrm{Var}(Y) + \mathrm{Var}(a_1X_1 + a_2X_2) - 2\,\mathrm{Cov}(Y, a_1X_1 + a_2X_2)]$$

From this last expression, and setting the derivatives with respect to a_1, a_2 and b equal to zero, it can be verified that $\hat{a}_1$, $\hat{a}_2$ and $\hat{b}$ are given by the equations:

$$\hat{a}_1 \mathrm{Var}(X_1) + \hat{a}_2 \mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(Y, X_1)$$

$$\hat{a}_1 \mathrm{Cov}(X_1, X_2) + \hat{a}_2 \mathrm{Var}(X_2) = \mathrm{Cov}\, Y, X_2)$$

$$\hat{b} = \overline{Y} - \hat{a}_1\overline{X}_1 - \hat{a}_2\overline{X}_2 \tag{3.56}$$

Letting:

$$S_X = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{bmatrix}$$

and

$$S_{YX} = \begin{bmatrix} \mathrm{Cov}(Y, X_1) \\ \mathrm{Cov}\,(Y, X_2 \end{bmatrix}$$

we can write:

$$\hat{\mathbf{a}} = S_X^{-1}\, S_{YX} \tag{3.57}$$

a formula which, when added to the third of the equations (3.56) above forms the analogue of formula (3.40). The minimum value of $D(a_1,a_2,b)$ can similarly be written with the same notation:

$$D\,(\hat{a}_1, \hat{a}_2, \hat{b}) = n\,[\mathrm{Var}\,(Y) + \mathrm{Var}\,(\hat{a}_1X_1 + \hat{a}_2X_2) - 2\,\mathrm{Cov}\,(Y, \hat{a}_1X_1 + \hat{a}_2X_2)]$$

$$D\,(\hat{a}_1, \hat{a}_2,\hat{b}) = n\,[\mathrm{Var}\,(Y) + \mathbf{S}'_{YX}\,\mathbf{S}_X^{-1}\,\mathbf{S}_{YX} - 2\,\mathbf{S}'_{YX}\,\mathbf{S}_X^{-1}\,\mathbf{S}_{YX}]$$

$$D\,(\hat{a}_1, \hat{a}_2, \hat{b}) = n\,[\mathrm{Var}\,(Y) - \mathbf{S}'_{YX}\,\mathbf{S}_X^{-1}\,\mathbf{S}_{YX}]$$

which leads to the relationship:

$$n\,\mathrm{Var}\,(Y) = n\,\mathbf{S}'_{YX}\,\mathbf{S}_X^{-1}\,\mathbf{S}_{YX} + D\,(\hat{a}_1, \hat{a}_2, \hat{b}) = n\,\mathbf{S}'_{YX}\,\hat{\mathbf{a}} + D\,(\hat{a}_1, \hat{a}_2, \hat{b}) \tag{3.58}$$

This formula, analogous to formula (3.41), shows how the total variance can be decomposed into two terms: the variance due to regression and the residual variance. As before, the quantity:

$$\hat{\rho}^2_{YX} = \frac{\mathbf{S}'_{YX}\,\hat{\mathbf{a}}}{\mathrm{Var}\,(Y)} = 1- \frac{D\,(\hat{a}_1, \hat{a}_2, \hat{b})}{n\,\mathrm{Var}\,(Y)}$$

is the percentage of variance explained by the regression. Its positive square root, called the *multiple correlation* between Y and X_1, X_2, is the correlation between Y and the function $\hat{a}_1 X_1 + \hat{a}_2 X_2$; it is equal to the maximum correlation that can be obtained between Y and all functions of the form a1X1 + a_2X_2. Table 3.8 uses data from Figure 3.10c (Table 3.3) to which is added a second predictor X_2 of Y.

Columns 4, 5 and 6 are obtained directly from columns 1, 2 and 3. The data from Table 3.3 combined with these results gives:

$$\sum_{i=1}^{11} (X_{2i} - \bar{X}_2)^2 = 104.3 - \frac{(29.5)^2}{11} = 25.14$$

$$\sum_{i=1}^{11} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 17.05 - \left(\frac{5.5 \times 29.5}{11}\right) = 2.30$$

$$\sum_{i=1}^{11} (Y_i - \bar{Y})(X_{2i} - \bar{X}_2) = 76.08 - \left(\frac{23.67 \times 29.5}{11}\right) = 12.60$$

the value of the other coefficients of equation (3.56) have been obtained previously (see Table 3.3). Ignoring the factor 1/11, the first two equations can be written:

$$1.10\hat{a}_1 + 2.30\hat{a}_2 = 2.55$$

$$2.30\hat{a}_1 + 25.14\hat{a}_2 = 12.60$$

leading to the estimates:

$$\hat{a}_1 = 1.57 \text{ and } \hat{a}_2 = 0.36$$

The third of these equations gives:

$$\hat{b} = 2.15 - 1.57(0.50) - 0.36\,(2.68) = 0.40$$

From (3.58), the component of variation explained by the regression can then be calculated:

$$n\mathbf{S}'_{YX}\,\hat{\mathbf{a}} = (2.55 \times 1.57) + (12.6 \times 0.36) = 8.54$$

and similarly the square of the multiple correlation coefficient:

$$\hat{\rho}^2_{YX} = \frac{8.54}{10.81} = 0.79$$

Table 3.8 An example of the calculation of multiple regression (Y and X_1 are columns 2 and 4 in Table 3.3)

Unit	X_1 (1)	X_2 (2)	Y (3)	X_1X_2 (4)	X_2Y (5)	X_2^2 (6)	$\hat{Y}$ (7)
1	0.0	0.0	0.65	0.00	0.00	0.00	0.40
2	0.1	1.6	1.21	0.16	1.94	2.56	1.14
3	0.2	1.2	1.36	0.24	1.63	1.44	1.13
4	0.3	2.8	1.77	0.84	4.96	7.84	1.88
5	0.4	5.4	2.55	2.16	13.77	29.16	2.97
6	0.5	4.5	2.63	2.25	11.83	20.25	2.80
7	0.6	1.6	1.21	0.96	1.94	2.56	1.92
8	0.7	3.7	3.22	2.59	11.91	13.69	2.83
9	0.8	2.3	2.50	1.84	5.75	5.29	2.48
10	0.9	3.9	4.23	3.51	16.50	15.21	3.22
11	1.0	2.5	2.34	2.50	5.85	6.25	2.87
Total	5.5	29.5	23.67	17.05	76.08	104.25	23.67

In the same way, from (3.58) it can be verified that[3]

$$D(\hat{a}_1, \hat{a}_2, \hat{b}) = 10.81 - 8.54 = 2.27$$

It is then important to be able to evaluate the role of each variable in the prediction of Y and in particular, the improvement in this prediction when the variable X_2 is added to the variable X_1 in the regression equation. The correlation of Y and X_2 (r = 0.76) shows that X_2 is a predictor of Y. However, as X_1 and X_2 are correlated (r = 0.44), it is probable that X_1 and X_2 provide partially the same information about Y. The independent relationship between X_2 and Y should therefore be evaluated.

When X_2 is added to the model equation, the deviance is reduced by 2.62, the difference between the deviance of the model not containing X_2 ($D_1 = D(\hat{a}_1, 0, \hat{b}) = 4.89$) and that of the model above ($D_2 = D(\hat{a}_1, \hat{a}_2, \hat{b}) = 2{,}27$). This reduction expresses the additional role of X_2 after taking X_1 into account. By expressing the reduction in relation to the deviance of the initial model, a measure of the specific contribution of X_2 is obtained:

$$\hat{\rho}^2_{YX_2 | X_1} = \frac{D_1 - D_2}{D_1} = \frac{2.62}{4.89} = 0.53 \tag{3.59}$$

Dividing by $\hat{\text{Var}}(Y)$ shows that:

$$\hat{\rho}^2_{YX_2 | X_1} = \frac{\hat{\rho}^2_{YX} - \hat{\rho}^2_{YX_1}}{1 - \hat{\rho}^2_{YX_1}} = \frac{0.79 - 0.55}{0.45} = 0.53 \tag{3.60}$$

where $\hat{\rho}^2_{YX_1} = 0.55$ is the square of the correlation of Y with X_1.

The square root of $\hat{\rho}^2_{YX_2 | X_1}$ with the same sign as $\hat{a}_2$ is called the *partial correlation* of Y with X_2, holding X_1 constant. Furthermore, it is the correlation between the residuals of the regressions of Y on X_1 and of X_2 on X_1, and is given by the formula:

$$\rho_{YX_2 | X_1} = \frac{\rho_{YX_2} - \rho_{YX_1}\rho_{X_1X_2}}{\sqrt{(1 - \rho^2_{YX_1})(1 - \rho^2_{X_1X_2})}} \tag{3.61}$$

from which we get the estimate

$$\hat{\rho}_{YX_2 | X_1} = \frac{0.76 - 0.74 \times 0.44}{\sqrt{0.45 \times 0.81}} = 0.72 \quad (= \sqrt{0.53} \text{ up to rounding errors})$$

Many authors have used the techniques of multivariate analysis to try to distinguish the roles of multiple factors or to better estimate the effect of a given factor by controlling for confounding effects. Two examples illustrating the use of these methods are given below.

[3] Note that the direct application of the formula $\Sigma(Y_i - \hat{Y})^2$ would lead to the value 2.29.

Blot and Fraumeni [53] estimated the effect of industrial exposure on lung cancer mortality using data from 3056 US counties, attempting to control for sociodemographic factors. Firstly, they estimated the total number of workers in each of 18 industrial sectors in each county, based on the census of industrial employment. For each of the 18 sectors, they then grouped the 3056 counties into three exposure categories: those in which less than 0.1% of the total residential population worked in the sector; those in which between 0.1 and 1% were so employed; and thirdly those with more than 1% of the population employed in the sector.

The estimation of risk associated with the 18 industrial sectors was carried out using a weighted multiple regression model including the exposure variable as well as the main factors to be controlled for. The dependent variable to be fitted was age-standardized lung cancer mortality for the period 1950 to 1969. The exposure variable was included in the model as a categorical variable with three levels defined as above by the proportion of the population working in the sector. The factors to be controlled for were population density, degree of urbanization and proportion of non-whites. A further indicator, situating the county in one of seven broad areas reflecting differences in lung cancer mortality in the USA, was introduced to take large-scale geographical variation into account. The model was thus intended to evaluate the risk associated with industrial activities after adjusting for potential confounding factors. Examination of the residuals, after initially fitting linear terms, led the authors to add quadratic factors to the regression. The weighted regression method was used, with weights defined by the square roots of the number of person-years accumulated in each county during the period under study, giving weights inversely proportional to the standard errors of the mortality rate estimates. The authors did not explain why they chose this type of weighting.

On the basis of the fitted models, the authors concluded that, after accounting for sociodemographic factors, the lung cancer mortality rate increased significantly for four of the 18 industrial sectors: paper, chemicals, petroleum and transport (Table 3.9)

Results of this kind should obviously be interpreted with caution. It is particularly advisable to question the ability of this multivariate analysis to effectively control for the known etiological factors for lung cancer. The authors considered that differ-

Table 3.9 Regression coefficients ([a]) of the standardized rate ([b]) of lung cancer by percentage of workers employed in four manufacturing industries

Industry	Percentage employed in industry	
	0.1-1%	≥ 1%
Paper	0.24 (0.36)	1.02 (0.50)
Chemical	1.49 (0.31)	2.26 (0.49)
Petroleum	0.98 (0.45)	1.32 (1.00)
Transportation	1.22 (0.32)	0.84 (0.46)

([a]) Estimated coefficients (standard error).

([b]) Standardized with respect to the white male population of USA.

ences in tobacco consumption between counties were partially associated with the degree of urbanization, which was accounted for in the model. It should also be noted that the classical approach adopted by Blot and Fraumeni considers each county as a statistically independent unit. It takes large-scale geographical variation into account in a way that differs from Gardner's approach described above. The integration of areas into non-contiguous zones, as in Gardner's method, can, to a certain extent, be thought of as a random assignment of spatially autocorrelated factors. On the other hand, the approach described here can be interpreted as an attempt to adjust the risk for confounding factors using large geographical zones in which they remain approximately constant; thus it indirectly accounts for the correlation in risk which might exist between geographically neighbouring units.

Other approaches which avoid the difficulties of interpretation created by spatial autocorrelation have been described; that proposed by Richardson [54] is described here. First, remember how confounding factors intervene in the equation relating the exposure of interest and the risk of disease in an ecological study.

It has been shown previously (3.53) that the relationship between risk and exposure, under general assumptions, can be written:

$$E(D_i) = M_i \mu_0 (1 + \alpha \bar{e}_i)$$

where $\bar{e}_i$ characterizes the average exposure in group i, M_i and D_i are the numbers of person-years and deaths in the group, and μ_0 is the baseline mortality rate.

If only this exposure plays a role in the determination of risk, the observations D_i would have independent Poisson distributions and estimation of the parameters μ_0 and α would not present any particular difficulty. In practice, other factors confound their effect with that of the exposure under study and should in principle be included in the equation. As they are generally not measured, the equation becomes:

$$E(D_i) = M_i \mu_0 (1 + \alpha \bar{e}_i) + f_i \tag{3.62}$$

where f_i is a random variable which is included as an error term, in the absence of more specific data on the confounding variables. Thus, we are led back to the estimation of a regression equation with correlated errors if, as is generally the case, the unmeasured confounding factors have spatial autocorrelation. If we do not take this correlation into consideration in the analysis, the result will be excessively liberal tests of significance, because the improvement in the deviance will be evaluated with respect to an underestimated error. This phenomenon will be systematic if the Poisson distribution is used as an error model. It will also occur in the situation of positive autocorrelation if the normal approximation for the distribution of incidence or mortality rates is used.

Some authors have proposed regression models with correlated errors [55,56]. However, fitting these models is often unduly complicated in relation to the importance of the results which are expected. In contrast, Richardson's approach is appealing because of its simplicity and the fact that it provides a rapid means of evaluating the significance of an association.

The test of the association is based on the variance σ_r^2 of the empirical correlation coefficient r between incidence (or mortality) and exposure, considered as two spatially autocorrelated processes X and Y. It can be shown that:

$$\sigma_r^2 \approx \frac{\text{Var}(S_{XY})}{E(S_{X^2})\,E(S_{Y^2})} \tag{3.63}$$

where S_{XY}, S_{X^2}, S_{Y^2} are the empirical covariance and variances of the two processes.

In the absence of autocorrelation, $\sigma_r^2 = 1/(N-1)$, where N is the number of observations X_i,Y_i. In the presence of autocorrelation, σ_r^2 is estimated from the observations and used to calculate $N^* = 1 + 1/\hat{\sigma}_r^2$ from which the significance of the correlation is tested with the statistic:

$$T = \frac{r\sqrt{N^* - 2}}{\sqrt{1 - r^2}} \tag{3.64}$$

considered as a Student's variable on $N^* - 2$ degrees of freedom. The method thus proceeds as if the number of autocorrelated observations made were equivalent to a smaller number N^* of independent observations. In the same article the author showed that the method can be extended to any number of variables. If, for example, the significance of the association between X and Y after adjustment for Z is to be evaluated, the correlation of residuals of the regressions of X and Y on Z could be assessed directly by the method.

In practice, S_{X^2} and S_{Y^2} are used to estimate their expected values. The calculation of the variance of S_{XY} requires an additional assumption; by calculating this variance conditional on X, we obtain:

$$\text{Var}(S_{XY}) = \frac{\sum_{i,j} (X_i - \overline{X})(X_j - \overline{X})\,\text{Cov}(Y_iY_j)}{N^2} \tag{3.65}$$

that is, S_{XY}^2 as an estimate of the variance of S_{XY}. When the Y_i are independent, $\text{Cov}(Y_iY_j) = 0$ if $i \neq j$. $\text{Var}(S_{XY})$ has the value $\frac{1}{N^2}\sum (X_i - \overline{X})^2\,\text{Var}(Y)$ and we find that $r\sqrt{N-1}$ is the standard normal variable corresponding to S_{XY}. When the Y_i are not independent, formula (3.65) is only informative under specific assumptions about the structure of the covariance of the Y_i. Accordingly, suppose that $N(N-1)/2$ pairs of geographical units can be stratified into subgroups in which the covariances of the X_i and the Y_i are constant. This grouping is generally based on the distance between the administrative centres of the geographical units being studied, under the assumption that the intensity of the autocorrelation only depends on distance.

The estimate of σ_r^2 is then written, using (3.63), (3.65) and the constancy of the covariances:

$$\hat{\sigma}_r^2 = \frac{\sum_k N_k C_{X_k} C_{Y_k}}{N^2 S_{X^2} S_{Y^2}}$$

where

$$C_{X_k} = \frac{\sum_{i,j} (X_{ik} - \overline{X}_k)(X_{jk} - \overline{X}_k)}{N_k}$$

and

$$C_{Y_k} = \frac{\sum_{i,j} (Y_{ik} - \overline{Y}_k)(Y_{jk} - \overline{Y}_k)}{N_k}$$

are the respective empirical covariances of the X_i and the Y_i in subgroup k and N_k is the number of pairs of units in this subgroup.

Applying these principles to the study of the association between lung cancer and occupational exposure, Richardson [54] showed that the percentage of men employed in the metal industry was correlated with lung cancer mortality across French departments (Table 3.10). The classical test overestimates the intensity of the association but the corrected test is highly significant and remains so even after adjustment for cigarette sales. Since adjustment for a confounding variable partially accounts for autocorrelation of errors, it should be expected that the total corrected

Table 3.10 Correlation between risk of dying from lung cancer ([a]) and employment in selected industries ([b]) in France [54]

	Correlation	Classical test (N = 82)		Corrected test		
	r	t	p	t	p	N*
Metal industry						
Crude	0.63	7.16	10^{-9}	3.00	0.010	16
Ajusted ([c])	0.52	5.46	10^{-6}	3.46	0.002	34
Mining Industry						
Crude	0.33	3.16	0.003	2.37	0.020	47
Adjusted ([c])	0.24	2.26	0.030	2.42	0.020	94
Textile industry						
Crude	0.28	2.57	0.010	1.52	0.140	30
Adjusted ([c])	0.26	2.40	0.020	1.91	0.070	53

([a]) Lung cancer mortality rate (35-74 truncated rate) for 1968-69.
([b]) As measured by percentage of men employed in the industry indicated.
([c]) Adjusted for the sales of cigarettes (number per inhabitant in 1953 ; source : SEITA).

number of observations in the test increases after this adjustment, which is in fact the case. Correlations with the mining and textile industries are weaker and the second is eliminated altogether by the corrected test. Richardson shows that the first of these two associations also disappears after adjustment for a geographical gradient. However, it might be questioned whether such a procedure might have led to overadjustment, and hence the elimination of the real associations, if the variable being studied has a large geographical autocorrelation, and possibly a strong covariation with the variable describing the geographical gradient.

Migrant studies

Migrant studies are based on the idea that immigrants are, by their life style and culture, exposed to risk factors which differ from those prevailing in the host country. Thus evidence for risk levels specific to immigrants can indirectly suggest or confirm etiological hypotheses. In general, the risk to which immigrants are subject is recognized by comparison with the risk in the host country, but it is sometimes compared with the risk in the country of origin.

Immigrants are identified by their nationality when they keep it, or by their place of birth. Some studies are exclusively based on surname. In certain situations, first-generation immigrants (born in the country of origin), can be distinguished from their children, often born in the host country, who are described as second-generation immigrants. This distinction sometimes provides information on the effects of behaviour changes resulting from the cultural integration, which act more profoundly on the second generation.

This technique has been used by Buell and Dunn [57] in their study of Japanese migrants living in California. The incidence of common forms of cancer in first and second-generation migrants was compared with the corresponding rates for California and Japan. The main results, shown in Figure 3.12 have been discussed by Cairns [58]. They show that the risk to which migrants are exposed converges towards the risk in the host country, passing through intermediate risk levels. These findings demonstrate the importance of environmental factors over factors linked to ethnicity. The change in risk is shown to differing degrees for cancers at four sites, the stomach, liver, colon and prostate. Incidence of colon cancer, much rarer in Japan than in the USA, increases markedly for first-generation migrants; the second generation has approximately the same rate as Californians. The transition is much slower for stomach cancer. The risk is extremely high in Japan, and remains much higher for Japanese migrants, even those of the second generation, than for Californians. This phenomenon can obviously be explained by the maintenance of risk behaviour or the failure to adopt protective behaviour, for example, dietary habits. On the other hand, based on these data, the hypothesis of an ethnic susceptibility for stomach cancer cannot be completely excluded.

The principle of migrant studies has been extended to cultural and religious minorities. Cancer risk has been studied among Mormons and Seventh Day Adventists, who are recognized as consuming little or no alcohol, tobacco, coffee or other

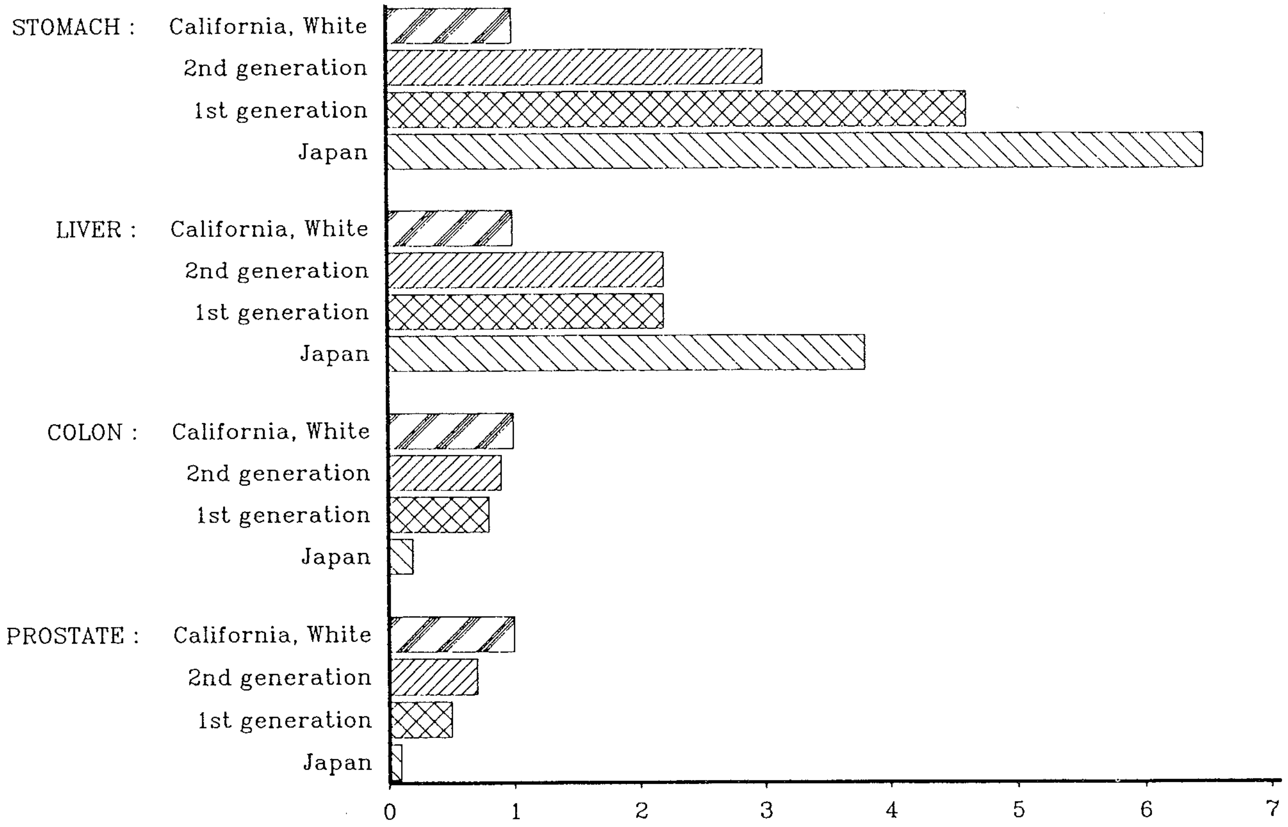


Figure 3.12 Relative risk of death from various cancers for male Japanese migrants to California compared to white Californian males
Source: Buell and Dunn [57]

stimulants. Research of this kind has largely confirmed the importance of lifestyle on cancer risk. For example, it has been shown that cancers of the upper aerodigestive tract were much less frequent in Californian Seventh Day Adventists than in the Californian population as a whole [59].

This type of study has sometimes allowed the effects of closely associated factors to be distinguished. In Chapter 1, it was noted that the apparent effect of urbanization on lung cancer disappeared when the association was studied in Mormons who were living in the same environment but were nonsmokers. The effect originally observed was thus largely due to the fact that smoking is more frequent in urban populations (see page 10).

In terms of methodology, migrant studies can be classified according to whether or not denominators are available. Given the numbers in each group for which risk is to be estimated, the appropriate analysis is the calculation of rates and their comparisons (see Chapter 2, page 85). In practice, the groups being studied are often small and indirect age standardization using the SMR or log-linear modelling based on the Poisson distribution is used. When denominators are not available, study of the relationship between risk and membership in specific groups can be carried out by the PMR method described on page 96 in Chapter 2. As we have seen, it is actually preferable to carry out the analysis using logistic regression identical to that used in case-control studies. A study of Italian migrants in Geneva illustrates this double approach[4].

The Geneva cancer registry has been operating since 1970 and identifies cases by nationality. Swiss nationality is not granted automatically after a certain length of residence, even for foreigners born on Swiss territory. Therefore most immigrants keep their original nationality for one or more decades, as do their descendants. Numbers of foreigners by sex, age and nationality living in Geneva have been estimated regularly since 1976.

Standardized morbidity ratios for immigrants of Italian nationality were first calculated for the main digestive tract cancers over the period during which denominators were available (1976-1987). The calculation was carried out by comparison with the incidence rates established for the total resident population of the Geneva canton. Although this population includes Italians who represented 9% of all residents, the potential diluting effect in the risks was not considered to be large. Table 3.11 shows that significant differences only emerged for gastric cancer, therefore subsequent investigations were restricted to this site alone.

Although the etiology of stomach cancer is not well understood, research has focused mainly on dietary factors. Consumption of salted or smoked food, particularly in places where refrigeration is not widely available, might be a risk factor; fresh fruit and vegetables, on the other hand, could have a protective effect. An often observed increase in risk in lower socioeconomic classes could simply be a marker of dietary practice associated with access to refrigeration. Relatively marked geographical differences have nevertheless been observed between countries, apparently independent of living standards. In Italy, in particular, differences in mortality are substantial from one province to another, with the highest rates observed in regions in the centre of the country. It is also widely accepted, notably from Buell

Table 3.11 Standardized incidence ratio ([a]) for Italians living in Geneva, Switzerland, by sex and site (1976-1987)

		Men		Women	
ICD-9	Site	Relative risk	95% CI	Relative risk	95% CI
150	Oesophagus	0.79	[0.43 ; 1.33]	0.83	[0.10 ; 2.99]
151	Stomach	1.61	[1.18 ; 2.14]	1.81	[0.92 ; 2.33]
153	Colon	0.97	[0.71 ; 1.30]	0.71	[0.45 ; 1.05]
154	Rectum	0.88	[0.56 ; 1.32]	0.94	[0.56 ; 1.49]
155	Liver	1.21	[0.75 ; 1.85]	0.85	[0.18 ; 2.48]
156	Gall-bladder	1.48	[0.60 ; 2.84]	1.12	[0.41 ; 2.55]

([a]) Geneva resident population incidence rates as standard rates.

[4] See Sarti et al. chapter 16 in [60]

and Dunn's study [57] described above, that the period of induction of gastric cancer is particularly long.

In Geneva, several characteristics routinely recorded for each registered case enable the role played by the above factors in gastric cancer carcinogenesis to be studied. They include socioeconomic status, country of birth and duration of residence since migration. Other information relating to the above hypotheses, such as province of birth and spouse's nationality, was obtained from local files of residents by *ad hoc* inquiries. Because it was not feasible to construct denominators for these additional variables, study of their effects could be carried out only by using an analysis of relative frequencies.

This analysis was carried out with 100 cases of stomach cancers occurring between 1970 and 1978 among Italian nationals residing in Geneva and 300 controls drawn randomly from 1161 cancers of other sites registered among Italian nationals during the same period. The number of controls was kept to three per case to minimize the manual investigation of data files. No matching was carried out. Analyses were carried out by unconditional logistic regression (see Chapter 2, page 98).

The evaluation of living standard was based on three socioeconomic categories (manual labourers; clerical workers; management and professional) and from a variable distinguishing five regions of birthplace (southern Italy; central Italy; northern Italy; Switzerland; other), which defined a gradient of socioeconomic status from most socioeconomically deprived to most socioeconomically privileged, that is, from southern Italy to Switzerland. For men, after taking age into account, neither of these variables significantly modified gastric cancer risk; a nonsignificant increase in risk was noted for central Italy. For women, no significant variation in risk was observed with social class, but the risk was significantly higher for women born in central Italy.

The degree of cultural integration was measured by the number of years of residence and by the fact of being married to a Swiss national. No significant association was found from the analysis of these two variables, despite a decreasing trend in risk with duration of residence (both sexes), and with a Swiss spouse (men only).

To investigate differences in risk with place of birth, the 95 Italian provinces were grouped by relative mortality rates, available for the period 1975 to 1977 into three categories: less than 80% of the national average (low); between 80 and 120% of the national average (medium); and more than 120% (high). Separate scales were constructed for both sexes. This breakdown was completed by a fourth class corresponding to cases born in Geneva, where stomach cancer mortality is particularly low, and this category was used as the reference.

This indicator was shown to be highly significantly associated with risk (after accounting for age). For provinces of birth characterized by the highest mortality rates, relative risk was estimated as 4.0 for men and 6.8 for women. The trend of increasing risk across categories was also significant.

In order to judge their effects in the presence of other factors, the variables under study were introduced simultaneously in the same model, with the exception of social class which was assumed to be represented largely by the place of birth. Because no interaction between these factors and sex was significant, an analysis was undertaken with both sexes combined. The results of this analysis (Table 3.12) confirmed the importance of province of birth as a risk factor for gastric cancer. There remained, however, an independent effect of region of birth (highest risk for central and southern Italy), which may reflect a residual role of the birth province, if this factor was too broadly categorized. The dominant role of birth province supports the results of analytical studies, which have demonstrated the local specificity of dietary habits in central Italy, suggesting that they play an important role in the mechanism of gastric cancer [61]. The apparent absence of effect of variables measuring the degree of integration (length of residence and Swiss spouse) is not surprising, considering that gastric cancer has a long latency period.

Table 3.12 Distribution of cases and controls and risk estimates associated with selected factors (both sexes combined) [60]

	Cases (100)	Controls (300)	Relative risk ([a])	p
Level of risk of Italian province of origin				< 0.001
Low ([b])	43	182	1	
Medium	23	76	1.4	
High	34	42	4.3	
Italian region of origin				< 0.05
North ([b])	68	228	1	
Central and South	32	72	2.3	
Length of residence	–	–	1.1	NS
Spouse				NS
Non-Swiss	84	238	1	
Swiss	16	62	0.8	

([a]) Adjusted for age and the other factors in the table.
([b]) Includes those born in Switzerland and elsewhere, except Italy.

Time trends

Objectives

In the context of descriptive epidemiology, there are many reasons for studying time trends. Firstly, information on the historical evolution of risk (incidence or mortality) can generate etiological hypotheses or confirmation of suspected associations between risk factors and disease. While the existence of geographical variation in incidence between populations might be explained by genetic differences, changes in incidence in single populations imply the introduction or disappearance of environmental risk factors much more clearly. Comparison of the development of environmental factors with the development of the frequency of different types of cancer should therefore be profitable. For example, the increase in lung cancer mortality parallels the progressive introduction of cigarette smoking, while its decrease quickly follows a decrease in the proportion of smokers.

However, in etiological research, the interpretation of chronological covariation remains delicate. It would be simple to show that the incidence of melanoma has undergone an increase identical to that of many changes in lifestyle which cannot be incriminated in the etiology of this cancer. Similarly, the general decrease in frequency of stomach cancer could be related to the modification of many environmental factors which accompany higher living standards; its etiology nevertheless remains largely unexplained. The existence of a direct link between the evolution in risk of a given cancer and that of a suspected etiological factor may be less questionable when they both show the same inversions of trend. For example, the parallel trends in incidence of larynx and oesophageal cancers (Figure 1.3) clearly suggests a common etiology, in this case alcohol consumption. Alcohol consumption has in fact declined substantially in the period when the generations at lowest risk of these cancers were between 20 and 25 years of age (Figure 3.13). When the joint evolution of a cancer and a risk factor are studied, it

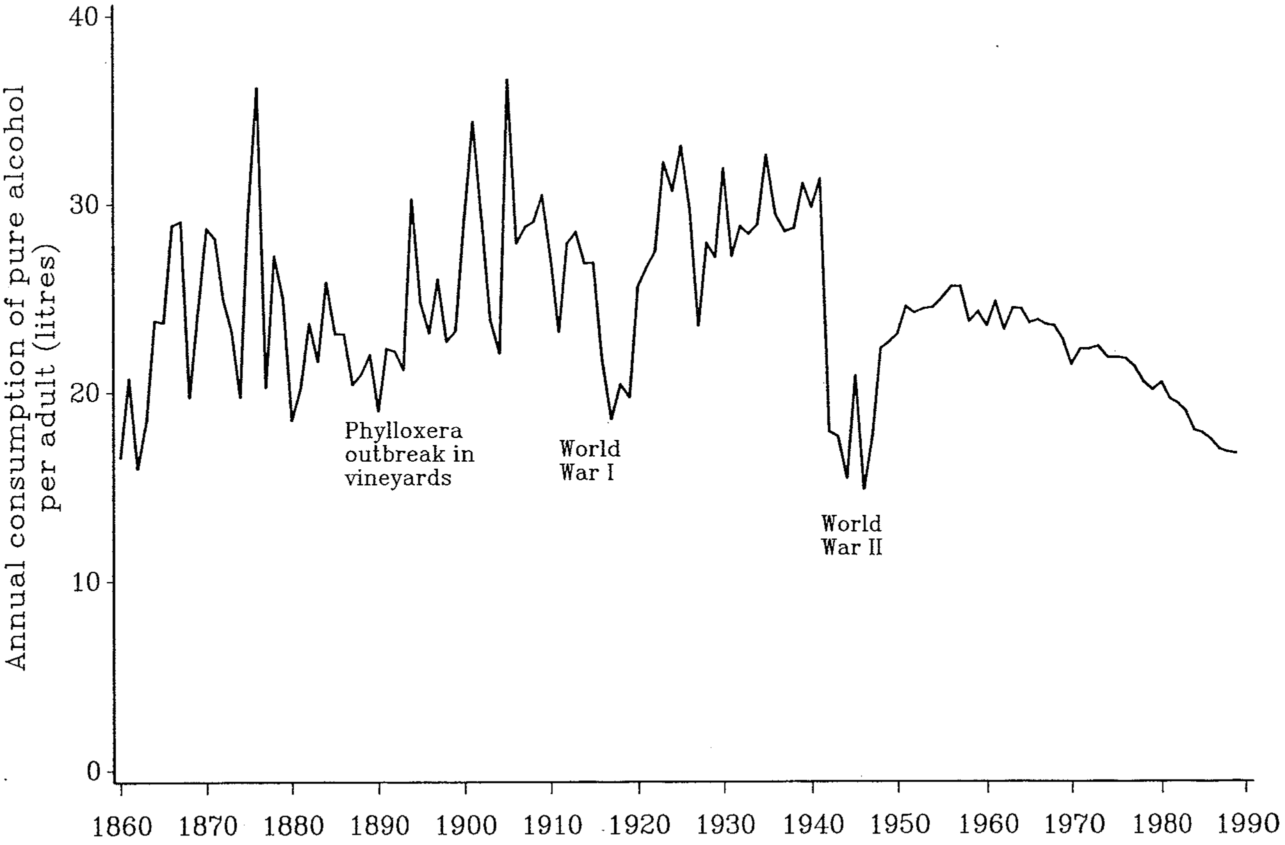


Figure 3.13 Change in alcohol consumption in France between 1860 and 1989
Source : Hill et al. [64]

is necessary to consider the mechanism of action of the risk factor and particularly the latency period. Thus, in contrast to the previous observation on alcohol-related cancers, the large decrease in tobacco consumption during the second world war did not have a marked effect on lung cancer mortality. In fact, it is difficult to detect joint evolution : risks and etiological factors generally undergo a slow, steady evolution.

The observation of time series can also be seen as an instrument for epidemiological surveillance of the population with the aim of detecting new risk factors. However, in addition to the difficulties created by the delayed effects of the latency period, it should be emphasized that rapid detection of changes in trends is not easily achieved. In particular, when monitoring relates to a relatively small population or a small risk, observed variations are often simply a reflection of purely random fluctuations.

The study of time trends is of particular interest in the evaluation of *primary prevention*, which involves the reduction in exposure to risk factors, and of *secondary prevention* (screening) which is aimed at reducing mortality. It is anticipated that the intervention will cause a more or less generalized shift in the existing trend in incidence or mortality. Before-and-after designs, aimed at identifying such shifts, have generally been used for this purpose.

The study of time trends is not limited to incidence or mortality. Descriptive epidemiology is increasingly concerned with the overall assessment of progress made through improved treatment or earlier detection of disease. This requires methods for quantifying the corresponding increase in survival rates calculated for all cases in the population in which the evaluation is being carried out.

Finally, from the public health viewpoint, the observation of changes in risk in the recent past leads naturally to a desire to predict its future development, in order to determine budget priorities and plan necessary services.

The following sections are devoted to definitions and basic concepts, which are of fundamental importance in the development of modelling methods, particularly those used in identifying age, period and cohort effects.

Methods

Components of temporal evolution

From 1955 to 1959, 417438 deaths from cancer were registered in France. Twenty-five years later, between 1980 and 1984, these deaths numbered 638 012. In other words, cancer deaths increased 53% over 25 years, or 1.7% per year (see formulae 3.68 and 3.69). To varying degrees, the same phenomena occurred in other Latin countries (Table 3.13). The increase concerned not only numbers of deaths for each type of cancer but also their proportion in all-cause mortality and crude rates.

Table 3.13 Changes in number of deaths ([a]) for cancer between 1955 and 1984 in selected European countries

	Men			Women		
	1955-59	1980-84 ([b])	Variation ([c]) (% per year)	1955-59	1980-84 ([b])	Variation ([c]) (% per year)
Spain						
Number	77 438	172 957	3.2	74 036	121 350	2.0
Crude rate	108.1	187.3	2.2	97.29	126.6	1.1
Proportion ([d])	11.2	22.6	3.0	10.9	17.6	2.0
France						
Number	212 718	382 883	2.4	204 720	255 129	0.9
Crude rate	198.4	288.2	1.5	179.0	183.6	0.1
Proportion ([d])	16.1	26.8	2.0	16.1	19.3	0.7
Italy						
Number	173 405	369 232	3.0	157 638	258 172	2.0
Crude rate	143.9	266.4	2.5	125.6	177.0	1.4
Proportion ([d])	14.1	25.4	2.4	14.2	19.9	1.4
Switzerland						
Number	25 800	41 552	1.9	23 213	32 860	1.4
Crude rate	207.8	267.1	1.0	175.2	200.4	0.5
Proportion ([d])	19.8	26.8	1.2	19.0	23.1	0.8

([a]) WHO mortality data bank.
([b]) Spain 1980-81; Italy: 1980-83.
([c]) Average annual rate of change over the period of n = t_1-t_0 years, calculated according to formula (3.69).
([d]) Proportion of deaths from cancer among deaths from all causes in the period.

These observations are important from the public health viewpoint. However, they do not reveal anything about the way in which cancer risk evolved over the course of the 25 years, and can even lead to errors in interpretation. The proportion of deaths due to cancer increases partly because of a decrease in the number of deaths from competing causes, while the increase in crude rates is largely explained by the ageing of the population. An examination of trends in the net risk of cancer mortality which leaves aside competing causes ends up with rather different conclusions (Table 3.14). In particular, net cancer mortality decreases when cancers associated with tobacco use are excluded. Similar conclusions were reached by a study carried out some years ago in the USA: while the number of cancer deaths increased 181% between 1930 and 1970, an analysis of the components of the increase shows that 10% was due to change of risk, 74% to population growth, 46% to the ageing of the population, 17% to the amplification of changes in risk resulting from demographic changes and finally 34% to interactions between demographic factors (62). A recent study carried out for the European Community predicted that cancer mortality would increase 48% for men and 20% for women between 1980 and 2000, with approximately half of this variation due to demographic changes expected during this time.

Table 3.14 Change in net risk (a) of dying from cancer between 1955 and 1984 in selected European countries

Country	Men			Women		
	1955-59	1980-84 (b)	Variation (c) (% per year)	1955-59	1980-84 (b)	Variation (c) (% per year)
Spain						
Tobacco-related (d)	3.5	7.2	3.1	0.6	0.7	0.7
Other	9.3	9.9	0.3	8.3	8.0	−0.2
Total	12.8	17.1	1.2	8.9	8.7	−0.1
France						
Tobacco-related (d)	6.2	10.8	2.2	0.7	0.9	1.0
Other	11.9	12.0	0.03	10.8	9.0	−0.7
Total	18.1	22.8	0.9	11.5	9.9	−0.6
Italy						
Tobacco-related (d)	4.3	10.2	3.5	0.8	1.1	1.3
Other	11.2	11.8	0.2	10.4	9.8	−0.2
Total	15.5	22.0	1.4	11.2	10.9	−0.1
Switzerland						
Tobacco-related (d)	6.9	8.9	1.0	0.8	1.1	1.3
Other	12.5	10.7	−0.6	12.3	9.9	−0.9
Total	19.4	19.6	0.04	13.1	11.0	−0.7

(a) Net risk is measured by the cumulative risk from 0 to 75 years ; source : WHO mortality data bank.
(b) Spain: 1980-81; Italy: 1980-83.
(c) Average annual rate of change over the period of $n = t_1 - t_0$ years, calculated according to formula (3.69).
(d) Sites for which the effect of tobacco use has been established (mouth and pharynx, oesophagus, lung, larynx and bladder).

In etiological research, the focus should be on the risk of disease and not only the risk of death. Unfortunately, trends in incidence can be studied in only a few countries, because of the relatively recent establishment of cancer registration. In addition, results can rarely be generalized because registries often cover subpopulations chosen by circumstance, not necessarily corresponding to regions that would have been selected for the study of specific hypotheses. Therefore, we are often forced to rely on mortality data, which are available over long time periods for both national and regional populations. Nevertheless, it should be kept in mind that the risk of death is only an indirect, and even a biased measure, of the risk of cancer occurrence, particularly because of the increase in survival.

The methods proposed in Chapter 2 for comparing incidence between populations should in principle be suitable for studying changes over time. However, most of these methods rely on the assumption that ratios of incidence (or mortality) remain more or less constant with age. In fact, it is far from certain that risk alters in the same way for all age groups in a changing environment. Indeed, there are, in general, good reasons to assume that different age groups behave in different ways.

The epidemic of lung cancer illustrates this point. At first, the older age groups were unaffected and the increase in risk was observed only in younger age groups. Signs that the epidemic is declining are now obvious, for example, in the UK and the USA, and again in the youngest age groups which are decreasingly exposed to the carcinogenic effects of tobacco. In the oldest groups, on the other hand, the increase in risk is sustained for much longer, as they are still experiencing the consequences of high tobacco consumption twenty years ago. In France, where the smoking epidemic occurred later, there is still an increase in risk in the younger age groups (Figure 3.14). In such circumstances, neither crude nor standardized rates can provide an appropriate assessment of trend. Calculations based on age-adjusted rates, which in principle control for the effects of population ageing, provide an incomplete picture of the phenomenon, and hide its more interesting components.

This example underlines the importance of observing changes in risk in young adults when the consequences of a new risk factor or protective agent are to be assessed (or predicted). For cancer, as for most non-transmissible diseases, etiological factors are often linked to forms of social behaviour which come and go with passing generations.

These considerations are illustrated in Figure 3.15, which shows cancer mortality over time in Scotland. If we only consider overall trends, the patterns in three usual standardized rates (African, European and world standards) are similar and indicate a regular and relatively small increase in risk. On the other hand, examination of rates calculated for less than and greater than 65 years of age shows that the trend in standardized rates is due to changes which diverge with age, with an

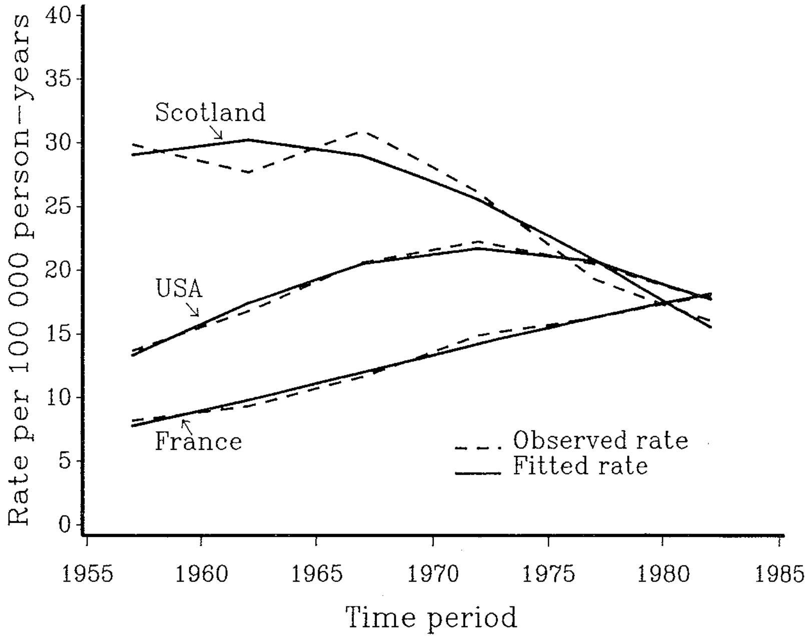


Figure 3.14 Lung cancer mortality trend in France, the USA and Scotland in 40–44-year-old men

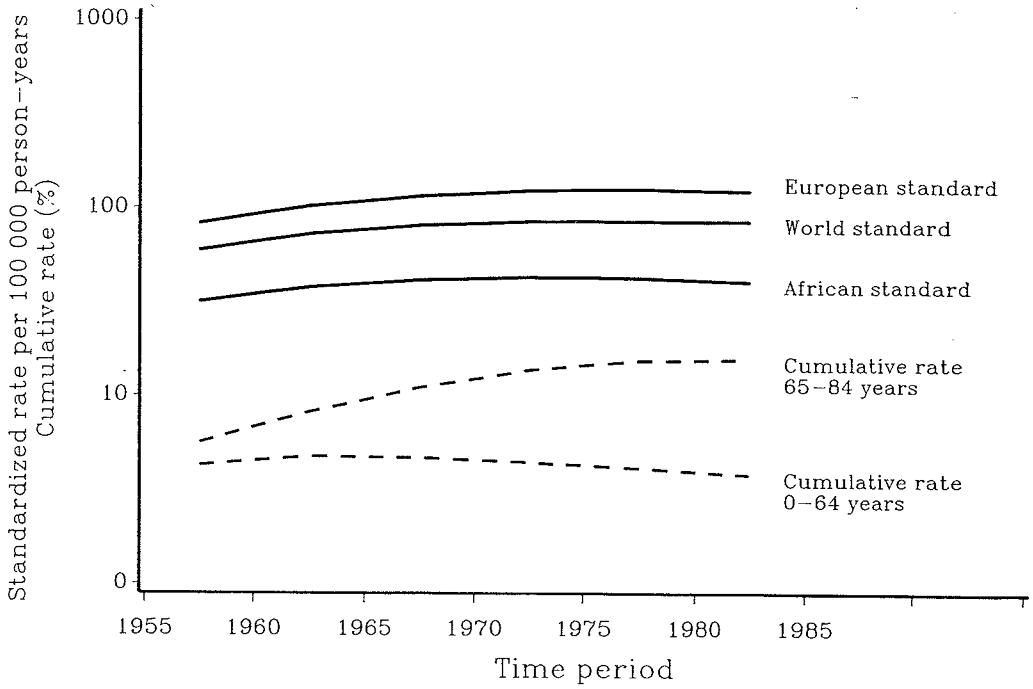


Figure 3.15 Lung cancer mortality trend in Scotland; men, 1955-1984

increase in risk for the oldest age groups and a decrease for the youngest. It is likely that this decline signals an inversion of trend which will ultimately affect other age groups.

Effect of data quality

In addition to real trends in risk and random variations, changes in data quality over time affect the observed trend in incidence or mortality. These effects can create apparent increases or decreases in risk, when the true risk is actually completely stable.

For incidence data, time series partially reflect progressive improvements in the registration rate, whether resulting from the development of diagnostic techniques or improved reporting systems for the registry. The newer the registry, the stronger this effect is likely to be. In some situations, the very existence of the registry creates an awareness which increases the proportion of cases diagnosed (such as through post-mortem examinations). In most registries, there has been a progressive decrease in the proportion of registered cases on the sole basis of death certificates. In Connecticut, the proportion declined from 35% in the first years after the registry was established (1935) to 1% in the 1980s [63]. This improvement in the rate of registration of cases during their lifetime has led to a temporary and artificial increase in the number of incident cases. It has been proposed that the standard indices calculated to assess the completeness of registration (proportion of cases registered from death certificates only and frequency of autopsy) be used to correct incidence rates.

As a registry develops, there is also an improvement in the quality of diagnostic information obtained for each cancer registered, and a consequent increase in the precision in coding of the site and type of the tumour. Codes corresponding to poorly defined sites are progressively less used as the percentage of histologically confirmed cases rises. An artificial increase in the frequency of well specified sites will therefore be seen. In the Connecticut registry, the percentage of histologically confirmed cases increased from 73% to 93% during the period mentioned above [63].

Finally, incidence can fluctuate as a result of changes in the stage at which cancer is detected, particularly for slow-growing tumours. For example, it is known that the incidence of *in situ* cervical tumours can increase explosively during cytological screening campaigns, because of the inclusion of prevalent cases which are not detected clinically. The detection of early stage disease has an even greater effect in the study of time trends in survival.

The quality of cancer mortality data has undoubtedly also improved over time, but the improvement has occurred more in the precision of diagnosis than in the number of registered deaths. As with incidence, there have probably been artificial increases in the number of deaths from better defined causes. Thus, increased mortality from ovarian cancer observed in France between 1950 and 1985 in women over 50 years might be due partially to the introduction of systematic surgical investigation of abdominal masses. Previously, some ovarian cancers discovered at an advanced stage were wrongly classified as peritoneal cancer [64].

Problems in classification have been discussed too extensively elsewhere to justify detailed review here. We simply note that all changes in classification, or even coding practices, can affect the number of cases at a given site or due to a specific cause of death and distort trends. The decision to register papillomas or non-infiltrating lesions has clearly played a role in the apparent increase in the incidence of bladder cancer. Also well known are the difficulties which arise in the study of trends in non-Hodgkin lymphoma, which is sometimes coded according to topographical site and other times as a tumour of the haematopoietic system.

The problem of imprecise data is accentuated by the differences in the evolution of precision with region or age. Errors in diagnosis are generally more serious in older people, and improvements in diagnostic precision can therefore have a fundamental effect on incidence rates in this age group. The phenomenon is probably a partial explanation for the recent increase in multiple myeloma in the elderly [65]. As a final point, it should be noted that chronological patterns in incidence or mortality rates depend on the quality of the denominators over time. Population estimates provided by statistical services may be increasingly distorted the further they are from the date of the census. This distortion often results in an underestimation of the denominators, because enumeration is not as accurate for persons leaving the population as it is for those arriving.

Role of modelling

Observed time trends should be evaluated in the context of the problem under study : sometimes it is sufficient to describe long-term trends; in other situations, interest might focus on variation over a more limited time period, in particular the recent past, if the goal is to predict new directions of the phenomenon. Apart from the simple description of changes in risk over time, the study of trends should therefore involve the search for models which can describe observed data via plausible hypotheses about the causes of observed changes. Under this approach, the relevant components of the time trend can effectively be separated from the random or systematic (e.g., seasonal fluctuations), allowing a more complete interpretation of the observed data.

Models of risk evolution over successive generations have a particularly important place in the study of cancer incidence, because of the long latency period between the start of exposure to a risk factor and the occurrence of the disease. When interest focuses on the *generation effect*, also known as the *cohort effect*, the inevitable presence of *period effects* created by, for example, changes in diagnostic practice or the appearance of an environmental risk factor which could simultaneously affect all age groups, necessitates the combined analysis of both the cohort and period components of risk. In other situations, the period effect may be of primary interest and the cohort effect is only a confounding factor that must be controlled for. An example of this situation is the evaluation of the effect of screening for cervical cancer (see page 202).

The use of models in the study of trends has not been widespread, because of two fundamental problems which will be discussed in this section.

The first is the difficulty of separating meaningful variations from those which can be considered to be random fluctuations. Simpler models might be discarded because the random component is in fact greater than that predicted by the Poisson distribution which is used to assess significance of the terms included in the model. In such a situation, it might be wrongly concluded that specific factors play a significant role in the explanation of the observed phenomenon.

The second difficulty lies in the impossibility of satisfactorily separating cohort and period effects from the data alone, when hypotheses on the nature of these components cannot be formulated *a priori*. It is for this reason that some authors have questioned the value of modelling over traditional graphical approaches to carry out this type of investigation [66]. This point of view, however, ignores the fact that exclusive use of graphical methods can also lead to subjective interpretations which an appropriate model may avoid.

The following section presents the tools required for the quantitative description of trends and the evaluation of the adequacy of the underlying models. Data on lung cancer in young adults are used to show how the analysis of trends in the logarithms of age-specific rates can display several types of time trend, and ultimately allow different components of this evolution to be revealed. This analysis naturally leads to a discussion of age-period, age-cohort and age-period-cohort models.

Description of trend by period

First, recall the concept of rate of change, which summarizes exponential increase in incidence or mortality. If $N(t_0)$ cancers were observed in year t_0, and $N(t_1)$ cancers in year $t_1 = t_0 + n$, the relative change is measured by:

$$\tau = \frac{N(t_1) - N(t_0)}{N(t_0)} \tag{3.66}$$

or by the corresponding percentage 100 xτ.

To derive the constant annual rate of change r that must apply each year to observe this relative change after n years, write:

$$N(t_1) = N(t_0)(1 + r)^n \tag{3.67}$$

or

$$1 + r = (1 + \tau)^{\frac{1}{n}} = \left(\frac{N(t_1)}{N(t_0)}\right)^{\frac{1}{n}} \tag{3.68}$$

given that $t_1 - t_0$ is equal to n, we have

$$\mathrm{Log}\,(1 + r) = \frac{\mathrm{Log}\,[N(t_1)] - \mathrm{Log}\,[N(t_0)]}{t_1 - t_0} \tag{3.69}$$

in other words, the slope of the line linking the logarithm of incidence at the two time points under consideration is practically equal to the average annual rate of change in incidence, since $\mathrm{Log}(1 + r) \approx r$ when the rate is small. If the rate is not small, and if ß denotes this slope, we have the relationship $r = e^{\beta} - 1$. The calculation above based on number of incident cases can obviously be carried out with all other indices of incidence or mortality.

When the numbers of cases occurring in the intervening years are known and if the logarithm of incidence varies linearly between the two dates, the rate of increase can be estimated by the slope of the line which best represents the logarithm of incidence as a linear function of year of diagnosis or death. Estimation of this regression line can be based on either maximum likelihood or weighted least squares.

As an example, we calculated the annual rate of change in lung cancer mortality among males in the USA, France and Scotland in the 40-44 years age group. The data for six successive five-year periods appear in Table 3.15 and in Figure 3.14.

Let k_t, m_t, λ_t be the numbers of cases and person-years and the incidence rate for the age group under consideration for the period t. As was described above, the rate of change is the value $e^{\beta_1} - 1 \approx \beta_1$ in the equation :

$$\mathrm{Log}(\lambda_t) = \beta_1 t + \beta_0 \tag{3.70}$$

Table 3.15 Change in lung cancer mortality over 25 years for men aged between 40 and 45 years

	USA			Scotland			France		
	Number	M x Y (a)	Rate (b)	Number	M x Y (a)	Rate (b)	Number	M x Y (a)	Rate (b)
1955-59	3 762	27 599	13.6	242	811.0	29.8	479	5 878	8.2
1960-64	4 900	29 249	16.8	222	803.2	27.6	612	6 586	9.3
1965-69	6 147	29 859	20.6	247	798.8	30.9	968	8 333	11.6
1970-74	6 318	28 416	22.2	195	747.7	26.1	1 265	8 507	14.9
1975-79	5 638	27 590	20.4	138	717.6	19.2	1 308	8 032	16.3
1980-84	5 413	30 569	17.7	116	724.4	16.0	1 349	7 484	18.0
Rate of change (c)		4.95 %			− 10.63 %			17.67 %	

(a) Man-years in thousands.
(b) Death rate per 100 000 man-years.
(c) Five-yearly percentage change $100(e^{\beta_1} - 1)$ estimated by the method of maximum likelihood using the linear model (3.70).

The parameter β_1 is estimated by maximum likelihood, supposing that k_t has a Poisson distribution with mean $m_t\, e^{\beta_1 t+\beta_0}$, or by using weighted least squares, minimizing:

$$\Delta(\beta_1, \beta_0) = \sum_t w_t \left[\text{Log}\left(\frac{k_t}{m_t}\right) - \beta_1 t - \beta_0 \right]^2 \tag{3.71}$$

where w_t is proportional to the inverse of the variance of the logarithm of the observed rate, that is :

$$w_t = \lambda_t m_t \approx k_t$$

The calculations were carried out with the software GLIM, using a program given in Appendix 2. Table 3.16 shows that the estimate of the rate of change and the deviance (an overall measure of the quality of the model's fit) are almost identical for the methods of maximum likelihood and weighted least squares when the model specifies a linear change in rates. On the other hand, the precision of the estimate, as indicated by the standard error, appears much greater when the method of maximum likelihood is used. In fact, this method assumes that the model is appropriate and that the variation observed around the values calculated for each period using equation (3.70) are those predicted by the Poisson distribution. In this situation, the deviance indicates that the differences between observed and expected numbers are too big for the model to be acceptable. This statistic should be of the order of 4 (the mean of a χ^2 distribution on four degrees of freedom), if the logarithms of the rates really varied linearly with time. Figure 3.14 suggests that the linear model

Table 3.16 Modelling of data from Table 3.15

Country	Method (a)	Model	Coefficients (b)		Deviance	d.f.
			β_2	β_1		
USA	ML	Linear	–	0.048 (0.003)	553.1	4
		Quadratic	−0.053 (0.002)	0.428 (0.017)	14.1	3
	LS	Linear	–	0.048 (0.041)	547.2	4
		Quadratic	−0.053 (0.005)	0.428 (0.036)	14.0	3
Scotland	ML	Linear	–	−0.112 (0.018)	15.5	4
		Quadratic	−0.041 (0.012)	0.162 (0.083)	4.1	3
	LS	Linear	–	−0.111 (0.036)	15.8	4
		Quadratic	−0.041 (0.014)	0.164 (0.095)	4.0	3
France	ML	Linear	–	0.163 (0.008)	14.9	4
		Quadratic	−0.016 (0.006)	0.282 (0.042)	6.5	3
	LS	Linear	–	0.163 (0.016)	15.0	4
		Quadratic	−0.016 (0.080)	0.282 (0.061)	6.5	3

(a) ML = maximum likelihood method based on the Poisson distribution; LS = method of weighted least squares.
(b) Standard error in brackets.

is quite good for France but not for the USA and Scotland. The measure of fit (deviation) is very bad for the USA (553.1 for a χ^2 on four degrees of freedom) but also poor for Scotland and France (approximately 15 on four degrees of freedom).

In the present situation, the poor fit observed for the USA and Scotland is partly due to the inversion of trends observed in these two countries during the period being studied. A linear model is therefore inadequate, and a second-order term must be added in the model to account for the concave curve representing this phenomenon:

$$\text{Log}(\lambda_t) = \beta_2 t^2 + \beta_1 t + \beta_0 \qquad (3.72)$$

Fitting this quadratic model, represented geometrically by a parabola, significantly improves the deviance compared to the linear model, as judged by maximum likelihood. This result suggests that the trend inversion is real.

The validity of this conclusion is difficult to challenge for Scotland because the second-degree model fits the data perfectly ($\chi^2 = 4.1$ for three degrees of freedom). The conclusion is also confirmed by the standard error of the quadratic term obtained from the method of least squares (t test = $-0.041/0.014 = -2.9$). This method assumes that $\log(k_t / m_t)$ has a normal distribution with mean $\beta_2 t^2 + \beta_1 t + \beta_0$ and variance σ^2/k_t. As σ^2 is estimated by the quotient of the deviance and its number of degrees of freedom, it will not be very different from 1 when the model with Poisson error is satisfactory. The result is that, in this situation, the standard error of the parameters obtained with the method of least squares will be close to the standard error estimated under the method of maximum likelihood, as can be seen from the Scottish data. Therefore, when the deviance suggests a good fit, the two methods are practically equivalent.

For the USA, and to a lesser extent for France, the problem of lack of fit remains. The test of the quadratic term based on the standard error obtained from the method of maximum likelihood is therefore not valid. For the French data, the coefficient of the quadratic term is not significant when evaluated by the method of least squares ($F_3^1 = (15.0 - 6.5)/(6.5/3) = 3.92$), but it is highly significant by the method of maximum likelihood ($\chi^2 = 14.9 - 6.5 = 8.4$ on one degree of freedom). Similarly, the standard error obtained using the method of maximum likelihood for the linear coefficient in the US data is obviously incorrect, while that obtained by the method of least squares correctly indicates the poor fit of this model. The two methods thus lead to contradictory results with neither being truly satisfactory.

For the USA and France, a large number of person-years of observation are available from populations that are *a priori* quite heterogeneous with respect to lung cancer risk. It is therefore likely that the randomness predicted by the Poisson distribution accounts only for a small portion of the random variation in the data. In particular, the assumption of a constant risk λ_t for all individuals is an oversimplification which masks a much more complex reality. For these two countries, the size of the populations being studied allows the rates to be estimated more precisely, showing that the observed variability is significantly greater than that predicted by the Poisson distribution.

The fit could certainly be improved by constructing a more complex model, especially by adding higher degree terms to describe observed variations more precisely; however, this approach is contrary to the principle of simplicity which is fundamental to all modelling, and can lead to a good but useless description of purely random variation.

In order to take the excess variability into account, it is preferable to conclude explicitly that λ_t, a fixed parameter to be estimated in the previous calculations, is in fact a random variable describing the distribution of risk in the population under study. Equations (3.70) and (3.72) are then only true on average. Effectively, we have:

$$\mathrm{Log}\,(\lambda_t) = f\,(t) + \varepsilon_t = \mathrm{Log}\,(v_t) + \varepsilon_t \qquad (3.73)$$

where f(t) is the model proposed for the change over time in the mean of Log(v_t), the logarithm of the rate, and ε_t is a random variable of unknown distribution and constant variance σ^2 [67,68]. Hinde assumes, in addition, that the distribution of ε_t is normal [67].

As a first approximation, Log(k_t / m_t) can be assumed to have a normal distribution with mean f(t) and variance $1/v_t m_t + \sigma^2$, the sum of the Poisson and extra-Poisson variance.

Calculations not shown here show that estimation of this model by the maximum likelihood method from data given in Table 3.15 gives σ^2 equal to 0.260×10^{-3} for the US data. This value corresponds to extra-Poisson variation of between 30 and 50%, but the likelihood is not significantly improved by the introduction of this additional parameter (χ^2 = 2.92 on one degree of freedom).

The estimate of σ^2 is null for the Scottish data, as would be expected given the excellent fit of the quadratic model without an extra Poisson variation obtained previously (Table 3.16).

The French data are as well described by a quadratic model without extra Poisson variation as by a linear model which includes variation of this type between 30 and 60% ($\hat{\sigma}^2 = 0.144 \times 10^{-2}$). This result proves that the slowing of the increase in lung cancer mortality, suggested by the more recent data, requires further confirmation before being unequivocally accepted.

From this discussion, it is clear that the rate of change alone is rarely sufficient to comprehensively describe the data, even within a single age group. *A fortiori*, a method which describes the evolution of the logarithm of a standardized rate using a linear regression can conceal interesting aspects of a time trend. In the Scottish data (Figure 3.15), it can be seen that standardization leads to an estimated increase of between 0.90 and 1.52% per year, depending on the standard population. However, the cumulative rate between 65 and 84 years of age increases by more than 4% per year, while the rate from 0 to 64 years decreases by nearly 0.6% per year, as shown in Table 3.17. Note that the trend in the cumulative rate between 0 and 84 years depends largely on the trend observed in the elderly and, consequently,

Table 3.17 Change in lung cancer mortality in men in Scotland ([a])

Standard population	Rate of change ([b])	Standard error
European	1.52	0.41
World	1.19	0.42
African	0.90	0.40
Cumulative rate 0-64 years	-0.61	0.32
65-84 years	4.10	0.65
0-84 years	2.70	0.47

([a]) Mortality data in six five-yearly periods from 1955 to 1984 (see Figure 3.15).WHO mortality data bank.

([b]) Estimated by the method of least squares assuming that the logarithm of the standardized rate varies linearly; the result is expressed as a percentage change per year.

completely disregards the important epidemiological fact that the lung cancer rate is decreasing in young people, as might be predicted by the changing smoking habits of this generation.

The preceding discussion underlines the importance of studying time trends with respect to age. Three examples corresponding to different epidemiological situations are shown in Figure 3.16. The first example concerns the incidence of bladder cancer in Birmingham, UK. The incidence of this cancer increased sharply for all age groups from the end of the 1960s, due to the inclusion of papillomas. The calculated rates of change are thus positive and of the same order of magnitude at each age; the curve obtained is approximately a horizontal line. The second example concerns the evolution of lung cancer mortality in Scotland, already discussed on several occasions. The graph shows that the rates of change increase strongly with age, and become positive after 65 years. The third example is provided by the incidence of cervical cancer in Birmingham, UK. The graph is a complex curve with a minimum at around 40 years. This shape could be partially explained by the progressive extension of screening to successive generations, and partly by increased exposure among young women to risk factors linked to sexual behaviour.

To obtain the data in Figure 3.16, rates of change have been calculated for each age group by fitting of the log-linear model :

$$\text{Log}\,(\lambda_{xt}) = \alpha_x + \beta_x t \tag{3.74}$$

where the rate of change β_x depends on the age group x.

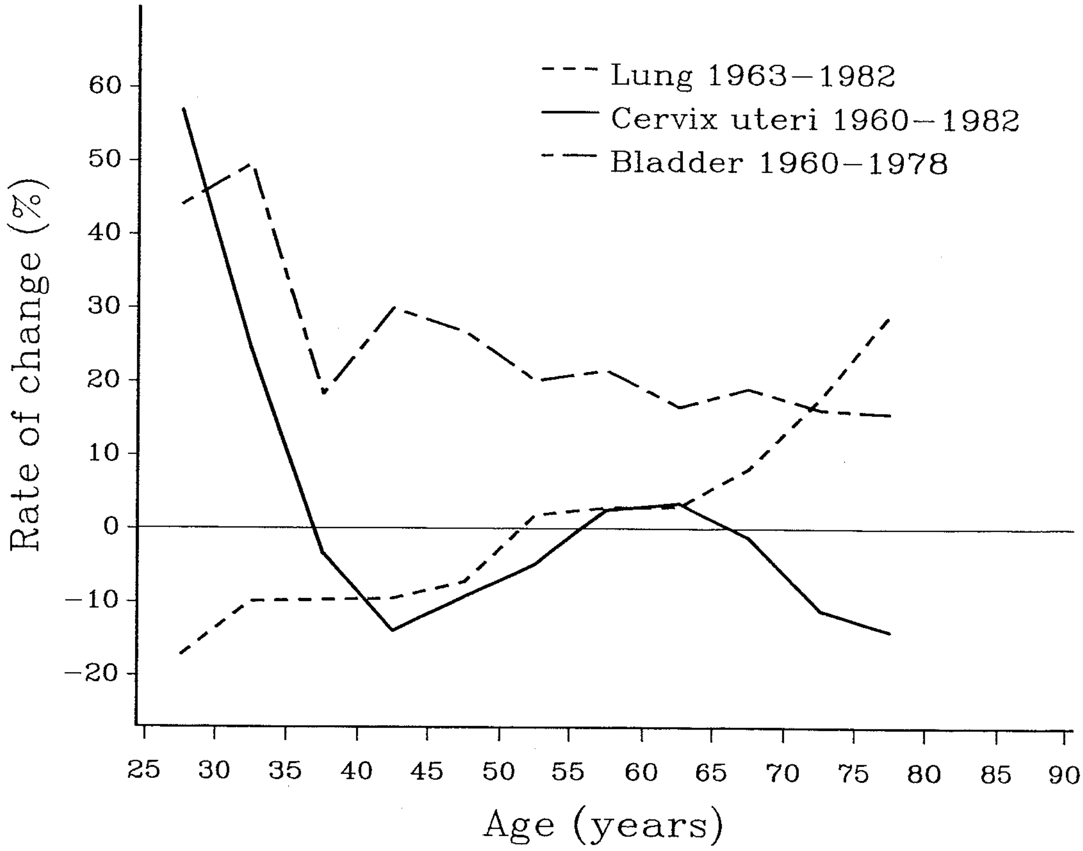


Figure 3.16 Trend in the age-specific incidence of bladder cancer in men and cervical cancer (Birmingham,UK), and of lung cancer in men (Scotland)

For bladder cancer in Birmingham, UK, the constant rate of change with age suggests that a model in which β is constant will provide an equally good description of the data. This model is written:

$$\text{Log}\,(\lambda_{xt}) = \alpha_x + \beta t \tag{3.75}$$

Fitting the models (3.74) and (3.75) gives χ^2 values of 70.28 and 80.69 respectively, on (4 × 11 – (2 × 11) = 22 and (4 × 11 – (11 + 1)) = 32 degrees of freedom, showing that the improvement in fit created by introducing a different slope for each age group is negligible. Nevertheless, the size of the deviance indicates that the linear model does not adequately describe the data.

A careful examination of the data given in Table 3.18 and Figure 3.17 shows that the increase, although similar in all age groups, was greater between the second and third time periods. The absence of linearity is not surprising in that it corresponds to a change in the case definition which occurred precisely between the second and third period, and resulted in the inclusion of papillomas, previously considered benign. The constant rate of change observed before indicates that this event has produced an effect which is proportional to the existing incidence. This finding was not obvious *a priori*: the relationship between papillomas and invasive cases could have varied with age. We therefore adopt a multiplicative model, in which the incidence rate is multiplied by a factor independent of age. In addition, the poor fit of the linear model leads us to calculate a relative rate for each period, rather than a single parameter summarizing the increase over the 15 years of reg-

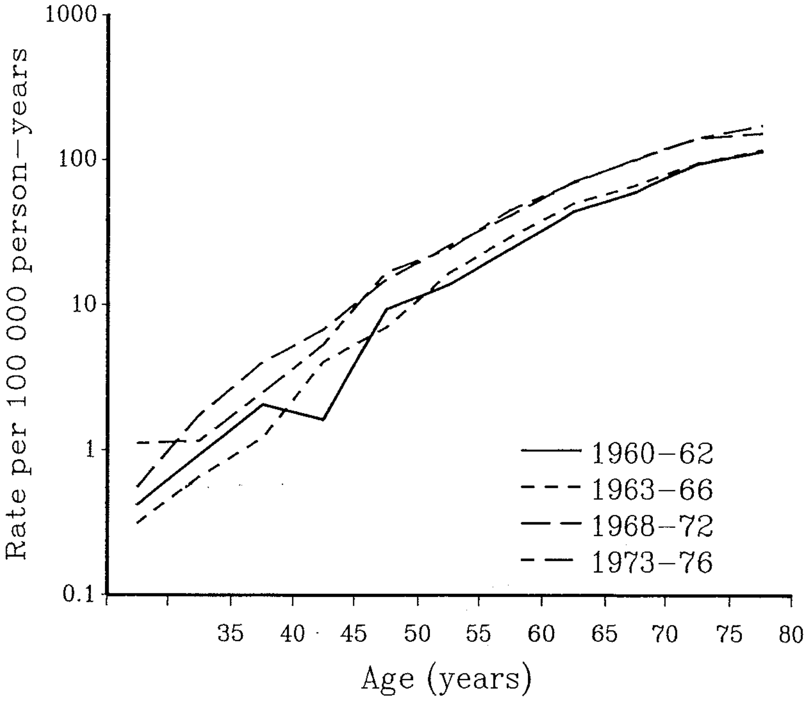


Figure 3.17 Bladder cancer incidence in Birmingham, UK; men, 1960-1976

istration. This way of describing the rates is usually called an *age-period model.* It is written:

$$\operatorname{Log}(\lambda_{xt}) = \alpha_x + \beta_t \qquad \begin{array}{l} x_1 \leq x \leq x_g \\ t_1 \leq t \leq t_h \\ \beta_{t_1} = 0 \end{array} \qquad (3.76)$$

where g and h are respectively the number of age groups and the number of study periods. The term ßt in the linear model is thus replaced by a term β_t, indicating changes of unspecified shape over time which are nevertheless identical in all age groups.

Maximum likelihood estimates of $\mu_x = 100\,000\,e^{\alpha_x}$ and $\rho_t = e^{\beta_t}$ are given in Table 3.18. The values of μ_x provide a smoothed incidence curve for the first registration period and ρ_t provide a description of the increase similar to that given by the SIR in the same Table.

The goodness-of-fit of the multiplicative age-period model can be assessed from the results. For example, incidence for the age group 60 to 64 years in the third period is estimated by:

$$100\,000\,\lambda_{xt} = 100\,000\,e^{\alpha_x}\,e^{\beta_t} = \mu_x\,\rho_t = 43.99 \times 1.62 = 71.26$$

Table 3.18 Incidence of bladder cancer; men, Birmingham, UK, 1960-1976 ([a])

Age (x)	Registration period (t)				Estimated rates ([b]) (μ_x)
	1960-62	1963-66	1968-72	1973-76	
25-29	0.42	0.31	0.55	1.10	0.45
30-34	0.00	0.65	1.73	1.15	0.71
35-39	2.06	1.21	4.02	2.49	1.85
40-44	1.62	4.03	6.74	5.29	3.43
45-49	9.40	7.02	14.95	16.80	8.97
50-54	13.90	16.65	25.73	24.41	15.12
55-59	24.25	29.15	41.06	44.81	25.94
60-64	44.50	50.51	71.39	70.25	43.99
65-69	60.47	66.97	100.69	101.97	61.62
70-74	94.84	95.73	141.96	142.70	87.88
75-79	116.08	118.16	154.19	174.42	103.43
Relative rate ([b]) : ρ_t	1.00	1.09	1.62	1.65	
SIR ([c])	71.47	78.00	115.91	117.98	
Observed cumulative rate 25-79 years	1.84	1.95	2.82	2.93	

[a] Rates as number of cases per 100 000.
[b] Estimated using an age-period model (3.76).
[c] Using observed incidence between 1970 and 1976 as standard.

as compared to the observed number of 71.39. The deviance of this model is 41.17 on $(4 \times 11 - (11 + 3)) = 30$ degrees of freedom. Despite being somewhat large for a χ^2 on 30 degrees of freedom, this value ($p = 0.08$) confirms that the multiplicative model is a good description of the data. As the SIRs have been designed for such a situation, they obviously provide a good description of the time trend. A detailed discussion of the adequacy of this model for the Birmingham incidence data can be found in a recent article which, to a large extent, inspired these developments [70].

At this stage, it is important to ask why an age-period model has been adopted to describe this data set. The presence of a clear change in rates for all ages between the second and third registration period excluded the model (3.75). In other words, it was necessary to introduce the effect of period as a non linear function of time, leading to model (3.76), which has an acceptable fit because of the proportionality of the observed incidence curves.

It is worth dwelling a little longer on model (3.75) which, as we will see below, can be equally well interpreted as an age-period or an age-cohort model. This model, known as an *age-drift* model [70, 71], implies the same linear change in the logarithms of incidence rates over time for all age groups. In this situation, the estimate of the rate of change β (or $e^\beta - 1$, if β is large) is a complete summary of the time trend. This model and an example of its application are presented below in detail.

Table 3.19 gives the incidence rate and the number of observed cases by five-year age group from 30 to 74 years for malignant melanoma in Norwegian women, for five time periods from 1960 to 1980. From the Table, it can be seen that incidence of this cancer has approximately quadrupled between 1960 and 1980 and that the increase has been very regular. This four-fold increase over 20 years corresponds to a growth of approximately 7% per year ($4^{1/20} = 1.07$).

We have seen that under model (3.75), $\text{Log}(\lambda_{xt})$ depends linearly on the period. On the other hand, the age effect is represented by separate parameters α_x for each age group, with no *a priori* assumptions about the shape of the age-incidence relationship. Just as we have considered other assumptions about the relationship with time, there are various ways of incorporating age in the model. Here, an age-drift model of the form:

$$\text{Log}\,(\lambda_{xt}) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \beta t \qquad (3.77)$$

where the logarithm of age-specific incidence is modelled by a polynomial of degree 3, provides a satisfactory fit for this data set. The deviance of the model fitted by maximum likelihood is 45.87 on 40 degrees of freedom ($p > 0.20$) and leads to an estimated annual rate of increase of 7.4%.

The age-drift model, shown in equation (3.75) in its age-period form, can be immediately transformed into an age-cohort model by writing:

$$\text{Log}\,(\lambda_{xu}) = (\alpha_x + \beta x) + \beta u = \alpha'_x + \beta u \qquad (3.78)$$

where $u = t - x$ is the year of birth of an individual aged x at time t. Thus, by adopting a different model of age-specific incidence, the age-drift model becomes an age-cohort model in which the change in risk depends linearly on the date of

Table 3.19 Incidence (a) of malignant melanoma in Norwegian women aged 30 to 74 years between 1959 and 1982

Age	Registration period				
	1959-61	1964-66	1968-72	1973-77	1978-82
30-34	3.10	4.84	8.07	12.14	11.71
	(10)	(14)	(42)	(79)	(89)
35-39	4.81	6.57	11.10	15.30	20.10
	(18)	(21)	(54)	(79)	(128)
40-44	6.47	7.84	12.01	20.65	21.01
	(25)	(29)	(64)	(101)	(108)
45-49	3.81	10.45	12.59	21.64	23.87
	(14)	(40)	(77)	(114)	(116)
50-54	4.36	6.07	10.17	18.23	22.30
	(15)	(22)	(64)	(112)	(118)
55-59	4.38	7.13	11.22	17.04	23.30
	(14)	(24)	(66)	(104)	(141)
60-64	4.48	8.74	8.85	15.18	21.52
	(13)	(27)	(48)	(86)	(127)
65-69	5.89	8.83	8.69	15.45	24.44
	(14)	(24)	(42)	(79)	(132)
70-74	10.39	7.97	12.86	15.83	24.90
	(19)	(17)	(52)	(69)	(116)
WTR (b)	5.27	7.52	10.75	15.35	21.93

(a) Rates per 100 000 person-years. Number of observed cases in brackets.

(b) Rates standardized to the truncated world population 30 to 74 years.

birth. If risk increases with time, incidence increases more rapidly with age if risk is measured longitudinally (intra-cohort); conversely, if incidence decreases, the cross-sectional incidence (intra-period) will have a steeper slope. The two curves differ by the quantity βx, a linear function of age (Figure 3.18), and serve to remind us that the real increase in risk of a given cancer with age cannot be determined when its incidence changes over time. Unless it is specified *a priori*, based on other observations, that the changes are due to either cohort or period effects, the increase in risk can only be measured up to a term βx.

Table 3.20 gives cross-sectional incidence estimated for the year 1975 based on model (3.77) and longitudinal incidence for the cohort born around 1925, calcu-

Table 3.20 Incidence of malignant melanoma by age for women in Norway (a)

	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Cross-sectional 1975 (b)	9.76	13.96	16.54	17.11	16.31	15.11	14.35	14.73	17.25
Longitudinal 1925 (c)	2.80	5.72	9.68	14.32	19.50	25.80	35.01	51.34	85.89

(a) Rate per 100 000 person-years.

(b) Incidence estimated for the year 1975 from model 3.77.

(c) Incidence estimated for the cohort born in 1925 from the model 3.78.

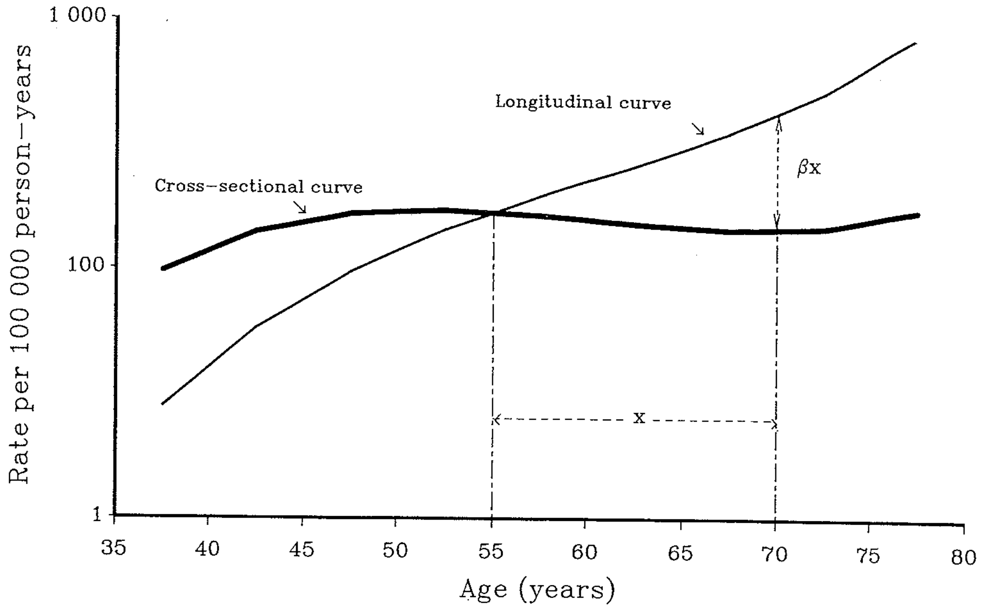


Figure 3.18 Skin melanoma incidence in Norway, women; 1975 (cross-sectional curve), and for the generation born in 1925 (longitudinal curve)

lated from the cohort version of the model given in formula (3.78). Here, the increase in risk with age is *a priori* better described by the longitudinal curve, insofar as sun exposure practices tend to change over generations. Furthermore, the cross-sectional curve gives a rather implausible description of the increase in risk with age. If this interpretation is correct, the cumulative risk of malignant melanoma for women aged between 30 and 74 years born in 1925 based on Table 3.20 is 1.25%. This risk has therefore increased from $(1.25/(1.077)^{25}) = 0.2\%$ for the generation born in 1900 to $(1.25 \times (1.077)^{15}) = 3.8\%$ for the generation of women born in 1940.

Table 3.19 can be reconstructed very accurately from the age-drift model using the data of Table 3.20 and a drift of 7.4%, except for the incidence over the first period in the age group 70-74 years, which is abnormally high. The estimated rate is in fact $17.25 \times e^{[0.074(1960-1975)]} = 5.68$.

Figure 3.19 shows rates estimated by cohort, under the longitudinal hypothesis. The change in shape observed between the oldest and youngest generations is quite likely to be mostly an artefact. This phenomenon once again shows how hard it is to model changes in risk with age: fitting a third-degree polynomial, which on average describes the data well in the observation period, undoubtedly leads to somewhat pessimistic estimates when extrapolated to young generations. Unfortunately, this uncertainty in the calculation of lifetime risk is inevitable, given that each cohort can only be observed over a limited age range.

Description of trend by cohort

Just as non-linear changes in risk with time leads to an age-period model, non-linear progression of risk with date of birth points to an age-cohort model. This model is satisfactory if the corresponding portions of the longitudinal incidence

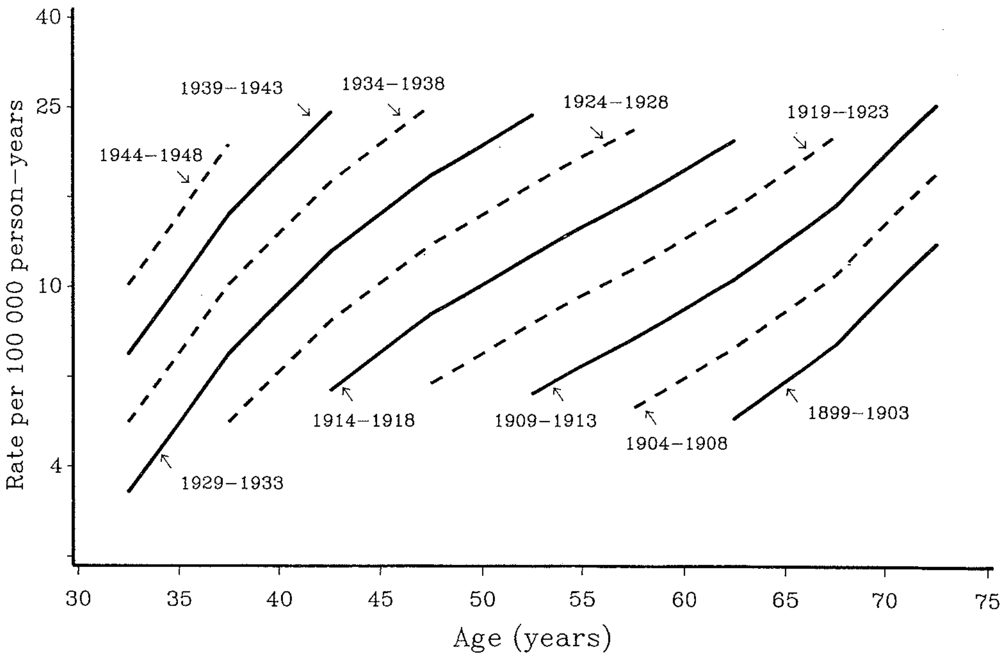


Figure 3.19 Skin melanoma incidence in Norway; estimated age-specific rate by birth cohort, women

curves are parallel. In certain situations, a graphical representation can often show to what extent this condition is fulfilled [72]. Thus, Figure 3.20 shows the time trend of lung cancer incidence in Scotland by age group according to calendar period (Figure 3.20a) and date of birth (Figure 3.20b). Diverging curves in Figure 3.20a clearly show the inadequacy of an age-period model. On the other hand, the parallel segments in the corresponding parts of the curves seen in Figure 3.20b suggest that an age-cohort model fits well.

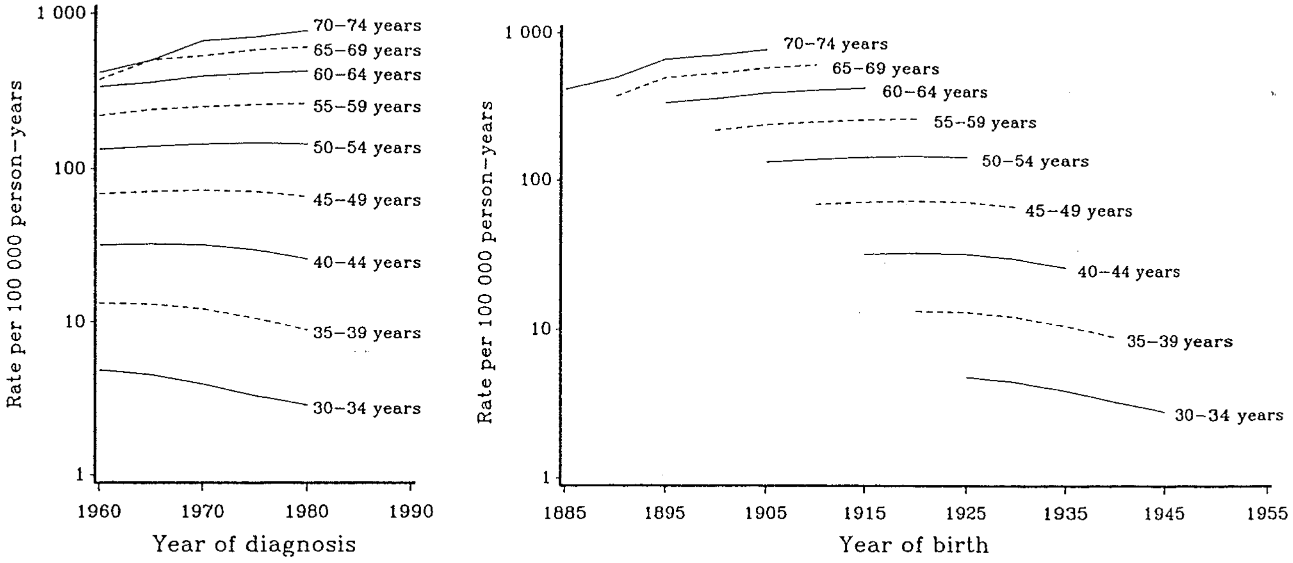


Figure 3.20 Trend in the age-specific incidence of lung cancer in Scotland; men

The age-cohort model is written:

$$\operatorname{Log}(\lambda_{xu}) = \alpha'_x + \gamma_u \qquad \begin{array}{l} x_1 \le x \le x_g \\ u_1 \le u \le u_\ell \end{array} \tag{3.79}$$

or, by writing explicitly the drift in the equation, as before (see formula 3.78):

$$\operatorname{Log}(\lambda_{xu}) = \alpha'_x + \beta u + \text{ non-linear terms in } u \tag{3.80}$$

As already mentioned, the use of this model is illustrated with data on lung cancer incidence in Scotland between 1964 and 1980 among men aged between 30 and 74 years (see Table 3.21).

Table 3.21 Incidence rates ([a]) of lung cancer in men in Scotland

Age	Registration period				Estimated rate ([b])
	1963-66	1970-72	1973-77	1978-82	u = 1925
30-34	4.07 (25)	4.29 (19)	3.68 (29)	3.08 (28)	4.80
35-39	15.14 (94)	9.55 (42)	11.00 (80)	7.12 (55)	13.09
40-44	29.94 (197)	29.21 (131)	26.41 (191)	21.97 (158)	32.22
45-49	73.38 (432)	73.40 (343)	69.52 (512)	59.71 (424)	71.59
50-54	143.91 (885)	140.38 (596)	143.31 (1070)	146.48 (1048)	143.54
55-59	245.12 (1483)	257.27 (1080)	265.39 (1729)	268.85 (1875)	259.72
60-64	372.42 (1923)	407.19 (1639)	410.38 (2618)	417.76 (2493)	424.10
65-69	448.37 (1654)	556.80 (1817)	589.29 (3305)	611.25 (3344)	624.92
70-74	407.39 (1015)	621.84 (1332)	743.46 (2969)	782.61 (3382)	831.00
WTR ([c])	5.27	7.52	10.75	15.35	21.93

[a] Observed rate per 100 000 person-years ; observed number in brackets.

[b] Age-specific rate estimated for the generation born in 1925. Rates underlined correspond approximately. to the ages for which this cohort is actually observed.

[c] Rates standardized to the truncated world population 30 to 74 years.

Note that the data used are not available at equidistant dates; it has therefore been necessary to reconstruct the cohorts, by dividing up the observation periods according to the cohorts that they include, and interpolating the corresponding person-years [73]. When there are three cohorts, the expectation of the observation k_{xt} for age x and time t can be written

$$E(k_{xt}) = \lambda_{x_1u_1} M_1 + \lambda_{x_2u_2} M_2 + \lambda_{x_3u_3} M_3$$

where:

- u_1, u_2, u_3 and x_1, x_2, x_3 are respectively the average birth dates and average ages in this period-time interval of the three cohorts spanning this observation period at age x;
- M_1, M_2, M_3 are the estimated person-years of observation in the corresponding sub-regions of the lexis diagram; and
- λ_{xu} is the incidence rate from the chosen model.

Estimation of the model is then straightforward using maximum likelihood as before. The likelihood based on the Poisson distribution is, apart from a constant term,

$$L = -\sum_{xt} \hat{k}_{xt} + \sum_{xt} k_{xt} \text{ Log}(\hat{k}_{xt}) \tag{3.81}$$

where $\hat{k}_{xt}$ is the value of k_{xt} estimated from the model.

For the data of table 3.21, the model:

$$\text{Log}(\lambda_{xu}) = \alpha(x) + \gamma(u) \tag{3.82}$$

where $\alpha(x)$ is a second-degree polynomial in x and $\gamma(u)$ a fifth-degree polynomial in u, provides a satisfactory fit (χ^2 = 24.8 on 28 degrees of freedom).

Incidence rates and observed numbers are given in Table 3.21, as well as age-specific rates estimated for the cohort born in 1925. Relative risks for other

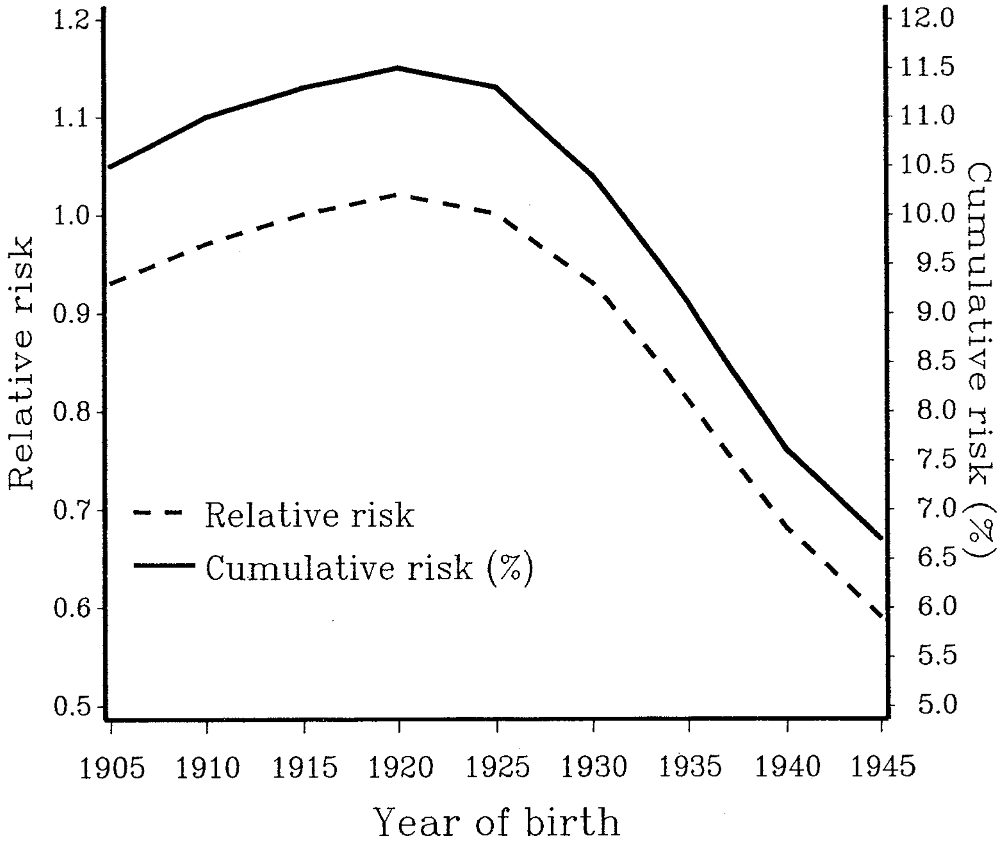


Figure 3.21 Trend in the risk of lung cancer in Scottish men born between 1905 and 1945

Table 3.22 Lung cancer risk ([a]) in Scotland by cohort, for men born between 1905 and 1945

	Year of birth								
	1905	1910	1915	1920	1925	1930	1935	1940	1945
Relative risk	0,93	0.97	1.00	1.02	1.00	0.93	0.81	0.68	0.59
Cumulative risk 30-74 years	10.50	11.00	11.30	11.50	11.30	10.40	9.10	7.60	6.70

([a]) Relative risk (reference 1925) and cumulative risk (%) are estimated from the age-cohort model

cohorts and corresponding cumulative risks from 30 to 74 years are given in Table 3.22 and Figure 3.21. Estimated rates corresponding to the observations are shown in Figure 3.22. This Figure shows the extent and the nature of the extrapolations carried out to obtain the cumulative risk for a given cohort.

In this example, a knowledge of the epidemiology of lung cancer would strongly suggest that risk has changed over successive cohorts. The fact remains that the fitting of a model, regardless of how good it is, does not prove whether an observed

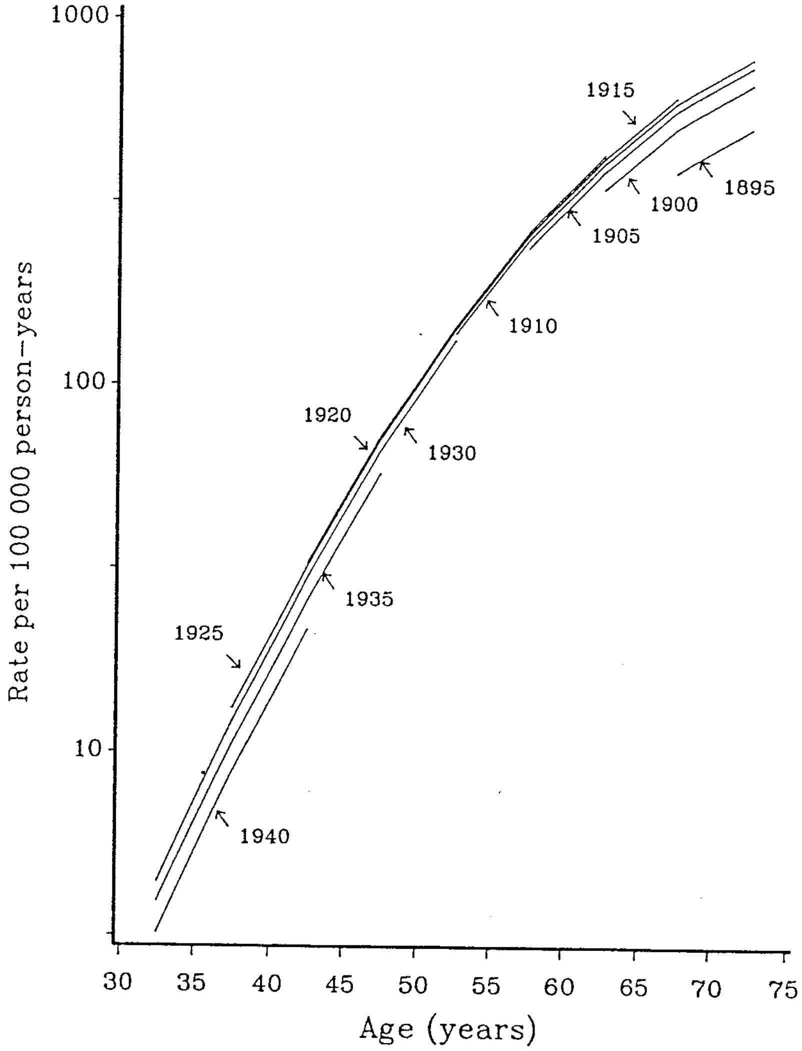


Figure 3.22 Estimated age-specific incidence of lung cancer in Scotland by birth cohort in men

effect is due to period or cohort. For instance, in this example, the absence of non-linear effects associated with period implies that the *a priori* hypothesis of a multiplicative age-cohort model (equation 3.82) can be accepted. Taken in isolation, the quality of the fit tells little about the validity of this last model.

Often, however, non-linear changes occur over time in factors related to period and cohort, necessitating their simultaneous introduction into the model; we are then led to consider age-period-cohort models.

Age-period-cohort models

We saw that when an age-period or age-cohort model describes the data well, it is possible to summarize the data simply, either by cross-sectional mortality or incidence rate and a series of standardized rates for each period, or by a longitudinal mortality or incidence rate and a series of cumulative risks for each cohort. Even if there is no ultimate proof of the models' validity, they provide a more or less full reconstruction of the information present in the data, and an accurate representation of the time trend. We have also seen that when the nature of the model is known *a priori,* estimates of the corresponding parameters can be obtained.

On the other hand, when neither of the two models is adequate, parameterization according to one or another of the time scales is no longer justified. Furthermore, even when it is known that an age-period-cohort model underlies the data, it is impossible to estimate all the parameters, because of the algebraic relationship between the three study factors (t = u + x). It has been proposed that the linear term, the drift, be partitioned according to the goodness of fit of the age-period and age-cohort models (74). Unfortunately, as has already been stated, goodness of fit only indicates the size of the contribution of the non-linear terms characterizing period or cohort changes, not their respective absolute size. Note, for example, that a perfectly linear cohort effect combined with a purely quadratic period effect leads to an age-period model with perfect fit.

To show its various forms, we write the age-period-cohort model in the form :

$$\text{Log}\,(\lambda_{xtu}) = \alpha_0 + \alpha x + a(x) + \beta t + p(t) + \gamma u + c(u) \qquad (3.83)$$

where a(x), p(t) and c(u) are the non-linear effects associated with age, period and cohort respectively. Thus written, this model is not identifiable, because t = u + x. It can be shown that two versions of this model are:

- the age-cohort model corrected for non-linear period effects, which, using the relationship $\beta t = \beta x + \beta u$, can be written:

$$\text{Log}\,(\lambda_{xtu}) = \alpha_0 + (\alpha + \beta)x + a(x) + (\gamma + \beta)u + c(u) + p(t) \qquad (3.84)$$

The linear coefficients of age and cohort are thus biased by β.

- the age-period model corrected for non-linear cohort effects, which, using the relationship $\gamma u = \gamma t - \gamma x$, can be written:

$$\text{Log}\,(\lambda_{xtu}) = \alpha_0 + (\alpha - \gamma)x + a(x) + (\beta + \gamma)t + p(t) + c(u) \qquad (3.85)$$

where the coefficient of the linear term in age is now biased by $-\gamma$. The coefficient of the linear term in period, $\beta + \gamma$, is the same as the coefficient corresponding to cohort in model (3.84). This coefficient (the drift) is the sum of the rates of change according to period and cohort: it is the linear approximation of the trend in the neighbourhood of the reference year of observation ($t = 0$) and year of birth ($u = 0$) respectively, when a, c and p are modelled by polynomials of degree greater than 1.

We illustrate the use of the age-period-cohort model with data on cervical cancer in Birmingham, UK between 1960 and 1982 (see Table 3.23). Figure 3.23, which shows that the trends in each age group are very different, shows that an age-period model is certainly not appropriate. Fitting the age-cohort model gives a deviance of 51.9 on 30 degrees of freedom, which becomes 38.0 on 27 degrees of freedom when period is added as a factor, a significant reduction ($p = 0.003$). Estimates from models (3.84) and (3.85) are given in Table 3.24. Following Holford [75], effects of each factor are presented by separating the overall linear trend from the 'non-linear' effects which correspond here to departures from linearity. This approach differs from the polynomial modelling used here in the age-cohort model, particularly

Table 3.23 Cervical cancer incidence ([a]) in women in Birmingham, UK, between 1960 and 1982

Age	Registration period					Estimated rate ([b]) u = 1920
	1960-62	1963-66	1968-72	1973-76	1979-82	
25-29	1.58 (7)	2.80 (17)	3.56 (31)	7.03 (55)	13.50 (91)	4.24
30-34	8.44 (40)	8.67 (51)	10.80 (82)	13.86 (90)	19.95 (149)	14.76
35-39	22.58 (117)	24.57 (153)	16.11 (118)	16.29 (98)	22.79 (151)	27.57
40-44	31.45 (154)	38.48 (263)	27.47 (214)	21.52 (126)	21.68 (128)	36.70
45-49	30.21 (150)	41.68 (265)	40.72 (338)	29.83 (183)	22.17 (125)	39.85
50-54	28.46 (136)	37.70 (243)	39.92 (312)	35.96 (239)	22.72 (134)	37.69
55-59	34.20 (148)	33.10 (202)	36.38 (277)	37.38 (209)	36.59 (225)	39.75
60-64	34.27 (130)	27.13 (146)	32.40 (231)	36.39 (208)	34.25 (189)	37.81
65-69	34.12 (106)	30.48 (137)	23.72 (146)	24.60 (127)	33.45 (173)	33.06
70-74	41.59 (101)	32.55 (111)	30.08 (148)	27.78 (118)	25.62 (114)	33.54
75-79	37.45 (65)	27.82 (67)	26.02 (89)	21.97 (66)	20.88 (70)	24.64

([a]) Observed rates for 100 000 person-years; observed numbers in brackets.

([b]) Age-specific rates estimated for the cohort born in 1920. Underlined rates correspond to the age intervals for which the cohort is actually observed.

with regard to the interpretation of the drift. In this case, it should be considered to be the best approximation to the linear change in incidence over the whole observation period. The drift is small ($\beta + \gamma = 0.01070$), because the decrease observed in some age groups is balanced by a substantial increase in other age groups. A polynomial model with cohorts centred around 1920 and periods around 1970 would give a much larger drift, given that the increase at these dates was already quite marked and that this version of the drift estimates local increases. It is important to note that, although it is identifiable, the drift depends essentially on the model selected, and it must be interpreted with care.

Fortunately, these subtleties are often irrelevant. In most situations, the structure of the time trend is much simpler and the different parameterizations are more or less equivalent. In the complex example considered here, change in risk across cohorts after correcting for linear effects of period (Table 3.25) still provides quite a satisfactory picture of the underlying epidemiological situation.

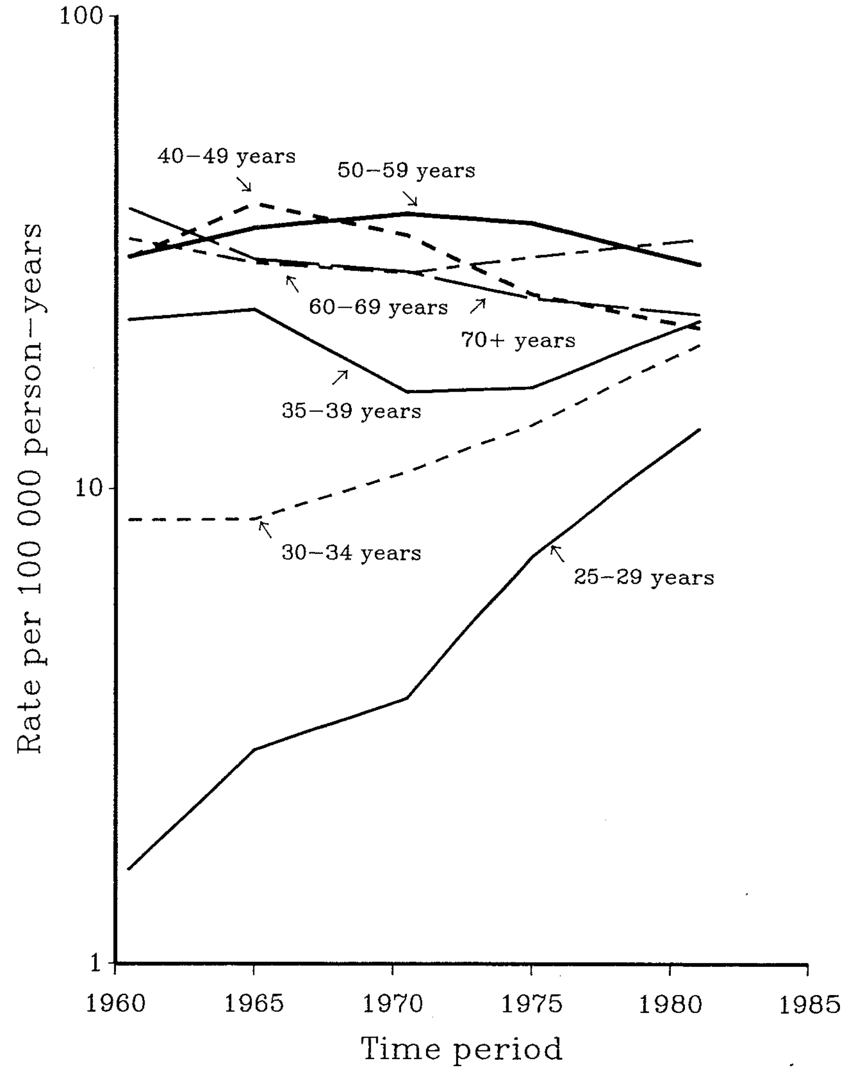


Figure 3.23 Trend in the age-specific incidence of cervical cancer in Birmingham, UK, between 1960 and 1982

**Table 3.24 Cervical cancer in Birmingham, UK.
Estimation of the age-period-cohort model**

Factor	Coding ([b])	Deviation from linearity	Total ([a])	
			Age-period	Age-cohort
Age	x	a(x)		
25	–5	–1.242	–1.765	–1.818
30	–4	–0.111	–0.529	–0.572
35	–3	0.399	0.085	0.053
40	–2	0.569	0.360	0.339
45	–1	0.537	0.432	0.422
50	0	0.366	0.366	0.366
55	1	0.304	0.409	0.419
60	2	0.138	0.347	0.368
65	3	–0.111	0.203	0.235
70	4	–0.212	0.206	0.249
75	5	–0.636	–0.113	–0.060
Linear effect			$\alpha - \gamma = 0.1046$	$\alpha + \beta = 0.1152$
Cohort	u	c(u)		
1885	–7	0.556		0.481
1890	–6	0.253		0.189
1895	–5	0.057		0.003
1900	–4	–0.070		–0.113
1905	–3	–0.224		–0.256
1910	–2	–0.211		–0.232
1915	–1	–0.071		–0.082
1920	0	–0.034		–0.034
1925	1	–0.052		–0.041
1930	2	–0.350		–0.329
1935	3	–0.599		–0.567
1940	4	–0.512		–0.469
1945	5	–0.186		–0.133
1950	6	0.317		0.381
1955	7	1.124		1.199
Period	t	p(t)		
1960-62	–2	–0.063	0.084	
1963-66	–1	0.054	0.043	
1968-72	0	0.040	0.040	
1973-76	1	0.012	0.023	
1979-82	2	–0.042	–0.021	
		$\alpha_0 = -8.249$		
Drift				$\beta + \gamma = 0.0107$

[a] The effect of the factor is obtained by summing the deviation from linearity and the linear effect corresponding to each of the models. Thus, the age effect at age 65 years (x = 3) in an age-cohort model corrected for nonlinear period effects is : 0.1152 × 3 – 0.111 = 0.235.

[b] Age, cohort and period variables are coded by corresponding integers, ignoring irregularities created by the observation periods. Age, cohort and period factors are centred around the categories 50-54, 1968-72 and 1920-25 respectively.

Table 3.25 Cervical cancer risk ([a]) in Birmingham, UK, by year of birth

	Year of birth												
	1890	1895	1900	1905	1910	1915	1920	1925	1930	1935	1940	1945	1950
Relative risk	1.20	1.00	0.89	0.77	0.79	0.92	0.97	0.96	0.72	0.57	0.63	0.87	1.46
Cumulative risk 25-79 years	1.99	1.66	1.47	1.28	1.31	1.52	1.59	1.58	1.19	0.94	1.03	1.44	2.41

([a]) Relative risk and cumulative risk (%) are estimated from the age-period-cohort model. Relative risks are normalized by the requirement that the sum of their logarithms is zero over the years considered.

Specific techniques and examples

Epidemiological evaluation of a secondary prevention campaign

The incidence of skin melanoma and associated mortality have shown a marked increase since the 1960s in most countries [73,76]. Some of this increase is most likely due to exposure to ultraviolet radiation, and another part can be attributed to improved diagnosis of these cancers. In theory, earlier detection of cases should limit the increase in mortality over time, or even reverse the trend. Accordingly, many countries or regions have developed intervention programmes, which in turn require evaluation. Even though secondary prevention programmes must ultimately be assessed on the basis of changes in mortality, the observation of larger increases in early-stage cases can also provide information on the effectiveness of the method of implementation of the programme.

A campaign conducted in Switzerland at the beginning of May 1988 had the twin objectives of primary prevention, aimed at educating the population about the dangers of prolonged exposure to the sun, and secondary prevention, through informing the public and the medical profession about the advantages of rapid and systematic examination (clinical and, if necessary, histological) of suspicious skin lesions. A year after this campaign was launched [77], only the second objective could be assessed. The ensuing analysis provides an example of the use of log-linear models to evaluate this type of chronological evolution.

The immediate objective of the campaign was to increase the number of cases diagnosed at an early stage, but it might also be expected that the number of advanced cases could also increase as a result of the intervention. The evaluation thus consisted of checking the assumption that the time trend prevailing before the campaign changed immediately after the launch of the campaign (that is, after June 1988), and that any increase was greater in early cases than in advanced cases.

For practical reasons, mainly related to the quality of cancer registration, data from before 1985 were not used to estimate the pre-campaign trend in incidence. Analysis was restricted to cases registered between 1 January 1985 and 30 April 1988 (three years and four months) and the campaign was assessed over the eight

Table 3.26 Skin melanomas by stage and calendar period in four Swiss registries

	Geneva		Neuchâtel		Vaud		St-Gall/Appenzell	
	Stage 1-2	Other ([a])	Stage 1-2	Other ([a])	Stage 1-2	Other ([a])	Stage 1-2	Other ([a])
1985								
Jan-Apr	6	6	–	3	12	7	7	5
May-Aug	9	3	3	–	14	17	7	4
Sep-Dec	9	4	3	1	13	8	8	5
1986								
Jan-Apr	9	4	1	1	13	14	5	5
May-Aug	11	5	–	–	10	13	12	9
Sep-Dec	7	7	4	2	14	13	10	7
1987								
Jan-Apr	8	5	3	1	15	12	4	6
May-Aug	17	7	1	4	22	10	6	12
Sep-Dec	5	4	3	–	7	16	7	6
1988								
Jan-Apr	5	3	4	3	9	8	4	4
May-Aug	20	12	9	3	23	12	17	12
Sep-Dec	12	4	4	4	24	17	5	5

([a]) Includes cases of unknown stage.

remaining months of 1988, when the effects of the intervention should have been apparent. In total, 734 skin melanomas were reported from January 1985 to December 1988 in the four participating regional registries (Geneva, Neuchâtel, St-Gall/Appenzell and Vaud). Given the short duration of the study period, it was not considered necessary to take denominators into account. On the other hand, monthly counts of cases were used, to allow for the effects of seasonal fluctuations.

In Switzerland, the melanoma incidence tends to increase markedly from the beginning of summer, and reach its lowest level during winter. It was decided *a priori* that a division of the year into three periods of four months (January to April, May to August and September to December) would provide a satisfactory description of the seasonal variation. Grouping into four-monthly periods also corresponded to the interval during which the effects of the campaign should have been noticeable, that is, the second and third periods of 1988. This grouping did not result in a significant loss of information compared to an analysis based on monthly data (χ^2 = 12.4 on nine degrees of freedom). All analyses were therefore carried out from data grouped in this way. For both practical and theoretical reasons, disease stages were also grouped. 'Early' cases were Breslow's stage 1 and 2 (up to and including 1.5 mm), while 'advanced' cases comprised those of stages 3 and 4 and unknown stage (7.9% of the total). Table 3.26 provides the data on which the analysis was based (see Table 3.27).

Table 3.27 Modelling of data from Table 3.26

Model	Estimate	Deviance	d.f.
Model A = Registry + Campaign + Year (continuous) + Four-month period		114.9	88
• Four-month period (a)			
Jan-Apr	1.00		
May-Aug	1.42 [1.16 ; 1.73]		
Sep-Dec	1.11 [0.90 ; 1.37]		
• Year (b)	2.30 [–6.60 ; 12.1]		
• Campaign			
Before campaign	1.00		
After campaign (c)	1.46 [1.13 ; 1.89]		
Model B = Model A + registry x campaign		111.0	85
Model C = Model A + stage		99.0	87
Model D = Model C + stage x campaign		96.5	86
Before campaign	1.00		
After campaign (c)			
– early stages	1.63 [1.22 ; 2.19]		
– autres stages	1.24 [0.90 ; 1.71]		

(a) Relative risk.
(b) Annual rate of increase (%).
(c) Relative increase in number of cases.

The first step in the analysis was to assess whether there had indeed been additional increase in incidence from the start of the campaign, taking into account the prior trend and seasonal variation. Trend was modelled using year of incidence as a continuous variable, with the four-monthly periods to represent seasonal changes. Region of registration was also introduced into this model as a factor to take into account both the differences between the size of the populations (denominators) and possible differences in the prevalence of the risk factors in the populations covered by the four registries (model A). The model expresses the logarithm of the expected number of cases as a linear function of the various factors:

$$\text{Log}\,[\mu_{r\,q\,c}\,(t)] = \alpha_r + \beta_q + \gamma_c + \delta t$$

where r, q, c are the indices of the region, the four-monthly periods and the campaign respectively, and where t is the year of incidence. The model was fitted by maximum likelihood assuming that the number of cases follows a Poisson distribution of mean $\mu_{r\,q\,c}(t)$. The result is an estimate of the overall effect of the campaign equal to 1.46 [1.13 ; 1.89], which means that incidence was 46% higher than expected on the basis of the pre-campaign trend and seasonal variation.

The second step was a comparison of the effectiveness of the prevention campaign in the four registry regions, by adding an interaction term (registry x campaign)

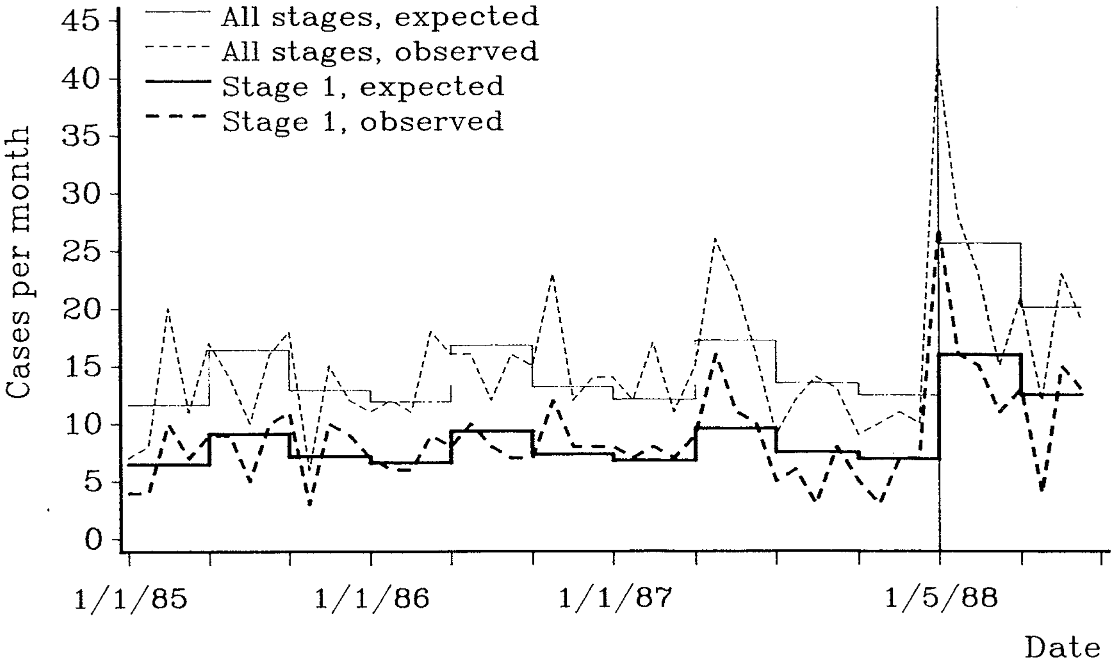


Figure 3.24 Observed and expected cases of skin melanoma before and after the start of a screening programme in Switzerland; both sexes combined; 1985–1988

to the above model, to allow for a possible different effect in each region (model B). The reduction in the deviance of 3.9 on three degrees of freedom led to the conclusion that there was no difference between the four regions with respect to the effect from the campaign.

The third step was to address the fundamental question as to whether the increase in incidence had been more marked for early stages. To test the hypothesis that the increase was identical for all stages, a model which included stage in addition to the other four factors included initially (registry, year, four-monthly period and campaign (model C)) was compared with a model augmented by an interactive term representing a campaign effect which differed for each stage (model D). The reduction in the deviance was 2.5 on one degree of freedom (p = 0.10). Despite the absence of a formal statistical significance at the 0.05 conventional level, the authors were convinced that the effect of the prevention campaign differed with respect to stage. The relative increase was estimated to be 1.63 [1.22 ; 2.19] in early cases and 1.24 [0.90 ; 1.71] in advanced cases, or 63% and 24% respectively. The campaign was therefore judged to be doubly effective on the basis of its first expected outcomes: (i) increased total incidence and (ii) a more marked increase in early cases.

The estimates obtained from fitting the final model (model D) provide the basis for calculating estimates which make up a smoothed curve (Figure 3.24). The number of expected cases can be calculated for any combination of values of the terms. For example, the number of cases over a whole year can be calculated by stage under the assumption that the prevention campaign either worked, or did not work.

In other words we can estimate the additional cases that were diagnosed during 1988, due to the effect of the screening campaign:

	Early stage	Other	Total
No screening	97	77	174
Screening	159	96	255
Additional cases	62	19	81
(% increase)	+ 64	+ 25	+ 47

Trends in cancer of the uterine cervix

In most western countries, the frequency of invasive cervical cancer has been decreasing for many years, almost certainly at least partly as a result of screening. However, a rise in incidence has recently been noted among young women in some countries. Various explanations have been offered, including an increase in sexual activity and the consequent increase in risk of infection by the human papilloma virus, an increase in the prevalence of smoking and decreased participation in screening programmes. Whatever the reasons for this phenomenon, it is of interest to examine the divergence by age of the time trend in different populations.

In Geneva, reliable incidence data are available from 1970. A study of time trends was first carried out on all invasive and microinvasive cases [78]. The time trend over the 18 years from 1970 to 1987 was analysed by modelling the logarithm of annual incidence rates by a linear function of year of diagnosis and estimating the parameters by maximum likelihood. Fitting the model

$$\text{Log}\,(\lambda_{xt}) = \alpha + \beta t$$

gave a rate of change of $\beta = -4.3\%$ per year [−6.0 ; −2.6], indicating a significant decrease in the crude incidence rate (Table 3.28, model B). The next step was to estimate the rate of change in the age-adjusted incidence from model (3.75):

$$\text{Log}\,(\lambda_{xt}) = \alpha_x + \beta t$$

which led to $\beta = -4.6\%$ [−6.3 ; −2.6] (model C).

The null hypothesis that the trends did not differ across age groups was tested by introducing a term for interaction between age group and year of diagnosis, which is equivalent to a different slope for each age group (model D) (see formula 3.74). Because of the significant improvement in the model's fit ($\chi^2 = 18.3$ on six degrees of freedom, $p < 0.05$), it was concluded that there was a real difference in trends between age groups, justifying different estimates of annual rates of change for each age group. These estimated rates of change are shown in Figure 3.25; estimates obtained by applying these rates of change to the incidence by age observed in 1970 are shown in Figure 3.26.

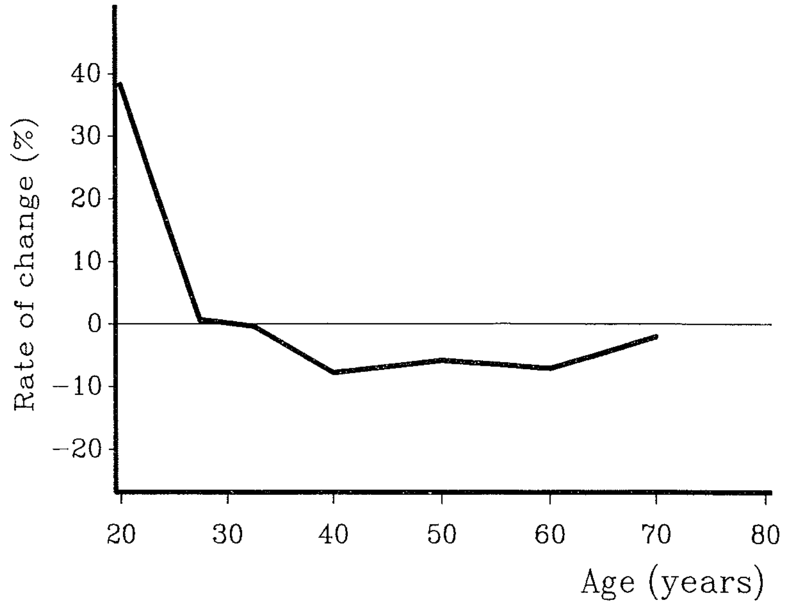


Figure 3.25 Age-specific rate of change in cervical cancer incidence in Geneva between 1970 and 1987

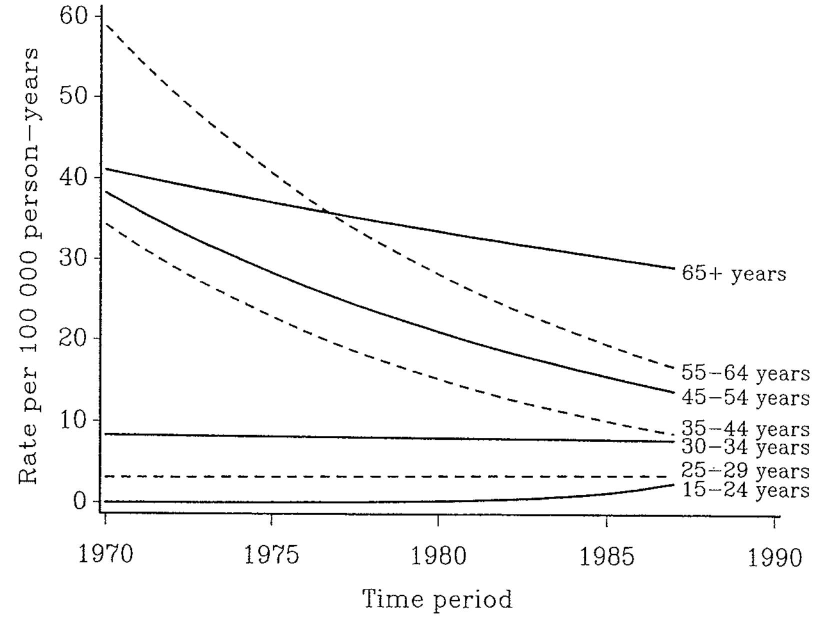


Figure 3.26 Trend in the age-specific incidence of cervical cancer in Geneva between 1970 and 1987

Although the numbers are small (480 cases observed in seven age groups over 18 years) and, consequently, the standard errors associated with the rates of change for each age group are high, the preceding analysis and data from Figures 3.25 and 3.26 suggest that there are three different types of time trend. The apparently increasing incidence for women less than 35 years could be a result of exposure to risk factors linked to sexual behaviour. In contrast, women aged 65 years and over, in whom incidence has only slightly decreased, might not have

Table 3.28 Modelling time trends by age group from annual rates of cancer of the uterine cervix in the canton of Geneva from 1970 to 1987 (all incident invasive and microinvasive cases)

Model	Rate of change %	95% CI	Deviance	d.f.
Model A = Constant			462.9	125
Log $(\lambda_{xt}) = \alpha$				
Model B = Year	−4.3	[− 6.0 ; −2.6]	438.3	124
Log $(\lambda_{xt}) = \alpha + \beta t$				
Model C = Year + Age	−4.6	[− 6.3 ; −2.6]	150.9	118
Log $(\lambda_{xt}) = \alpha_x + \beta t$				
Model D = Age * Year				
Log $(\lambda_{xt}) = \alpha_x + \beta_x t$				
15-24 (4 cases)	38.3	[− 2.0 ; 95.1]		
25-29 (9 cases)	0.7	[−11.2 ; 14.2]		
30-34 (22 cases)	−0.4	[− 8.4 ; 8.3]		
35-44 (84 cases)	−7.8	[−11.8 ; −3.8]		
45-54 (92 cases)	−5.8	[− 9.7 ; −1.9]		
55-64 (101 cases)	−7.1	[−10.7 ; −3.5]		
65+ (168 cases)	−2.0	[− 4.9 ; 1.0]	132.6	112

benefited from screening as much as younger women either because screening for this cohort was not yet routine or, more likely, because they stopped being screened after menopause. Incidence decreases substantially and relatively uniformly only in women aged between 35 and 65 years. Most of this change can reasonably be attributed to screening.

Bibliographical notes

A more detailed discussion of the concepts and methods of graphs and spatial analysis can be obtained from Cliff and Haggett's *Atlas of disease distributions : analytical approaches to epidemiological data* [79], effectively a manual of statistical ecology. While mainly using examples from the field of transmissible diseases, including the historical data of John Snow, the book also deals with problems relevant to cancer epidemiology, such as nasopharyngeal cancer in China, clusters of mesothelioma cases in the USA and monitoring risk around nuclear power plants or in the region of Chernobyl. The book reviews the principal techniques used to define regions and to smooth data, and also considers the problem of detecting outliers and clusters, both spatial and spatio-temporal. Also discussed are methods for detecting autocorrelation, estimating spatial patterns and regression involving exposure factors in ecological analyses.

Another text brings together a series of papers on cancer mapping, including presentations of the principal mortality atlases published at the time [80]. Several articles of this latter monograph discuss the various problems raised in the geographical representation of epidemiological data on cancer, or comment on methodological issues, such as the choice of colour.

The recent article by Walter and Birnie [81] provides a survey of the 49 atlases which appear during the fifteen-year period ending in 1989. The atlases are examined and classified by population and disease, and by the mapping and statistical techniques adopted. The authors emphasize the diversity of methods used and the consequent difficulty in making comparisons across atlases.

Research into the analysis of the spatial distribution of cancer, and in particular on the detection of clusters, has been published recently; two publications of note are the proceedings of the meetings organized by the Royal Statistical Society of the UK, on cancer incidence near nuclear installations [82], and the review of Marshall [83].

Applications of the empirical Bayes and grouping methods proposed by Huel [43] are presented in a thesis by Colonna on geographical studies in the situation where incidence is low [21]. This paper also deals with autocorrelation and its measurement. The thesis by Mollie includes a detailed mathematical discussion of smoothing based on the Bayesian approach, with an application to cancer mortality in France [48]. On the same subject, articles by Clayton and Bernardinelli [84] and Bernardinelli and Montomoli [85] provide an original point of view and practical examples.

The epidemiological literature includes many studies which have tried to link risk and exposure at the level of groups, mainly defined geographically. These studies provide examples of the methods dealt with on page 141 of this chapter. Of particular note are three studies on dietary factors which appeared at the time when ecological correlation analysis first became widely used, and which clearly illustrate the methodological problems raised by measurement of exposure at the group level. The first study relates to the geographical correlation observed in the USA (across states) and in Europe (across countries) between alcohol and tobacco consumption on one hand and various cancer sites on the other [86]. The second study examines the relationship between dietary factors and the various types of cancer, using national statistics from 32 countries [87]. The third article also considers dietary factors, but includes diseases other than cancer [88].

A thesis by Viel compares the results of published case-control and cohort studies on the effect of pesticides with those that he obtained from ecological analysis of French data from the departements. These analyses are carried out using the method proposed by Gardner [51], a Poisson regression adjusting for latitude and longitude, and a correlation test modified to take into account autocorrelation, proposed by Clifford and coworkers [90]. The work provides a good example of geographical correlation methods applied to the study of an association involving an exposure which is difficult to quantify at the individual level.

There have been a number of studies published on cancer risk in migrants. Recent monographs published by the International Agency for Research on Cancer have considered Jewish migrants to Israel [91] and Italian migrant populations [60]. The first of these monographs is an excellent example of the use of information on the country of birth and time since arrival, in a country characterized by immigration from many countries. The second only considers one country of origin, Italy, but studies their outcome in a range of host countries.

Data on time trends in cancer incidence and mortality are essential for the development of public health policy. For this reason, it is surprising that the literature in this area is relatively poor. There has been little research on the simultaneous estimation of rates of change having variable precision. There is however a need for methods to allow data of this type to be presented in a more convincing manner. The only work in this area has been based on empirical Bayes methods, particularly in the estimation of cohort effects in the youngest and oldest cohorts. Breslow and Clayton have proposed the estimation of random effects based on autoregressive models, in which the estimate for each cohort is based on some information from earlier and later cohorts [92]. In contrast, Desouza has used data on the trend in several geographical areas, to estimate cohort effects in each area by making use of information from other study areas [93]. These methods have nonetheless been used very little, and their value in practice is still unknown. The current rate of progress in the analysis of longitudinal data suggests that there will be a rapid improvement in this situation [94].

The majority of research on time trends has involved relatively simple methods. This lack of sophistication is undoubtedly justified both by the lack of suitable computer software, and by the desire to publish observed data with only a minimum of smoothing compatible with the needs of graphical presentation. Research in this area has been published by Hakulinen and coworkers, on trends in cancer incidence in Nordic countries [95]; by Osmond and coworkers for trends in cancer mortality in England and Wales during 1951-80 [96], by Devesa and coworkers who carried out a fairly complete survey of trends of cancer incidence and mortality in the USA [97], by Lee and coworkers for trends in cancer incidence in Singapore [98], Hill and coworkers for those in France [64], La Vecchia and coworkers for Europe [99] and, finally, Coleman and coworkers who reviewed trends in cancer incidence and mortality using the data available from all five continents [73].

For a general discussion of methodological problems in the study of time trends, in particular those which are not statistical, two meeting reports may be of value [100,101].

REFERENCES

[1] Atlas graphique et statistique de la suisse. Berne, Bureau de statistique du Département fédéral de l'intérieur, 1914. (Statistique de la Suisse; 191^{e} livraison)

[2] CISLAGHI C, DECARLI A, LA VECCHIA C, LAVERDA N, MEZZANOTTE G, SMANS M. *Dati. Indicatori e mape di mortalita tumorale : Italia 1975/1977.* Bologna, Pitagora Editrice,1986

[3] Smans M, Muir CS, Boyle P. *Atlas of cancer mortality in the European Community.* (IARC Scientific Publications, No 107), Lyon, IARC, 1992

[4] Mason TJ, McKay FW, Hoover R, et al. *Atlas of cancer mortality for US counties : 1950-1969.* Bethesda, US Department of Health, Education and Welfare, National Cancer Institute, 1975, (DHEW Publication ; No (NIH)75-780)

[5] *In* Cliff AD, Ord JK, (eds). *Spatial processes : models & applications.* London, Pion, 1981, pp. 1-6

[6] Teppo L, Pukkala M, Hakama M *et al. Way of life and cancer incidence in Finland.* Helsinki, Finnish Cancer Registry, 1980, p. 18 ; pp. 40-44

[7] Carstairs V, Lowe M. Small area analysis : creating an area base for environmental monitoring and epidemiological analysis. *Community Med* 1986, **8** :15-28

[8] Dirichlet GL. Über die Preduction der positiven quadritschen Formen mit drei unbestimmten ganzen Zahlen. *J Reine Angew Math* 1850, **40** : 209-34

[9] Cislaghi C, Decarli A, La Vecchia C, Mezzanotte G, Vigotti MA. Trends surface and models applied to the analysis of geographical variations in cancer mortality. *Rev Epidémiol Santé Publ* 1990, **38** : 57-69

[10] Atlas of cancer mortality in the People's Republic of China : China map press n° 1358. Shanghai, Yan'an Xilu, 1979

[11] Kemp I, Boyle P, Smans M, Muir C. *Atlas of cancer in Scotland 1975-1980 : incidence and epidemiological perspectives.* (IARC Scientific Publications, No 72), Lyon, IARC, 1985

[12] Smans M, Estève J. Practical approaches to disease mapping. *In* P Elliott, J Cuzick, D English, R Stern (eds) : *Geographical and environmental epidemiology : methods for small area studies.* Oxford, Oxford University Press, 1992, pp. 141-150

[13] Rezvani A, Doyon F, Flamant R. *Atlas de la mortalité par cancer en France.* Paris, Les Éditions INSERM, 1986

[14] Ryckerboer R, Janssens G, Thiers G. *Atlas de la mortalité par cancer en Belgique (1969-1976).* Brussels, Ministère de la santé publique et de la famille, Institut d'hygiène et d'épidémiologie, 1983

[15] Gardner MJ, Winter PD, Taylor CP, Acheson ED. *Atlas of cancer mortality in England and Wales 1968-1978.* New York, John Wiley, 1983

[16] Moran PAP. The interpretation of statistical maps. *J R Statist Soc B* 1948, **10** : 243-51

[17] Geary RC. The contiguity ratio and statistical mapping. *Incorpor Statist* 1954, **5** : 115-45

[18] Ohno Y, Aoki K. Cancer deaths by city and county in Japan 1959-1971 : a test of significance for geographic clusters of disease. *Soc Sci Med* 1981, **15** : 251 : 8

[19] Smans M. Analysis of spatial aggregation. *In* P Boyle, CS Muir, E Grundmann (eds) : *Cancer mapping.* Berlin, Springer, 1989, pp. 83-6. (*Recent Results in Cancer Research,* **114**)

[20] Knox G. Epidemiology of childhood leukaemia in Northumberland and Durham. *Br J Prev Soc Med* 1964, **18** : 17-24

[21] Colonna M. *Analyse de la distribution spatiale du cancer : problème posé par l'étude de faibles incidences* (Dissertation). Grenoble, Université des sciences sociales de Grenoble, Département informatique et mathématiques en sciences sociales, Laboratoire de statistique et analyse de données, 1991, 224 p.

[22] Gardner MJ. Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the UK. *J R Statist Soc A* 1989, **152** : 307-25

[23] Pocock SJ, Cook DG, Beresford SAA. Regression of area mortality rates on explanatory variables : what weighting is appropriate ? *Appl Statist* 1981, **30** : 286-95

[24] Schulman J, Selvin S, Merrill DW. Density equalized map projections : a method for analysing clustering around a fixed point. *Stat Med* 1988, **7** : 491-505

[25] Stone RA. Investigations of excess environmental risks around putative sources : statistical problems and a proposed test. *Stat Med* 1988, **7** : 649-60

[26] Cook-Mozaffari PJ, Darby SC, Doll R, *et al.* Geographical variation in mortality from leukaemia and other cancers in England and Wales in relation to proximity to nuclear installations : 1969-78. *Br J Cancer* 1989, **59** : 476-85

[27] Jablon S, Hrubec Z, Boice JD. Cancer in populations living near nuclear facilities. *J Am Med Ass* 1991, **265** : 1403-8

[28] Kinlen LJ. Evidence for an infective cause of childhood leukaemia : comparison of a Scottish new town with nuclear reprocessing sites in Britain. *Lancet* 1988, **2** : 1323-27

[29] Kinlen LJ. The relevance of population mixing to the aetiology of childhood leukaemia. *In* WA Crosbie, JH Gittus (eds) : *Medical response to effects of ionising radiation.* London, Elsevier, 1989, pp. 272-8

[30] Potthoff RF, Whittinghill M. Testing for homogeneity. I. The binomial and multinomial distributions. *Biometrika* 1966, **53** : 167-82

[31] Potthoff RF, Whittinghill M. Testing for homogeneity. II. The Poisson distribution. *Biometrika* 1966, **53** : 183-90

[32] Muirhead CR, Ball AM. Contribution to the discussion of the Royal Statistical Society meeting on cancer near nuclear installations. *J R Statist Soc A* 1989, **152** : 376-77

[33] Muirhead CR, Butland BK. Testing for over-dispersion using an adapted form of the Pothoff-Wittinghill method. *In* FE Alexander, P Boyle (eds) : *Detecting localized clusters of disease.* Lyon : IARC. (IARC Scientific Publications, in preparation)

[34] *In* Cliff AD, Ord JK (eds). *Spatial processes : models & applications.* London, Pion, 1981 : p. 97

[35] Urquhart J, Black R, Buist E. Exploring small area methods. *In Methodology of enquiries into disease clustering.* London, London School of Hygiene and Tropical Medicine, 1989 : pp. 41-52

[36] Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. *J R Statist Soc B* 1990, **52** : 73-104

[37] Besag J, Newell J. The detection of clusters in rare diseases. *J R Statist Soc A* 1991, **154** : 143-55

[38] David FN, Barton DE. Two space-time interaction tests for epidemicity. *Br J Prev Soc Med* 1966, **20** : 44-8

[39] Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967, **27** : 209-20

[40] *In* Cliff AD, Ord JK (eds). *Spatial processes : models & applications.* London, Pion, 1981 : 22-4

[41] Suomen Syöpäkartasto. *Atlas of cancer incidence in Finland : 1953-82.* Helsinki, Finnish Cancer Registry, 1987

[42] Ménégoz F, Colonna M, Lutz JM, Schaerer R. *Atlas du cancer dans le département de l'Isère.* Grenoble, Registre des cancers de l'Isère, 1989

[43] Huel G, Petiot JF, Lazar P. Algorithm for the grouping of contiguous geographical zones. *Stat Med* 1986, **5** : 171-81

[44] Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987, **43** : 671-81

[45] Besag J. Spatial interaction and the statistical analysis of lattice systems. *J R Statist Soc B* 1974, **36** : 192-936

[46] Mollie A, Richardson S. Empirical Bayes estimates of cancer mortality rates using spatial models. *Stat Med* 1991, 10 : 95-112

[47] Besag J, Kempton R. Statistical analysis of field experiments using neighbouring plots. Biometrics 1986, 42 : 231-51

[48] Mollié A. Représentation géographique des taux de mortalité : modélisation spatiale et méthodes bayesiennes [Dissertation]. Paris, Université Paris 6, 1990

[49] Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health* 1982, **72** : 1336-44

[50] Durkheim E. *Suicide: a study in sociology.* New York, Free Press, 1951, pp. 153-80

[51] Gardner MJ, Winter PD. Extensions to a technique for relating mortality and environment – exemplified by nasal cancer and industry. *Scand J Work Environ Health* 1984, **10** : 219-23

[52] Hakama M, Hakulinen T, Pukkala E, Saxén E, Teppo L. Risk indicators of breast and cervical cancer on ecologic and individual levels. *Am J Epidemiol* 1982, **116** : 990-1000

[53] Blot WJ, Fraumeni JF. Geographic patterns of lung cancer : industrial correlations. *Am J Epidemiol* 1976, **103** : 539-50

[54] Richardson S. A method for testing the significance of geographical correlations with application to industrial lung cancer in France. *Stat Med* 1990, **9** : 515-28

[55] Cook DG, Pocock SJ. Multiple regression in geographical mortality studies with allowance for spatially correlated errors. *Biometrics* 1983, **39** : 361-71

[56] Mardia KV, Marshall RJ. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrica* 1984, **71** : 135-46

[57] Buell P, Dunn JE. Cancer mortality among Japanese Issei and Nisei of California. *Cancer* 1965, **8** : 656-64

[58] Cairns J. The cancer problem. *Sci Am* 1975, 233 : 64-72 and 77-8

[59] Philips RL. Role of lifestyle on dietary habits in risk of cancer among seventh days adventists. *Cancer Res* 1975, **35** : 3513-22

[60] Geddes M, Parkin DM, Khlat M, Balzi D, Buiatti E. *Cancer in Italian migrant populations.* (IARC Scientific Publications, No 123), Lyon, IARC, 1992

[61] Buiatti E, Palli D, Decarli A et al. A case-control study of gastric cancer and diet in Italy. *Int J Cancer* 1989, 44 : 611-6

[62] Devesa SS, Schneiderman MA. Increase in the number of cancer deaths in the United States. *Am J Epidemiol* 1977, **106** : 1-5

[63] Yale University. Connecticut Cancer Epidemiology Unit. Forty-five years of cancer incidence in Connecticut : 1935-79. Bethesda, US Department of Health and Human Services, National Cancer Institute, 1986 (National Cancer Institute monograph, 70)

[64] Hill C, Benhamou E, Doyon F, Flamant R. *Évolution de la mortalité par cancer en France : 1950-1985.* Paris, Les Editions INSERM, 1989

[65] Cuzick J, Velez R, Doll R. International variations and temporal trends in mortality from multiple myeloma. *Int J Cancer* 1983, **32** : 13-9

[66] Kupper LL, Janis JM, Karmous A. Statistical age-period-cohort analysis : a review and critique. *J Chron Dis* 1985, **38** : 811-30

[67] Hinde J. Compound Poisson regression models. *In* : R Gilchrist, (ed) : *GLIM 82 : Proceedings of the international conference on generalised linear models.* New York , Springer Verlag, 1982, pp. 109-21

[68] Breslow N. Extra-Poisson variations in log-linear models. *Appl Statist* 1984, **33** : 38-44

[69] Breslow N. Score tests in overdispersed GLM's. *In* : M Decarli et al. (eds) : *Proceedings of GLIM 89 and the Fourth international workshop on statistical modelling.* New York, Springer, 1989, pp. 64-74

[70] Clayton D, Schifflers E. Models for temporal variation in cancer rates. I. Age-period and age-cohort models. *Stat Med* 1987, **6** : 449-67

[71] Clayton D, Schifflers E. Models for temporal variation in cancer rates. II. Age-period-cohort models. *Stat Med* 1987, **6** : 469-81

[72] Roush GC, Schymura MJ, Holford TR, White C, Flannery JT. Time period compared to birth cohort in Connecticut incidence rates for twenty-five malignant neoplasms. *J Nat Cancer Inst* 1985, **74** : 779-88

[73] Coleman M, Esteve J, Damiecki P. et al. Time trends in cancer incidence and mortality. (IARC Scientific Publications, No 121) Lyon : IARC, 1993

[74] Osmond C, Gardner MJ. Age-period and cohort models applied to cancer mortality rates. *Stat Med* 1982, **1** : 245-59

[75] Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics* 1983, **39** : 311-24

[76] Muir CS, Nectoux J. Time trends : malignant melanoma of skin. *In* K Magnus (ed) : *Trends in cancer incidence : causes and practical implications.* Washington, Hemisphere Publ., 1982 : pp. 365-85

[77] Bulliard JL, Raymond L, Levi F et al. Évaluation épidémiologique préliminaire de la campagne suisse pour la prévention du mélanome malin. *Med Hyg* 1990, **48** : 370-4

[78] Bouchardy C, Fioretta G, Raymond L, Vassilakos P. Age differentials in trends of uterine cervical cancer incidence from 1970 to 1987 in Geneva. *Rev Epidémiol Santé Publ* 1990, **38** : 261-2

[79] Cliff AD, Haggett P. *Atlas of disease distributions : analytic approaches to epidemiological data.* Oxford : Basil Blackwell, 1988

[80] Boyle P, Muir, CS, Grundmann E (eds) *Cancer Mapping.* Berlin, Springer, 1989. (*Recents results in cancer research*, **114**)

[81] Walter SD, Birnie SE. Mapping mortality and morbidity patterns. *Int J Epidemiol* 1991, **20** : 678-689

[82] Royal Statistical Society meeting on cancer near nuclear installations. *J R Statist Soc A* 1989, **152** : 305-89

[83] Marshall RJ. A review of methods for the statistical analysis of spatial patterns of diseases. *J R Statist Soc A* 1991, **154** : 421-41

[84] Clayton D, Bernardinelli L. Bayesian methods for mapping disease risk. *In* P Elliott, J Cuzick, D English, R Stern (eds) : *Geographical and environmental epidemiology : methods for small area studies.* Oxford, Oxford University Press, 1992

[85] Bernardinelli L, Montomoli C. Empirical Bayes versus fully Bayes analysis of geographical variation in disease risk. *Stat Med* 1992, **11** : 983-1007

[86] Breslow NE, Enstrom JE. Geographic correlations between cancer mortality rates and alcohol-tobacco consumption in the United States. *J Natl Cancer Inst* 1974, **53** : 631-9

[87] Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *Int J Cancer* 1975, 15 : 617-31

[88] Knox EG. Foods and diseases. *Br J Prev Soc Med* 1977, 31 : 71-80

[89] Viel JF. Étude des associations géographiques entre mortalité par cancers en milieu agricole et exposition aux pesticides. [Dissertation]. Paris : Université Paris XI, Faculté de Médecine Paris Sud, 1992

[90] Clifford P, Richardson S, Hémon D. Assessing the significance of the correlation between two spatial processes. *Biometrics* 1989, 45 : 123-34

[91] Steinitz R, Parkin DM, Young JL, Bieber CA, Katz L. *Cancer incidence in Jewish migrants to Israel, 1961-1981.* (IARC Scientific Publications, No 98), Lyon, IARC, 1989

[92] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Technical Report n° 106. Seattle WA : Department of Biostatistics School of Public Health, 1991 (Technical report; n° 106)

[93] Desouza CM. An empirical Bayes formulation of cohort models in cancer epidemiology. *Stat Med* 1991, 10 : 1241-56

[94] Zeger SL, Karim MR, Generalized linear models with random effects – A Gibbs sampling approach. *J Am Stat Assoc* 1991, **86** : 79-86

[95] Hakulinen T, Andersen AA, Malker B et al. *Trends in cancer incidence in the Nordic countries.* Helsinki : The Nordic Cancer Registries, 1986

[96] OSMOND C, GARDNER MJ, ACHESON ED. Analysis of trends in cancer mortality in England and Wales during 1951-80 separating changes associated with period of birth and period of death. *Br Med J* 1982, **284** : 1005-8

[97] DEVESA S, SILVERMAN D, YOUNG JL et al. Cancer incidence and mortality trends among whites in the United States, 1947-84. *J Natl Cancer Inst* 1987, **79** : 701-70

[98] LEE HP, DAY NE, SHANMUGARATNAM K. Trends in cancer incidence in Singapore 1968-82. (IARC Scientific Publications, No 91), Lyon, IARC, 1988

[99] LA VECCHIA C, LUCCHINI F, NEGRI E, BOYLE P, MAISONNEUVE P, LEVI F. Trends of cancer mortality in Europe, 1955-1989. I. Digestive sites. *Eur J Cancer* 1992, **28** : 132-235

[100] MAGNUS K, ed. *Trends in cancer incidence. Causes and practical implications.* Washington DC : Hemisphere Publishing Corp, 1982

[101] DAVIS DL, HOEL D. Trends in cancer mortality in industrial countries. *Ann NY Acad Sci* 1990, **609**, New York