Paper 29-25

An Application for the Analysis and Visualization of Disease Incidence Data

Paul Nicholson - University of Leeds

Co-authors: Deborah Allon, Estelle Gilman
University of Leeds

Abstract

Descriptive Epidemiology is concerned with the analysis of the incidence of diseases both geographically and over time. A characteristic of this field of study is the need for access to diverse types of data. These include data on disease incidence, population data from the region of study, data relating to diseases and map data sets at various geographical levels. The ability to manage such diverse data in a way which allows easy and efficient access to different subsets of data is an essential preliminary requirement to subsequent analysis.

Preliminary investigations usually consist of the tabulation of age-specific incidence rates for each geographical region and the calculation of standardised rates, typically presented as standardised morbidity ratios (SMRs). Subsequent investigations include the analysis of variations in incidence rates with age, sex and geographical region, tests for spatial clustering and the use of maps to display SMRs. Thus, in addition to data management, an equally important requirement is access to a variety of analytical tools, ranging from simple tabulation facilities to sophisticated statistical and numerical techniques.

This paper describes the development of a general application for use in Descriptive Epidemiology and illustrates its use by epidemiologists based at the University of Leeds.

Introduction

The development of this application arose following interest expressed by the Leukaemia Research Fund Centre for Clinical Epidemiology at Leeds (LRF), in the development of an integrated system for the management, analysis and display of data relating to the incidence of leukaemias and lymphomas in England and Wales.

The LRF have been engaged in the production of disease atlases of leukaemia in England and Wales since 1990. The majority of their work has been based on a DEC Alpha/AXP 2100 running under OpenVMS, using a variety of programs written in C, FORTRAN, PASCAL and GENSTAT for analysis and ARC/INFO for mapping.

This system provides a purely batch style of working. Also, communication between the different programs is via intermediate results stored in external files. The availability of the SAS® system at the University was seen by the LRF as providing a means of integrating the operations involved in the production of the Leukaemia Disease Atlases under the control of a single system. In addition, SAS was seen as having the potential for opening up interactive ways of working, more relevant to exploratory analysis than the batch style of working.

Accordingly, a development project was started in 1996, with backing from SAS Institute, with the aim of transferring the Atlas production to the SAS system. Early on in the discussions with the users, it became evident that many of their requirements are common to epidemiologists working on other disease studies. It was decided, therefore, that the aims of the project should be widened - that the primary aim should be the development of an application suitable for use in any disease study whilst addressing the needs of the LRF in particular.

Data

A variety of types of data are required in descriptive epidemiological studies. The major types are:

- Case data
  - this could be either morbidity data (disease occurrences) or mortality data.

- Population data
  - this is usually derived from the national census and provides population estimates broken down by sex and age group for areas included in the study.
- Coverage data
  - indicating the time periods during which regions participated in the study and details of any sub-regions not participating.
- Disease data
  - consisting of names of diseases and details of disease group hierarchies.
- Spatial data
  - consisting of digitised map boundary data for the study region.
- Area data
  - consisting of names and codes for geographical areas and other area-related items of data.
- Study data
  - descriptive information for each study handled by the application.

Data Model

The choice of an efficient and flexible data model was seen as a central requirement of this application. Extensive discussions with the users revealed a number of particular features of the data and operational requirements for accessing data which placed practical constraints on the choice of data model.

A number of data models, based on relational database theory, were considered. In a perfect relational model, only the lowest stratum, from which other summary tables can be calculated, should be stored. This approach has serious shortcomings in this application since, typically, population figures are made available at the highest level first. It can take considerable time after a census to release full ward and enumeration district (ED) level population figures. A perfect relational model would therefore result in unacceptable delays in research.

Data modelling was complicated further by the existence in the LRF Data Collection Study of so called 'part areas' - regions which either did not fully contribute to the study or which contributed for only a part of the study period. In order to obtain population data based upon the same contributing areas as the case data, it was, therefore, necessary to include in the model an entity to permit the construction of population data sets adjusted for the existence of part areas. Thus, the model chosen is based on relational theory but is expanded to include summary tables at all geographical levels for all area related entities.

For a detailed description of the data model and details of the implementation of data access mechanisms see Allon and Nicholson (1997).

The Disease Registry Application

The Disease Registry Application is an application based upon SAS® software for the management and analysis of epidemiological data stored in a disease registry. The application employs the specially constructed data model, described above, to store and access the many categories of data common to descriptive epidemiological studies and provides facilities for analysis, graphics and mapping. The application is data set driven and may be applied to any disease study.

Current capabilities of the application include:

- Calculation of age-specific disease rates
- Plots of age-specific disease rates
- Calculation of directly standardised morbidity rates
- Calculation of indirectly standardised morbidity rates (SMRs)
- Calculation of directly standardised morbidity rates
- Tests for geographical variation in SMRs
- Smoothing of SMRs using a variety of methods
- Mapping of raw and smoothed estimates of SMRs
- Tests for spatial clustering

User Interfaces

Figure 1 shows the Primary Interface of the application. (The screen has been designed specifically for the LRF but may easily be customised for other end-users).

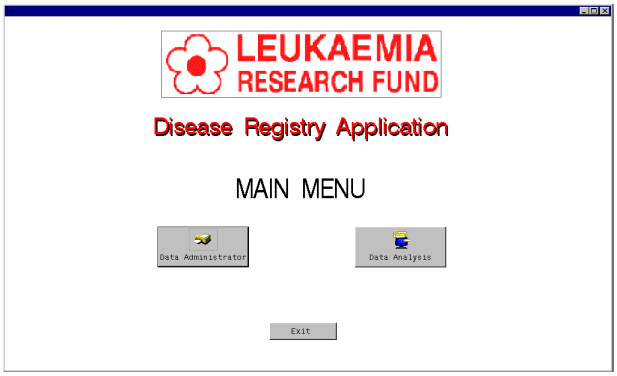


Figure 1. Primary Menu

Before the application can be used, the data needs to be loaded into SAS permanent data sets. The **Data Administrator** menu (under development) provides access to tools for loading data and modifying existing data. This facility is designed for use solely by a Data Administrator.

The **Analysis** menu, shown in Figure 2, provides access to all of the analysis and graphics tools.

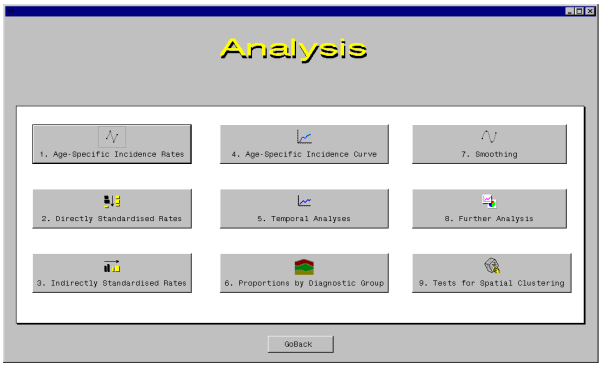


Figure 2. Analysis Menu

Apart from the high-level menus, the majority of the user interfaces make extensive use of the TAB object. An early version of the application (see Nicholson and Allon (1997)), made use of a hierarchical sequence of frames for data selection. However, the introduction of the TAB object in SAS 6.12 prompted a re-think in the design of the user interface. The use of the Tab Layout allows the variety of screens required for data selection to be organised within a single frame. For further details of the use of the TAB object in user interfaces, see Allon (1999).

Performing an Analysis

Most analyses consist of two distinct stages - a **data selection** stage and an **analysis** stage. The data selection stage is entered automatically on selection of the required analysis. Data selection screens serve two purposes. They serve to identify the data required for analysis and they are also used to determine a unique storage location for the results of the analysis. If the results of an analysis already exist, the data selection stage serves to locate those results.

Data Selection

Data selection consists of a first phase, common to all analyses, followed by a possible second phase specific to the analysis being performed.

The first phase requires the user to specify a study, a case data set, a disease group of interest, demographic data such as age-groups and study dates and (optionally) a geographical subset. A single TAB object is used for this phase.

The order in which data is selected is controlled by the application. The first step is to specify a study required for analysis. Having selected a study, the next step is to specify a case dataset. The first two tabs labelled **Study** and **Case Dataset** cater for these selections.

Having selected a study and a case data set, the tab labelled **Disease Group**, displayed in Figure 3, will become un-grayed, allowing a disease group to be specified.

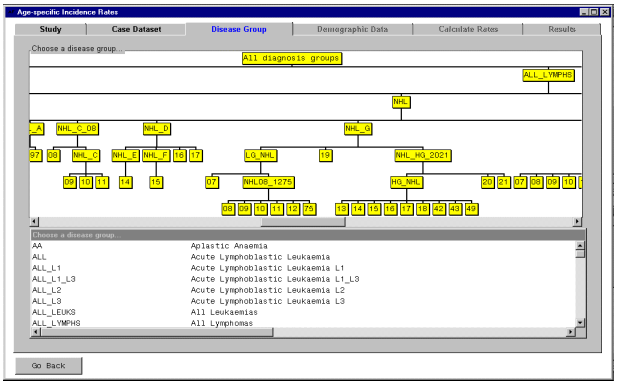


Figure 3. Disease Group Selection Screen

The list box displays the available set of disease groups, in alphabetical order, with their full descriptions alongside. Linked to the listbox is an organisational chart displaying the hierarchical structure of the disease groups. The numeric codes represent individual diseases. The named nodes represent disease groups.

The user is free to select the required group from either the listbox or the chart. The value selected is automatically communicated and displayed in both objects.

Once a disease group has been selected, the tab labelled **Demographic Data** becomes un-grayed. The user must specify the demographic data required before the analysis can proceed. The demographic data selection screen is displayed in Figure 4.

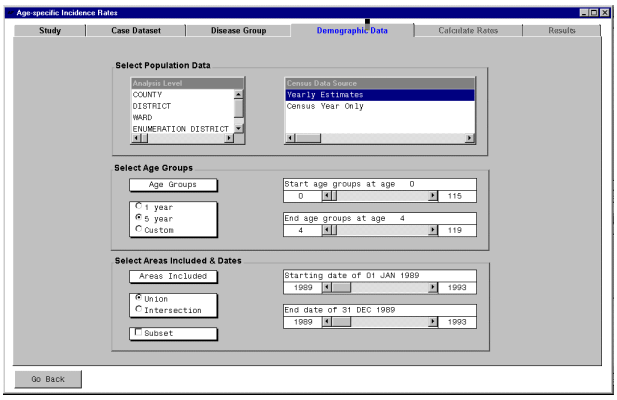


Figure 4. Demographic Data Selection Screen

Select Population Data requires the user to specify the geographic level of analysis (e.g. County) and the Census Data Source. The latter determines the type of population estimates to be used. Currently, two types of population estimates are supported - **Yearly** and **Census Year Only.**

Yearly estimates are population estimates for individual years in the study period, derived from the Census by demographic modelling. Census Year Only figures are population counts observed for the actual year of the census.

Select Age Groups is used to specify the size of the age group (default is 5 years) and the age range for the study. A 'Custom' option caters for the specification of non-standard age-groups.

Select Areas Included & Dates is used to specify the time period for the analysis and the criterion for inclusion of data. Two criteria are available - referred to as Union and Intersection. A union includes data for all areas contributing to the study at any time between the specified start and end dates. An intersection restricts analysis to data taken from the largest possible coverage remaining constant between the specified start and end dates. The inclusion option is relevant only if Part Areas exist in the study.

Finally, the **Subset** option allows data selection to be restricted further by geographical area. This step is optional. If a geographical subset of data is requested, an optional further screen, shown in Figure 5, is displayed. This uses a MAP object, as an interactive device to allow the user to select the regions from which data is required for analysis.

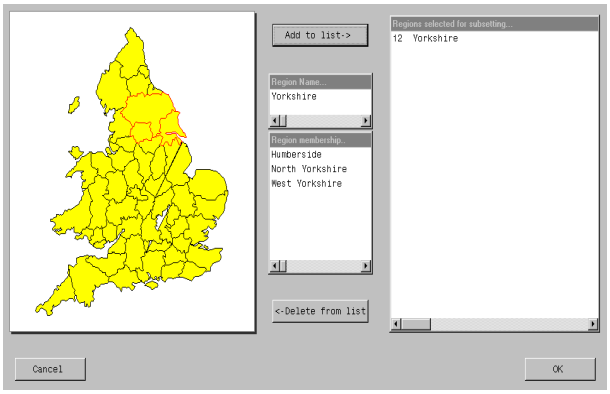


Figure 5. Geographical Subsetting Screen

The Tab Layout object inherently provides a summary of the choices made and gives the user the chance to change any of them if necessary. Once the choices are accepted, the application will progress to the analysis stage.

Age-Specific Disease Rates

The first step for many analyses is the creation of a table of age-specific disease rates. The incidence rate for a particular combination of age-group and disease group is the number of disease occurrences in the age-group divided by the total number of *person-years* accumulated by the age-group during a specified time period. A typical table of age-specific disease rates is shown in Figure 6.

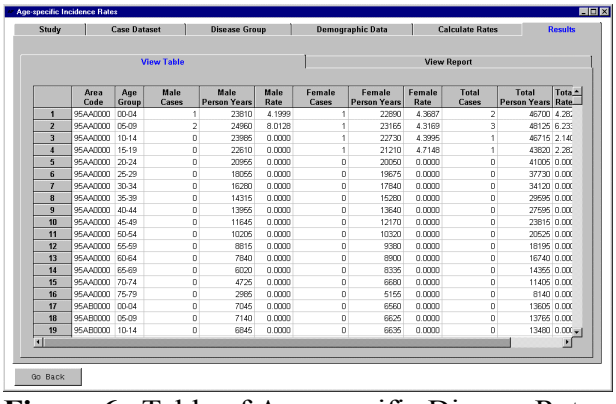


Figure 6. Table of Age-specific Disease Rates

If rates have been computed previously, a dialogue screen will advise the user of this and offer the choice of either viewing the existing values using the **Results** tab, or re-computing them. Rates may also be displayed in the form of a report more suitable for publication purposes.

Plots of Age-Specific Disease Rates

An alternative mechanism for reporting age-specific disease rates is to plot them in the form of a curve. Figure 7, below, shows a typical curve. The rates displayed are aggregated over the region of study (cf. the tables of age-specific rates broken down by geographic area). The pop-up menu allows the user to copy the plot to the clipboard for inclusion in another application, such as a word-processing package.

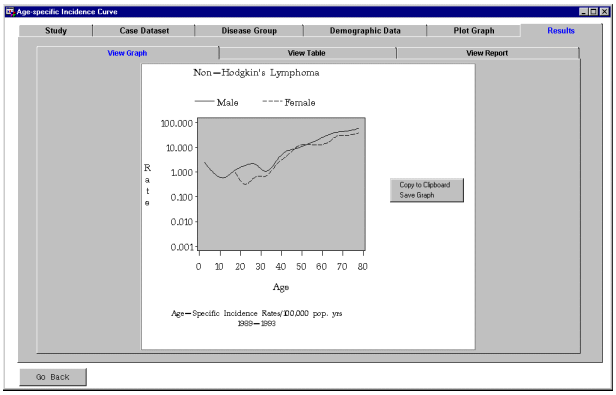


Figure 7. Age-specific Disease Rates

The rates may also be viewed in Table and Report format using the two tabs provided.

Directly Standardised Rates

Disease rates may vary with factors such as age and sex. Standardisation is used to reduce the effects on disease rates of differences between populations in the distribution of these factors.

Directly standardised rates are normally used to facilitate comparisons of disease rates between studies carried out in different countries. They are computed by multiplying the disease rates, by age and sex, of a given area by the numbers in each age and sex group of a standard reference population whose age and sex distribution is fixed. The directly standardised rate indicates the number of cases which would have occurred in the study population, if that population had the same age distribution as the standard population.

The Disease Registry Application provides a set of **Standard Populations** in common use by epidemiologists world-wide. A standard population distribution is the expected distribution of 100 individuals across a set of standard age-groups. Standard populations provided include *African*, *European*, *World*, *World Truncated* and *Uniform* populations. Details of these populations may be found in Esteve *et al.* (1994).

On selection of Directly Standardised Rates, the user is requested to specify the data required. If an age-specific rates table has previously been created, it will be located. Otherwise it will be computed. The user is allowed to specify multiple standard populations. The standardised rates produced are stored in a SAS data set and displayed in both Table and Report formats.

Indirectly Standardised Rates (SMRs)

Indirect standardisation allows disease rates for individual areas (e.g. counties) to be compared with the rate shown by some overall reference area (e.g. an entire country). For each area, the number of people in each age and sex group is multiplied by the relevant disease rate of the reference population, and the results are summed to give the number of cases expected in that area if its disease rates were the same as the reference population. The ratio of observed to expected number of cases in the area is then multiplied by 100 to give the standardised incidence ratio. This measure is referred to by various authors as either **SIR** for *standardised incidence ratio* or as **SMR** in which the letter M may represent either morbidity or mortality. The convention SMR is used in this application. Areas whose disease rates are higher than those of the reference population will have SMRs greater than 100 and those whose rates are lower than the reference population will have SMRs less than 100.

The method used to compute SMRs is that proposed by Mantel and Stark (1968) and simplified further by Breslow and Day (1975). The iterative scheme described by Breslow and Day has been implemented using SAS/IML, with Base SAS used to manipulate data sets and produce reports.

For a comparison of Direct and Indirect Standardisation see Esteve *et al*. (1984).

Figure 8 shows the Results screen with a map of the SMRs in the foreground. The map is displayed using a MAP object. To identify a particular geographic area, the user may click on the map using the mouse. The area code, area name and associated SMR value will be displayed in the boxes alongside the map.

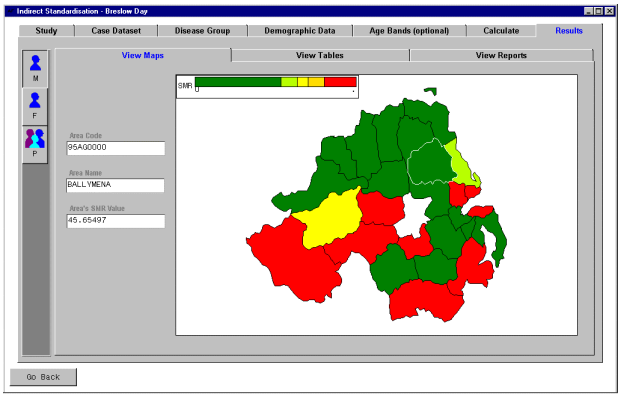
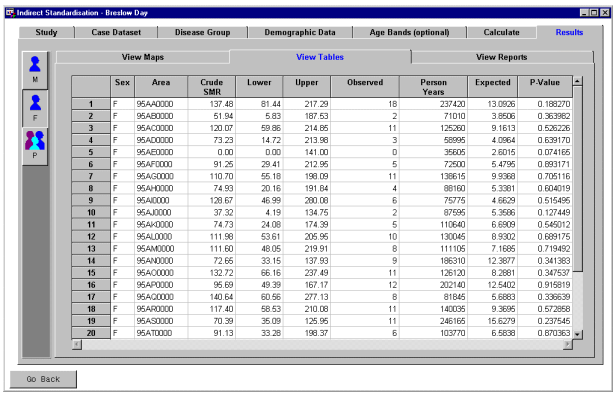


Figure 8. Map of SMRs

The numerical values of the SMRs are also available for inspection in tabular and report format using the tabs **View Table** and **View Report**. Figure 9 shows the Table view of the SMRs. A toolbar on the left-hand side of the Results screen allows for scrolling between the different sexes in all of the output displays.



	Sex	Area	Crude SMR	Lower	Upper	Observed	Person Years	Expected	P-Value
1	F	95AA0000	137.48	81.44	217.29	18	237420	13.0926	0.188270
2	F	95AB0000	51.94	5.83	187.53	2	71010	3.8506	0.363982
3	F	95AC0000	120.07	59.86	214.85	11	125360	9.1613	0.526226
4	F	95AD0000	73.23	14.72	213.98	3	58995	4.0964	0.639170
5	F	95AE0000	0.00	0.00	141.00	0	35605	2.6015	0.074165
6	F	95AF0000	91.26	29.41	212.95	5	72500	5.4795	0.893171
7	F	95AG0000	110.70	55.18	198.09	11	138615	9.9368	0.705116
8	F	95AH0000	74.93	20.16	191.84	4	98160	5.3381	0.604019
9	F	95AI0000	128.67	46.99	280.08	6	75775	4.6629	0.515495
10	F	95AJ0000	37.32	4.19	134.75	2	87595	5.3586	0.127449
11	F	95AK0000	74.73	24.08	174.39	5	110640	6.6909	0.545012
12	F	95AL0000	111.98	53.61	205.95	10	130045	8.9302	0.689175
13	F	95AM0000	111.60	48.05	219.91	8	111105	7.1685	0.719492
14	F	95AN0000	72.65	33.15	137.93	9	186310	12.3877	0.341383
15	F	95AO0000	132.72	66.16	237.49	11	126120	8.2881	0.347537
16	F	95AP0000	95.69	49.39	167.17	12	202140	12.5402	0.915819
17	F	95AQ0000	140.64	60.56	277.13	8	81845	5.6883	0.338639
18	F	95AR0000	117.40	58.53	210.08	11	140035	9.3695	0.572858
19	F	95AS0000	70.39	35.09	125.95	11	246165	15.6279	0.237545
20	F	95AT0000	91.13	33.28	198.37	6	103770	6.5838	0.870363

Figure 9. Table View of SMRs

Tests for Geographical Variation in SMRs

Poisson regression is used to compare SMRs between different geographical regions and between different age-groups. The Poisson regression model has been implemented using PROC GENMOD in SAS/STAT. Confidence intervals and tests of heterogeneity are computed using results saved in output data sets by PROC GENMOD. The scheme described by Breslow (1984), implemented using SAS/IML and PROC GENMOD, is used to adjust for extra-Poisson variation.

Mapping of SMRs

The mapping of SMR values alone can misrepresent the geographical distribution of the incidence of a disease. Apparently extreme SMRs may be based on only a few cases and give a biased picture of disease incidence. The use of p-values has been adopted in some disease atlases in an attempt to overcome this problem. However, regions of high population with SMRs deviating only slightly from the overall mean SMR may have a high level of statistical significance despite having no corresponding biological importance. An alternative approach adopted by many epidemiologists is to smooth the SMRs prior to mapping.

The Disease Registry Application supports a variety of methods for smoothing SMRs. They include:

- Empirical Bayes Estimation
- Maximum A Posteriori Estimation (MAP)
- Median Smoothing

The use of empirical Bayes estimation for smoothing SMRs was proposed by Clayton and Kaldor (1987). The approach rests upon hypothesising a prior distribution for the SMR and, assuming that the observed number of cases follows a Poisson distribution, exploits Bayes' theorem to compute *a posteriori* estimates of the SMRs.

The estimates obtained are referred to as 'shrinkage estimates', reflecting the contraction of the range of SMRs towards the overall mean exhibited by this method. Values based upon small numbers of cases shrink towards the overall mean relative risk whereas values based upon large numbers of cases remain close to the original SMR. Shrinkage estimates, therefore, provide a compromise between individual SMRs and the overall mean relative risk.

The method has been implemented in the Disease Registry Application using SAS/IML. Following Clayton and Kaldor, a choice of gamma and log-normal prior distributions is offered. The effect of using a gamma prior is illustrated in Figure 10.

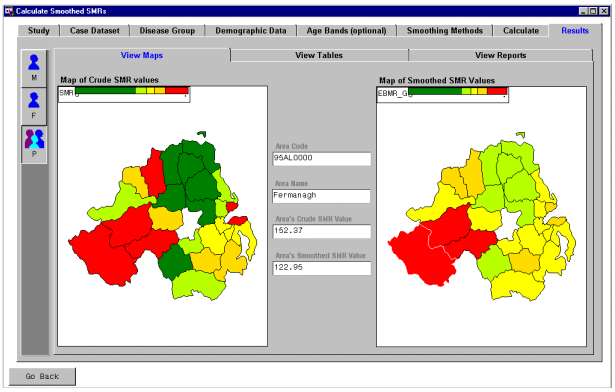


Figure 10. SMRs and 'Shrinkage' estimates

The shrinkage estimate is based upon a model that assumes independence in the observed SMRs. However, it may be more realistic to expect neighbouring areas to have similar rates, particularly if the areas under study are very small. It then becomes more meaningful to take into account possible correlations between SMRs in neighbouring areas.

A *spatial* model, also based upon empirical Bayes estimation, in which smoothing is restricted to adjacent areas was proposed by Besag (1974). This model, referred to as the Maximum A-Posteriori (MAP) model, has been implemented in the Disease Registry Application using an external procedure written in C and called using the SCL MODULE function.

MAP estimates provide a compromise between the individual SMR and the local mean SMR from adjacent areas. The effect of using a MAP estimate is illustrated in Figure 11.

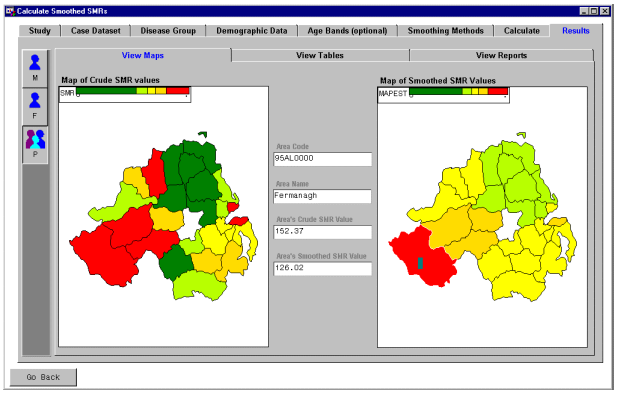


Figure 11. SMRs and MAP estimates

The third smoothing method offered by the Disease Registry Application uses a technique referred to as *Head-Banging*. This method, described by Hansen (1991) and based upon an original idea by J.Tukey, uses robust techniques which eliminate individual outliers whilst preserving important features such as discontinuities between whole regions.

This technique has been favoured by Jones, *et al*, (1996), in the production of the new US Mortality Atlas, who observe that it overcomes the tendency of the empirical Bayes approach towards 'over shrinkage' in the estimates of relative risk.

The Head-Banging method has been implemented using a procedure written in C, called via the MODULE function.

Tests for Spatial Clustering

A number of tests for spatial clustering are being developed. Currently, only the Smans test has been implemented. This tests for a tendency of areas with similar disease risks to be adjacent. Other tests being developed include the Potthoff - Whittinghill test and Knox's test.

Future Plans

A number of enhancements are planned. In the area of analysis, these include the incorporation of further tests for spatial clustering, tests for seasonality of disease occurrence and facilities for post-hoc cluster analysis. For the latter, an interface is being developed using the SAS/GIS object, experimental in SAS 6.12, to facilitate the selection of geographical areas for study.

General application development will include the addition of a facility to allow analyses to be deferred for batch processing and the development of facilities for data administration.

Summary

The application has been in use by the Leukaemia Research Fund Centre for Clinical Epidemiology at Leeds since March 1998. The users have reported a significant improvement in productivity using the new system. The application will shortly be used to produce a revised atlas describing the leukaemias and lymphomas in Northern Ireland. More recently the application has also been used successfully by the Department of Paediatric Epidemiology at Leeds in a study of childhood diabetes in North Yorkshire.

An important practical aspect of the application, cited by the users, is that it provides a *protocol* for conducting descriptive epidemiological studies. Using the old system, the LRF encountered a number of inconsistencies generated either by different people using different conventions or as a result of the multitude of programs and file formats used. By using procedures which have been thoroughly checked and by retaining tight application control, the user is prevented from performing non-standard, or, worse, error-prone operations. This provides consistency both within a particular research group and between different research groups.

The SAS system meets the majority of the user requirements directly. The adoption of an appropriate data model addressed the limitations affecting the previous system and the integration of applications has eliminated the need for multiple software systems and the inter-communication of results via external files. The use of the MODULE function has enabled computationally intensive procedures, developed using external codes written in C, to be integrated into the application.

References

1. Allon, D. and Nicholson, P. (1997). 'Data Modelling for an Epidemiological Database'. SEUGI 15 Conference Proceedings.
2. Allon, D. (1999) Guiding the User: Using the Tab Layout Object in a SAS/AF Application. SEUGI 17 Conference Proceedings.
3. Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion). Journal of the Royal Statistical Society Series B, **36**, 192-236.
4. Breslow, N.E. (1984) Extra-Poisson variation in linear models. Appl. Statist., **33**, 38-44.
5. Breslow, N.E. and Day, N.E. (1975) Indirect standardisation and the multiplicative model for rates with reference to the age adjustment of cancer incidence and relative frequency data. J. Chron. Dis., **28**, 289-303.
6. Clayton, D.C. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. Biometrics; **43**, 617-681.
7. Esteve J, Benhamou, E and Raymond, L. Statistical Methods in Cancer Research : (Vol. IV) : Descriptive Epidemiology. (IARC Scientific Publications, No, 128), Lyon, IARC, 1994.
8. Hansen, K.M. (1991) Head-Banging: Robust smoothing in the plane. IEEE Transactions on Geoscience and Remote Sensing, Vol. **29**, No 3.
9. Jones, Gretchen K., Linda W. Pickle and Michael Mungiole, (1996). The Use of the SAS Programming Language and Procedures to Model and Process Data Used in Creating the New U.S. Mortality Atlas. SUGI 96 Conference Proceedings.
10. Mantel, N. and Stark, Charles R., (1968) Computation of Indirect-Adjusted Rates in the Presence of Confounding. Biometrics, **24**, 997-1005.
11. Nicholson, P., Allon, D., McNally, R and Rowland, D. (1997). 'Analysing the Incidence of Leukaemia in England and Wales'. SEUGI 15 Conference Proceedings.
12. SAS Institute Inc., SAS/IML® Software: Usage and Reference, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1989. 501 pp.
13. SAS Institute Inc., SAS/STAT® Software: Changes and Enhancements through Release 6.11, Cary, NC: SAS Institute Inc., 1996. 1104 pp.

Acknowledgements

We would like to thank SAS Institute for providing both financial and technical support for this project and the Leukaemia Research Fund for providing financial support for the Data Collection Study. This work is based on data provided with the support of the ESRC. Thanks are due to the Office of National Statistics for the provision of population estimates, which remain the property of the Crown, and also to the staff of UKBORDERS, Edinburgh, for the supply of the 1991 Census digitised boundary data sets for England and Wales.

Addresses for Correspondence

The author may be contacted at the following addresses:

Information Systems Services
University of Leeds
Leeds, LS2 9JT, UK
E-mail: P.Nicholson@leeds.ac.uk
Tel: +44 113 233 5405
Fax: +44 113 233 5411

A project web page is located at:
http://www.leeds.ac.uk/iss/projects/sas/epi.html.